Geometrically Constrained Blind Moving Source Extraction based on Constant Separation Vector and Auxiliary Function Technique

Ruifeng Zhang, Tetsuya Ueda and Shoji Makino
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan

***Abstract*—This paper focuses on the permutation problem in Blind Source Extraction (BSE) by employing beamforming-based Geometrical Constraints (GC). Specifically, we focus on the recently proposed auxiliary function-based Independent Vector Extraction (Aux-IVE) with Constant Separation Vector (CSV) mixing model. Building upon this mixing model, we provide the source signals' spatial information, utilizing GC based on beamforming. We facilitate the extraction of moving target source and improve the extraction performance. Furthermore, we discuss the effect of the number of constraints on the extraction performance. Experimental results confirmed that our proposed method, GC-CSV-Aux-IVE, exhibits superior extraction performance and achieves nearly 100% accuracy in extracting the moving target source.**

I. Introduction

Recently, the applications of Blind Source Separation (BSS) [1]–[3] span across a wide range of fields, contributing to advancements in speech recognition, audio processing, and speech enhancement. BSS algorithms can separate individual speech signals in scenarios where multiple speeches overlap, improving speech recognition accuracy and estimating the speaker's location.

In the field of BSS, Independent Component Analysis (ICA) [4], [5] and Independent Vector Analysis (IVA) [6], [7] have emerged as prominent techniques for separating mixed signals into their constituent sources. Auxiliary function-based Independent Vector Analysis (Aux-IVA) [8], [9] is an extension of IVA that incorporates auxiliary functions to improve the separation performance.

Moreover, in environments with noisy or degraded speech, Blind Source Extraction (BSE) methods [10]–[12] offer the capability to enhance the target speech by extracting it from the mixture, leading to improved speech intelligibility and quality. BSE techniques have emerged as effective methods for extracting specific source signals from mixtures. By utilizing various algorithms and signal processing methods [13], BSE focuses on extracting the Source of Interest (SOI) while suppressing or removing interference from other sources in the mixture.

In recent years, the Constant Separating Vector (CSV) model [14], [15] proposed in the BSE domain has demonstrated significant advancements in the moving source extraction. Based on the CSV model, [16] proposed CSV-Aux-IVE, which estimates a time-varying mixing vector and a time-invariant filter and utilizes an auxiliary function technique to enhance the extraction performance and the convergence speed of Independent Vector Extraction (IVE).

However, since the permutation problem exists, it is often difficult to determine which source will be extracted in the field of BSE. An important consideration is ensuring that the extracted signal corresponds to SOI. In [14], they addressed this problem by employing the pilot signal which is a signal that is mutually dependent with the corresponding source signal. This method uses prior information about the sound source to resolve the permutation ambiguity.

On the other hand, we propose to solve this ambiguity using Geometric Constraints (GC) [17]. Moreover, we aim to give several constraints for the moving speaker. By employing GC method [18]–[21], we can enhance the accuracy of signal extraction by only requiring approximate location information of the target signal and sensors positions. Consequently, it can be widely applied in scenarios where the signal's moving range is known, such as conferences and speech monitoring. GC method is based on beamforming, which utilizes spatial information and sensor positions to optimize BSS problems. It exploits spatial information to guide the separation matrix to obtain the signal in the desired direction. This method optimizes the constraint weights to ensure that extracted source signal keeps unit energy. Therefore, it is our objective to address the permutation problem and improve the performance of moving target source extraction by applying the GC method.

II. Problem Formulation

A. Blind Source Extraction

Considering the situation in BSE, we assume there are M microphones in the room. After the Short Time Fourier Transform (STFT), we obtain microphone signals $\boldsymbol{x}_{f,l}$ and source signals $\boldsymbol{s}_{f,l}$, where f and l represent the frequency bin and time frame index. Our main task is to extract SOI $s_{f,l}$ from M microphones, and $\boldsymbol{z}_{f,l}$ represent the remaining background signals. Generally, we can assume that $s_{f,l}$ corresponds to the

