On Joint Dereverberation and Source Separation with Geometrical Constraints and Iterative Source Steering

Kaien Mo*, Xianrui Wang*†, Yichen Yang†, Tetsuya Ueda*,
Shoji Makino* and Jingdong Chen†
* Waseda University, Japan
† Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, China

***Abstract*—In order to improve both the separation performance and the convergence speed, several geometrically constrained independent vector analysis (GC-IVA) algorithms have been developed. Those algorithms are based on the multiplicative transfer function model, which assumes that the analysis window length is longer than the effective part of the room impulse responses. However, this assumption does often not hold in reverberant environments, particularly if the reverberation is strong, which makes the algorithms suffer from significant performance degradation. To circumvent this issue, an algorithm was developed, which jointly optimizes the weighted prediction error (WPE) dereverberation method and GC-IVA (GC-WPE-IVA). While it has demonstrated promising performance, this joint optimization method involves matrix inversion; so it is computationally very expensive. This work attempts to improve the efficiency and stability of GC-WPE-IVA. We develop an iterative source steering (ISS) updating algorithm in the framework of GC-WPE-IVA. The experimental results show that the developed method is computationally much more efficient yet it can achieve comparable separation performance in reverberation environments as compared to GC-WPE-IVA.**

I. Introduction

Speech enhancement is vital in speech communications systems as the source signals are always contaminated by interference, reverberation, and background noise [1]–[4]. When multiple sources are active simultaneously and no prior information is available, blind source separation (BSS) is usually used to separate source signals [1]. Independent component analysis (ICA) [5], one of the most adopted BSS methods, achieves source separation by exploiting the statistical independence among the sources. But it suffers from the problem of permutation. By considering the relationship among different frequency components, independent vector analysis (IVA) [6], [7], which is an extension of ICA, was developed to deal with the frequency permutation problem [8]. To accelerate the convergence and improve the robustness of IVA, the auxiliary function technique was developed [9] and the so-called iteration projection (IP) [9] and iterative source steering (ISS) [10] rules were proposed to optimize the auxiliary function. Although IVA has attracted great attention for its performance, the separated signals are in a random output order, which is known as the global permutation problem. In many applications, such as speech recognition, it is of great importance to match the separated sources and the corresponding speakers.

In real-world applications, the geometry of the array is fixed and the locations of the sources are often known or can be estimated [11]. To make use of such information, geometrically constrained independent vector analysis (GC-IVA) [12]–[16] algorithms were proposed. Recently, GC-IVA with the ISS-based updating rules (GC-IVA-ISS) was introduced, which demonstrated promising performance [17]; but its performance suffers from dramatic degradation if the reverberation is strong, which happens often in practice.

In order to mitigate the impact of reverberation on BSS, the weighted prediction error (WPE) [18], [19], one of the most widely used blind dereverberation methods, is usually implemented before separation [20]. Recently, convolutional beamformer (CBF) [21], which jointly updates the WPE and separation filters, was developed to improve the source extraction performance in strong reverberant environments. The source-wise factorization of the CBF [22] was further developed to reduce the computational complexity. ILRMA-T-ISS [23] is another work that deals with source separation in heavy reverberant environments with low computational cost. To cope with the problem of global permutation, the spatial information of the target source is used in the CBF framework (GC-WPE-IVA) [24], resulting in better performance than the conventional methods. Although GC-WPE-IVA is the state-of-the-art source separation method with dereverberation, it is still computationally costly due to the IP-based updating rules for the separation filters.

This work attempts to improve the efficiency and stability of the algorithm developed in [24]. We develop an ISS updating algorithm in the framework of GC-WPE-IVA, leading to a new algorithm called GC-WPE-IVA-ISS. Experiments are carried out and the results show that GC-WPE-IVA-ISS is computationally much more efficient yet it can achieve better or at least comparable separation performance in reverberation environments as compared to GC-WPE-IVA. Besides, since the spatial information of the sources is used, the global permutation problem is naturally solved with GC-WPE-IVA-ISS.

