TF-CrossNet: Leveraging Global, Cross-Band, Narrow-Band, and Positional Encoding for Single- and Multi-Channel Speaker Separation

Vahid Ahmadi Kalkhorani and DeLiang Wang, *Fellow, IEEE*

Abstract—**We introduce TF-CrossNet, a complex spectral mapping approach to speaker separation and enhancement in reverberant and noisy conditions. The proposed architecture comprises an encoder layer, a global multi-head self-attention module, a cross-band module, a narrow-band module, and an output layer. TF-CrossNet captures global, cross-band, and narrow-band correlations in the time-frequency domain. To address performance degradation in long utterances, we introduce a random chunk positional encoding. Experimental results on multiple datasets demonstrate the effectiveness and robustness of TF-CrossNet, achieving state-of-the-art performance in tasks including reverberant and noisy-reverberant speaker separation. Furthermore, TF-CrossNet exhibits faster and more stable training in comparison to recent baselines. Additionally, TF-CrossNet's high performance extends to multi-microphone conditions, demonstrating its versatility in various acoustic scenarios.**

Index Terms—**Complex spectral mapping, multi-channel, single-channel, speaker separation, time-frequency domain.**

I. Introduction

In human and machine speech communication, the presence of acoustic interference, such as background noise or competing speakers, presents a considerable challenge for speech understanding. To address these challenges, speech separation systems have been developed to separate target speech signals from noisy and reverberant environments. Speech separation includes speaker separation and speech enhancement [1]. The task of speaker separation is to separate the speech signals of multiple speakers and speech enhancement aims to separate a single speech signal from nonspeech background noise. Both tasks are essential for various applications, including hearing aids, teleconferencing, and voice-controlled assistants.

Significant strides have been made in monaural talker-independent speaker separation with the introduction of deep clustering [2] and permutation invariant training (PIT) [3]. By effectively tackling the permutation ambiguity issue inherent in talker-independent training, these approaches have substantially elevated speaker separation performance. Subsequent developments have produced impressive performance gains.

For example, deep CASA [4] breaks down the speaker separation task into two phases: simultaneous grouping and sequential grouping. Conv-TasNet [5] operates on short windows of signals and performs end-to-end masking-based separation. DPRNN [6] segments a time-domain signal into fixed-length blocks, where intra- and inter-block recurrent neural networks (RNNs) are applied iteratively to facilitate both local and global processing. SepFormer [7] replaces RNNs with a set of multi-head self-attentions (MHAs) and linear layers. Like Conv-TasNet, SepFormer is a masking approach in the time domain. The availability of spatial information from multiple microphones allows for location-based training to resolve the permutation ambiguity issue, which further improves speaker separation results [8].

While most of the effective monaural speaker separation algorithms operate in the time domain, recently, deep neural networks (DNNs) operating in the frequency domain have gained prominence by harnessing various forms of spectral information, including full-band/cross-band and sub-band/narrow-band for both single- and multi-channel speech separation. The representative model of TF-GridNet [9] employs cross-band and narrow-band long short-term memory (LSTM) networks in conjunction with a cross-frame self-attention module to perform complex spectral mapping [10], [11], [12], [13]. The most effective TF-GridNet model comprises a two-stage DNN with a neural beamformer positioned in the intermediate stage. This model has strongly improved speech separation results in a variety of single-channel and multi-channel tasks. SpatialNet [14] shares a foundational framework with TF-GridNet, but employs a combination of a Conformer narrow-band block and a convolutional-linear cross-band block. Notably, SpatialNet excludes any LSTM or RNN layers. Furthermore, SpatialNet operates as a single-stage network and exhibits a more stable training trajectory, especially under conditions involving half-precision (16-bit) training. SpatialNet demonstrates very competitive results in multi-channel speaker separation. But its utility is primarily tailored for multi-channel scenarios, given its substantial reliance on spatial information afforded by microphone arrays;

Received 11 April 2024; revised 31 August 2024 and 22 October 2024; accepted 27 October 2024. Date of publication 6 November 2024; date of current version 25 November 2024. This work was supported in part by National Science Foundation under Grant ECCS-2125074, in part by Ohio Supercomputer Center, in part by NCSA Delta Supercomputer Center under Grant OCI 2005572, and in part by Pittsburgh Supercomputer Center under Grant NSF ACI-1928147. The associate editor coordinating the review of this article and approving it for publication was Dr. Marc Delcroix.*(Corresponding author: Vahid Ahmadi Kalkhorani.)*

Vahid Ahmadi Kalkhorani is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: ahmadikalkhorani.1@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA, and also with the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2024.3492803

as shown later, its performance in the single-channel scenario is limited. Another notable limitation of SpatialNet, in comparison to TF-GridNet, is its performance degradation with increasing sequence length, as recently reported in [15].

To overcome the aforementioned shortcomings and further enhance the performance of complex spectral mapping for speaker separation, we examine the underlying reasons behind the observed performance differences between TF-GridNet and SpatialNet, particularly in scenarios involving monaural separation and long utterances. We attribute the observed performance degradation of SpatialNet relative to TF-GridNet to two primary factors. First, the self-attention module within TF-GridNet operates as a global attention mechanism, whereas SpatialNet processes each frequency independently, unable to benefit from cross-frequency and hidden features. We believe that the lack of such global attention contributes to SpatialNet's diminished performance in processing long sequences. Second, RNNs as exemplified by LSTM possess the capability to implicitly extract positional information [16], [17], [18]. Therefore, even though neither SpatialNet nor TF-GridNet architecture explicitly incorporates positional encoding, the use of RNNs captures positional cues in TF-GridNet implicitly.

In this study, we propose a new DNN architecture, called TF-CrossNet, for single- and multi-channel speaker separation. Building upon complex spectral mapping and the SpatialNet framework, we make the following contributions:

- We present a new DNN architecture for both single- and multi-channel speaker separation tasks. This architecture employs a global multi-head self-attention module to capture cross-frequency and cross-embedding correlations.
- We introduce a novel positional encoding method to TF-CrossNet to address the out-of-distribution problem of common positional encoding methods.
- TF-CrossNet advances the state-of-the-art speaker separation performance on multiple benchmark datasets. In addition, superior results are achieved with a reduced computational overhead in terms of both inference and training time.

The rest of the paper is organized as follows. Section II describes the single- and multi-channel speaker separation problem in the time-frequency (T-F) domain. The detailed description of TF-CrossNet is given in Section III. Section IV presents the experimental setup. Evaluation and comparison results are provided in Section V. Concluding remarks are given in Section VI.

