REAL-TIME SINGLE-CHANNEL SPEAKER-CONDITIONED TARGET SPEAKER EXTRACTION USING TCN-CONFORMER WITH EFFICIENT SELF-ATTENTION MECHANISMS

Ragini Sinha*, Christian Rollwage*, Simon Doclo*†
* Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany
† Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

***Abstract*—Speaker-conditioned target speaker extraction systems aim to extract the target speaker from a mixture of speakers by utilizing auxiliary information about the target speaker. Typically, such systems consist of a speaker embedder network and a speaker separator network. While self-attention mechanisms have demonstrated remarkable performance in speech processing tasks, including target speaker extraction, their high memory usage and computational complexity pose challenges for real-time applications. To address these limitations, we integrate a linear self-attention mechanism into the separator network, significantly reducing memory and computational costs, and thereby making the system more suitable for real-time applications. Furthermore, we evaluate the performance of this linear self-attention-based speaker extraction system against a system using memory-efficient self-attention. Experimental results on two-speaker, three-speaker, and noisy two-speaker mixtures show that linear self-attention not only improves speaker extraction performance compared to both traditional and memory-efficient self-attention but also significantly reduces the real-time factor and computational cost.**

***Index Terms*—Target speaker extraction, efficient self-attention, real-time, conformer, TCN**

I. Introduction

In many applications, it is important to extract a target speaker from overlapping speech recordings in a noisy environment. Traditional approaches like blind source separation (BSS) [1]–[5] can be utilized to first estimate all sources from the mixture of speakers and then select the target speaker. As a more direct approach, speaker-conditioned target speaker extraction algorithms have been proposed, which estimate the target speaker from the mixture utilizing auxiliary information about the target speaker [6]–[18]. Commonly used auxiliary information includes reference speech [6]–[8], [10]–[12], video [13], [14], speech activity [15] or directional cues [16], [17]. In this paper, we focus on single-channel target speaker extraction utilizing reference speech as auxiliary information.

The Oldenburg Branch for Hearing, Speech and Audio Technology HSA is funded in the program ≫Vorab≪ by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation for its further development.

Typically, speaker-conditioned target speaker extraction systems consist of two networks: a speaker embedder network and a speaker separator network, which are trained either separately or jointly. The speaker embedder network generates a speaker embedding from the reference speech of the target speaker. This speaker embedding guides the separator network in extracting the target speaker from the mixture. Some separator networks estimate a time-frequency mask [7]–[9], while other separator networks directly estimate the target speaker in the time domain [10]–[12]. Various architectures have been explored for both networks. For instance, LSTM-based architectures have been used for the speaker embedder network in [7], [9], [10], whereas ResNet-based architectures have been used in [8], [11], [12]. For the separator network, a convolutional long short-term memory (CNN-LSTM) architecture has been used in [7], [9], whereas an attention-based architecture has been in [8], [12], and temporal convolutional neural network (TCN)-based architectures have been used in [10], [11].

Recently, self-attention (SA) [19] has received significant attention for target speaker extraction [8], [12], [13] due to its ability to capture complex dependencies, support parallel processing, and its flexibility in handling diverse input features. Despite its impressive performance, traditional SA often comes with high memory and computational costs [19]. The memory and computational costs of traditional SA scale quadratically with the length of the mixture signal, making it resource-intensive and challenging for real-time applications. To overcome these limitations, various SA variants have been proposed [20]–[22], which aim to reduce the memory and computational costs without affecting the performance. Some of these variants have been successfully applied to speech enhancement and separation [5], [23], [24]. However, to the best of our knowledge, the potential benefits of these SA variants have not been explored specifically for real-time speaker-conditioned target speaker extraction.

