Low-complexity Multi-Channel Speaker Extraction with Pure Speech Cues

Bang Zeng*†, Hongbin Suo‡, Yulong Wan‡, Ming Li*†
* Data Science Research Center, Duke Kunshan University, Kunshan, China
† School of Computer Science, Wuhan University, Wuhan, China
E-mail: bangzeng@whu.edu.cn, ming.li369@dukekunshan.edu.cn
‡ Data&AI Engineering System, OPPO, Beijing, China
E-mail: suohongbin@oppo.com, wanyulong@oppo.com

***Abstract*—Most multi-channel speaker extraction schemes use the target speaker's location information as a reference, which must be known in advance or derived from visual cues. In addition, memory and computation costs are enormous when the model deals with the fusion input. In this paper, we propose Speaker-extraction-and-filter Network (SeafNet), which is a low-complexity multi-channel speaker extraction network with only speech cues. Specifically, the SeafNet separates the mixture by utilizing the correlation between an estimation of target speaker on reference channel and the mixed input on rest channels. Experimental results show that compared with the baseline, the SeafNet model achieves 6.4% relative SISNRi improvement on the fixed geometry array and 8.9% average relative SISNRi improvement on the ad-hoc array. Meanwhile, the SeafNet achieves 60% relative reduction in the number of parameters and 42% relative reduction in the computational cost.**

I. Introduction

The cocktail party problem [1] shows the fact that the performance of the speech applications, such as speech recognition, is significantly affected by noise and reverberation in the real world. Speech separation is widely used to deal with this problem by improving the noise robustness of the back-end speech systems. Although most methods work well in the single-channel scenario [2]–[5], the performance declines significantly when the separation system faces a reverberation background. The multi-channel speech separation model with a beamformer is designed to address this issue. Conventional beamforming algorithms can be broadly classified as fixed and adaptive. The fixed beamformer [6] first calculates the time difference of arrival (TDOA) between the reference microphone and the remaining microphones to derive a time-shifted signal for each microphone and then sums up the time-shifted signals to obtain the final beamformer output. The adaptive beamforming algorithms determine the optimal weights of the beamformer through different criteria [7]–[9], the main one of which is the minimum variance distortionless response [8]. In recent years, multi-channel speech separation adopting a learning-based beamformer has become dominant. There are two main routes for these multi-channel speech separation schemes. One class [10]–[12] uses deep neural networks (DNNs) to pre-separate the signals of each channel first, and then estimates the beamformer coefficients based on the pre-separated outputs, or filters the pre-separated outputs directly with a conventional beamformer. The other class [13]–[16], which implicitly employs beamforming in DNN, directly models the mapping relationship between multi-channel inputs and target outputs. Among the aforementioned methods, Filter-and-Sum Network with Transform-Average-Concatenate (FaSNet-TAC) [16] belongs to the latter, and achieves good performance. FaSNet-TAC is a microphone permutation and number invariant model due to the use of pair-wise features. Moreover, in order to overcome the shortcomings of using pair-wise features, FaSNet-TAC adopts TAC operation to fully utilizes the information from all microphones. Experimental results in [16] shows that FaSNet-TAC works well in both fixed geometry array and ad-hoc array. However, the unknown number of speakers and the global permutation problem are still the main obstacles for multi-channel blind speech separation algorithms.

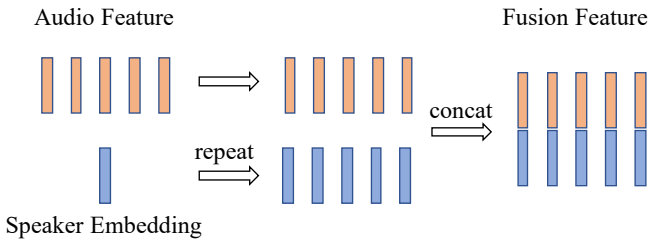

Fig. 1. The tranditional way to use the speaker embedding in target speaker extraction.

