Learnable Cross-Correlation based Filter-and-Sum Networks for Multi-channel Speech Separation

Xianrui Wang*†, Shiqi Zhang†‡, Bo He†, Shoji Makino†, and Jingdong Chen*
* CIAIC and Shaanxi Provincial Key Laboratory of Artificial Intelligence,
Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
† Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 808-0135, Japan
‡ Audio Research Group, Tampere University, Tampere, Finland

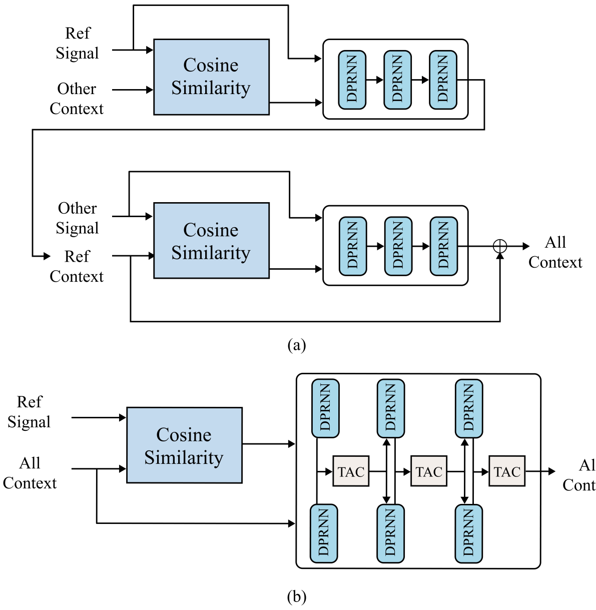
***Abstract*—Multichannel source separation plays an important role in audio and speech signal processing. With recent advancements in deep neural networks (DNN), numerous DNN-based beamforming algorithms have been developed. To leverage spatial information, a time domain filter-and-sum network (FaSNet) was introduced, and the transform average concatenate (TAC) technique was subsequently adopted to further enhance separation performance. FaSNet captures spatial information by assessing cosine similarity between different channels; but this approach may have limited spatial resolution and could exhibit bias in noisy, reverberant environments, thereby potentially compromising performance. Motivated by the efficacy of the generalized cross-correlation (GCC) method in achieving reliable source localization in adverse environments, this paper introduces a learnable cross-correlation (LCC) module for FaSNet and FaSNet-TAC. By offering improved flexibility and robustness across diverse environments, LCC enhances source separation performance, which is validated by several simulations.**

***Index Terms*—Multichannel source separation, neural network based beamfroming, spatial information, learnable cross-correlation.**

I. Introduction

Separating different source signals in noisy and reverberant environments [1]–[3] is of vital importance. Beamforming is one of the most widely used methods to address this problem [4]–[7]. With recent advancements in deep neural networks (DNN), numerous DNN-based beamforming algorithms have been proposed [8]–[10]. As shown in [3], [7], incorporating spatial information can help significantly improve the separation performance. To this end, a filter-and-sum network (FaSNet) [11] was proposed. The original FaSNet comprises several Temporal Convolutional Network (TCN) modules [12], which are employed to separate source signals. Additionally, cosine similarity modules are utilized to capture spatial information. To further enhance separation performance, TCN modules are replaced with dual-path recurrent neural network (DPRNN) modules [13], [14]. Moreover, the transform average concatenate (TAC) technique was adopted [14]. TAC first transforms each channel's feature into a latent space and then the average of transformed features is concatenated into following blocks.

While FaSNet [11] and FaSNet-TAC [14] have demonstrated promising separation performance, they rely on cosine similarity to extract spatial information, which may suffer from limited spatial resolution and bias in noisy, reverberant environments. Generalized cross-correlation (GCC) is one of the most widely used methods in source localization [15]–[19]. GCC modifies the cross-correlation through a frequency domain weight function. There is a number of member algorithms in the GCC family, which consider different weight functions and their proprieties depend on the weight functions. For example, phased transform (PHAT) [20] is robust to reverberation while Roth processor [21] and smoothed coherence transform (SCOT) [22] are robust to noise. Inspired by the principles in GCC, we propose to design in this paper a learnable cross-correlation (LCC) module, in which a time domain convolution function is learned as the weight function from input data adaptively. By replacing the cosine similarity with the proposed LCC module, we derive two new frameworks, i.e., LCC-FaSNet and LCC-FaSNet-TAC. These frameworks offer enhanced separation performance, particularly in scenarios where two target speakers are closely distributed, due to LCC's ability to achieve higher spatial resolution. Moreover, as the LCC module dynamically learns the weight function from input data, the proposed frameworks



