Switching Constant Separating Vector for Moving Source Extraction with Geometric Constraints

Changda Chen*, Yichen Yang*†, Yuehao Zhao*, Shoji Makino*, and Jingdong Chen†
*: Waseda University, Japan
†: Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China
E-mail: changda.chen@toki.waseda.jp, yang_yichen@mail.nwpu.edu.cn, zhaoyuehao@akane.waseda.jp,
s.makino@ieee.org, jingdongchen@ieee.org

Abstract—**Blind source extraction (BSE) becomes particularly challenging when the source of interest (SOI) is moving. Constant separating vector-based independent vector extraction with auxiliary function optimization (CSV-AuxIVE) uses a block-varying mixing model and estimates a time-invariant separating vector to extract the moving SOI. However, its performance deteriorates when the number of microphones is limited. To overcome this limitation, a switching mechanism was introduced in SwCSV-IVE, allowing different separating vectors across time blocks and frequency bins. Despite this improvement, SwCSV-IVE still suffers from instability due to inter-state permutation ambiguity. In this paper, we propose GC-SwCSV-IVE, a geometrically constrained switching CSV-based IVE method that incorporates geometric constraints (GC). By exploiting direction-of-arrival (DOA) information of the SOI, GC-SwCSV-IVE guides the optimization of switching separating vectors for improved robustness. Additionally, we introduce a state regularization technique to further enhance numerical stability during optimization. Experimental results show that the proposed method significantly improves both extraction performance and accuracy across various microphone configurations.**

I. Introduction

Blind source separation (BSS) is a fundamental technique in acoustic signal processing [1], aiming to separate individual source signals from observed mixtures in scenarios with multiple concurrent sources. A widely adopted approach is independent component analysis (ICA) [2]–[4], which achieves separation by maximizing the statistical independence among the sources. For speech signals, frequency-domain ICA (FD-ICA) [5] is commonly employed to handle convolutive mixtures with long room impulse responses (RIRs). However, FD-ICA suffers from the inner permutation problem, caused by the independent processing of each frequency bin. To address this problem, independent vector analysis (IVA) was proposed [6]–[8], which models higher-order dependencies across frequency components to align sources consistently. However, conventional IVA is limited to the determined case where the number of microphones equals the number of sources and is incapable of selectively extracting a specific source of interest (SOI), due to the outer permutation problem.

To enable focused extraction of a single target source, blind source extraction (BSE) has been developed to isolate a single SOI, while treating all other signals as interference or background noise. Independent vector extraction (IVE), a specialized variant of IVA, achieves the extraction by leveraging the non-Gaussian nature of the SOI in contrast to an approximately Gaussian background [9], [10]. However, these methods rely on a time-invariant mixing model [9], [10], which limits their effectiveness when the SOI is moving. To address this limitation, the constant separating vector (CSV) concept was introduced into the IVE framework [11], resulting in the CSV-AuxIVE method [12]. This approach allows the mixing system to change across time blocks. Despite this improvement, CSV-AuxIVE struggles when only a few microphones are available, as estimating a single constant separating vector that effectively spans the entire motion range of the SOI demands more spatial degrees of freedom than the array can provide.

To reduce reliance on the number of microphones, a switching mechanism [13]–[15] has been incorporated into IVE, resulting in SwIVE [16]. SwCSV-IVE [17] represents an early attempt to integrate this mechanism into the CSV model. It extends CSV-AuxIVE by assigning multiple separating vectors across time blocks and frequency bins. However, SwCSV-IVE is hindered by the inter-state permutation problem [18], where the extracted source may not consistently correspond to the SOI across different block-frequency regions. This inconsistency can prevent SwCSV-IVE from outperforming the original method or achieving reliable extraction.

To address this issue, in this paper, we propose a geometrically constrained switching CSV-based IVE (GC-SwCSV-IVE), which incorporates direction-of-arrival (DOA) based spatial information via geometric constraints (GC) [19]. Geometric constraints have shown effectiveness in both blind source separation and extraction [20]–[23]. Building on this, GC-SwCSV-IVE leverages prior knowledge of DOA of the SOI to guide the optimization of switching separating vectors, effectively mitigating the inter-state permutation issue. Additionally, a state regularization technique is introduced to enhance numerical stability, especially for states assigned to fewer block-frequency regions. Experimental results confirm that GC-SwCSV-IVE significantly improves both the extraction performance and accuracy across varying numbers of microphones.

II. Signal Model and Problem Formulation

Consider an acoustic scenario involving a target source along with multiple interference and noise sources. A microphone array with M elements is used to capture the signals. In the short-time Fourier transform (STFT) domain, the observed

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

signals can be represented as

$$\mathbf{x}_{f,l} = \mathbf{A}_f \begin{bmatrix} S_{f,l} \\ \mathbf{z}_{f,l} \end{bmatrix}, \tag{1}$$

where $f = 1, \ldots, F$ and $l = 1, \ldots, L$ index the frequency bins and time frames, respectively, F and L denote the total number of frequency bins and time frames, $\mathbf{A}_f \in \mathbb{C}^{M \times M}$ represents the mixing matrix, $\mathbf{x}_{f,l} \in \mathbb{C}^{M \times 1}$ represents the observed signals, $S_{f,l}$ is the SOI, and $\mathbf{z}_{f,l} \in \mathbb{C}^{(M-1) \times 1}$ represents the background signals.

