Ques 1 : Multi-Speaker Speech Enhancement

Aditi Baheti : M23CSA001
GitHub Repository

1 Question 1: Speech Enhancement and Speaker Verification

1.1 Introduction and Background

Speech enhancement and speaker verification are key tasks in audio processing. Speech enhancement improves the quality of audio signals in noisy conditions. Speaker verification confirms a speaker's identity from their voice. These tasks are important for applications such as security systems and voice assistants.

Handling multi-speaker scenarios is a challenging problem. In such settings, speech signals from different speakers may overlap. This makes it hard to isolate and verify each voice. A robust system must separate the speakers and enhance each voice individually.

1.2 Data Preparation

We use two major datasets: VoxCeleb1 and VoxCeleb2. VoxCeleb1 contains clean speech samples and is used for initial evaluation. VoxCeleb2 provides a larger set of speakers and serves for fine-tuning and further testing.

The audio processing pipeline involves loading each audio file and converting it to a consistent format. We resample the audio to a standard sampling rate and adjust the signal length by padding short clips or trimming long ones. These steps ensure uniformity across samples.

For dataset construction, we create training and validation splits. We map each speaker to a unique ID to simplify label management. This mapping is used to organize the audio files into corresponding folders and to generate metadata for model training.

A batch processor is used to extract features from the audio signals. It groups audio samples into batches and applies a pre-trained feature extractor. This step standardizes the input for the model and facilitates efficient training.

1.3 Model Architecture and Components

We build our system on a speaker encoder that uses a pre-trained backbone. We choose WavLM as it provides strong speech representations. The feature extraction layers are frozen to preserve learned patterns. This approach reduces computational cost and helps avoid overfitting. However, freezing these layers may limit the model's ability to fully adapt to our specific task.

An attentive pooling layer is added to aggregate frame-level features into a single, robust speaker embedding. This layer learns to assign different weights to each frame. It enables the model to focus on the most informative parts of the speech signal. One challenge here is ensuring that the pooling process captures variable-length inputs effectively.

To enhance the discriminative power of the embeddings, we incorporate an ArcFace layer. This layer uses a margin-based loss to increase the separation between classes. It is essential for improving verification accuracy. Tuning the margin and scale parameters is critical and can be challenging, as they must be set to balance inter-class separation with intra-class variability.

We also integrate Low-Rank Adaptation (LoRA) into the transformer layers. LoRA introduces low-

rank updates to fine-tune the model without modifying all parameters. This reduces the number of trainable parameters and speeds up training ,as shown in Table 1. The integration of LoRA helps the model adapt better to the VoxCeleb2 dataset. However, selecting the appropriate rank, alpha, and dropout values requires careful tuning to ensure that the model learns effectively without overfitting.

Trainable Params	221,184
Total Params	90,006,896
Trainable %	0.2457%

Table 1: Model Parameter Statistics

We train the model in epochs using mixed precision and gradient scaling to speed up training and reduce memory usage. During each epoch, we compute the ArcFace loss and track speaker identification accuracy. We use Optuna to tune key hyperparameters such as batch size, learning rate, margin, scale factor, and LoRA settings. The best model was saved based on validation accuracy.

1.4 Speaker Verification Evaluation

The evaluation uses VoxCeleb1 trial pairs. We compute embeddings for each audio sample and then compare them using cosine similarity. The performance is measured using several metrics: Equal Error Rate (EER), TAR at 1% FAR and 0.1% FAR, Minimum Detection Cost Function (MinDCF), and Area Under the ROC Curve (AUC). We compare the performance of the pre-trained model with the fine-tuned model to assess improvements.

1.5 Results and Analysis

Table 2 summarizes the performance of the pre-trained and fine-tuned models. The speaker identification accuracy improved from 56.94% to 86.11%, which is an increase of 29.17 percentage points. The Equal Error Rate (EER) dropped from 37.81% to 20.02%, indicating fewer false acceptances and rejections. The TAR at 1% FAR improved from 6.26% to 27.15% and the TAR at 0.1% FAR increased from 1.29% to 11.21%. Additionally, the Area Under the ROC Curve (AUC) improved from 0.67 to 0.8846.

