

TF-LOCOFORMER: TRANSFORMER WITH LOCAL MODELING BY CONVOLUTION FOR SPEECH SEPARATION AND ENHANCEMENT

Saijo, Kohei; Wichern, Gordon; Germain, François G; Pan, Zexu; Le Roux, Jonathan

TR2024-126 September 11, 2024

Abstract

Time-frequency (TF) domain dual-path models achieve high-fidelity speech separation. While some previous state-of-the-art (SoTA) models rely on RNNs, this reliance means they lack the parallelizability, scalability, and versatility of Transformer blocks. Given the wide-ranging success of pure Transformer-based architectures in other fields, in this work we focus on removing the RNN from TF-domain dual-path models, while maintaining SoTA performance. This work presents TF-LoCoformer, a Transformer-based model with Local-modeling by Convolution. The model uses feed-forward networks (FFNs) with convolution layers, instead of linear layers, to capture local information, letting the self-attention focus on capturing global patterns. We place two such FFNs before and after self-attention to enhance the local-modeling capability. We also introduce a novel normalization for TF-domain dual-path models. Experiments on separation and enhancement datasets show that the proposed model meets or exceeds SoTA in multiple benchmarks with an RNN-free architecture.

International Workshop on Acoustic Signal Enhancement (IWAENC) 2024

TF-LOCOFORMER: TRANSFORMER WITH LOCAL MODELING BY CONVOLUTION FOR SPEECH SEPARATION AND ENHANCEMENT

Kohei Saijo^{1,2*}, Gordon Wichern¹, François G. Germain¹, Zexu Pan¹, Jonathan Le Roux¹

¹Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA ²Waseda University, Japan

ABSTRACT

Time-frequency (TF) domain dual-path models achieve high-fidelity speech separation. While some previous state-of-the-art (SoTA) models rely on RNNs, this reliance means they lack the parallelizability, scalability, and versatility of Transformer blocks. Given the wide-ranging success of pure Transformer-based architectures in other fields, in this work we focus on removing the RNN from TF-domain dual-path models, while maintaining SoTA performance. This work presents TF-Locoformer, a Transformer-based model with Local-modeling by COnvolution. The model uses feed-forward networks (FFNs) with convolution layers, instead of linear layers, to capture local information, letting the self-attention focus on capturing global patterns. We place two such FFNs before and after self-attention to enhance the local-modeling capability. We also introduce a novel normalization for TF-domain dual-path models. Experiments on separation and enhancement datasets show that the proposed model meets or exceeds SoTA in multiple benchmarks with an RNN-free architecture.

Index Terms— speech separation, self-attention, convolution

1. INTRODUCTION

The past decade has witnessed dramatic progress in speech separation thanks to advancements in neural networks (NNs). Early work such as deep clustering [1] trains NNs to estimate time-frequency (TF) masks in the TF-magnitude domain. In contrast, time-domain audio separation network (TasNet)-style NNs have improved separation performance drastically by introducing learnable encoders and decoders [2]. Dual-path modeling, which chunks the features and conducts local and global modeling alternately for efficient sequence modeling, is now one of the mainstream approaches in time-domain end-to-end (E2E) networks [3–5]. More recently, dual-path modeling in the TF domain, where temporal and frequency modeling are done alternately, has shown impressive performance improvements [6, 7]. TF-domain models have the potential to perform better in realistic reverberant conditions, as the FFT window size is usually much longer than the kernel size of a typical learnable encoder [8].

While Transformer-based architectures [9] have shown great success in time-domain E2E separation [5, 10, 11], the current state-of-the-art (SoTA) TF-domain dual-path separation model relies on RNNs [7]. Although RNNs have some advantages such as smaller memory cost in inference, the training is typically very time-consuming since it cannot be parallelized. In contrast, Transformer-based models have been eagerly investigated in other fields [9, 12, 13] because of their multiple advantages: the training process can be parallelized, models can accept prompts (e.g., to specify the task [14, 15]), and they may scale well [16–18]; these features are either unattainable or at least unconfirmed in RNNs. Before

*This work was done during an internship at MERL.