Speaker extraction [17], [18], also known as target speech separation, can extract a specific speaker's speech from the mixture with reference information. The speaker's location is commonly used as auxiliary reference information in multi-channel scenarios [19], [20]. However, the speaker's location must be known in advance or extracted from the visual features [21]. Using the speaker embedding from the target speaker's enrollment speech can deal with this issue, yet this will lead to another problem of bigger model size and computational costs due to the relatively high dimensionality of fusion features as shown in Figure 1. This problem is even more serious on multi-channel tasks. In this paper, we extend the FaSNet-TAC [16] model to the multi-channel speaker extraction task. Moreover, to resolve the aforementioned limitations, we propose a relatively low-complexity speaker extraction

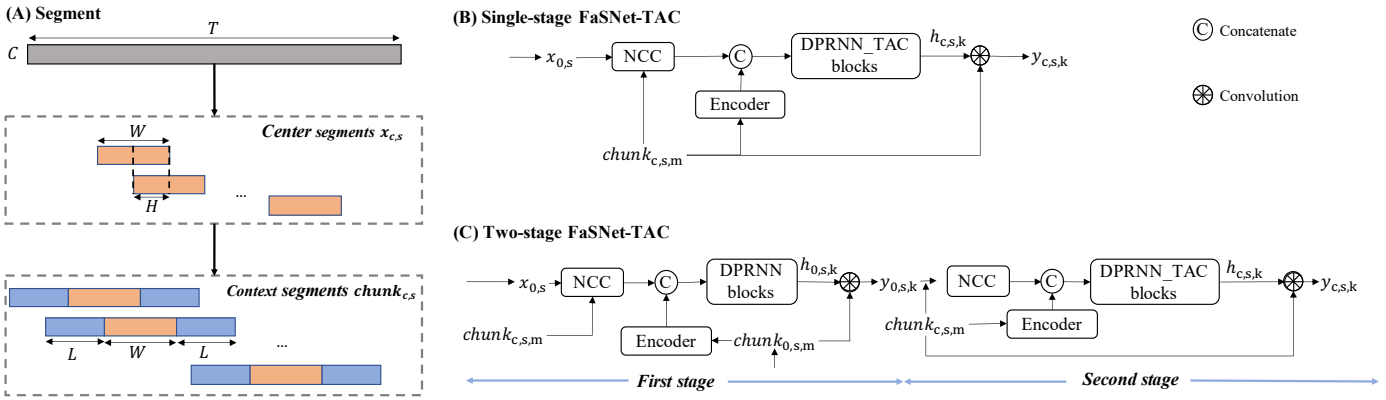

Fig. 2. System flowchart of the origin FaSNet-TAC models. (A) The FaSNet-TAC first splits the input into center segments and context segments. (B) The single-stage FaSNet-TAC model. (C) The two-stage FaSNet-TAC model

network only using speech cues, called Speaker-extraction-and-filter Network (SeafNet). Specifically, the SeafNet first extracts the target speaker's wave of the reference microphone signal by utilizing a reference speech. Then, the SeafNet model computes the Normalized Cross-Correlation (NCC) of this estimation and mixed waveform. The NCC feature contains the content difference between the reference microphone's target speaker estimation and the multi-channel mixed input. Finally, the SeafNet derives the separation result in the same way as the FaSNet-TAC model. In this way, the speaker features only need to be concatenated with the time-domain features of the reference microphone, rather than with the ones of all microphones' signal, significantly reducing the number of parameters and computational costs of the model.

II. Methods

A. Original FaSNet with TAC

FaSNet-TAC is an effective multi-channel blind speech separation solution. Figure 2 (B) and (C) shows the two-stage [15] and single-stage [16] FaSNet-TAC models. As shown in Figure 2 (A), FaSNet-TAC first split the input into center segments with the length of W:

$$\mathbf{x}_{c,s} = \mathbf{x}_c[sH : sH + W - 1] \tag{1}$$

where $\mathbf{x}_{c,s}$ is the center segments, $c \in 0, 1, ..., C-1$ is the c-th microphone of the input, $s \in \mathbf{Z}$ is the index of the segments, and $H \in [0, W-1]$ is the hop size. Then, $\mathbf{x}_{c,s}$ is concatenated with a context window with the length of L on both sides and results $\mathbf{chunk}_{c,s}$:

$$\mathbf{chunk}_{c,s} = \mathbf{x}_c[sH - L : sH + W + L - 1] \tag{2}$$

The two-stage FaSNet-TAC first calculates the Normalized Cross-Correlation (NCC) feature, defined as the cosine similarity between the reference microphone signal and the signal of each remaining microphones. The NCC feature can be formulated as:

$$\begin{cases} \mathbf{chunk}_{c,s,m} = \mathbf{x}_{c,s}[m : m + W - 1] \\ \mathbf{q}_{c,s,m} = \frac{\mathbf{x}_{0,s}(\mathbf{chunk}_{c,s,m})^T}{||\mathbf{x}_{0,s}||_2||\mathbf{chunk}_{c,s,m}||_2} \end{cases}, m = 1, ..., 2L + 1 \tag{3}$$

