INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Implicit Filter-and-sum Network for End-to-end Multi-channel Speech Separation

Yi Luo, Nima Mesgarani

Department of Electrical Engineering, Columbia University, USA
yl3364@columbia.edu, nima@ee.columbia.edu

Abstract

Various neural network architectures have been proposed in recent years for the task of multi-channel speech separation. Among them, the filter-and-sum network (FaSNet) performs end-to-end time-domain filter-and-sum beamforming and has shown effective in both ad-hoc and fixed microphone array geometries. However, whether such explicit beamforming operation is a necessary and valid formulation remains unclear. In this paper, we investigate the beamforming operation and show that it is not necessary. To further improve the performance, we change the explicit waveform-level filter-and-sum operation into an implicit feature-level filter-and-sum operation around a context of features. A feature-level normalized cross correlation (fNCC) feature is also proposed to better match the implicit operation for an improved performance. Experiment results on a simulated ad-hoc microphone array dataset show that the proposed modification to the FaSNet, which we refer to as the implicit filter-and-sum network (iFaSNet), achieve better performance than the explicit FaSNet with a similar model size and a faster training and inference speed.

Index Terms: Speech separation, speech enhancement, multi-channel, end-to-end

1. Introduction

The design of multi-channel speech separation systems is one of the active topics in the speech separation community in the past years. Despite the advances in time-frequency domain neural beamformers where a neural network is used to assist the conventional beamformers for better robustness and performance [1–11], time-domain architectures have also earned the attention from the community due to their ability to perform purely end-to-end optimization towards the target signals. Moreover, as conventional beamformers often require a temporal context for better estimation of spatial features [12, 13], time-domain systems have the potential to be operated at frame-level with a lower theoretical system latency.

Recent time-domain systems can be classified into three categories. The first category reformulates the multi-channel separation problem as a single-channel separation problem on a selected reference microphone, with the help of additional cross-channel features. Various cross-channel features have been proposed and analyzed in versatile datasets [14–18]. The second category processes the multi-channel mixtures with a convolutional encoder where the channels are treated as different feature maps in a convolutional operation. Such systems learn a direct mapping between the mixtures and the target signals [19, 20]. The third category performs end-to-end beamforming without solving the optimization problems required in conventional beamformers. The beamforming filters can be compared to the masks in single-channel separation systems as they both operate at frame-level [21, 22].

One of the systems in the third category is the filter-and-sum network (FaSNet), which performs explicit time-domain filter-and-sum beamforming at frame level [21, 22]. FaSNet attempts to directly estimate the beamforming filters via a neural network, and previous results have proven its effectiveness in simulated noisy reverberation datasets comparing with single-channel methods and baseline methods in the first category. However, the problem formulation of FaSNet, i.e., the explicit filter-and-sum beamforming operation, has not been verified yet. Two core issues come from the problem formulation of FaSNet. Firstly, when the reverberant clean targets are used as the training objective, the beampatterns of the beamforming filters are implicitly constrained so that not only the signals coming from the direction of the direct path signals are reconstructed, but also all the reverberation signals have to be recovered. Such beampattern can be hard or impossible to be properly defined. Although it is possible to perform joint dereverberation and separation by setting the direct-path signals as the training target, it may significantly increase the difficulty of the task and lead to a performance degradation. It is thus important to validate the beamforming formulation of FaSNet. Secondly, it has been shown that oracle beamformers in frequency domain can achieve better performance than those in time domain [21, 23], hence it is necessary to revisit the formulation of the explicit waveform-level filter-and-sum operation and explore an implicit feature-level operation.

In this paper, we empirically explore the two issues by designing multiple ablation experiments on the standard FaSNet. We first show that such explicit filter-and-sum problem formulation is not necessary, and the model can achieve on-par performance by replacing the filter-and-sum operation applied on all channels to a filtering operation applied only on the reference channel. Since the standard FaSNet generates a set of beamforming filters for each channel and can thus be defined as a multi-input-multi-output (MIMO) system, the modification can be defined as a multi-input-single-output (MISO) system since it only generates the filter for the reference channel. Moreover, we investigate whether a feature-level filtering operation can lead to better separation performance than a waveform-level filtering operation. By designing a new cross-channel feature, namely the feature-level normalized cross correlation (fNCC), we show that such implicit filtering operation can result in an improved separation performance than the explicit waveform-level filter-and-sum operation with a faster training and inference speed. Finally, we explore the use of a context-aware module to extend the filtering operation to a context-level filter-and-sum operation. Combining all the modifications to the standard FaSNet gives us the *implicit filter-and-sum network (iFaSNet)*. Ablation experiments show that such modifications allow iFaSNet to outperform the standard FaSNet with a similar level of model size and complexity.

