2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Spatially-Regularized Switching Independent Vector Analysis

Tetsuya Ueda*, Tomohiro Nakatani†, Rintaro Ikeshita†, Shoko Araki†, and Shoji Makino*
* Waseda University, Japan and † NTT Coporation, Japan

***Abstract*—This paper proposes a novel algorithm that uses spatial regularization and a switching filter to enhance the separation accuracy of Independent Vector Analysis (IVA) with specified source permutation based on prior knowledge of sources' Directions-of-Arrival (DOAs). We call this algorithm Spatially-Regularized Switching Independent Vector Analysis (SR-SwIVA). Switching IVA (SwIVA) is a blind source separation (BSS) algorithm that builds on IVA and improves its separation accuracy with a switching filter. However, SwIVA requires relatively reliable estimates of acoustic transfer functions (ATFs) from sources to microphones for its initialization, which are difficult to obtain from DOAs in real reverberant environments. To overcome this limitation, we introduce spatial regularization to SwIVA, which has been shown to be effective for conventional BSS techniques to align permutation (or order) of separated sources with given DOAs. We conduct simulation experiments to show that our proposed SR-SwIVA can significantly improve the source separation accuracy using ATFs estimated from DOAs while completely eliminating the source permutation alignment errors. The achieved source separation accuracy is comparable to that of SwIVA with oracle ATFs.**

I. Introduction

Multichannel audio source separation has been actively studied as preprocessing of various speech applications to improve their performance. To develop effective source separation techniques, researchers often assume that speech signals' Directions of Arrival (DOAs) are available or can be estimated, e.g., from human look directions or using a camera. A promising source separation method for such a situation is Spatially-Regularized Source Separation (SRSS) that utilizes Blind Source Separation (BSS) techniques under regularization by DOA information [1], [2], [3], [4], [5], [6].

Independent Component Analysis (ICA) [7] is a BSS algorithm that estimates separation filters as ones that maximize the independence between the separated signals. To apply ICA to time-frequency domain audio signals, Independent Vector Analysis (IVA) [8], [9], [10], [11] has been proposed; it can group sources separated in different frequencies into individual sources using frequency-independent probabilistic source models. Several algorithms based on IVA have also been developed, including Independent Vector Extraction (IVE) [12], [13] and Convolutional beamformer extension of IVA (CIVA) [14], [15]. IVE can extract signals under the situations where microphones outnumber sources of interest in a computationally efficient way.

While BSS does not aim to determine the sources' order (or permutation) in the separated outputs, SRSS uses DOAs of sources as auxiliary information to align the separated source signals [1], [2], [3], [4], [5], [6]. We call this process source permutation alignment. In concrete, SRSS introduces to the BSS cost function a spatial regularization (also known as a geometric constraint [1]) using Acoustic Transfer Functions (ATFs) from sources to microphones estimated using DOAs (hereafter denoted as DOA-based ATFs). Note that DOA-based ATFs contain substantial estimation errors, e.g., due to reverberation in a real environment. Thus it is hard for classical beamformers like a Minimum Power Distortionless Response (MPDR) beamformer to extract speech signals accurately. In contrast, SRSS (e.g., Spatially-Regularized IVA (SR-IVA) [3], [4], [5], [6]) can estimate filters much more accurately by combining the BSS cost function with spatial regularization.

Although the above-mentioned multichannel audio source separation can separate signals effectively, the separation accuracy is limited, especially when the number of available microphones is small. A switching filter has recently been proposed to overcome this limitation [16], [17], [18], [19], [20]. With a switching filter, we cluster time frames of the microphone signals at each frequency so that each cluster contains relatively a small number of sources. Then, we estimate and apply different separation filters to respective clusters separately. It can achieve more accurate separation in noisy environments than simply applying a single separation filter to a whole captured signal because each cluster contains a smaller number of sources than a whole signal. The switching filter was initially introduced into a beamformer [16], [17] to enable it to handle underdetermined situations. It was then successfully introduced to IVA/CIVA [19] and IVE/CIVE [20] to improve their source separation accuracy. We denote IVA with switching filters as SwIVA, IVE with switching filters as SwIVE, and CIVA with switching filters as SwCIVA hereafter.

Despite the accurate source separation using a switching filter, SwIVA and algorithms based on SwIVA have a problem in that they are often trapped by a bad local optimum unless we carefully initialize their parameters at appropriate values. A reason for the problem to occur is that the permutation of separated sources can differ over different clusters. This problem is called inter-state permutation. To avoid the problem, previous research has proposed Blind Single-State initialization (*BSS init*) and SPatially-Guided initialization (*SPG init*) techniques [19]. Especially, *SPG init* avoids the inter-state permutation problem and aligns the source permutation when using relatively reliable ATFs estimated by a neural network. However, our preliminary experiments showed that using DOA-based ATFs was insufficient for *SPG init* in achieving

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

that.