979-8-3503-0067-3/23/$31.00 ©2023 IEEE

first source of $\boldsymbol{s}_{f,l}$ without loss of generality. So the relationship of microphone signals $\boldsymbol{x}_{f,l} = [x_{f,l,1}, \ldots, x_{f,l,M}]^{\mathsf{T}} \in \mathbb{C}^M$ and source signals $\boldsymbol{s}_{f,l} = [s_{f,l}, \boldsymbol{z}_{f,l}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^N$ where $\boldsymbol{z}_{f,l} = [z_{f,l,1}, \ldots, z_{f,l,N-1}]$ can be written as

$$\boldsymbol{x}_{f,l} = \boldsymbol{A}_{f,l}\boldsymbol{s}_{f,l} = \boldsymbol{A}_{f,l}\begin{bmatrix} s_{f,l} \\ \boldsymbol{z}_{f,l} \end{bmatrix}, \tag{1}$$

where $\boldsymbol{A}_{f,l}$ stands for $M \times N$ mixing matrix and $(\cdot)^{\mathsf{T}}$ denotes the transpose. If M is equal to N, the inverse matrix of $\boldsymbol{A}_{f,l}$ exists and is called separation matrix $\boldsymbol{W}_{f,l} \in \mathbb{C}^{M\times M}$, so (1) can be written as:

$$\boldsymbol{s}_{f,l} = \boldsymbol{W}_{f,l}\boldsymbol{x}_{f,l} = \begin{bmatrix} \boldsymbol{w}_f^{\mathsf{H}} \\ \boldsymbol{B}_{f,l} \end{bmatrix}\boldsymbol{x}_{f,l}. \tag{2}$$

Here $(\cdot)^{\mathsf{H}}$ denotes Hermit transpose. Then we divide mixing matrix as $\boldsymbol{A}_{f,l} = [\boldsymbol{a}_{f,l}\ \boldsymbol{D}_{f,l}]$ which implies that $\boldsymbol{a}_{f,l}$ is the first column of $\boldsymbol{A}_{f,l}$ while $\boldsymbol{D}_{f,l}$ is the remaining part. Similarly, $\boldsymbol{w}_f^{\mathsf{H}}$ refers to the first row of the matrix, and $\boldsymbol{B}_{f,l}$ refers to the rest. Then the mixing and separation model (1) and (2) can be written as:

$$\boldsymbol{x}_{f,l} = \boldsymbol{a}_{f,l}s_{f,l} + \boldsymbol{D}_{f,l}\boldsymbol{z}_{f,l}, \tag{3}$$

and

$$\begin{bmatrix} s_{f,l} \\ \boldsymbol{z}_{f,l} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{x}_{f,l} \\ \boldsymbol{B}_{f,l}\boldsymbol{x}_{f,l} \end{bmatrix}. \tag{4}$$

B. CSV Mixing Model

In the CSV model, the signal after STFT is divided into T blocks, and the length of each block is defined as L_b. At this time, the index of each block is $t \in \{1, \ldots, T\}$. Thus, the microphone signal is divided into several blocks. We consider the signal in each block approximately stationary. Accordingly, the expressions of the corresponding mixing matrix $\boldsymbol{A}_{f,t}$ in (1) and separation matrix $\boldsymbol{W}_{f,t}$ in (2) will be rewrite as [12]

$$\boldsymbol{A}_{f,t} = [\boldsymbol{a}_{f,t}, \boldsymbol{D}_{f,t}] = \begin{bmatrix} \gamma_{f,t} & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & \frac{1}{\gamma_{f,t}}(\boldsymbol{g}_{f,t}\boldsymbol{h}_f^{\mathsf{H}} - \boldsymbol{I}_{M-1}) \end{bmatrix}, \tag{5}$$

$$\boldsymbol{W}_{f,t} = \begin{bmatrix} \boldsymbol{w}_f^{\mathsf{H}} \\ \boldsymbol{B}_{f,t} \end{bmatrix} = \begin{bmatrix} \beta_f^* & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & -\gamma_{f,t}\boldsymbol{I}_{M-1} \end{bmatrix}. \tag{6}$$