(a)

(b)

Fig. 1. Structures of: (a). two-stage FaSNet, (b). single-stage FaSNet-TAC.

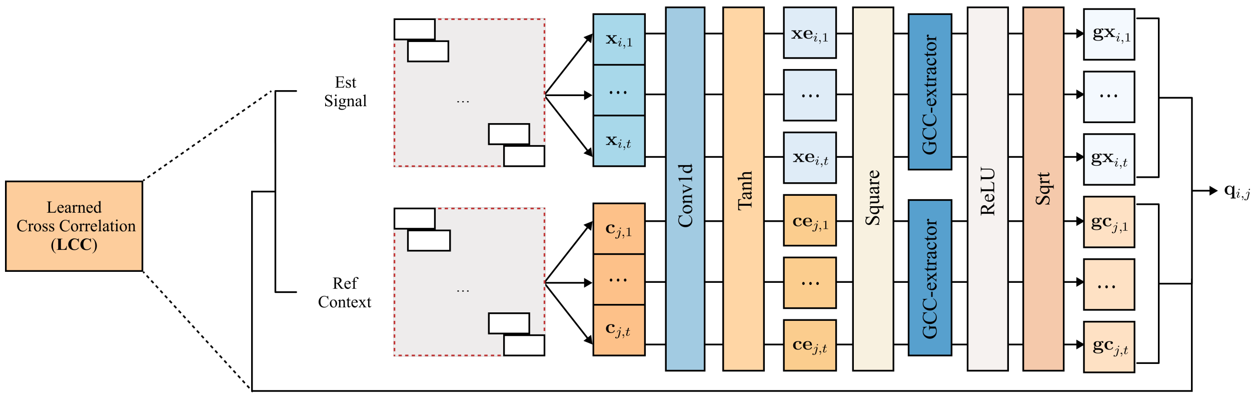


Fig. 2. The structure of the proposed LCC module.

are anticipated to outperform conventional approaches across diverse environments.

The remainder of the paper is organized as follows. In Section II, we present a brief overview of the cosine similarity feature used in the original FaSNet and FaSNet-TAC methods. We then introduce the proposed LCC module in Section III. In Section IV, we present the training configuration and compare the proposed methods with their conventional counterparts. Finally, we draw the conclusion in Section V.

II. Cosine similarity feature

Incorporating spatial information properly is shown to be an effective approach to improve source separation performance [3], [7]. In this paper, we consider the original two-stage FaSNet structure [11] and single-stage FaSNet-TAC structure [14], both of which incorporate DPRNN modules. As depicted in Fig. 1 (a) and Fig. 1 (b), FaSNet and FaSNet-TAC extract the spatial information using the cosine similarity feature [11], [14], which is named normalized cross-correlation (NCC). Mathematically, NCC is defined as

$$\mathrm{NCC}_{\mathbf{s}_1,\mathbf{s}_2} = \frac{\mathbf{s}_1^T\mathbf{s}_2}{||\mathbf{s}_1||_2||\mathbf{s}_1||_2}, \tag{1}$$

where $\mathbf{s}_1$, $\mathbf{s}_2$ are two signal vectors and $||\cdot||_2$ denotes ℓ_2 norm.

For the two-stage FaSNet, in the initial stage, a single microphone is selected as the reference. The NCCs between the signals picked up by the other microphones and the reference microphone are computed to estimate the target signal. In the subsequent stage, the NCCs between the microphone signals and the estimated target signal are employed to further enhance the separation performance. While for the single-stage FaSNet-TAC, NCCs are only calculated at the beginning and the spatial information are maintained with TAC blocks.