where $\mathbf{q}_{c,s,m}$ is the NCC feature between $\mathbf{x}_{0,s}$ and $\mathbf{chunk}_{c,s,m}$. $\mathbf{x}_{0,s}$ and $\mathbf{chunk}_{c,s,m}$ is the center segments of the reference microphone and the context segments of the c-th microphone, respectively. The mean of the obtained NCC features $M(\mathbf{q}_{(0,...,C-1),s,m})$ is then concatenated with the context segments' encoding of reference microphone $E(\mathbf{chunk}_{0,s,m})$. Then, the results are fed into N_1 Dual-Path RNN (DPRNN) [5] blocks $\mathscr{F}(*)$ to calculate the beamformer coefficients of the reference microphone $\mathbf{h}_{0,s,k}$ where $k = 0, 1, ..., K-1$. K is the number of the speakers. The estimation of the reference microphone $\mathbf{y}_{0,s,k}$ is then obtained by filtering $\mathbf{chunk}_{0,s,m}$ with the beamformer coefficients $\mathbf{h}_{0,s,k}$:

$$\mathbf{h}_{c,s,k} = \mathscr{F}_{TAC}([E(\mathbf{chunk}_{c,s,m}), \mathbf{g}_{c,s,m}]) \tag{4}$$

$$\mathbf{y}_{s,k} = \mathbf{y}_{0,s,k} + \sum_{c=1}^{C-1} \mathbf{chunk}_{c,s,m} \circledast \mathbf{h}_{c,s,k} \tag{5}$$

Unlike the above two-stage architecture, the single-stage FaSNet-TAC model jointly estimates the beamformer coefficients for all input channels and applies the TAC module between all DPRNN blocks.

B. FaSNet-TAC for Speaker Extraction

We propose two main options for using the FaSNet-TAC to extract the target source. The first solution, called FaSNet-TAC+Sim, is to use the original FaSNet-TAC to obtain the separation results of multiple speakers and then filter the target source through the cosine similarity in the inference stage. The second solution, called Tar-FaSNet-TAC, is to utilize the target speaker embedding during the training stage to obtain a speaker extraction model based on FaSNet-TAC, which can directly extract the target source.

Figure 3 (A) shows the structure schematic Tar-FaSNet-TAC model, which consists of a speaker embedding extraction module and a speaker extraction module. The ResNet34 used as the speaker embedding extraction module in this work is based on [22]. The speaker extraction model consists of N DPRNN blocks with TAC. The speaker embedding extraction module first extract the target speaker embedding $\mathbf{e} \in \mathbb{R}^{1 \times D}$, where D is the dimension of the frequency bins. Then the

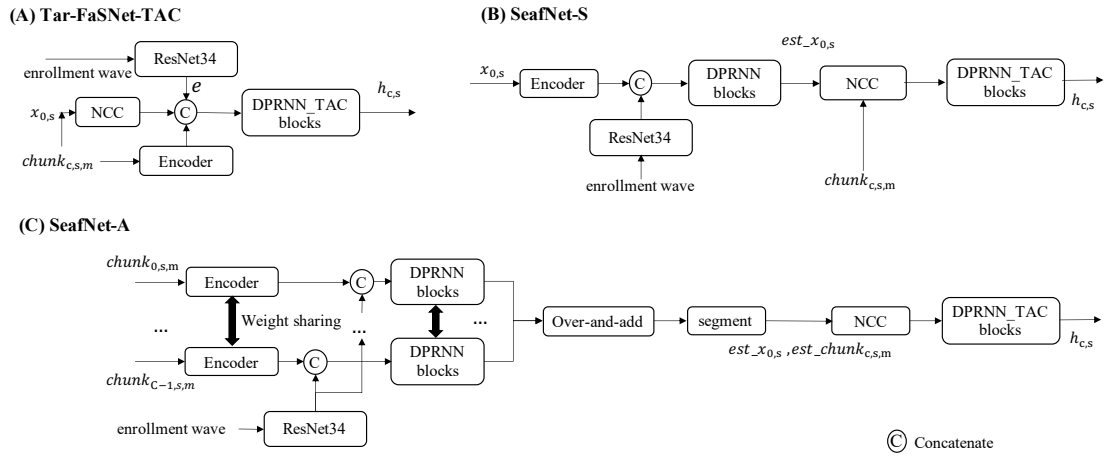

Fig. 3. System flowchart of the SeafNet models. (A) Multi-channel speaker extraction model based on FaSNet-TAC. (B) The SeafNet-S model applies speaker extraction on each microphone. (C) The SeafNet-A model only applies speaker extraction on reference microphone.