II. Problem Statement

For a mixture of C speakers in a noisy-reverberant environment captured by an array of M microphones, the recorded mixture in the time domain $\mathbf{y}(n) \in \mathbb{R}^M$ can be modeled in terms of the direct-path signals $\mathbf{s}_c(n) \in \mathbb{R}^M$, their reverberations $\mathbf{h}_c(n) \in \mathbb{R}^M$, and reverberant background noises $\mathbf{v}(n) \in \mathbb{R}^M$ [9], [19]

$$\mathbf{y}(n) = \sum_{c=1}^{C} (\mathbf{s}_c(n) + \mathbf{h}_c(n)) + \mathbf{v}(n), \tag{1}$$

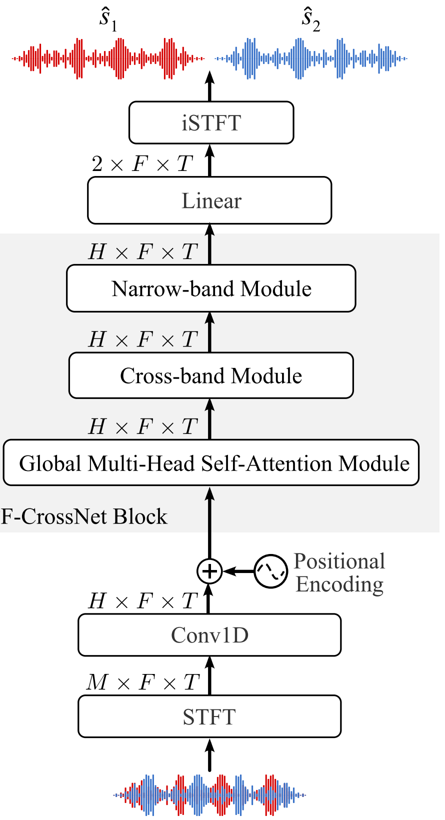

Fig. 1. Diagram of the proposed TF-CrossNet architecture, with $\hat{s}_1$ and $\hat{s}_2$ denoting separated speaker signals.

where n denotes discrete time and c indexes speakers. In the short-time Fourier transform (STFT) domain, the model is expressed as:

$$\mathbf{Y}(t, f) = \sum_{c=1}^{C} (\mathbf{S}_c(t, f) + \mathbf{H}_c(t, f)) + \mathbf{V}(t, f), \tag{2}$$

where t indexes time frames and f frequency bins. $\mathbf{Y}(t, f)$, $\mathbf{S}_c(t, f)$, $\mathbf{H}_c(t, f)$, and $\mathbf{V}(t, f) \in \mathbb{C}^M$ denote the complex spectrograms of the mixture, the direct-path signal and its reverberation of speaker c, and background noise, respectively.

The goal of complex spectral mapping based speaker separation is to train a DNN to estimate the real and imaginary parts of the direct-path signal of each speaker at a reference microphone from the mixture $\mathbf{Y}(t, f)$. We can turn the general formulation in (2) to more specific forms by restricting certain parameters and terms. In the case of monaural, anechoic speaker separation, $C > 1$, $M = 1$, both $\mathbf{H}_c(t, f)$ and $\mathbf{V}(t, f)$ are absent. In reverberant speaker separation, $C > 1$ and $\mathbf{V}(t, f)$, if present, represents a weak noise. In the case of noisy-reverberant speaker separation, $C > 1$ and $\mathbf{V}(t, f)$ includes significant background noise.

III. TF-CrossNet

The diagram of the proposed system is provided in Fig. 1. TF-CrossNet comprises an encoder layer, a global multi-head self-attention (GMHSA) module, a cross-band module, a narrow-band module, and a decoder layer. To ensure comparable energy levels for all signals processed by TF-CrossNet, we normalize the input signal by its variance before processing its samples. In the multi-channel setup, we normalize the signals from all microphones by the variance of the reference microphone; the same variance is applied to restore the scale of a predicted signal. Then, we apply STFT to the normalized signal and stack the real and imaginary (RI) parts. For the multi-channel setup, we stack the RI parts from all microphones as done in neural spectrospatial filtering [20]. The stacked RI parts are sent to the encoder layer, which learns to extract acoustic features from the input in the STFT domain. The global multi-head self-attention module captures global correlations, while the cross-band module captures cross-band correlations. The narrow-band module focuses on capturing information at neighboring frequency bins. Finally, the output layer maps the separated features to a T-F representation, which is then converted back to the time domain using inverse short-time Fourier transform (iSTFT).

A. Encoder Layer

The encoder is a 1D convolutional layer (Conv1D) layer with a kernel size of k and a stride of 1. The encoder layer converts the input T-F domain signal from $M \times F \times T$ to $H \times F \times T$ where H is the number of hidden channels. F is the number of frequency bins and T is number of frames.

B. Random Chunk Positional Encoding

To address the limitation of separation methods in dealing with long utterances, we introduce a positional encoding method, called random-chunk positional encoding (RCPE), to tackle the out-of-distribution problem in positional encoding approaches. RCPE is inspired by random positional encoding recently proposed for natural language processing [21]. Transformers demonstrate impressive generalization capabilities on learning tasks with a fixed context length. However, their performance degrades when tested on longer sequences than the maximum length encountered in training. This degradation is attributed to the fact that positional encoding becomes out-of-distribution for longer sequences, even for relative positional encoding [21]. RCPE selects a contiguous chunk of positional embedding vectors from a pre-computed positional encoding matrix during training. For RCPE, we start by defining PE as a combination of sine and cosine functions [22] as

$$\mathrm{PE}(t, 2i) = \sin\left(\frac{t}{10000^{2i/(F\cdot H))}}\right), \tag{3a}$$

$$\mathrm{PE}(t, 2i+1) = \cos\left(\frac{t}{10000^{2i/(F\cdot H)}}\right), \tag{3b}$$

where $i \in [1, F \cdot H]$ and $t \in [1, T]$ index the feature and time dimensions, respectively.

When the model is in the training mode, we select a random chunk from index τ to index $\tau + T - 1$, where τ is drawn randomly from $[1, T^{\max} - T + 1]$, with $T^{\max}$ denoting the maximum desired sequence length during inference. When the model is in test or validation mode, we select the first T embedding vectors. Finally, we reshape and add the selected positional embeddings to the input features. We obtain positional encoding vectors as

$$RCPE(t, i) = \begin{cases} PE(t+\tau, i) & \text{if training,} \\ PE(t, i) & \text{otherwise.} \end{cases} \tag{4}$$

This technique allows the TF-CrossNet model to see all possible positional embedding vectors during the training stage while maintaining the relative distance between embeddings, thus improving generalization to longer sequences. Additionally, RCPE has no learnable parameter and has a negligible computational cost.

C. Global Multi-Head Self-Attention Module

Fig. 2(a) shows the diagram of the global multi-head self-attention module. This module resembles TF-GridNet's cross-frame self-attention mechanism, but with modifications to enhance efficiency. In TF-GridNet [9], the cross-frame self-attention module employs three point-wise convolution layers for frame-level feature extraction of queries, keys, and values. In contrast, we utilize a single convolution layer with $L(2E + H/L)$ output channels to extract frame-level features from T-F embeddings. Increasing the output dimension, rather than performing sequential convolutions, increases parallel computation and accelerates the operation. Subsequently, we split the result into L queries $Q^l \in \mathbb{R}^{E\times F\times T}$, keys $K^l \in \mathbb{R}^{E\times F\times T}$, and values $V^l \in \mathbb{R}^{H/L\times F\times T}$. Here, E represents the output channel dimension of the point-wise convolution and l indexes the head number. This method avoids the sequential operations of the three Conv1D layers, which can be computationally expensive. Subsequently, a self-attention layer is applied to these embeddings to capture global correlations. The results of all heads are concatenated and passed to another point-wise convolution with an output dimension of D followed by a parametric rectified linear unit (PReLU) activation function and layer normalization (LN). We add this value to the input of the GMHSA module to obtain the output of the module. Note that, compared to [14] where the MHA module acts on each frequency bin separately, we first merge all frequency features into the channel dimension and then apply MHA. This method allows each frame to attend to any frame of interest in all feature channels, facilitating the exploitation of long-range correlations in both frequency and hidden feature channels.