The rest of the paper is organized as follows. Section 2

http://dx.doi.org/10.21437/Interspeech.2021-1158

briefly goes over the original FaSNet architecture and introduces the proposed iFaSNet. Section 3 provides the experiment configurations. Section 4 presents the results and discussions. Section 5 concludes the paper.

2. Implicit Filter-and-sum Network

2.1. Filter-and-sum Network Recap

Filter-and-sum network (FaSNet) performs time-domain filter-and-sum beamforming at frame level and directly estimates the beamforming filters with a neural network. For each frame of input mixtures from the M channels $\{\mathbf{y}^i\}_{i=1}^M \in \mathbb{R}^{1\times L}$, a context window of W samples in both past and future is concatenated into $\mathbf{y}^i$, resulting in a context frame $\hat{\mathbf{y}}^i \in \mathbb{R}^{1\times(L+2W)}$. For each target source, M time-domain beamforming filters $\{\mathbf{h}^i\}_{i=1}^M \in \mathbb{R}^{1\times(1+2W)}$ are estimated from $\{\mathbf{y}^i\}_{i=1}^M$ by a neural network, and the filters are applied to the input to obtain the separated outputs $\hat{\mathbf{x}} \in \mathbb{R}^{1\times L}$:

$$\hat{\mathbf{x}} = \sum_{i=1}^{M} \hat{\mathbf{y}}^i \circledast \mathbf{h}^i \tag{1}$$

where $\circledast$ represents the convolution operation. The estimation of the filters rely on both the channel-wise features and the cross-channel features, and FaSNet applies a linear fully-connected (FC) layer to extract the channel-wise features for each input channel:

$$\mathbf{s}^i = \hat{\mathbf{y}}^i \hat{\mathbf{W}} \tag{2}$$

where $\hat{\mathbf{W}} \in \mathbb{R}^{(L+2W)\times N}$ is a learnable parameter matrix. The linear FC layer can be also regarded as a linear 1-D convolutional layer. The cross-channel feature used by FaSNet is the normalized cross correlation (NCC) feature $\mathbf{q}^i \in \mathbb{R}^{1\times(1+2W)}$ calculated between the center frame at reference microphone $\mathbf{y}^1$ and the context frame at all microphones $\{\hat{\mathbf{y}}^i\}_{i=1}^M$:

$$\begin{cases} \hat{\mathbf{y}}_j^i = \hat{\mathbf{y}}^i[j : j+L-1] \\ q_j^i = \frac{\mathbf{y}_1 \hat{\mathbf{y}}_j^{iT}}{\|\mathbf{y}_1\|_2 \|\hat{\mathbf{y}}_j^i\|_2} \end{cases}, \quad j = 1, \ldots, 2W+1 \tag{3}$$

The channel-wise features $\mathbf{s}^i$ are concatenated with the cross-channel features $\mathbf{q}^i$ to serve as the input to the filter estimation module. The filter estimation modules contains stacked dual-path RNN (DPRNN) blocks [24] with the transform-average-concatenate (TAC) module for microphone number and permutation (location) invariance. The TAC takes the features from all the channels as input, and its output is send to a DPRNN block shared by all channels. We recommend the readers to refer to the original literature for the detailed design [22]. Such architectures have shown effective in both ad-hoc and fixed microphone array geometries.

2.2. Implicit Filter-and-sum Network

2.2.1. MIMO versus MISO

The training target for the standard FaSNet is typically the reverberant clean signals. In the problem configuration where the time-domain filter-and-sum operation is applied, it implies that the beamforming filters should not only enhance the signal coming from a certain direction, but also reconstruct all the reverberation components. However, as reverberation may come from all possible directions, the ideal beampattern of such beamforming filters might be hard or even impossible to define. Although FaSNet applies frame-level beamforming where infinite optimal frame-level filters may exist since the linear equation in equation 1 is underdetermined (L equations and $M \times (1+2W)$ unknowns), finding such reverberation-preserving filters for all channels may still be unnecessary.