III. Proposed Learnable Cross-Correlation Module

While FaSNet and FaSNet-TAC have demonstrated promising separation performance, they rely on cosine similarity to extract spatial information, which may suffer from limited spatial resolution and bias in noisy, reverberant environments. To circumvent this limitation, we introduce a new, lightweight LCC module in this work. The LCC module, as illustrated in Fig. 2, is a dual-layer convolutional neural network operating at the time-domain frame level. It first partitions the input signals $\mathbf{x}_i$, $i = 1, ..., N$, into short frames (named "chunk") with a frame length of L and a hop size of $H \in [0, L-1]$, i.e.,

$$\mathbf{x}_{i,t} = \mathbf{x}_i[tH : tH + L - 1], \tag{2}$$

where t is the frame index. We additionally form a signal vector at the corresponding time index, referred to as the "**context**", consisting of extra W immediate future and past samples relative to $\mathbf{x}_{i,t}$, i.e.,

$$\mathbf{c}_{j,t} = \mathbf{x}_j[tH - W : tH + L + W - 1]. \tag{3}$$

For simplicity, we omit the frame index t in subsequent discussions, as it is unambiguous. Following input signal segmentation, we utilize a pair of convolutional layers with shared parameters, labeled as $Conv_1(\cdot)$, to encode both the context signal $\mathbf{c}_j$ and the chunk signal $\mathbf{x}_i$. Subsequently, in order to preserve the dynamic range of subsequent neural network inputs, we scale their outputs to the range $[-1, 1]$ using the $\tanh(\cdot)$ function. Mathematically, this process is expressed as

$$\begin{aligned} \mathbf{ce}_j &= \tanh\left[Conv_1(\mathbf{c}_j)\right], \\ \mathbf{xe}_i &= \tanh\left[Conv_1(\mathbf{x}_i)\right], \end{aligned} \tag{4}$$

where $\mathbf{ce}_j \in \mathbb{R}^{O_{c1}\times(2W+L-I_{c1}+1)}$ denotes the encoded context signal, I_{c1} and O_{c1} denote, respectively, the number of input and output channels of the first convolutional layer (In this paper, we set $I_{c1} = O_{c1} = L$ for simplicity in computation), $\mathbf{xe}_i \in \mathbb{R}^{O_{c1}\times 1}$ is the encoded chunk signal.

Subsequently, the encoded vector is squared and passed through the GCC extractor (a single convolutional layer), denoted as $Conv_2(\cdot)$, to dynamically adjust its information in the O_{c1} dimension. The ReLU activation function is then applied to remove negative values, and the square root operation is performed to yield the corresponding Pearson coefficient-like

TABLE I

EXPERIMENTAL RESULTS BASED ON A 6-ELEMENT CIRCULAR MICROPHONE ARRAY WITH DIFFERENT SPEAKER ANGLES.

Model	# Parameters	SDRi (dB)					SI-SNRi (dB)				
		Speaker angle				Overall	Speaker angle				Overall
		<15°	15-45°	45-90°	>90°		<15°	15-45°	45-90°	>90°	
FaSNet	3.70M	4.77	5.50	6.18	6.59	5.81	2.93	3.58	4.15	4.48	3.83
+LCC (proposal)	+4.22K	**5.02**	**5.96**	**6.75**	**7.05**	**6.26**	**3.99**	**4.84**	**5.57**	**5.79**	**5.11**
FaSNet-TAC	2.76M	8.48	10.55	12.19	12.88	11.06	7.75	9.85	11.49	12.23	10.48
+LCC (proposal)	+4.22K	**9.69**	**11.58**	**12.98**	**13.60**	**12.09**	**8.94**	**10.87**	**12.28**	**12.94**	**11.39**