To capture the time-varying mixing process resulting from the movement of the SOI, we adopt a block-varying mixing model, as used in CSV-based IVE methods [11], [12]. This model partitions the total L time frames into T ($\geq$1) consecutive blocks, each containing $L_b = L/T$ frames, and assumes that the mixing matrix remains constant within each block but can vary across blocks. Mathematically, this model is formulated as

$$\mathbf{x}_{f,t,l'} = \mathbf{A}_{f,t} \begin{bmatrix} S_{f,t,l'} \\ \mathbf{z}_{f,t,l'} \end{bmatrix}, \tag{2}$$

where $t = 1, \ldots, T$ denotes the block index, and $\mathbf{A}_{f,t} \in \mathbb{C}^{M \times M}$ represents the block-dependent mixing matrix. Based on (2), the original time frame-based signals $S_{f,(t-1)L_b+l'}$, $\mathbf{z}_{f,(t-1)L_b+l'}$, and $\mathbf{x}_{f,(t-1)L_b+l'}$ can be reformulated as $S_{f,t,l'}$, $\mathbf{z}_{f,t,l'}$, and $\mathbf{x}_{f,t,l'}$, respectively, where $l' = 1, \ldots, L_b$ denotes the frame index within the t-th block. Following the approach in [9], one can parameterize the mixing matrix $\mathbf{A}_{f,t}$ as

$$\begin{aligned} \mathbf{A}_{f,t} &= [\mathbf{a}_{f,t} \quad \mathbf{Q}_{f,t}] \\ &= \begin{bmatrix} \gamma_{f,t} & \mathbf{h}_f^{\mathsf{H}} \\ \mathbf{g}_{f,t} & \dfrac{1}{\gamma_{f,t}}(\mathbf{g}_{f,t}\mathbf{h}_f^{\mathsf{H}} - \mathbf{I}_{M-1}) \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathbf{a}_{f,t} \in \mathbb{C}^{M \times 1}$ denotes the block-dependent mixing vector for the SOI, and $\mathbf{Q}_{f,t} \in \mathbb{C}^{M \times (M-1)}$ represents the mixing subspace associated with the background signals. Let us assume that $\mathbf{A}_{f,t}$ is invertible. The corresponding block-varying demixing model is given by

$$\begin{bmatrix} \hat{S}_{f,t,l'} \\ \hat{\mathbf{z}}_{f,t,l'} \end{bmatrix} = \mathbf{W}_{f,t}\mathbf{x}_{f,t,l'}, \tag{4}$$

where $\mathbf{W}_{f,t} = \mathbf{A}_{f,t}^{-1}$.

Similarly, the demixing matrix $\mathbf{W}_{f,t}$ can be parameterized as

$$\mathbf{W}_{f,t} = \begin{bmatrix} \mathbf{w}_f^{\mathsf{H}} \\ \mathbf{B}_{f,t} \end{bmatrix} = \begin{bmatrix} \beta_f^* & \mathbf{h}_f^{\mathsf{H}} \\ \mathbf{g}_{f,t} & -\gamma_{f,t}\mathbf{I}_{M-1} \end{bmatrix}, \tag{5}$$

where $\mathbf{w}_f \in \mathbb{C}^{M \times 1}$ is the so-called constant separating vector, assumed to be time-invariant across all frames, $(\cdot)^{\mathsf{H}}$ denotes the conjugate transpose, and $\mathbf{B}_{f,t} \in \mathbb{C}^{(M-1) \times M}$ represents the demixing submatrix for the background signals.

It is important to note that the above parameterization is derived under the constraint that $\mathbf{B}_{f,t}\mathbf{a}_{f,t} = \mathbf{0}_{M-1}$ [9], where $\mathbf{0}_{M-1}$ denotes the $(M-1) \times 1$ zero vector. The parameters $\gamma_{f,t}$, $\mathbf{h}_f^{\mathsf{H}}$, $\mathbf{g}_{f,t}$ and β_f^* are introduced to simplify the derivation, where $(\cdot)^*$ denotes the complex conjugate, and $\mathbf{I}_{M-1}$ is the $(M-1) \times (M-1)$ identity matrix. With this formulation, the SOI can be directly estimated as

$$\hat{S}_{f,t,l'} = \mathbf{w}_f^{\mathsf{H}}\mathbf{x}_{f,t,l'}. \tag{6}$$

III. Proposed GC-SwCSV-IVE

A. Switching mechanism-based CSV model

The existence of a time-invariant separating vector generally requires a large number of microphones to ensure sufficient spatial degrees of freedom [12]; otherwise, the extraction performance may degrade. To address this limitation, we incorporate a switching mechanism [13]–[16] into the CSV model, as proposed in our previous work [17], to mitigate performance degradation in scenarios with a limited number of microphones. With this mechanism, the mixing model in (2) can be reformulated as

$$\mathbf{x}_{f,t,l'} = \sum_{j=1}^{J} \delta_{f,t}^{(j)} \mathbf{A}_{f,t}^{(j)} \begin{bmatrix} S_{f,t,l'} \\ \mathbf{z}_{f,t,l'} \end{bmatrix}, \tag{7}$$

