GEOMETRICALLY CONSTRAINED JOINT MOVING SOURCE EXTRACTION AND DEREVERBERATION BASED ON CONSTANT SEPARATING VECTOR MIXING MODEL

Mingxue Song, Tetsuya Ueda, Ruifeng Zhang, Jiahui Hu, Shoji Makino

Waseda University, Japan

ABSTRACT

This paper proposes a joint optimization algorithm that simultaneously achieves geometrically constrained moving source extraction and dereverberation. Recently, Constant Separating Vector (CSV) mixing model has been proposed, which is useful for moving source extraction. Based on this model, Independent Vector Extraction with auxiliary function (CSV-AuxIVE) has been proposed as a fast and stable moving source extraction method. When applying CSV-AuxIVE to a more realistic situation, we need to avoid the uncertainty of which signal CSV-AuxIVE extracts. Moreover, its source extraction performance will be limited when the reverberation time is long. Thus, to extract the moving target source signal in a highly reverberant environment, we derive an algorithm that jointly optimizes Weighted Prediction Error (WPE)-based dereverberation and CSV-AuxIVE with Geometric Constraint (GC), which we call GC-CSV-WPEIVE. By applying several GCs into the range of the moving target speaker, our proposed algorithm can extract the target signal, even in a highly reverberant environment. We show the efficacy of GC-CSV-WPEIVE using a simulation experiment. Furthermore, we discuss the robustness of its extraction performance against the moving range estimation error.

***Index Terms*—** Blind Source Extraction, Constant Separating Vector, Weighted Prediction Error, Geometric Constraint, Dereverberation

1. INTRODUCTION

When speeches are captured by a distant microphone array, diffuse noises, directional interferences, and long reverberation significantly degrade the performance of speech applications such as teleconferences and speech recognition. In more realistic scenarios, the speakers are likely to move in a specific range. Thus, moving source extraction methods have been extensively researched recently.

To extract the moving Source of Interest (SOI), we focus on a frequency domain approach using Short Time Fourier Transform (STFT) and use Indenpendent Component Analysis (ICA) [1, 2]-based source extraction methods. ICA is a basis of Blind Source Separation (BSS), which separates observed mixtures into independent source signals without any prior information. Independent Vector Analysis (IVA) [3, 4] simultaneously solves source separation and frequency-domain permutation problems by utilizing higher-order correlations between frequency bins in each source. Independent Vector Extraction (IVE) [5] is an extension of IVA, which extracts only N non-Gaussian components using $M(> N)$ microphones. It skips separating the background noises and thus reduces the computational cost when we use many microphones. Recently, it has been further extended to auxiliary function-based IVE (AuxIVE) [6, 7], which utilizes an anxiliary function technique [8, 9] for rapid convergence and stable calculation.

While the aforementioned methods apply a static mixing and separating model, their source extraction performance is limited when the sound sources are moving. For this situation, online source extraction (e.g., [10]) tackles it by estimating a time-varying separating model using forgetting coefficients. On the other hand, this paper focuses on the Constant Separating Vector (CSV) mixing model-based IVE for moving source extraction [11, 12]. It handles the moving situation by assuming a time-varying mixing model while applying a time-invariant separating vector. Its extension using the auxiliary function-based algorithm (CSV-AuxIVE) has been proposed recently [13]. It is reported that CSV-AuxIVE shows robust source extraction performance against the source movement and outperforms the online source extraction methods.

In this paper, we aim to address the following issues simultaneously based on CSV-AuxIVE: 1) uncertainty in extracting the SOI and 2) degradation of source extraction due to reverberation. Many studies have addressed each of these two problems. For the former, researchers have applied Geometric Constraint (GC) [14] to control the IVE's output signal based on spatial prior information [15, 16, 17]. Among them, GC-CSV-AuxIVE [17] applied several constraints pointing to directions in the moving range and showed a robust extraction of the SOI. For the latter, researchers have proposed a joint optimization algorithm with IVE and Weighted Prediction Error (WPE) [18]-based dereverberation [19, 20]. Recently, research [21] applied WPE into CSV-AuxIVE (CSV-WPEIVE) [21] and improved the moving source extraction performance by jointly optimizing them. In non-CSV-based methods, joint optimization of WPE and IVE with GC has been proposed (GC-WPEIVE) [22], which shows high extraction performance for fixed source situations. However, CSV-based methods that perform dereverberation and geometrically constrained moving source extraction have yet to be proposed.