This paper newly introduces spatial regularization to SwIVA to solve the above-mentioned inter-state permutation problem as well as to align the source permutation only using DOA-based ATFs. We call the algorithm Spatially-Regularized SwIVA (SR-SwIVA)[1]. In this algorithm, we use spatial regularization to align the source permutation in each cluster, thus the algorithm avoids the inter-state permutation problem during its filter optimization. Furthermore, we propose an initialization technique using SR-IVA with no switching mechanism. Because SR-IVA is free from the inter-state permutation problem, we can effectively avoid the problem at the filter initialization of SR-SwIVA. Our experiment shows that when using DOA-based ATFs, SR-SwIVA can perform as accurate source permutation alignment as SR-IVA and as much separation accuracy as SwIVA. In addition, the separation accuracy obtained by SR-SwIVA based on DOA-based ATFs is comparable to that obtained by SwIVA using accurately estimated ATFs.

In the remainder of this paper, we describe the problem formulation in Section II and the baseline algorithms in Section III. Our proposed algorithm is presented in Section IV. Experiment and conclusion are given in Sections V and VI.

II. Problem formulation

Suppose that N source signals are mixed and captured by M ($\geq N$) microphones. We represent observed signals $\boldsymbol{x}(f,t)$ and source signals $\boldsymbol{s}(f,t)$ at each time $t = 1,\ldots,T$ and frequency $f = 1,\ldots,F$ in the Short-Time Fourier Transform (STFT) domain as

$$\boldsymbol{x}(f,t) = [x_1(f,t),\ldots,x_M(f,t)]^{\mathsf{T}} \in \mathbb{C}^M, \tag{1}$$

$$\boldsymbol{s}(f,t) = [s_1(f,t),\ldots,s_N(f,t)]^{\mathsf{T}} \in \mathbb{C}^N, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes the transpose. We model the observed signals $\boldsymbol{x}(f,t)$ as a mixture of signals defined by

$$\boldsymbol{x}(f,t) = \sum_{n=1}^{N} \boldsymbol{d}_n(f,t) + \boldsymbol{r}(f,t), \tag{3}$$

$$\boldsymbol{d}_n(f,t) = \boldsymbol{h}_n(f)s_n(f,t), \tag{4}$$

where $\boldsymbol{d}_n(f,t) = [d_{n,1}(f,t),\ldots,d_{n,M}(f,t)] \in \mathbb{C}^M$ is the microphone image of the nth source and $\boldsymbol{h}_n(f)$ is the transfer function of the nth source. $\boldsymbol{r}(f,t)$ is the diffuse noise composed of the background noise.

We also suppose that we have an estimate of the transfer functions (hereafter called a steering vector). Assuming that the DOAs of the sources and the locations of microphones are given or estimated, we obtain a steering vector $\boldsymbol{a}_n(f)$ based on the plane wave assumption and the relative time-delay-of-arrival (TDOA) $\boldsymbol{\tau}_n \in \mathbb{R}^M$ from the nth source to M microphones:

$$\begin{aligned}\boldsymbol{a}_n(f) &= [a_{n,1}(f),\ldots,a_{n,M}(f)]^{\mathsf{T}} \in \mathbb{C}^M \\ &= \frac{1}{\sqrt{M}} \exp\left(-2\pi \frac{f}{N_{\mathrm{F}}} \boldsymbol{\tau}_n \sqrt{-1}\right) \text{ for } 1 \leq n \leq N. \end{aligned} \tag{5}$$

[1]Because it is straightforward to incorporate spatial regularization into SwIVE and SwCIVA, this paper omits this explanation for conciseness.

where each element $a_{n,m}(f)$ is an estimate of a transfer function from the nth source to the mth microphone and N_{F} is the length of Fourier transform. Note that $\boldsymbol{a}_n(f)$ contains substantial errors from the true transfer function $\boldsymbol{h}_n(f)$ in (4) because $\boldsymbol{a}_n(f)$ only contains a direct path component but does not include reflection paths.

The aim of this paper is to develop an effective algorithm to estimate a microphone image of each source at the first microphone as $[\hat{s}_1(f,t),\ldots,\hat{s}_N(f,t)]^{\mathsf{T}}$ given $\boldsymbol{x}(f,t)$ and $\{\boldsymbol{a}_n(f)\}_{n=1}^N$. It is important to note that we need to align the estimated sources with the same permutation as the sources in (2) during the estimation. We call this process source permutation alignment.

III. Baseline algorithms

This section briefly describes two of our baseline algorithms, SR-IVA and SwIVA.

A. Spatially-Regularized IVA (SR-IVA) [3], [4], [5], [6]

SR-IVA is an algorithm to estimate signals $[\hat{s}_1(f,t),\ldots,\hat{s}_N(f,t)]^{\mathsf{T}}$ with the same source permutation as in (2) by applying separation matrices $\{\boldsymbol{W}(f)\}_f$ to the observed signal:

$$\hat{\boldsymbol{s}}(f,t) = \boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,t), \tag{6}$$

where

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f),\ldots,\boldsymbol{w}_M(f)] \in \mathbb{C}^{M\times M}, \tag{7}$$

$$\hat{\boldsymbol{s}}(f,t) = [\hat{s}_1(f,t),\ldots,\hat{s}_M(f,t)]^{\mathsf{T}} \in \mathbb{C}^M. \tag{8}$$