where $\mathbf{A}_{f,t}^{(j)}$ denotes the mixing matrix corresponding to the j-th state, $\delta_{f,t}^{(j)} \in \{0, 1\}$ is a binary switching variable that satisfies $\sum_{j=1}^{J} \delta_{f,t}^{(j)} = 1$, the index $j = 1, \ldots, J$ represents the switching state with J being the total number of switching states. Based on the switching mixing model in (7), the corresponding demixing system can be rewritten as

$$\begin{bmatrix} \hat{S}_{f,t,l'} \\ \hat{\mathbf{z}}_{f,t,l'} \end{bmatrix} = \sum_{j=1}^{J} \delta_{f,t}^{(j)} \mathbf{W}_{f,t}^{(j)} \mathbf{x}_{f,t,l'}, \tag{8}$$

where the demixing matrix $\mathbf{W}_{f,t}^{(j)}$ can be reparameterized, following (5), as

$$\mathbf{W}_{f,t}^{(j)} = \begin{bmatrix} (\mathbf{w}_f^{(j)})^{\mathsf{H}} \\ \mathbf{B}_{f,t} \end{bmatrix} = \begin{bmatrix} (\beta_f^{(j)})^* & (\mathbf{h}_f^{(j)})^{\mathsf{H}} \\ \mathbf{g}_{f,t} & -\gamma_{f,t}\mathbf{I}_{M-1} \end{bmatrix}. \tag{9}$$

Consequently, the SOI can be estimated as

$$\hat{S}_{f,t,l'} = \sum_{j=1}^{J} \delta_{f,t}^{(j)} (\mathbf{w}_f^{(j)})^{\mathsf{H}} \mathbf{x}_{f,t,l'}, \tag{10}$$

where the term $\sum_{j=1}^{J} \delta_{f,t}^{(j)} (\mathbf{w}_f^{(j)})^{\mathsf{H}}$ acts as a selector, choosing the appropriate separating vector from the set $\{\mathbf{w}_f^{(j)}\}_j$ for each block-frequency bin. This allows different separating vectors to model distinct time-varying characteristics across blocks, instead of relying on a single separating vector as in the original CSV model. As a result, the method relaxes the need for a large number of microphones.

B. Probabilistic model

To incorporate prior DOA information into the optimization of the switching separating vectors, we adopt the maximum a posteriori (MAP) principle to derive the cost function [20]. Assuming that the observed signals are independently and identically distributed (i.i.d.) across all frequency bins and time frames, and that $S_{f,t,l'}$ and $\mathbf{z}_{f,t,l'}$ are mutually independent, we derive the negative MAP-based cost function. This cost is averaged over the blocks assigned to $\mathbf{w}_f^{(j)}$ for $1 \leq j \leq J$ and is given by

$$\begin{aligned}\mathcal{J} = &-\sum_{j=1}^{J}\sum_{f=1}^{F}\frac{1}{T_f^{(j)}}\sum_{t=1}^{T}\delta_{f,t}^{(j)}\Big(\log\mathbb{E}\left[p(S_{f,t,l'})\right]+\log\mathbb{E}\left[p(\mathbf{z}_{f,t,l'})\right]\\ &+\log\left|\det\mathbf{W}_{f,t}^{(j)}\right|^2\Big)-\sum_{j=1}^{J}\sum_{f=1}^{F}\log p(\mathbf{w}_f^{(j)}),\end{aligned}\tag{11}$$

where $T_f^{(j)} = \sum_{t=1}^{T}\delta_{f,t}^{(j)}$ represents the number of time blocks assigned to $\mathbf{w}_f^{(j)}$, $\mathbb{E}[\cdot]$ denotes the expectation over time frames within the t-th block, $p(S_{f,t,l'})$ and $p(\mathbf{z}_{f,t,l'})$ are the probability density functions (PDFs) of the SOI and background signals, respectively, and $p(\mathbf{w}_f^{(j)})$ represents the prior distribution of the separating vector from $\mathbf{W}_{f,t}^{(j)}$ as we focus only on modeling the prior of the SOI. Note that the first two terms of (11) correspond to the negative log-likelihood (NL). The background signals are assumed to follow a zero-mean circular complex Gaussian distribution, i.e., $p_{\mathbf{z}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{\mathbf{z}})$, and $\mathbf{C}_{\mathbf{z}}$ is the covariance matrix of $\mathbf{z}_{f,t}$. The PDF of the SOI is given by

$$p_{\mathrm{s}}(S_{f,t,l'}) = f\left(\frac{S_{f,t,l'}}{\sigma_{f,t}}\right)\sigma_{f,t}^{-2},\tag{12}$$

where $\sigma_{f,t}^2$ is the block-varying variance of $S_{f,t,l'}$, which is unknown but can be approximated by the sample-based variance $\hat{\sigma}_{f,t}^2 = \sum_{j=1}^{J}\delta_{f,t}^{(j)}(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{C}_{f,t}\mathbf{w}_f^{(j)}$ with $\mathbf{C}_{f,t} = \mathbb{E}[\mathbf{x}_{f,t,l'}\mathbf{x}_{f,t,l'}^{\mathsf{H}}]$ being the covariance matrix of $\mathbf{x}_{f,t,l'}$, and $f(\cdot)$ is assumed to be a PDF modeling a time-varying Gaussian variable, enabling the switching mechanism across frequency bins. This formulation differs from that in [12] and is expressed as

$$f\left(\frac{S_{f,t,l'}}{\hat{\sigma}_{f,t}}\right) \propto \frac{1}{r_{f,t,l'}}\exp\left(-\frac{|S_{f,t,l'}|^2}{r_{f,t,l'}\hat{\sigma}_{f,t}^2}\right),\tag{13}$$

where $r_{f,t,l'}$ is the time-varying variance of $S_{f,t,l'}/\hat{\sigma}_{f,t}$.

