Target Speaker Extraction using Multi-Stage Cross-Attention and Frequency-wise State Initialization

Hyeonseung Kim, *Student Member, IEEE,* and Jong Won Shin, *Senior Member, IEEE*

***Abstract*—Several recent target speaker extraction (TSE) models directly utilize enrollment speech without explicitly extracting low-dimensional speaker embeddings. However, these methods typically inject the speaker information only once at the input of the speaker extraction network, which may be insufficient because the conditioning information can become diluted as it propagates through repeated separator blocks. In this letter, we propose a TSE model built upon the TF-GridNet, which is a speech separation model performing dual-path modeling in the time-frequency domain with cross-frame self-attention modules. In the proposed TSE model, the self-attention modules in the first M separator blocks are replaced by cross-attention between the enrollment speech and the mixture signal, providing speaker information in multiple stages without introducing additional parameters or computation compared with the original TF-GridNet blocks. In addition, the initial hidden and cell states of the inter-frame long short-term memory (LSTM) modules are determined for each frequency from the enrollment speech. As the pattern of the temporal correlation may be different for each frequency depending on the pitch and speaking style, speaker-dependent frequency-wise state initialization would be helpful. Experimental results showed that the proposed TSE model demonstrated the best PESQ scores and comparable SI-SDRs with lower computational complexity.**

***Index Terms*—Target speaker extraction, cross-attention, initial hidden state, inter-frame LSTM**

I. Introduction

TARGET speaker extraction (TSE) aims to isolate the speech of the enrolled speaker from the input mixture signal. The performance of TSE has been significantly improved by virtue of the recent advancements in speech separation. For instance, SpEx [1] and its variants [2], [3], [4], [5] were based on Conv-TasNet [6], and other TSE models such as X-DPRNN [7], VEVEN [8], X-SepFormer [9], and X-TF-GridNet [10] can also be regarded as extensions of the corresponding speech separation models. Recently, time-frequency (TF) domain speech separation models [11], [12] have been proposed using a dual-path structure [13], [14]. Unlike time-domain dual-path models which utilize local and global sequence modules, TF domain models employ intra-frame full-band modules and sub-band temporal modules. Many recent TSE models utilize these TF-domain speech separation models as backbones [10], [15], [16], [17], [18] or time-domain dual-path models in the TF-domain [19], [20].

Early approaches to TSE typically obtain low-dimensional speaker embeddings from the enrollment speech and utilize them to extract target speeches [1], [2], [3], [4], [5]. Recently, several works [16], [18], [19], [20], [21] have been proposed to directly model the relationship between the enrollment and input mixture signals using cross-attention modules and demonstrated superior performances. These approaches, however, rely on a single fusion module at the input of the extractor network, which may make the enrollment speech information diluted in the later stages in the extractor network.

Another limitation of the existing TSE models with TF-domain dual-path structures is that they adopt identical sub-band temporal modules for all frequencies and treat the exactly in the same way. As the speech has a different pattern of temporal correlation for each frequency depending on the pitch frequency and speaking style, it would be beneficial to process each sub-band differently. One possible way is to set initial hidden and cell states of LSTMs differently. Initial hidden and cell states were exploited in several previous works [22], [23]. For instance, skipping memory LSTM (SkiM) [22] demonstrated that providing adequate initial hidden and cell states to LSTM layers can improve the speech separation performance. Predictive SkiM (pSkiM) [23] further enhanced these initial hidden and cell states by using contrastive predictive coding to predict future states, successfully improving speech separation quality relative to SkiM.

In this letter, we propose a TSE method based on TF-GridNet [12] that addresses the aforementioned limitations. First, to integrate enrollment information in multiple stages more efficiently, we replace some of the cross-frame self-attention modules within the TF-GridNet blocks with cross-attention that directly utilizes the enrollment signal. Second, we introduce frequency-dependent, speaker-aware conditioning by extracting initial hidden and cell states for the sub-band temporal LSTM modules from the enrollment speech. Both the cross-attention embeddings and the initial hidden and cell states are obtained from a single module. This strategy results in only a slight increase in complexity over the baseline separation model. Experimental results demonstrated that our proposed model achieved superior PESQ scores and comparable scale-invariant signal-to-distortion ratios (SI-SDRs) compared to existing TSE models on various datasets with lower computational complexity.

