On Joint Dereverberation and Single Moving Source Separation with Online Source Steering

Yiting Zhang†, Kaien Mo†, Tetsuya Ueda†, Yichen Yang† and Shoji Makino†
† Waseda University, Japan

***Abstract*—To solve the Blind Source Separation (BSS) problems in time-variant and highly reverberant environments, an online joint optimization algorithm for Weighted Prediction Error (WPE)-based dereverberation and Independent Vector Analysis (IVA)-based source separation (online WPE-IVA) has been reported. On the other hand, Online Source Steering (OSS) was recently introduced for filter updating in online IVA with low computational cost for a target-source tracking task where only one speaker is moving. Inspired by this work, we integrate OSS into online WPE-IVA, aiming to reduce the computational cost of updating separation filters. Experimental results show that our proposed method reduces computational costs while achieving comparable separation performance to conventional online WPE-IVA in target-source tracking tasks in highly reverberant environments.**

I. INTRODUCTION

Blind Source Separation (BSS) is a group of methods aiming to extract clean source signals from mixtures without any prior information [1]. Short-time Fourier transform (STFT)-domain Independent Component Analysis (ICA) is a popular multi-channel BSS method, which estimates separation filters based on the statistical independence between the sources [2]. Since the STFT-domain ICA calculates the separation filters in each frequency bin independently, the output permutation of sources may not be aligned in different frequency bins. Independent Vector Analysis (IVA) is proposed to address the above permutation problem by leveraging the high-order dependencies among frequency bins [3], [4]. Furthermore, the majorization-minimization algorithm is used for IVA's separation filter updating, which is widely known as Auxiliary-function-based IVA (AuxIVA) [5].

IVA-based methods are initially used in offline processing, which will cause large latency and cannot adapt to time-variant environments. As a solution for these problems, online IVA-based methods are proposed [6]–[8]. Unlike the offline IVA which processes a long batch of data at once, the online IVA updates separation filters incrementally after every new frame is received. This makes it suitable for real-time processing under time-variant environments. Moreover, to accelerate computational efficiency, previous research proposed an algorithm, Online Source Steering (OSS) [9] for online IVA, which updates the separation filters under the task where a certain source moves and other sources are spatially stationary (target-source tracking task).

The methods discussed above suppose that the analysis window of STFT is longer than the length of room impulse responses [10]. This assumption is hard to be satisfied when reverberation is strong and the IVA-based methods suffer from significant performance degradation. To mitigate the impact of strong reverberation, Weighted Prediction Error (WPE)-based dereverberation is usually used before separation [11], [12]. Furthermore, an online algorithm that jointly optimizes WPE and IVA (online WPE-IVA) [13], [14] has been proposed to achieve highly accurate source separation in highly reverberant environments. However, the highly efficient implementation for online WPE-IVA has not been developed yet for the target-source tracking task.

This work aims to further promote the computational efficiency of existing online WPE-IVA [13], [14] for the target-source tracking task. We introduce OSS into updating separation filters efficiently for online WPE-IVA, which is named "WPE-IVA-OSS". Since matrix operation for updating separation filters is equally converted to the scalar operation like [9], the computational cost of WPE-IVA-OSS is lower than existing online WPE-IVA. Simulation experiments are carried out to validate the efficiency of our proposed method in highly reverberant environments.

II. MODELS

A. Separation model

Suppose that there are N speech sources in a reverberant environment and M microphones are used to capture the signals. The captured signals in the STFT domain can be expressed as

$$\boldsymbol{x}_{f,t} = \sum_{\tau=0}^{L_a-1} \boldsymbol{A}_{f,t,\tau}\boldsymbol{s}_{f,t-\tau}, \tag{1}$$

where $t = 1, \cdots, T$ and $f = 1, \cdots, F$ are the indexes of time frames and frequency bins, respectively. T denotes the total number of time frames and F denotes the total number of frequency bins. $\boldsymbol{s}_{f,t} \in \mathbb{C}^{N\times 1}$ and $\boldsymbol{x}_{f,t} \in \mathbb{C}^{M\times 1}$ are the vectors containing the source and microphone signals, respectively. $\boldsymbol{A}_{f,t,\tau} \in \mathbb{C}^{M\times N}$ is the convolutional mixing matrix at time lag τ, and L_a is the number of time lagged matrices. This paper assumes a determined case ($M = N$). Then, the separation rule can be expressed using a convolutional beamformer (CBF) [15], [16] as

