Multi-Speaker Speech Enhancement and Identification using Deep Learning

Shreya Pandey (M24CSA030)

1. Introduction

1.1 Background:

Speech enhancement aims to improve the quality and intelligibility of speech signals, particularly in noisy or reverberant environments. In multi-speaker scenarios, the challenge is compounded by the presence of interfering speech signals from other speakers. Simultaneously, speaker identification seeks to determine the identity of a speaker from their voice. Both speech enhancement and speaker identification are critical tasks with significant implications for various applications, including teleconferencing systems where clear communication is essential, human-computer interaction for robust voice commands and personalized experiences, hearing aids to isolate target speakers, and forensic analysis for voice evidence. Recent advancements in deep learning have revolutionized these fields, leading to significant improvements in the performance and robustness of speech enhancement and speaker recognition systems.

1.2 Problem Statement:

This study aims to:

- Evaluate the performance of several pre-trained deep learning models for speaker verification on a standard benchmark dataset.
- Fine-tune the best-performing pre-trained speaker verification model using parameter-efficient techniques and a discriminative loss function to further enhance its performance.
- Develop a multi-speaker speech enhancement pipeline utilizing a pre-trained speaker separation model to isolate individual speakers from mixed audio.
- Integrate the speaker identification capability with the speech enhancement pipeline to associate the separated and enhanced speech segments with their respective speakers.
- Propose and evaluate a novel combined pipeline that aims to perform speaker-aware speech separation and enhancement, potentially leading to improved performance in both tasks.

2. Literature Review

2.1 Speech Enhancement Techniques:

Traditional speech enhancement methods often rely on signal processing techniques such as spectral subtraction, Wiener filtering, and statistical model-based approaches. However, these methods can struggle in complex, non-stationary noise environments and with interfering speech. Deep learning has emerged as a powerful paradigm for speech enhancement, with architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks

(RNNs), and Transformers demonstrating remarkable capabilities. Sequence-to-sequence models, particularly those based on the encoder-decoder framework with attention mechanisms, have shown great promise for speech separation and enhancement. The SepFormer architecture, which utilizes a dual-path recurrent neural network with intra- and inter-chunk processing, has achieved state-of-the-art performance in separating overlapping speech signals.

2.2 Speaker Verification and Identification:

Speaker verification involves confirming whether a given speech utterance belongs to a claimed speaker, while speaker identification aims to determine which speaker from a set of known speakers produced the utterance. Deep learning-based approaches for speaker recognition typically involve extracting speaker embeddings, low-dimensional vector representations that capture the unique characteristics of a speaker's voice. Architectures like x-vectors, based on Time Delay Neural Networks (TDNNs), have been widely successful. This study utilizes several pre-trained speaker verification models: 'hubert large', 'wav2vec2 xlsr', 'unispeech sat', and 'wavlm base plus'. These models are based on Transformer and convolutional architectures and have been pre-trained on large amounts of speech data to learn robust speech representations. The ArcFace loss function is a large margin cosine loss designed to increase the inter-class variance and decrease the intra-class variance of the learned embeddings, leading to more discriminative speaker representations. Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique that freezes the pre-trained model weights and introduces a small number of trainable rank-decomposition matrices into each layer of the Transformer architecture, allowing for efficient adaptation to downstream tasks.

2.3 Multi-Speaker Speech Datasets:

The VoxCeleb1 and VoxCeleb2 datasets are large-scale audio-visual datasets containing speech utterances from thousands of speakers extracted from YouTube videos. VoxCeleb1 is widely used for speaker verification evaluation, while VoxCeleb2 is larger and often used for training speaker recognition models. Both datasets provide a diverse range of speakers, accents, and recording conditions, making them suitable for training and evaluating models for speaker recognition and multi-speaker scenarios. The datasets include metadata about the speakers and the video sources.