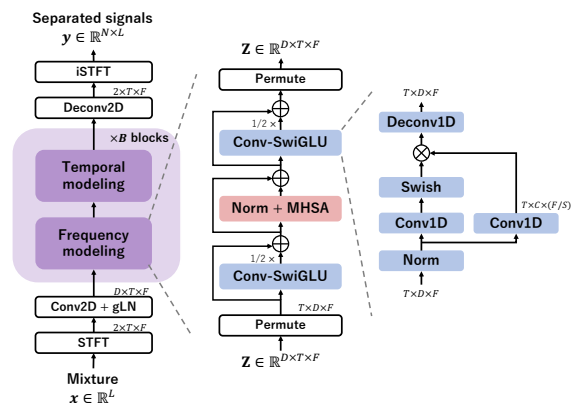


Fig. 1. Overview of the proposed TF-Locoformer. The temporal modeling block is the same as the frequency modeling block with a permutation of the time and frequency dimensions.

realizing these potential benefits of Transformer-based models, the first step, which is the goal of this paper, is to investigate whether we can obtain comparable or better performance as RNN-based SoTA models with an RNN-free model of similar complexity.

Global and local modeling often both play an important role in speech processing [19]. There are strong hints in the literature that this is also true for speech separation, and that Transformer blocks lack an intrinsic local modeling ability. Indeed, Transformer blocks, which excel at global modeling thanks to self-attention, have worked well in the context of time-domain dual-path models [5], where self-attention in the intra-chunk path is limited to local modeling by construction, while failing to compete in TF-domain dual-path architectures, where local modeling is not explicitly enforced. On the other hand, RNNs, which can capture local information, do attain strong performance in TF-domain dual-path models. Notably, the current SoTA model in TF-domain separation, TF-GridNet [7], exploits both RNNs and self-attention. As an alternative to RNNs for capturing local information, we consider inserting convolution layers within a Transformer-based dual-path architecture.

We propose TF-Locoformer (TF-domain Transformer with Local modeling by COnvolution), a simple extension of the Transformer-based model that alternates global and local modeling. Starting from the normal Transformer block, we first replace the two linear layers in the feed-forward network (FFN) with 1d-convolution and 1d-deconvolution layers, respectively. In addition, we leverage swish gated linear unit (SwiGLU) activations in the FFNs and place such FFNs both before and after self-attention, inspired by the success of the *macaron-style* architecture [19, 20]. To further improve the performance, we introduce a novel normalization layer for TF-domain dual-path models. We empirically demonstrate that these extensions boost the local-modeling capability and enable TF-Locoformer to

achieve comparable or better performance than the current state-of-the-art RNN-based models. Our source code is available online¹.

2. TF-LOCOFORMER

2.1. Overview of TF-Locoformer

Figure 1 shows the overview of the proposed TF-Locoformer model. The model is based on TF-domain dual-path modeling [6, 7], where frequency and temporal modeling are done alternately. It separates each source by complex spectral mapping, where their real and imaginary (RI) components are directly estimated.

Let us denote a monaural mixture of N speech signals $\mathbf{s} \in \mathbb{R}^{N \times L}$ and a noise signal $\mathbf{b} \in \mathbb{R}^L$ as $\mathbf{x} = \sum_n \mathbf{s}_n + \mathbf{b} \in \mathbb{R}^L$, where L is the number of samples in the time domain and $n = 1, \dots, N$ is the speech source index. In short-time Fourier transform (STFT) domain, the mixture is written as $\mathbf{X} \in \mathbb{R}^{2 \times T \times F}$, where T and F are the number of frames and frequency bins, and 2 corresponds to real and imaginary parts.