In this paper, we consider a TCN-Conformer-based baseline system proposed in [12], which utilizes a ResNet-based

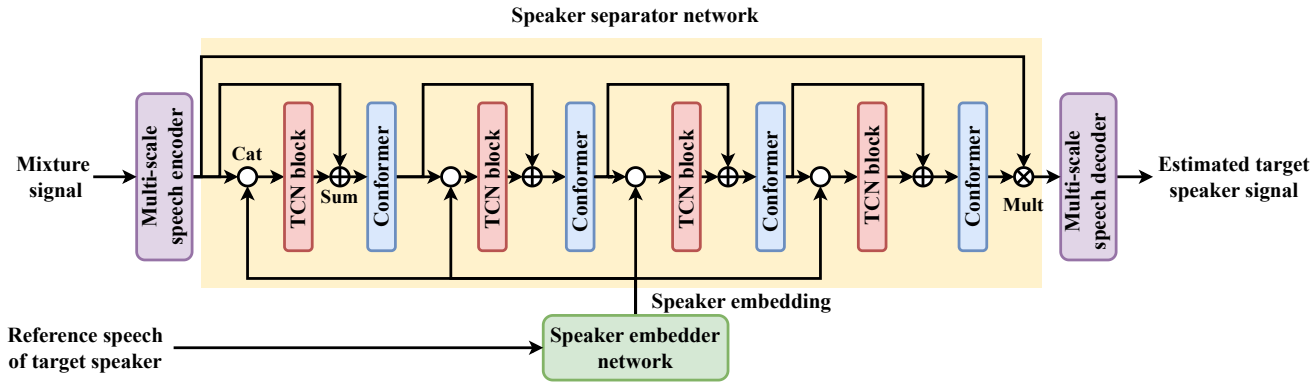

Figure 1. TCN-Conformer system for single-channel target speaker extraction, either using traditional or memory-efficient or linear self-attentions in the conformer blocks. "Cat" represents the concatenation operation.

speaker embedder network and a TCN-Conformer-based separator network to perform target speaker extraction in the time domain (see Fig. 1). The separator network is composed of TCN and conformer blocks, where each conformer block utilizes traditional multi-head self-attention (MHSA). Aiming at reducing both memory and computational costs, we propose two key modifications. First, we replace the traditional MHSA in each conformer block with a linear MHSA [21] (resulting in a linear TCN-Conformer system), which scales linearly with input length. Second, we explore several system variants by progressively reducing the total number of parameters by a factor of 2 (Medium), 4 (Small), and 8 (XSmall). Additionally, we investigate the effect of a memory-efficient MHSA [22], which specifically targets memory reduction in the considered baseline system. Experimental results on both clean and noisy mixtures show that all causal variants of the proposed linear TCN-Conformer system significantly improve target speaker extraction performance compared to both the corresponding causal baseline system with traditional and memory-efficient MHSA. Furthermore, the proposed linear TCN-Conformer systems achieve a substantial reduction in the real-time factor (RTF) and computational costs, making them more efficient for real-time application.

II. Target Speaker Extraction

In this section, we first review the baseline TCN-Conformer-based system with traditional MHSA [12] and memory-efficient MHSA [22] in Section II-A, and then discuss the proposed linear TCN-Conformer system in Section II-B.

A. Baseline TCN-Conformer System

Fig. 1 depicts the separator network of the considered baseline system, consisting of 4 identical stacks of TCN and conformer blocks. Each TCN block exploits the local context, while each conformer block exploits both local and global context features of the mixture signal. The first TCN block receives a concatenation of the encoded features obtained from the multi-scale speech encoder and the target speaker embedding as input, while the first conformer block receives the sum of the encoded features and the output from the first TCN block as input. For the remaining blocks, each TCN block receives a concatenation of the target speaker embedding and the output from the previous conformer block as input. Conversely, each conformer block receives the sum of the outputs from the previous TCN and conformer blocks as input.