D. Cross-Band Module

To capture cross-band correlations within the input signal, we adopt the cross-band module proposed in [14]. This module, illustrated in Fig. 2(b), integrates two frequency-convolutional modules and a full-band linear module. The frequency-convolutional module aims to capture correlations between neighboring frequencies. This module includes an LN layer, a grouped convolution layer along the frequency axis (F-GConv1d), and a PReLU activation function. In the full-band linear module, we first employ a linear layer followed by sigmoid-weighted linear unit (SiLU) activation function to

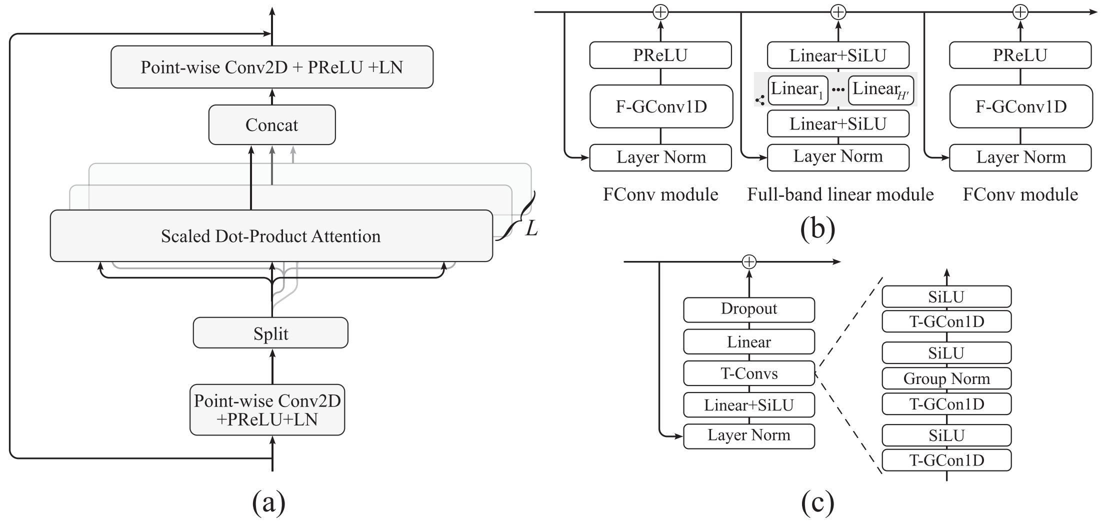

Fig. 2. TF-CrossNet building blocks. (a) Global multi-head self-attention module. (b) Cross-band module. (c) Narrow-band module.

reduce the number of hidden channels from H to H'. Then, we apply a set of linear layers along the frequency axis to capture full-band features. Each feature channel has a dedicated linear layer denoted as Linear$_i$ for $i = 1, \ldots, H'$, as shown in Fig. 2(b). Note that the parameters of these layers are shared among all TF-CrossNet blocks. Finally, the output of the module is obtained by increasing the number of channels back to H using a linear layer with SiLU activation and adding to the original input of this module.

E. Narrow-Band Module

As illustrated in Fig. 2(c), the narrow-band module is composed of a layer normalization (LN), a linear layer followed by a SiLU activation, a time-convolutional (T-Conv) layer, and a final linear layer. The first linear layer in this module increases the number of features in the input from H to H'' and the last linear layer converts the feature dimension back to H.

T-Conv is composed of three grouped 1 d convolution (T-GConv1D) layers followed by a SiLU activation function. The second T-GConv1D is followed by a grouped normalization layer. The narrow-band module is a modified version of the Conformer convolutional block [23]. Compared to SpatialNet's narrow-band block, we remove the MHA module as narrow-band correlations are captured in the GMHSA module of TF-CrossNet.

F. Output Layer

We use a linear output layer to map the processed features from the final TF-CrossNet block to the predicted RI parts of each talker. Subsequently, we obtain the time-domain separated speech signals by performing the iSTFT. As mentioned at the beginning of Section III, we multiply the estimated target signals by the variance of the input mixture to ensure that their energy levels are consistent with the mixture level.

G. Loss Functions

We use the scale-invariant signal-to-distortion ratio (SI-SDR) [24] loss function $\mathcal{L}_{\text{SI-SDR}}$ to train TF-CrossNet on the WSJ0-2mix dataset [2]. For training on other datasets, we employ a combination of magnitude loss $\mathcal{L}_{\text{Mag}}$ and SI-SDR loss $\mathcal{L}_{\text{SI-SDR}}$, similar to [9]. We find that the combined loss function improves time-domain metrics such as SI-SDR, as well as more magnitude-based metrics like PESQ and word error rate (WER). We use the standard form of SI-SDR where the target signal is scaled to match the scale of the estimated signal. Also, we scale the magnitude loss by the L_1 norm of the magnitude of the target signal in the STFT domain similar to [25]. These loss functions are defined below

$$\mathcal{L} = \mathcal{L}_{\text{Mag}} + \mathcal{L}_{\text{SI-SDR}}, \tag{5a}$$

$$\mathcal{L}_{\text{Mag}} = \frac{\| |\text{STFT}(\hat{\boldsymbol{s}}_c)| - |\text{STFT}(\boldsymbol{s}_c)| \|_1}{\| |\text{STFT}(\boldsymbol{s}_c)| \|_1}, \tag{5b}$$

$$\mathcal{L}_{\text{SI-SDR}} = -\sum_{c=1}^{C} 10 \log_{10} \frac{\|\boldsymbol{s}_c\|_2^2}{\|\hat{\boldsymbol{s}}_c - \alpha_c \boldsymbol{s}_c\|_2^2}, \tag{5c}$$

$$\alpha_c = \frac{\boldsymbol{s}_c^T \hat{\boldsymbol{s}}_c}{\boldsymbol{s}_c^T \boldsymbol{s}_c}. \tag{5d}$$

In the above equations, $\|\cdot\|_1$ is the L_1 norm, $|\cdot|$ is the magnitude operator, α_c is the scaling factor, and $(\cdot)^T$ denotes the transpose operation. We employ utterance-level PIT [3] to resolve the permutation ambiguity problem during training.

IV. Experimental Setup

A. Datasets

We assess the efficacy of the proposed TF-CrossNet model for speaker separation under anechoic, reverberant, and noisy-reverberant environments. We use publicly available datasets, and compare with previously published results to document the relative performance.