Note that $\hat{s}_{N+1}(f,t),\ldots,\hat{s}_M(f,t)$ correspond to noise components. With BSS techniques such as IVA, we can obtain $\boldsymbol{W}(f)$ based on the maximum likelihood estimation assuming the statistical independence between each source $\hat{s}_n(f,t)$. However, we cannot determine the source permutation uniquely based only on the independence between sources. Thus, SR-IVA introduces a spatial regularization term to force each separation filter $\boldsymbol{w}_n(f)$ to extract a signal corresponding to $\boldsymbol{a}_n(f)$. The following is an example of a regularization term [5], which we use throughout this paper.

$$\mathcal{L}_{\mathrm{regl}}(\{\boldsymbol{W}(f)\}_f) = \sum_{f=1}^{F}\sum_{n=1}^{N} \|\boldsymbol{w}_n(f) - \boldsymbol{a}_n(f)\|_2^2, \tag{9}$$

where $\|\cdot\|_2$ stands for l-2 norm. This regularization term enforces the nth separation filter $\boldsymbol{w}_n(f)$ to have a value close to the nth steering vector $\boldsymbol{a}_n(f)$. SR-IVA estimates signals $\hat{s}_n(f,t)$ by combining the likelihood function of BSS with the spatial regularization, thus simultaneously realizing source separation and source permutation alignment.

B. Switching IVA (SwIVA) [19]

Switching filter has recently been proposed to improve source separation accuracy. This section gives an overview of switching IVA (SwIVA). Fig. 1 shows the processing flow of SwIVA.

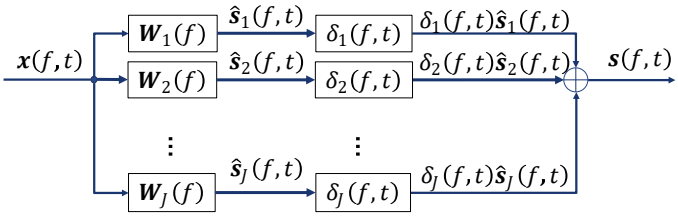


Fig. 1: Processing flow of switching IVA

SwIVA uses a multi-state separation model. It applies $J \geq 2$ separation matrices, $\boldsymbol{W}_j(f)$ for $1 \leq j \leq J$, to the observed signal to yield J different sets of separated sources $\hat{\boldsymbol{s}}_j(f,t) \in \mathbb{C}^M$:

$$\hat{\boldsymbol{s}}_j(f,t) = \boldsymbol{W}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,t) \text{ for } 1 \leq j \leq J, \tag{10}$$

$$\boldsymbol{W}_j(f) = [\boldsymbol{w}_{j,1}(f), \ldots, \boldsymbol{w}_{j,M}(f)] \in \mathbb{C}^{M \times M}. \tag{11}$$

Here j is an index of a switching state and J is the number of the switching states (or the number of separation matrices). Then, it uses a switch to select one of J separated sources at each time frequency and yields its final output:

$$\hat{\boldsymbol{s}}(f,t) = \sum_{j=1}^{J} \delta_j(f,t)\hat{\boldsymbol{s}}_j(f,t), \tag{12}$$

$$\sum_{j=1}^{J} \delta_j(f,t) = 1 \quad \text{and} \quad \delta_j(f,t) \in \{0,1\}. \tag{13}$$

The switch mechanism is implemented using a time-frequency dependent switching weight $\{\delta_j(f,t)\}_{j,f,t}$ in (12). In this paper, we only consider hard switches and allow $\delta_j(f,t)$ to take only binary values, 0 or 1. It has been reported that SwIVA achieved more accurate source separation than IVA that applies a single set of separation matrix in (6).

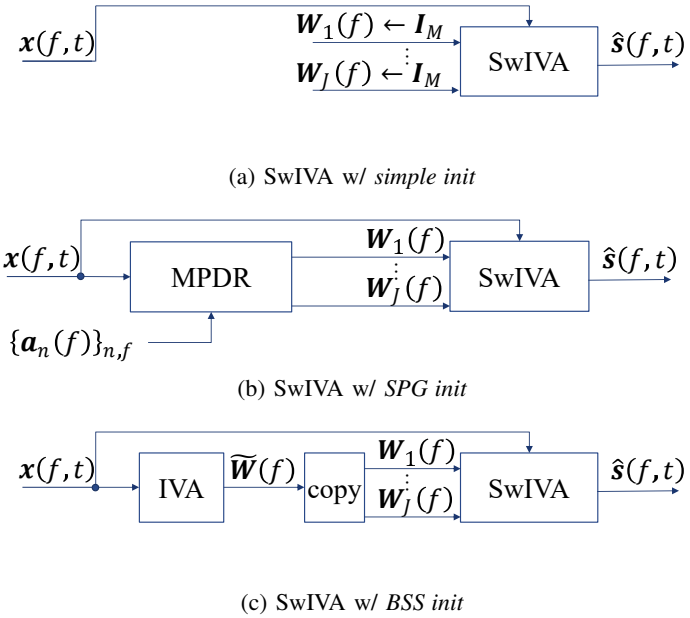
One important issue with SwIVA is that the separation matrices $\{\boldsymbol{W}_j(f)\}_j$ are often trapped by a bad local optimum that yields relatively less accurate source separation. This problem occurs because the permutations of separated sources can differ over different states. Thus, the order of sources changes when the switching state changes, resulting in large estimation errors. We call this problem inter-state permutation problem.