II. Signal model and problem formulation

Assume that there are N sources and M microphones. The source signals and observed signals can be expressed in a vector form as

$$\boldsymbol{s}_{f,t} = [s_{1,f,t} \quad s_{2,f,t} \quad \cdots \quad s_{N,f,t}]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}, \tag{1}$$

$$\boldsymbol{x}_{f,t} = [x_{1,f,t} \quad x_{2,f,t} \quad \cdots \quad x_{M,f,t}]^{\mathsf{T}} \in \mathbb{C}^{M\times 1}, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes transpose, $s_{n,f,t}$ and $x_{m,f,t}$ represent, respectively, the nth source signal and mth microphone observation signal in the short-time Fourier transformation (STFT) domain, $f = 1 \cdots F$ and $t = 1 \cdots T$ are frequency bin and time frame indexes with F and T being, respectively, the numbers of frequency bins and frames. $\boldsymbol{x}_{f,t}$ can be expressed as

$$\boldsymbol{x}_{f,t} = \sum_{\tau=0}^{L_A-1} \boldsymbol{A}_{f,\tau}\boldsymbol{s}_{f,t-\tau}, \tag{3}$$

where $\boldsymbol{A}_{f,\tau} \in \mathbb{C}^{M\times N}$ is the convolutional mixing matrix at time lag τ, and L_A is the order of time-lagged mixing filters. In this work, we only consider the determined case where $N = M$. Based on the multi-input multi-output (MIMO) CBF formulation [21], the beamformer's output can be expressed as

$$\boldsymbol{y}_{f,t} = \boldsymbol{W}_{f,0}\boldsymbol{x}_{f,t} + \sum_{\tau=D}^{L+D-1} \boldsymbol{W}_{f,\tau}\boldsymbol{x}_{f,t-\tau}, \tag{4}$$

where $\boldsymbol{W}_{f,0} \in \mathbb{C}^{N\times M}$ and $\boldsymbol{W}_{f,\tau} \in \mathbb{C}^{N\times M}$ are the coefficient matrix of CBF, $\boldsymbol{y}_{f,t} = [y_{1,f,t} \quad y_{2,f,t} \quad \cdots \quad y_{N,f,t}]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}$ consists of the N separated signals, D and L are the time delay and the length of the CBF filters, respectively. According to source-wise factorization of CBF [22], the nth separated signal in $\boldsymbol{y}_{f,t}$ can be rewritten into two following form

$$\boldsymbol{z}_{n,f,t} = \boldsymbol{x}_{f,t} - \boldsymbol{G}_{n,f}^{\mathsf{H}}\overline{\boldsymbol{x}}_{f,t}, \tag{5}$$

$$y_{n,f,t} = \boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{z}_{n,f,t}, \tag{6}$$

where $(\cdot)^{\mathsf{H}}$ represents conjugate transpose, $\overline{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}_{f,t-D}^{\mathsf{T}} \quad \boldsymbol{x}_{f,t-D-1}^{\mathsf{T}} \quad \cdots \quad \boldsymbol{x}_{f,t-L-D+1}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML\times 1}$ contains past observed signals, $\boldsymbol{G}_{n,f} \in \mathbb{C}^{ML\times M}$ is the MIMO WPE filter, $\boldsymbol{z}_{n,f,t}$ is the dereverberated signal of nth source and $\boldsymbol{q}_{n,f}$ is the demixing filter. Note that (5) and (6) are strictly equal to (4) when $\boldsymbol{q}_{n,f}^{\mathsf{H}} = \boldsymbol{w}_{n,f,0}$ and $-\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{G}_{n,f}^{\mathsf{H}} = [\boldsymbol{w}_{n,f,D} \quad \boldsymbol{w}_{n,f,D+1} \quad \cdots \quad \boldsymbol{w}_{n,f,L+D-1}]$, where $\boldsymbol{w}_{n,f,\tau}$ is the nth row of $\boldsymbol{W}_{f,\tau}$.

To exploit the dependency between different frequency bins and thereby avoid the frequency permutation problem as IVA, CBF assumes that the sources follow a multivariate Gaussian distribution with time dependent variance $r_{n,t} = \sum_f |y_{n,f,t}|^2/F$. Therefore, the negative log-likelihood cost function can be derived as [22]

$$\mathcal{L} = -2\sum_f \log|\det \boldsymbol{Q}_f| + \frac{1}{T}\sum_{n,f,t}\left(\log r_{n,t} + \frac{|y_{n,f,t}|^2}{r_{n,t}}\right), \tag{7}$$

where $\boldsymbol{Q}_f = [\boldsymbol{q}_{1,f} \quad \boldsymbol{q}_{2,f} \quad \cdots \quad \boldsymbol{q}_{N,f}]^{\mathsf{H}}$ is the demixing matrix.

