INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-Channel Speech Separation with Cross-Attention and Beamforming

Ladislav Mošner, Oldřich Plchot, Junyi Peng, Lukáš Burget, Jan "Honza" Černocký

Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

{imosner, iplchot, pengjy, burget, cernocky}@fit.vutbr.cz

Abstract

Originally, single-channel source separation gained more research interest. It resulted in immense progress. Multi-channel (MC) separation comes with new challenges posed by adverse indoor conditions making it an important field of study. We seek to combine promising ideas from the two worlds. First, we build MC models by extending current single-channel time-domain separators relying on their strength. Our approach allows reusing pre-trained models by inserting designed lightweight reference channel attention (RCA) combiner, the only trained module. It comprises two blocks: the former allows attending to different parts of other channels w.r.t. the reference one, and the latter provides an attention-based combination of channels. Second, like many successful MC models, our system incorporates beamforming and allows for the fusion of the network and beamformer outputs. We compare our approach with the SOTA models on the SMS-WSJ dataset and show better or similar performance.

Index Terms: multi-channel source separation, cross-channel attention, beamforming

1. Introduction

Over the past years, single-channel time-domain source separation progressed tremendously, achieving remarkable output quality in clean-audio conditions [1, 2, 3, 4].

Recently, distant speech processing devices, such as home assistants or meeting transcription systems, have gained increased popularity. Such devices are often equipped with multiple microphones that, in addition to an increased number of channels, provide spatial information due to the sensors' placement. Among speech-related tasks, multi-channel source separation is important as it is often used as a pre-processing step for multi-speaker ASR and also has other applications. As multiple channels and adverse conditions (including reverberation) pose new challenges, active research in the field is underway.

Approaches to multi-channel speech source separation can be broadly divided into (1) those extending existing single-channel models and (2) those specifically targeted for the task (potentially providing inductive bias).

Extensions of the pioneering time-domain source separation Conv-TasNet [1] model were presented in [5, 6]. A common feature of the architectures is that they aim to process the reference microphone while providing spatial clues as an additional input concatenated with the encoded version of the reference microphone. In [5], the authors provide the network with inter-channel phase difference (IPD) features. Instead of relying on hand-crafted features, MC-Conv-TasNet lets an auxiliary network learn suitable spatial features in a data-driven way [6].

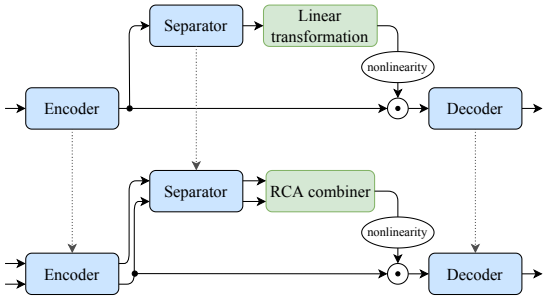


Figure 1: *The proposed framework replacing the linear transformation of separation models with designed reference channel attention (RCA) combiner. Dotted arrows represent a transfer of learned weights.*

FaSNet is a notable representative of time-domain multi-channel source separation models tailored toward the task [7]. It is based on a network predicting time-domain filters applied to channels, effectively implementing a version of adaptive filter-and-sum beamforming. Multiple new architectures extend the original FaSNet [8, 9, 10, 11]. Among them, [10] and [11] incorporate a self-attention mechanism, however, none of them uses an attention mechanism across channels.

Apart from time-domain models, time-frequency modeling is explored in [12, 13]. The networks map complex input spectra to complex spectra of separated outputs and provide noteworthy results.

As noted, utilization of attention across channels is not widespread in multi-channel source separation. Multi-head attention across channels, as part of the source mask estimating network, was presented in [14, 15]. The proposed networks were supposed to provide separated outputs, but the final focus was on speech recognition and continuous speech separation (usual separation metrics were not evaluated in the studies). In ASR, a self-attention channel combiner (jointly trained pre-processing) was presented in [16]. In [17, 18], a multi-channel encoder-decored ASR model was proposed and refined. The repeated encoder block contains a succession of channel-wise self attention and cross-channel attention (CCA) layers. Multi-channel processing has become popular also in the diarization field. Transformer encoders of EEND-EDA [19] were

The work was supported by Czech Ministry of Interior project No. VJ01010108 "ROZKAZ", Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, Czech Ministry of Education, Youth and Sports project no. LTAIN19087 "Multi-linguality in speech technologies", and Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports through the e-INFRA CZ (ID:90140).