Manuscript received –
This research was supported by the by the IITP (Institute of Information Communications Technology Planning Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2021-II211835).

Hyeonseung Kim and Jong Won Shin are with the Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Korea (e-mail: kimhs355@gm.gist.ac.kr; jwshin@gist.ac.kr)

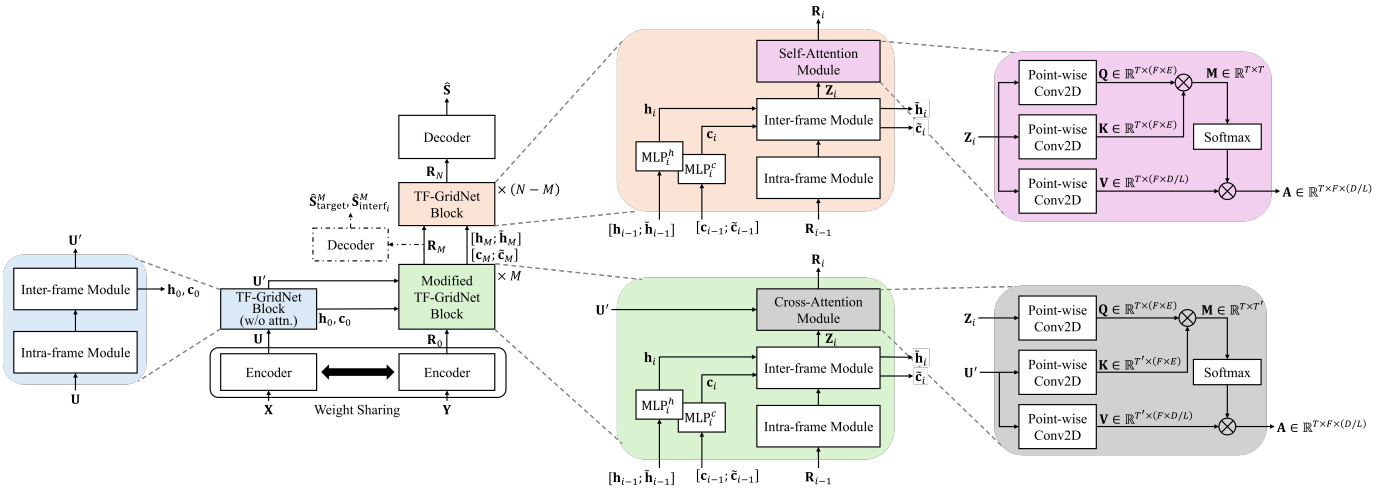

Fig. 1: Overall structure of of the proposed model. A single, shared encoder is used for both the enrollment network and the speaker extraction network (SEN). The enrollment network employs a single TF-GridNet block without an attention module (blue), while the SEN uses the modified TF-GridNet blocks (green) with cross-attention modules (gray) and the original TF-GridNet blocks (orange) which contain self-attention modules (purple). MLP indicates multi-layer perceptron.

II. Related Works

A. TF-GridNet

The TF-GridNet [12] processes the stacked real and imaginary components of the complex STFT as input. It employs a two-dimensional convolution layer as the encoder to generate embeddings for each T-F unit. Before applying the intra-frame LSTM to the spectral sequence, an unfold operation with kernel size of I and stride of J is applied to the frequency sequence. This operation stacks I neighboring frequency bins into one frequency band. The stacked frequency band sequence is then passed through the intra-frame LSTM, followed by a deconvolution layer with the kernel size of I and stride of J instead of a fold operation. The inter-frame LSTM is applied similarly to the intra-frame LSTM, with the difference being that the unfold operation and LSTM are applied to the temporal sequence. The output of the inter-frame LSTM is then passed through the cross-frame multi-head self-attention module with L heads to capture long-range contextual information. A combination of the intra-frame and inter-frame LSTMs, along with the cross-frame self-attention module, forms a TF-GridNet Block.