III. Proposed method

A. A Probabilistic model

In order to deal with the global permutation problem in strong reverberation conditions, we introduce the geometrical constraints [25] to the conventional CBF, which is expressed as

$$\mathcal{L}_{\mathrm{GC}} = \sum_{n,\varphi,f} \lambda_{n,\varphi}|\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2, \tag{8}$$

where $\lambda_{n,\varphi}$ is a non-negative weighting coefficient, $\boldsymbol{d}_{\varphi,f}$ is the steering vector along the direction φ and $c_{n,\varphi}$ is a non-negative-valued constraint. Note that if $c_{n,\varphi} = 1$, optimizing (8) gives the conventional delay-and-sum beamformer, which is steered to φ to extract the target signal whereas setting $c_{n,\varphi}$ to a small value will create a spatial null to suppress interference incident from the direction φ [14]. Combining (8) and (7) gives the cost function for the proposed method, i.e.,

$$\begin{aligned}\mathcal{L}(\Theta) = &-2\sum_f \log|\det \boldsymbol{Q}_f| + \frac{1}{T}\sum_{n,f,t}\left(\log r_{n,t} + \frac{|y_{n,f,t}|^2}{r_{n,t}}\right)\\ &+ \sum_{n,\varphi,f}\lambda_{n,\varphi}|\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2,\end{aligned} \tag{9}$$

where $\Theta = \{\Theta_{\boldsymbol{G}}, \Theta_{\boldsymbol{Q}}, \Theta_r\}$ is the parameter set to be estimated, $\Theta_{\boldsymbol{G}} = \{\boldsymbol{G}_{n,f}\}$, $\Theta_{\boldsymbol{Q}} = \{\boldsymbol{Q}_f\}$, and $\Theta_r = \{r_{n,t}\}$.

B. Optimization algorithm

According to the coordinate ascent method [20], each parameter in (9) can be updated iteratively by fixing the others until convergence.

1) Update of $\Theta_{\boldsymbol{G}}$: To update the WPE filter $\boldsymbol{G}_{n,f}$, the cost function (9) can be rewritten by fixing other parameters and ignoring the constant terms as

$$\mathcal{L}(\Theta_{\boldsymbol{G}}) = \frac{1}{T}\sum_{n,t,f}|\boldsymbol{q}_{n,f}^{\mathsf{H}}(\boldsymbol{x}_{f,t} - \boldsymbol{G}_{n,f}^{\mathsf{H}}\overline{\boldsymbol{x}}_{f,t})|^2/r_{n,t}. \tag{10}$$

Note that (10) is a quadratic function with respect to $\Theta_{\boldsymbol{G}}$. So, the parameter that minimizes (10) is [22]

$$\boldsymbol{G}_{n,f} \leftarrow \boldsymbol{R}_{n,f}^{-1}\boldsymbol{P}_{n,f}, \tag{11}$$

where

$$\boldsymbol{R}_{n,f} = \sum_t \frac{\overline{\boldsymbol{x}}_{f,t}\overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{r_{n,t}}, \tag{12}$$

$$\boldsymbol{P}_{n,f} = \sum_t \frac{\overline{\boldsymbol{x}}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}}{r_{n,t}}, \tag{13}$$

are spatio-temporal covariance matrices.