For single-channel speaker separation in anechoic conditions, we employ the WSJ0-2mix dataset [2], which is widely used for benchmarking monaural talker-independent speaker separation algorithms. The WSJ0-2mix dataset consists of 20,000 ($\sim$30.4 hours), 5,000 ($\sim$7.7 hours), and 3,000 ($\sim$ 4.8 hours) two-speaker mixtures for training, validation, and test sets, respectively. In WSJ0-2mix, the two utterances in each mixture are fully overlapped, and their relative energy level is sampled from the range of $[-5, 5]$ dB. Speech is sampled at a rate of 8 kHz. To make a fair comparison, similar to TF-GridNet, we do not utilize any data augmentation techniques such as dynamic-mixing [26] or speed-perturbation [7].

For joint speaker separation, denoising, and dereverberation, we employ the WHAMR! dataset [27] and the single-channel SMS-WSJ dataset [28]. WHAMR! utilizes the two-speaker mixtures from WSJ0-2mix, but introduces reverberation to each clean anechoic signal and non-stationary background noises. The dataset includes 20,000 ($\sim$30.4 hours), 5,000 ($\sim$7.7 hours), and 3,000 ($\sim$ 4.8 hours) mixtures for training, validation, and testing, respectively.

Furthermore, for both monaural and multi-channel separation in noisy and reverberant environments, we employ the SMS-WSJ dataset [28]. This simulated two-speaker mixture dataset incorporates clean speech signals from the WSJ0 corpus and simulates a six-microphone circular array with a radius of 10 cm. room impulse responsess (RIRs) are generated using the image method [29], with T60 uniformly sampled between 0.2 s and 0.5 s. Additionally, white sensor noise is added to speech mixtures with signal-to-noise ratios (SNRs) uniformly sampled in the range of 20 dB to 30 dB. The source positions are randomly sampled within 1 m to 2 m away from the array center. The signals are sampled at a rate of 8 kHz, and the dataset includes a baseline automatic speech recognition (ASR) model built from Kaldi [30].

We also assess TF-CrossNet on the REVERB challenge dataset [31], which includes both simulated and recorded signals of a speaker sampled at 16 kHz. We employ the REVERB evaluation set including simulated (SimData) and real (RealData) recordings for single- and multi-channel speech dereverbration and enhancement assessment. SimData consists of 2176 utterances from the WSJCAM0 corpus [32], convolved with measured RIRs from three rooms of different sizes and two near and far microphone distances. The background noise in the recordings is primarily stationary diffuse noise generated by the air-conditioning systems in the rooms. RealData consists of 372 utterances from the MC-WSJ-AV corpus [33], recorded in a different room from those is SimData, with speaker-to-microphone distances of 1.0 cm and 2.5 m. For the training set, similar to [13], [14], we increase the number of RIRs by simulating rectangular rooms using the image method [29], [34], with room length and width randomly chosen between 5 and 10 m, respectively, and height between 3 and 4 m. An 8-channel circular microphone array is positioned within this space, with its height randomly chosen between 1 and 2 m. The array's center is displaced from the room center by values randomly sampled between -1.0 and 1.0 m. The array radius is randomly chosen between 3 and 10 cm. The target speech source is placed at a distance from the array center between 0.5 and 3.0 m. The reverberation time (T60) is randomly sampled between 0.2 and 1.5 seconds. Using this configuration, we generate 40,000 RIRs and convolve them with source signals to obtain the reverberated signals. Similar to [13], we utilize the direct-path signal for both training and metric computation. Specifically, we use the samples within a 5-ms window around the peak from the measured RIRs to estimate the direct-path signal for metric calculations.

B. Network Configuration

For our proposed TF-CrossNet architecture, we make use of the hyperparameters in [9] and [14]. We set the kernel size of encoder layer k, time-dimension group convolution (T-GConv1d), and frequency-dimension group convolution (F-GConv1d) to 5, 5, and 3, respectively. The number of groups for T-GConv1d, F-GConv1d, and group normalization is all set to 8. The proposed model architecture comprises $B = 12$ blocks, with hidden channel sizes set to $H = 192$, $H' = 16$, and $H'' = 384$. We employ $N = 4$ self-attention heads in the GMHSA module with an embedding dimension of $D = 64$ and $E = \lceil 512/F \rceil$, where $\lceil \cdot \rceil$ denotes ceiling operation.

To process the input data, we apply STFT using a Hanning window with frame length of 256 samples (32 ms) and frame shift of 128 samples (16 ms). The length of training utterances is fixed at 3 seconds for the WSJ0-2mix and REVERB datasets and 4 seconds for the WHAMR! and SMS-WSJ datasets [14]. We assume a maximum utterance length of 30 seconds to calculate T^{max}, introduced in Section III-B.

We utilize the Adam optimizer with a maximum learning rate of 0.001. We start with a cosine warm-up scheduler that increases the learning rate from 10^{-6} to 10^{-3} over the first 10 epochs. Following this, we switch to the PyTorch ReduceLROnPlateau scheduler, setting the patience to 3 epochs and the reduction factor to 0.9. We found that this learning rate scheduler is more stable and results in faster convergence than the exponential decay or ReduceLROnPlateau schedulers used in [14] and [9], respectively. In our experiments, we employ the half-precision (mixed-16) training strategy to reduce the memory footprint and accelerate training. We train the model until the validation loss does not improve for 10 epochs consecutively. In each case, we use the maximum number of batches that fit into the GPU memory (NVIDIA A100 GPU with 40 GB).

C. Evaluation Metrics

We employ a set of widely used objective metrics to assess the performance of TF-CrossNet. These metrics include, SI-SDR and its improvement (SI-SDRi) [24], SDR and its improvement (SDRi) [35], narrow-band perceptual evaluation of speech quality (PESQ) [36], and extended short-time objective intelligibility (eSTOI) [37]. To compute these metrics, we utilize the TorchMetrics[audio] package [38], which offers a comprehensive set of evaluation tools specifically designed for audio tasks.

TABLE I
ABLATION STUDY ON THE WHAMR! DATASET

Row	Positional encoding	GMHSA	NB-MHSA	Params (M)↓	GFLOPs↓	SI-SDR↑	PESQ↑
1	✗	✗	✓	6.50	118.84	10.2	2.54
2	✗	✓	✓	8.35	143.51	11.5	2.85
3	LSTM	✓	✓	9.41	159.59	11.9	2.94
4	LSTM	✗	✓	7.56	135.05	11.6	2.89
5	RCPE	✓	✓	8.35	143.50	11.8	2.92
6	RCPE	✓	✗	6.57	96.14	11.8	2.91

For assessment on the REVERB challenge dataset, we employ the official objective measures.[1] These include three intrusive speech metrics: cepstrum distance (CD), log likelihood ratio (LLR), frequency-weighted segmental signal-to-noise ratio (FWSegSNR), and one non-intrusive metric, speech-to-reverberation modulation energy ratio (SRMR). As done in [14], we use the best pre-trained model[2] from ESPnet [39] to evaluate the ASR performance on the Reverb challenge dataset.