Previous research has proposed several initialization techniques for SwIVA to avoid the inter-state permutation problem [19]. We explain three of them in the following using Fig. 2.

1) Simple initialization (simple init): In this initialization, we set each separation matrix $\boldsymbol{W}_j(f)$ as an $M \times M$ identity matrix $\boldsymbol{I}_M$ as shown in Fig. 2a:

$$\boldsymbol{W}_j(f) \leftarrow \boldsymbol{I}_M \text{ for } 1 \leq j \leq J \text{ and } 1 \leq f \leq F. \tag{14}$$

It is reported that *simple init* does not avoid the inter-state permutation problem well for SwIVA.



(a) SwIVA w/ *simple init*

(b) SwIVA w/ *SPG init*

(c) SwIVA w/ *BSS init*

Fig. 2: Flow of SwIVA with three different initialization techniques

2) Spatially-guided initialization (SPG init): In this initialization, we employ the conventional MPDR beamformer to initialize the separation matrices $\boldsymbol{W}_j(f)$ as shown in Fig. 2b. In concrete, after randomly initializing $\delta_j(f,t)$, each filter $\boldsymbol{w}_{j,n}(f)$ is initialized:

$$\boldsymbol{w}_{j,n}(f) = \begin{cases} \frac{a^*_{n,1}(f)\boldsymbol{\Xi}_j^{-1}(f)\boldsymbol{a}_n(f)}{\boldsymbol{a}_n^{\mathsf{H}}(f)\boldsymbol{\Xi}_j^{-1}(f)\boldsymbol{a}_n(f)} & \text{for } 1 \leq n \leq N, \\ \boldsymbol{e}_n, & \text{for } N+1 \leq n \leq M, \end{cases} \tag{15}$$

where

$$\boldsymbol{\Xi}_j(f) = \frac{\sum_t \delta_j(f,t)\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)}{\sum_t \delta_j(f,t)}, \tag{16}$$

is a sample covariance matrix using time-frequency bins specified by the jth switch $\delta_j(f,t)$. $(\cdot)^*$ denotes complex conjugate, and $\boldsymbol{e}_n$ denotes the nth column of $\boldsymbol{I}_M$. The effectiveness of *SPG init* has been shown when we can obtain relatively accurate ATFs using a neural network.

3) Blind single-state initialization (BSS init): In this initialization, we use IVA to initialize the separation matrices $\boldsymbol{W}_j(f)$ as shown in Fig. 2c. Let $\tilde{\boldsymbol{W}}(f) \in \mathbb{C}^{M \times M}$ be a separation matrix estimated by IVA (or SwIVA with $J = 1$), then we initialize $\boldsymbol{W}_j(f)$ by $\tilde{\boldsymbol{W}}(f)$:

$$\boldsymbol{W}_j(f) \leftarrow \tilde{\boldsymbol{W}}(f) \text{ for } 1 \leq j \leq J \text{ and } 1 \leq f \leq F, \tag{17}$$

and randomly initialize $\delta_j(f,t)$. *BSS init* can avoid the inter-state permutation but is not intended to align the source permutation.

C. Drawbacks of SwIVA

Although using *SPG init* for SwIVA has been shown effective in avoiding the inter-state permutation problem when accurately estimated ATFs are available, it is not certain whether we can correctly align the source permutation based on only the initialization. In addition, our preliminary experiments showed that *SPG init* was insufficient when ATFs are estimated based on DOAs. This is probably because DOA-based ATFs contain substantial errors when estimated in real reverberant environments.

IV. Proposed algorithm: Spatially-Regularized SwIVA (SR-SwIVA)

This section proposes Spatially-Regularized SwIVA (SR-SwIVA) that controls SwIVA using DOA-based ATFs. To solve the problem discussed in Section III-C, we utilize DOA-based spatial regularization, which was shown effective for SR-IVA to align the source permutation. We first explain the estimation model and parameter optimization of SR-SwIVA in Section IV-A, and then present a new initialization technique for SR-SwIVA in Section IV-B. We expect that the algorithm effectively avoids the inter-state permutation problem and aligns the source permutation at the same time.

A. Estimation model and parameter optimization

First, we explain the estimation model and develop the parameter optimization of SR-SwIVA. SR-SwIVA uses the same separation model as SwIVA in (10). Then, similar to SR-IVA, we define the cost function as the sum of the negative log likelihood function $\mathcal{L}_{\text{NL}}(\mathcal{X};\Theta)$ used for SwIVA [19] and a spatial regularization term $\mathcal{L}_{\text{regl}}(\mathcal{W})$ that we design for SR-SwIVA:

$$\mathcal{L}(\Theta) = \mathcal{L}_{\text{NL}}(\mathcal{X};\Theta) + \lambda_{\text{regl}}\mathcal{L}_{\text{regl}}(\mathcal{W}). \tag{18}$$