The input \mathbf{X} is first encoded into an initial feature \mathbf{Z} with feature dimension D :

$$\mathbf{Z} = \text{gLN}(\text{Conv2D}(\mathbf{X})) \in \mathbb{R}^{D \times T \times F}, \quad (1)$$

where gLN is global layer normalization [2]. For frequency modeling, we view the feature \mathbf{Z} as a stack of T arrays of shape $D \times F$, where D is the feature dimension and F the sequence length (i.e., we permute the dimension order of \mathbf{Z} to $T \times D \times F$). Frequency modeling is then performed as:

$$\mathbf{Z} \leftarrow \mathbf{Z} + \text{ConvSwiGLU}(\mathbf{Z})/2, \quad (2)$$

$$\mathbf{Z} \leftarrow \mathbf{Z} + \text{MHSA}(\text{Norm}(\mathbf{Z})), \quad (3)$$

$$\mathbf{Z} \leftarrow \mathbf{Z} + \text{ConvSwiGLU}(\mathbf{Z})/2, \quad (4)$$

where MHSA stands for multi-head self-attention [9]. MHSA has H heads and each head processes D/H -dimensional feature. We use rotary positional encoding [21] for encoding the relative position of each frequency bin. The ConvSwiGLU module and the Norm layer will be described in Sections 2.2 and 2.3, respectively. Temporal modeling is done in the same way by viewing the feature \mathbf{Z} as a stack of F arrays of shape $D \times T$, where T is the sequence length (i.e., we permute the dimension order of \mathbf{Z} to $F \times D \times T$). After alternating between frequency and temporal modeling B times, the final feature \mathbf{Z} is used to estimate the RI components of N sources:

$$\hat{\mathbf{S}} = \text{DeConv2D}(\mathbf{Z}) \in \mathbb{R}^{2 \times N \times T \times F}. \quad (5)$$

Finally, we obtain the time-domain signals by inverse STFT.

2.2. ConvSwiGLU module

RNN-based models have shown strong performance in TF-domain dual-path models [7], which we attribute to the inherent ability of their hidden state updates to better capture local information. In contrast, it is expected to be hard for the MHSA to always have locally smooth attention weights. This consideration motivates us to design a model that has strong local-modeling capability.

To this end, we introduce a new FFN, named ConvSwiGLU. ConvSwiGLU boosts the local-modeling capability by utilizing 1d-convolution and 1d-deconvolution layers instead of linear layers. We

Table 1. Summary of hyper-parameter notations and default model configurations for three model sizes: Small (S), Medium (M), and Large (L).

Symbol	Description	S	M	L
D	Embedding dimension of each TF bin	96	128	128
B	Number of Locoformer blocks	4	6	9
C	Hidden dimension in Conv-SwiGLU	256	384	384
K	Kernel size in Conv1D and Deconv1D	4	4	4
S	Stride in Conv1D and Deconv1D	1	1	1
H	Number of heads in self-attention	4	4	4
G	Number of groups in RMSGroupNorm	4	4	4
-	Number of parameters [M]	5.0	15.0	22.5

also exploit the SwiGLU activation, which has shown better performance than the Swish activation in the NLP field [22]. Formally, its processing is written as:

$$\mathbf{Z} \leftarrow \text{Norm}(\mathbf{Z}), \quad (6)$$

$$\mathbf{Z} \leftarrow \text{Swish}(\text{Conv1D}(\mathbf{Z})) \otimes \text{Conv1D}(\mathbf{Z}), \quad (7)$$

$$\mathbf{Z} \leftarrow \text{Deconv1D}(\mathbf{Z}). \quad (8)$$

As described in Section 2.1, we place this ConvSwiGLU both before and after the self-attention, which enables stronger local modeling.

2.3. Group normalization for TF-domain dual-path models

Typically, TF-domain dual-path models utilize layer normalization or root mean square normalization (RMSNorm) [23] as Norm layer, to normalize the D -dimensional vector of each TF bin. However, since the goal is the separation of multiple speakers, we believe that it may be beneficial to encourage each D -dimensional vector $\mathbf{Z}_{t,f}$ to be split into groups corresponding to disentangled concepts such as speaker IDs.

In RMSGroupNorm, we view each D -dimensional vector $\mathbf{Z}_{t,f}$ as a stack of G vectors of dimension D/G , where G is the group size, and we normalize each D/G -dimensional vector separately. This encourages the model to disentangle each D -dimensional vector into some groups, which could be helpful for speech separation. Note that we normalize each TF bin, unlike the group normalization in image processing [24]. As in RMSNorm, RMSGroupNorm features an affine transform with two D -dimensional learnable parameters. Note that $G = 1$ corresponds to the original RMSNorm. In experiments, we demonstrate that RMSGroupNorm gives slightly but consistently better performance than RMSNorm in our model.