Permutation inconsistencies arise when different separating vectors from the set $\{\mathbf{w}_f^{(j)}\}_j$ are applied across frequency bins and time blocks, leading to ambiguity in the extracted signal, i.e., it may correspond to either the SOI or one of the background signals. This ambiguity negatively impacts separation performance. While CSV-AuxIVE aims to estimate a single separating vector that covers the entire movement range of the SOI, our approach instead guides the optimization of multiple separating vectors, encouraging each to produce strong beam responses within that range after switching. To achieve this, we introduce a prior in (11) that constrains the far-field responses of the switching separating vectors at specified DOAs. We refer to this prior as the GC, which is defined as

$$\begin{aligned}&-\sum_{j=1}^{J}\sum_{f=1}^{F}\log p(\mathbf{w}_f^{(j)})\\ &\qquad= \lambda_{\mathrm{GC}}\sum_{j=1}^{J}\sum_{f=1}^{F}\sum_{\phi\in\Phi}\left|(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{d}_{f,\phi}-c_\phi\right|^2,\end{aligned}\tag{14}$$

where λ_{GC} is the weight of GC, Φ is the set of steering directions, $\mathbf{d}_{f,\phi}$ is the steering vector corresponding to direction ϕ, and c_ϕ is a non-negative constraint. A value of $c_\phi \geq 1$ indicates signal enhancement from direction ϕ, while smaller values correspond to suppression. Since the objective is to preserve only the SOI, we set $c_\phi = 1$ within the angular range of the SOI, which is assumed to be known (i.e., the start and end directions of the movement of the SOI are given). The set $\Phi = \{\phi_1, \ldots, \phi_p\}$ consists of p uniformly spaced directions covering this range. Note that each direction in Φ is weighted equally, as the precise trajectory of the SOI is unknown.

By substituting the assumed PDFs of $S_{f,t,l'}$ and $\mathbf{z}_{f,t,l'}$ into (11), we derive the final cost function, given by

$$\begin{aligned}\mathcal{J}(\Theta) = &\sum_{j=1}^{J}\sum_{f=1}^{F}\frac{1}{T_f^{(j)}}\sum_{t=1}^{T}\delta_{f,t}^{(j)}\mathcal{J}_{\mathrm{NL}}^{(j)}(\theta_{f,t}^{(j)})\\ &+\lambda_{\mathrm{GC}}\sum_{j=1}^{J}\sum_{f=1}^{F}\sum_{\phi\in\Phi}\left|(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{d}_{f,\phi}-1\right|^2,\end{aligned}\tag{15}$$

where

$$\begin{aligned}\mathcal{J}_{\mathrm{NL}}^{(j)}(\theta_{f,t}^{(j)}) = &\,\mathbb{E}\left[\frac{|(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{x}_{f,t,l'}|^2}{r_{f,t,l'}\hat{\sigma}_{f,t}^2}\right]+\log\mathbb{E}\left[r_{f,t,l'}\right]\\ &+\log\hat{\sigma}_{f,t}^2+\mathbb{E}\left[\mathbf{z}_{f,t,l'}^{\mathsf{H}}\mathbf{C}_{\mathbf{z}}^{-1}\mathbf{z}_{f,t,l'}\right]\\ &-\log|\gamma_{f,t}|^{2(M-2)}-\log|(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{a}_{f,t}|^2,\end{aligned}\tag{16}$$

$\Theta = \{\mathcal{S}, \mathcal{R}, \mathcal{A}, \mathcal{W}, \mathcal{D}\}$, $\mathcal{S} = \{\hat{\sigma}_{f,t}^2\}_{f,t}$, $\mathcal{R} = \{r_{f,t,l'}\}_{f,t,l'}$, $\mathcal{A} = \{\mathbf{a}_{f,t}\}_{f,t}$, $\mathcal{W} = \{\mathbf{w}_f^{(j)}\}_{j,f}$, $\mathcal{D} = \{\delta_{f,t}^{(j)}\}_{j,f,t}$, and $\theta_{f,t}^{(j)} = \left\{\hat{\sigma}_{f,t}^2, \{r_{f,t,l'}\}_{l'}, \mathbf{a}_{f,t}, \mathbf{w}_f^{(j)}, \delta_{f,t}^{(j)}\right\}$.

C. Optimization algorithm

Since (15) does not have a closed-form solution, we employ a coordinate descent algorithm [24] to minimize it by iteratively updating each parameter set: $\mathcal{S}$, $\mathcal{R}$, $\mathcal{A}$, $\mathcal{W}$, and $\mathcal{D}$, while keeping the others fixed. Proper initialization is particularly important for updating $\mathcal{W}$ with the switching mechanism [18]. As illustrated in Fig. 1, we first describe the initialization strategy, followed by the parameter update procedure.

