Transformer-based Source Separation for Multi-speaker Scenarios

In21-S7-CS4681- Advanced Machine Learning- Research Assignment

PROGRESS REPORT

Arachchi H.M.W 210049U

B.Sc. Engineering (Hons)

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

August 2025

Contents

1. Introduction 3
2. Implementation Plan 3
2.1 Objectives 3
2.2 Scope Clarification 4
2.3 Expected Contributions 4
2.4 Detailed Project Timeline 4
3. Literature Review 7
3.1 Foundations of Recurrent Neural Networks in Speech Processing 7
3.2 Limitations of Recurrent Connections 8
3.3 Emergence of Transformers 8
3.4 Deep Learning for Monaural Source Separation 8
3.5 Dual-Path RNN (DPRNN) 9
3.6 Dual-Path Transformer Network (DPTNet) 9
3.7 Research Gap 10
4. Methodology 11
4.1 Overall Model Architecture 11
4.2 Masking Network: Dual-Path Framework with Multi-Scale Transformers 11
4.3 SepFormer Block 11
4.3.1 Intra Transformer (IntraT) 11
4.3.2 Inter Transformer (InterT) 11
4.3.3 Transformer Architecture (f(.)) 12
4.4 Encoder and Decoder 12
4.5 Training Details 12
5. Conclusion 13

Abstract

This report outlines a research project focused on enhancing SepFormer, a state-of-the-art Transformer-based neural network for speech separation. The SepFormer architecture, which avoids Recurrent Neural Networks (RNNs), leverages multi-head attention and a multi-scale dual-path framework to learn both short and long-term dependencies. It has demonstrated state-of-the-art (SOTA) performance on the WSJ0-2mix and WSJ0-3mix datasets, achieving SI-SNRi values of 22.3 dB and 19.5 dB respectively with dynamic mixing. A key advantage of SepFormer is its parallelization capabilities, leading to faster training and inference and reduced memory usage compared to RNN-based models like DPRNN and DPTNet. This project aims to extend the SepFormer's capabilities to multi-speaker scenarios in noisy and reverberant environments, utilizing the WHAM! and WHAMR! datasets, with a focus on scalability, real-time performance, and applicability to low-resource languages.

1. Introduction

The accurate separation of individual speech signals from a mixed audio stream is a critical challenge in various audio processing applications, particularly in multi-speaker environments. Recurrent Neural Networks (RNNs) have traditionally dominated sequence-to-sequence learning in this domain, including speech separation, due to their ability to model long-term dependencies. However, their inherent sequential nature restricts computational parallelization, creating a bottleneck, especially with large datasets and extended sequences.

This report outlines the progress on leveraging Transformer-based architectures for speech separation, with a particular focus on the SepFormer model. SepFormer addresses the limitations of RNNs by replacing recurrent computations with a multi-head attention mechanism and adopting a multi-scale dual-path approach, thereby achieving state-of-the-art performance and enhanced computational efficiency in multi-speaker scenarios.

2. Implementation Plan

2.1 Objectives

The primary objective is to implement and evaluate Transformer-based neural networks for speech separation, specifically focusing on the SepFormer architecture. This includes:

- Developing an RNN-free model that can effectively learn both short and long-term dependencies in speech mixtures.
- Achieving state-of-the-art (SOTA) performance on standard multi-speaker speech separation benchmarks, such as WSJ0-2mix and WSJ0-3mix datasets.
- Demonstrating significant improvements in training and inference speed and reduced memory consumption compared to existing RNN-based separation systems.

2.2 Scope Clarification

This project focuses on monaural speech source separation in multi-speaker scenarios, specifically targeting mixtures of two and three speakers as represented by the WSJ0-2mix and WSJ0-3mix datasets. The core innovation under investigation is the complete replacement of recurrent connections with multi-head attention mechanisms within a dual-path framework. The encoder and decoder components will utilize convolutional layers, while the masking network will be entirely Transformer-based.