TABLE II

EXPERIMENT RESULTS BASED ON A 6-ELEMENT CIRCULAR MICROPHONE ARRAY WITH DIFFERENT OVERLAP RATES

Model	# Parameters	SDRi (dB)				SI-SNRi (dB)			
		Overlap rate				Overlap rate			
		<25%	25-50%	50-75%	>75%	<25%	25-50%	50-75%	>75%
FaSNet	3.70M	7.98	5.87	4.57	3.54	5.60	3.88	2.85	1.94
+LCC (proposal)	+4.22K	**8.74**	**6.21**	**4.85**	**3.80**	**7.31**	**5.11**	**3.86**	**2.84**
FaSNet-TAC	2.76M	15.82	10.70	8.74	7.17	14.97	9.98	8.02	6.48
+LCC (proposal)	+4.22K	**16.63**	**11.51**	**9.59**	**8.18**	**15.98**	**10.79**	**8.86**	**7.45**

term [23], [24], i.e.,

$$\begin{aligned} \mathbf{gc}_j &= \sqrt{\mathrm{ReLU}\big[Conv_2(\mathbf{ce}_j^2)\big]}, \\ \mathbf{gx}_i &= \sqrt{\mathrm{ReLU}\big[Conv_2(\mathbf{xe}_i^2)\big]}. \end{aligned} \tag{5}$$

Finally, the learnable cross-correlation feature is calculated with following equation:

$$\begin{cases} \mathbf{c}_{j,p} = \mathbf{c}_i[p : p + L - 1] \\ \mathbf{f}_{i,j,p} = \dfrac{\mathbf{x}_i \mathbf{c}_{j,p}^T}{\|\mathbf{gc}_{j,p}\|_2 \|\mathbf{gx}_i\|_2} \end{cases}, \quad j = 1, ..., 2W + 1, \tag{6}$$

where $p \in [0, 2W - I_{c1} + 1]$ represents the position of the focused same-length segment in $\mathbf{c}_j$. It is worthy note that the proposed module modifies the cross correlation through a time domain convolution, which is equivalent to the frequency domain multiplication, therefore it can be regarded as a learnable GCC module.

IV. EXPERIMENT

In this section, we provide a detailed description of the training configuration and conduct several simulations in both noisy and reverberant environments. To validate the effectiveness of the proposed LCC module, we conduct a comparative analysis of the proposed networks exclusively against their conventional counterparts, i.e., FaSNet and FaSNet-TAC. The proposed module could also be integrate into other networks.

A. Dataset

The target signals are randomly selected from LibriSpeech [25]. Noise signals from CHiME3 [26] are chosen to control the signal-to-noise ratio (SNR). The sampling rate for both target and noise signals is $16\,\mathrm{kHz}$. We consider scenarios where there are two target speakers with the relative signal-to-interference ratio (SIR) ranging from $0\,\mathrm{dB}$ to $5\,\mathrm{dB}$. One point noise is added and the corresponding SNR is set to range from $-5\,\mathrm{dB}$ and $30\,\mathrm{dB}$.

We consider a room with randomly sampled dimensions, where the length and width fall between 3 and 10 m, and the height ranges from 2.5 to 4 m. The reflection coefficients of all the room surfaces are assumed identical and they are randomly generated to achieve a reverberation time T_{60} ranging from $100\,\mathrm{ms}$ to $500\,\mathrm{ms}$. A uniform circular array with a radius of $5\,\mathrm{cm}$ and consisting of 6 omnidirectional microphones, is utilized. The array is positioned randomly within the room, ensuring that the minimum distance between the array center and each wall is $0.5\,\mathrm{m}$. Speaker positions are then sampled to ensure the average speaker angle relative to the microphone center is uniformly distributed between $0°$ and $180°$. Noise source position is sampled without additional constraints. Subsequently, the microphone observation signals are generated by convolving the speech and noise source signals with their corresponding room impulse responses (RIRs) generated using the toolbox gpuRIR [27], which is based on the image model [28] and adding the results together with the specified SIR and SNR values. In total, 100,000, 10,000, and 5,000 4-second-long utterances are generated as the training, validation, and evaluation sets, respectively.