It is natural to consider an alternative problem formulation rather than the standard filter-and-sum formulation. Since the standard FaSNet estimates a set of filters for each of the channels and can be viewed as a multi-input-multi-output (MIMO) system, a simple way to bypass the issue of bad beampatterns is to change it into a multi-input-single-output (MISO) system where only the filter for the reference channel is estimated. The features from all the other channels are thus viewed as additional information to assist the separation on a (randomly) selected reference channel. This reformulates the multi-channel separation problem back to the single-channel separation problem as in the first category discussed in Section 1 and maintains the underdetermined nature in equation 1 (L equations and $1+2W$ unknowns, where $1+2W$ is typically larger than L). Figure 1 (A) shows the MISO module.

2.2.2. Feature-level Implicit Filtering

As discussed in Section 1, most existing neural beamformers are mainly designed in the frequency domain due to the fact that oracle frequency-domain beamformers typically have better performance than those in time domain. As frequency-domain beamformers are typically formulated as a multiplication operation on the spectrums, a similar operation can be defined in time-domain systems as a multiplication operation on learnable features. Note that recent single-channel speech separation systems have widely applied a set of learnable encoder and decoder similar to the one in equation 2 to replace the short-time Fourier transform (STFT) and estimated a multiplicative mask on the encoder outputs to match the formulation in time-frequency masking systems [25]. The formulation of implicit filtering can thus be connected to the masking operation in such systems.

The extraction of the channel-wise features in equation 2 is calculated on the context of input mixture frame $\hat{\mathbf{y}}^i$. In standard single-channel systems, the multiplicative mask is estimated for a learnable feature calculated from the center frame $\hat{y}^i$ only without the context. To match the formulation of such configuration, we modify the encoder weight to $\mathbf{W} \in \mathbb{R}^{L\times N}$. If we treat each column as a basis filter, then the length of the basis filters in the modified encoder is L instead of the original $L+2W$ in $\hat{\mathbf{W}}$. The encoder is applied to the entire context $\hat{\mathbf{y}}^i$ with a corresponding hop size (which is empirically set to $L/2$), which results in a sequence of encoder outputs $[\mathbf{f}_{t-C}^i, \ldots, \mathbf{f}_t^i, \ldots, \mathbf{f}_{t+C}^i] \in \mathbb{R}^{(1+2C)\times N}$ where t denotes the frame index and C denotes the context size. The estimated filter with shape $\mathbb{R}^{1\times N}$ is only applied to $\mathbf{f}_t^i$, while all encoder outputs of the context are used as the input the filter estimation modules. A decoder with its weight $\mathbf{U} \in \mathbb{R}^{N\times L}$ is applied to transform the filtered feature back to waveforms. Figure 1 (B) shows the linear decoder module.

2.2.3. Feature-level Normalized Cross Correlation

The cross-channel feature in FaSNet is calculated by time-domain normalized cross correlation (tNCC) defined in equation 3. The rationale behind tNCC is to capture both the delay information across channels and the source-dependent in-

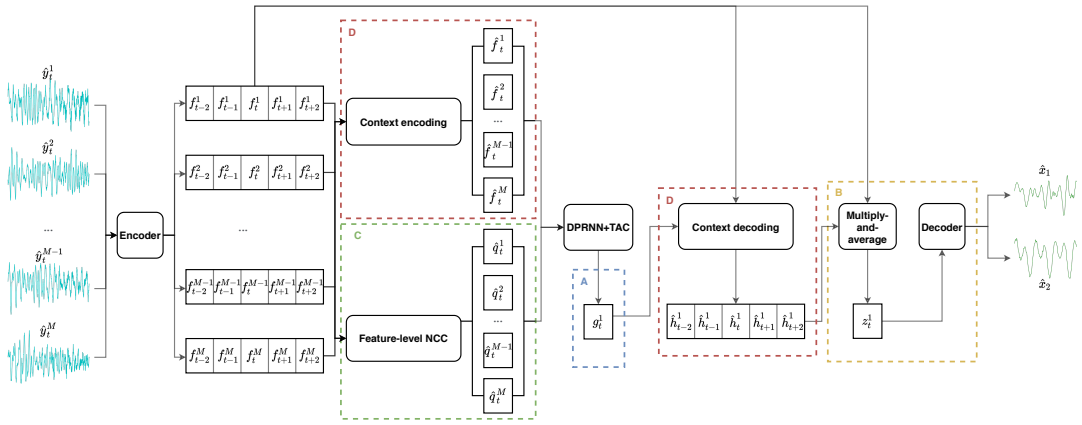

Figure 1: *Flowchart for the proposed iFaSNet architecture. The modifications to the original FaSNet are highlighted, which include (A) the use of MISO design instead of the original MIMO design, (B) the use of implicit filtering in the latent space instead of the original explicit filtering on the waveforms, (C) the use of feature-level NCC feature for cross-channel information instead of the original sample-level NCC feature, and (D) the use of context-aware filtering instead of the original context-independent filtering.*