1) Initialization: We begin by applying GC-CSV-AuxIVE, a geometrically constrained extension of CSV-AuxIVE [23], which uses the same DOA information as our proposed method. This is used to update $\{\mathbf{w}_f\}_f$ over K iterations. The separating vectors $\{\mathbf{w}_f^{(j)}\}_f$ for each switching state j are then initialized with the resulting values of $\{\mathbf{w}_f\}_f$. The switching weights $\mathcal{D}$ are initially assigned random values in the range $[0, 1]$, and are later binarized during the update process.

2) Update of $\mathcal{S}$ and $\mathcal{R}$: The sample-based variance $\hat{\sigma}_{f,t}^2$ is updated as

$$\hat{\sigma}_{f,t}^2 \leftarrow \sum_{j=1}^{J}\delta_{f,t}^{(j)}(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{C}_{f,t}\mathbf{w}_f^{(j)}.\tag{17}$$

Then, $r_{f,t,l'}$ is updated as

$$r_{f,t,l'} \leftarrow \frac{\sum_{j=1}^{J}\delta_{f,t}^{(j)}\left|(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{x}_{f,t,l'}\right|^2}{\hat{\sigma}_{f,t}^2}.\tag{18}$$

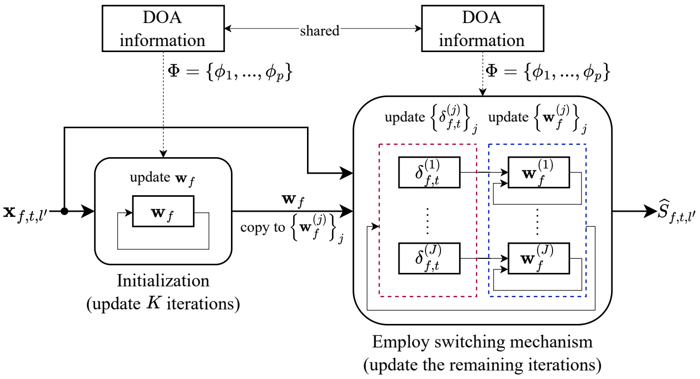

DOA information
shared
DOA information
$\Phi = \{\phi_1, ..., \phi_p\}$
$\Phi = \{\phi_1, ..., \phi_p\}$
update $\left\{\delta_{f,t}^{(j)}\right\}_j$
update $\left\{\mathbf{w}_f^{(j)}\right\}_j$
update $\mathbf{w}_f$
$\mathbf{w}_f$
$\mathbf{x}_{f,t,l'}$
$\mathbf{w}_f$
copy to $\left\{\mathbf{w}_f^{(j)}\right\}_j$
$\delta_{f,t}^{(1)}$
$\mathbf{w}_f^{(1)}$
$\delta_{f,t}^{(J)}$
$\mathbf{w}_f^{(J)}$
$\widehat{S}_{f,t,l'}$
Initialization
(update K iterations)
Employ switching mechanism
(update the remaining iterations)

Fig. 1: Flowchart of the GC-SwCSV-IVE Optimization Procedure.

To prevent the frequency permutation problem during the subsequent update of $\mathcal{W}$, we adopt a frequency-independent source model [15] and define $r_{t,l'}$ as

$$r_{t,l'} \leftarrow \frac{1}{F}\sum_{f=1}^{F} r_{f,t,l'}. \tag{19}$$

3) Update of $\mathcal{A}$ and $\mathcal{W}$: To ensure reliable estimation of $\mathbf{w}_f^{(j)}$, the parameter vectors are subject to a distortionless constraint and an orthogonality condition. Specifically, we impose the distortionless constraint $\sum_{j=1}^{J}\delta_{f,t}^{(j)}(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{a}_{f,t} = 1$, along with the orthogonality constraint [12], which assumes that the SOI is uncorrelated with the background signals, i.e., $\mathbb{E}[\hat{\mathbf{z}}_{f,t,l'}\hat{S}_{f,t,l'}] = \mathbf{0}_{M-1}$. By combining these constraints, the expression for $\mathbf{a}_{f,t}$ as a function of $\mathbf{w}_f^{(j)}$ can be derived following the method in [9], yielding

$$\mathbf{a}_{f,t} \leftarrow \frac{\sum_{j=1}^{J}\delta_{f,t}^{(j)}\mathbf{C}_{f,t}\mathbf{w}_f^{(j)}}{\hat{\sigma}_{f,t}^2}. \tag{20}$$

Next, we isolate the terms in (15) that depend solely on $\mathbf{w}_f^{(j)}$ and deduce their derivative with respect to $(\mathbf{w}_f^{(j)})^{\mathsf{H}}$, yielding

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial(\mathbf{w}_f^{(j)})^{\mathsf{H}}} = &\frac{1}{T_f^{(j)}}\sum_{t=1}^{T}\delta_{f,t}^{(j)}\left(\frac{\mathbf{V}_{f,t}\mathbf{w}_f^{(j)}}{\hat{\sigma}_{f,t}^2} - \frac{(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{V}_{f,t}\mathbf{w}_f^{(j)}}{\hat{\sigma}_{f,t}^2}\mathbf{a}_{f,t}\right)\\ &+ \lambda_{\text{GC}}\sum_{\phi\in\Phi}\left(\mathbf{d}_{f,\phi}\mathbf{d}_{f,\phi}^{\mathsf{H}}\mathbf{w}_f^{(j)} - \mathbf{d}_{f,\phi}\right),\end{aligned} \tag{21}$$

where $\mathbf{V}_{f,t} = \mathbb{E}[(\mathbf{x}_{f,t,l'}\mathbf{x}_{f,t,l'}^{\mathsf{H}})/r_{t,l'}]$.