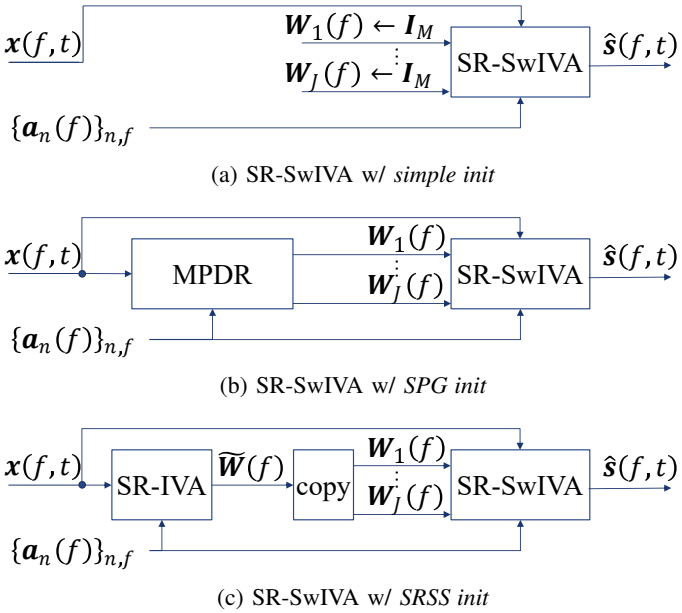
$$\mathcal{L}_{\text{NL}}(\mathcal{X};\Theta) = \sum_{j,f,t}\delta_j(f,t)\left\{\sum_{n=1}^{M}\left(\log v_n(f,t) + \frac{|\hat{s}_{j,n}(f,t)|^2}{v_n(f,t)}\right) - 2\log\left|\det \boldsymbol{W}_j(f)\right|\right\} \tag{19}$$

where $\hat{s}_{j,n}(f,t)$ is the nth element of $\hat{\boldsymbol{s}}_j(f,t)$ in (10), $\mathcal{X} = \{\boldsymbol{x}(f,t)\}_{f,t}$ is the set of observed signals, and $\Theta = \{\mathcal{D},\mathcal{W},\mathcal{V}\}$ is a set of model parameters composed of switching weights $\mathcal{D} = \{\delta_j(f,t)\}_{j,f,t}$, separation matrices $\mathcal{W} = \{\boldsymbol{W}_j(f)\}_{j,f}$, and the time-frequency dependent variance of the sources[2] $\mathcal{V} = \{v_n(f,t)\}_{n,f,t}$.

As for the spatial regularization term, $\mathcal{L}_{\text{regl}}(\mathcal{W})$, this paper adopts almost the same definition as that introduced for SR-IVA in (9). A major difference from SR-IVA is that we duplicate the same term to regularize the separation filters in all states. In concrete, the regularization term is formulated:

$$\mathcal{L}_{\text{regl}}(\mathcal{W}) = \sum_{f=1}^{F}\sum_{j=1}^{J}\sum_{n=1}^{N}\|\boldsymbol{w}_{j,n}(f) - \boldsymbol{a}_n(f)\|_2^2. \tag{20}$$

[2]Following the previous work for SwIVA [19], we use the frequency-independent source model only for updating the separation matrices as a practical technique to group the separated sources over different frequencies.



(a) SR-SwIVA w/ *simple init*

(b) SR-SwIVA w/ *SPG init*

(c) SR-SwIVA w/ *SRSS init*

Fig. 3: Flow of SR-SwIVA with three different initialization techniques

This term forces separation filters $\boldsymbol{w}_{j,n}(f)$ that share the same source index n at different states $1 \le j \le J$ to be close to the same steering vector of the nth source $\boldsymbol{a}_n(f)$. Therefore, as long as the spatial regularization successfully aligns the source permutation in the same way as SR-IVA, all the filters $\boldsymbol{w}_{j,n}(f)$ are expected to extract the same nth source for $1 \le j \le J$. As a consequence, we can also solve the inter-state permutation problem and align the source permutation.

The parameter optimization of SR-SwIVA can be straightforwardly derived by combining the existing algorithms (i.e., SR-IVA [5] and SwIVA [19]). We describe the derivation in Appendix A.

B. Spatially-Regularized single-state initialization (SRSS init)

Next, we introduce the DOA-based spatial regularization to the initialization for SR-SwIVA, referred to as Spatially-Regularized Single-State initialization (*SRSS init*). The aim of using *SRSS init* is to improve the estimation accuracy of SR-SwIVA better than *SPG init* that was used for SwIVA.

Fig. 3 shows flows of SR-SwIVA with three different initialization techniques that we compare in our experiments. In Figs. 3a and 3b, we use the same initialization techniques used in Figs. 2a and 2b. Fig. 3c shows our proposed *SRSS init*, which is a modification of *BSS init* in Fig. 2c for SwIVA. In *SRSS init*, we use SR-IVA (or SR-SwIVA with $J = 1$) instead of IVA to initialize all the parameters of SR-SwIVA. Because SR-IVA aligns the source permutation effectively, it can be a desirable initialization for SR-SwIVA using DOA-based ATFs.

V. EXPERIMENT

In this section, we compare the source separation and alignment accuracy of IVA, SwIVA, and SR-SwIVA combined with various initialization techniques, including *SPG init* and *SRSS init*. Our experiment shows the effectiveness of SR-SwIVA when using DOA-based ATFs.