Thus, we derive a joint optimization algorithm that comprises the benefits of GC-CSVAuxIVE [17] and CSV-WPEIVE [21], which we call GC-CSV-WPEIVE. In this research, we assume to obtain the range of the moving SOI from, e.g., human look directions or using a camera. Then, we optimize WPE and CSV-AuxIVE with several GCs toward the range of the moving SOI. Our simulation experiment shows that our proposed algorithm can effectively extract the moving SOI, even in a highly reverberant environment. In addition, we evaluate its extraction performance when the range estimation error occurs. While the previous research used only the exact range [17], our experiment demonstrates the extraction performance's robustness against the moving range estimation error.

2. PROBLEM FORMULATION

Let us consider the situation where we extract the moving SOI from M microphones. We denote the observed signals as $\boldsymbol{x}_{f,n} = [x_{1,f,n}, \ldots, x_{M,f,n}]^{\mathsf{T}} \in \mathbb{C}^M$, the SOI as $s_{f,n} \in \mathbb{C}$, and the back-

ground signals as $\boldsymbol{z}_{f,n} \in \mathbb{C}^{M-1}$ at each time frame $n = 1, \ldots, N$ and frequency bin $f = 1, \ldots, F$ in the STFT domain. Here, $(\cdot)^{\mathsf{T}}$ represents the transpose. Note that the background signals $\boldsymbol{z}_{f,n}$ include not only background noises but also an interference signal.

We assume that the relation between $s_{f,n}$, $\boldsymbol{z}_{f,n}$, and $\boldsymbol{x}_{f,n}$ can be written using time-varying convolutive transfer functions:

$$\boldsymbol{x}_{f,n} = \sum_{l=0}^{L-1} \boldsymbol{A}_{f,n,l} \begin{bmatrix} s_{f,n-l} \\ \boldsymbol{z}_{f,n-l} \end{bmatrix} \tag{1}$$

where the past frames $l = 1, \ldots, L$ are convolved. $\boldsymbol{A}_{f,n,l}$ is the convolutive transfer matrix from the source and noises to the microphones. We also assume that the moving range of the SOI is given or estimated as ϕ in advance. This paper aims to extract the SOI $s_{f,n}$ by using this moving range ϕ and observed signals $\boldsymbol{x}_{f,n}$.

3. MODELS

3.1. CSV mixing model with dereverberation

Before extracting the SOI from observed signals $\boldsymbol{x}_{f,n}$, we first apply a dereverberation filter $\boldsymbol{D}_f \in \mathbb{C}^{ML \times M}$ to obtain a dereverberated signals $\boldsymbol{y}_{f,n} \in \mathbb{C}^{M}$, as modeled in [21]:

$$\boldsymbol{y}_{f,n} = \boldsymbol{x}_{f,n} - \boldsymbol{D}_f^{\mathsf{H}} \overline{\boldsymbol{x}}_{f,n}, \tag{2}$$

where $\overline{\boldsymbol{x}}_{f,n} = [\boldsymbol{x}_{f,n-1}^{\mathsf{T}}, \ldots, \boldsymbol{x}_{f,n-L}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML}$ is a vector containing a past observation sequence for L frames and $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose.

Next, we extract the SOI from the dereverberated signals $\boldsymbol{y}_{f,n}$ using CSV mixing model [12, 13]. In this model, we let the frames be divided into $T \geq 1$ time intervals called blocks, and each block includes N_b frames for simplicity, hence $N = TN_b$. Hereafter, we treat the frame index $\{(t-1)N_b + 1, ..., tN_b\}$ as the same block index t for $t = 1, \ldots, T$. For example, we denote $\boldsymbol{y}_{f,(t-1)N_b+n'}$ as $\boldsymbol{y}_{f,t,n'}$ for $t = 1, \ldots, T$ and $n' = 1, \ldots, N_b$. Then, we obtain both $s_{f,t,n'}$ and $\boldsymbol{z}_{f,t,n'}$ by using a semi-time-varying separating model [12, 13]:

$$\begin{bmatrix} s_{f,t,n'} \\ \boldsymbol{z}_{f,t,n'} \end{bmatrix} = \boldsymbol{W}_{f,t}^{\mathsf{H}} \boldsymbol{y}_{f,t,n'}, \tag{3}$$

where $\boldsymbol{W}_{f,t}^{\mathsf{H}} = \boldsymbol{A}_{f,t}^{-1}$ is a separating matrix. We can parameterize both separating matrix $\boldsymbol{W}_{f,t}$ and mixing matrix $\boldsymbol{A}_{f,t}$ as proposed by [5]:

$$\boldsymbol{W}_{f,t} = [\boldsymbol{w}_f \ \boldsymbol{B}_{f,t}] = \begin{bmatrix} \beta_f & \boldsymbol{g}_{f,t}^{\mathsf{H}} \\ \boldsymbol{h}_f & -\gamma_{f,t}^{*} \boldsymbol{I}_{M-1} \end{bmatrix}, \tag{4}$$

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

The time-invariant separating vector $\boldsymbol{w}_f \in \mathbb{C}^M$ in (4) enables us to extract one moving source stably. Hereafter, we omit the index n' to reduce redundancy.

3.2. Source model

Next, we derive a likelihood function to find the optimum $\boldsymbol{W}_{f,t}$ and $\boldsymbol{D}_f$. Let $\boldsymbol{s}_t = [s_{1,t}, \ldots, s_{F,t}] \in \mathbb{C}^F$ denote the vector component of the SOI, and its joint pdf is $p(\boldsymbol{s}_t)$. $p(\boldsymbol{z}_{f,t})$ denotes the pdf of the background signals $\boldsymbol{z}_{f,t}$. We assume $\boldsymbol{s}_t$ and $\boldsymbol{z}_{f,t}$ are mutually independent over all times and frequencies and their joint pdf can be represented by the product of marginal pdfs as

$$p(\{\boldsymbol{s}_t, \boldsymbol{z}_{f,t}\}_{f,t}) = \prod_t p(\boldsymbol{s}_t) \prod_{f,t} p(\boldsymbol{z}_{f,t}). \tag{6}$$

The variance of SOI can be different from block to block; $p(\boldsymbol{s}_t)$ can be set as the following pdf:

$$p(\boldsymbol{s}_t) = g\left(\left\{ \frac{s_{f,t}}{\hat{\sigma}_{f,t}} \right\}_f \right) \left(\prod_{f=1}^{F} \hat{\sigma}_{f,t} \right)^{-2}, \tag{7}$$

where $\hat{\sigma}_{f,t} = \sqrt{\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{C}_{f,t} \boldsymbol{w}_f}$ is the variance of $s_{f,t}$ and $\boldsymbol{C}_{f,t} = \mathbb{E}\left[\boldsymbol{y}_{f,t} \boldsymbol{y}_{f,t}^{\mathsf{H}}\right]$ is a frame-based covariance matrix of $\boldsymbol{y}_{f,t}$. $g(\cdot)$ is a pdf corresponding to a normalized non-Gaussian random variable with its time-varying variance r_t:

$$g\left(\left\{ \frac{s_{f,t}}{\hat{\sigma}_{f,t}} \right\}_f \right) = C \exp(-G_R(r_t)), \tag{8}$$

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 $r \geq 0$. In our implementation, we use $\psi(r) = r^{-1}$, which represents the super-Gaussian signals well. The pdf of the background is assumed to be circular Gaussian with zero mean and covariance matrix $\boldsymbol{C}_{\boldsymbol{z}_{f,t}} = \mathbb{E}[\boldsymbol{z}_{f,t} \boldsymbol{z}_{f,t}^{\mathsf{H}}]$:

$$p(\boldsymbol{z}_{f,t}) = \mathcal{N}(\boldsymbol{0}_{M-1}, \boldsymbol{C}_{\boldsymbol{z}_{f,t}}), \tag{9}$$

where $\boldsymbol{0}_M \in \mathbb{R}^M$ is a zero vector.