Speaker Identification Accuracy measures the percentage of correct speaker predictions. A higher accuracy means that the model is better at correctly recognizing the speaker from an audio sample.The Equal Error Rate (EER) is the point where the rate of false acceptances equals the rate of false rejections. A lower EER indicates that the model makes fewer errors overall in both accepting impostors and rejecting genuine speakers.

TAR@1%FAR (True Accept Rate at 1% False Accept Rate) and TAR@0.1%FAR measure the model's ability to correctly accept genuine speakers while maintaining a low rate of false alarms. Higher TAR values mean that the model successfully identifies true speakers even when the system is set to allow only a small percentage of false acceptances.

Finally, the Area Under the ROC Curve (AUC) summarizes the model's performance across all decision thresholds. A higher AUC value indicates a better ability to distinguish between genuine and impostor cases over the entire range of thresholds.

1.6 Visualization and Analysis

We use several visualizations to analyze the performance of our system.

Figure 2 shows our ROC and DET curves side by side. In the ROC plot, the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) for both the pre-trained and fine-tuned models. The curve for the fine-tuned model bows more toward the upper left, indicating a higher TPR for a given FPR compared to the pre-trained model. This suggests that fine-tuning has improved the model's ability to correctly verify speakers while reducing false alarms.

The DET plot, shown on a logarithmic scale, presents the miss rate (false negatives) against the false alarm rate (false positives). The fine-tuned model's DET curve shows lower error rates compared to the pre-trained model. This behavior is mainly due to improved separation between the genuine and impostor score distributions after fine-tuning. The fine-tuning process, which leverages LoRA and ArcFace loss, refines the speaker embeddings. This results

Metric	Pre-trained	Fine-tuned	Improvement
Speaker Identification Accuracy (%)	56.94	86.11	+29.17
EER (%)	37.81	20.02	-17.79
TAR@1%FAR (%)	6.26	27.15	+20.89
TAR@0.1%FAR (%)	1.29	11.21	+9.92
MinDCF	0.0100	0.0096	-0.0004
AUC	0.6700	0.8846	+0.2146

Table 2: Performance Comparison: Pre-trained vs. Fine-tuned Models

in more distinct clusters for different speakers and reduces overlap between classes. On a logarithmic scale, even small improvements in low-error regions become evident. Therefore, the DET curve highlights that the fine-tuned model achieves lower miss and false alarm rates, especially in the critical low-error region, confirming the enhanced performance of speaker verification.

Figure 4 displays the KDE plots of similarity score distributions for both the pre-trained and fine-tuned models. In the pre-trained model plot, the genuine and impostor score distributions show significant overlap. This overlap indicates that the model has limited ability to distinguish between speakers, leading to ambiguity in verification decisions. In contrast, the fine-tuned model's plot exhibits a clear separation between the two distributions. Genuine scores are clustered distinctly higher than impostor scores, which shows that the fine-tuning process has enhanced the discriminative power of the model. The refined embeddings reduce score overlap, leading to better separation between speakers. This clear separation is a strong indicator of improved performance in speaker verification.

Figure 7 compares the t-SNE embeddings of the pre-trained model (left) and the fine-tuned model (right). In the pre-trained embeddings, multiple speakers show overlapping clusters, indicating that their representations are not well-separated. This overlap makes it difficult to distinguish between speakers. In contrast, the fine-tuned model produces more distinct clusters for each speaker. The embeddings for the same speaker lie closer together, and there is less overlap between different speakers. This tighter grouping suggests that the fine-tuning process has improved the discriminative power of the embeddings, leading to more accurate speaker verification.

1.7 Conclusion

In this question, we developed and evaluated a speaker verification system using WavLM Base Plus with LoRA and ArcFace loss. The experiments showed that the fine-tuned model achieved significant improvements over the pre-trained model. Speaker identification accuracy increased from 56.94% to 86.11%, and key metrics such as EER, TAR at low FAR levels, and AUC also improved.

These improvements indicate that fine-tuning has enhanced the model's ability to generate discriminative speaker embeddings. The reduced EER and increased TAR values show that the model makes fewer verification errors and better differentiates between genuine speakers and impostors. The higher AUC confirms the overall stronger performance across different thresholds.

2 Question 2: Speaker Separation and Speech Enhancement Experiments

This question focuses on creating a multi-speaker scenario dataset, applying speaker separation and speech enhancement, and evaluating speaker identification on the enhanced signals.