3. EXPERIMENTS

3.1. Dataset and experimental setup

To evaluate the model, we used three speech separation corpora, WSJ0-2mix [1], Libri2mix [25], and WHAMR! [26], and a speech enhancement corpus, the Interspeech DNS-challenge 2020 dataset (denoted as DNS) [27]. In the speech separation datasets, we always used the fully overlapped `min` version with a sampling rate of 8 kHz, while the sampling rate of the DNS dataset was 16 kHz.

WSJ0-2mix contains two-speaker mixtures of utterances from the WSJ0 corpus. The total lengths of the training, validation, and test sets are 30 h, 10 h, and 5 h, respectively.

Libri2Mix contains two-speaker mixtures of utterances from Librispeech [28]. The total lengths of the training, validation, and test sets are 212 h, 11 h, and 11 h, respectively.

¹<https://github.com/merlresearch/tf-locoformer>

Table 2. Comparison with previous models on WSJ0-2mix. Methods with * use speed perturbation when doing dynamic mixing. “-” denotes unavailable result in original work. Results in [dB].

System	Domain	#params[M]	w/o DM		w/ DM	
			SI-SNRi	SDRi	SI-SNRi	SDRi
DPRNN [3]	T	2.6	18.8	19.0	-	-
DPTNet [4]	T	2.7	20.2	20.6	-	-
Wavesplit [37]	T	29.0	21.0	21.2	22.2	22.3
SepFormer* [5]	T	25.7	20.4	20.5	22.3	22.4
TFPSNet [6]	TF	2.7	21.1	21.3	-	-
QDPN* [10]	T	200.0	22.1	-	23.6	-
TF-GridNet [7]	TF	14.4	23.5	23.6	-	-
MossFormer2* [11]	T	55.7	-	-	24.1	-
SepTDA ₂ [38]	T	21.2	24.0	23.9	-	-
TF-LoCoformer (S)	TF	5.0	22.0	22.1	22.8	23.0
TF-LoCoformer (M)	TF	15.0	23.6	23.8	24.6	24.7
TF-LoCoformer (L)	TF	22.5	24.2	24.3	25.1	25.2

WHAMR! is a noisy reverberant version of WSJ0-2mix. The model is trained to jointly perform dereverberation, denoising, and separation.

DNS has 2700 h of training data and 300 h of validation data, which are simulated using the official script [27]. The non-blind anechoic test set is used for testing.

All experiments are done using the ESPnet-SE pipeline [29]. A summary of hyper-parameters and the default model configurations are shown in Table 1. We mainly investigate the Medium model. On some datasets, we also evaluate the Small and Large models to show fairer comparisons with previous models. Only on WHAMR!, which has strong reverberation, we set $K = 8$ and halve C , mimicking TF-GridNet. We set the STFT window and hop sizes to 16 ms and 8 ms for all datasets, except for WHAMR! where the window size is set to 32 ms. We use the AdamW optimizer [30] with a weight decay of $1e-2$. We first linearly increase the learning rate from 0 to $1e-3$ over the first 4000 training steps. The learning rate is then decayed by 0.5 if the validation loss does not improve for 3 epochs. The training is conducted for up to 150 epochs² with early stopping if the validation loss does not improve for 10 epochs. When using dynamic mixing (DM), training is done up to 200 epochs, and the learning-rate decay and early stopping are applied after 75 epochs in the Small and Medium models, and 65 epochs in the Large model. The batch size is 4 and input samples are 4-second long. The L_2 norm of the gradient is clipped to 5. Each input mixture is normalized by dividing it by its standard deviation.

We use permutation-invariant [1, 31] scale-invariant signal-to-noise ratio (SI-SNR) [32] as the loss function for speech separation. For speech enhancement, we used a time-domain L_1 loss plus a TF-domain multi-resolution L_1 loss [33]. We used four FFT window sizes {256, 512, 768, 1024} with 50% overlap.