Based on the assumptions above, we can obtain the negative log-likelihood function of the observed signals $\mathcal{X} = \{x_{m,f,t}\}_{m,f,t}$:

$$\mathcal{L}(\mathcal{X}) \overset{c}{=} \frac{1}{T} \sum_t^T \Bigg\{ \mathbb{E}\left[G_R\left(r_t\right)\right] + \sum_f \Bigg(\log \hat{\sigma}_{f,t}^2 + \mathbb{E}\left[\boldsymbol{z}_{f,t}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{z}_{f,t}}^{-1} \boldsymbol{z}_{f,t}\right] - \log |\gamma_{f,t}|^{2(M-2)} \Bigg) \Bigg\}, \tag{10}$$

where $\overset{c}{=}$ denotes equality up to the constant terms. By applying auxiliary function techniques [8] into (10), we can obtain the following auxiliary function to be minimized:

$$\mathcal{L}_{\text{aux}}(\mathcal{X}) \overset{c}{=} \frac{1}{T} \sum_{t=1}^{T} \sum_{f=1}^{F} \Bigg\{ \frac{1}{2} \frac{\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{V}_{f,t} \boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2} + \log \hat{\sigma}_{f,t}^2 + \mathbb{E}\left[\boldsymbol{z}_{f,t}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{z}_{f,t}}^{-1} \boldsymbol{z}_{f,t}\right] - (M-2) \log |\gamma_{f,t}|^2 \Bigg\}, \tag{11}$$

where

$$\boldsymbol{V}_{f,t} = \mathbb{E}\left[\psi(r_t) \boldsymbol{y}_{f,t} \boldsymbol{y}_{f,t}^{\mathsf{H}}\right]. \tag{12}$$

3.3. Geometric constraint

We introduce GC [14] as a regularization of the auxiliary function in (11). The regularization term is designed so that the separating vector $\boldsymbol{w}_f$ extracts the SOI. This research uses the following regularization term, which restricts the far-field response of the separating vector $\boldsymbol{w}_f$ at several directions θ in [15, 16, 17]:

$$\mathcal{L}_{\text{GC}}(\mathcal{W}) = \lambda_{\text{GC}} \sum_{f=1}^{F} \sum_{\theta \in \Theta} |\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{u}_{f,\theta} - 1|^2. \tag{13}$$

where $\mathcal{W} = \{\boldsymbol{w}_f\}_f$ represents all the separating vectors and Θ is the set of different directions θ corresponding to the SOI. $\boldsymbol{u}_{f,\theta}$ is the steering vector pointing to the direction θ, and λ_{GC} is a parameter that weighs the importance of the constraint. This term in (13) forces the separating vector $\boldsymbol{w}_f$ to respond 1 to the direction θ, to preserve the SOI.

4. PROPOSED METHOD: GC-CSV-WPEIVE

Based on the above models, we introduce our proposed algorithm. To handle a highly reverberant environment for moving source extraction, CSV-WPEIVE [21] has been proposed, which optimizes both $\boldsymbol{W}_{f,t}$ and $\boldsymbol{D}_f$ for moving source extraction. However, it has an uncertainty problem of which signal is being extracted. On the other hand, GC-CSV-AuxIVE [17] has been proposed to extract the SOI. Although this algorithm works well, it does not involve a dereverberation scheme (i.e., it treats $\boldsymbol{D}_f$ in (2) as a zero matrix and does not optimize $\boldsymbol{D}_f$). Thus, the ability of the source extraction will be limited under a highly reverberant environment. Thus, in this section, we propose a geometrically constrained CSV-WPEIVE (GC-CSV-WPEIVE), which utilizes the features of GC and dereverberation.

To realize moving source extraction according to the informed range ϕ under a highly reverberant environment, we use the following objective function:

$$\mathcal{L}(\mathcal{W}) = \mathcal{L}_{\text{aux}}(\mathcal{X}) + \mathcal{L}_{\text{GC}}(\mathcal{W}). \tag{14}$$

We use a coordinate descent method to reduce the objective function (14) by repeatedly updating $\mathcal{W}$, $\mathcal{D} = \{\boldsymbol{D}_f\}_f$, and $\mathcal{R} = \{r_t\}_t$ one by one as shown below.