V. EVALUATION RESULTS

A. Ablation Study on WHAMR!

Table I presents an ablation study conducted on the single-channel WHAMR! dataset. Each row represents a different configuration of the model. The columns in this table provide information about the presence of RCPE, GMHSA, and narrow-band multi-head self-attention (NB-MHSA), along with the number of trainable parameters in millions or Params (M), and the number of Giga floating point operations (GFLOPs) per second of input audio, as well as the separation performance metrics of SI-SDR, SDR, and PESQ. For the computation of GFLOPs, we use the official tool provided by PyTorch[3]. The absence of RCPE and GMHSA (Row 1) results in lower SI-SDR and PESQ scores compared to the configurations where these components are present. Note that Row 1 corresponds to the architecture of SpatialNet [14]. Adding the GMHSA module in Row 2 improves SI-SDR by 1.3dB and PESQ by 0.31, highlighting the important role of GMHSA. In the third row, we include an LSTM encoder before TF-CrossNet blocks, which performs positional encoding implicitly. The LSTM encoder comprises two bidirectional long short-term memory (BLSTM) layers similar to TF-GridNet's intra-frame full-band and sub-band temporal modules. Although this configuration exhibits the highest SI-SDR and PESQ values among the tested configurations, it has the largest number of parameters and the lowest computational efficiency. In the fourth row, we exclude the GMHSA module. This decreases both SI-SDR and PESQ scores, demonstrating the contribution of GMHSA even with the LSTM encoder. Including RCPE in the fifth row improves SI-SDR by 0.3 dB and PESQ by 0.07 compared to the second row, demonstrating the utility of the proposed positional encoding. Finally, in Row 6, we remove NB-MHSA and obtain speaker separation results with only a 0.01 PESQ reduction compared to Row 5. But the configuration with no NB-MHSA has about 20% fewer trainable parameters and 33% fewer GFLOPs. This shows that the narrow-band correlations are already captured in the GMHSA module and there is little need to include both modules in the network.

TABLE II
ABLATION STUDY COMPARING DIFFERENT POSITIONAL ENCODING METHODS

Row	Positional encoding	SI-SDR↑	PESQ↑
1	✗	11.4	2.84
2	SPE	11.4	2.84
3	LSTM	11.9	2.92
4	RCPE	11.8	2.91

Table II presents an ablation study comparing different positional encoding methods, evaluating their impact on SI-SDR and PESQ metrics. The baseline model without positional encoding achieves 11.4dB SI-SDR and 2.84 PESQ, similar to the performance using standard positional encoding (SPE) with no random selection. Both RCPE and LSTM outperform the baseline and SPE. Specifically, LSTM improves SI-SDR to 11.9dB and PESQ to 2.92, while RCPE yields 11.8dB SI-SDR and 2.91 PESQ. Note that RCPE achieves comparable performance to LSTM, with far fewer parameters, which require no training.

B. WSJ0-2mix Results

We first evaluate the performance of TF-CrossNet for monaural anechoic speaker separation. The mixture SI-SDR is 0dB, and the mixture SDR is 0.2dB. The results are provided in Table III along with 16 other baselines. The table includes two versions of TF-GridNet, one with 8.2M parameters and another with 14.5M parameters. The original study [9] reports the 14.5M parameter version on the WSJ0-2mix dataset. To compare models of comparable sizes, we include the 8.2M variant as well. The performance of this smaller TF-GridNet model is based on a model checkpoint trained by its original first author [40]. CrossNet surpasses the performance of state-of-the-art methods, including TF-GridNet (8.2M) [9] by 0.5dB SI-SDR and 0.6dB SDR. Moreover, our proposed model has around 20% fewer trainable parameters compared to TF-GridNet and faster training and inference as presented in Section V-G later. Furthermore, our proposed model underwent half-precision floating-point training rather than full-precision training done in TF-GridNet, effectively reducing memory requirements and expediting the training process.

[1] [Online]. Available: https://reverb2014.audiolabs-erlangen.de/tools/REVERB-SPEENHA.Release04Oct.zip

[2] Transformer ASR + Transformer LM + SpeedPerturbation + SpecAug + applying RIR and noise data on the fly

[3] torch.utils.flop_counter.FlopCounterMode

TABLE III
SPEAKER SEPARATION RESULTS OF TF-CROSSNET AND COMPARISON METHODS ON THE WSJ0-2MIX DATASET

Method	Params (M)↓	SI-SDRi↑	SDRi↑
Conv-TasNet [5]	5.1	15.3	15.6
Deep CASA [4]	12.8	17.7	18.0
FurcaNeXt [41]	51.4	-	18.4
SUDO RM-RF [42]	2.6	18.9	-
DPRNN [6]	7.5	20.1	20.4
DPTNet [17]	2.7	20.2	20.6
DPTCN-ATPP [43]	4.7	19.6	19.9
SepFormer [7]	26.0	20.4	20.5
Sandglasset [44]	2.3	20.8	21.0
Wavesplit [26]	29.0	21.0	21.2
TFPSNet [45]	2.7	21.1	21.3
DPTNet [17]	4.0	21.5	21.7
SFSRNet [46]	59.0	22.0	22.1
QDPN [47]	200.0	22.1	-
TF-GridNet* [9]	8.2	22.8	22.9
TF-GridNet [9]	14.5	23.5	23.6
TF-CrossNet	6.6	23.2	23.4

*checkpoint from [40]

TABLE IV
SPEAKER SEPARATION RESULTS OF TF-CROSSNET AND COMPARISON METHODS ON THE WHAMR! DATASET

Method	SI-SDR↑	SDR↑	PESQ↑	eSTOI↑
Unprocessed	-6.1	-3.5	1.41	0.317
Sepformer [7]	7.9	-	-	-
MossFormer [48]	10.2	-	-	-
SpatialNet (large) [14]	10.2	11.2	2.54	0.772
TF-GridNet (1-stage) [9]	10.6	11.7	2.75	0.793
DasFormer [49]	11.2	12.2	-	-
TF-GridNet (2-stage) [9]	11.2	12.3	2.79	0.808
TF-CrossNet (2-stage)	12.0	13.1	2.91	0.824
TF-CrossNet	11.8	12.9	2.91	0.823

TABLE V
SPEAKER SEPARATION AND ASR RESULTS ON SINGLE-CHANNEL SMS-WSJ DATASET

Method	SI-SDR↑	SDR↑	PESQ↑	eSTOI↑	WER↓
Unprocessed	-5.5	-0.4	1.50	0.441	78.40
Oracle direct-path	∞	∞	4.50	1.000	6.16
DPRNN-TasNet [6]	6.5	-	2.28	0.734	38.10
SISO_1 [19]	5.7	-	2.4	0.748	28.70
$\text{DNN}_1\text{+(FCP+DNN}_2)\times 2$ [19]	12.7	14.1	3.25	0.899	12.80
$\text{DNN}_1\text{+(msFCP+DNN}_2)\times 2$ [50]	13.4	-	3.41	-	10.90
TF-GridNet [9] (1-stage)	16.2	17.2	3.45	0.924	9.49
TF-GridNet [9] (2-stage)	18.4	19.6	3.70	0.952	7.91
TF-CrossNet	19.2	20.2	3.74	0.953	8.35