2) Update of $\Theta_{\boldsymbol{Q}}$: By fixing $\Theta_{\boldsymbol{G}}$ and Θ_r, the cost function (9) can be simplified to

$$\begin{aligned}\mathcal{L}(\Theta_{\boldsymbol{Q}}) = &-2\sum_f \log|\det \boldsymbol{Q}_f| + \sum_{n,f} \boldsymbol{q}_{n,f}^{\mathsf{H}} \boldsymbol{U}_{n,f} \boldsymbol{q}_{n,f} \\ &+ \sum_{n,\varphi,f} \lambda_{n,\varphi} |\boldsymbol{q}_{n,f}^{\mathsf{H}} \boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2,\end{aligned} \tag{14}$$

where

$$\boldsymbol{U}_{n,f} = \frac{1}{T}\sum_t \frac{\boldsymbol{z}_{n,f,t}\boldsymbol{z}_{n,f,t}^{\mathsf{H}}}{r_{n,t}} \tag{15}$$

is the weighted covariance matrix of $\boldsymbol{z}_{n,f,t}$. To accelerate the convergence of updating $\Theta_{\boldsymbol{Q}}$, instead of using the IP-based method that updates each row of $\boldsymbol{Q}_f$ iteratively [14] (which requires N times matrix inversion), the proposed method adopts the ISS updating rule [10] in which the whole filter is updated with a rank-1 matrix as

$$\boldsymbol{Q}_f \leftarrow \boldsymbol{Q}_f - \boldsymbol{v}_{k,f}\boldsymbol{q}_{k,f}^{\mathsf{H}}, \tag{16}$$

where $\boldsymbol{v}_{k,f} \in \mathbb{C}^{N\times 1}$ is a vector to be estimated. The update of $\boldsymbol{Q}_f$ will be repeated for $k = 1, ..., N$. Substituting (16) into the cost function (14) gives

$$\begin{aligned}\mathcal{L}_{\mathrm{ISS}}(\boldsymbol{v}_{k,f}) = &-2\sum_f \log|\det(\boldsymbol{Q}_f - \boldsymbol{v}_{k,f}\boldsymbol{q}_{k,f}^{\mathsf{H}})| \\ &+ \sum_{n,f} (\boldsymbol{q}_{n,f} - v_{n,k,f}^*\boldsymbol{q}_{k,f})^{\mathsf{H}} \boldsymbol{U}_{n,f} (\boldsymbol{q}_{n,f} - v_{n,k,f}^*\boldsymbol{q}_{k,f}) \\ &+ \sum_{n,\varphi,f} \lambda_{n,\varphi} |(\boldsymbol{q}_{n,f} - v_{n,k,f}^*\boldsymbol{q}_{k,f})^{\mathsf{H}}\boldsymbol{d}_{\varphi,f} - c_{n,\varphi}|^2,\end{aligned} \tag{17}$$

where $v_{n,k,f}$ is the nth element of $\boldsymbol{v}_{k,f}$ and $(\cdot)^*$ denotes complex conjugate. Solving the equation $\partial\mathcal{L}_{\mathrm{ISS}}(\boldsymbol{v}_{k,f})/\partial v_{n,k,f}^* = 0$, which is inspired by the work in [17], we obtain the update rules for $v_{n,k,f}$, which is divided into the following two cases:

- $n \neq k$: In this case, the update rule is

$$v_{n,k,f} = \frac{\boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{U}_{n,f}\boldsymbol{q}_{k,f} + \sum \lambda_{n,\varphi}(g_{n,\varphi,f} - c_{n,\varphi})g_{k,\varphi,f}^*}{\boldsymbol{q}_{k,f}^{\mathsf{H}}\boldsymbol{U}_{n,f}\boldsymbol{q}_{k,f} + \sum \lambda_{n,\varphi}|g_{k,\varphi,f}|^2}; \tag{18}$$

- $n = k$: In this case, $v_{k,k,f}$ is updated according to

$$v_{k,k,f} = \begin{cases} 1 - (\alpha_{k,f})^{-1/2} & \text{if } \beta_{k,f} = 0, \\ 1 - \beta_{k,f}^* \frac{|\beta_{k,f}| + \sqrt{|\beta_{k,f}^2| + 4\alpha_{k,f}}}{2\alpha_{k,f}|\beta_{k,f}|} & \text{otherwise,} \end{cases} \tag{19}$$

where

$$g_{n,\varphi,f} = \boldsymbol{q}_{n,f}^{\mathsf{H}}\boldsymbol{d}_{\varphi,f}, \tag{20}$$

$$\alpha_{k,f} = \boldsymbol{q}_{k,f}^{\mathsf{H}}\boldsymbol{U}_{k,f}\boldsymbol{q}_{k,f} + \sum_{\varphi} \lambda_{k,\varphi}|g_{k,\varphi,f}|^2, \tag{21}$$

$$\beta_{k,f} = \sum_{\varphi} \lambda_{k,\varphi} c_{k,\varphi} g_{k,\varphi,f}. \tag{22}$$

Once $\boldsymbol{v}_{k,f}$ is obtained, $\boldsymbol{Q}_f$ can be updated through (16).