4.1. Update of separating matrices $\mathcal{W}$

Before updating the separating vector $\boldsymbol{w}_f$, we update $\boldsymbol{a}_{f,t}$ instead of $\boldsymbol{B}_{f,t}$ (Note that they share the same parameters $\gamma_{f,t}$ and $\boldsymbol{g}_{f,t}$) under the distortionless constraint and the orthogonal constraint [5]. Orthogonal constraint restricts the zero sample correlation between the SOI $s_{f,t}$ and the noise signals $\boldsymbol{z}_{f,t}$, i.e., $\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{B}_{f,t} = \boldsymbol{0}_{1\times(M-1)}$. Under the above constraints, we can obtain the t-th mixing vector $\boldsymbol{a}_{f,t}$:

$$\boldsymbol{a}_{f,t} = \frac{\boldsymbol{C}_{f,t}\boldsymbol{w}_f}{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{w}_f}. \tag{15}$$

Then, we calculate the derivative of the objective function (14) with respect to $\boldsymbol{w}_f$:

$$\frac{\partial}{\partial \boldsymbol{w}_f^{\mathsf{H}}}\mathcal{L}_{\text{aux}} \overset{c}{=} \frac{1}{2T}\sum_{t=1}^{T}\left\{\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\boldsymbol{w}_f - \frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right\} + \lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{u}_{f,\theta}\boldsymbol{u}_{f,\theta}^{\mathsf{H}}\boldsymbol{w}_f - \lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{u}_{f,\theta}, \tag{16}$$

where we use the same technique as the previous research [5, 13] to replace the derivative from the second to fourth terms in (11) as $\boldsymbol{0}_M$. Then, we get the linearized solution of $\boldsymbol{w}_f$ by fixing $\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f$ and $\hat{\sigma}_{f,t}^2$ as constant terms [13, 17]:

$$\boldsymbol{w}_f \leftarrow \left(\lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{u}_{f,\theta}\boldsymbol{u}_{f,\theta}^{\mathsf{H}} + \frac{1}{2T}\sum_{t=1}^{T}\frac{\boldsymbol{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1}\left(\lambda_{\text{GC}}\sum_{\theta\in\Theta}\boldsymbol{u}_{f,\theta} + \frac{1}{2T}\sum_{t=1}^{T}\frac{\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2}\boldsymbol{a}_{f,t}\right). \tag{17}$$

4.2. Update of dereverberation filter $\mathcal{D}$

We update the dereverberation filter $\boldsymbol{D}_f$ based only on the SOI part like in [21]. By ignoring the third term in (11) and dropping the constant terms with respect to $\mathcal{D}$, we obtain

$$\mathcal{L}(\mathcal{D}) \overset{c}{=} \sum_{f=1}^{F}\left\|\left(\boldsymbol{D}_f - \boldsymbol{R}_f^{-1}\boldsymbol{P}_f\right)\boldsymbol{w}_f\right\|_{\boldsymbol{R}_f}^2, \tag{18}$$

where $\|\boldsymbol{x}\|_{\boldsymbol{R}} = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{R}\boldsymbol{x}$. Spatio-temporal covariance matrices $\boldsymbol{R}_f$ and $\boldsymbol{P}_f$ are calculated as

$$\boldsymbol{R}_f = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\psi(r_t)\frac{\overline{\boldsymbol{x}}_{f,t}\overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{\hat{\sigma}_{f,t}^2}\right], \tag{19}$$

$$\boldsymbol{P}_f = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\psi(r_t)\frac{\overline{\boldsymbol{x}}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}}{\hat{\sigma}_{f,t}^2}\right]. \tag{20}$$

Because $\boldsymbol{R}_f$ is positive definite, we can minimize the cost function in (18) by solving:

$$\boldsymbol{D}_f \leftarrow \boldsymbol{R}_f^{-1}\boldsymbol{P}_f. \tag{21}$$

4.3. Update of variances $\mathcal{R}$

After updating $\boldsymbol{y}_{f,t}$ and $s_{f,t}$ using (2) and (3), we update the variance r_t. This paper uses the coarse-fine source variance model [19] to improve the extraction performance of GC-CSV-WPEIVE further. To avoid the frequency-domain permutation problem, we use the frequency-invariant variance r_t for updating the separating vector $\boldsymbol{w}_f$ in (17) through (12):

$$r_t \leftarrow \sqrt{\sum_{f=1}^{F}\left|\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right|^2}. \tag{22}$$

On the other hand, we replace the variance r_t in (19) and (20) as a frequency-variant variance $r_{f,t}$:

$$r_t \leftarrow r_{f,t} = \left|\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right|. \tag{23}$$

5. EXPERIMENT

This section shows two experiments: 1) comparison of source extraction performance between conventional and proposed methods. 2) investigation of the source extraction performance against the moving range estimation error.