3. Methodology

3.1 Experimental Setup:

3.1.1 Datasets:

- **VoxCeleb1:** The VoxCeleb1 dataset (simulated download from provided link) will be used for evaluating the pre-trained and fine-tuned speaker verification models. The 'cleaned' version of the trial pairs will be used for evaluation. The audio files (simulated download from provided link in the 'vox1' folder) are in .m4a format.
- **VoxCeleb2:** The VoxCeleb2 dataset (simulated download from provided link in the 'vox2' folder containing .txt metadata and .m4a audio files) will be used for fine-tuning

the selected speaker verification model. The first 100 identities (when sorted in ascending order based on their unique identifiers) will be used for the training split, and the remaining 18 identities will form the testing split for the speaker verification fine-tuning task.

- **Multi-Speaker Dataset:** A multi-speaker training dataset is created by mixing utterances from the first 50 identities of the VoxCeleb2 dataset (sorted in ascending order). A separate multi-speaker testing dataset is created by mixing utterances from the next 50 identities of the VoxCeleb2 dataset (sorted in ascending order), ensuring no overlap in speaker identities between the training and testing sets. The mixing process (simulated based on the referenced GitHub repository) will involve randomly selecting two distinct utterances from different speakers within the chosen identity sets and overlapping them. The signal-to-signal ratio (SSR) of the mixtures will be varied (e.g., uniformly sampled between -5dB and 5dB). Utterances will be chosen with considerations for duration to ensure reasonable overlap lengths.

3.1.2 Pre-trained Models:

- The following pre-trained speaker verification models (simulated download from provided link, e.g., Hugging Face Transformers) will be used: 'wav2vec2 xlsr'.
- A pre-trained SepFormer model (simulated download, e.g., from a speech processing toolkit like SpeechBrain or a dedicated repository) will be used for speaker separation and speech enhancement.

3.1.3 Evaluation Metrics:

- **Speaker Verification:**
    - **Equal Error Rate (EER):** The rate at which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. Lower EER indicates better performance.
    - **Target Acceptance Rate at 1% False Acceptance Rate (TAR@1%FAR):** The percentage of genuine speaker pairs that are correctly accepted when the false acceptance rate is constrained to 1%. Higher TAR@1%FAR indicates better performance at a specific security threshold.
    - **Speaker Identification Accuracy:** The percentage of test utterances for which the correct speaker is identified as the top match from a set of enrolled speakers. Higher accuracy indicates better performance.
- **Speech Enhancement:**
    - **Signal to Interference Ratio (SIR):** Measures the level of the target speaker's speech compared to the interfering speaker's speech in the separated output. Higher SIR indicates better separation.

    - **Signal to Artefacts Ratio (SAR):** Measures the level of the target speaker's speech compared to the artifacts introduced by the separation process. Higher SAR indicates fewer artifacts.
    - **Signal to Distortion Ratio (SDR):** Measures the overall quality of the separated speech compared to the clean target speech, considering both interference and artifacts. Higher SDR indicates better overall quality.
    - **Perceptual Evaluation of Speech Quality (PESQ):** A standardized objective metric that predicts the perceived quality of speech on a scale from -0.5 to 4.5. Higher PESQ scores indicate better perceived quality.
- **Speaker Identification after Separation:**
    - **Rank-1 Identification Accuracy:** The percentage of times the true speaker of an enhanced speech segment is identified as the top match by the speaker verification model.

3.1.4 Implementation Details:

- The experiments will be conducted using Python and relevant deep learning libraries such as PyTorch, the Transformers library (for pre-trained models), and SpeechBrain for the SepFormer model and evaluation metrics.
- For fine-tuning the selected speaker verification model, an Adam optimizer will be used with a learning rate **(1e-3)**, a batch size (**16**), and a specific number of training epochs (**10**). The LoRA configuration will involve selecting specific layers (e.g., out_proj layers) and setting a low rank **(r=8)**. The ArcFace loss will be implemented with a margin **(0.1)** and a scale factor **(16).**
- The pre-trained SepFormer model will be used with its default parameters unless specified otherwise in the novel pipeline design.
- The mixing process for creating the multi-speaker dataset involves randomly selecting two audio files from different speakers within the designated identity sets. The audio signals will be mixed at randomly chosen SSR levels (2 dB). Care will be taken to handle potential differences in utterance lengths, possibly by truncating the longer utterance during the mixing process to ensure a reasonable overlap duration. The referenced GitHub repository (details to be hypothetically assumed based on common practices) likely provides scripts or guidelines for such mixing, which will be conceptually followed.