C. Results on WHAMR! and Single-Channel SMS-WSJ Datasets

The single-channel WHAMR! results are summarized in Table IV. TF-CrossNet achieves an SI-SDR of 11.8dB, an SDR of 12.9dB, and a PESQ of 2.91, outperforming the previous best of TF-GridNet (1-stage) [9] and TF-GridNet (2-stage) [9] by 0.16 and 0.12 PESQ, respectively. The 2-stage TF-GridNet consists of the first DNN followed by a single-channel multi-frame Wiener filter (SCMFWF) and then the second DNN. This comparison is significant as TF-CrossNet is a single-stage model, and a 2-stage model not only has more parameters but also takes more effort to train and deploy. Our advantage can be attributed to the use of more convolutional layers, which enables TF-CrossNet to learn filtering operations. Note that SpatialNet is not designed for single-channel separation tasks even though it can be applied to monaural separation. We include its results in Table IV for reference purposes only. Without spatial cues, the performance of SpatialNet is reduced significantly. TF-CrossNet leverages the strengths of both SpatialNet and TF-GridNet while avoiding LSTM layers in TF-GridNet. RCPE in our model serves to capture the positional information encoded in the recurrent layers of TF-GridNet. Consequently, TF-CrossNet remains effective for single-channel separation without computationally expensive recurrent connections. Compared to the results in Table III, these results underscore the advantage of TF-CrossNet over TF-GridNet for single-channel speaker separation in noisy-reverberant conditions.

To examine the impact of SCMFWF on model performance, we train TF-CrossNet with a similar setup to the two-stage TF-GridNet, and 2-stage TF-CrossNet results are included in Table IV. We observe a very small improvement. Thus, we conclude that two stages are not necessary and will not be further assessed for TF-CrossNet. This observation shows that, compared to TF-GridNet where SCMFWF improves the performance, Wiener filtering is not essential for TF-CrossNet.

Table V presents evaluation and comparison results on the single-channel SMS-WSJ dataset, including ASR results in terms of WER in percentage (%) evaluated on the official ASR model [28], [30]. TF-CrossNet outperforms TF-GridNet (1-stage) [9] by the substantial margin of 3.0 dB in SI-SDR and 0.29 in PESQ. Notably, TF-CrossNet outperforms the two-stage TF-GridNet aside from the WER score. The better WER score of the two-stage TF-GridNet is likely due to its use of neural beamformers which can significantly reduce WERs in both single- and multi-talker scenarios [14].

D. Results on Multi-Channel SMS-WSJ

Table VI reports the performance of the six-channel speaker separation and ASR on the SMS-WSJ corpus, along with the oracle WER results. The table reveals large improvements in speech quality and ASR performance thanks to speaker separation. The time-domain end-to-end models FaSNet+TAC [51] and MC-ConvTasNet [52] show inferior performance compared to other methods, especially on the ASR task. Time-frequency methods such as $\text{MISO}_1\text{-BF-MISO}_3$ [19] and TFGridNet [9] incorporate neural beamforming and post-processing, and demonstrate significantly better separation and ASR performances. Among the comparison methods, SpatialNet is the top performer, and it leverages an advanced full-band and sub-band combination network and extensively employs convolutional and linear layers that can act as a large filter. TF-CrossNet

TABLE VI
SPEAKER SEPARATION AND ASR RESULTS ON THE 6-CHANNEL SMS-WSJ DATASET

Method	SI-SDR↑	SDR↑	PESQ↑	eSTOI↑	WER↓
Unprocessed	-5.5	-0.4	1.50	0.441	78.40
Oracle direct-path	∞	∞	4.50	1.000	6.16
FasNet+TAC [51]	8.6	-	2.37	0.771	29.80
MC-ConvTasNet [52]	10.8	-	2.78	0.844	23.10
MISO$_1$ [19]	10.2	-	3.05	0.859	14.0
LBT [8]	13.2	14.8	3.33	0.910	9.60
MISO$_1$-BF-MISO$_3$ [19]	15.6	-	3.76	0.942	8.30
TF-GridNet (1-stage) [9]	19.9	21.2	3.89	0.966	6.92
TF-GridNet (2-stage) [9]	22.8	24.9	4.08	0.980	6.76
SpatialNet [14]	25.1	27.1	4.08	0.980	6.70
SpatialNet$^\diamond$	24.8	26.9	4.15	0.985	6.66
TF-CrossNet	25.8	27.6	4.20	0.987	6.30

$^\diamond$ Trained with the same setup as TF-CrossNet

surpasses SpatialNet and TFGridNet in speaker separation performance, e.g. by larger than 0.1 PESQ improvement. Since the original SpatialNet is trained using a different setup, to make a fair comparison, we also train the SpatialNet with the same setup as TF-CrossNet, including the loss function and learning rate scheduler, and report the results in Table VI. Even though the use of the same training setup improves SpatialNet performance in terms of PESQ, WER, and eSTOI, it still underperformers TF-CrossNet, e.g. by 1 dB in SI-SDR and 5.7% relative WER. TF-CrossNet's WER of 6.30% is remarkably close to the oracle score of 6.16%. As TF-CrossNet has a similar architecture to SpatialNet, the superior performance of TF-CrossNet can be attributed to the proposed positional encoding and the GMHSA module.

E. Results on REVERB Challenge

Table VII reports the performance of speech dereverberation, enhancement, and recognition on the single- and 8-channel REVERB dataset. On the single-channel SimData, TF-CrossNet shows notable gains, achieving a PESQ score of 3.80 and substantially outperforming Wang and Wang [13] (3.29) and WPE (2.51). Additionally, it obtains the best scores in FWSegSNR (17.26) and SRMR (6.85), surpassing comparison methods. On the 8-channel SimData, TF-CrossNet exhibits strong performance, outperforming all the other methods in terms of CD (1.39) and LLR (0.16). Its PESQ score (4.04) is very close to that of SpatialNet (4.05), and its FWSegSNR score (19.95) is a little lower than that of SpatialNet (21.80). TF-CrossNet demonstrates superior dereverberation performance on the RealData, achieving the highest SRMR scores in both single- and 8-channel cases, representing large improvements over the unprocessed signals from 3.18 to 6.85 and 7.05 in single- and 8-channel respectively.

In terms of ASR results on SimData, TF-CrossNet produces strong performance on both single- and 8-channel datasets. On the single-channel SimData, it achieves a WER of 3.6/3.6 for far/near subsets, outperforming WPE (4.6/3.7). On the 8-channel dataset, TF-CrossNet further improves WER to 3.4/3.4, surpassing all other methods including SpatialNet (3.6/3.6) and WPE+BeamformIt (3.7/3.5). Our WER scores of 3.4/3.4 are close to those on the clean dataset [14]. On RealData, TF-CrossNet also achieves the best WER scores of 3.1/3.0 on the far/near 8-channel subsets, outperforming all other methods including SpatialNet (3.1/3.2), as well as the best WERs of 3.7/3.8 for the far/near single-channel datasets.