10.21437/Interspeech.2023-2537

replaced by multi-channel versions in [20]. The first version followed [15], the second version was a co-attention encoder.

In this work, we extend the first group of multi-channel source separation models building upon architectures validated in single-channel scenarios. Recently designed models for time-domain speech source separation share a common high-level structure depicted in the upper part of Figure 1. As opposed to previous works [5, 6], we do not provide an extra input to the separator. Instead, we replace linear transformation within the network with the proposed *reference channel attention* (RCA) combiner (Figure 1). It was motivated by the potential ability to align information from various channels and, eventually, recover the information occluded in the reference channel from other channels. Our scheme allows the adoption of parameters of a pre-trained single-channel systems and updates only the RCA block on multi-channel data. It results in efficient training (especially when the number of microphones is large) as the output error is not back-propagated to the network preceding our module. Our contributions can be summarized as follows:

- The proposed approach is compatible with various time-domain models following the encoder-separator-decoder structure, including Conv-TasNet [1], DPRNN [2], DPTNet [3], Sepformer [4], and others.
- Our cross-frame reference channel attention (Section 2.3) represents an alternative to CCA [17], which is designed to provide outputs wrt. the reference channel.
- We show that the resulting networks can be conveniently used to estimate beamforming weights. Our final best-performing models benefit from the synergy of network-based separation and beamforming. They provide about 15% improvement in SI-SNRi over single-channel models.

2. Method

2.1. Single-channel time-domain speech separation

As shown in Figure 1, the time domain source mixture signal is transformed into an internal representation via a trainable encoder (usually implemented by a convolutional layer). The encoder output is passed to a separating network specific to each model. It is based on a temporal convolutional network (TCN) [1], LSTM layers [2], or transformer encoder layers [3, 4]. The network produces one mask per each source. We assume two speech sources. The masks are applied to the encoder output by multiplication. Resulting frames are subject to decoding through a trainable decoder, which performs transformation back to the time domain.

For the purpose of this paper, it is convenient to point out commonalities of separating networks of well-known models. The input to the network is normalized, for instance, by layer normalization [21] or global layer normalization [1]. The normalization is followed by a specific structure, often employing dual-path processing [2, 3, 4]. Subsequently, a nonlinearity is applied to the resulting representation. This is the output $\mathbf{R} \in \mathbb{R}^{T \times f}$ of the *separator* block in Figure 1, where T represents frames, and f is the dimensionality of features. Importantly, $\mathbf{R}$ is linearly projected to per-source features:

$$\mathbf{P}^{(s_1)} = \mathbf{R}\mathbf{W}^{(s_1)}, \quad \mathbf{P}^{(s_2)} = \mathbf{R}\mathbf{W}^{(s_2)}. \tag{1}$$

Projection matrices $\mathbf{W}^{(\cdot)} \in \mathbb{R}^{f \times d}$ are unique for sources s_1 and s_2. The dimensionality of the output features is d. Masks for individual sources are obtained by

$$\mathbf{M}^{(s_1)} = \sigma(\mathbf{P}^{(s_1)}; \theta), \quad \mathbf{M}^{(s_2)} = \sigma(\mathbf{P}^{(s_2)}; \theta). \tag{2}$$

The σ transformation can be as simple as nonlinearity (sigmoid, ReLU) or can use a gating mechanism parametrized by θ [3].

2.2. Extension to multi-channel separation through reference channel attention

In this section, we describe a straightforward extension of existing pre-trained source separation networks to multichannel ones that provide outputs aligned with the reference microphone. The alignment is important (1) to be able to compute time-domain loss correctly, (2) since it allows to obtain predictions for all channels by changing the reference one (it will be used in the beamforming weights estimation).