$$\boldsymbol{y}_{f,t} = \boldsymbol{Q}_{f,t,0}\boldsymbol{x}_{f,t} + \sum_{\tau=D}^{L+D-1} \boldsymbol{Q}_{f,t,\tau}\boldsymbol{x}_{f,t-\tau}, \tag{2}$$

where $\boldsymbol{Q}_{f,t,0} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{Q}_{f,t,\tau} \in \mathbb{C}^{N\times M}$ are the coefficient matrix of CBF, $\boldsymbol{y}_{f,t} \in \mathbb{C}^{N\times 1}$ is the vector containing the separated signals, D is the prediction delay and L is the length of CBF filters. By the source-wise factorization of CBF [16], the nth separated signal in $\boldsymbol{y}_{f,t}$ can be rewritten into two equations:

$$\boldsymbol{z}_{n,f,t} = \boldsymbol{x}_{f,t} - \boldsymbol{G}^{\mathsf{H}}_{n,f,t}\overline{\boldsymbol{x}}_{f,t}, \tag{3}$$

$$y_{n,f,t} = \boldsymbol{w}^{\mathsf{H}}_{n,f,t}\boldsymbol{z}_{n,f,t}, \tag{4}$$

where $\overline{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}^{\mathsf{T}}_{f,t-D} \ \cdots \ \boldsymbol{x}^{\mathsf{T}}_{f,t-L-D+1}]^{\mathsf{T}} \in \mathbb{C}^{ML\times 1}$ represents a vector containing past microphone signals, $(\cdot)^{\mathsf{T}}$ and $(\cdot)^{\mathsf{H}}$ denote transpose and conjugate transpose, respectively. The first sub-filter $\boldsymbol{G}_{n,f,t} \in \mathbb{C}^{ML\times M}$ is a single-target dereverberation filter for the nth source. The output of this filter, $\boldsymbol{z}_{n,f,t}$, is the dereverberated signal of the nth source. The second sub-filter $\boldsymbol{w}_{n,f,t} \in \mathbb{C}^{M\times 1}$ is a separation filter extracting the nth source.

B. Probabilistic model

Same as WPE-IVA [16], we assume that each separated signal independently follows a zero-mean multivariate complex Gaussian distribution with a time-dependent variance $r_{n,t}$:

$$p(\boldsymbol{y}_{n,t}) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_F, r_{n,t}\boldsymbol{I}_F), \tag{5}$$

where $\boldsymbol{y}_{n,t} = [y_{n,1,t} \ \cdots \ y_{n,F,t}]^{\mathsf{T}} \in \mathbb{C}^{F\times 1}$, $\boldsymbol{0}_F \in \mathbb{C}^{F\times 1}$ is a zero vector, and $\boldsymbol{I}_F$ denotes a $F \times F$ identity matrix. By introducing the forgetting factor $0 < \beta < 1$, the negative log-likelihood function for online processing can be derived:

$$\mathcal{L}(\Theta_t) = \frac{1}{\sum_{t'\le t}\beta^{t-t'}} \sum_{n,f,t'\le t} \beta^{t-t'}\left(\log r_{n,t'} + \frac{|y_{n,f,t'}|^2}{r_{n,t'}}\right) - 2\sum_f \log|\det \boldsymbol{W}_{f,t}|, \tag{6}$$

where $\Theta_t = \{\mathcal{R}_t, \mathcal{G}_t, \mathcal{W}_t\}$ is the parameter set to be estimated, $\mathcal{R}_t = \{r_{n,t}\}_n$, $\mathcal{G}_t = \{\boldsymbol{G}_{n,f,t}\}_{n,f}$, $\mathcal{W}_t = \{\boldsymbol{W}_{f,t}\}_f$, and $\boldsymbol{W}_{f,t} = [\boldsymbol{w}_{1,f,t} \ \cdots \ \boldsymbol{w}_{N,f,t}]^{\mathsf{H}}$ is the separation matrix.