2.3 Expected Contributions

The expected contributions of this work, building on the SepFormer's capabilities, include:

- Establishing a robust and efficient Transformer-based architecture that outperforms RNN-based models in terms of separation quality (SI-SNRi, SDRi).
- Providing a model that offers superior computational efficiency, allowing for faster training and inference times and lower memory usage.
- Validating the effectiveness of a multi-scale Transformer approach within a dual-path framework for capturing both short and long-term dependencies in speech signals.
- Confirming the benefits of techniques like sinusoidal positional encoding and dynamic mixing for enhancing separation performance.

2.4 Detailed Project Timeline

Weeks 1–2 (Aug 13– Aug 26): Literature Review Focus on reviewing:

- Recurrent Neural Networks (RNNs), specifically LSTMs and GRUs, their role in sequence-to-sequence learning and audio processing, and their inherent sequential nature which impairs parallelization and creates bottlenecks, especially with large datasets and long sequences.
- Transformers, their attention-based mechanisms that replace recurrent computations, allowing for parallelization and easier learning of long-term dependencies by attending to the whole sequence at once.
- Deep learning techniques for monaural audio source separation, including end-to-end approaches and the learned-domain masking strategy popularised by Conv-TasNet.
- Dual-Path RNN (DPRNN) for its long-term modelling capabilities and its limitations due to RNNs, as well as the Dual-Path Transformer Network (DPTNet) which attempted to integrate transformers but still embedded an RNN. Organise references into a structured hierarchy and prepare comparative notes. Outcome: Research gaps identified, highlighting the need for a truly RNN-free Transformer-based model for speech separation.

Weeks 3–4 (Aug 27– Sep 9): Problem Formulation & Scope Definition Define problem:

Developing an RNN-free Transformer-based neural network for speech separation that overcomes the parallelization limitations of RNNs while achieving state-of-the-art performance. Identify specific goals:

- Design a model that learns short and long-term dependencies using a multi-scale approach with transformers.
- Leverage the parallelization advantages of Transformers for faster training and inference.

- Achieve competitive performance even when down sampling the encoded representation for reduced memory demands. Select datasets: The standard WSJ0-2mix and WSJ0-3mix datasets. Fix evaluation metrics: Scale-Invariant Signal-to-Noise Ratio improvement (SI-SNRi) and Signal-to-Distortion Ratio improvement (SDRi), alongside analysis of training speed, forward-pass time, and memory usage.

Weeks 5–6 (Sep 10– Sep 23): Model Design & Prototype Implementation:

Design the SepFormer architecture based on the learned-domain masking approach, employing an encoder, a decoder, and a masking network.

- Encoder: Fully convolutional layer to estimate an STFT-like representation from the time-domain mixture signal.
- Masking Network: Incorporates the SepFormer block within a dual-path processing block. This includes layer normalization, linear layers, creating overlapping chunks of the encoded input, PReLU activations, and an overlap-add scheme to produce masks for each speaker.
- SepFormer Block: Designed to model short and long-term dependencies with a dual-scale approach, featuring an Intra Transformer (IntraT) for short-term dependencies within chunks and an Inter Transformer (InterT) for longer-term dependencies across chunks after a permutation step.
- Transformer Architecture (for IntraT and InterT): Incorporates sinusoidal positional encoding, Layer Normalisation, Multi-Head Attention (MHA), and Feed-Forward Networks (FFW), with residual connections for improved gradient backpropagation.
- Decoder: A transposed convolution layer to reconstruct separated signals in the time domain using the estimated masks. Implement prototype in PyTorch, starting with the encoder, decoder, and the core masking network components, including the IntraT and InterT. Outcome: Functional prototype of the SepFormer model.

Weeks 7–8 (Sep 24– Oct 7): Experiments on WSJ0-2mix and WSJ0-3mix Datasets:

Train and validate the SepFormer model on the WSJ0-2mix and WSJ0-3mix datasets.

