Research on models for speech separation

Timofei Senin
Faculty Of Computer Science, HSE, Moscow
tpsenin@edu.hse.ru

Alexander Matosyan
Faculty Of Computer Science, HSE, Moscow
aamatosyan@edu.hse.ru

Last Edit Date: November 22, 2024

***Abstract*—This paper presents a comparative analysis of two state-of-the-art approaches in blind speech separation: Conv-TasNet and Dual-Path Recurrent Neural Network (DPRNN). Both methods represent significant advances in time-domain speech separation. Conv-TasNet introduces a fully-convolutional architecture with a temporal convolutional network (TCN), utilizing an encoder-decoder framework optimized for speaker separation. DPRNN offers an alternative approach by implementing a novel dual-path structure that efficiently handles long sequential inputs through intra- and inter-chunk processing. We examine the architectural differences and performance characteristics of both methods. Our analysis demonstrates that Conv-TasNet provides better speech separation quality than DPRNN in our training scenario. This comparison provides valuable insights into the evolution of speech separation architectures and their practical applications in real-world scenarios. The findings suggest that both methods have distinct advantages, with Conv-TasNet being more suitable for real-time applications.**

Index Terms—Speech separation, deep learning, time-domain processing, Conv-TasNet, DPRNN, single-channel separation

I. Introduction

Speech separation, also known as the ”cocktail party problem,” remains one of the most challenging tasks in audio signal processing. The ability to separate individual speakers from a mixed audio signal has numerous practical applications, ranging from speech recognition systems to telecommunications and audio production. Some approaches to this problem have primarily relied on time-frequency domain representations, such as spectrograms, which have inherent limitations including phase reconstruction issues and suboptimal representation for separation tasks.

Recent advances in deep learning have revolutionized the field of speech separation, particularly with the emergence of time-domain approaches. These methods process the raw waveform directly, avoiding the limitations of time-frequency representations and potentially achieving better separation quality. Two notable architectures that have demonstrated significant improvements in this domain are Conv-TasNet and DPRNN.

Conv-TasNet (1) introduced a novel approach using a fully-convolutional architecture with a learnable encoder-decoder framework. This method achieved unprecedented performance in speaker separation tasks while maintaining relatively low latency and only 5 million parameters, making it suitable for real-time applications. The system's temporal convolutional network (TCN) effectively captures long-term dependencies in the speech signal while keeping the model size managable.

DPRNN (2), on the other hand, addressed the challenges of processing long sequential inputs by introducing a unique dual-path structure. This architecture efficiently handles long sequences by splitting them into smaller chunks and processing them through intra- and inter-chunk operations. The result is also a more parameter-efficient model.

In this paper, we analyze and compare these two influential methods, examining their architectural designs, computational requirements, and performance characteristics.

We also provide training code and resulting metrics for both architectures. Everything can be obtained in our public GitHub repo. [1]

II. Related Work

The field of speech separation has witnessed significant evolution with the advent of deep learning approaches. This section reviews key developments in neural network architectures for speech separation, focusing on convolutional, recurrent, and transformer-based approaches.

Convolutional Neural Networks (CNNs)

Early deep learning approaches for speech separation primarily utilized CNNs operating in the time-frequency domain. Deep Attractor Source Separation (3) introduced the concept of learned attractors in the embedding space. A significant breakthrough came with Conv-TasNet (1), which moved processing to the time domain using a fully-convolutional architecture. This was followed by improvements like Gated Conv-TasNet (4), which incorporated gating mechanisms to enhance separation quality. These CNN-based approaches demonstrated superior performance while maintaining reasonable computational efficiency.

Recurrent Neural Networks (RNNs)

[1] https://github.com/Timofon/speech_separation

RNNs have been particularly effective in modeling dependencies in speech signals. Permutation Invariant Training with RNN (5) was one of the first successful applications of RNNs to speech separation. DPRNN (2) introduced a novel dual-path structure that efficiently handled long sequences.