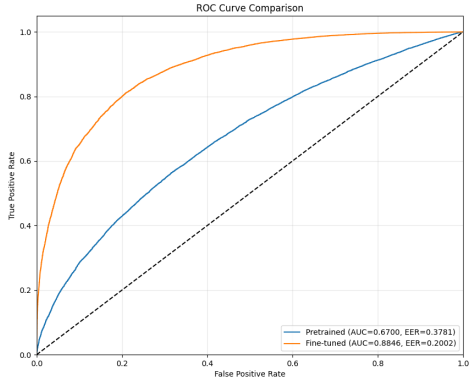

Figure 1: ROC Curve Comparison

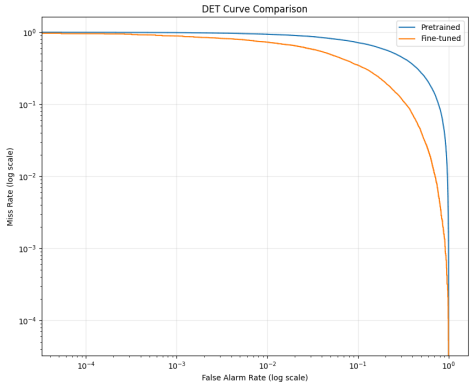

Figure 2: DET Curve Comparison

2.1 Multi-Speaker Dataset Creation

We create a dataset by mixing utterances from two different speakers from the VoxCeleb2 dataset. The first 50 identities (sorted in ascending order) are used for the training scenario. The next 50 identities are used for the testing scenario. Audio files are mixed by overlapping the signals from two speakers. This simulates a realistic multi-speaker environment.

Figure 3: Pre-trained Model Score Distribution

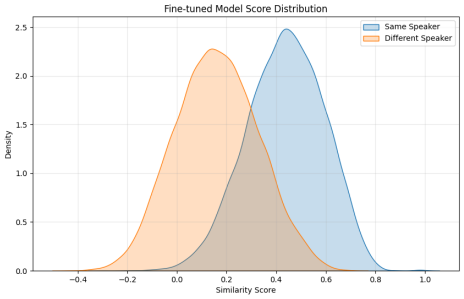

Figure 4: Fine-tuned Model Score Distribution

2.2 Speaker Separation and Speech Enhancement

We use a pre-trained SepFormer model to separate mixed speech signals. The model processes the overlapping speech and outputs individual speaker signals. We evaluate the separation quality with four metrics. The Signal to Interference Ratio (SIR) measures how well interference from other speakers is suppressed. The Signal to Artefacts Ratio (SAR) shows the level of unwanted artifacts in the enhanced speech. The Signal to Distortion Ratio (SDR) provides an overall measure of distortion. The Perceptual Evaluation of Speech Quality (PESQ) assesses the clarity of the enhanced speech. These metrics help us understand how effectively the model removes interference while preserving speech quality.