B. TSE with Cross-Attention using Enrollment Signal

While conventional TSE methods rely on speaker embeddings generated by pre-trained or jointly trained extractor networks, some recent researches exploit the enrollment utterance directly via cross-attention [16], [18], [19], [20], [24], [25], [26]. For instance, CIENet [19], [20] applies cross-attention for real and imaginary parts of the compressed spectrum for the enrollment and mixture. Similarly, USEF-TSE [18] performs cross-attention between the outputs of the convolutional encoders for the enrollment and mixture signal. In both of the methods, the output of the cross-attention is concatenated with the encoded mixture and then passed to the speech extraction network (SEN).

C. Predictive Skipping Memory LSTM

SkiM [22] is a variant of the DPRNN [13] architecture that replaces the standard inter-chunk LSTM with a module called Mem-LSTM. This Mem-LSTM module receives the sequence of hidden and cell states from the preceding intra-chunk LSTM block and generate the initial states for the subsequent intra-chunk LSTM. Building upon this, pSkiM [23] introduces an additional contrastive predictive coding (CPC) loss. This loss function is designed to train the initial hidden and cell states generated by the Mem-LSTM to be predictive of the final states of subsequent segments. This objective encourages the model to learn representations that capture long-range temporal dependencies more effectively, thereby enhancing its overall temporal modeling capabilities.

III. Proposed Method

The overall structure of the proposed method is shown in Fig. 1, which is based on the TF-GridNet [12] explained in Section II-A. It uses a single TF-GridNet block without the self-attention module to extract speaker information from an enrollment speech, which is used to generate embeddings for cross-attention and initial hidden and cell states of the inter-frame LSTMs in the SEN based on the TF-GridNet. The cross-frame multi-head self-attention modules in the first M TF-GridNet blocks of the SEN are replaced by multi-head cross-attention modules exploiting speaker information. In this way, we can provide speaker information multiple times without introducing excessive computation and model parameters.

Let $\mathbf{X} \in \mathbb{R}^{T' \times F \times 2}$ and $\mathbf{Y} \in \mathbb{R}^{T \times F \times 2}$ be the real and imaginary parts of the short-time Fourier transform (STFT) of the enrollment speech and the mixture, respectively, where T' and T are the number of frames in the enrollment and mixture and F is the number of frequency bins. $\mathbf{X}$ and $\mathbf{Y}$ are processed by the same encoder, which was proven to be effective in [2], to produce $\mathbf{U} \in \mathbb{R}^{T' \times F \times D}$ and $\mathbf{R}_0 \in \mathbb{R}^{T \times F \times D}$, respectively.

$\mathbf{U}$ is then processed by the intra-frame module and the inter-frame module of the TF-GridNet block to produce the output $\mathbf{U}'$ along with the frequency-wise hidden and cell states for the last frame in each direction, $\mathbf{h}_0 \in \mathbb{R}^{F\times 2\times H}$ and $\mathbf{c}_0 \in \mathbb{R}^{F\times 2\times H}$, in which '2' corresponds to the bidirectional output. The initial hidden and cell states for the LSTMs in the speaker encoder are set to zero.

The SEN consists of an encoder, M modified TF-GridNet blocks, $(N-M)$ TF-GridNet blocks, and a decoder. In the first M blocks, the cross-frame self-attention module is replaced by a cross-attention module as shown in Fig. 1. In a head of the cross-attention module, the output from the inter-frame module $\mathbf{Z}_i$ is transformed into query $\mathbf{Q} \in \mathbb{R}^{T\times(F\times E)}$ as in the self-attention module, while $\mathbf{U}'$ is used to obtain key $\mathbf{K} \in \mathbb{R}^{T'\times(F\times E)}$ and value $\mathbf{V} \in \mathbb{R}^{T'\times(F\times D/L)}$. With this cross-attention, the parts of $\mathbf{Z}_i$ that have similar patterns with $\mathbf{U}'$ would get higher weights, which would be beneficial to extract target speech. In the later $(N-M)$ blocks, the self-attention modules were used as in the original TF-GridNet to focus on the reconstruction of the target speech itself, exploiting global temporal correlation.