speaker extraction module calculates the NCC feature $\mathbf{g}_{c,s,m}$ between the reference signal and multi-channel inputs. The obtained $\mathbf{e}$ and $\mathbf{g}_{c,s,m}$ are then concatenated with the encoder of the context segments $\mathbf{chunk}_{c,s,m}$ to derive the target speaker's filter coefficients $\mathbf{h}_{c,s}$, where $c \in 0, ..., C-1$:

$$\mathbf{h}_{c,s} = \mathscr{F}_{TAC}([E(\mathbf{chunk}_{c,s,m}), \mathbf{g}_{c,s,m}, \mathbf{e}]) \tag{6}$$

Finally, the model generate the estimated target source by filtering multi-channel inputs with $\mathbf{h}_{c,s}$.

C. SeafNet

Although the solutions in Section II-B work well, the model's size and computation costs are significantly increased when processing fusion features. In order to address this problem, we further propose the SeafNet framework. Figure 3 shows the two different structures of SeafNet, named SeafNet-A and SeafNet-S, respectively.

1) SeafNet-A: SeafNet-A in Figure 3 (C) first applies single-channel speaker extraction on the context segments of all input microphones:

$$\mathbf{est}_{c,s,m} = \mathscr{F}([E(\mathbf{chunk}_{c,s,m}), \mathbf{e}]), \quad c = 0, ..., C-1 \tag{7}$$

where $\mathscr{F}(*)$ indicates the DPRNN blocks, C indicates the number of microphones, and $\mathbf{e}$ is the target speaker embedding. Next, we obtain the center segments $\mathbf{est\_x}_{c,s}$ and context segments $\mathbf{est\_chunk}_{c,s,m}$ of the pre-separated output through overlap-and-add and split operation. Then SeafNet-A calculates the NCC feature between $\mathbf{est\_x}_{0,s}$ and $\mathbf{est\_chunk}_{c,s,m}$ and estimates the target speaker's filter coefficients:

$$\mathbf{g}_{c,s,m} = \frac{\mathbf{est\_x}_{0,s}(\mathbf{est\_chunk}_{c,s,m})^T}{||\mathbf{est\_x}_{0,s}||_2||\mathbf{est\_chunk}_{c,s,m}||_2} \tag{8}$$

$$\mathbf{h}_{c,s} = \mathscr{F}_{TAC}([E(\mathbf{est\_chunk}_{c,s,m}), \mathbf{g}_{c,s,m}]) \tag{9}$$

where $\mathscr{F}_{TAC}(*)$ indicates the DPRNN with TAC blocks. Finally, $\mathbf{h}_{c,s}$ are convolved with $\mathbf{est\_chunk}_{c,s,m}$ and generate the separation results like the original FaSNet-TAC.

2) SeafNet-S: SeafNet-S in Figure 3 (B) first applies single-channel speaker extraction only on the reference microphone's center segments, not on the context segments of all microphones as in the SeafNet-A model:

$$\mathbf{est}_{0,s} = \mathscr{F}([E(\mathbf{x}_{0,s}), \mathbf{e}]) \tag{10}$$

where $E(\mathbf{x}_{0,s})$ indicates the encoder of the reference microphone's center segments. $\mathbf{e}$ is the target speaker embedding. Then SeafNet-S calculates the NCC feature $\mathbf{g}_{c,s,m}$ between $\mathbf{est}_{0,s}$ and all microphones' context segments $\mathbf{chunk}_{c,s,m}$. The target speaker's filter coefficients can be calculated as:

$$\mathbf{g}_{c,s,m} = \frac{\mathbf{est}_{0,s}(\mathbf{chunk}_{c,s,m})^T}{||\mathbf{est}_{0,s}||_2||\mathbf{chunk}_{c,s,m}||_2} \tag{11}$$

$$\mathbf{h}_{c,s} = \mathscr{F}_{TAC}([E(\mathbf{chunk}_{c,s,m}), \mathbf{g}_{c,s,m}]) \tag{12}$$

At last, $\mathbf{h}_{c,s}$ are convolved with $\mathbf{chunk}_{c,s,m}$ and generate the final estimation.