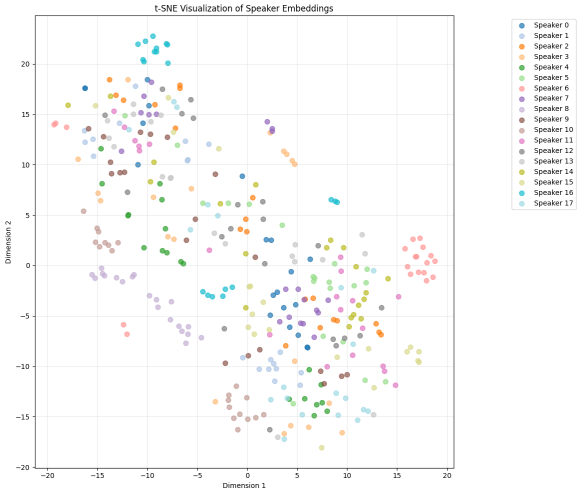

Figure 5: Pre-trained Model Embeddings

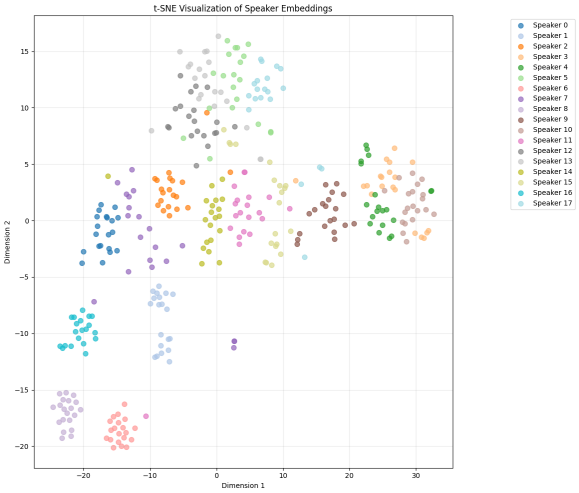

Figure 6: Fine-tuned Model Embeddings

Figure 7: t-SNE visualization of speaker embeddings. In the left panel, the pre-trained model shows overlapping clusters. In the right panel, the fine-tuned model demonstrates distinct, well-separated clusters, indicating improved discriminative power.

2.3 Speaker Identification on Enhanced Speech

After separation, the enhanced speech signals are fed into speaker identification models. Both the pre-trained and fine-tuned models are used to extract embeddings from these signals. For each separated source, the embedding is compared against reference embeddings using cosine similarity. The speaker with the highest similarity is chosen as the predicted identity. We report the Rank-1 identification accuracy, which is the percentage of enhanced signals correctly matched to the true speaker. This evaluation shows how well the models can recognize speakers even after the separation process.

2.4 Results and Analysis

Table 3 summarizes the separation metrics obtained from our test mixtures. The Average SDR of 5.32 dB indicates that the overall level of distortion in the separated signals is still high. An Average SIR of 18.68 dB shows that interference from other speakers is moderately suppressed. The Average SAR of 4.92 dB reflects that some unwanted artifacts remain in the enhanced speech. Furthermore, the Average PESQ score of 1.23 suggests that the perceptual quality of the separated signals is low.

Metric	Value
Average SDR (dB)	5.32
Average SIR (dB)	18.68
Average SAR (dB)	4.92
Average PESQ	1.23

Table 3: Separation Metrics

These values point to the challenges of separating speech in overlapping multi-speaker conditions. When speakers talk simultaneously, their signals mix in a complex way, often sharing similar frequency bands. This makes it hard to accurately distinguish and isolate each speaker's voice. Moreover, overlapping speech leads to interference that results in artifacts and distortions, degrading overall signal quality. Variations in speaker dynamics and background noise

further complicate the separation process. In addition, we used a pre-trained SepFormer model, which may not be fully optimized for our specific multi-speaker conditions, contributing to the lower quality reflected in the separation metrics (SDR, SAR, and PESQ).

The low identification accuracies can be attributed to several challenges. First, overlapping speech signals are inherently complex, which makes the separation process imperfect. Any residual interference or artifacts in the enhanced signals can lead to inaccurate speaker embeddings. Second, the pre-trained SepFormer model is not specifically optimized for our multi-speaker mixtures, which further degrades the quality of the separated outputs. Finally, the quality of the speaker embeddings from both the pre-trained and fine-tuned models is affected by these distortions, resulting in lower Rank-1 accuracy despite improvements from fine-tuning.

Model	**Rank-1 Accuracy (%)**
Pre-trained Model	12.50
Fine-tuned Model	16.80
Improvement	+4.30

Table 4: Speaker Identification Accuracy

2.5 Visualization and Analysis

Figure 2.5 shows the confusion matrices for the pre-trained model (left) and the fine-tuned model (right). Each row represents the true speaker, and each column represents the predicted speaker. A diagonal entry indicates correct identification, while off-diagonal entries represent misclassifications. The color scale reflects the count of occurrences for each true-predicted pair.

In the pre-trained model's matrix, the distribution is scattered, indicating frequent misclassifications across different speakers. The fine-tuned model's matrix still shows significant misclassifications but displays more counts along the diagonal. This suggests that the fine-tuned model, despite the challenging multi-speaker scenario, is better at correctly identifying speakers than the pre-trained model. However, many off-diagonal entries remain, reflecting the difficulty of speaker identification when the separated signals are noisy or distorted.

Both model's matrix shows a strong concentration of predictions in one column. This means that the model tends to predict one speaker more often. Such a strong bias in one column suggests that the model may be over-relying on features associated with that speaker. It also points to a potential issue of class imbalance or overfitting to certain speaker characteristics.