A. Conditions of experiment

We conducted an experiment using TIMIT-ConvMix, which is composed of simulated noisy reverberant mixtures. To generate the mixtures, we first concatenated utterances extracted from the TIMIT corpus [21] to obtain a set of single-speaker clean utterance sequences, each of which is 30 s long. Then we mixed three utterance sequences and five different additive noise signals extracted from the CHiME-3 dataset [22] after individually reverberating them. We randomly selected 8 Room Impulse Responses (RIRs) out of 9 RIRs from the RWCP E2A dataset [23], corresponding to the azimuths of $10^\circ, 30^\circ, 50^\circ, \ldots, 170^\circ$. We then performed the convolution by assigning the RIRs to the utterances and noises. Here, we set the minimum angle difference between each utterance at 40°. RT60 of the RIRs was 0.3 s. We used four microphones labeled in the dataset: the 21st, 22nd, 23rd, and 24th. All four types of noise signals were used, BUS, STR, PED, and CAF, although each mixture contained only a single type. We set the power ratio of each reverberant utterance sequence to the sum of the additive noise signals to 10 dB.

We used a Hann window for a short-time analysis, where frame length N_{F} and shift size were set to 64 ms and 32 ms, respectively. The sampling frequency f_{s} was set to 16 kHz. We initialized switching weights $\delta_j(f,t)$ at a random value in a range of 1 ± 10^{-3}, and normalized it to satisfy $\sum_{j=1}^{J} \delta_j(f,t) = 1$. Note that after initialization, the weights $\delta_j(f,t)$ are updated to binary values during the parameter optimization. For all algorithms, including IVA, SR-IVA, SwIVA, and SR-SwIVA, we updated the separation filters 50 times and applied projection back [24] post-processing to solve the scale ambiguity.

Because we used linear arrays, we set relative TDOA $\boldsymbol{\tau}_n = [\tau_{n1}, \ldots, \tau_{nM}]$ in (5) for the DOA-based steering vectors:

$$\tau_{nm} = f_{\mathrm{s}} \frac{d(m-1)}{c} \cos\left(2\pi \frac{\theta_n}{360^\circ}\right), \tag{21}$$

where $c = 343$ m/s is the speed of sound and $d = 0.0281$ meter is the distance between adjacent microphones. We used the angle labels for RIRs in the RWCP dataset as the speakers' directions θ_n. Note that even using the angle labeled at the dataset, the steering vector in (5) should contain substantial errors because it does not include the effects of reflection paths. To confirm the performance using oracle information, we also evaluated the source separation accuracy of SRSS algorithms when replacing DOA-based ATFs $\boldsymbol{a}_n(f)$ with oracle steering vectors $\boldsymbol{h}_n(f)$ in (4). We created $\boldsymbol{h}_n(f)$ as the primary eigenvector of the spatial covariance matrix of the noiseless reverberant source image of $s_n(f,t)$.

In this evaluation, we adopted signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) as the metrics of source separation accuracy [25], which is widely used in source separation research. We used the MUSEVAL V4 toolkit [26] with its bss_eval_images configuration. As reference signals, we used clean utterance sequences that were convolved with the RIRs for generating the corresponding mixtures. We evaluated SDRs and SIRs using a correct permutation of estimated sources, not depending on the actual permutation estimated by each separation algorithm. We determined the correct permutation as the order of estimated sources that gave the best SIR score with the reference signals in their original order. In addition to SDR and SIR, we evaluated the correctness of the estimated source permutation alignment using the permutation error (permE) defined.

$$\text{permE} = \frac{\text{\# of mixtures separated with incorrect permutation}}{\text{Total \# of mixtures (=100)}}. \tag{22}$$

The estimated source permutation alignment was determined to be incorrect when it was not identical to the correct source permutation.

B. Result

We compared the SDR improvement, SIR improvement, and permE for each source separation algorithm in Table I. We conducted two experiments, one with three microphones and the other with four microphones, and show the results in Tables Ia and Ib. For each experiment, we set the regularization weight λ_{regl} as a value, with which SR-SwIVA achieved the highest SDRs while reducing permE to zero.

We first show the result of the baseline algorithms, SR-IVA and SwIVA. In both Tables Ia and Ib, although SR-IVA ($J = 1$) reduced permE to zero and achieved comparable SDRs and SIRs as IVA, SwIVA with *SPG init* showed 5 % permE when $M = 3$. Moreover, SwIVA with *SPG init* showed 100 % permE when $M = 4$. This is because DOA-based ATFs contained a lot of errors, as shown by the low source separation accuracy (3.83 dB SIR improvement) by MPDR beamformer. The above results show that when using DOA-based ATFs, SwIVA with *SPG init* cannot correctly align separated sources.

Next, we show the result of SR-SwIVA with several initialization techniques presented in Fig. 3 using the same tables. From the tables, SR-SwIVA could not achieve 0 % permE when using *simple init* or *SPG init* except for using *simple init* with $M = 4$. Also, their SDRs are comparable to or even worse than the baseline algorithms. On the other hand, SR-SwIVA with *SRSS init* achieved the highest SDR and 0 % permE when using DOA-based ATFs. Thus, we conclude that SR-SwIVA with *SRSS init* can separate signals and align the source permutation more effectively than the baseline algorithms.