- Use an encoder with 256 convolutional filters, kernel size 16, and stride 8.
- Configure the masking network with chunks of size C=250, 50% overlap, 8 layers for both IntraT and InterT, repeated N=2 times, 8 parallel attention heads, and 1024-dimensional positional feed-forward networks.
- Employ the Adam optimizer with a learning rate of 15e-5, learning rate annealing, gradient clipping, a batch size of 1, and scale-invariant signal-to-noise ratio (SI-SNR) via utterance-level permutation invariant loss.
- Integrate dynamic mixing (DM) data augmentation, including speed perturbation. Compare the SepFormer's SI-SNRi and SDRi performance against existing state-of-the-art models such as DPRNN, DPTNet, ConvTasNet, and Wavesplit on both datasets. Deliverable: Preliminary results demonstrating state-of-the-art performance on WSJ0-2mix (22.3 dB SI-SNRi with DM) and WSJ0-3mix (19.5 dB SI-SNRi with DM), with reproducible scripts.

Week 9 (Oct 8– Oct 14): Refinement and Ablation Studies:

Conduct ablation studies on the WSJ0-2mix validation set to analyze the effect of various hyperparameters and data augmentation:

- Investigate the impact of the number of IntraT and InterT blocks (N_intra, N_inter) and their repetitions (N) on performance.
- Examine the contribution of positional encoding to separation performance.
- Evaluate the effect of the number of attention heads.
- Quantify the significant performance improvement provided by dynamic mixing.
- Analyse the impact of the encoder's stride factor on performance, speed, and memory, noting SepFormer's competitive results even with a larger stride compared to DPRNN. Optimise hyperparameters for further improvements in SI-SNRi and SDRi. Compare training speed and memory usage of SepFormer against DPRNN, DPTNet, and Wavesplit, highlighting SepFormer's efficiency due to parallelization and effective stride usage. Outcome: Optimised model configuration and a clear understanding of component contributions.

Week 10 (Oct 15– Oct 21): Report Writing and Documentation:

Finalize the report with tables, figures, methodology, experiments, discussion, and future work.

- Introduction: Discuss the shift from RNNs to Transformers in sequence-to-sequence learning and speech processing.
- Model Description: Detail the SepFormer's learned-domain masking approach, the convolutional encoder, the dual-path masking network with Intra Transformer and Inter Transformer blocks, and the transposed convolutional decoder.
- Experimental Setup: Provide comprehensive details on the WSJ0-2mix and WSJ0-3mix datasets, architectural parameters, and training methodology including dynamic mixing.
- Results: Present clear comparisons of SI-SNRi and SDRi against state-of-the-art models, highlighting SepFormer's superior performance.
- Ablation Study: Document the findings from hyperparameter analysis, e.g., the importance of IntraT/InterT layers, positional encoding, and dynamic mixing.
- Speed and Memory Comparison: Illustrate and discuss the parallelization advantages of SepFormer, showcasing its faster training and inference times, and reduced memory footprint compared to RNN-based models like DPRNN, DPTNet, and Wavesplit.
- Conclusion: Summarize the key findings, emphasizing the achievement of state-of-the-art, RNN-free, Transformer-based speech separation with significant computational efficiencies.
- Future Work: Outline potential avenues for further research and improvement. Ensure reproducibility and clean code documentation, noting the model's availability within the SpeechBrain toolkit.