Transformer-based Approaches

More recently, transformer architectures have shown promising results in speech separation. SepFormer (6) adapted the transformer architecture specifically for separation tasks, leveraging self-attention mechanisms to capture both local and global dependencies. Dual-path transformer networks (7) combined DPRNN's chunk-based processing with transformer's attention mechanism. More recently, Dual-Path Mamba (8) introduced selective state spaces instead of self-attention, achieving comparable performance with reduced computational cost.

III. Methodology

A. *ConvTasNet*

Conv-TasNet is a time-domain speech separation model that consists of three main components, as shown in Figure 1.

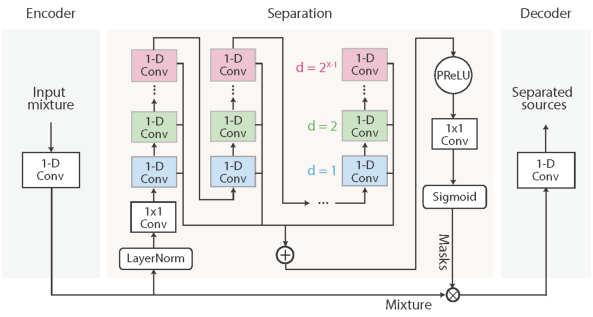

Fig. 1. Architecture of Conv-TasNet

Encoder

The encoder transforms the time-domain signal into a higher-dimensional representation using a 1-D convolution operation. The encoded representation is then normalized using Layer Normalization.

Separator

Separator consists of multiple repeating modules each of several 1-D conv blocks. Inside one module there are n 1-D conv blocks. The outputs inside each module are summed up and also outputs of all modules are summed up.

1-D conv block

The one-dimensional convolutional block 2 serves as the fundamental processing unit in the Conv-TasNet encoder, operating directly on audio signals in the time domain. The block architecture consists of three sequential components: a convolutional layer, layer normalization, and parametric ReLU activation. The convolutional layer applies learnable filters that slide across the input signal with a specified kernel size and stride, effectively extracting local temporal features from the audio. Layer normalization stabilizes the training process by maintaining zero mean and unit variance across the features. PReLU activation introduces nonlinearity.

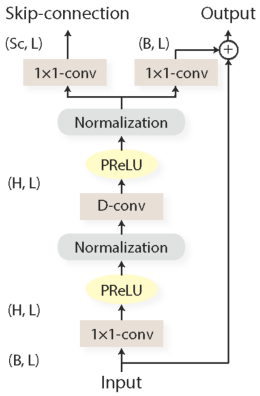

Fig. 2. Architecture of 1-D conv block

It is worth noting that each 1x1-conv has dilation and stride parameter equal to two in the power of the order of the block in the module. This method increases receptive field of a model.

The separator module generates C masks corresponding to C target sources, with each mask having the same dimensions as the encoded mixture. These masks, obtained through the TCN network and sigmoid activation, represent source-specific patterns in the latent space. The encoded mixture is then element-wise multiplied with each mask to produce C separated representations, which are subsequently fed into the decoder for time-domain signal reconstruction.

Decoder

The decoder reconstructs the separated speech signals through transposed convolution operations. It projects the separated features back to the time domain using the same number of basis functions as the encoder. The output consists of C channels, corresponding to the number of speakers to be separated.

B. *DualPathRNN*

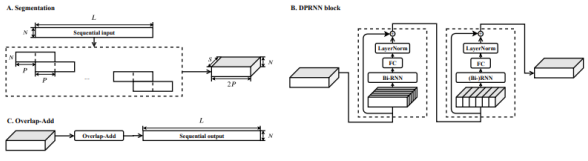

Fig. 3. Architecture of dprnn

The paper (2) introduces a novel dual-path recurrent neural network (DPRNN) architecture designed for efficient processing of extremely long sequences. DPRNN employs a strategy of dividing the input sequence into overlapping segments, followed by alternating and iterative processing within each segment (local) and across segments (global) using two distinct RNNs. DPRNN can be applied to any systems that require longterm sequential modeling, including time-domain audio separation systems. DPRNN itself consists of three stages, as

shown in 3: (1) Segmentation, (2) DPRNN-block processing, (3) Overlap-add. But to solve the source separation task few more parts are needed: (4) Encoder, (5) Masker and (6) Decoder.