In evaluation, we use as weights the averages over the five model checkpoints that had the best losses on the validation set. We use the following evaluation metrics: SI-SNR improvement (SI-SNRi), signal-to-distortion ratio improvement (SDRi) [34], short-time objective intelligibility (STOI) [35], and wide-band perceptual evaluation of speech quality (PESQ-WB) [36].

3.2. Anechoic speech separation and enhancement

Table 2 compares the performance of the proposed model with the results reported in the literature on the WSJ0-2mix dataset. First, we focus on the TF-domain models, TFPSNet and TF-GridNet, both of

²On the DNS dataset, we define 21,000 training steps as one epoch because the training data is very large.

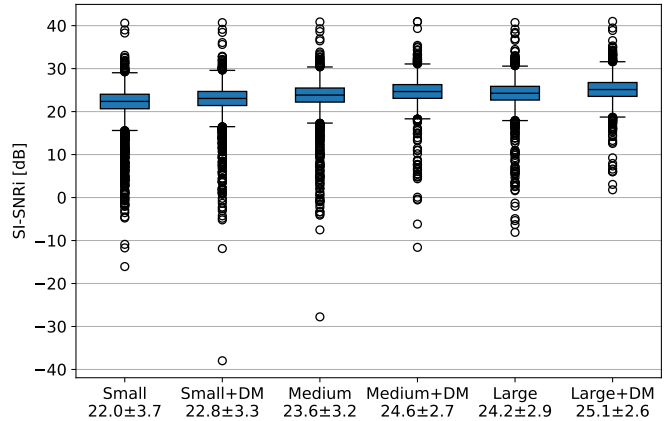


Fig. 2. Box-plots of SI-SNRi [dB] for models with different sizes on WSJ0-2mix test set. The numbers below the model size indicate average and standard deviations of SI-SNRi.

Table 3. Comparison with previous models on Libri2Mix. DM and speed perturbation were not used. Results in [dB].

System	Domain	#params[M]	SI-SNRi	SDRi
Conv-TasNet [2]	T	5.1	14.7	-
Wavesplit [37]	T	29.0	19.5	20.0
SepFormer [5]	T	25.7	19.2	19.4
MossFormer2 [11]	T	55.7	21.7	-
TF-LoCoformer (M)	TF	15.0	22.1	22.2

which have RNNs. We compare them with the Medium model because all of them have $B = 6$ dual-path blocks. The result demonstrates that TF-LoCoformer achieves comparable or better performance than RNN-based TF-domain models, implying that RNN-free models can work well in the TF-domain by introducing strong local modeling. Next, we compare TF-LoCoformer with the other models. Since the current SoTA model on WSJ0-2mix, SepTDA, has nine multi-path blocks, the Large model is suitable for this comparison. TF-LoCoformer again achieves comparable or better performance than previous SoTA models. We also trained the models using dynamic mixing and observed noticeable improvements.

Although it is often argued that the performance on WSJ0-2mix is saturated, we find that there is still room for improvement on samples which give low SI-SNRi. Figure 2 shows the boxplots of SI-SNRi given by the proposed models on the WSJ0-2mix test set. It can be observed that the larger models reduce the number of failures and work more robustly. Dynamic mixing results in a similar effect by augmenting the data.

On the larger-scale speech separation and enhancement datasets, TF-LoCoformer works well too. Table 3 and Table 4 show the performance of TF-LoCoformer and competing models on the Libri2Mix and DNS datasets, respectively. The proposed TF-LoCoformer gives the best performance on both datasets, which demonstrates its potential scalability. In addition, TF-LoCoformer is also effective for denoising.

3.3. Noisy reverberant speech separation

Table 5 compares the performance of the proposed model with previous models on the WHAMR! dataset. We evaluated the Small model to fairly compare with the current SoTA model, TF-GridNet, because it was configured to have $B = 4$ dual-path blocks on WHAMR!

Table 4. Comparison with previous non-causal models on the DNS2020 non-blind test dataset. SI-SNR results in [dB].