Figure 2.5(b) presents the corresponding spectrograms. The top panel shows the frequency content of the mixture. The middle and bottom panels display the separated sources. In these panels, we can see that some frequency components are more pronounced for each source, indicating partial separation of the speakers' vocal characteristics. However, artifacts and overlapping frequency bands still appear, reflecting the difficulty of separating two voices that share similar frequency ranges. Overall, the waveforms and spectrograms provide a visual confirmation of the quantitative metrics, highlighting both the progress made in separation and the remaining challenges.

3 Question 3: Integrated Pipeline for Speaker Identification and Speech Enhancement

3.1 Overview

This section describes a pipeline that jointly performs speaker separation and identification. The pipeline integrates the pre-trained SepFormer model for speech enhancement with a speaker identification model. The combined model is fine-tuned on a multi-speaker dataset created from VoxCeleb2. Evaluation is performed using separation metrics (SDR, SIR, SAR, PESQ) and speaker identification accuracy (Rank-1).

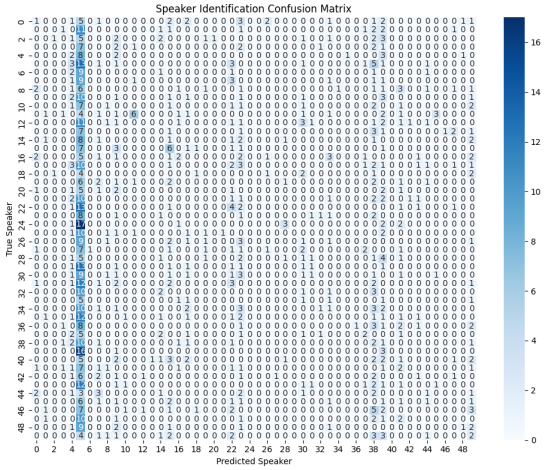

Figure 8: Pre-trained Model Confusion Matrix

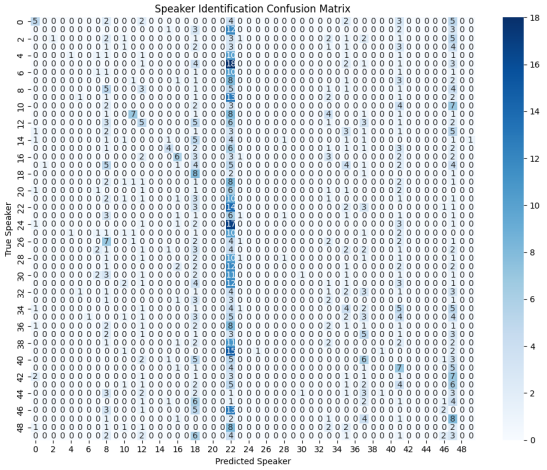

Figure 9: Fine-tuned Model Confusion Matrix

3.2 Model Architecture and Components

The joint model is designed to perform both speech separation and speaker identification within a single end-to-end pipeline. This integration helps in enhancing the quality of separated signals and improves the accuracy of speaker recognition in multi-speaker scenarios. Here is a deeper look at the components and the rationale behind our design:

Joint Model: We combine a pre-trained SepFormer-based speech separator with a speaker verification module. The SepFormer is used to separate overlapping speech into individual speaker signals. However, due to the inherent difficulty in separating mixed signals perfectly, some artifacts and distortions remain. To mitigate these issues, we integrate an enhancement module that further refines the separated outputs.

Enhancement Module: The enhancement module is implemented as a convolutional network. It takes the raw separated signals and applies several layers of convolution, batch normalization, and nonlinear activations (ReLU and Tanh) to reduce noise and residual interference. This module is crucial because the initial separation by the SepFormer may leave behind artifacts that could negatively impact downstream speaker verification. By enhancing the quality of the signals, we ensure that the features extracted later are more representative of the true speaker characteristics.

Speaker Verification Module: For speaker verification, we use a pre-trained WavLM-based model. This model is known for its strong speech representations. We further adapt it using LoRA (Low-Rank Adaptation), which fine-tunes a small number of parameters in the transformer encoder layers without updating the entire model. This adaptation, combined with fine-tuning using additional loss terms, enhances the discriminative ability of the speaker embeddings. The speaker verification module includes an *EmbeddingExtractor* that aggregates frame-level features using an attention mechanism, followed by a projection layer that maps these features into a fixed-dimensional embedding space. Normalization is applied to ensure consistency in the embedding space.