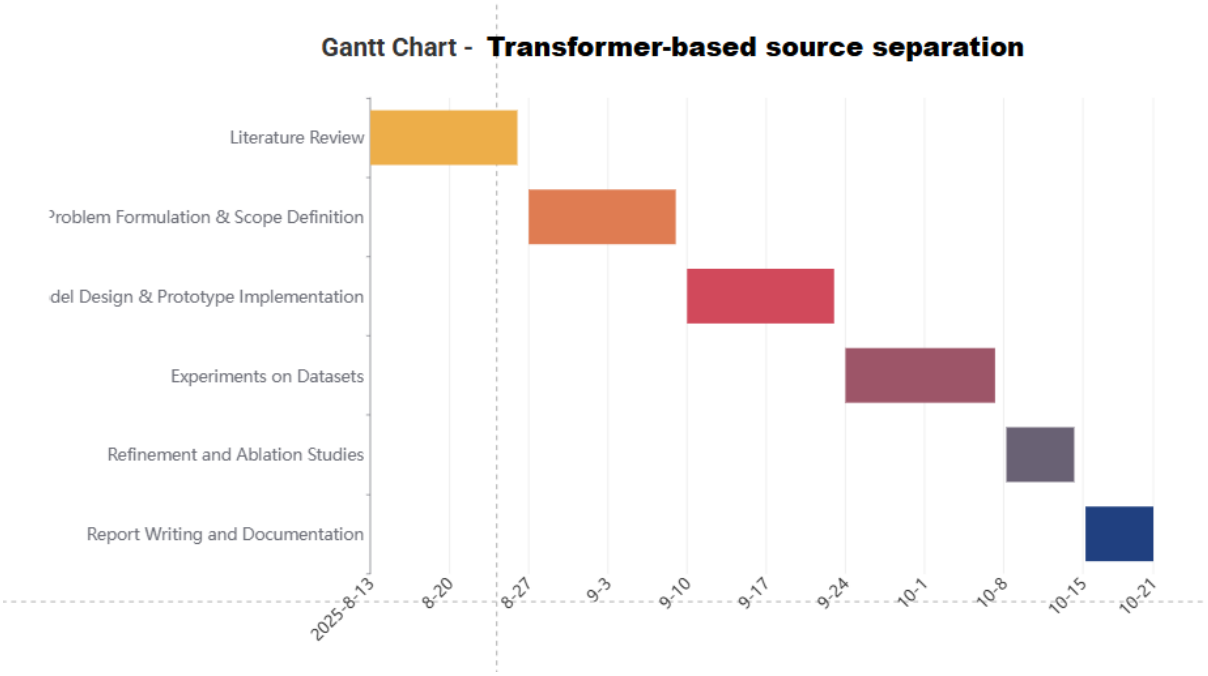

Figure 1: Timeline

3. Literature Review

3.1 Foundations of Recurrent Neural Networks in Speech Processing

Recurrent Neural Networks (RNNs) have long been a cornerstone in speech processing, enabling the modeling of sequential dependencies that are inherent to audio signals. Unlike feedforward architectures, which operate on fixed-length inputs without memory, RNNs incorporate recurrent connections that allow information to persist across time steps. This capability makes them particularly suited for speech-related tasks, where contextual information and temporal continuity are crucial for understanding and generation [1], [2].

A major advancement in RNNs came with the introduction of gated mechanisms, such as Long Short-Term Memory (LSTM) [3] and Gated Recurrent Units (GRU) [4]. These architectures address the vanishing and exploding gradient problems encountered in vanilla RNNs, thereby enabling the capture of long-term temporal dependencies in speech signals. The gating structures dynamically regulate how information is stored, updated, and forgotten, which is essential for effectively modeling complex acoustic patterns.

RNN-based architectures have been widely adopted across speech domains, including automatic speech recognition (ASR), speech enhancement, speech separation, and text-to-speech (TTS) synthesis. For instance, in ASR, bidirectional LSTMs coupled with Connectionist Temporal Classification (CTC) loss significantly improved recognition accuracy by learning context from both past and future frames [5]. Similarly, in speech enhancement and separation, RNNs are employed to estimate clean signals or time-frequency masks from noisy and overlapped mixtures, exploiting their ability to capture dependencies over extended time scales [6], [7].

Although more recent models based on attention mechanisms, such as Transformers, have surpassed RNNs in many benchmarks, RNNs remain highly relevant in low-resource and real-time applications due to their relative simplicity, efficiency, and lower computational footprint. As such, they continue to provide a strong foundation for both research and practical deployments in speech processing systems [8].

3.2 Limitations of Recurrent Connections

Despite their effectiveness in modeling temporal dependencies, Recurrent Neural Networks (RNNs) face significant limitations due to their inherently sequential processing nature. By definition, RNNs compute each time step based on the hidden state from the previous one, making them difficult to parallelize across sequence length [9]. This sequential bottleneck leads to inefficiencies when training on large datasets or handling long input sequences, resulting in slower convergence and increased computational demands.