The initial hidden and cell states for the inter-frame LSTMs in the i-th modified or normal TF-GridNet block, $\mathbf{h}_i \in \mathbb{R}^{F\times 2\times H}$ and $\mathbf{c}_i \in \mathbb{R}^{F\times 2\times H}$, are computed from the initial and final output states, $(\mathbf{h}_{i-1}, \mathbf{c}_{i-1})$, $(\tilde{\mathbf{h}}_{i-1}, \tilde{\mathbf{c}}_{i-1})$ of the preceding $(i-1)$-th block, using separate multi-layer perceptrons (MLPs), each with one hidden layer:

$$\mathbf{h}_i = \text{MLP}_i^h(\left[\mathbf{h}_{i-1}; \tilde{\mathbf{h}}_{i-1}\right]), \tag{1}$$

$$\mathbf{c}_i = \text{MLP}_i^c([\mathbf{c}_{i-1}; \tilde{\mathbf{c}}_{i-1}]), \tag{2}$$

where $[;]$ denotes the concatenation operation. For the first block, $\mathbf{h}_0$ and $\tilde{\mathbf{c}}_0$ are set as zero tensors. The output of the final TF-GridNet block, $\mathbf{R}_N$, is passed through the decoder to generate the complex spectrogram of the extracted speech, $\hat{\mathbf{S}} \in \mathbb{R}^{T\times F\times 2}$.

As a loss function to train the proposed TSE system, we employ the negative SI-SDR [27] evaluated on $\hat{\mathbf{S}}$ and the intermediate separation results along with the CPC loss proposed in [23]. The complex spectrograms for separated signals from the first M modified TF-GridNet blocks are obtained by applying another decoder to the output of the M-th block, $\mathbf{R}_M$. Among the C outputs when considering $C-1$ interfering speakers, the first output, $\hat{\mathbf{S}}_{\text{t}}^M \in \mathbb{R}^{T\times F\times 2}$, is always regarded as the estimate for the target speech because the model is informed by the enrollment speech, while other outputs, $\hat{\mathbf{S}}_{\text{i}_c}^M \in \mathbb{R}^{T\times F\times 2}$, $c \in \{1, 2, ..., C-1\}$, need to be mapped to the interfering speeches using PIT [28]. Both the final output $\hat{\mathbf{S}}$ and the intermediate outputs $\hat{\mathbf{S}}_{\text{t}}^M$ and $\hat{\mathbf{S}}_{\text{i}_c}^M$ are transformed into the time domain signals $\hat{\mathbf{s}}$, $\hat{\mathbf{s}}_{\text{t}}^M$, and $\hat{\mathbf{s}}_{\text{i}_c}^M$, $c \in \{1, 2, ..., C-1\}$ using inverse STFT to compute the negative SI-SDR loss. The overall loss function is given by

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{S}}(\mathbf{s}, \hat{\mathbf{s}}) + \frac{\lambda_1}{C}\left(\mathcal{L}_{\text{S}}(\mathbf{s}, \hat{\mathbf{s}}_{\text{t}}^M) + \min_{\pi\in P}\sum_{c=1}^{C-1}\mathcal{L}_{\text{S}}\left(\mathbf{s}_{\text{i}_c}, \hat{\mathbf{s}}_{\text{i}_{\pi(c)}}^M\right)\right) \\ &+ \frac{\lambda_2}{FN}\sum_{f=1}^{F}\sum_{i=1}^{N}\left(\mathcal{L}_{\text{C}}\left(\mathbf{h}_i^f, \tilde{\mathbf{h}}_i^f\right) + \mathcal{L}_{\text{C}}\left(\mathbf{c}_i^f, \tilde{\mathbf{c}}_i^f\right)\right),\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{S}}(\mathbf{x}, \mathbf{y})$ indicates the negative SI-SDR loss between $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{s}$ and $\mathbf{s}_{\text{i}_c}$ are the target and interfering speeches. P is a set of all possible permutations, and λ_1 and λ_2 are the weights of the loss for the intermediate outputs and CPC loss, respectively. The CPC loss $\mathcal{L}_{\text{C}}(\mathbf{x}, \tilde{\mathbf{x}})$ is defined as