Segmentation

The segmentation stage divides the input sequence into chunks, allowing for overlap, and arranges these chunks into a three-dimensional tensor.

DPRNN-block processing

A DPRNN block comprises two RNNs with distinct recurrent connection patterns. Initially, a bidirectional RNN processes each chunk independently to extract local features. Subsequently, RNN processes these chunks collectively to capture global dependencies between them. Multiple blocks can be stacked to increase network depth.

Overlap-add

The three-dimensional output of the final DPRNN block is transformed back into a sequential representation using an overlap-add operation on the chunks.

Encoder

The encoder converts the input waveform into a front-end sequential representation. This block is the same as in the ConvTasNet model. The details can be seen in the ConvTasNet section.

Masker

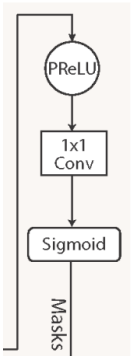

Fig. 4. Architecture of masker

Masker converts the output of overlap-add operation into two same shape sequential representations for two speakers. Two different masking methods were tried: as shown in Figure 4 there is a PReLU activation layer followed by a Conv1D layer which doubles the number of channels and a Sigmoid activation layer at the end to get the masks as shown in Figure 5 there are two branches with an activation layer followed by a Conv1D layer each, outputs of the branches are then connected by an element-wise product operation the result of which is forwarded through a Sigmoid activation layer.

We then multiply the ouput of the encoder layer by the masks and pass the result into the decoder layer.

Decoder The decoder converts the masked sequential representations back into two waveforms. This block is the same as in the ConvTasNet model. The details can be seen in the ConvTasNet section.

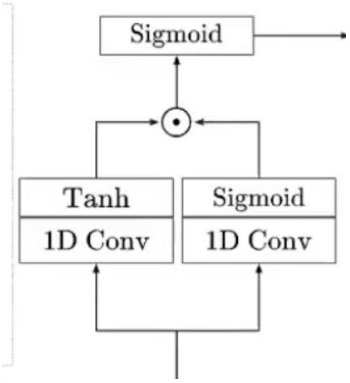

Fig. 5. Architecture of masker

IV. Experimental Setup

For our experiments, we used a custom dataset provided by the teaching team, which contains mixed audio signals and their corresponding source components. The dataset consists of 20,000 training samples and 5,000 validation samples, with all audio signals sampled at 16 kHz and having a duration of 2 seconds.

All experiments were conducted on a single NVIDIA V100 GPU with 32GB of memory. For Conv-TasNet, we trained the model using Adam optimizer with an initial learning rate of $1e-3$ and ReduceLROnPlateau scheduler, using a batch size of 16. For the DPRNN model, we employed Adam optimizer with an initial learning rate of $1e-3$, combined with StepLR scheduler that applied a decay rate of 0.98 every 2 epochs, and used a larger batch size of 64. Also we clipped gradients to the maximum value of 5.

For both models, we used Scale-Invariant Signal-to-Distortion Ratio improvement (SI-SDRi) as our training objective and evaluation metric. SI-SDRi measures the quality of separation by comparing the separated sources with their corresponding ground truth.

V. Experiments

A. ConvTasNet

1) Basic training: In our best experiment we achieved $loss = -12.92$ which can be seen on 1. SI-SDRi was 8.61.