formation for different targets. However, whether tNCC is still a good feature in the implicit filtering formulation is unknown. Since implicit filtering operates in the feature space and does not explicitly requires the information of sample-level delay, it is necessary to modify tNCC such that it better explores the cross-channel information in the feature level. Here we modify the tNCC to a feature-level NCC (fNCC) feature. Denote the context feature $[\mathbf{f}^i_{t-C}, \dots, \mathbf{f}^i_t, \dots, \mathbf{f}^i_{t+C}]$ as $\mathbf{F}^i_t$, fNCC calculates the cosine similarity between the contextual feature in the reference channel $\mathbf{F}^1_t$ and the contextual feature in all channels $\{\mathbf{F}^i_t\}^M_{i=1}$:

$$\hat{\mathbf{q}}^i_t = \bar{\mathbf{F}}^1_t \bar{\mathbf{F}}^{iT}_t \tag{4}$$

where $\bar{\mathbf{F}}^i_t$ denotes the column-normalized feature of $\mathbf{F}^i_t$ where each column has a unit length, and $\hat{\mathbf{q}}^i_t \in \mathbb{R}^{(1+2C)\times(1+2C)}$ denotes the fNCC feature at time t for channel i. $\hat{\mathbf{q}}^i_t$ is then flatten to a vector of shape $1\times(1+2C)^2$. For the default setting in FaSNet where $W = L = 16$ ms $= 256$ samples with a 50% hop size, we have $C = 2$ and $(1+2C)^2 = 25 \ll 1+2W = 513$. Figure 1 (C) shows the fNCC calculation module.

2.2.4. *Context-aware Filter-and-sum*

Utilizing context information to improve the modeling of local frame is very common in various systems [26, 27]. To make use of the contextual feature $\mathbf{F}^i_t$, a straightforward way is to concatenate all of them and pass to the filter estimation module. However, here we propose a context encoder and decoder to perform both dimension reduction on the features. A context encoder is applied to $[\mathbf{f}^i_{t-C}, \dots, \mathbf{f}^i_t, \dots, \mathbf{f}^i_{t+C}]$ to model the intra-context dependencies, and the output is averaged across time to squeeze into a single feature vector $\hat{\mathbf{f}}^i_t \in \mathbb{R}^{1\times N}$. $\hat{\mathbf{f}}^i_t$ together with the fNCC feature $\hat{\mathbf{q}}^i_t$ are concatenated and used as the input to the filter estimation modules. The output of the MISO filter estimation modules $\mathbf{g}^1_t \in \mathbb{R}^{1\times N}$ is then concatenated to each feature in the contextual encoder outputs and passed to a context decoder to generate a set of contextual filters $[\hat{\mathbf{h}}^1_{t-C}, \dots, \hat{\mathbf{h}}^1_t, \dots, \hat{\mathbf{h}}^1_{t+C}] \in \mathbb{R}^{(1+2C)\times N}$. The filters are then applied to the contextual encoder outputs to form an implicit, intra-context "filter-and-sum" operation:

$$\mathbf{z}^1_t = \frac{1}{1+2C} \sum_{j=0}^{2C} \mathbf{f}^1_{t-C+j} \odot \hat{\mathbf{h}}^1_{t-C+j} \tag{5}$$

where $\odot$ denotes the Hadamard product. In this paper we use a bidirectional LSTM (BLSTM) layer for both the contextual encoder and decoder. Figure 1 (D) shows the context encoder and decoder.

3. Experiment Configurations

3.1. Dataset

We evaluate our approach on a simulated ad-hoc multi-channel two-speaker noisy speech dataset. 20000, 5000 and 3000 4-second long utterances are simulated for training, validation and test sets, respectively. All utterances are generated at 16K Hz sample rate. For each utterance, two speech signals and one noise signal are randomly selected from the 100-hour Librispeech subset [28] and the 100 Nonspeech Corpus [29], respectively. The overlap ratio between the two speakers is uniformly sampled between 0% and 100%, and the two speech signals are shifted accordingly and rescaled to a random relative SNR between 0 and 5 dB. The relative SNR between the power of the sum of the two clean speech signals and the noise is randomly sampled between 10 and 20 dB. The transformed signals are then convolved with the room impulse responses simulated by the image method [30] using the gpuRIR toolbox [31]. The length and width of all the rooms are randomly sampled between 3 and 10 meters, and the height is randomly sampled between 2.5 and 4 meters. The reverberation time (T60) is randomly sampled between 0.1 and 0.5 seconds. After convolution, the echoic signals are summed to create the mixture for each microphone. The ad-hoc array dataset contains utterances with 2 to 6 microphones, where the number of utterances for each microphone configuration is set equal.