$$\mathcal{L}_{\text{C}}(\mathbf{x}, \tilde{\mathbf{x}}) = -\mathbb{E}\left[\log\frac{\exp\left(\tilde{\mathbf{x}}^\top \mathbf{P}_{\mathbf{x}}\mathbf{x}\right)}{\sum_{\mathbf{z}\in\mathcal{Z}}\exp\left(\tilde{\mathbf{z}}^\top \mathbf{P}_{\mathbf{z}}\mathbf{x}\right)}\right], \tag{4}$$

where $\mathbf{P}_{\mathbf{x}}$ is a projection matrix for the initial state $\mathbf{x}$ to estimate the corresponding final state $\tilde{\mathbf{x}}$, and $\mathcal{Z}$ is a set including $\mathbf{x}$ and negative samples from different frequencies, layers, or training samples. While the CPC loss in the previous work [23] focused on the capability to predict the future states, $\mathcal{L}_{\text{C}}(\mathbf{x}, \tilde{\mathbf{x}})$ in (4) is on the predictability of the final state from the initial state. As the frequency-dependent speaker information within $(\mathbf{h}_i^f, \mathbf{c}_i^f)$ should remain in the last frame, it is helpful to enforce the final states to be predictable to a certain extent from the initial state.

IV. Experiments

A. Experimental configurations

We performed experiments on the *min* version of the Libri2Mix [29] dataset, with all audio sampled at 16 kHz. The models were trained on the `train-100` subset, which consisted of 13,900 mixtures generated from 251 speakers. The validation and test sets are 3,000 utterances each, spoken by non-overlapping 40 speakers, respectively. The enrollment speech for each mixture was prepared as in [37]. To compare with a wider variety of TSE models, we also evaluated our model on the WSJ0-2mix [30], WHAM! [31], and WHAMR! [32] datasets with a sampling rate of 8 kHz. The WSJ0-2mix dataset is a widely-used benchmark for speech separation comprising 20,000, 5,000, and 3,000 samples for training, validation, and evaluation. The training and validation sets share the same 101 speakers, while the test data are from other 18 speakers. We used *min* version with the enrollment speech used in [1]. The WHAM! [31] and WHAMR! [32] datasets are the extensions of the WSJ0-2mix dataset designed to assess the robustness of speech separation to noise and reverberation.

The configuration for the proposed model followed the 7th row of Table XIII in [12]. Specifically, we used a 16 ms window with an 8 ms hop size for the STFT, and F was 129 and 65 for 16 kHz and 8 kHz sampling rate, respectively. The output dimension of the encoder, D, was set to 32. The kernel size I and stride size J for the unfold operation were both set to 4, and the number of hidden units in the LSTM layers, H, was set to 128. The number of heads L in a self- or cross-attention module was 4 with an embedding dimension E of 4 for 16 kHz and 8 for 8 kHz experiments, which made the corresponding hidden dimension for the attention module approximately 512. The total number of blocks N and the number of modified TF-GridNet blocks M were set to 6 and 4, respectively. C was set to 2 as the dataset we tested contained two speakers. The models for the Libri2Mix dataset were trained for 86 epochs, while those for the WSJ0-2mix, WHAM!, and WHAMR! datasets were trained for 153 epochs. The weight of the loss for the intermediate outputs and the

TABLE I: Performance of the target speaker extraction models on the *min* version of the Libri2Mix dataset sampled at 16 kHz.