System	#params[M]	SI-SNR	STOI	PESQ-WB
Noisy	-	9.1	91.5	1.58
MFNet [39]	6.1	20.3	98.0	3.43
USES [15]	3.1	21.2	98.1	3.46
TF-Locoformer (M)	15.0	23.3	98.8	3.72

Table 5. Comparison with previous models on WHAMR!. DM* indicates DM with speed perturbation. Results in [dB].

System	Domain	#params[M]	SI-SNRi	SDRi
Wavesplit+DM [37]	T	29.0	13.2	12.2
SepFormer+DM* [5]	T	25.7	14.0	13.0
QDPN+DM* [10]	T	200.0	14.4	-
MossFormer2+DM* [11]	T	55.7	17.0	-
TF-GridNet [7]	TF	5.5	17.1	15.6
TF-Locoformer (S)	TF	5.0	17.4	15.9
TF-Locoformer (M)	TF	15.0	18.5	16.9

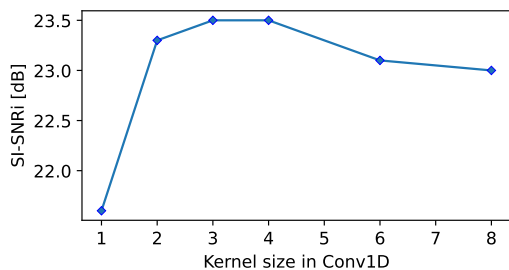


Fig. 3. Average SI-SNRi with different kernel sizes on WSJ0-2mix test set. Medium model is shown.

evaluation [7]. The result shows that the proposed model again outperforms the SoTA models. The Medium model scores even higher, implying that the proposed model could still achieve even better performance by using larger configurations.

The comparison not only shows that TF-Locoformer is effective, but also demonstrates that the current best TF-domain models perform better than the best available time-domain approaches in reverberant conditions. While some models such as QDPN [10] or MossFormer2 [11] gave comparable or better performance than the Small TF-Locoformer on WSJ0-2mix, the latter achieves better performance on WHAMR!, even without dynamic mixing. This observation is in line with [8], which argues that time-domain models struggle with reverberation due to their short kernel size in the encoder and decoder (e.g. 2 ms). The results clearly show the superiority of TF-domain models in reverberant conditions, which are more representative of real-world applications.

3.4. Ablation study

We conducted an ablation study to examine the effectiveness of each module in our model: convolution layers, ConvSwiGLU FFN, Macaron-style architecture, and RMSGroupNorm. Figure 3 shows the average SI-SNRi of the Medium models with different kernel sizes on the WSJ0-2mix test set. Since the default model configuration was $K = 4$ and $C = 384$, we set $C = 1536/K$ to make all the models roughly the same size. We used here the RMSNorm as the normalization layer to avoid potential influence of RMSGroup-

Table 6. Ablation study on WSJ0-2mix. Results in [dB].

System	SI-SDRi	SDRi
A0 TF-Locoformer (M)	23.6	23.8
A1 Macaron-style \rightarrow Single ConvSwiGLU	22.8	22.9
A2 SwiGLU \rightarrow Swish activation	22.2	22.4

Table 7. Comparison of normalization layers on WSJ0-2mix. Results in [dB].

System	RMSNorm		RMSGroupNorm	
	SI-SNRi	SDRi	SI-SNRi	SDRi
TF-Locoformer (S)	21.7	21.9	22.0	22.1
TF-Locoformer (M)	23.5	23.6	23.6	23.8
TF-Locoformer (L)	24.0	24.1	24.2	24.3

Norm’s grouping on the results. $K = 1$ is equivalent to a normal linear layer, in which case each block is almost the same as the pure Transformer block. The results clearly demonstrate the importance of the convolution layer: the models with $K \geq 2$ perform much better than that with $K = 1$. At the same time, models with longer kernels are more computationally efficient because the input to the Swish activation and the following gating have smaller hidden dimension C . However, we observe a trade-off between efficiency and performance: models with longer kernels (e.g., $K = 8$) do not give the best performance. Instead, we find that $K = 3$ and $K = 4$ lead to the best results.