Here $(\cdot)^*$ denotes the complex conjugate, and $\boldsymbol{I}_M$ denotes to $M \times M$ identity matrix. In the remaining sections of this paper, we omit the frame index $l = \{1, \ldots, L\}$. Given that the mixing matrix and separation matrix are inverse to each other, we can obtain $\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{a}_{f,t} = 1$, which is called distortionless constraint. Similarly, we can get $\boldsymbol{B}_{f,t}\boldsymbol{a}_{f,t} = 0$, where $\boldsymbol{B}_{f,t}$ represents the block matrix. Using the assumption above, we can rewrite (1) and (2) as follows:

$$\boldsymbol{x}_{f,t} = \begin{bmatrix} \gamma_{f,t} & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & \frac{1}{\gamma_{f,t}}(\boldsymbol{g}_{f,t}\boldsymbol{h}_f^{\mathsf{H}} - \boldsymbol{I}_{M-1}) \end{bmatrix}\begin{bmatrix} s_{f,t} \\ \boldsymbol{z}_{f,t} \end{bmatrix}, \tag{7}$$

and

$$\begin{bmatrix} s_{f,t} \\ \boldsymbol{z}_{f,t} \end{bmatrix} = \begin{bmatrix} \beta_f^* & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & -\gamma_{f,t}\boldsymbol{I}_{M-1} \end{bmatrix}\boldsymbol{x}_{f,t}. \tag{8}$$

We assume that $\boldsymbol{s}_{f,t}$ and $\boldsymbol{z}_{f,t}$ are mutually independent across all time frequency bins. Based on this assumption, we can express the joint pdf of them as follows:

$$p(\{\boldsymbol{s}_t, \boldsymbol{z}_{f,t}\}_{f,t}) = \prod_t p(\boldsymbol{s}_t)\prod_{f,t} p(\boldsymbol{z}_{f,t}), \tag{9}$$

here $p(\boldsymbol{s}_t)$ denote the joint pdf of the SOI vector component $\boldsymbol{s}_t = [s_{1,t}, \ldots, s_{F,t}] \in \mathbb{C}^F$ and $p(\boldsymbol{z}_{f,t})$ denote the pdf of $\boldsymbol{z}_{f,t}$, respectively.

We consider the block-dependent variance of SOI, which can vary from block to block. We introduce pdf, denoted as $p(\boldsymbol{s}_t)$, that captures this block-dependent variance and accurately represents the characteristics of the source signal. Specifically, the pdf $p(\boldsymbol{s}_t)$ is defined as follows:

$$p(\boldsymbol{s}_t) = h\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right)\left(\prod_{f=1}^{F}\hat{\sigma}_{f,t}\right)^{-2}, \tag{10}$$

where $\hat{\sigma}_{f,t} = \sqrt{\boldsymbol{w}_f^{\mathsf{H}}\hat{\boldsymbol{C}}_{f,t}\boldsymbol{w}_f}$ and $\hat{\boldsymbol{C}}_{f,t} = \mathbb{E}\left[\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]$. $h(\cdot)$ is a pdf corresponding to a normalized non-Gaussian random variable:

$$h\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right) = C\exp(-G_R(r_t)), \tag{11}$$

where C is a coefficient and G_R is a continuous and differentiable function of a real variable r satisfying that $\psi(r) = \frac{G_R'(r)}{r}$ is continuous and monotonically decreasing in $r \geq 0$. The background noise $\boldsymbol{z}_{f,t}$ is assumed to follow a circular Gaussian distribution with zero mean and a variance matrix $\hat{\boldsymbol{C}}_{\boldsymbol{z}_{f,t}} = \mathbb{E}\left[\boldsymbol{z}_{f,t}\boldsymbol{z}_{f,t}^{\mathsf{H}}\right]$.