The cost function (6) can be optimized iteratively at every time frame by using a recursive coordinate descent method [13], [17] as

$$\mathcal{R}_t \leftarrow \underset{\mathcal{R}_t}{\operatorname{argmin}}\ \mathcal{L}\left(\Theta_t; \mathcal{R}_{t-1}, \mathcal{G}_{t-1}, \mathcal{W}_{t-1}\right), \tag{7}$$

$$\mathcal{G}_t \leftarrow \underset{\mathcal{G}_t}{\operatorname{argmin}}\ \mathcal{L}\left(\Theta_t; \mathcal{R}_t, \mathcal{G}_{t-1}, \mathcal{W}_{t-1}\right), \tag{8}$$

$$\mathcal{W}_t \leftarrow \underset{\mathcal{W}_t}{\operatorname{argmin}}\ \mathcal{L}\left(\Theta_t; \mathcal{R}_t, \mathcal{G}_t, \mathcal{W}_{t-1}\right). \tag{9}$$

Based on the above update rule, online WPE-IVA [13], [14] have been proposed.

III. PROPOSED METHOD

In this section, we propose WPE-IVA-OSS by introducing the update rule of OSS [9] into online WPE-IVA [13], [14]. We consider the target-source tracking task [9], where only one target source is moving and other $N-1$ sources are stationary. Hereafter, we introduce each update rule for the target-source tracking task. Note that we briefly introduce the updates of $\mathcal{R}_t$ and $\mathcal{G}_t$ for a better reading experience because they are the same as [13].

A. Update of $\mathcal{R}_t$ [13]

After calculating $\boldsymbol{y}_{f,t}$ through (3) and (4) with the filters $\boldsymbol{G}_{n,f,t-1}$ and $\boldsymbol{W}_{f,t-1}$ updated in the previous time frame, $r_{n,t}$ can be updated by

$$r_{n,t} \leftarrow \frac{1}{F}\sum_f |y_{n,f,t}|^2. \tag{10}$$

B. Update of $\mathcal{G}_t$ [13]

According to [13], the update rule for $\boldsymbol{G}_{n,f,t}$ can be written:

$$\boldsymbol{k}_{n,f,t} = \frac{\boldsymbol{R}^{-1}_{n,f,t-1}\overline{\boldsymbol{x}}_{f,t}}{\beta r_{n,t} + \overline{\boldsymbol{x}}^{\mathsf{H}}_{f,t}\boldsymbol{R}^{-1}_{n,f,t-1}\overline{\boldsymbol{x}}_{f,t}}, \tag{11}$$

$$\boldsymbol{R}^{-1}_{n,f,t} = \frac{1}{\beta}\left(\boldsymbol{R}^{-1}_{n,f,t-1} - \boldsymbol{k}_{n,f,t}\overline{\boldsymbol{x}}^{\mathsf{H}}_{f,t}\boldsymbol{R}^{-1}_{n,f,t-1}\right), \tag{12}$$

$$\boldsymbol{G}_{n,f,t} = \boldsymbol{G}_{n,f,t-1} + \boldsymbol{k}_{n,f,t}\boldsymbol{z}^{\mathsf{H}}_{n,f,t}, \tag{13}$$

where $\boldsymbol{R}_{n,f,t} = \sum_{t'\le t}\beta^{t-t'}\overline{\boldsymbol{x}}_{f,t'}\overline{\boldsymbol{x}}^{\mathsf{H}}_{f,t'}/r_{n,t'}$ is a spatio-temporal covariance matrix and $\boldsymbol{k}_{n,f,t}$ is the Kalman gain.