5.1. Experimental condition

We considered a situation where one target speech signal, one interference speech signal, and 4 point noises were mixed and observed by six microphones. Ten mixtures were generated for the experiment. We randomly selected two different speakers and obtained point-source speech signals from the test set of the CMU ARCTIC Corpus [23]. Then we concatenated them so that the length of each signal became 20 seconds. Point-source noise signals were recorded in a cafe (CAF) from the third 'CHiME' Speech Separation and Recognition Challenge [24]. We obtained room impulse response (RIR) data using the image method [25]. And we used the `signal generator`[1] to simulate RIRs. Figure 1 illustrates

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

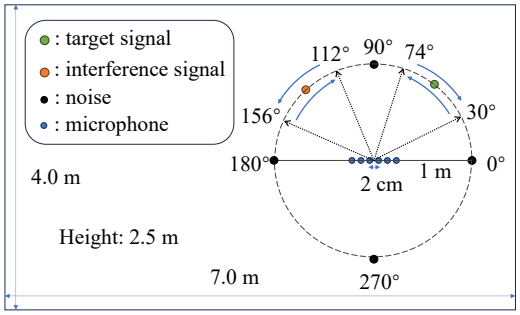

Fig. 1. Experimental sound source and microphone layout

the mixing condition. To investigate the trends of extraction performance, we considered two situations, one with a fixed interference speech signal at 135° and the other with an interference moving in $112° \leq \theta \leq 156°$ at a uniform angular velocity of $9°/s$. We let SOI move in $30° \leq \theta \leq 74°$ at the same velocity for both situations. Four noise signals were fixed at 0°, 90°, 180°, 270°. Before mixing SOI, interference, and noises, we adjusted input source-to-interference ratio (inputSIR) $= 10 \log \lambda_1/\lambda_2 = 0$ dB and input source-to-noise ratio (inputSNR) $= 10 \log \frac{\lambda_1}{\lambda_3+\ldots+\lambda_M} = 5$ dB. Here, λ_1 corresponds to the sample variance of SOI, λ_2 corresponds to that of the interference, and $\{\lambda_3, \ldots, \lambda_M\}$ correspond to those of noises. We set the sampling frequency to 16 kHz and the reverberation time $RT_{60} = 600$ ms.

For source extraction, we set the STFT window length and shift as 512 and 256 samples, respectively. We set the block length $N_b = 100$ frames and the dereverberation filter length $L = 4$. Dereverberation filters $\mathcal{D}$ were updated once every five iterations for computational efficiency. In total, we updated $\mathcal{W}$ 100 times and $\mathcal{D}$ 20 times. We set 10 geometric constraints at equal intervals in the moving range of the SOI 30° to 74° and the weight of GC λ_{GC} to 40.

We calculated the improvement of source-to-distortion ratio (iSDR) and source-to-interference ratio (iSIR), to show the extraction accuracy [26]. We used the MUSEVAL V4 toolkit with its *bss_eval_images* configuration and set the length of the bss_eval filter at 1 tap. In this experiment, all SOI, interference, and noise signals were set as the reference signals of bss_eval. Thus, we renamed SIR outputted from bss_eval as source-to-interference-and-noise Ratio (SINR).

5.2. Results

The moving source extraction performance of the proposed method and the baselines are presented in Table 1. In both situations, CSV-AuxIVE and CSV-WPEIVE cannot extract the SOI in every group of the dataset, which results in low iSINR. On the other hand, GC-CSV-AuxIVE yields high iSINR by incorporating GCs. Moreover, the proposed method shows the highest iSDR by further incorporating WPE for dereverberation. In the situation where the interference is static, although GC-WPEIVE in [22] can create a spatial null towards the interference, the proposed method matches the moving SOI better.