Time-domain separating networks extract representation $\mathbf{R}$. Rows of $\mathbf{R}$ (i.e., vectors $\mathbf{r}_t$) live in such a space that a linear transformation takes them to a source-specific feature space. Therefore, we hypothesize that $\mathbf{R}$ represents a suitable level where the information from microphones can be fused.

A simple approach would be to average channel-specific $\mathbf{R}_c$. However, due to sound propagation delay, misalignment could occur. It would result in unwanted smoothing. Instead, we propose a *reference channel attention* (RCA) module that allows to attend to different parts of different channels to improve the estimation of source signals at the reference microphone. Our framework reuses the pre-trained model to predict $\mathbf{R}_c$ for all channels $c = 1, \ldots, C$. Subsequently, the RCA module substitutes linear transformation in (1) as follows:

$$\mathbf{P}^{(s_n)} = \mathcal{F}_{\mathrm{RCA}}\left(\{\mathbf{R}_c\}_{c=1}^{C}; \phi^{(s_n)}\right), \tag{3}$$

with $\mathcal{F}_{\mathrm{RCA}}$ representing the RCA combiner parametrized by source-specific $\phi^{(s_n)}$. Subsequent computation steps follow the original architecture the module is inserted into.

2.3. Reference channel attention combiner

As shown in Figure 2a, the RCA combiner $\mathcal{F}_{\mathrm{RCA}}$ is composed of two consecutive blocks — *cross-frame reference channel attention* and *attentive channel combination*. The former allows attending to other parts of other channels aiming to pick up information that is corrupted or obscured in the reference channel. The latter combines channels with respect to the reference one.

To simplify the description, we will omit normalization, and we will focus only on one source, noting that an analogy holds for the second source.

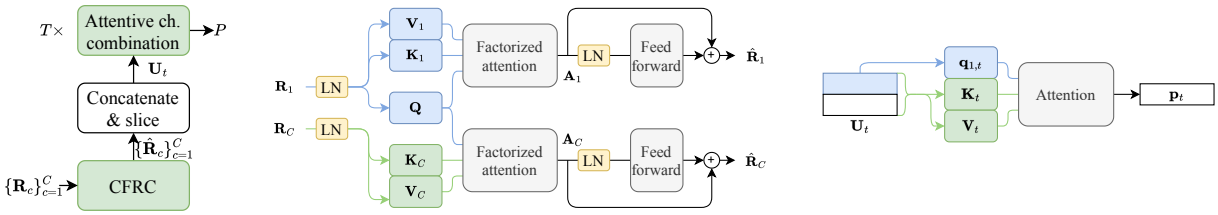
Cross-frame reference channel attention (CFRC): This attention module was partially inspired by CCA from [17]. However, it was modified to fit the purpose of extracting sources at the reference channel.

To allow the model to separate sources at the reference microphone (with index 1), we extract per-frame queries (from that particular channel): $\mathbf{Q} = \mathbf{R}_1\mathbf{W}^{(Q)}$. $\mathbf{W}^{(Q)} \in \mathbb{R}^{f \times d}$ is a query projection matrix. Channel-specific keys and values are obtained by projecting $\mathbf{R}_c$ with weight matrices that are shared across channels:

$$\begin{aligned} \mathbf{K}_c &= \mathbf{R}_c\mathbf{W}^{(K)}, \\ \mathbf{V}_c &= \mathbf{R}_c\mathbf{W}^{(V)}, \end{aligned} \tag{4}$$

where $\mathbf{W}^{(K)} \in \mathbb{R}^{f \times d}$ and $\mathbf{W}^{(V)} \in \mathbb{R}^{f \times d}$. The weight-sharing approach makes the module agnostic to the number of sensors.

In the following step, depending on the channel, self- or cross- attention is performed. We make use of relative positional encoding [22] (omitted from formulas for brevity) and a version of factorized attention [23], where we allow the query



(a) *RCA combiner* (b) *Cross-frame reference channel attention (CFRC)* (c) *Attentive channel combination*

Figure 2: *(a) The designed multi-channel RCA combiner comprises two main blocks: (b) CFRC attending across frames using queries extracted from the reference channel, (c) attentive channel combination, which aggregates per-frame information from all channels wrt. the reference one.*

frame to attend to keys from a limited time context. The reason is twofold: 1) The CFRC attention module should be able to align information from various microphones. Since the time delay between signals of two microphones is limited by sound wave propagation, only a limited time span is required. 2) By allowing constrained context for attention, the resulting system can process input sequences of various lengths without the problem of quadratic growth of memory with the sequence length (known for standard self-attention [24]).