C. Update of $\mathcal{W}_t$

If $\mathcal{G}_t$ and $\mathcal{R}_t$ are fixed, the cost function (6) is equivalent to the one used in online AuxIVA:

$$\mathcal{L}(\mathcal{W}_t) = \sum_{n,f}\boldsymbol{w}^{\mathsf{H}}_{n,f,t}\boldsymbol{V}_{n,f,t}\boldsymbol{w}_{n,f,t} - 2\sum_f \log|\det \boldsymbol{W}_{f,t}|, \tag{14}$$

where

$$\boldsymbol{V}_{n,f,t} = \alpha\boldsymbol{V}_{n,f,t-1} + (1-\alpha)\frac{\boldsymbol{z}_{n,f,t}\boldsymbol{z}^{\mathsf{H}}_{n,f,t}}{r_{n,t}} \tag{15}$$

is the recursive form of the spatial covariance matrix at every time frame, and α is the forgetting factor which is different from β in (6) for more practical parameter estimation [13], [18]. In WPE-IVA-OSS, we use the same rank-1 update for separation matrix $\boldsymbol{W}_{f,t}$ frame by frame [8], [9], [19]. Specifically, we let ith separation filter $\boldsymbol{w}_{i,f,t}$ corresponds to the moving speaker and update $\boldsymbol{W}_{f,t}$ using $\boldsymbol{W}_{f,t} \leftarrow \boldsymbol{\eta}_{f,t}\boldsymbol{w}^{\mathsf{H}}_{i,f,t-1}$ where $\boldsymbol{\eta}_{f,t}$ is the coefficients to be optimized. The update rule can be summarized with the following equations [8]:

$$u_{n,f,t} = \boldsymbol{w}^{\mathsf{H}}_{i,f,t-1}\boldsymbol{V}_{n,f,t}\boldsymbol{w}_{n,f,t-1}, \tag{16}$$

$$d_{n,f,t} = \boldsymbol{w}^{\mathsf{H}}_{i,f,t-1}\boldsymbol{V}_{n,f,t}\boldsymbol{w}_{i,f,t-1}, \tag{17}$$

$$\boldsymbol{w}_{n,f,t} = \begin{cases} d^{-1/2}_{n,f,t}\boldsymbol{w}_{i,f,t-1} & \text{if } n = i, \\ \boldsymbol{w}_{n,f,t-1} - \frac{u_{n,f,t}}{d_{n,f,t}}\boldsymbol{w}_{i,f,t-1} & \text{otherwise.} \end{cases} \tag{18}$$

Then, by using the same technique used for original OSS [9], we can obtain $u_{n,f,t}$ and $d_{n,f,t}$ as

$$u_{n,f,t} = \frac{(1-\alpha)\hat{y}_{i,f,t}y^*_{n,f,t}}{r_{n,t}}, \tag{19}$$

$$d_{n,f,t} = \alpha\frac{d_{n,f,t-1}}{d_{i,f,t-1}} + \frac{(1-\alpha)|\hat{y}_{i,f,t}|^2}{r_{n,t}}, \tag{20}$$

where $\hat{y}_{i,f,t} = \boldsymbol{w}^{\mathsf{H}}_{i,f,t-1}\boldsymbol{z}_{n,f,t}$ is the auxiliary estimated signal just for obtaining $u_{n,f,t}$ and $d_{n,f,t}$. By using the above rules into (18), this modification results in smaller computational complexity for updating $\mathcal{W}_t$ from $O(FN^3)$ (for online WPE-IVA with ISS [14]) to $O(FN^2)$ at every time frame t.

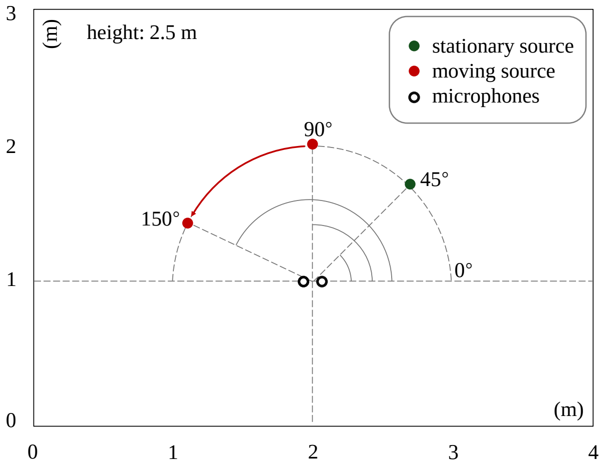

Fig. 1. Simulation layout. The stationary source is fixed at 45° for all 60 s and the moving source is fixed at 90° for the first 20 s. Then it started moving on an arc from 90° to 150° for the next 20 s and finally fixed at 150° for the last 20 s. The distance between the center of the microphone array and the source was 1 m.