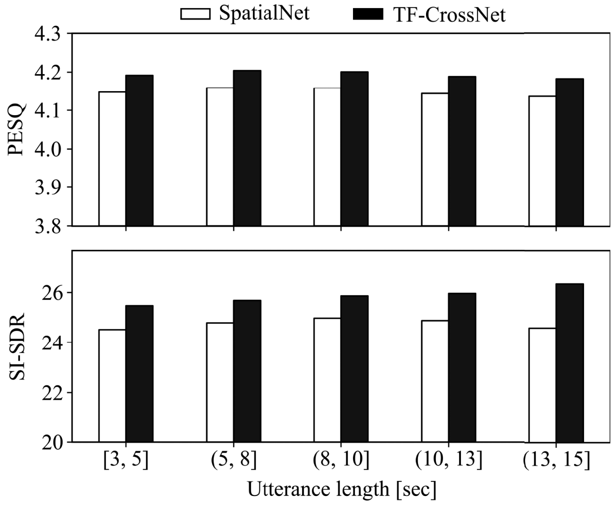

Fig. 3. Effects of sequence length on the performance of TF-CrossNet and SpatialNet. Speaker separation performance is plotted for different intervals of mixture lengths (in seconds).

The strong performance on both SimData and RealData indicates TF-CrossNet's generalizability and effectiveness in real-world conditions.

F. Performance Over Different Utterance Lengths

To assess the impact of utterance length on TF-CrossNet's performance, we plot the SI-SDR and PESQ scores across various sequence lengths on the 6-channel SMS-WSJ dataset. The results are depicted in Fig. 3. Note that, as described in Section IV-B, the length of training utterances is fixed at 3 seconds for this dataset. So the evaluated lengths are untrained. TF-CrossNet yields better performance than SpatialNet [14] across all sequence lengths. Both models have relatively consistent PESQ scores across different utterance lengths. TF-CrossNet also shows stable, even increasing, SI-SDR performance as sequence lengths increase, whereas SpatialNet exhibits slight degradation for sequences longer than 10 seconds, in line with the findings reported in [15].

To further assess the impact of positional encoding on TF-CrossNet's performance, we plot the SI-SDR and PESQ scores across different utterance lengths on the WHAMR! dataset in Fig. 4. Like in Fig. 3, the evaluated lengths are not included during training except for the shortest range of 1-4 seconds. The figure shows that TF-CrossNet with RCPE (black bars) consistently outperforms the model without positional encoding (white bars) across all utterance length ranges. Notably, the performance gap between the two models becomes more pronounced for longer utterances, with RCPE providing more benefit, particularly in SI-SDR. With LSTM (hatched bars), the model shows slightly

TABLE VII
SPEECH DEREVERBERATION, ENHANCEMENT AND RECOGNITION RESULTS ON THE SINGLE- AND 8-CHANNEL REVERB DATASETS

	SimData					RealData	
Method	CD↓	LLR↓	FWSegSNR↑	PESQ↑	WER↓	SRMR↑	WER↓
Unprocessed	5.08	0.67	8.32	2.37	4.9 / 3.7	3.18	6.5 / 5.9
Results on single-channel REVERB dataset							
WPE [53], [54]	4.95	0.63	9.38	2.51	4.6 / 3.7	3.83	5.8 / 5.5
Wang and Wang [13]	3.16	0.53	15.61	3.29	- / -	6.69	- / -
TF-CrossNet	2.18	0.16	17.26	3.80	3.6 / 3.6	6.85	3.7 / 3.8
Results on 8-channel REVERB dataset							
WPE [53], [54]	4.75	0.53	11.42	2.83	3.8 / 3.5	5.04	4.6 / 4.9
WPE+BeamformIt [39]	3.94	0.49	12.52	3.12	3.7 / 3.5	5.62	4.4 / 3.6
WPD [55], [56]	4.15	0.49	10.50	2.80	3.8 / 3.5	5.72	4.5 / 4.1
Wang and Wang [13]	2.78	0.39	18.75	3.71	- / -	6.30	- / -
SpatialNet [14]	2.26	0.29	21.80	4.05	3.6 / 3.6	6.88	3.1 / 3.2
TF-CrossNet	1.39	0.16	19.95	4.04	3.4 / 3.4	7.05	3.1 / 3.0

In WER columns the first number is for the far subset and the second is for the near subset.

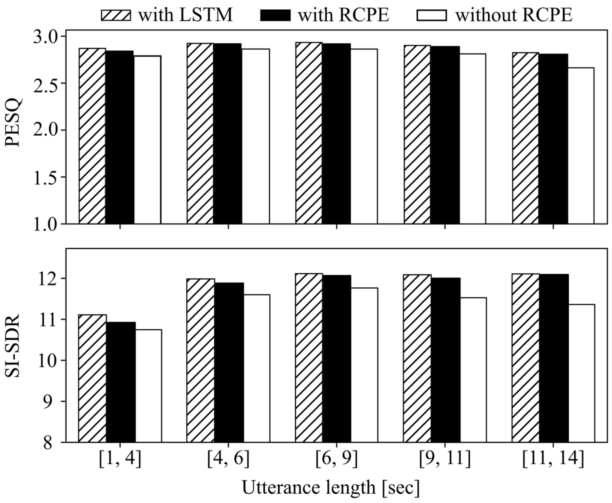

Fig. 4. Effects of positional encoding with respect to utterance lengths on the performance of TF-CrossNet on WHAMR! dataset.

better results. Both RCPE and LSTM exhibit relatively stable performance across various utterance lengths. This performance profile highlights the contribution of the proposed positional encoding, and is important for real-world applications where the length of mixture utterances may vary significantly.

G. Computational Complexity

Finally, we document computational load in terms of GFLOPs and the number of trainable parameters in millions (Params) of TF-CrossNet and several other methods. The complexities are tabulated in Table VIII for two sampling rates of 8 and 16 kHz. The computation of GFLOPs is as outlined in [14], where GFLOPs are quantified based on a four-second audio signal captured by a 6-channel microphone array speaker. The complexities of the comparison methods are obtained from [14].

TABLE VIII
COMPUTATIONAL COMPLEXITY AND MODEL SIZE OF THE PROPOSED MODEL AND COMPARISON METHODS

	8 kHz		16 kHz	
Model	GFLOPs↓	Params (M)↓	GFLOPs↓	Params (M)↓
FasNet+TAC [51]	26.7	2.7	27.9	2.8
MISO$_1$ [19]	-	-	81.9	8.58
LBT [8]	-	-	221.6	6.55
DasFormer [49]	33.3	2.2	76.4	2.2
TFGridNet [9]	348.4	11.0	695.6	11.2
SpatialNet [14]	119.0	6.5	237.9	7.3
TF-CrossNet	96.3	6.6	191.7	8.2

As clear from the table, TF-CrossNet exhibits much lower complexity than TF-GridNet. Compared to SpatialNet, TF-CrossNet has smaller GFLOPs and comparable numbers of trainable parameters.