After applying the auxiliary function method described in [8], we can obtain

$$\hat{\mathbb{E}}\left[-\log h\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right)\right] \leq \frac{1}{2}\frac{1}{\hat{\sigma}_{f,t}^2}\boldsymbol{w}_f^{\mathsf{H}}\mathbb{E}\left[\psi(r_t)\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]\boldsymbol{w}_f + R_t, \tag{12}$$

where r_t is an auxiliary variable and R_t depends purely on r_t. Based on the above assumptions, we proceed to apply the auxiliary function method for updating the objective function. Consequently, we obtain the auxiliary function for CSV as follows:

$$\mathcal{L}_{\text{Aux}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{f=1}^{F}\left\{\frac{1}{2}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2} + \log\hat{\sigma}_{f,t}^2 + \mathbb{E}\left[\boldsymbol{z}_{f,t}^{\mathsf{H}}\boldsymbol{C}_{\boldsymbol{z}_{f,t}}^{-1}\boldsymbol{z}_{f,t}\right] - (M-2)\log|\gamma_{f,t}|^2\right\} + R_t, \tag{13}$$

where

$$\boldsymbol{V}_{f,t} = \mathbb{E}\left[\psi(r_t)\boldsymbol{x}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}\right]. \tag{14}$$

To minimize the auxiliary function with respect to the normal variables, we calculate the partial derivative with respect

to $\boldsymbol{w}_f$. In order to minimize the auxiliary function (13), we derivate it and get

$$\frac{\partial \mathcal{L}_{\text{Aux}}}{\partial \boldsymbol{w}_f^*} = \frac{1}{2T}\sum_{t=1}^{T}\left\{\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\boldsymbol{w}_f - \frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right\}, \quad (15)$$

where we used the same technique as the previous researches [12], [16] to replace the derivative of the third and fourth term in (13) as $\sum_{f=1}^{F}\boldsymbol{a}_{f,t}$. Using conventional technique [16], we get an update rule for the separation filter $\boldsymbol{w}_f$ by taking linearized solution by fixing $\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}$ and $\hat{\sigma}_{f,t}^2$ as constant terms. Then the update rule of $\boldsymbol{w}_f$ can be written as

$$\boldsymbol{w}_f = \left(\sum_{t=1}^{T}\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1}\sum_{t=1}^{T}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}. \quad (16)$$

III. Proposed Method

So far, the CSV-Aux-IVE in Section II-B still encounters the permutation problem. In other words, without prior information, it is not determined whether the separate output corresponds to the SOI or Interference (IR). Therefore, we introduce GC [17] to ensure the extraction of the SOI. Then, let us consider GC that restrict the far-field response of the seperation filter$\boldsymbol{w}_f$ at the direction θ, which is described as

$$\mathcal{L}_{\text{GC}} = \lambda_{\text{GC}}\sum_{f=1}^{F}\sum_{\theta\in\Theta}|\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{d}_{f\theta} - 1|^2, \quad (17)$$

where $\boldsymbol{d}_{f\theta}$ is the steering vector pointing to the direction θ, λ_{GC} is a parameter weighting the importance of GC in the objective function, and we can preserve the SOI by designing this constraint when θ is the DOA (Direction of Arrival) of the target, which we call Unit Response (UR) constraints. The set Θ represents the collection of angles that will be taken into account.

The target source we aim to extract is moving, while the filter $\boldsymbol{w}_f$ remains time-invariant. Thus, a challenge lies in how to provide the filter with information regarding the movement of SOI. In this study, we make the assumption that the range of movement for the signals is known. We divide this range into p equally spaced angles and subsequently apply UR constraints to the signals at each angle.

We combine two loss functions (13) and (17) to apply UR constraints to SOI by summing them together to create a composite objective function, denoted as

$$\mathcal{Q}_{\text{CSV+GC}} = \mathcal{L}_{\text{Aux}} + \mathcal{L}_{\text{GC}}. \quad (18)$$

Then we calculate the partial derivative of the objective function. Thus the resulting derivative is as follows:

$$\frac{\partial \mathcal{Q}_{\text{CSV+GC}}}{\partial \boldsymbol{w}_f^*} = \frac{1}{2T}\sum_{t=1}^{T}\left\{\frac{\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2} - \frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right\} + \lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta}\boldsymbol{d}_{f\theta}^{\mathsf{H}}\boldsymbol{w}_f - \lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta}. \quad (19)$$

Subsequently, we take the linearized solution of $\boldsymbol{w}_f$:

$$\boldsymbol{w}_f = \left(\lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta}\boldsymbol{d}_{f\theta}^{\mathsf{H}} + \frac{1}{2T}\sum_{t=1}^{T}\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1}\left(\lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{d}_{f\theta} + \frac{1}{2T}\sum_{t=1}^{T}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right). \quad (20)$$

IV. Experiment

To evaluate the effectiveness of GC-CSV-Aux-IVE, we conducted a speech separation experiments. The evaluation criteria included extraction performance and order of the output signal, which we calculate its average as its accuracy. We compared our proposed method, GC-CSV-Aux-IVE, with the existing CSV-Aux-IVE method without being supervised.