B. Experimental configuration

Using the generated datasets, we trained and tested the proposed LCC-FaSNet and LCC-FaSNet-TAC, as well as the original baseline models without LCC. Models were trained using the Adam [29] optimizer with a momentum of 0, batch size of 8, and 120 epochs. The initial learning rate

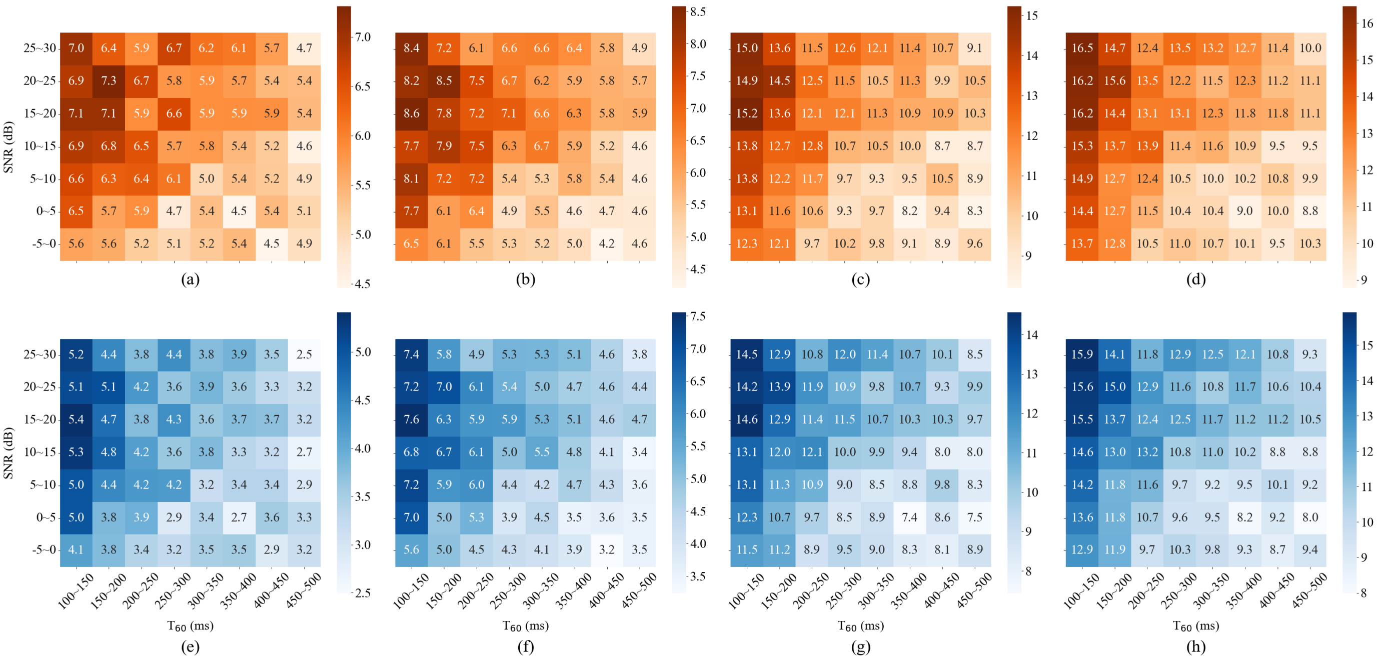


Fig. 3. Performance of the compared methods in noisy and reverberant environments: (a) SDRi of FaSNet, (b) SDRi of LCC-FaSNet, (c) SDRi of FaSNet-TAC, (d) SDRi of LCC-FaSNet-TAC, (e) SI-SNRi of FaSNet, (f) SI-SNRi of LCC-FaSNet, (g) SI-SNRi of FaSNet-TAC, and (h) SI-SNRi of LCC-FaSNet-TAC. The horizontal axis represents T_{60} and the vertical axis represents SNR (dB).

was set to 2×10^{-3}. If the loss did not decrease after 3 consecutive epochs, the learning rate was halved. Training was halted either when the learning rate did not decrease for two successive epochs or when the maximum number of iterations was reached. The model with the lowest loss on the validation set was chosen for testing and results recording. All experiments were conducted on an NVIDIA A40 with 48 GB of GPU memory.

C. Results and Discussions

Before examining performance, we analyze the additional memory cost of the proposed LCC module. As shown in Tab. I and Tab. II, the proposed LCC module contributes only $4.22\,\mathrm{k}$ to the parameter count, which is less than 0.2% of the parameter volume of the baseline models. Specifically, for FaSNet, LCC utilizes an additional $4.22\,\mathrm{k}$ (0.11%) parameters, and for FaSNet-TAC, LCC uses only 0.15% more parameters.