3.2. Model Configurations

The standard FaSNet with DPRNN blocks and TAC modules is used as the backbone architecture as well as the baseline. The

Table 1: *Experiment results with various model configurations. SI-SDRi is reported on decibel scale.*

Model	# of param.	# of mics	Overlap ratio				Average
			<25%	25-50%	50-75%	>75%	
FaSNet	2.9M	2 / 4 / 6	15.0 / 15.3 / 14.8	10.7 / 11.1 / 11.6	8.6 / 9.2 / 9.3	5.4 / 7.0 / 7.0	9.7 / 10.8 / 10.9
+MISO	2.9M		14.8 / 15.5 / 15.7	10.4 / 11.3 / 11.9	8.5 / 9.0 / 9.4	5.0 / 6.8 / 7.1	9.5 / 10.8 / 11.2
+fNCC	3.0M		14.5 / 14.8 / 14.4	10.1 / 11.0 / 11.3	8.3 / 8.9 / 9.0	4.9 / 6.7 / 6.9	9.3 / 10.4 / 10.6
+MISO+fNCC	2.9M		15.0 / 15.7 / 15.7	10.6 / 11.4 / 12.2	8.4 / 9.4 / 9.6	5.3 / 7.4 / 8.0	9.7 / 11.1 / 11.6
+MISO+implicit	2.9M		14.2 / 14.9 / 15.2	9.8 / 10.9 / 11.3	7.7 / 8.1 / 8.7	4.6 / 5.7 / 6.1	8.9 / 10.0 / 10.6
+MISO+implicit+fNCC	3.0M		15.3 / 16.0 / 16.1	10.9 / 11.8 / 12.5	8.5 / 9.6 / 10.1	5.7 / 7.6 / 8.3	9.9 / 11.4 / 12.0
+MISO+implicit+fNCC+context	3.3M		**15.6 / 16.4 / 16.5**	**11.2 / 12.4 / 12.9**	**9.0 / 10.1 / 10.3**	**5.8 / 7.9 / 8.8**	**10.2 / 11.8 / 12.3**

Table 2: *Training and inference speeds of different model configurations. The speeds are measured on a single NVIDIA TITAN Pascal graphic card with a batch size of 4.*

Model	Training speed	Inference speed
FaSNet	268.0 ms	98.1 ms
+MISO	222.6 ms	93.2 ms
+fNCC	331.7 ms	105.7 ms
+MISO+fNCC	333.0 ms	105.8 ms
+MISO+implicit	187.9 ms	96.6 ms
+MISO+implicit+fNCC	**132.2 ms**	**44.2 ms**
+MISO+implicit+fNCC+context	156.7 ms	52.5 ms

frame size L and the context size W are both set to 16 ms (256 points), and the hop size is set to 50%. In iFaSNet, we use the same configuration of L and W. The implementations of both the standard FaSNet and the iFaSNet are available online[1].

3.3. Training Configurations

All models are trained for 100 epochs with the Adam optimizer [32] with an initial learning rate of 0.001. Negative signal-to-noise ratio (SNR) is used as the training objective for all models. The learning rate is decayed by 0.98 for every two epochs. Gradient clipping by a maximum gradient norm of 5 is always applied. Early stopping is applied when no best validation model is found for 10 consecutive epochs. No other training tricks or regularization techniques are used. Auxiliary autoencoding training (A2T) is applied to enhance the robustness on this reverberant separation task [33].

4. Results and Discussions

Table 1 presents the ablation experiment results of the standard FaSNet with different modifications applied. We first explore the effect of changing the original MIMO configuration into a MISO configuration. It can be observed that removing the beamforming filters for all other channels except for the reference channel does not harm the overall performance, and the performance in the 6 microphone setting is even improved. This results supports our discussion in Section 2.2.1 that jointly using explicit filter-and-sum formulation and setting reverberant clean signals as training targets might not be a proper configuration.