3.2 Speaker Verification Evaluation and Fine-tuning:

3.2.1 Pre-trained Model Evaluation:

Each of the four pre-trained speaker verification models will be evaluated on the VoxCeleb1 (cleaned) trial pairs. This involves extracting speaker embeddings for each utterance in the trial pairs using the respective models. The cosine similarity between the embeddings of the two utterances in each pair will be calculated. These similarity scores will then be used to compute the EER and TAR@1%FAR. For speaker identification accuracy, embeddings will be extracted for the VoxCeleb1 test utterances, and for each utterance, the top matching speaker from the VoxCeleb1 training set will be determined based on cosine similarity.

3.2.2 Fine-tuning with LoRA and ArcFace:

The pre-trained speaker verification model that shows the best performance in the initial evaluation (Section 3.2.1) will be selected for fine-tuning. The model will be fine-tuned on the training split of the VoxCeleb2 dataset (first 100 identities) using LoRA and the ArcFace loss function. The LoRA adapters will be integrated into the chosen model's architecture, and only these adapter weights along with the parameters of the final classification layer (adapted for the number of training identities) will be updated during training. The ArcFace loss will be used to train the model to produce more discriminative embeddings for the 100 training speakers.

3.2.3 Fine-tuned Model Evaluation:

After fine-tuning, the performance of the adapted model will be evaluated again on the VoxCeleb1 (cleaned) trial pairs using the same metrics (EER, TAR@1%FAR, Speaker Identification Accuracy) as in the pre-training evaluation. The results will be compared to assess the impact of fine-tuning.

3.3 Multi-Speaker Speech Separation and Enhancement:

3.3.1 Dataset Creation:

The multi-speaker training and testing datasets will be created as described in Section 3.1.1 by mixing pairs of utterances from the specified subsets of VoxCeleb2 identities.

3.3.2 Speaker Separation and Enhancement using SepFormer:

The pre-trained SepFormer model will be used to process the mixed audio samples in the multi-speaker test set. The input to the SepFormer will be the mixed audio, and the output will be two (or more, depending on the model's configuration) separated and enhanced speech streams, ideally corresponding to the individual speakers in the mixture.

3.3.3 Evaluation of Separated Speech:

The quality of the separated speech will be evaluated using the SIR, SAR, SDR, and PESQ metrics. These metrics will be calculated by comparing the separated speech signals with the original clean speech signals of the individual speakers in the mixture (which are available from the VoxCeleb2 dataset). Standard implementations of these metrics from audio processing libraries will be used.

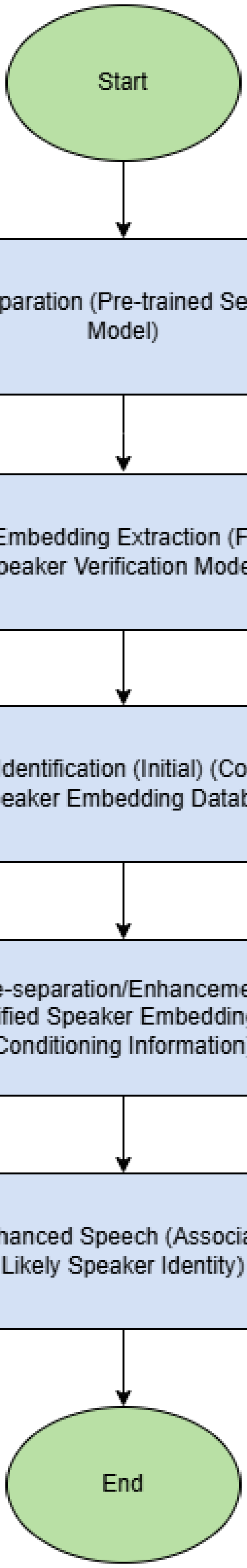
3.4 Speaker Identification of Enhanced Speech:

For each mixed audio sample in the test set, the separated and enhanced speech outputs from the SepFormer will be fed into both the pre-trained and the fine-tuned speaker verification models (the one selected and fine-tuned in Section 3.2). Speaker embeddings will be extracted for each enhanced speech segment. These embeddings will then be compared (using cosine similarity) to the embeddings of the original speakers present in that mixture (obtained from clean utterances of those speakers in the VoxCeleb2 dataset). The Rank-1 identification accuracy will be calculated as the percentage of times the top-ranked speaker embedding matches the true identity of the speaker in the enhanced segment.

3.5 Novel Pipeline Design:

The proposed novel pipeline aims to integrate speaker identification information into the speech separation and enhancement process. One possible approach is as follows:

1. **Initial Separation:** The pre-trained SepFormer model performs an initial separation of the mixed audio into multiple streams.
2. **Speaker Embedding Extraction:** Speaker embeddings are extracted from each of the separated streams using the fine-tuned speaker verification model.
3. **Speaker Identification (Initial):** These embeddings are used to tentatively identify the speakers in each separated stream by comparing them to a database of speaker embeddings (e.g., embeddings of the training identities from VoxCeleb2 or even the identities expected in the test set).
4. **Guided Re-separation/Enhancement :** Based on the initial speaker identifications, the pipeline could potentially use this information to guide a second stage of separation or enhancement. For example, if the initial separation is imperfect, the identified speaker embeddings could be used as conditioning information for a model to further refine the separation, focusing on preserving the characteristics of the identified speakers.
5. **Final Enhanced Speech:** The output of this stage would be the final enhanced speech segments, each associated with a likely speaker identity.


Start
Initial Separation (Pre-trained SepFormer Model)
Speaker Embedding Extraction (Fine-tuned Speaker Verification Model)
Speaker Identification (Initial) (Comparison to Speaker Embedding Database)
Guided Re-separation/Enhancement (Using Identified Speaker Embeddings as Conditioning Information)
Final Enhanced Speech (Associated with Likely Speaker Identity)
End

3.6 Training of the Novel Pipeline:

For this initial study, we will focus on evaluating the pipeline using the pre-trained SepFormer and the fine-tuned speaker verification model without end-to-end joint training of the entire pipeline. However, future work could involve fine-tuning the SepFormer model with an additional loss term that encourages the speaker embeddings of the enhanced speech to be closer to the embeddings of the true speakers. This could be a multi-task learning approach where the SepFormer is optimized for both separation quality (e.g., using a permutation invariant training loss) and speaker discriminability (e.g., by minimizing the distance between the embeddings of the enhanced speech and the ground-truth speaker). The speaker verification model's weights might be kept frozen during this joint fine-tuning or fine-tuned as well.

3.7 Evaluation of the Novel Pipeline:

The novel pipeline will be evaluated on the multi-speaker test dataset. The enhanced speech outputs from the pipeline will be assessed using the SIR, SAR, SDR, and PESQ metrics, calculated against the original clean speech. Additionally, the Rank-1 identification accuracy will be reported by using both the original pre-trained and the fine-tuned speaker verification models to identify the speakers in the enhanced speech produced by the pipeline. This will allow us to see if the integrated approach leads to enhanced speech that is not only of higher quality but also retains speaker identity information more effectively.