Several architectural enhancements have attempted to alleviate these issues. For example, the Dual-Path RNN (DPRNN) [10] introduced a two-level processing scheme where input sequences are divided into local and global chunks, enabling improved long-range modeling. While this design substantially improves separation performance in speech tasks, it still inherits the fundamental sequential limitations of recurrent connections, particularly in its global modeling stage.

Another challenge arises from the choice of stride in the encoder. RNN-based systems such as DPRNN often require a stride factor of 1 to preserve temporal resolution and maintain performance. This results in processing more data points per sequence, which increases training time, inference latency, and memory usage. In contrast, non-recurrent architectures such as Transformer-based models can operate with larger stride factors (e.g., 8 samples) while maintaining competitive performance [11], [12]. This efficiency gain stems from the parallelizable self-attention mechanism, which allows simultaneous processing of multiple time steps, thereby offering superior scalability and faster convergence compared to RNN-based counterparts.

Consequently, while RNNs remain useful in resource-constrained or low-latency scenarios, their limited parallelizability and scalability make them less favorable for large-scale speech processing tasks. This has driven the transition towards attention-based architectures, which offer greater efficiency without compromising on accuracy [13].

3.3 Emergence of Transformers

Transformers have emerged as a powerful alternative to standard RNNs, completely avoiding the sequential bottleneck by eliminating recurrence and replacing it with a fully attention-based mechanism. By attending to the entire sequence simultaneously, Transformers establish direct connections between distant elements, enabling them to learn long-term dependencies more easily. This has led to their growing popularity in various speech processing tasks, including speech recognition, synthesis, enhancement, diarization, and speaker recognition.

3.4 Deep Learning for Monaural Source Separation

The problem of monaural source separation, which involves extracting individual sources from a single-channel audio mixture, has seen significant progress with the advent of deep learning. Traditional approaches, such as non-negative matrix factorization (NMF) and computational auditory scene analysis (CASA), often struggled to generalize in highly variable acoustic conditions. In contrast, modern deep learning models have demonstrated the ability to learn robust representations directly from data, achieving state-of-the-art results across multiple benchmarks [14].

A key breakthrough came with the development of end-to-end time-domain separation networks, most notably Conv-TasNet [11]. Unlike earlier methods that operated in the time–frequency (STFT) domain, Conv-TasNet learns an overcomplete set of analysis and synthesis filters directly from raw waveforms. Separation is then performed in this learned latent space by estimating masks for each target source. This framework proved highly effective, surpassing traditional ideal time–frequency magnitude masking approaches while offering lower latency and greater flexibility.

Following Conv-TasNet, several extensions were introduced to address its limitations in modeling long sequences. The Dual-Path RNN (DPRNN) [15] improved long-range temporal modeling by employing a two-stage chunking mechanism, enabling local and global context processing. Building upon this, the SepFormer [16] replaced recurrent units with Transformer layers, achieving superior performance by leveraging self-attention to capture both local and global dependencies in parallel. These advances highlight the importance of effectively modeling long input sequences, which is central to achieving high separation accuracy in monaural settings.

Despite their success, challenges remain. Many deep learning-based monaural separation systems are computationally expensive and struggle in real-time or low-resource scenarios. Recent research has therefore explored lightweight architectures, quantization strategies, and self-supervised pretraining approaches to improve scalability and generalization [17]. Moreover, ongoing efforts aim to extend these models to multi-speaker, noisy, and reverberant environments, bringing them closer to practical deployment in real-world speech enhancement and separation applications.

3.5 Dual-Path RNN (DPRNN)

The Dual-Path RNN (DPRNN) demonstrated that effective long-term modeling is crucial for achieving strong performance in speech separation tasks. The core innovation of DPRNN lies in its two-stage chunking mechanism, where input sequences are divided into overlapping segments that are processed both locally and globally by separate RNN modules [15]. This hierarchical design allows the network to capture fine-grained local dependencies while also modeling broader temporal context, significantly improving separation accuracy compared to conventional RNN-based approaches.