III. Experiment Setup

A. Dataset

Datasets for speaker embedding extraction module: We use the development set of the VoxCeleb2 [23], which contains over 1 million utterances from 5991 speakers, to train a ResNet34-based speaker verification model. We use the test set of the VoxCeleb1 [23] for testing. The test set contains 4715 utterances from 40 celebrities and has no overlap with the identities in the VoxCeleb2.

Datasets for speaker extraction module: We use the same dataset in [16], a multi-channel two-speaker noisy reverberant dataset with both ad-hoc and fixed geometry microphone arrays. Each class of microphone array includes 20000, 5000, and 3000 4-second long utterances for training, validation and test, respectively. The source utterances in these multi-channel mixed waveform are selected from the 100-hour Librispeech [24], and we randomly select an additional utterance as the enrollment waveforms for each source.

TABLE I

RESULTS ON THE 6-MIC FIXED GEOMETRY ARRAY. **SV**: WITH OR WITHOUT A SPEAKER VERIFICATION MODEL IN THE INFERENCE STAGE. **EMB**: WITH OR WITHOUT A LOCAL TARGET SPEAKER EMBEDDING EXTRACTED IN ADVANCE. **FASNET-TAC+SIM**: ORIGINAL SINGLE-STAGE FASNET-TAC COMBINED WITH TARGET SOURCE SELECTION THROUGH COSINE SIMILARITY IN THE INFERENCE STAGE. **TAR-FASNET-TAC**: MULTI-CHANNEL TARGET SPEAKER EXTRACTION MODEL BASED ON THE SINGLE-STAGE FASNET-TAC AND RESNET34 MODELS DESCRIBED IN SECTION II-B.

Methods	SV	Emb	# of Params(M)	FLOPS(G)	SISNRi(dB)
Baseline: FaSNet-TAC + Sim	w/	w/	9.0(2.9+6.1)	64.31(55.09 + 2*4.61)	10.92
Proposed: Tar-FaSNet-TAC	w/o	w/	6.3	148.10	**12.15**
Proposed: SeafNet-A	w/o	w/	3.6	82.30	11.64
Proposed: SeafNet-S	w/o	w/	**3.6**	**37.03**	11.62

TABLE II

RESULTS ON THE AD-HOC ARRAY. THE PARAMS, FLOPS, AND SISNRI OF THE METHODS ARRAYS WITH 2 / 4 / 6 MICROPHONES ARE REPORTED.

Methods	SV	Emb	# of Params(M)	FLOPs(G)	SISNRi(dB)
Baseline: FaSNet-TAC + Sim	w/	w/	9.0	29.2 / 46.75 / 64.31	9.72 / 9.67 / 11.74
Proposed: Tar-FaSNet-TAC	w/o	w/	6.5	50.98 / 99.54 / 148.10	**11.01 / 12.55 / 12.94**
Proposed: SeafNet-A	w/o	w/	3.6	28.24 / 55.27 / 82.30	10.5 / 11.59 / 12.26
Proposed: SeafNet-S	w/o	w/	**3.6**	**19.18 / 28.11 / 37.03**	10.10 / 11.55 / 12.15

B. Implementation details

Speaker embedding extraction module: We use the default ResNet34V2[1] to extract a 256-dimensional embedding for the target speaker. Specifically, we use the SAP [25] encoder and the ge2e [26] loss function in our training.

Speaker extraction module: We have compared the performance of four models: 1) FaSNet-TAC[2]+Sim, 2) Tar-FaSNet-TAC, 3) proposed SeafNet-A, and 4) proposed SeafNet-S. The results in 1) are the baseline. We use 4 DPRNN blocks in all models and the hyperparameters are the same as [16]. Specifically, we apply TAC between all the DPRNN blocks in 1) and 2), while in 3) and 4) we use it only in the last two blocks. The length of the center segment W is 4ms and the length of context segment L is 16ms. The scale-invariant SNR (SISNR) [27] is used as the training target. The SISNR improvement (SISNRi) is used as the separation performance metric. In addition, we evaluate the FLOPs and Params information of all models.

IV. RESULTS AND DISCUSSIONS

Table I shows the experimental results on the 6-mic fixed geometry array. Among all schemes, the Tar-FaSNet-TAC model obtained the maximum SI-SNRi improvement, which achieved 1.23 dB SISNRi improvement over the baseline. However, its FLOPs' cost increases significantly because the dimensionality of the fusion features is much larger than the one of the original model inputs. The SeafNet-A further reduces the number of model parameters, but the FLOPs cost is still higher than the baseline due to the iterative pre-separation of each microphone. The SeafNet-S model has a comparable separation performance to the SeafNet-A model, which achieves a 6.4% relative improvement over the FaSNet-TAC+Sim baseline model. It can be noticed that the proposed SeafNet-S method has the lowest model size and computational cost.