The (traditional) SA mechanism in [19] transforms each input feature vector $\mathbf{y}_i \in \mathbb{R}^d$ with dimension d into three vectors: query ($\mathbf{q}_i \in \mathbb{R}^{d_k}$), key ($\mathbf{k}_i \in \mathbb{R}^{d_k}$), and value ($\mathbf{v}_i \in \mathbb{R}^{d_v}$). The query and key have the same feature dimension d_k, while d_v denotes the feature dimension of the value. The similarity between the i-th query and the j-th key is computed as $softmax\left(\mathbf{q}_i^T \mathbf{k}_j\right)$. SA focuses on finding similarities between all pairs of positions, where for n positions, the queries, keys and values are represented as matrices $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, and $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, respectively. The output of a traditional SA layer is given as:

$$Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sigma\left(\mathbf{Q}\mathbf{K}^T\right)\mathbf{V}, \tag{1}$$

where σ represents the softmax operation. Since traditional SA issues a separate query for each position, it exhibits overall memory and computational costs of $O(n^2)$. MHSA consists of the concatenation of several parallel SA layers.

In [22], a memory-efficient MHSA mechanism was proposed, which preserves closely the same mathematical equivalence of traditional MHSA while optimizing memory usage. Unlike traditional MHSA, which stores the full attention matrix in eq (1), memory-efficient MHSA dynamically recomputes the attention matrix during backpropagation instead of retaining it in memory, which significantly reduces memory consumption while maintaining the accuracy and functionality of traditional MHSA.

B. Proposed Linear TCN-Conformer System

The proposed linear TCN-Conformer system uses the same block diagram as in Fig. 1, where the traditional MHSA in each conformer block is replaced with a linear MHSA [21] to reduce both memory and computational costs. Similar to traditional SA, linear SA also transforms input features

into queries, keys, and values through linear transformations. However, instead of treating the keys as n feature vectors in $\mathbb{R}^{d_k}$, linear SA interprets them as d_k feature maps [21]. Each feature map acts as a weight across all positions and aggregates the corresponding values through a weighted sum. The output of this linear SA layer is given as:

$$Att_{lin}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sigma_q(\mathbf{Q}) \left(\sigma_k(\mathbf{K})^T \mathbf{V}\right), \quad (2)$$

where σ_q and σ_k represent row-wise and column-wise softmax operations, respectively. Although these softmax operations on the query and key matrices ($\mathbf{Q}$, $\mathbf{K}$) are not the same as performing a single softmax on $\mathbf{Q}\mathbf{K}^T$ as in (1), they may closely approximate the overall effect. The property of $\sigma(\mathbf{Q}\mathbf{K}^T)$ is that each row sums to 1, representing a normalized attention distribution over all positions. The matrix $\sigma_q(\mathbf{Q})\sigma_k(\mathbf{K})^T$ retains this property. Therefore, the linear SA mechanism in eq (2) offers a close approximation to traditional SA in eq (1) while significantly reducing memory and computational requirements, with both scaling linearly with input length.

III. Experiments

In this section, we discuss the datasets for training, validation, and testing, the used (hyper)parameters, and the training procedure of the considered baseline and proposed target speaker extraction systems.

A. Datasets

Similarly to [12], we have simulated three different types of mixtures at a sampling rate of 16 kHz: 2 speakers (2-mix), 3 speakers (3-mix), and 2 speakers with noise (noisy-mix) from the WSJ0 dataset [25] and the WHAM dataset [26]. For the training and validation sets, we have used the subset si_tr_s from the WSJ0 dataset, while we have used the subsets si_dt_05 and si_et_05 to create the test set. The test set contains entirely different speakers than the training and validation sets. To generate the 2-mix dataset, two different speakers are randomly selected and mixed at an SNR between 0 and 5 dB, where the first selected speaker is considered the target speaker and the second speaker is considered the interfering speaker. A different utterance of the target speaker is chosen as the reference speech to compute the speaker embedding. The 3-mix dataset is generated using a similar procedure, where two interfering speakers with equal power are mixed with the target speaker, also at an SNR between 0 and 5 dB. For the noisy-mix dataset, both the target and the interfering speaker are selected from the WSJ0 dataset, while the noise samples are selected from the training, validation, and test splits of the WHAM dataset. In total, the training and validation sets contain $47,926$ and $12,792$ utterances, respectively, while the test set contains $7,478$ utterances for all 2-mix, 3-mix, and noisy-mix.