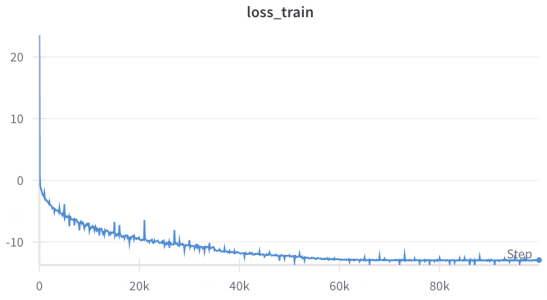

Fig. 6. Main training loss

Looking at the graph 7, we can see that the training is actually quite noisy. The gradient norm graph can serve as a confirmation of this fact.

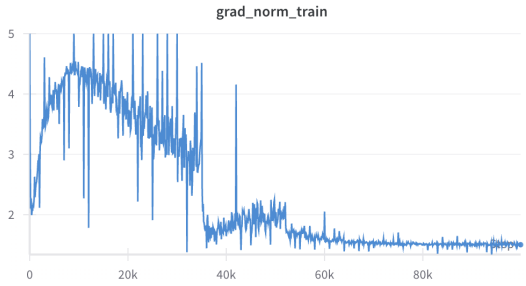

Fig. 7. Gradient norms

2) Training using SI-SNRi loss: In this experiment we tried to train our model on SI-SNRi loss. By validation loss 8 we can see that training is very unstable so we decided to stick to SI-SDRi loss.

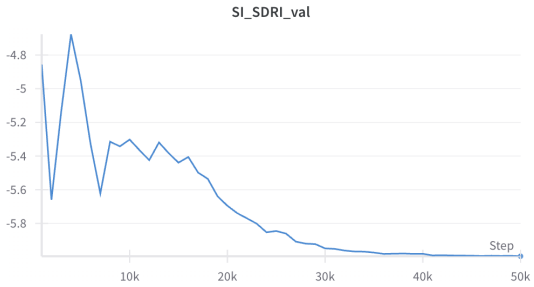

Fig. 8. SI-SNRi loss

3) Training with augmented target: In this expriment we decided to apply gain augmentation to target audios. Surprisingly, training with this approach increased the final metric by 0.2 9

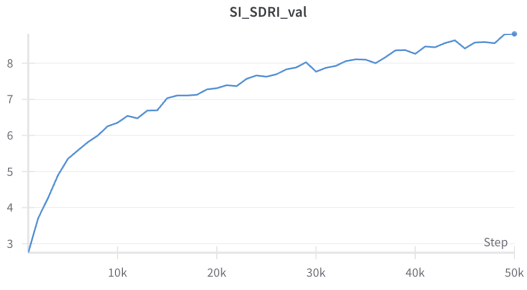

Fig. 9. SI-SDRi metric

B. DualPathRNN

1) First full training: For the first training we chose to use the first method of masking in the model architecture. The best SI-SDRI 10 on validation was 6.98.

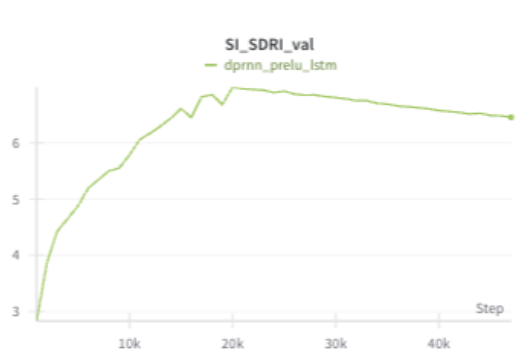

Fig. 10. SI-SDRi metric

2) Full training with another masker: Then we decided to try a more complex masking method. The model lost the ability to learn 11 as the loss immediately reached a plateau near zero. It seems that the model got too complex because of many activation layers.

Fig. 11. SI-SDRi metric

3) Fine-tuning with mix-audio augmentations: We decided to take our best model and fine-tune it on our dataset with augmentations of mixed audios. We used the Gain augmentation with probability of application equal to 0.52. The best SI-SDRI on validation 12 before reaching plateau was 7.77. As it then occurred the resulting model of this experiment was our best model based on DPRNN.