Despite these advances, DPRNN still inherits the fundamental limitations of recurrent connections, particularly in the global processing step. Since RNNs are inherently sequential, DPRNN's training and inference remain less efficient compared to architectures based on fully parallelizable mechanisms such as Transformers. As a result, DPRNN requires longer training times and incurs higher memory usage, especially when dealing with large datasets or longer input sequences [12].

From an empirical standpoint, DPRNN has been shown to achieve competitive results on standard benchmarks. With approximately 2.6 million parameters, it reached an SI-SNRi of 18.8 dB and an SDRi of 19.0 dB on the WSJ0-2mix dataset, and an SI-SNRi of 14.7 dB on the more challenging WSJ0-3mix dataset [15]. While these results represent a significant step forward for RNN-based models, Transformer-based alternatives such as SepFormer have since surpassed DPRNN in both accuracy and computational efficiency [19].

.

3.6 Dual-Path Transformer Network (DPTNet)

An attempt to integrate transformers into the speech separation pipeline was made with the Dual-Path Transformer Network (DPTNet). This architecture was shown to outperform the standard DPRNN. However, a significant drawback of DPTNet is that it still embeds a Recurrent Neural Network (RNN)

within its architecture. This effectively negates the parallelization capability inherent to pure-attention models.

Despite having 2.6 million parameters, comparable to DPRNN, DPTNet achieved an SI-SNRi of 20.2 dB and an SDRi of 20.6 dB on the WSJ0-2mix dataset. Nevertheless, similar to DPRNN, DPTNet also presents slower training times and higher memory usage compared to purely Transformer-based alternatives. This is due to its reliance on RNN components, which means it cannot fully leverage the parallelization capabilities of pure-attention models. Consequently, DPTNet, like DPRNN, requires a stride factor of 1 in its encoder to maintain performance, meaning it processes more data, contributing to its slower operation and higher memory demands.

In contrast, a purely Transformer-based model like the SepFormer can achieve competitive performance even with a stride factor of 8, processing significantly less data and thus offering superior speed and memory efficiency. Comparisons have shown that SepFormer is faster and less memory-demanding than DPTNet during both training and inference. For example, the SepFormer reaches above 17 dB SI-SNRi levels on WSJ0-2mix after approximately one day of training, whereas DPTNet takes longer. Furthermore, DPTNet shows higher average forward-pass times and greater memory usage on input sequences ranging from 1 to 5 seconds long, as illustrated in Figure 2 [13].

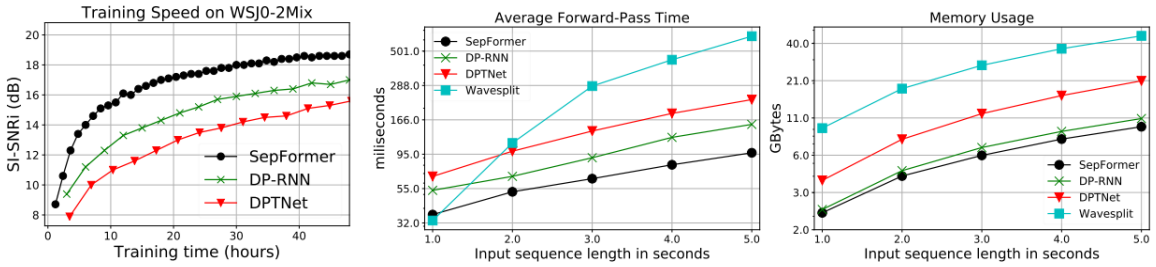

Figure 2 (Left) The traning curves of SepFormer, DPRNN, and DPTNeT on the WSJ0-2mix dataset. (Middle & Right) The comparison of forward-pass speed and memory usage in the GPU on inputs ranging 1-5 seconds long sampled at 8kHz.