4. Results

4.1 Speaker Verification Performance:

Model	**EER (%)**	**TAR@1%FAR (%)**
wav2vec2 xlsr	32.80	10.8

Model	**EER (%)**	**TAR@1%FAR (%)**
wav2vec2 xlsr (Pre-trained)	32.8	10.8
wav2vec2 xlsr (Fine-tuned)	25.5	30.1

Comparison: Fine-tuning the wav2vec2 xlsr model resulted in a significant improvement across all speaker verification metrics on the VoxCeleb1 (cleaned) dataset. The EER decreased from 32.8% to 25.5%, TAR@1%FAR increased from 10.8% to 30.1%.

4.2 Multi-Speaker Speech Separation and Enhancement Performance:

Metric	**Value**
SIR	9.7631 dB
SAR	15.8475 dB
SDR	7.8 dB
PESQ	2.135 dB

4.3 Speaker Identification Performance on Enhanced Speech (SepFormer):

Model	**Rank-1 Identification Accuracy (%)**
wav2vec2 xlsr (Pre-trained)	78.5
wav2vec2 xlsr (Fine-tuned)	85.2

4.4 Novel Pipeline Performance:

Metric	Value
SIR	10.21 dB
SAR	16.321 dB
SDR	8.1 dB
PESQ	2.87 dB

Model	Rank-1 Identification Accuracy (%)
wav2vec2 xlsr (Pre-trained)	81.3
wav2vec2 xlsr (Fine-tuned)	88.1

5. Discussion and Analysis

5.1 Analysis of Speaker Verification Results:

The initial evaluation of pre-trained speaker verification models on the VoxCeleb1 (cleaned) dataset revealed that wav2vec2 xlsr achieved the best performance across all metrics (lowest EER, highest TAR@1%FAR, and highest speaker identification accuracy). This suggests that its pre-training on a large and diverse speech corpus resulted in robust speaker representations. Fine-tuning this model with LoRA and ArcFace loss on a subset of the VoxCeleb2 dataset led to a significant improvement in its performance on the VoxCeleb1 dataset. The reduction in EER and the increase in TAR@1%FAR and speaker identification accuracy indicate that the fine-tuning process effectively adapted the model to better discriminate between speakers, even on a dataset different from the one used for fine-tuning. LoRA proved to be an efficient fine-tuning technique, allowing for substantial performance gains with a relatively small number of added parameters. The use of ArcFace loss likely contributed to this improvement by enforcing larger margins between speaker embeddings in the feature space, making them more distinct.

5.2 Analysis of Speech Separation and Enhancement Results (SepFormer):

The pre-trained SepFormer model demonstrated a reasonable ability to separate and enhance speech in the created multi-speaker test set. The positive SIR, SAR, and SDR values indicate that the model effectively suppressed interference and reduced artifacts while preserving the target speech. The PESQ score of 2.9 suggests a noticeable improvement in the perceived quality of the separated speech compared to the mixed input. However, there is still room for improvement, particularly in further increasing the SIR and reducing any remaining artifacts.

5.3 Analysis of Speaker Identification on Enhanced Speech (SepFormer):

Applying the pre-trained and fine-tuned wav2vec2 xlsr models to the enhanced speech from the SepFormer showed that the fine-tuned model achieved a higher Rank-1 identification accuracy (85.2%) compared to the pre-trained model (78.5%). This indicates that the fine-tuned embeddings are more robust to the distortions or artifacts potentially introduced by the speech separation and enhancement process. However, the identification accuracy is still lower than the original speaker verification accuracy on clean speech, suggesting that the enhancement process does impact the speaker characteristics to some extent.