A. Setup

In this experiment, we utilized a subset of the ATR Japanese Speech Database [22], consisting of recordings from three male speakers and three female speakers. We randomly selected two distinct speeches from this dataset to create mixtures. The indoor impulse response is generated through image method [23], with the observation signal generated using signal generator[1]. The layout of the simulated room is illustrated in Fig. 1. In total, we generated 100 pairs of original signals and mixed signals. The length of each signal is about 20 seconds. Among these, the first 50 signals had the SOI moving from 90° to 150°, while the remaining 50 signals had the SOI moving from 150° to 90°. Here SOI was set to move at a speed of 3° per second, while the IR is fixed at 45°. We have 6 microphones, 1 moving target source, and 1 fixed interference source.

To make this experiment more convincing, we added a white noise signal to the mixing signal. We adjusted the SNR to input SNR $= 10\log\frac{\lambda_1}{\lambda_2}$ as specified value, where λ_1 corresponds to the sample variance of mixing signal and λ_2 corresponds to that of noises. The other experiment conditions are shown in Table I.

In this experiment, we assumed that the movement range of SOI is known. We divided this movement range into p equal parts to set the collection of angles Θ. Then we applied UR constraints to SOI at each angle. Here is an example of how the Θ can be set:

- when p = 4, Θ = {90°, 110°, 130°, 150°},
- when p = 6, Θ = {90°, 102°, 114°, 126°, 138°, 150°},
- when p = 16, Θ = {90°, 94°, 98°, ..., 142° , 146°, 150°}.

For GC-CSV-Aux-IVE, we updated $\boldsymbol{w}_f$ 50 times while for CSV-Aux-IVE, we updated $\boldsymbol{w}_f$ 100 times. After applying GC, the convergence speed of the iterations is greatly improved, which results in the reduction of the number of convergence iterations required for convergence, compared to CSV-Aux-IVE. As a result, the convergence can be achieved in significantly fewer iterations.

[1]https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator

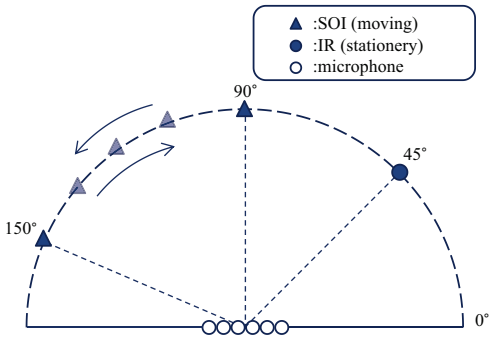

Fig. 1. Configurations of sources and microphones

TABLE I
EXPERIMENTAL CONDITIONS

Sampling rate	16 kHz
STFT window function	Hann
STFT window length	512 samples
STFT shift size	128 samples
Reverberation time	100 ms
Microphone spacing	2 cm
Input SNR	15 dB
L_b	100
λ_{GC}	0.2

TABLE II
AVERAGE SDR, SIR, AND ACCURACY OF EXTRACTING SOI (M=6)

method	SDR [dB]	SIR [dB]	Accuracy.
CSV-Aux-IVE	7.89	10.53	40.4%
GC-CSV-Aux-IVE (p=4)	11.42	15.73	97.4%
GC-CSV-Aux-IVE (p=6)	11.79	16.50	98.2%
GC-CSV-Aux-IVE (p=16)	**12.36**	**17.87**	**99.6%**

We exhibited the source extraction performance using the average source-to-distortion ratio (SDR), the source-to-interference ratio (SIR), and the accuracy of target source extraction. These evaluations were calculated by using BSS-eval [24] and setting the dry source as a reference. In order to obtain a more accurate extraction accuracy, we divided the signals into 20 blocks for comparison purposes. Hence we applied the BSS-eval to each block individually and calculated the average of the results from the 20 blocks. And we evaluated the accuracy of the extraction by assessing the output signal order.