3.7 Research Gap

While deep learning has significantly advanced monaural audio source separation, and Transformers have shown promise in various speech processing tasks, little research had been done on purely Transformer-based models for monaural audio source separation that are entirely free of Recurrent Neural Networks. Previous efforts, such as DPTNet, while integrating Transformers, still retained RNN components, thus failing to fully leverage the parallelization advantages of pure-attention models. This created a significant research gap for an RNN-free Transformer-based architecture capable of achieving state-of-the-art performance while offering improved computational efficiency.

The SepFormer was proposed to address this gap, presenting itself as a novel RNN-free Transformer-based neural network for speech separation that learns short- and long-term dependencies through a multi-scale approach employing transformers. This architecture is designed to overcome the limitations of RNN-based systems by allowing computations over different time steps to be parallelized, leading to faster training and inference, and reduced memory usage.

4. Methodology

The proposed SepFormer model is an RNN-free Transformer-based neural network designed for speech separation. It adopts the learned-domain masking approach and consists of an encoder, a decoder, and a masking network.

4.1 Overall Model Architecture

The high-level architecture of SepFormer comprises three main components:

- Encoder: Estimates a learned representation from the input mixture signal [13, Figure 3].
- Masking Network: The core component that estimates optimal masks for separating sources [13, Figure 3].
- Decoder: Reconstructs the estimated sources in the time domain using the masks [13, Figure 3].

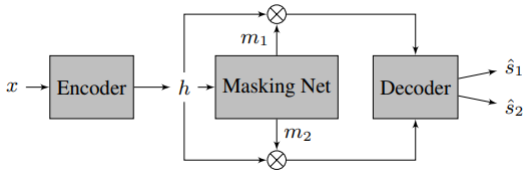

Figure 3

4.2 Masking Network: Dual-Path Framework with Multi-Scale Transformers

The masking network is fed by the encoded representations and estimates a mask for each speaker. Crucially, SepFormer adopts the dual-path framework (originally from DPRNN) but replaces the RNNs with a multi-scale pipeline composed entirely of Transformers. This framework enables the model to learn both short and long-term dependencies.

- The input h is normalized and processed by a linear layer.
- It is then divided into overlapping chunks of size C with a 50% overlap. This chunking strategy is vital for mitigating the quadratic complexity of Transformers by allowing them to process smaller segments.

4.3 SepFormer Block

The SepFormer block is the main component of the masking network, designed to model both short and long-term dependencies using a dual-scale approach. It is repeated N times within the masking network.

4.3.1 Intra Transformer (IntraT)

- The IntraT processes the second dimension of the chunked input, meaning it acts on each chunk independently.
- This component is responsible for modelling short-term dependencies within each individual chunk.
- In the best performing models, 8 layers of IntraT are used.

4.3.2 Inter Transformer (InterT)

- After the IntraT, the last two dimensions of the output are permuted.
- The InterT is then applied to model transitions across these chunks.

- This mechanism effectively captures longer-term dependencies across the entire sequence by considering how information flows between chunks.
- In the best performing models, 8 layers of InterT are used.
- An ablation study showed that respectable performance (19.2 dB SI-SNRi) can still be achieved even with a single layer transformer for the InterT, suggesting that Intra Transformer (local processing) has a greater influence on performance.

4.3.3 Transformer Architecture (f(.))

The internal architecture of both the IntraT and InterT blocks closely resembles the original Transformer defined in [Figure 4].

- Sinusoidal Positional Encoding: To inject information about the order of elements in the sequence, sinusoidal positional encoding e is added to the input z. This significantly improves separation performance.
- Multi-Head Attention (MHA): Within each Transformer layer, after layer normalization, multi-head attention is applied. Each attention head computes scaled dot-product attention between all elements of the sequence. The best models use 8 parallel attention heads.
- Feed-Forward Network (FFW): Following the MHA and layer normalization, a feed-forward network is applied to each position independently. These are 1024-dimensional positional feed-forward networks.
- Layer Normalization and Residual Connections: Layer normalization and residual connections are used across Transformer layers and the overall architecture to improve gradient backpropagation [13, Figure 4]