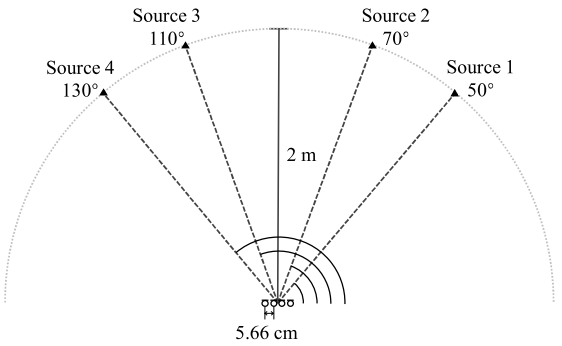

Fig. 1. Simulation layout.

3) Update of Θ_r: After updating the dereverberation ($\boldsymbol{G}_{n,f}$) and separating ($\boldsymbol{Q}_f$) filters, new separated signals $\boldsymbol{y}_{f,t}$ can be obtained through (5) and (6). Then, the time-varying variance $r_{n,t}$ is updated as

$$r_{n,t} \leftarrow \frac{1}{F}\sum_f |y_{n,f,t}|^2. \tag{23}$$

IV. Experiment

A. Experimental setup

The observation signals are generated by convolving the speech signals from the TIMIT database [26] with the room impulse responses (RIRs) from the RWCP dataset [27]. For every mixed signal, 4 speech segments, which are arbitrarily selected from different speakers, are concatenated to form 10 s long clean speech signals. The room reverberation time T_{60} is selected to 300 ms and 600 ms. The white Gaussian noise is added to control the signal-to-noise ratio (SNR) to 30 dB. The layout of sources and microphones is illustrated in Fig. 1, where a 4-element uniform linear microphone array (ULA) with an inter-element spacing of 5.66 cm is used. There are four sources in the sound field and the DOAs of the four sources are 50°, 70°, 110°, and 130°, respectively. The distance between the ULA center and the sources is 2 m. To measure the separation performance, twenty-five Monte Carlo simulations are carried out. All observed signals are sampled at 16 kHz and the STFT is conducted with the Hann window of 64 ms (1024 samples) and a window shift of 16 ms (256 samples).

We define two matrices, i.e., $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \ \ \boldsymbol{\lambda}_2 \ \cdots \ \boldsymbol{\lambda}_N]^{\mathsf{T}}$ and $\boldsymbol{C} = [\boldsymbol{c}_1 \ \ \boldsymbol{c}_2 \ \cdots \ \boldsymbol{c}_N]^{\mathsf{T}}$, with the parameters in (8) where $\boldsymbol{\lambda}_n = [\lambda_{n,\varphi_1} \ \ \lambda_{n,\varphi_2} \ \cdots \ \lambda_{n,\varphi_N}]^{\mathsf{T}} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{c}_n = [c_{n,\varphi_1} \ \ c_{n,\varphi_2} \ \cdots \ c_{n,\varphi_N}]^{\mathsf{T}} \in \mathbb{R}^{N\times 1}$ and $\Phi = \{\varphi_1, \varphi_2, \cdots, \varphi_N\}$ is the set of DOAs. According to the work in [17], we set $\boldsymbol{C} = \boldsymbol{E}$ so the output order is the same as the order of the input DOA and we use *NULL* constraints $\boldsymbol{\Lambda} = \Lambda(\boldsymbol{J} - \boldsymbol{E})$ where Λ is a non-negative number, $\boldsymbol{J}$ is an all-one matrix and $\boldsymbol{E}$ is the identity matrix. We initialize Λ as 8000 and decrease it over iterations according to [13], i.e.,