Let $\mathbf{q}_t \in \mathbb{R}^d$ be a query vector corresponding to the frame index t (i.e., the t-th row of matrix $\mathbf{Q}$). Analogically, let vectors $\mathbf{k}_{c,t} \in \mathbb{R}^d$ and $\mathbf{v}_{c,t} \in \mathbb{R}^d$ be t-th rows of $\mathbf{K}_c$ and $\mathbf{V}_c$, respectively. Then the output $\mathbf{a}_{c,t}$ of the utilized sparse attention for frame index t and channel c is defined as

$$\mathbf{a}_{c,t} = \mathrm{softmax}\left(\frac{\mathbf{q}_t^\intercal \mathbf{K}_{c,\mathcal{S}_t}^\intercal}{\sqrt{d}}\right)\mathbf{V}_{c,\mathcal{S}_t}, \qquad \mathbf{K}_{c,\mathcal{S}_t} = (\mathbf{k}_{c,\tau})_{\tau\in\mathcal{S}_t}, \quad \mathbf{V}_{c,\mathcal{S}_t} = (\mathbf{v}_{c,\tau})_{\tau\in\mathcal{S}_t}. \tag{5}$$

$\mathcal{S}_t = \{t-x, \dots, t+x\}$ is a set of frame indices around t within a context of x. Concatenation of vectors $\mathbf{a}_{c,t}$ over the time dimension results in $\mathbf{A}_c \in \mathbb{R}^{T\times d}$. As shown in Figure 2b, the output $\hat{\mathbf{R}}_c$ of the CFRC attention module is obtained from $\mathbf{A}_c$ by feed-forward network and residual connection.

Attentive channel combination: The last block of the RCA module performs a frame-wise combination of channels in an attentive fashion (being invariant to the number of channels). To this end, the outputs of the CFRC attention module are altered. First, they are concatenated to yield tensor $[\hat{\mathbf{R}}_1, \hat{\mathbf{R}}_2, \dots, \hat{\mathbf{R}}_C] \in \mathbb{R}^{C\times T\times d}$. Then the same processing steps are repeated for every frame index (i.e., slice $\mathbf{U}_t \in \mathbb{R}^{C\times d}$ of the aforementioned tensor). We note that in practice, computation is performed in parallel thanks to independence.

As detailed in Figure 2c, the combination is performed wrt. the reference channel (marked in blue). A query vector is extracted only from the reference channel: $\mathbf{q}_{1,t} = \mathbf{W}^{(Q,\mathrm{comb})}\mathbf{u}_{1,t}$, where $\mathbf{u}_{1,t} \in \mathbb{R}^d$ is a row of $\mathbf{U}_t$ corresponding to the reference microphone. The output of the block is

$$\mathbf{p}_t = \mathrm{softmax}\left(\frac{\mathbf{q}_{1,t}^\intercal(\mathbf{U}_t\mathbf{W}^{(K,\mathrm{comb})})^\intercal}{\sqrt{d}}\right)(\mathbf{U}_t\mathbf{W}^{(V,\mathrm{comb})}), \tag{6}$$

where $\mathbf{W}^{(Q,\mathrm{comb})}$, $\mathbf{W}^{(K,\mathrm{comb})}$, and $\mathbf{W}^{(V,\mathrm{comb})}$ are projection matrices of shapes $d \times d$. Vectors $\mathbf{p}_t$ are finally concatenated, yielding a source-specific matrix $\mathbf{P}$ corresponding to that in (1).

We note that this module is a version of *multi-head attention across channels* [14] (also used in [20]). Contrary to [14], a query is extracted only from the reference channel (not from all).