By treating the terms $(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{V}_{f,t}\mathbf{w}_f^{(j)}$ and $\hat{\sigma}_{f,t}$ as constants, we can solve the linearized equation obtained by setting (21) to zero. The resulting solution is given by

$$\begin{aligned}\mathbf{w}_f^{(j)} \leftarrow &\left(\lambda_{\text{GC}}\sum_{\phi\in\Phi}\mathbf{d}_{f,\phi}\mathbf{d}_{f,\phi}^{\mathsf{H}} + \frac{1}{T_f^{(j)}}\sum_{t=1}^{T}\frac{\delta_{f,t}^{(j)}\mathbf{V}_{f,t}}{\hat{\sigma}_{f,t}^2}\right)^{-1}\cdot\\ &\left(\lambda_{\text{GC}}\sum_{\phi\in\Phi}\mathbf{d}_{f,\phi} + \frac{1}{T_f^{(j)}}\sum_{t=1}^{T}\frac{\delta_{f,t}^{(j)}(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{V}_{f,t}\mathbf{w}_f^{(j)}}{\hat{\sigma}_{f,t}^2}\mathbf{a}_{f,t}\right).\end{aligned} \tag{22}$$

4) Update of $\mathcal{D}$: To update the switching weights $\mathcal{D}$, we assign $\delta_{f,t}^{(j)} = 1$ to the state j that minimizes the cost function in (15), for each frequency bin and time block. The update rule for $\delta_{f,t}^{(j)}$ is given by

$$\delta_{f,t}^{(j)} \leftarrow \begin{cases}1, & \text{if } j = \underset{j'}{\operatorname{argmin}}\ \mathcal{J}_{\text{NL}}^{(j')}(\theta_{f,t}^{(j')})\\ 0, & \text{otherwise}\end{cases}. \tag{23}$$

Note that in updating $\delta_{f,t}^{(j)}$, the second, third, and fourth terms in (16) are ignored, and the GC term is omitted as it does not depend on $\delta_{f,t}^{(j)}$. Once $\delta_{f,t}^{(j)}$ is updated, the SOI is estimated using (10).

D. State regularization technique

Due to the instability in the distribution of $\delta_{f,t}^{(j)}$ across different time blocks at each frequency bin, the value of $T_f^{(j)}$ in (22) can become very small, or even approach zero, which may lead to numerical instability during the update process. This issue was observed in our preliminary experiments, highlighting the need for caution to avoid potential division-by-zero errors.

Inspired by a related work in [16], we rewrite the first term in (15) as

$$\mathcal{J}_{\text{NL}}(\Theta) = \sum_{j=1}^{J}\sum_{f=1}^{F}\frac{1}{(T_f')^{(j)}}\sum_{t=1}^{T}(\delta'_{f,t})^{(j)}\mathcal{J}_{\text{NL}}^{(j)}(\theta_{f,t}^{(j)}), \tag{24}$$

where $(\delta'_{f,t})^{(j)} = \delta_{f,t}^{(j)} + \lambda_{\text{state}}$ and $(T_f')^{(j)} = \sum_{t=1}^{T}(\delta'_{f,t})^{(j)}$ with λ_{state} being a small positive regularization term introduced to stabilize the state assignment. In the proposed method, we set

$$\lambda_{\text{state}} = \frac{1}{5J} \text{ for } J > 1. \tag{25}$$

For states spanning many time blocks, the regularization term has a minimal effect on the cost. Conversely, for states with fewer assigned time blocks, this regularization allows the unassigned block-frequency bins to exert a mild influence on optimizing their corresponding separating vectors, thereby improving numerical stability. By incorporating (24) in the final cost function (15), only minor adjustments to the original optimization algorithm are required, i.e., we replace $(\delta'_{f,t})^{(j)}$ with $\delta_{f,t}^{(j)}$ in (20) and (22), and replace $(T_f')^{(j)}$ with $T_f^{(j)}$ in (22). Finally, (17) and (18) are modified as

$$\hat{\sigma}_{f,t}^2 \leftarrow \frac{\sum_{j=1}^{J}(\delta'_{f,t})^{(j)}(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{C}_{f,t}\mathbf{w}_f^{(j)}}{1 + J\lambda_{\text{state}}}, \tag{26}$$

$$r_{f,t,l'} \leftarrow \frac{\sum_{j=1}^{J}(\delta'_{f,t})^{(j)}\left|(\mathbf{w}_f^{(j)})^{\mathsf{H}}\mathbf{x}_{f,t,l'}\right|^2}{\hat{\sigma}_{f,t}^2(1 + J\lambda_{\text{state}})}. \tag{27}$$

Note that the binary switching weights are still used to estimate the SOI.