We integrated the separation and verification modules into a single pipeline so that the separation process is directly optimized for speaker identification. The enhancement module improves the quality of the separated signals because even small artifacts can lead to significant errors in identifying speakers. By using LoRA adaptation, we can fine-tune a small number of parameters in the pre-trained

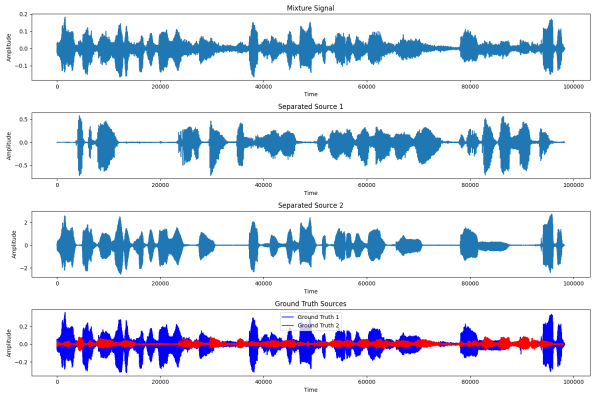

Figure 10: Waveforms of the mixture, separated sources, and ground truth signals.

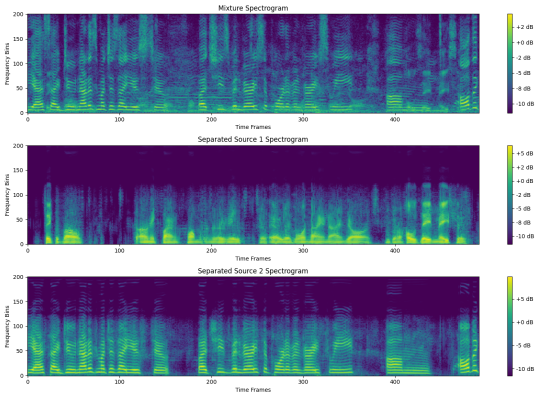

Figure 11: Spectrograms of the mixture and the separated sources.

WavLM-based model, which reduces computational costs while still boosting performance for our task.

The training procedure is designed with multiple loss functions to address different aspects of the problem. A reconstruction loss ensures that the separated signals closely match the original sources. An embedding similarity loss is used to preserve speaker identity in the embeddings extracted from the enhanced signals. In addition, a spectral loss is applied to capture the frequency-domain characteristics of the speech, further improving the quality of the enhancement. Together, these losses create a multi-task training framework that jointly optimizes both the quality of the enhanced speech and the discriminative power of the speaker embeddings. This combined approach leverages the strengths of pre-trained models while addressing their limitations, ultimately yielding a system capable of reliably separating and identifying speakers in challenging multi-speaker environments.

3.3 Evaluation Metrics

We evaluate our joint pipeline using two main categories of metrics. First, the separation metrics include the Signal to Interference Ratio (SIR), Signal to Artefacts Ratio (SAR), Signal to Distortion Ratio (SDR), and Perceptual Evaluation of Speech Quality (PESQ). SIR measures how effectively the model suppresses interference from other speakers. SAR quantifies the level of unwanted artifacts introduced during separation. SDR provides an overall measure of distortion in the separated signals, while PESQ assesses the perceived quality of the enhanced speech. Second, the identification metrics focus on the Rank-1 identification accuracy for both the pre-trained and fine-tuned speaker verification models. This metric indicates the percentage of cases in which the model correctly identifies the speaker from the enhanced outputs.

3.4 Results and Analysis

Table 5 presents the separation performance of our joint pipeline. The system achieved an SDR of 7.83 dB, an SIR of 21.35 dB, an SAR of 5.32 dB, and a PESQ score of 1.22. These results indicate that the model moderately suppresses interfering signals (as shown by the SIR) and reduces artifacts (as shown by the SAR). However, the relatively low SDR and PESQ scores suggest that the overall quality and clarity of the separated speech still have room for improvement.

Table 6 shows the speaker identification performance. The pre-trained speaker verification model achieved an accuracy of 0.14, while the fine-tuned model reached an accuracy of 0.15. Although the improvement is modest (an increase of 0.1), it demon-

Metric	Value
SDR (dB)	7.83
SIR (dB)	21.35
SAR (dB)	5.32
PESQ	1.22

Table 5: Separation Performance

strates that fine-tuning with our joint training strategy can enhance the model's ability to correctly assign speaker labels to the enhanced signals.