3. Experimental setup

As already discussed, the proposed module is insertable into various time-domain source separation models. In this study, we focus on two of them — Conv-TasNet (a pioneering model in time-domain separation) and DPTNet (a representative of architectures based on transformer-encoder blocks). As is common in signal-based separation, all models were trained to optimize SI-SNR in an utterance-level permutation invariant training (PIT) fashion [25]. We compare models in terms of SI-SNR improvement (SI-SNRi [dB]), short-time objective intelligibility (STOI), and perceptual evaluation of speech quality (PESQ).

3.1. Extension of Conv-TasNet

Insertion of the RCA combiner into Conv-TasNet follows Figure 1 exactly. Linear transformation in (1) is commonly implemented by 1×1 convolution. Hence, our module seamlessly replaces the convolution layer in the original model.

We set hyperparameters according to the best non-causal model in [1] following Asteroid[1] recipe. ReLU is used as a nonlinearity providing the final values of source masks.

3.2. Extension of DPTNet

DPTNet utilizes a dual-path scheme dividing a sequence into overlapping segments and performing consecutive attention within and across segments. In [3], linear transformation (Figure 1) is applied to all features of all segments. Then, the segments are combined in an overlap-and-add fashion. To reduce computational and memory burden associated with a simple replacement of the linear transformations, we first merge segments via overlap-and-add. This yields $\mathbf{R}$ — suitable input to the RCA combiner. We argue that this is a minimal change. As opposed to Conv-TasNet, the σ function in (2) applied to resulting $\mathbf{P}$ is not a simple nonlinearity. A gating approach is adopted instead.

Contrary to the original work [3], we set hyperparameters according to Asteroid. The encoder/decoder window size is set to 16. The overlap of frames is 50%. The dimensionality of encoder features is 64. We use only 2 DPT layers (not 6), where each transformer encoder block has 4 heads. LSTM output dimensionality is 256, and segments in DPT have a length of 100.

3.3. Data

For both training and evaluation, we utilize the SMS-WSJ dataset [26] — a spatialized multi-speaker version of WSJ. It provides 33,561 (87.4h) training, 982 (2.5h) validation, and

[1] https://github.com/asteroid-team/asteroid

1,332 (3.4h) evaluation mixtures. The mixtures were created by simulating reverberant room conditions drawing T60 randomly from the interval [0.2, 0.5] s. The employed virtual circular microphone array, with a radius of 10 cm, comprises six sensors. Recordings also contain a low-energy additive white noise.

For the experiments with two channels, opposite microphones were selected. In training, random pairs were used. Only the pair of channels (1,4) was employed in the evaluation.

To focus solely on the separation ability of models, reverberant images of sources at the reference microphone serve as targets in training and testing.

4. Experiments

4.1. Direct model outputs

As a reference, we first trained single-channel Conv-TasNet and DPTNet models on the SMS-WSJ training data (picking random channels from the array). The results obtained with the first microphone are presented in Table 1, with *ch* being 1. The weights learned in this stage were reused in multi-channel experiments, where they were kept fixed without fine-tuning.

Next, we replaced linear projections of the interest with RCA combiners and trained them using signals of either two or six microphones. We note that it is possible to train RCA combiners on two microphones and use them when testing on six channels and vice versa (due to channel count invariance). However, in this study, we match the number of training and test channels. In Table 1 (net ✓, BF ✗), we observe slight improvements compared to single-channel models. It is noteworthy that the performance is bounded by pre-trained fixed weights. The improvement comes solely from a lightweight RCA combiner. It makes use of other channels to separate sources in the reference one.

4.2. Integration with beamforming

The designed approach can be seamlessly combined with beamforming. It represents an extension with no additional trainable parameters. The advantage of our approach is that after the first propagation of channels through the separator (i.e., after obtaining $\{\mathbf{R}_c\}_{c=1}^{C}$), no further forward propagation through it is required. By switching the reference microphone, the RCA combiner consecutively provides outputs aligned with all the channels. It is needed for subsequent beamforming. Since $\mathbf{R}_c$ are extracted by a shared network, the source permutation problem does not exist at the stage of RCA. It makes it easy to align sources in channels at the network output.