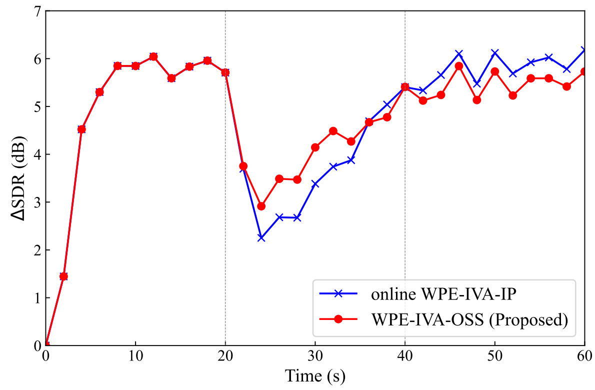

Fig. 2. SDR improvements (ΔSDR) measured every 2 s. Online WPE-GCIVA is used in the first 20 s, and WPE-IVA-OSS is applied in the last 40 s compared with online WPE-IVA-IP and online WPE-IVA-ISS. Since the SDR behavior of WPE-IVA-OSS is theoretically the same as online WPE-IVA-ISS, we omit drawing the latter.

TABLE I
TOTAL RUNTIME OF UPDATING THE SEPARATION FILTERS IN THE LAST 40 S

Method	Runtime
online WPE-IVA-IP [13]	2.4124 s
online WPE-IVA-ISS [14]	1.2775 s
WPE-IVA-OSS (Proposed)	0.4808 s

IV. EXPERIMENT

To evaluate the effectiveness of the proposed WPE-IVA-OSS, several target-source tracking experiments were conducted. We evaluate each method in terms of separation performance and computational efficiency.

A. Experimental Setup

20 mixture signals with a length of 60 s were generated with 2 source signals randomly selected from 2 speakers from the ATR Japanese Speech Database [20]. We concatenated the clean signals so that the signal length becomes 60 s. For mixing simulation, we used Matlab *signal_generator* based on the image method [21]. The position of sound sources and microphones is shown in Fig. 1, where the room boundary is controlled for a 600 ms room reverberation time T_{60}. A 2-element linear microphone array with an inter-element spacing of 2 cm was used. All the signals were sampled at 16 kHz. The STFT was computed with the Hann window of 64 ms (1024 samples) and a window shift length of 16 ms (256 samples). Other parameters α, β, D, L were set at 0.99, 0.96, 2, 10, respectively. The matrices $\boldsymbol{W}_{f,0}, \boldsymbol{G}_{n,f,0}, \boldsymbol{V}_{n,f,0}, \boldsymbol{R}_{n,f,0}^{-1}$ were initialized as $\boldsymbol{I}_M, \mathcal{O}_{ML\times M}, 10^{-5}\boldsymbol{I}_M, \boldsymbol{I}_{ML}$, where $\mathcal{O}_{ML\times M}$ is a zero matrix with the size as $ML \times M$.

Since the output permutation of separated signals is random when using online WPE-IVA [13], [14], we used Geometric Constraint (GC) [22] to detect which separated signal corresponds to the moving source. Specifically, we used online WPE-IVA with GC (online WPE-GCIVA) [23], in which we added the following penalty term derived from GC to the cost function in (6):

$$\mathcal{L}_{\mathrm{GC}}(\boldsymbol{W}_{f,t}) = \sum_{n} \sum_{\phi\in\Theta_n} \lambda_{f,\phi,t} |\boldsymbol{w}_{n,f,t}^{\mathsf{H}} \boldsymbol{d}_{\phi,f}|^2, \tag{21}$$