IV. Experiments

A. Experimental setup

We create 20-second speech signals with a sampling rate of 16 kHz by concatenating speech clips from the CMU ARCTIC dataset [25]. The signal from the male speaker *bdl* serves as the

TABLE I: Average SDR (dB), SIR (dB), and extraction accuracy under different numbers of microphones.

Method	J	$M=3$			$M=4$			$M=5$			$M=6$		
		SDR	SIR	Acc.	SDR	SIR	Acc.	SDR	SIR	Acc.	SDR	SIR	Acc.
CSV-AuxIVE [12]	–	4.94	6.68	90.5%	6.09	8.72	93.8%	6.68	9.78	94.8%	6.81	10.19	94.8%
SwCSV-IVE [17]	2	5.35	7.31	92.3%	6.81	9.81	95.0%	7.51	11.02	96.3%	7.73	11.56	98.0%
SwCSV-IVE [17]	3	5.49	7.56	92.5%	6.70	9.66	96.0%	7.38	10.82	96.3%	7.78	11.60	97.0%
GC-CSV-AuxIVE [23]	–	5.74	7.89	95.8%	7.08	10.22	97.3%	7.22	10.48	96.5%	7.45	11.04	96.0%
GC-SwCSV-IVE	2	7.12	**10.35**	99.5%	8.30	**12.62**	**100%**	8.58	**13.41**	99.8%	8.94	**13.75**	**100%**
GC-SwCSV-IVE	3	**7.36**	10.23	**99.8%**	**8.49**	12.08	**100%**	**8.79**	12.91	**100%**	**9.32**	13.51	**100%**

SOI, while the female speaker *slt* serves as a single interference source. Background noise is simulated using recordings from the cafe (CAF) environment in CHiME-3 [26], modeled as point sources. The mixing setup, illustrated in Fig. 2, features the SOI moving between 80° and 120° at a constant angular speed of 8°/s, the interference fixed at 50° with an input SIR being 0 dB, and the noise sources positioned at 20°, 150° and 170°, respectively with an input SNR being 10 dB, where the noise power is the sum of the variances of the three noise signals. We use the image method [27] to generate RIRs for a room of size $5 \times 6 \times 3$ m. Mixed signals with a moving source were generated using the `signal generator`[1]. The reverberation time, T_{60}, was set to 200 ms. To obtain signals in the STFT domain, a 512-sample (32 ms) Hanning window with a 128-sample (8 ms) shift was applied.

We compare four CSV-based IVE methods in our experiments: CSV-AuxIVE, SwCSV-IVE, GC-CSV-AuxIVE, and the proposed GC-SwCSV-IVE. Each time block consists of $L_b = 100$ frames. As introduced in Sec. III-B, the GC weight is set to $\lambda_{\text{GC}} = 0.2$, and the number of constrained DOAs is $p = 6$, with $\Phi = \{80°, 88°, 96°, 104°, 112°, 120°\}$ in (14). The switching weights $\mathcal{D}$ are randomly initialized within the range 1 ± 10^{-3} and normalized such that $\sum_{j=1}^{J} \delta_{f,t}^{(j)} = 1$. After initialization, these weights are updated to binary values. All methods perform 50 iterations of the separating vector. For SwCSV-IVE and GC-SwCSV-IVE, the first $K = 25$ iterations serve as initialization, followed by switching vector updates in the remaining 25 iterations. Finally, the estimated SOI is obtained and scale ambiguity is resolved using projection back [28].

To evaluate the proposed methods, we use the BSS-EVAL toolbox [29] to compute SDR and SIR, using the clean signals of the SOI and interference as references. Performance is assessed in a block-wise manner to capture temporal variations in extraction accuracy. Each 20-second output signal is divided into 20 non-overlapping 1-second blocks. A block is considered correctly extracted if its SIR with respect to the SOI exceeds that with respect to the interference; otherwise, it is marked as incorrect. The accuracy for each signal is computed as the average of these block-level decisions. This evaluation is repeated over 20 test speech signals. We report the mean SDR, SIR, and accuracy (Acc.) across all utterances for uniform linear arrays with 3, 4, 5, and 6 microphones, respectively.

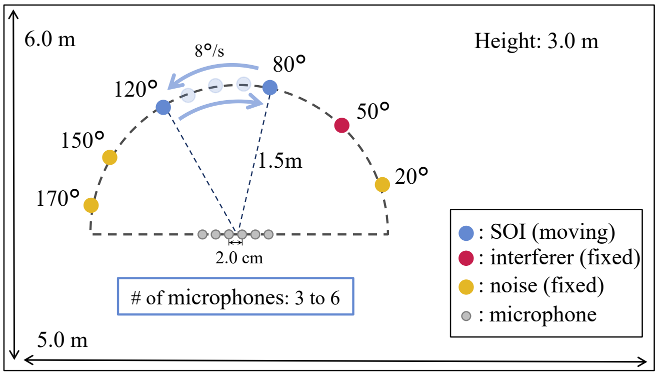


Fig. 2: Simulation layout for signal mixing.