B. Training settings

Similarly to the baseline system [12], the speaker embedder network consists of 3 residual blocks. Each residual block of the embedder network consists of two 1-D CNN layers, followed by a batch-normalization and a PReLU activation function. A residual connection is employed between the input and the second CNN with batch normalization. The dimensionality of the speaker embedding is fixed to 256. The separator network consists of a multi-scale encoder and decoder having 3 different filter lengths (2.5 ms, 10 ms, and 20 ms). Each TCN block of the separator network consists of two 1-D CNN layers, two PReLU activations with layer normalization, and one dilated depth-wise separable convolutional layer (DDS-CNN). Each conformer block [27] consists of four different blocks: two feed-forward, one MHSA, and a CNN arranged in the same structure as in [12].

Table I
(Hyper)Parameter settings for the different variants of the baseline and proposed systems.

Variants	Number of filters (speech encoder)	Attention-dimension	Number of filters (DDS-CNN)
Large	256	256	512
Medium	1024	64	920
Small	512	64	512
XSmall	256	64	256

We train the baseline system with traditional and memory-efficient MHSA, and the proposed linear TCN-Conformer system in causal mode, ensuring that all systems have a similar number of parameters as the baseline system with traditional MHSA in non-causal mode. For all systems, we consider different variants with a different number of parameters (see Table I). A large variant with about 12.8 M parameters, a medium variant (factor 2 reduction), a small variant (factor 4 reduction), and an extra small variant (factor 8 reduction). To achieve the different variants, we varied the number of filters in the speech encoder, the dimension of the MHSA in the conformer block, and the number of filters in DDS-CNN layers (see Table I), while keeping all other (hyper)parameters the same. For each variant, 4 stacks of TCN and conformer blocks are used in the separator network. All considered systems were trained using the ADAM optimizer [28] with a learning rate of 0.001 using a weighted combination of the multi-scale SI-SNR loss and the cross-entropy loss [11] for all mixture types together. All systems were trained considering 4-s segments of audio signals for 150 epochs with an early stopping criterion of 6 epochs.

IV. Results and Discussion

The performance of all considered systems is evaluated separately on the 2-mix, 3-mix, and noisy-mix test sets[1]. Besides using the scale-invariant signal-to-distortion ratio (SI-SDR) [29] as performance metric for speaker extraction, we have also considered the computational and memory costs measured by the total number of multiplications and additions (MACs) per seconds, the real-time factor (RTF), and the total number of parameters (#Param). MACs and #Param have been computed using the `Torchinfo` library of PyTorch. The RTF

[1]Audio examples: https://github.com/raginisinha/LinearTSEexamples

Table II

MEAN SI-SDR (DB) (↑), REAL-TIME FACTOR (RTF) (↓), TOTAL NUMBER OF MACS PER SECONDS (↓), AND NUMBER OF PARAMETERS FOR THE DIFFERENT VARIANTS OF THE BASELINE SYSTEMS (NON-CAUSAL AND CAUSAL MODES) WITH TRADITIONAL MHSA, THE BASELINE SYSTEM WITH MEMORY-EFFICIENT MHSA (CAUSAL MODE), AND THE PROPOSED EFFICIENT TCN-CONFORMER SYSTEMS (CAUSAL MODE). ALL SYSTEMS ARE TRAINED WITH ALL TYPES OF MIXTURES (2-MIX, 3-MIX, AND NOISY-MIX) TOGETHER. ↑ INDICATES HIGHER IS BETTER, WHILE ↓ INDICATES LOWER IS BETTER.