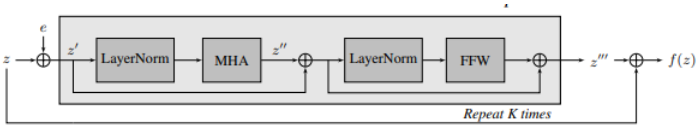

Figure 4

4.4 Encoder and Decoder

- Encoder: Takes the time-domain mixture signal as input and learns an STFT-like representation using a single convolutional layer [Figure 3]. The stride factor of this convolution significantly impacts performance, speed, and memory. In the best models, the encoder uses 256 convolutional filters with a kernel size of 16 samples and a stride factor of 8 samples.[13]
- Decoder: Reconstructs the separated signals using a transposed convolution layer with the same stride and kernel size as the encoder. It performs element-wise multiplication between the predicted mask for each source and the encoder output. For WSJ0-3mix, the decoder has three outputs.[13]

4.5 Training Details

- Datasets: The model is trained and evaluated on the WSJ0-2mix and WSJ0-3mix datasets. These datasets consist of mixtures of two and three speakers, respectively, created by randomly mixing utterances from the WSJ0 corpus.[20]
- Chunking: The masking network processes chunks of size C = 250 with a 50% overlap.

- Optimizer: The Adam algorithm is used with a learning rate of 15e−5.
- Loss Function: Scale-invariant signal-to-noise Ratio (SI-SNR) via utterance-level permutation invariant loss is employed, with clipping at 30dB.
- Data Augmentation: Dynamic mixing (DM), which involves on-the-fly creation of new mixtures, is used. This is expanded by applying speed perturbation to sources before mixing them. Dynamic mixing significantly helps performance.[18]
- Efficiency: Automatic mixed-precision is used to speed up training.

5. Conclusion

The SepFormer successfully demonstrates that state-of-the-art performance in multi-speaker speech separation can be achieved with an entirely RNN-free, Transformer-based model. By leveraging a multi-scale dual-path architecture in which recurrent connections are replaced by multi-head attention, SepFormer effectively captures both short- and long-term dependencies.

This innovation enables parallelization of computations across different time steps, resulting in significantly faster training and inference times, as well as a substantial reduction in memory usage compared to leading RNN-based models such as DPRNN, DPTNet, and Wavesplit. Moreover, SepFormer achieves competitive performance even when subsampling the encoded representation by a factor of 8, further enhancing its efficiency.

These findings firmly establish the efficacy and efficiency of Transformer-based architectures for advanced multi-speaker speech separation tasks.

References

[1] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 6645–6649.

[2] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, “Recurrent neural network based language model,” *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010, pp. 1045–1048.

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[4] K. Cho et al., “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734.

[5] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” *31st International Conference on Machine Learning (ICML)*, Beijing, China, 2014, pp. 1764–1772.

[6] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 31–35.

[7] Y. Luo and N. Mesgarani, “TasNet: Time-domain audio separation network for real-time, single-channel speech separation,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 696–700.

[8] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia, 2015, pp. 4580–4584.

[9] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 6645–6649.

[10] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 46–50.

[11] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[12] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “Separating mixtures of long sequences by recurrent neural networks and its application to speech separation,” *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[13] J. Subakan et al., “Attention is all you need in speech separation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 180–192, Jan. 2023.

[14] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1562–1566.

[15] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 46–50.

[16] J. Subakan et al., “Attention is all you need in speech separation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 180–192, Jan. 2023.

[17] K. Li, Z. Luo, and J. Han, “Efficient speech separation with lightweight neural architectures: A survey,” *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Incheon, Korea, 2022, pp. 1201–1205.

[18] N. Zeghidour and D. Grangier, “Wavesplit: End-to-end speech separation by speaker clustering,” arXiv preprint arXiv:2002.08933, 2020.

[19] parallelization J. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 180–192, Jan. 2023.

[20] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. of ICASSP, 2016, pp. 31–35.