We next evaluated the iSDR of GC-CSV-WPEIVE by changing the given moving range ϕ. Figure 2 shows the change in iSDR when changing the lower and higher bounds of the moving range. The red area in the figure indicates that iSDR is higher than the highest iSDR of GC-WPEIVE [22] (= 4.93 dB). As shown in the figure, we obtained the highest iSDR in a range different from the true range

Table 1. Improvement of SDR (iSDR) and SINR (iSINR) [dB]

method (moving interference)	iSDR	iSINR
CSV-AuxIVE [13]	3.21	-1.58
CSV-WPEIVE [21]	3.92	-2.98
GC-CSV-AuxIVE [17]	4.11	10.57
GC-WPEIVE [22]	4.66	10.37
GC-CSV-WPEIVE (Proposed)	**5.19**	**11.82**
method (fixed interference)	iSDR	iSINR
CSV-AuxIVE [13]	3.13	-3.98
CSV-WPEIVE [21]	4.10	-2.12
GC-CSV-AuxIVE [17]	4.41	11.06
GC-WPEIVE [22]	5.01	10.63
GC-CSV-WPEIVE (Proposed)	**5.72**	**12.70**

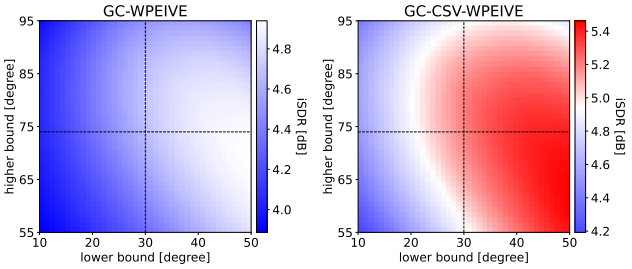

Fig. 2. iSDR of GC-WPE [22] and GC-CSV-WPEIVE with different GC ranges. Vertical and horizontal lines mean the true minimum and maximum angles of the moving range according to Fig. 1, respectively. White area represents the highest iSDR of GC-WPEIVE (= 4.93 dB).

$(30°, 74°)$. And our proposed method can result in higher iSDR than 4.93 dB as long as we set the moving range appropriately. This result confirms the robustness of the extraction performance against the moving range estimation error.

6. CONCLUSIONS

In this paper, we proposed GC-CSV-WPEIVE that simultaneously achieved geometrically constrained moving source extraction and dereverberation. We optimized WPE and CSV-AuxIVE with several GCs toward the range of the moving speaker. The experimental results demonstrated that the proposed GC-CSV-WPEIVE effectively extracts the moving SOI, even in a highly reverberant environment. Moreover, the investigation of the moving range exhibited its robustness against the estimation error of the range.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[3] 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.

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

[5] 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.

[6] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. WASPAA*, 2019, pp. 185–189.

[7] 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.

[8] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *Proc. LVA/ICA*, 2010, pp. 165–172.

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

[10] 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.

[11] 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.

[12] Z. Koldovskỳ, V. Kautskỳ, P. Tichavskỳ, J. Čmejla, and J. Málek, "Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers," *IEEE Trans. SP*, vol. 69, pp. 2158–2173, 2021.

[13] 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.

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

[15] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. ICASSP*, 2020, pp. 846–850.

[16] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. SP*, vol. 68, pp. 3545–3558, 2020.

[17] R. Zhang, T. Ueda, and S. Makino, "Geometrically constrained blind moving source extraction based on constant separation vector and auxiliary function technique," in *Proc. APSIPA*, 2023, pp. 2008–2012.

[18] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. ICASSP*, 2008, pp. 85–88.

[19] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. ICASSP*, 2021, pp. 6129–6133.

[20] R. Ikeshita and T. Nakatani, "Independent vector extraction for fast joint blind source separation and dereverberation," *IEEE SP Letters*, vol. 28, pp. 972–976, 2021.

[21] T. Ueda and S. Makino, "Constant separating vector-based blind source extraction and dereverberation for a moving speaker," in *Proc. EUSIPICO*, 2023, pp. 930–934.

[22] Y. Yang, X. Wang, A. Brendel, Z. Wen, W. Kellermann, and J. Chen, "Geometrically constrained source extraction and dereverberation based on joint optimization," in *Proc. EUSIPCO*, 2023, pp. 41–45.

[23] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004, pp. 223–224.

[24] 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.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943–950, 1979.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.