Variants	Systems	MHSA	Mode	2-mix	3-mix	noisy-mix	RTF	MACs	#Param
-	Input mixture	-	-	2.5	-1.3	-3.2	-	-	-
Large	Baseline system [12]	Traditional	Non-causal	17.5	10.7	9.3	-	17.92 G	12.8 M
	Baseline system [12]	Traditional	Causal	12.6	7.1	6.0	2.31	16.72 G	12.6 M
	Baseline system [12]	Memory-efficient	Causal	11.8	6.6	5.8	2.00	16.66 G	12.3 M
	Proposed system	Linear	Causal	**12.9**	**7.3**	**6.7**	**1.60**	14.05 G	12.3 M
Medium	Baseline system	Traditional	Non-causal	15.8	9.5	8.7	-	15.24 G	6.4 M
	Baseline system	Traditional	Causal	11.2	6.4	**6.1**	1.83	13.09 G	6.4 M
	Baseline system	Memory-efficient	Causal	11.0	6.1	5.7	1.62	13.00 G	6.3 M
	Proposed system	Linear	Causal	**11.7**	**6.6**	5.9	**0.23**	9.27 G	6.3 M
Small	Baseline system	Traditional	Non-causal	14.6	8.6	8.1	-	11.74 G	3.1 M
	Baseline system	Traditional	Causal	10.9	6.3	5.9	1.41	9.27 G	3.1 M
	Baseline system	Memory-efficient	Causal	10.7	6.0	5.7	1.07	9.19 G	3.0 M
	Proposed system	Linear	Causal	**11.4**	**6.4**	**6.1**	**0.12**	5.03 G	3.0 M
XSmall	Baseline system	Traditional	Non-causal	14.0	8.0	7.9	-	8.92 G	1.7 M
	Baseline system	Traditional	Causal	10.6	6.1	5.2	0.90	7.71 G	1.7 M
	Baseline system	Memory-efficient	Causal	10.2	6.0	5.2	0.72	7.66 G	1.6 M
	Proposed system	Linear	Causal	**11.3**	**6.3**	**5.9**	**0.08**	3.31 G	1.6 M

has been measured on an Intel Core i7-10850H CPU (2.7 GHz) as the time required to process an audio signal divided by its duration, where we have conducted 100 passes with 4-s segments of audio signals.

For the different variants, Table II shows the mean SI-SDR and the computational and memory costs of the baseline systems with traditional MHSA (non-causal and causal modes), the baseline system with memory-efficient MHSA (causal mode), and the proposed linear TCN-Conformer systems (causal mode). First, it can be observed that all target speaker extraction systems significantly improve the SI-SDR for all mixture types compared to the input mixtures. As expected for all variants, the performance of the baseline system with traditional MHSA degrades in causal mode compared to non-causal mode. Second, it can be observed that the baseline system with memory-efficient MHSA shows a slight reduction in RTF compared to the baseline system with traditional MHSA, however with a small reduction in the speaker extraction performance. Third, it can be observed that the proposed linear TCN-Conformer system outperforms the corresponding causal baseline system with traditional and memory-efficient MHSA at a significantly reduced RTF for all variants and mixture types (except for the Medium variant and noisy-mix). Furthermore, it can also be observed that the number of MACs per seconds for both baseline systems with traditional MHSA and memory-efficient MHSA (causal mode) remains approximately similar, while for the proposed system with linear MHSA the number of MACs reduces significantly compared to each baseline system for each variant. Although it may appear rather surprising that the proposed system with linear MHSA improves the speaker extraction performance compared to both baseline systems, a similar performance improvement has also been observed in [21]. One possible reason is that linear MHSA generates smoother, more globally coherent temporal attention weights that suppress artifacts more effectively than traditional or memory-efficient MHSA. Fourth, it can be observed that except for the XSmall causal baseline systems, none of the other baseline systems is suitable for real-time processing (i.e., RTF larger than 1). On the contrary, except for the Large variant of the proposed linear TCN-Conformer system with 12.3 M parameters, all other variants are suitable for real-time processing. For example, compared to the corresponding XSmall causal baseline system with traditional MHSA, the proposed XSmall linear TCN-Conformer system with 1.6 M parameters improves the performance by 0.7 dB for 2-mix, 0.2 dB for 3-mix, and 0.7 dB for noisy-mix with a reduction of approximately 91% in RTF. Additionally, it can be observed that the RTF reduction improves for smaller variants of the system.