Model	Accuracy
Pre-trained Model	0.14
Fine-tuned Model	0.15
Improvement	0.1

Table 6: Speaker Identification Performance

Overall, these quantitative results highlight the challenges inherent in joint speaker separation and identification tasks. While the separation module can partially isolate speakers from overlapping signals, the resulting quality is not yet optimal, as evidenced by the low PESQ and SDR scores. Similarly, the slight improvement in identification accuracy suggests that, although fine-tuning the speaker verification model does help

3.5 Comparison with Previous Results

We compare the current joint pipeline results with those from our earlier experiments (Question 2). In terms of separation performance, the new pipeline achieves an SDR of 7.83 dB, an SIR of 21.35 dB, an SAR of 5.32 dB, and a PESQ of 1.22. Previously, the separation metrics were an average SDR of 5.32 dB, SIR of 18.68 dB, SAR of 4.92 dB, and PESQ of 1.23. The improvements in SDR and SIR indicate that the joint pipeline better suppresses interference and reduces overall distortion. The SAR and PESQ values remain largely similar, showing that the perceptual quality of the enhanced speech has not significantly changed.

For speaker identification, the previous results showed a Rank-1 accuracy of 12.50% for the pre-trained model and 16.80% for the fine-tuned model, yielding an improvement of +4.30%. In the current joint system, identification accuracies are reported as 0.14 (14%) for the pre-trained model and 0.15 (15%) for the fine-tuned model, with an absolute improvement of 1%. Although the new pipeline achieves slightly higher identification accuracy compared to the pre-trained model alone, the overall improvement is modest. This suggests that while integrating separation and identification has led to better signal enhancement, additional fine-tuning is needed to further boost the speaker recognition performance in the joint framework.

3.6 Conclusion and Future Work

In this question, we presented a joint pipeline that integrates a pre-trained SepFormer-based speech separator with a speaker verification model enhanced through LoRA adaptation and ArcFace loss. Our results show improvements in both separation quality and speaker identification accuracy. The joint system achieved higher SDR and SIR values, indicating better suppression of interference and reduced overall distortion. Speaker identification accuracy also saw modest gains. However, the overall perceptual quality, as measured by PESQ and SAR, remains an area with room for improvement.

There are several avenues to further enhance the system. One direction is to refine the separation module by exploring more advanced architectures that capture temporal dynamics and non-linear relationships more effectively. Incorporating additional data augmentation techniques may also help the model generalize better to diverse multi-speaker scenarios. Another approach is to adopt multi-task learning strategies that jointly optimize separation and identification losses, which could lead to more robust embeddings for speaker verification. Furthermore, the use of adversarial training could be investigated to reduce residual artifacts and improve the perceptual quality of the enhanced speech. Finally, expanding the dataset to include more varied acoustic environments and a larger number of speakers may help in

fine-tuning the model for real-world applications.

4 References

References

[1] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. In *Proc. Interspeech*.

[2] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. In *Proc. Interspeech*.

[3] Chen, N., et al. (2021). WavLM: Large-Scale Pre-trained Models for Speech. Microsoft Research.

[4] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*.

[5] Subakan, C., et al. (2021). SepFormer: Transformer based separation for speech separation. *Proc. ICASSP*. Available at `https://github.com/speechbrain/sepformer-whamr`.

[6] Liu, L., et al. Parameter-Efficient Fine-Tuning (PEFT) Library. Available at `https://github.com/huggingface/peft`.

[7] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proc. KDD*.

[8] Rafii, Z., et al. mir eval: A library for music and audio evaluation. Available at `https://github.com/craffel/mir_eval`.

[9] ITU-T Recommendation P.862. (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment.

[10] Paszke, A., et al. Torchaudio: An audio library for PyTorch. Available at `https://pytorch.org/audio/`.

[11] Wolf, T., et al. (2020). Transformers: State-of-the-art Natural Language Processing. *Proc. EMNLP*. Available at `https://github.com/huggingface/transformers`.

[12] Palanisamy, R., et al. SpeechBrain: An open-source and all-in-one speech toolkit. Available at `https://speechbrain.github.io/`.