In terms of actual time, training TF-GridNet on the SMS-WSJ dataset takes approximately 14 days on a single NVIDIA A100 GPU, whereas TF-CrossNet takes around 6 days. In terms of inference time, we measure the real-time factor (RTF), defined as the ratio of processing time to input signal duration, for a 6-channel, 4-second utterance on an NVIDIA V100 GPU. The resulting RTF values are 0.192 for SpatialNet, 0.657 for TF-GridNet, and 0.155 for TF-CrossNet.

VI. CONCLUDING REMARKS

We have introduced TF-CrossNet, a novel DNN architecture for single- and multi-channel speaker separation in noisy-reverberant environments. TF-CrossNet includes an encoder layer, a global multi-head self-attention module, cross-band and narrow-band modules, and an output layer, to leverage both global and local information in an audio signal to enhance speaker separation and speech enhancement performance. The global multi-head self-attention module allows the model to attend to any frame of interest in all feature and frequency channels, facilitating the exploitation of long-range dependencies. We introduce a novel random chunk positional encoding technique to improve generalization to longer sequences. The cross-band module captures cross-band correlations within the input signal, while the narrow-band module focuses on capturing correlations at neighboring frequency bins. The evaluation experiments conducted on multiple open datasets demonstrate that TF-CrossNet achieves state-of-the-art performance for single- and multi-channel speaker separation tasks. Moreover, TF-CrossNet exhibits stable performance in separating multi-talker mixtures of variable lengths, and is computationally efficient compared to recently-established strong baselines.

Acknowledgment

The authors thank the members of the Perception and Neurodynamics Lab (PNL), especially Hassan Taherian and Heming Wang, for insightful discussions and their suggestions during the development of TF-CrossNet. They also thank the three anonymous reviewers and the associate editor for their helpful comments, which have improved the manuscript. They also thank the three anonymous reviewers and the associate editor for their helpful comments, which have improved the manuscript.

References

[1] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[2] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.

[3] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[4] Y. Liu and D. L. Wang, “Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[5] 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, Aug. 2019.

[6] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 46–50.

[7] C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi, “Exploring self-attention mechanisms for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2169–2180, 2023.

[8] H. Taherian, K. Tan, and D. L. Wang, “Multi-channel talker-independent speaker separation through location-based training,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2791–2800, 2022.

[9] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, “TF-GridNet: Integrating full-and sub-band modeling for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[10] D. S. Williamson, Y. Wang, and D. L. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[11] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[12] K. Tan and D. L. Wang, “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[13] Z.-Q. Wang and D. L. Wang, “Deep learning based target cancellation for speech dereverberation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[14] C. Quan and X. Li, “SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1310–1323, 2024.

[15] H. Taherian and D. L. Wang, “Multi-channel conversational speaker separation via neural diarization,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[16] J. Rosendahl, V. A. K. Tran, W. Wang, and H. Ney, “Analysis of positional encodings for neural machine translation,” in *Proc. Spoken Lang. Transl.*, 2019. [Online]. Available: https://aclanthology.org/2019.iwslt-1.20/

[17] 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, pp. 2642–2646.

[18] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, “Hybrid LSTM-transformer model for emotion recognition from speech audio files,” *IEEE Access*, vol. 10, pp. 36018–36027, 2022.

[19] Z.-Q. Wang, P. Wang, and D. L. Wang, “Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[20] K. Tan, Z.-Q. Wang, and D. L. Wang, “Neural spectrospatial filtering,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[21] A. Ruoss et al., “Randomized positional encodings boost length generalization of transformers,” *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1889–1903.

[22] A. Vaswani et al., “Attention is all you need,” *Adv. Neural Info. Process. Sys.*, vol. 30, pp. 1–11, 2017.

[23] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. Interspeech*, 2020, pp. 5036–5040.

[24] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.

[25] Z. Pan, M. Ge, and H. Li, “A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction,” in *Proc. Interspeech*, 2022, pp. 1786–1790.

[26] 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.

[27] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, “WHAMR!: Noisy and reverberant single-channel speech separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 696–700.

[28] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, “SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition,” 2019, *arXiv:1910.13934*.

[29] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[30] D. Povey et al., “The kaldi speech recognition toolkit,” in *Proc. Workshop Autom. Speech Recognit. Understanding*, 2011. [Online]. Available: https://publications.idiap.ch/downloads/papers/2012/Povey_ASRU2011_2011.pdf

[31] K. Kinoshita et al., “A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP J. Adv. Signal Process.*, vol. 7, pp. 1–19, 2016.

[32] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 81–84.

[33] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, “The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2005, pp. 357–362.

[34] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,* 2018, pp. 351–355.

[35] 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, Jul. 2006.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 2001, pp. 749–752.

[37] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Tran. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[38] N. S. Detlefsen et al., "TorchMetrics-measuring reproducibility in pytorch," *J. Open Source Softw.*, vol. 7, 2022, Art. no. 4101.

[39] C. Li et al., "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 785–792.

[40] Z.-Q. Wang, "ESPnet2 pretrained model," 2023, doi: 10.5281/zenodo.7565926. [Online]. Available: https://zenodo.org/records/7565926

[41] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Proc. MultiMedia Model,* 2020, pp. 653–665.

[42] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM-RF: Efficient networks for universal audio source separation," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process.*, 2020, pp. 1–6.

[43] Y. Zhu, X. Zheng, X. Wu, W. Liu, L. Pi, and M. Chen, "DPTCN-ATPP: Multi-scale end-to-end modeling for single-channel speech separation," in *Proc. 2021 5th Int. Conf. Commun. Inf. Syst.*, 2021, pp. 39–44.

[44] M. W. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5759–5763.

[45] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-frequency domain path scanning network for speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6842–6846.

[46] J. Rixen and M. Renz, "SFSRNet: Super-resolution for single-channel audio source separation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 11220–11228.

[47] J. Rixen and M. Renz, "QDPN: Quasi-dual-path network for single-channel speech separation," in *Proc. Interspeech*, 2022, pp. 5353–5357.

[48] S. Zhao and B. Ma, "MossFormer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[49] S. Wang, X. Kong, X. Peng, H. Movassagh, V. Prakash, and Y. Lu, "DASFormer: Deep alternating spectrogram transformer for multi/single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[50] Z.-Q. Wang, G. Wichern, and J. L. Roux, "Convolutive prediction for reverberant speech separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 56–60.

[51] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6394–6398.

[52] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6389–6393.

[53] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[54] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Proc. IEEE 13th Speech Commun.; ITG- Symp.*, 2018, pp. 1–5.

[55] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, Jun. 2019.

[56] W. Zhang, A. S. Subramanian, X. Chang, S. Watanabe, and Y. Qian, "End-to-end far-field speech recognition with unified dereverberation and beamforming," in *Proc. Interspeech*, 2020, pp. 324–328.