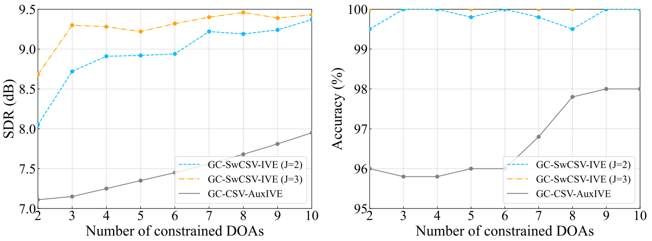
B. Experimental results

Table I presents the extraction performance and accuracy of all methods across different numbers of microphones, i.e., different values of M. When the number of microphones is limited, such as when $M = 3$, SwCSV-IVE can only offer slightly better average performance than CSV-AuxIVE, which is considered unsatisfactory due to the limited gains resulting from the inter-state permutation problem. In contrast, the proposed GC-SwCSV-IVE consistently delivers substantial improvements. Compared to GC-CSV-AuxIVE, which does not utilize a switching mechanism, GC-SwCSV-IVE achieves over 1 dB higher SDR, over 2 dB higher SIR, and nearly 100% accuracy across all microphone configurations.

Notably, the SDR and SIR achieved by GC-SwCSV-IVE with $M = 3$ are close to those of GC-CSV-AuxIVE with $M = 5$. This suggests that the switching mechanism can partially offset performance degradation caused by a smaller number of microphones. Furthermore, increasing the number of total switching states J from 2 to 3 in GC-SwCSV-IVE leads to improved SDR, although a slight drop in SIR indicates minor interference leakage. In contrast, increasing the value of J in SwCSV-IVE without spatial guidance does not necessarily improve the SDR, reflecting the potential risk of introducing more distortions.

Additionally, we vary the number of constrained DOAs (i.e., the value of p) in the GC framework to assess its effect on extraction performance and accuracy, and the results are presented in Fig. 3. As seen, GC-SwCSV-IVE consistently outperforms GC-CSV-AuxIVE across different values of p, and it can approach or reach 100% extraction accuracy in all cases. The advantage becomes more evident when p is smaller.

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



(a) Effect on SDR (dB)

(b) Effect on accuracy (%)

Fig. 3: Effect of the number of constrained DOAs on source extraction performance and accuracy. Condition: $M = 6$.

This demonstrates that GC-SwCSV-IVE is capable of handling complex source movements and maintains robust performance even when the assumed range of DOAs deviates from the actual source trajectory.

V. Conclusions

In this paper, we propose GC-SwCSV-IVE, a method that integrates GC with a state regularization mechanism to address the instability of SwCSV-IVE. Experimental results demonstrate substantial improvements in both extraction performance and accuracy compared to several baseline methods across varying numbers of microphones. Furthermore, the results highlight that the switching mechanism effectively reduces the reliance on a large number of microphones in CSV-based IVE methods.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23K28113.

References

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

[2] P. Comon, “Independent component analysis, a new concept?” *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[3] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, June 2000.

[4] S. Makino, S. Araki, R. Mukai, and H. Sawada, “Audio source separation based on independent component analysis,” in *Proc. ISCAS*, vol. 5, 2004, pp. V-668–V-671.

[5] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomput.*, vol. 22, no. 1-3, pp. 21–34, Nov. 1998.

[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, H. T. Attias, S.-Y. Lee, and T.-W. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Dec. 2006.

[8] 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.

[9] Z. Koldovskỳ and P. Tichavskỳ, “Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence,” *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, Dec. 2018.

[10] R. Scheibler and N. Ono, “Fast independent vector extraction by iterative SINR maximization,” in *Proc. IEEE ICASSP*, 2020, pp. 601–605.

[11] Z. Koldovskỳ, J. Málek, and J. Janskỳ, “Extraction of independent vector component from underdetermined mixtures through block-wise determined modeling,” in *Proc. IEEE ICASSP*, 2019, pp. 7903–7907.

[12] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, “Auxiliary function-based algorithm for blind extraction of a moving speaker,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2022, no. 1, pp. 1–16, Jan. 2022.

[13] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, “Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations,” in *Proc. IEEE ICASSP*, 2019, pp. 7908–7912.

[14] K. Yamaoka, N. Ono, and S. Makino, “Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3461–3475, Nov. 2021.

[15] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, “Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1032–1047, Mar. 2022.

[16] T. Ueda, T. Nakatani, R. Ikeshita, S. Araki, and S. Makino, “DOA-informed switching independent vector extraction and beamforming for speech enhancement in underdetermined situations,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 52, pp. 1–20, Oct. 2024.

[17] Y. Zhao, T. Ueda, and S. Makino, “Moving blind source extraction based on constant separating vector and switching mechanism,” in *Proc. NCSP*, 2025, pp. 53–56.

[18] T. Ueda, T. Nakatani, R. Ikeshita, S. Araki, and S. Makino, “Spatially-regularized switching independent vector analysis,” in *Proc. APSIPA ASC*, 2023, pp. 2024–2030.

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

[20] 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. Signal Process.*, vol. 68, pp. 3545–3558, June 2020.

[21] 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, pp. 41–45.

[22] 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.

[23] 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 ASC*, 2023, pp. 2008–2012.

[24] S. J. Wright, “Coordinate descent algorithms,” *Math. Program.*, vol. 151, no. 1, pp. 3–34, June 2015.

[25] J. Kominek and A. W. Black, “The CMU arctic speech databases,” in *Proc. ISCA SSW*, 2004, pp. 223–224.

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

[27] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[28] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomput.*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[29] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, July 2006.