Finally, we show the results using oracle ATFs. In both tables, the MPDR beamformer worked much better using oracle ATFs than using DOA-based ATFs, and SwIVA with *SPG init* attained high SDR and SIR and reduced permE to 0 %. It means that *SPG init* is effective for SwIVA only when

TABLE I: The improvements of SDR (SDRi), SIR (SIRi) [dB], and permutation error (permE) [%]. SDRs and SIRs were calculated using a correct source permutation that achieved the highest SIRs among all possible source permutations.

(a) $N = 3, M = 3, \lambda_{\text{regl}} = 0.4$

	initialization	J	SDRi	SIRi	permE
Blind Source Separation					
IVA [10]	-	1	7.73	19.97	87.5
SwIVA [19]	*BSS init*	2	8.55	22.43	87.5
DOA-based ATFs informed Source Separation					
MPDR	-	-	3.21	12.26	0.0
SR-IVA [3, .etc]	-	1	7.68	19.77	0.0
SwIVA [19]	*SPG init*	2	8.28	21.78	5.0
SR-SwIVA	*simple init*	2	7.99	20.64	7.5
	SPG init	2	7.34	19.44	6.0
	SRSS init	2	8.52	22.38	0.0
Oracle ATFs informed Source Separation					
MPDR	-	-	6.55	15.01	0.0
SwIVA [19]	*SPG init*	2	8.67	22.73	0.0
SR-SwIVA	*SRSS init*	2	8.50	22.32	5.0

(b) $N = 3, M = 4, \lambda_{\text{regl}} = 1.0$

	initialization	J	SDRi	SIRi	permE
Blind Source Separation					
IVA [10]	-	1	8.67	23.65	87.5
SwIVA [19]	*BSS init*	2	9.27	25.33	100.0
DOA-based ATFs informed Source Separation					
MPDR	-	-	2.31	3.83	80.0
SR-IVA [3, .etc]	-	1	8.70	23.63	0.0
SwIVA [19]	*SPG init*	2	8.75	23.52	100.0
SR-SwIVA	*simple init*	2	8.81	23.47	0.0
	SPG init	2	8.93	24.28	37.5
	SRSS init	2	9.30	25.31	0.0
Oracle ATFs informed Source Separation					
MPDR	-	-	7.63	18.22	0.0
SwIVA [19]	*SPG init*	2	9.05	24.47	0.0
SR-SwIVA	*SRSS init*	2	9.27	25.22	0.0

we have relatively accurate ATFs. In contrast, SR-SwIVA with *SRSS init* can achieve the same high scores regardless of using oracle ATFs or DOA-based ATFs. Thus we showed that SR-SwIVA could achieve as much accurate source separation and source permutation alignment as SwIVA based on relatively accurate ATFs.

VI. Conclusions

In this paper, we have proposed a novel algorithm called SR-SwIVA that uses spatial regularization to enhance SwIVA for accurate source separation and permutation alignment based on DOA-based ATFs. We have shown that DOA-based ATFs have significant estimation errors in real reverberant environments, which prevent conventional SwIVA from achieving high separation accuracy and correct permutation alignment. To address this challenge, we have developed SR-SwIVA by incorporating spatial regularization into SwIVA and presented a new effective initialization technique called *SRSS init* that uses DOA-based ATFs. We have conducted an experiment using noisy reverberant sound mixtures and demonstrated that our proposed SR-SwIVA can significantly improve the SDRs and SIRs compared to our baseline algorithms, SR-IVA and SwIVA, with DOA-based ATFs, and eliminate the source permutation alignment errors completely. Moreover, we have shown that the source separation accuracy achieved by SR-SwIVA is similar to that of SwIVA with oracle ATFs.

Acknowledgment

This work was supported by JSPS KAKENHI 23H03423 and JST SPRING JPMJSP2128.

Appendix A. parameter optimization of SR-SwIVA

We optimize the parameters Θ for SR-swIVA by minimizing the cost function in (18). Because no closed form solution has been obtained, we use iterative estimation based on a coordinate descent method. It alternately updates one of $\mathcal{W}$, $\mathcal{D}$, and $\mathcal{V}$ by fixing the other parameters, and iterates the update until convergence is obtained.

Because the update of $\mathcal{D}$ and $\mathcal{V}$ is not related with the regularization term, we can update $\mathcal{D}$ and $\mathcal{V}$ using the same update rules written in [19]. Hereafter we only explain the update of $\mathcal{W}$. For the update of the separation matrices $\mathcal{W}$, we adopt a frequency-independent source model following a practical technique proposed in [19] to group separated sources over different frequencies. In concrete, instead of using the frequency-dependent variance $v_n(f,t)$, we use a frequency-independent variance $v_n(t)$ for the update, which is obtained by $v_n(t) \leftarrow \frac{1}{F}\sum_{f=1}^{F} v_n(f,t)$. Then, fixing $\delta_j(f,t)$ and $v_n(t)$, we can rewrite the function $\mathcal{L}(\Theta)$ in (18) as