Now, we investigate the separation performance with variations in the angle between two target speakers. The results are shown in Tab. I. As seen, the proposed methods consistently outperform their conventional counterparts across all cases. Furthermore, the TAC technique significantly improves separation performance. Notably, as the angle between the two speakers decreases, especially below $15°$, the proposed frameworks achieve notably superior separation performance. This improvement is attributed to the limited spatial resolution of cosine similarity, which leads to performance degradation in FaSNet and FaSNet-TAC. Conversely, the proposed LCC module offers higher spatial resolution, thereby enhancing separation performance. Additionally, as seen in Tab. II, the proposed methods achieve superior separation performance across various overlap ratios between the two speakers, indicating the suitability of the LCC module for diverse applications.

Moreover, segmented tests were conducted across different SNR levels and T_{60} values. The results are presented in Fig. 3 where darker colors indicate larger gain under the current conditions, while lighter colors signify less gain. It is evident that the proposed methods consistently outperform their counterparts across all configurations, which validates the robustness and flexibility of the proposed LCC module.

V. Conclusion

Separating source signals in noisy and reverberant environments is an important task in audio signal processing. In the so-called FaSNet and FaSNet-TAC, cosine similarity is chosen to represent the inter-channel feature. However, this choice possesses limited spatial resolution and may not be optimal across varied environments. To take full advantage of spatial information, in this paper, we designed a learnable cross-correlation (LCC) module, which dynamically generates a weighted correlation from input observations. We integrated the proposed module into FaSNet and FaSNet-TAC. Simulations confirm that the proposed LCC module enhances separation performance, particularly in scenarios with closely distributed target speakers, owing to its high resolution. Furthermore, improvements are observed across different overlap ratios, SNR levels, and reverberation levels, underscoring the robustness and flexibility of the proposed LCC module.

References

[1] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 715–726, 2007.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays: signal processing techniques and applications*, Springer, 2001, pp. 157–180.

[3] X. Wang, A. Brendel, G. Huang, Y. Yang, W. Kellermann, and J. Chen, "Spatially informed independent vector analysis for source extraction based on the convolutive transfer function model," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[5] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[6] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 2008, vol. 1.

[7] Y. Yang, X. Wang, W. Zhang, and J. Chen, "Independent vector analysis assisted adaptive beamfomring for speech source separation with an acoustic vector sensor," in *Proc. IWAENC*, 2022, pp. 1–5.

[8] X. Xiao, S. Watanabe, H. Erdogan, *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 5745–5749.

[9] 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. IEEE ICASSP*, 2017, pp. 271–275.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, 2017.

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

[12] 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, pp. 1256–1266, 2019.

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

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

[15] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *J. Signal Process. Sys.*, vol. 63, pp. 265–275, 2011.

[16] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, 2019.

[17] X. Wang, G. Huang, J. Benesty, J. Chen, and I. Cohen, "Time difference of arrival estimation based on a kronecker product decomposition," *IEEE Signal Process. Lett.*, vol. 28, pp. 51–55, 2020.

[18] Y. Zou and H. Liu, "TDOA localization with unknown signal propagation speed and sensor position errors," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1024–1027, 2020.

[19] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1270–1281, 2020.

[20] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, 2012.

[21] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE spectrum*, vol. 8, no. 4, pp. 62–70, 1971.

[22] E. A. Cohen and A. T. Walden, "A statistical study of temporally smoothed wavelet coherence," *IEEE Trans. on Signal Process.*, vol. 58, no. 6, pp. 2964–2973, 2010.

[23] J. Benesty, J. Chen, and Y. Huang, "On the importance of the pearson correlation coefficient in noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 757–765, 2008.

[24] I. Cohen, Y. Huang, J. Chen, *et al.*, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 37–440, 2009.

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

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.

[27] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "GpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimed. Tools and Appl.*, vol. 80, no. 4, pp. 5653–5671, Oct. 2020.

[28] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

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