Model	PESQ	SDR (dB)	SI-SDR (dB)
Mixture	1.15	0.1	0.0
TF-GridNet (separation)	2.84	16.3	17.5
SSL-TD-SpeakerBeam [37]	2.45	15.2	14.7
ResNet34-BSRNN [17]	-	15.2	14.6
SHuBERT-BSRNN [15]	-	16.0	15.4
EcapaTDNN-BSRNN [16]	-	-	15.9
Proposed	**3.19**	18.1	**17.9**
w/o hc initialization	3.14	17.6	17.5
Proposed (M=1)	3.16	**18.2**	17.6

TABLE II: Performance of the target speaker extraction models on the WSJ0-2mix dataset.

Model	GMAC/s	PESQ	SDR (dB)	SI-SDR (dB)
Mixture	-	2.02	0.2	0.0
X-TF-GridNet [10]	68.32	3.70	20.4	19.7
Large [10]	113.24	3.77	21.7	20.7
CIENet-C2F				
-mDPRNN [20]	22.96	3.88	21.6	21.2
-mDPTNet [20]	25.88	<u>3.94</u>	22.3	21.9
USEF-Sepformer [18]	45.8	3.67	20.1	19.9
USEF-TFGridNet [18]	125.3	3.92	**23.7**	**23.3**
Proposed	10.81	3.86	21.0	20.7
w/o hc initialization	10.80	3.87	20.9	20.6
Proposed (high resol.)	22.16	**3.98**	<u>22.6</u>	<u>22.3</u>

CPC loss, λ_1 and λ_2, were set to 0.5 and 1.0, respectively. The Adam optimizer and cosine annealing scheduler with a learning rate ranging from 1e-3 to 1e-7 were used for optimization. The models with the lowest validation losses were chosen. A separation model with the same configuration was also implemented for performance comparison.

We evaluated the performance of the model using SDR [33], SI-SDR [27], and Perceptual Evaluation of Speech Quality (PESQ) [34], [35] scores. Additionally, we compared the computational complexity in terms of Giga multiply-accumulate operations per second (GMAC/s) using ptflops [36].

B. Experimental results

Table I presents the performance of TSE models on the Libri-2mix dataset. The proposed method outperformed the TF-GridNet speech separation model with oracle target speech selection and other TSE models in all metrics. The performance of the proposed method without hidden and cell state initialization is also given to compare the performance improvement brought by the cross-attention and the frequency-dependent states initialization. Compared with the separation model with oracle target speech selection, the proposed cross-attention module showed the 0.3 higher average PESQ score, and the state initialization further improve all three metrics. We have also measured the performances with $M = 1$ to verify if the cross-attention should be applied across multiple stages. Slightly higher performances were observed with $M = 4$, without introducing any additional computation.

The performances on the WSJ0-2mix, WHAM! and WHAMR! datasets sampled at 8 kHz are compared in Table II, III, and IV, respectively. The best performance in each metric is marked as bold, and the second one is underlined. As the recently proposed models compared in the tables require much higher computational loads than the proposed method as shown in Table II, we have also present the performances of a high temporal resolution variant of the proposed model for which the stride of the unfold operation was set to $J = 2$ instead of $J = 4$, doubling the temporal resolution. The high resolution variant of the proposed TSE model outperformed the X-TF-GridNet [10] and CIENet-C2F [20] in all metrics except the SDR and SI-SDR for the WHAMR! dataset with a lower computational cost. While the USEF-TFGridNet [18] showed higher SDR and SI-SDR on the WSJ0-2mix and WHAM! datasets, the proposed method provides a lightweight and efficient alternative. We can also observe that the speaker-dependent frequency-wise initialization of the hidden and cell states only had a marginal effect on the performances for these three datasets with the sampling rate of 8 kHz, compared with the experimental results on the Libri2Mix dataset sampled at 16 kHz. It suggests that the benefits of this frequency-dependent conditioning were more pronounced on wideband signals, in which more frequency bins in a wider frequency range may have more diverse patterns of temporal correlation. It is noted that the computational cost of the proposed model with $J = 4$ for the experiments on the Libri2Mix dataset was 20.18 GMAC/s. The code is available at https://github.com/kimhs355/MCFS-TSE.

TABLE III: Performance of the target speaker extraction models on the WHAM! dataset.