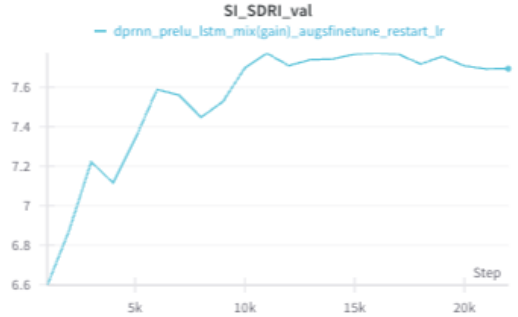

Fig. 12. SI-SDRi metric

4) Fine-tuning with target augmentations: This experiment 13 is very similar to the last one. The only change is that we apply the Gain augmentation to the target audios. The best SI-SDRI on validation before reaching plateau was 7.49. That means that the model from the third experiment was our best.

Fig. 13. SI-SDRi metric

VI. Results

We evaluated both Conv-TasNet and DPRNN models on our speech separation task. The comparison of model performance and complexity is presented in Table I.

Model	SI-SDRi	Parameters
Conv-TasNet	8.86	5M
DPRNN	7.70	10M

TABLE I
Performance comparison of speech separation models

As shown in Table I, Conv-TasNet achieves better separation performance with SI-SDRi of 8.86 while having fewer parameters (5M) compared to DPRNN. Despite having twice as many parameters (10M), DPRNN shows lower separation quality with SI-SDRi of 7.70 dB. This suggests that the temporal convolutional architecture of Conv-TasNet is more efficient in capturing the necessary features for speech separation in our experimental setup.

During our experiments, we observed that DPRNN exhibited rapid overfitting tendencies. This behavior might be attributed to the fact that DPRNN was originally designed and optimized for a different dataset.

VII. Conclusion

In this work, we conducted a comparative study of two prominent speech separation architectures: Conv-TasNet and DPRNN. Our experiments were performed on a dataset comprising 20,000 training samples and 5,000 validation samples, with audio signals sampled at 16 kHz and 2-second duration.

The results demonstrate that Conv-TasNet outperforms DPRNN in terms of both separation quality and model efficiency. With only 5M parameters, Conv-TasNet achieved an SI-SDRi of 8.86 dB, while DPRNN, despite having 10M parameters, reached an SI-SDRi of 7.70 dB. Furthermore, we observed that DPRNN showed a tendency to overfit quickly, possibly due to its architecture being originally optimized for a different dataset.

These findings suggest that architectural efficiency might be more crucial than model capacity for speech separation tasks. Conv-TasNet's superior performance with fewer parameters indicates that its temporal convolutional design might be better suited for capturing the relevant features in speech separation problems.

Future work could explore ways to adapt DPRNN's architecture to be more generalizable across different datasets or investigate hybrid approaches that combine the strengths of both architectures.

References

[1] 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, 2019.

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

[3] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 246–250.

[4] K. Tan and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," in *INTERSPEECH 2019*, 2019, pp. 3663–3667.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[6] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.

[7] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *INTERSPEECH 2020*, 2020, pp. 2642–2646.

[8] C. Wang, W. Shuai, Z. Yang, and D. Wang, "Dual-path mamba for speech separation," *arXiv preprint arXiv:2401.05455*, 2024.

TABLE II
EVALUATION CRITERIA FOR EACH SECTION.