$$\Lambda_i = \Lambda_0\alpha^i, \tag{24}$$

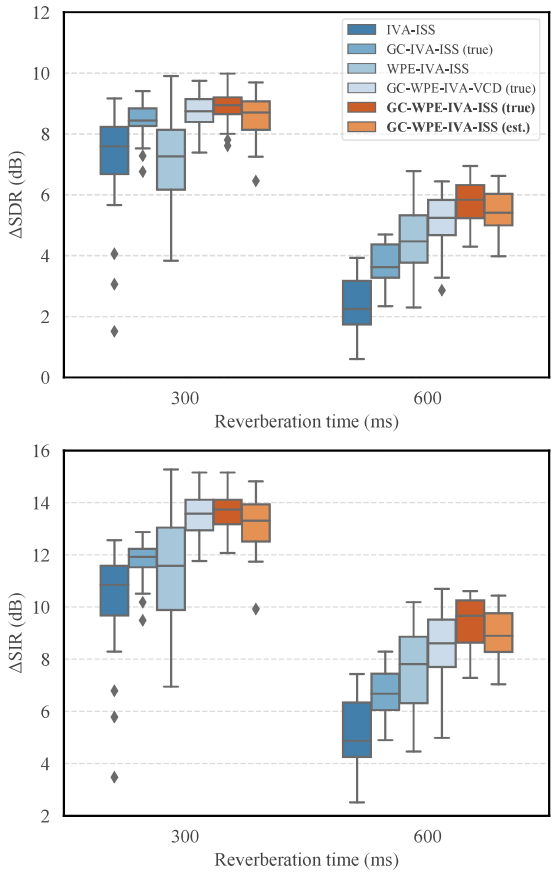

Fig. 2. Performance of the studied methods under different reverberation conditions.

where i represents the current number of iterations, Λ_0 is the initial value of Λ and the factor α is set to 0.8. For the WPE filter, the time delay D and filter length L are set to 2 and 10, respectively. The results for all the methods are obtained with 50 iterations and the WPE filter is updated every 10 iterations so both the convergence speed and computational cost are jointly considered. All the algorithms are implemented on a workstation powered by Intel Xeon E3-1505M. The improvement of signal-to-distortion ratio (ΔSDR) and signal-to-interference ratio (ΔSIR) [28] are used as the metrics to evaluate the separation performance.

The separation performance of the proposed algorithm (denoted as GC-WPE-IVA-ISS) is compared with IVA-ISS [10], GC-IVA-ISS [17], the joint optimization algorithm of WPE and IVA [22] with the ISS update method (WPE-IVA-ISS) and the geometrically constrained WPE-IVA with the vectorwise coordinate descent update method (GC-WPE-IVA-VCD) [24].

B. Experimental results

The SDR and SIR improvements of all the studied methods under different reverberation conditions are shown in Fig. 2. For the proposed GC-WPE-IVA-ISS, both the ground truth DOAs $\Phi_{\text{True}} = \{50°, 70°, 110°, 130°\}$ (denoted as true) and the DOAs with observation error $\Phi_{\text{True}} + \varepsilon$ (est.) are considered,

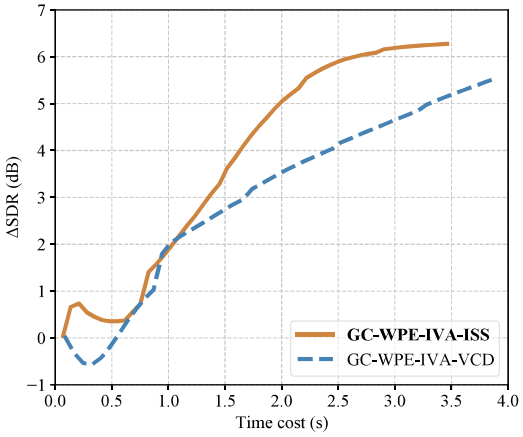

Fig. 3. Convergence speed of GC-WPE-IVA-ISS and GC-WPE-IVA-VCD.

where the error ε is computer generated random number with a uniform distribution in $[-15°, 15°]$ for every source. As seen from Fig 2, in the light reverberation condition with T_{60} = 300 ms, the ΔSDR of GC-IVA-ISS is close to that of the proposed method even without dereverberation. However, as the reverberation becomes stronger (with T_{60} = 600 ms), the separation performance of GC-IVA-ISS decreases significantly. In comparison, with the joint optimization of dereverberation filters and separation filters and the ISS-based updating rules, the proposed method generates a much better ΔSDR and ΔSIR than the compared conventional methods. In the presence of DOA errors, which is inevitable in practical applications with multi-sources and reverberation, the separation performance of the developed method decreases slightly. But its performance is still better than, or at least comparable to the compared conventional methods.