Model	PESQ	SDR (dB)	SI-SDR (dB)
Mixture	1.43	-4.2	-4.5
X-TF-GridNet [10]	-	11.6	10.8
CIENet-C2F-mDPRNN [20]	2.59	12.4	11.8
CIENet-C2F-mDPTNet [20]	<u>2.77</u>	13.4	12.8
USEF-Sepformer [18]	-	11.3	10.6
USEF-TFGridNet [18]	-	**13.7**	**13.1**
Proposed	<u>2.77</u>	12.7	12.2
w/o hc initialization	2.76	12.7	12.2
Proposed (high resol.)	**2.86**	<u>13.5</u>	<u>12.9</u>

TABLE IV: Performance of the target speaker extraction models on the WHAMR! dataset.

Model	PESQ	SDR (dB)	SI-SDR (dB)
Mixture	1.41	-3.5	-6.1
X-TF-GridNet [10]	-	10.3	8.5
CIENet-C2F-mDPRNN [20]	2.60	11.7	10.6
CIENet-C2F-mDPTNet [20]	<u>2.72</u>	**12.5**	**11.4**
USEF-Sepformer [18]	-	6.8	5.1
USEF-TFGridNet [18]	-	11.4	10.0
Proposed	2.66	11.2	10.0
w/o hc initialization	2.63	11.0	9.8
Proposed (high resol.)	**2.79**	<u>12.2</u>	<u>11.0</u>

V. CONCLUSION

In this letter, we proposed a TSE model that efficiently integrates speaker information into a TF-GridNet backbone. This was achieved through two main contributions: replacing self-attention with cross-attention in multiple stages, and generating speaker-dependent initial states for frequency-wise LSTM guidance. Our proposed model outperformed previous methods for the Libri2Mix dataset sampled at 16 kHz, and exhibited the best PESQ scores and comparable SDRs and SI-SDRs to other leading ones for the WSJ0-2mix, WHAM!, and WHAMR! dataset, with a lower computational complexity.

REFERENCES

[1] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," in *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 1370-1384, Apr. 2020.

[2] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A Complete Time Domain Speaker Extraction Network," in *Proc. Interspeech 2020*, Oct. 2020, pp. 1406-1410.

[3] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, H. Li, "Multi-Stage Speaker Extraction with Utterance and Frame-Level Reference Signals," in *2021 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Jun. 2021, pp. 6109-6113.

[4] W. Wang, C. Xu, M. Ge, H. Li, "Neural Speaker Extraction with Speaker-Speech Cross-Attention Network," in *Proc. Interspeech 2021*, Sep. 2021, pp. 3535-3539.

[5] J. Chen *et al.*, "MC-SpEx: Towards Effective Speaker Extraction with Multi-Scale Interfusion and Conditional Speaker Modulation," in *Proc. Interspeech 2023*, Aug. 2023, pp. 4034-4038.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 8, pp. 1256-1266, Aug. 2019.

[7] Y. Hao, J. Xu, J. Shi, P. Zhang, L. Qin, and B. Xu, "A Unifiend Framework for Low-Latency Speaker Extraction in Cocktail Party Environments," in *Proc. Interspeech 2020*, Oct. 2020, pp. 1431-1435.

[8] L. Yang, W. Liu, L. Tan, J. Yang, H.-G. Moon, "Target Speaker Extraction with Ultra-Short Reference Speech by VE-VE Framework," in *2023 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Jun. 2023, pp. 1-5.

[9] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-Sepformer: End-to-End Speaker Extraction Network with Explicit Optimization on Speaker Confusion," in *2023 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Jun. 2023, pp. 1-5.

[10] F. Hao, X. Li, and C. Zheng, "X-TF-GridNet: A Time-Frequency Domain Target Speaker Extraction Network with Adaptive Speaker Embedding Fusion," in *Information Fusion*, vol. 112, pp. 1-16, Jun. 2024.

[11] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-Freuqnecy Domain Path Scanning Network for Speech Separation," in *2022 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 2022, pp. 6842-6846.