In Table 6, we evaluate the contribution of our other modifications to the original Transformer block while keeping the model size constant. A1 removes the first ConvSwiGLU module while increasing the hidden dimension in the second ConvSwiGLU to $2C$. A2 further swaps the SwiGLU activation for Swish (i.e., removes the right Conv1D branch in Fig. 1) while increasing the hidden dimension to $3C$. The result demonstrates that both designs help improving the separation performance.

Finally, we compare the performance of the models with RMSNorm and RMSGroupNorm in Table 7. Although the improvement is slight, the proposed RMSGroupNorm consistently leads to better performance. This result demonstrates that, as discussed in Section 2.3, encouraging the model to make groups in each TF bin can be effective for TF-domain dual-path separation models.

4. CONCLUSION

We presented TF-Locoformer, a speech separation model that effectively integrates global and local modeling for TF-domain dual-path modeling. While previous SoTA TF-domain dual-path models are based on RNNs, we developed a model based on the Transformer block, considering its advantages such as parallelizable architecture, potential scalability, and versatility (e.g., prompting to specify the task). Inspired by the success of RNNs, which perform both local and global modeling, we designed a model where self-attention does global modeling and convolution handles local modeling. We also proposed an effective normalization layer for TF-domain models. The experimental comparison demonstrated that the proposed model gives comparable or better performance than previous SoTA models on four commonly-used benchmarks. Through the ablation study, we have shown the importance of local modeling. Finally, the experiments implied that TF-domain models deal with reverberation much better than time-domain models. In the future, we will investigate the scalability of TF-Locoformer, as well as its effectiveness on music and general sound separation.

5. REFERENCES

- [1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016.
- [2] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020.
- [4] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020.
- [5] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021.
- [6] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-frequency domain path scanning network for speech separation," in *Proc. ICASSP*, 2022.
- [7] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim *et al.*, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.
- [8] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilá, R. Doddipatla *et al.*, "Monaural source separation: From anechoic to reverberant environments," in *Proc. IWAENC*, 2022.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," *Proc. NeurIPS*, 2017.
- [10] J. Rixen and M. Renz, "QDPN - quasi-dual-path network for single-channel speech separation," in *Proc. Interspeech*, 2022.
- [11] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang *et al.*, "MossFormer2: Combining transformer and RNN-free recurrent network for enhanced time-domain monaural speech separation," in *Proc. ICASSP*, 2024.
- [12] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma *et al.*, "A comparative study on Transformer vs RNN in speech applications," in *Proc. ASRU*, 2019.
- [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.
- [15] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, "Toward universal speech enhancement for diverse input conditions," in *Proc. ASRU*, 2023.
- [16] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
- [17] J. Droppo and O. Elibol, "Scaling laws for acoustic models," in *Proc. Interspeech*, 2021.
- [18] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. CVPR*, 2022.
- [19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [20] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong *et al.*, "Understanding and improving transformer from a multi-particle dynamic system point of view," in *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.
- [21] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.
- [22] N. Shazeer, "GLU variants improve Transformer," *arXiv preprint arXiv:2002.05202*, 2020.
- [23] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Proc. NeurIPS*, 2019.
- [24] Y. Wu and K. He, "Group normalization," in *Proc. ECCV*, 2018.
- [25] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.
- [26] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. ICASSP*, 2020.
- [27] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020.
- [28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.
- [29] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang *et al.*, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for asr integration," in *Proc. SLT*, 2021.
- [30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2018.
- [31] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017.
- [32] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR — half-baked or well done?" in *Proc. ICASSP*, 2019.
- [33] Y.-J. Lu, S. Cornell, X. Chang, W. Zhang, C. Li *et al.*, "Towards low-distortion multi-channel speech enhancement: The ESPNET-SE submission to the L3DAS22 challenge," in *Proc. ICASSP*, 2022.
- [34] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [36] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.
- [37] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.
- [38] Y. Lee, S. Choi, B.-Y. Kim, Z.-Q. Wang, and S. Watanabe, "Boosting unknown-number speaker separation with transformer decoder-based attractor," in *Proc. ICASSP*, 2024.
- [39] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang *et al.*, "A mask free neural network for monaural speech enhancement," in *Proc. Interspeech*, 2023.