Figure 3 plots the convergence curves of GC-WPE-IVA-ISS and GC-WPE-IVA-VCD with ground true DOA. Note that the updating of the WPE filter is time-consuming. To better illustrate the acceleration of the convergence speed brought by the ISS-based optimization, the results in Fig. 3 show only the time cost for updating the separation filters. The results demonstrate that, due to the efficiency and stability of ISS, the developed method gets convergence with a much lower time cost than GC-WPE-IVA-VCD.

The accuracy of the output channel order of the proposed GC-WPE-IVA-ISS as well as GC-IVA-ISS, GC-WPE-IVA-VCD is listed in Table I. Note that the results of IVA-ISS and WPE-IVA-ISS are not presented here since their output orders are random. From Table I, it is seen that in the light reverberation condition, all the studied methods are able to produce accurate output order. But as the reverberation time becomes longer, the output accuracy of the conventional methods is affected while the proposed method can still guarantee the accuracy of the output order.

V. Conclusion

To make efficient use of the *a priori* spatial information of the sound sources, thereby improving the source separation

TABLE I
THE ACCURACY OF THE OUTPUT CHANNEL ORDER OF THE STUDIED METHODS UNDER DIFFERENT REVERBERATION CONDITIONS.

T_{60}	GC-IVA-ISS	GC-WPE-IVA-VCD	GC-WPE-IVA-ISS
300 ms	100%	100%	100%
600 ms	96%	96%	100%

performance in reverberant environments, we presented in this work a geometrically constrained algorithm, which jointly optimizes source separation and dereverberation filters with the ISS-based optimization. The experimental results show that the developed method is computationally much more efficient yet it can achieve better or at least comparable separation performance in reverberation environments as compared to the studied baseline algorithms. It is also robust to DOA estimation error. Another merit of the developed algorithm is that it is able to deal with the global permutation problem inherently.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. Makino, *Audio Source Separation*. Switzerland: Springer, 2018.
[2] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Singapore: Wiley-IEEE Press., 2018.
[3] G. Huang, J. Benesty, and J. Chen, "Fundamental approaches to robust differential beamforming with high directivity factors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3074–3088, 2022.
[4] X. Wang, N. Pan, J. Benesty, and J. Chen, "On multiple input/binaural output antiphasic speaker signal extraction," in *Proc IEEE ICASSP*, 2023, pp. 1–5.
[5] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, 1st ed. Oxford, UK: Academic press/Elsevier, 2010.
[6] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.
[7] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.
[8] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Proc. IEEE ISCAS*, 2007, pp. 3247–3250.
[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, 2011, pp. 189–192.
[10] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.
[11] 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, Dec. 2020.
[12] A. H. Khan, M. Taseska, and E. A. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.
[13] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. IEEE ICASSP*, 2018, pp. 746–750.
[14] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 846–850.
[15] 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. IEEE IWAENC*, 2022, pp. 1–5.
[16] 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, in press.
[17] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering," in *Proc. EUSIPCO*, 2022, pp. 757–761.
[18] 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, Sep. 2010.
[19] 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.
[20] 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.
[21] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2267–2282, Jul. 2020.
[22] 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.
[23] T. Nakashima, R. Scheibler, M. Togami, and N. Ono, "Joint dereverberation and separation with iterative source steering," in *Proc. IEEE ICASSP*, 2021, pp. 216–220.
[24] Y. Yang, X. Wang, A. Brendel, W. Zhang, W. Kellermann, and J. Chen, "Geometrically constrained source extraction and dereverberation based on joint optimization," in *Proc. EUSIPCO*, 2023, in press.
[25] 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.
[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, Feb. 1993.
[27] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," pp. 965–968, 2000.
[28] 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.