V. CONCLUSION

In this paper, we demonstrated the advantages of using linear self-attention for real-time speaker-conditioned target speaker extraction using a TCN-Conformer architecture. To reduce both memory and computational costs, we replaced the traditional multi-head self-attention in each conformer block of the separator network with a linear multi-head self-attention, which requires linear memory and computational costs. Additionally, we reduced the overall number of parameters by factors of 2, 4, and 8 to enhance the capability for real-time processing. Experimental results on different mixtures show that the proposed system with linear self-attention outperforms the corresponding causal baseline system with traditional and memory-efficient self-attention, considerably improving computational cost and real-time factor.

REFERENCES

[1] Shoji Makino, Te-Won Lee, and Hiroshi Sawada, *Blind speech separation*, vol. 615, Springer, 2007.

[2] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio Source Separation and Speech Enhancement*, John Wiley & Sons, 2018.

[3] Yang Sun, Wenwu Wang, Jonathon Chambers, and Syed Mohsen Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 125–139, 2018.

[4] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, IEEE, pp. 46–50.

[5] Cem Subakan, Mirco Ravanelli, Samuele Cornell, François Grondin, and Mirko Bronzi, "Exploring self-attention mechanisms for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 2169–2180, 2023.

[6] Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černockỳ, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[7] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 2728–2732.

[8] Tingle Li, Qingjian Lin, Yuanyuan Bao, and Ming Li, "Atss-Net: Target speaker separation via attention-based neural network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1411–1415.

[9] Ragini Sinha, Christian Rollwage, and Simon Doclo, "Variants of lstm cells for single-channel speaker-conditioned target speaker extraction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–13, 2024.

[10] Zining Zhang, Bingsheng He, and Zhenjie Zhang, "X-TaSNet: Robust and accurate time-domain speaker extraction network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1421–1425.

[11] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li, "Spex+: A complete time domain speaker extraction network," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1406–1410.

[12] Ragini Sinha, Marvin Tammen, Christian Rollwage, and Simon Doclo, "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, Sept. 2022, pp. 1–5.

[13] Jiuxin Lin, Xinyu Cai, Heinrich Dinkel, Jun Chen, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Zhiyong Wu, Yujun Wang, and Helen Meng, "Av-Sepformer: Cross-attention sepformer for audio-visual target speaker extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, June 2023, pp. 1–5.

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[15] Marc Delcroix, Kateřina Žmolíková, Tsubasa Ochiai, Keisuke Kinoshita, and Tomohiro Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June 2021, pp. 6099–6103.

[16] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jimeng Zheng, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 4290–4294.

[17] Ali Aroudi and Sebastian Braun, "DBnet: Doa-driven beamforming network for end-to-end reverberant sound source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June 2021, pp. 211–215.

[18] Kateřina Žmolíková, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černockỳ, and Dong Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.

[21] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li, "Efficient attention: Attention with linear complexities," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, Jan. 2021, pp. 3531–3539.

[22] Markus N Rabe and Charles Staats, "Self-attention does not need $O(n^2)$ memory," *arXiv:2112.05682*, 2021.

[23] Koen Oostermeijer, Qing Wang, and Jun Du, "Lightweight causal transformer with local self-attention for real-time speech enhancement.," in *Proc. Interspeech*, Brno, Czech Republic, 2021, pp. 2831–2835.

[24] Yuma Koizumi, Shigeki Karita, Scott Wisdom, Hakan Erdogan, John R Hershey, Llion Jones, and Michiel Bacchiani, "DF-Conformer: Integrated architecture of Conv-TasNet and Conformer using linear complexity self-attention for speech enhancement," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, Oct. 2021, IEEE, pp. 161–165.

[25] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 31–35.

[26] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech 2019*, Graz, Austria, 2019, pp. 1368–1372.

[27] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Shanghai, China, Oct. 2020.

[28] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd International Conference for Learning Representations*, San Diego, USA, July 2015, pp. 1–15.

[29] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 626–630.