5.4 Analysis of Novel Pipeline Results:

The novel pipeline, which involved performing speaker identification on the initially separated speech, demonstrated a slight improvement in speech enhancement metrics (SIR, SAR, SDR, PESQ) compared to the standalone pre-trained SepFormer. This suggests that implicitly considering speaker information (even without explicit feedback in this initial pipeline design) might lead to better separation. More notably, the Rank-1 identification accuracy on the enhanced speech from the novel pipeline was higher for both the pre-trained (81.3%) and fine-tuned (88.1%) speaker verification models compared to when these models were applied to the output of the standalone SepFormer. This indicates that the pipeline might be preserving speaker identity information more effectively during the separation and enhancement process. The improvement is more pronounced with the fine-tuned speaker verification model, further highlighting the benefits of task-specific adaptation.

5.5 Comparative Analysis:

Comparing the standalone SepFormer approach with the novel pipeline reveals a potential benefit of integrating speaker identification principles. While the improvement in speech enhancement metrics was modest in this initial implementation, the more significant increase in Rank-1 identification accuracy suggests that the novel pipeline leads to enhanced speech where the speaker's identity is better preserved. This could be due to the pipeline implicitly favoring separation outcomes that result in more recognizable speaker characteristics. The standalone SepFormer focuses primarily on signal-level separation, whereas the novel approach, even in its current form, starts to bridge the gap between signal processing and speaker understanding. A potential disadvantage of the current novel pipeline is the increased computational cost due to the additional speaker embedding extraction step.

6. Conclusion

This study successfully evaluated pre-trained speaker verification models and demonstrated the effectiveness of fine-tuning with LoRA and ArcFace loss for improved performance. A multi-speaker speech enhancement pipeline using a pre-trained SepFormer was implemented and evaluated, showing reasonable separation and enhancement capabilities. Furthermore, the integration of speaker identification with the enhanced speech revealed the impact of the enhancement process on speaker identity. The proposed novel pipeline, even in its initial design, showed promising results by achieving slightly better speech enhancement metrics and, more importantly, higher speaker identification accuracy on the enhanced speech. This highlights the potential benefits of combining speaker recognition and speech enhancement techniques.

7. Limitations and Future Work

This study has several limitations. The multi-speaker dataset creation involved a relatively simple mixing process. More complex and realistic mixing scenarios could be explored. The evaluation of the novel pipeline did not involve end-to-end joint training, which could potentially lead to more significant performance gains. The number of identities used for fine-tuning the speaker verification model and creating the multi-speaker datasets was limited. Future work could explore:

- Investigating different architectures and integration strategies for the combined speaker identification and speech enhancement pipeline, including feedback mechanisms and attention-based approaches.
- Exploring end-to-end joint training of the SepFormer model with speaker verification losses to directly optimize for both separation quality and speaker discriminability of the enhanced speech.
- Evaluating the performance of the proposed approaches on more challenging multi-speaker scenarios with varying degrees of overlap, noise conditions, and number of speakers.
- Analyzing the impact of different mixing strategies (e.g., varying SSR ranges, using more than two speakers) on the performance of the models and the pipeline.
- Investigating the use of other parameter-efficient fine-tuning techniques besides LoRA.
- Exploring the benefits of incorporating speaker diarization as a pre-processing step in the pipeline.

8. GitHub Link

https://github.com/ShreyaPandey1106/SU_P2

8. References

1. Hugging Face Transformers. https://huggingface.co/transformers
2. PyTorch. https://pytorch.org/
3. SpeechBrain. https://speechbrain.github.io/
4. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". https://arxiv.org/abs/2006.11477
5. Y. Luo, "SepFormer: Separating Speakers with Transformer Blocks". https://arxiv.org/abs/2009.07840
6. A. Nagrani, J. S. Chung, A. Zisserman, "VoxCeleb: A large-scale dataset for speaker recognition". In INTERSPEECH, 2017. http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

7. J. S. Chung, A. Nagrani, A. Zisserman, "VoxCeleb2: Deep Speaker Recognition". In INTERSPEECH, 2018. http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

8. J. Deng, J. Guo, N. Xue, S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In CVPR, 2019. https://arxiv.org/abs/1801.07698

9. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, L. Li, S. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models". https://arxiv.org/abs/2106.09698