Given the time-domain separated sources, we follow [27] to compute source masks per channel, aggregate them over channels, and eventually use them to estimate per-source spatial covariance matrices (SCM). This way of SCMs prediction is suitable in our scenario because the outputs of separation networks can have a different dynamic range compared to the input (due to SI-SNR objective). Since the network outputs are used to estimate source prevalence in time-frequency bins (by computing ratio), a dynamic range is of no importance. Finally, MVDR [28, 29] beamformer weights are computed using SCMs.

As shown in Table 1 (net ✗, BF ✓), beamformed signals do not outperform separation network outputs when two microphones are used. On the other hand, beamforming with six channels provides outputs with better intelligibility and perceptual quality as measured by STOI and PESQ. This is likely due to the fact that beamforming does not introduce artifacts (which are audible in separation network outputs). However, residual interfering sources arguably still harm the SI-SNRi metric.

Table 1: *Multi-channel source separation results on SMS-WSJ. *The output of FaSNet-TAC is a beamformed (BF) signal.*

base model	# param.	ch	output net	output BF	SI-SNRi	STOI	PESQ
Oracle MVDR	–	2	✗	✓	6.84	0.86	2.34
Oracle MVDR	–	6	✗	✓	10.75	0.94	3.04
FaSNet-TAC [9]	2.7M	6	✓	✓*	11.52	0.91	2.85
MC-C-TasNet [6]	5.0M	6	✓	✗	11.82	0.92	2.94
Conv-TasNet	5.0M	1	✓	✗	10.34	0.89	2.75
	5.5M	2	✓	✗	10.43	0.90	2.77
	5.5M	2	✗	✓	6.62	0.86	2.31
	5.5M	2	✓	✓	9.98	0.90	2.62
	5.5M	6	✓	✗	10.43	0.90	2.77
	5.5M	6	✗	✓	9.65	0.93	2.91
	5.5M	6	✓	✓	**11.98**	**0.94**	**3.17**
DPTNet	2.8M	1	✓	✗	10.92	0.90	2.77
	3.0M	2	✓	✗	11.05	0.90	2.82
	3.0M	2	✗	✓	6.74	0.86	2.32
	3.0M	2	✓	✓	10.30	0.90	2.64
	3.0M	6	✓	✗	11.13	0.91	2.83
	3.0M	6	✗	✓	9.84	0.94	2.93
	3.0M	6	✓	✓	**12.40**	**0.94**	**3.22**

Finally, we propose an approach based on a fusion of network and beamformer outputs. It comes at a very low computational cost compared to beamforming since network outputs are already available during beamformer weights estimation. Let $\mathbf{y}^{(s_n,\text{net})}$ and $\mathbf{y}^{(s_n,\text{BF})}$ be time-domain outputs of the network and beamformer for the source s_n, respectively. Considering scale invariance, the fusion $\mathbf{y}^{(s_n,\text{fus})}$ is obtained as

$$\mathbf{y}^{(s_n,\text{fus})} = \frac{\langle \mathbf{y}^{(s_n,\text{net})}, \mathbf{y}^{(s_n,\text{BF})} \rangle}{2||\mathbf{y}^{(s_n,\text{net})}||_2^2} \mathbf{y}^{(s_n,\text{net})} + \frac{\mathbf{y}^{(s_n,\text{BF})}}{2}. \quad (7)$$

In the case of six channels, the fusion benefits from network outputs (with high SI-SNRi) and beamformer outputs (providing perceptually better signals) — Table 1 (net ✓, BF ✓).

We compare our six-channel versions of models with representatives of the two aforementioned time-domain model groups trained from scratch — MC-Conv-TasNet [6] and FaSNet-TAC+joint+4ms [9] (Asteroid version), as well as with the oracle MVDR using ideal binary masks [30]. We note that to facilitate a fair comparison with our approach, the employed MC-Conv-TasNet does not use WPE [31] to preprocess inputs. Its spatial encoder does not use pairs of channels, but for consistency with our approach, it employs all at once. We observe that six-channel versions of our models are competitive as both tend to outperform or perform at least on par with all the baselines.

5. Conclusion

In this paper, we proposed a reference channel attention (RCA) combiner — a module that extends various time-domain single-channel source separation models into multi-channel ones while reusing pre-trained parameters (adding only 0.6M and 0.2M parameters to Conv-TasNet and DPTNet, respectively). We showed improvements in separation metrics brought by RCA. Importantly, our framework allows for beamforming integration and fusion with network outputs providing considerable gains.