We then evaluate the role of the cross-channel features in different model settings. We observe that using fNCC together with the original MIMO configuration leads to worse performance than using the original tNCC feature, while using fNCC together with the MISO configuration can improve the separation performance in high overlapped utterances. This shows that proper cross-channel features should be selected to keep inline with the system's problem formulation in order to achieve a good performance.

Next we test whether implicit filtering can further improve the performance. Applying MISO configuration together with tNCC feature and implicit filtering achieves even worse performance than the baseline FaSNet, while changing the tNCC feature to fNCC feature results in a performance boost. Since fNCC is calculated on the contextual encoder outputs and the implicit filtering configuration estimates multiplicative filters on the center frame of the contextual encoder outputs, this further verifies that matching the problem formulation with a proper cross-channel feature is crucial for a good overall performance.

The last ablation experiment verifies the effectiveness of the context encoder and decoder modules. With a slightly increased model size, the intra-context filter-and-sum formulation can further improve the performance upon the implicit filtering formulation, which shows that exploring contextual information is beneficial. This can also be related to recent literature in multi-tap beamformers where the beamforming filters are estimated for a context of frames [10].

Table 2 compares the training and inference speed for different model configurations. We can see that when the MISO, implicit filtering, and fNCC feature are applied, the training and inference speeds are 2 times and 2.2 times faster, respectively. Adding the context encoder and decoder modules makes the network slightly slower, but the training and inference speeds are still 1.7 times and 1.9 times faster than the standard FaSNet, respectively. The results show that iFaSNet does not sacrifice complexity for performance.

5. Conclusion

In this paper, we revisited the necessity for explicit filter-and-sum beamforming in an end-to-end multichannel speech separation network, the FaSNet, and proposed multiple modifications to improve the overall performance. We considered the MISO problem formulation, the implicit feature-level processing, the feature-level NCC feature, and the context-aware filtering as directions to improve the model. We named our modification to FaSNet as the implicit FaSNet (iFaSNet), and designed multiple ablation experiments to show that the proposed iFaSNet can lead to a better overall performance with a faster training and inference speed.

6. Acknowledgments

This work was funded by a grant from the National Institute of Health, NIDCD, DC014279; and a grant from Marie-Josée and Henry R. Kravis.

[1] https://github.com/yluo42/TAC

7. References

[1] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 196–200.

[2] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Proc. Interspeech*, 2016, pp. 1981–1985.

[3] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3246–3250.

[4] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[5] X. Zhang, Z.-Q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 276–280.

[6] J. Heymann, M. Bacchiani, and T. N. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 6722–6726.

[7] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Deep learning based speech beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 5389–5393.

[8] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust asr," in *Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*. IEEE, 2019, pp. 6745–6749.

[9] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *Acoustics, Speech and Signal Processing (ICASSP), 2020 IEEE International Conference on*. IEEE, 2020, pp. 6384–6388.

[10] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," *Proc. Interspeech 2020*, pp. 56–60, 2020.

[11] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning mvdr beamformer for target speech separation," *arXiv preprint arXiv:2008.06994*, 2020.

[12] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 4, pp. 780–793, 2017.

[13] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018, pp. 531–535.

[14] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[15] R. Gu and Y. Zou, "Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation," *arXiv preprint arXiv:2001.00391*, 2020.

[16] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2020 IEEE International Conference on*. IEEE, 2020, pp. 7319–7323.

[17] Y. Koyama, O. Azeez, and B. Raj, "Efficient integration of multi-channel information for speaker-independent speech separation," *arXiv preprint arXiv:2005.11612*, 2020.

[18] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, "Spatial and spectral deep attention fusion for multi-channel speech separation using deep embedding features," *arXiv preprint arXiv:2002.01626*, 2020.

[19] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[20] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 1888–1900, 2020.

[21] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Automatic Speech Recognition and Understanding (ASRU), 2019 IEEE Workshop on*. IEEE, 2019, pp. 260–267.

[22] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2020 IEEE International Conference on*. IEEE, 2020, pp. 6394–6398.

[23] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, W. D. O'Brien Jr, B. C. Wheeler, and A. S. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 379–391, 2004.

[24] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2020 IEEE International Conference on*. IEEE, 2020, pp. 46–50.

[25] 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 (TASLP)*, vol. 27, no. 8, pp. 1256–1266, 2019.

[26] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2014.

[27] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[29] G. Hu, "100 Nonspeech Sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[30] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[31] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, pp. 1–19, 2020.

[32] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] Y. Luo, C. Han, and N. Mesgarani, "Distortion-controlled training for end-to-end reverberant speech separation with auxiliary autoencoding loss," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.