Section	Goal	Evaluation		
		Good	Ok	Not Good
Title	To give content information to reader.	Engaging.	Appropriate.	Not enough content information or too much.
Abstract	To concisely summarize the experimental question, general methods, major findings, and implications of the experiments in relation to what is known or expected	Key information is presented completely and in a clear, concise way. All information is correct. Organization is logical. Captures any reader's interest	Sufficient information is presented in proper format. Would benefit from some reorganization. Understandable with some prior knowledge of experiment.	Some key information is omitted or tangential information is included. Some information is misrepresented. Some implications are omitted. Incorrect format is used.
Introduction	To identify central experimental questions, and appropriate background information. To present a plausible hypothesis and a means of testing it.	Relevant background information is presented in balanced, engaging way. Your experimental goals and predictions are clear and seem a logical extension of existing knowledge. Writing is easy to read. All background information is correctly referenced.	Relevant background information is presented but could benefit from reorganization. Your experiment is well described and a plausible hypothesis is given. With some effort, reader can connect your experiments to background information. Writing is understandable. Background information is correctly referenced.	Background information is too general, too specific, missing and/or misrepresented. Experimental question is incorrectly or not identified. No plausible hypothesis is given. Writing style is not clear, correct or concise. References are not given or properly formatted
Related Work	To provide information about state-of-the-art solutions, baselines, and useful/required algorithms/techniques from the literature	All baselines are presented in a brief format. Required background algorithms/techniques are explained. All citations are provided and correct. Only useful information is written.	All baselines and algorithms/techniques are explained but have unnecessary too deep details.	Not relevant/not used baselines. None or missing references. Rephrasing of original paper instead of brief summary.
Methodology and Experimental Setup	To describe procedures correctly, clearly, and succinctly. Included a correctly formatted citation of the lab manual.	Sufficient for another researcher to repeat your experiment. Steps presented.	Procedures could be pieced together with some effort. Steps presented.	Procedures incorrectly or unclearly described or omitted. Steps not presented.
Results	To present your data using text AND figures/tables.	Text tells story of your major findings in logical and engaging way. Figures and tables are formatted for maximum clarity and ease of interpretation. All figures and tables have numbers, titles and legends that are easy for the reader to follow.	Text presents data but could benefit from reorganization or editing to make story easier for reader. Text includes interpretation of results that is better suited for discussion section. Figures and tables are formatted to be clear and interpretable. All figures and tables have numbers, titles and legends.	Text omits key findings, inaccurately describes data, or includes irrelevant information. Text difficult to read due to style or mechanics of writing. Text difficult to read due to logic or organization. Figures and tables missing information, improperly formatted or poorly designed. Figures and tables have inadequate or missing titles or legends.
Discussion	To evaluate meaning and importance of major findings.	Appropriate conclusions drawn from findings. Connections made between experimental findings. Connections made between findings and background information. Future directions considered. Writing is compelling.	Appropriate conclusions drawn from findings. Experimental limitations considered. Writing is clear.	Conclusions omitted, incorrectly drawn or not related to hypothesis. Relationship between experimental findings and background information is missing or incorrectly drawn. Writing style and mechanics make argument difficult to follow.
References	To give credit work on which your own is based.	Complete list of reliable sources, including peer-reviewed journal article(s). Properly formatted in body of report and in reference section.	Adequate list or reliable sources. With minor exceptions, properly formatted in body of report and in reference section.	List is incomplete or includes sources not cited in body of report. List includes inappropriate sources. List not properly formatted. References not properly cited in body of report.

TABLE III
EVALUATION OF WRITING STYLE

Writing Style and Mechanics	Evaluation	
	Good	Not Good
Verb Voice	Appropriate for audience. Consistent passive or active voice.	Too simple or too advanced. Irregular use of passive and active voice.
Word choice	Concise. Says what you mean. Scientific vocabulary used correctly.	Verbose. Ambiguous or incorrect. Scientific vocabulary misused.
Fluency	Sentences and paragraphs are well structured. Punctuation is correct or has only minor errors. Grammar is correct or has minor errors. Spelling is correct.	Sentences are repetitive or awkward. Paragraphs are not logical. Periods, commas, colons, and semicolons are misused. Significant number of run-on sentences, sentence fragments, misplaced modifiers, subject/verb disagreements. Significant number of spelling errors.
Scientific format	Past tense for describing new findings. Present tense used for accepted scientific knowledge and figure legends. All sections are included and properly formatted. Formal language	Misleading verb tenses. Some sections are missing. Figures miss legends. References are not properly formatted. Informal language: contractions, slang, etc.