In our experiments, pre-trained weights remained fixed. In the future, we might release this constraint and fine-tune the whole network on multi-channel data. By allowing the RCA module to provide outputs for all channels, we might base the whole separation network on it.

6. References

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, 2019.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *ICASSP*, 2020, pp. 46–50.

[3] 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.

[4] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention Is All You Need In Speech Separation," in *ICASSP*, 2021, pp. 21–25.

[5] R. Gu, J. Wu, S. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-End Multi-Channel Speech Separation," *CoRR*, vol. abs/1905.06286, 2019.

[6] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On End-to-end Multi-channel Time Domain Speech Separation in Reverberant Environments," in *ICASSP*, 2020, pp. 6389–6393.

[7] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 260–267.

[8] Y. Luo and N. Mesgarani, "Implicit Filter-and-Sum Network for End-to-End Multi-Channel Speech Separation," in *Proc. Interspeech*, 2021, pp. 3071–3075.

[9] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end Microphone Permutation and Number Invariant Multi-channel Speech Separation," in *ICASSP*, 2020, pp. 6394–6398.

[10] Z. Wang, Y. Zhou, L. Gan, R. Chen, X. Tang, and H. Liu, "DE-DPCTnet: Deep Encoder Dual-path Convolutional Transformer Network for Multi-channel Speech Separation," in *2022 IEEE Workshop on Signal Processing Systems (SiPS)*, 2022, pp. 1–5.

[11] T. Zhao, C. Bao, X. Yang, and X. Zhang, "DPTNet-based Beamforming for Speech Separation," in *2022 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2022, pp. 1–5.

[12] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone Complex Spectral Mapping for Utterance-wise and Continuous Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[13] 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," *CoRR*, vol. abs/2209.03952, 2022.

[14] D. Wang, Z. Chen, and T. Yoshioka, "Neural Speech Separation Using Spatially Distributed Microphones," in *Proc. Interspeech*, 2020, pp. 339–343.

[15] D. Wang, T. Yoshioka, Z. Chen, X. Wang, T. Zhou, and Z. Meng, "Continuous Speech Separation with Ad Hoc Microphone Arrays," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1100–1104.

[16] R. Gong, C. Quillen, D. Sharma, A. Goderre, J. Laínez, and L. Milanović, "Self-Attention Channel Combinator Frontend for End-to-End Multichannel Far-Field Speech Recognition," in *Proc. Interspeech*, 2021, pp. 3840–3844.

[17] F.-J. Chang, M. Radfar, A. Mouchtaris, B. King, and S. Kunzmann, "End-to-End Multi-Channel Transformer for Speech Recognition," in *ICASSP*, 2021, pp. 5884–5888.

[18] F.-J. Chang, M. Radfar, A. Mouchtaris, and M. Omologo, "Multi-Channel Transformer Transducer for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 296–300.

[19] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," in *Proc. Interspeech*, 2020, pp. 269–273.

[20] S. Horiguchi, Y. Takashima, P. García, S. Watanabe, and Y. Kawaguchi, "Multi-Channel End-To-End Neural Diarization with Distributed Microphones," in *ICASSP*, 2022, pp. 7332–7336.

[21] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *CoRR*, vol. abs/1607.06450, 2016.

[22] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-Attention with Relative Position Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 464–468.

[23] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating Long Sequences with Sparse Transformers," *CoRR*, vol. abs/1904.10509, 2019.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[25] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, Oct. 2017.

[26] 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," *arXiv preprint arXiv:1910.13934*, 2019.

[27] L. Mošner, O. Plchot, L. Burget, and J. Černocký, "Multi-Channel Speaker Verification with Conv-Tasnet Based Beamformer," in *ICASSP*, 2022, pp. 7982–7986.

[28] J. Capon, "High-resolution Frequency-wavenumber Spectrum Analysis," *Proceedings of the IEEE*, vol. 57, no. 8, 1969.

[29] M. Souden, J. Benesty, and S. Affes, "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, 2010.

[30] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.

[31] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.