B. Result

Table II shows the average SDR, SIR, and accuracy of the output signal. After applying GC, the accuracy of extracting SOI significantly improved from the initial 40% to over 98%. Additionally, the SDR experienced an improvement of approximately 4 dB, while the SIR showed an improvement of around 6 dB. Furthermore, we examined the performance of the extraction method in the different cases of p: $p > M$, $p = M$, and $p < M$ in Table II. Then we evaluated the extraction results under these three scenarios. Interestingly,

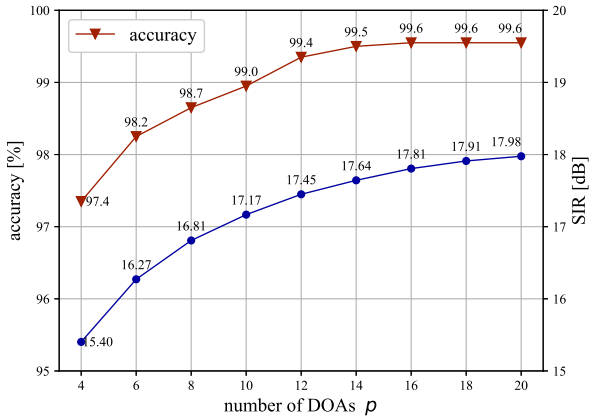

Fig. 2. Effect of increasing p on extraction performance

as the number of UR constraints p was increased, a notable improvement in the extraction performance was observed. The trend of this improvement is depicted in Fig. 2. The increasing trend of both the accuracy and SIR were observed with the increment of p.

V. CONCLUSION

This paper discovered a novel and effective BSE method for moving source. We apply a GC-based approach to address the permutation problem inherent in BSE. Our approach utilizes prior spatial information to aid in the filter updates. We extract the moving SOI by constraining multiple angles within the known range. Then we explore the effect of the changing number of constraint angles on the results. When the number of DOAs p is increased, the extraction effect is also improved. The results also demonstrated that, under similar conditions to the moving target source, GC-CSV-Aux-IVE algorithm achieved an average improvement of approximately 7dB in terms of SIR compared to the original algorithm and achieved an accuracy rate of nearly 100%.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] S. Makino, *Audio source separation*. Springer, 2018.
[2] S. Y. Low, R. Togneri, and S. Nordholm, "Spatio-temporal processing for distant speech recognition," in *Proc. ICASSP*, vol. 1, 2004, pp. I–1001.
[3] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.
[4] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
[5] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. ISSPA*, vol. 2, 2003, pp. 411–414.
[6] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.
[7] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IWAENC*, 2012, pp. 1–4.

[8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.
[9] 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.
[10] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. WASPAA*, 2019, pp. 185–189.
[11] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. SP*, vol. 69, pp. 3252–3267, 2021.
[12] Z. Koldovskỳ and P. Tichavskỳ, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. SP*, vol. 67, no. 4, pp. 1050–1064, 2018.
[13] J. Malek, J. Cmejla, and Z. Koldovsky, "Blind extraction of target speech source: Three ways of guidance exploiting supervised speaker embeddings," in *Proc. IWAENC*, 2022, pp. 1–5.
[14] Z. Koldovskỳ, J. Málek, and J. Janskỳ, "Extraction of independent vector component from underdetermined mixtures through block-wise determined modeling," in *Proc. ICASSP*, 2019, pp. 7903–7907.
[15] V. Kautský, Z. Koldovský, P. Tichavský, and V. Zarzoso, "Cramér–rao bounds for complex-valued independent component extraction: Determined and piecewise determined mixing models," *IEEE Trans. SP*, vol. 68, pp. 5230–5243, 2020.
[16] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP JASMP*, vol. 2022, no. 1, pp. 1–16, 2022.
[17] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.
[18] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.
[19] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. ICASSP*, 2020, pp. 846–850.
[20] 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.
[21] A. H. Khan, M. Taseska, and E. A. P. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.
[22] 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.
[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.