[12] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating Ful- and Sub-Band Modeling for Speech Separation," in *IEEE/ACM Trans. Aduio, Speech, and Lang. Process.*, vol. 31, pp. 3221-3236, Aug. 2023.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *2020 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 2020, pp. 46-50.

[14] 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*, Oct. 2020, pp. 2642-2646.

[15] J. Lin, M. Ge, W. Wang, and H. Li, "Selective HuBERT: Self-Supervised Pre-Training for Target Speaker in Clean and Mixture Speech," in *IEEE Signal Process. Lett.*, vol. 31, pp. 1014-1018, Apr. 2024.

[16] K. Zhang *et al.*, "Multi-Level Speaker Representation for Target Speaker Extraction," in *2025 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 2025, pp. 1-5.

[17] J. Li *et al.*, "On the Effectiveness of Enrollment Speech Augmentation for Target Speaker Extraction," in *IEEE Spoken Lang. Technol. Workshop*, Dec. 2024, pp. 325-332.

[18] B. Zeng and M. Li, "USEF-TSE: Universal Speaker Embedding Free Target Speaker Extraction," in *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 33, pp. 2110-2124, May, 2025.

[19] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target Speaker Extraction by Directly Exploiting Contextual Information in the Time-Frequency Domain," in *2024 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 2024, pp. 10476-10480.

[20] X. Yang, C. Bao, and X. Chen, "Coarse-to-Fine Target Speaker Extraction Based on Contextual Information Exploitation," in *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 32, pp. 3795-3810, Aug. 2024.

[21] K. Xue, R. Fan, C. Sun, P. Zhao, and J. An, "DCF-Net: Efficient Target Speaker Extraction by Leveraging Mixture and Enrollment Interactions," in *IEEE Signal Process. Lett.*, vol. 32, pp. 3240-3244, Aug. 2025.

[22] C. Li, L. Yang, W. Wang, and Y. Qian, "SkiM: Skipping Memory LSTM for Low-Latency Real-Time Continuous Speech Separation," in *2022 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 2022, pp. 681-685.

[23] C. Li, Y. Wu, and Y. Qian, "Predictive SkiM: Contrastive Predictive Coding for Low-Latency Online Speech Separation," in *2023 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Jun. 2023, pp. 1-5.

[24] X. Xiao *et al.*, "Signel-Channel Speech Extraction Using Speaker Inventory and Attention Network," in *2019 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 2019, pp. 86-90.

[25] B. Zeng, H. Suo, Y. Wan, and M. Li, "SEF-Net: Speaker Embedding Free Target Speaker Extraction Network," in *Proc. Interspeech 2023*, Aug. 2023, pp. 3452-3456.

[26] Y. Hu, H. Xu, Z. Guo, H. Huang, and L. He, "SMMA-Net: An Audio Clue-Based Target Speaker Extraction Network with Spectrogram Matching and Mutual Attention," in *2024 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 2024, pp. 1496-1500.

[27] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?," in *2019 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 2019, pp. 626-630.

[28] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks," in *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 10, pp. 1901-1913, Jul. 2017.

[29] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," *arXiv preprint arXiv:2005.11626*, 2020.

[30] Y. Isik, J. R. Hershey, Z. Chen, S. Watanabe, and J. L. Roux "Single-Channel Multi-Speaker Separation Using Deep Clustering," in *Proc. Interspeech 2016*, Sep. 2016, pp. 545-549.

[31] G. Wichern *et al.*, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech 2019*, Sep. 2019, pp. 1368-1372.

[32] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *2020 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May. 2020, pp. 696-700.

[33] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," in *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, Jun. 2006.

[34] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Rec. ITU-T P. 862, ITU, Geneva, Switzerland, 2001.

[35] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, Rec. ITU-T P. 862.2, ITU, Geneva, Switzerland, 2007.

[36] V. Sovrasov. ptflops: a flops counting tools for neural networks in pytorch framework. [Online]. Available: http://github.com/sovrasov/flops-counter.pytorch

[37] J. Peng *et al.*, "Target Speech Extraction with Pre-Trained Self-Supervised Learning Models," in *2024 IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 2024, pp. 10421-10425.