where $\lambda_{f,\phi,t}$ is a weighting coefficient and $\Theta_n = \{\phi_1, \dots, \phi_N\} \backslash \{\phi_n\}$ represents DOAs for all speakers excluding the nth speaker. $\boldsymbol{d}_{\phi,f}$ is the steering vector pointing to the ϕ direction. By applying and decreasing this penalty term, we can force the nth separation filter $\boldsymbol{w}_{n,f,t}$ to create a spatial null in the $\phi \in \Theta_n$ directions. We set angles for GC $\{\phi_1, \phi_2\} = \{45°, 90°\}$. We set $\lambda_{f,\phi,t} = \lambda_{f,\phi,0}\sigma^t$ where σ is the decreasing factor set to 0.8 and we set $\lambda_{f,\phi,0} = 8,000$. Since the detailed update rule has been discussed in [23] and is not directly related to our main discussion, we skip explaining it. In this paper, we used the above online WPE-GCIVA for the first 20 s. Since the correspondence between separated signals and sources was obtained in the first 20 s, there was no need to use GC for the last 40 s. After that, we used three different methods for the last 40 s: online WPE-IVA using Iterative Projection (online WPE-IVA-IP) [13], online WPE-IVA using Iterative Source Steering (online WPE-IVA-ISS) [14], and the proposed WPE-IVA-OSS. The improvement of signal-to-distortion ratio (ΔSDR) [24] was used as the metric to evaluate the separation performance.

B. Results

The SDR improvements of all the studied methods are shown in Fig. 2. In the period of 20 to 40 s when there was a source moving, WPE-IVA-OSS converged faster than online WPE-IVA-IP. However, after the convergence, WPE-IVA-OSS performed slightly worse than online WPE-IVA-IP. These results might be because WPE-IVA-OSS updates the separation matrix by rank-1 updating instead of the intact updating in online WPE-IVA-IP.

Table I shows the average separation filter updating time for the target-source tracking task for the last 40 s signal over 20 trials. To better illustrate the acceleration brought by the OSS-based optimization, the results in Table I only show the time cost for updating the separation filters instead of both the dereverberation and separation filters. Additionally, we did not consider the runtime of the first 20 s because the same method was used during the first 20 s. The results demonstrate that WPE-IVA-OSS spends a lower runtime than online WPE-IVA-IP [13] and online WPE-IVA-ISS [14].

V. CONCLUSION

This paper proposed a new algorithm named WPE-IVA-OSS for the target-source tracking task in highly reverberant environments. Thanks to the OSS algorithm, the separation matrix updating in WPE-IVA-OSS is simplified from matrix operations into scalar operations, and the computational complexity is reduced from $O(FN^3)$ to $O(FN^2)$. The experimental results show that our proposed method is computationally more efficient while achieving the same or at least comparable separation performance compared to the baseline methods.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] S. Makino, *Audio Source Separation*. Switzerland: Springer, 2018.

[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[4] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.

[5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, 2011, pp. 189–192.

[6] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 7, pp. 1431–1438, 2010.

[7] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc HSCMA*, 2014, pp. 107–111.

[8] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *Proc. APSIPA ASC*, 2022, pp. 749–753.

[9] T. Nakashima, R. Ikeshita, N. Ono, S. Araki, and T. Nakatani, "Fast online source steering algorithm for tracking single moving source using online independent vector analysis," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[10] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.

[11] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sept. 2010.

[12] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[13] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proc IEEE ICASSP*, 2021, pp. 506–510.

[14] K. Mo, X. Wang, Y. Yang, S. Makino, and J. Chen, "Low algorithmic delay implementation of convolutional beamformer for online joint source separation and dereverberation," *arXiv preprint arXiv:2406.09821*, 2024.

[15] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2267–2282, 2020.

[16] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.

[17] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2010.

[18] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Blind and spatially-regularized online joint optimization of source separation, dereverberation, and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1157–1172, 2024.

[19] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.

[20] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

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

[22] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 10, no. 6, pp. 352–362, Dec. 2002.

[23] K. Mo, X. Wang, Y. Yang, T. Ueda, S. Makino, and J. Chen, "On joint dereverberation and source separation with geometrical constraints and iterative source steering," in *Proc. APSIPA ASC*, 2023, pp. 1138–1142.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.