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_j(f)) \overset{c}{=} &-2\log\left|\det \boldsymbol{W}_j(f)\right| \\ &+ \sum_{n=1}^{M}\big(\boldsymbol{w}_{j,n}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f)\boldsymbol{w}_{j,n}(f) \\ &- \boldsymbol{w}_{j,n}^{\mathsf{H}}(f)\boldsymbol{p}_n(f) - \boldsymbol{p}_n^{\mathsf{H}}(f)\boldsymbol{w}_{j,n}(f)\big), \end{aligned} \tag{23}$$

where

$$\boldsymbol{\Pi}_{j,n}(f) = \begin{cases} \boldsymbol{\Sigma}_{j,n}(f) + \lambda_{\text{regl}}\boldsymbol{I}_M & \text{for } 1 \le n \le N, \\ \boldsymbol{\Sigma}_{j,n}(f) & \text{for } N+1 \le n \le M, \end{cases} \tag{24}$$

$$\boldsymbol{p}_n(f) = \begin{cases} \lambda_{\text{regl}}\boldsymbol{a}_n(f) & \text{for } 1 \le n \le N, \\ \boldsymbol{0}_M & \text{for } N+1 \le n \le M, \end{cases} \tag{25}$$

$$\boldsymbol{\Sigma}_{j,n}(f) = \frac{1}{T_j(f)}\sum_{t=1}^{T}\frac{\delta_j(f,t)}{v_n(t)}\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t), \tag{26}$$

$$T_j(f) = \sum_{t=1}^{T}\delta_j(f,t), \tag{27}$$

and $\boldsymbol{0}_M \in \mathbb{C}^M$ is a zero vector. Because the above objective has the same form as SR-IVA [4], [5], we can apply iterative optimization techniques proposed for it, such as Vectorwise Coordinate Decent (VCD) [2], and Iterative Source Steering [11], [6]. This paper employs VCD that updates $\boldsymbol{W}_j(f)$ with a sequence of $\boldsymbol{w}_{j,1}(f) \rightarrow \boldsymbol{w}_{j,2}(f), \ldots, \rightarrow \boldsymbol{w}_{j,M}(f)$ for $j =$

$1, \ldots, J$:

$$\boldsymbol{u}_{j,n}(f) = (\boldsymbol{W}_j^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f))^{-1}\boldsymbol{e}_n, \tag{28}$$

$$\tilde{\boldsymbol{u}}_{j,n}(f) = \boldsymbol{\Pi}_{j,n}^{-1}(f)\boldsymbol{p}_n(f), \tag{29}$$

$$h_{j,n}(f) = \boldsymbol{u}_{j,n}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f)\boldsymbol{u}_{j,n}(f), \tag{30}$$

$$\hat{h}_{j,n}(f) = \boldsymbol{u}_{j,n}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,n}(f)\tilde{\boldsymbol{u}}_{j,n}(f), \tag{31}$$

$$\tilde{h}_{j,n}(f) = \frac{\hat{h}_{j,n}(f)}{2h_{j,n}(f)}\left[-1 + \sqrt{1 + \frac{4h_{j,n}(f)}{|\hat{h}_{j,n}(f)|^2}}\right], \tag{32}$$

$$\boldsymbol{w}_{j,n}(f) = \begin{cases} \frac{1}{\sqrt{h_{j,n}(f)}}\boldsymbol{u}_{j,n}(f) + \tilde{\boldsymbol{u}}_{j,n}(f) & \text{if } \hat{h}_{j,n}(f) = 0, \\ \hat{h}_{j,n}(f)\boldsymbol{u}_{j,n}(f) + \tilde{\boldsymbol{u}}_{j,n}(f) & \text{otherwise.} \end{cases} \tag{33}$$

REFERENCES

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

[2] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *Proc. ICASSP*, 2018, pp. 746–750.

[3] A. H. Khan, M. Taseska, and E. A. Habets, "A geometrically constrained independent vector analysis algorithm for online source extraction," in *Proc. LVA/ICA*, 2015, pp. 396–403.

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

[5] 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 Processing*, vol. 68, pp. 3545–3558, 2020.

[6] 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.

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

[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2006.

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

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

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

[12] Z. Koldovsky and P. Tichavsky, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Processing*, vol. 67, no. 4, pp. 1050–1064, 2018.

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

[14] 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.

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

[16] K. Yamaoka, A. Brendel, N. Ono, S. Makino, M. Buerger, T. Yamada, and W. Kellermann, "Time-frequency-bin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios," in *Proc. EUSIPCO*, 2018, pp. 1582–1586.

[17] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 3461–3475, 2021.

[18] R. Ikeshita, N. Kamo, and T. Nakatani, "Blind signal dereverberation based on mixture of weighted prediction error models," *IEEE Signal Processing Letters*, vol. 28, pp. 399–403, 2021.

[19] 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. ASLP*, vol. 30, pp. 1032–1047, 2022.

[20] ——, "Switching independent vector extraction and its joint optimization with weighted prediction error dereverberation," in *Proc. ICA*, 2022.

[21] L. Consortium *et al.*, "Timit acoustic-phonetic continuous speech corpus," 1993.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[23] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," 2000.

[24] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, 2001.

[25] 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.

[26] "Museval," *https://github.com/sigsep/sigsep-mus-eval*.