Table II shows the results on ad-hoc array and we only report the results on 2, 4, and 6 microphones to compare with [16]. The results in Table II shows similar trends as in Table I. Specifically, the Tar-FaSNet-TAC model achieves the largest SISNR improvement, but accompanied with an increase in FLOPs' costs. The SeafNet-S model achieves 3.5% relative SISNRi improvement, 42% relative FLOPs reduction, and 60% relative Params reduction against the baseline (6-mic), respectively. We note that in the 2-mic scenario, the SeafNet-S still outperforms the baseline, but there is a slight degradation in the performance compared to the SeafNet-A model. It may be because the output of pre-separation account for a notable contribution to the calculation of the NCC features, which could greatly affect the separation performance of the model.

Although the SISNRi performance of our proposed SeafNet-A and SeafNet-S is slightly lower than that of Tar-FaSNet-TAC, the SeafNet model outperforms the Tar-FaSNet-TAC model by a large margin in terms of Params' and FLOPs' cost. In addition, the results in Table I and Table II show that SeafNet-S outperforms the baseline on both fixed geometry array and ad-hoc array with lower Params' and FLOPs' cost. It means that for the speaker extraction task, the improvement brought by our proposed SeafNet models is quite effective.

V. CONCLUSION

In this work, we extend the FaSNet-TAC model to multi-channel target speaker extraction task and propose a Speaker extraction-and-filter Network (SeafNet). SeafNet first extracts the target source from the signal of the reference microphone by utilizing auxiliary speech. Then, SeafNet calculates the NCC feature between the estimated target source and multi-channel input. Finally, SeafNet derives a set of filter coefficients using the fusion feature of the input and NCC feature. The filter result of the input is our final output. Experimental results show that our proposed SeafNet achieves a favorable separation performance improvement while maintaining relatively small model size and low computational costs.

VI. ACKNOWLEDGEMENT

This research is funded in part by the Science and Technology Program of Suzhou City (SYC2022051) and OPPO. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

[1]https://github.com/clovaai/voxceleb_trainer

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

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. of ICASSP*, IEEE, 2017, pp. 246–250.

[3] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. of ICASSP*, IEEE, 2017, pp. 241–245.

[4] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of ICASSP*, IEEE, 2018, pp. 696–700.

[5] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. of ICASSP*, IEEE, 2020, pp. 46–50.

[6] T. Chou, "Frequency-independent beamformer with low response error," in *Proc. of ICASSP*, IEEE, vol. 5, 1995, pp. 2995–2998.

[7] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[8] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2009.

[9] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[10] 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 *Proc. of ICASSP*, IEEE, 2017, pp. 276–280.

[11] 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. of Interspeech*, 2016, pp. 1981–1985.

[12] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Dnn-based speech mask estimation for eigenvector beamforming," in *Proc. of ICASSP*, IEEE, 2017, pp. 66–70.

[13] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. of ICASSP*, IEEE, 2017, pp. 271–275.

[14] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Processing Letters*, vol. 28, pp. 1370–1374, 2021.

[15] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. of ASRU*, IEEE, 2019, pp. 260–267.

[16] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. of ICASSP*, IEEE, 2020, pp. 6394–6398.

[17] T. Li, Q. Lin, Y. Bao, and M. Li, "Atss-net: Target speaker separation via attention-based neural network," in *Proc. of Interspeech*, 2020.

[18] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1370–1384, 2020.

[19] 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," in *Proc. of ICASSP*, IEEE, 2021, pp. 6089–6093.

[20] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized spatio-temporal rnn beamformer for target speech separation," in *Proc. of Interspeech*, 2021.

[21] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-spex: Localized target speaker extraction," in *Proc. of ICASSP*, IEEE, 2022, pp. 7287–7291.

[22] Y. Kwon, H. S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: Lessons from VoxSRC 2020," in *Proc. of ICASSP*, IEEE, 2021.

[23] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2020.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. of ICASSP*, IEEE, 2015, pp. 5206–5210.

[25] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Proc. of Interspeech*, 2017, pp. 1517–1521.

[26] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. of ICASSP*, IEEE, 2018, pp. 4879–4883.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *Proc. of ICASSP*, IEEE, 2019, pp. 626–630.