Ueda *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:52
https://doi.org/10.1186/s13636-024-00373-3

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

DOA-informed switching independent vector extraction and beamforming for speech enhancement in underdetermined situations

Tetsuya Ueda[1*], Tomohiro Nakatani[2], Rintaro Ikeshita[2], Shoko Araki[2] and Shoji Makino[1]

Abstract

This paper proposes novel methods for extracting a single Speech signal of Interest (SOI) from a multichannel observed signal in underdetermined situations, i.e., when the observed signal contains more speech signals than microphones. It focuses on extracting the SOI using prior knowledge of the SOI's Direction of Arrival (DOA). Conventional beamformers (BFs) and Blind Source Separation (BSS) with spatial regularization struggle to suppress interference speech signals in such situations. Although Switching Minimum Power Distortionless Response BF (Sw-MPDR) can handle underdetermined situations using a switching mechanism, its estimation accuracy significantly decreases when it relies on a steering vector determined by the SOI's DOA. Spatially-Regularized Independent Vector Extraction (SRIVE) can robustly enhance the SOI based solely on its DOA using spatial regularization, but its performance degrades in underdetermined situations. This paper extends these conventional methods to overcome their limitations. First, we introduce a time-varying Gaussian (TVG) source model to Sw-MPDR to effectively enhance the SOI based solely on the DOA. Second, we introduce the switching mechanism to SRIVE to improve its speech enhancement performance in underdetermined situations. These two proposed methods are called Switching weighted MPDR (Sw-wMPDR) and Switching SRIVE (Sw-SRIVE). We experimentally demonstrate that both surpass conventional methods in enhancing the SOI using the DOA in underdetermined situations.

Keywords Speech enhancement, Underdetermined situations, Switching mechanism, Time-varying Gaussian distribution

1 Introduction

This paper addresses multichannel speech enhancement methods that can extract a single Speech signal of Interest (SOI) from multiple microphone inputs when the inputs may be contaminated by an unknown number of interference speech signals. This task becomes particularly challenging in underdetermined situations where the captured speech signals outnumber the microphones. In such a case, the SOI's Direction of Arrival (DOA) serves as a useful clue for identifying the SOI. We refer to speech enhancement based on the DOA as DOA-informed speech enhancement. One scenario for applying DOA-informed speech enhancement is recording an SOI using a limited number of microphones attached to a mobile device, such as a smartphone or smart glasses, and estimating the DOA from a camera embedded in the device. The objective of this paper is to develop a superior DOA-informed speech enhancement technique in underdetermined situations.

Two major approaches can be used for DOA-informed speech enhancement: beamforming [1, 2] and Blind Source Separation (BSS) with spatial regularization [3].

*Correspondence:
Tetsuya Ueda
t.ueda@ieee.org
[1] Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kita-Kyushu 808-0135, Fukuoka, Japan
[2] NTT Corporation, Kyoto, Japan

Beamforming is a useful approach for DOA-informed speech enhancement [2]. Among various beamformers (BFs), a Minimum Variance Distortionless Response BF (MVDR) [1] is one state-of-the-art example. It estimates a spatial filter that can suppress signal space dominated by interference speech signals. The estimated filter keeps the SOI unchanged according to a distortionless constraint based on given Acoustic Transfer Functions (ATFs) from the SOI to microphones. Two methods are widely used to estimate the ATFs: one using SOI's DOA based on the plane-wave assumption (hereafter denoted as DOA-based ATFs) [1, 2] and the other using voice activity detection [4]. This paper adopts the former method because the latter method struggles to detect the SOI's voice activity when there are several interference speakers. To estimate the spatial filter, however, MVDR requires Spatial Covariance Matrices (SCMs) of the interferences (interference SCMs). When the interferences contain speech signals, it is challenging to accurately estimate the interference SCMs by distinguishing the interferences from the SOI in the mixture.

In contrast, a Minimum Power Distortionless Response BF (MPDR) [1] can estimate spatial filters without using the interference SCMs. MPDR estimates a filter that minimizes the power of the microphone observation while retaining the SOI using a distortionless constraint similar to MVDR.

Although MPDR performs speech enhancement without interference SCMs, their interference suppression is limited in underdetermined situations [5, 6]. For example, MPDR struggles to effectively suppress all the $N-1$ interference speech signals when we have only $M(<N)$ microphones, where N is the number of speech signals. Researchers have proposed a switching mechanism [5–10] to overcome this limitation. With the mechanism, we cluster time frames of the microphone signals so that each cluster contains a relatively small number of sources. Then, we estimate and apply different spatial filters to the respective clusters separately. This can also be interpreted as using a time-varying filter, where we estimate and apply different spatial filters to individual time frames based on the clustering results. This approach can achieve interference suppression more effectively than simply applying a single time-invariant filter to a whole captured signal because each cluster contains fewer sources than a whole signal. The switching mechanism was originally introduced into MPDR to handle underdetermined situations [6]. This paper refers to the method as switching MPDR (Sw-MPDR).

One serious drawback of Sw-MPDR [6] (and MPDR) is that it is sensitive to errors in the estimated ATFs. In particular, the speech enhancement performance degrades when the ATFs are estimated using the SOI's DOA. No matter how accurate the DOA is, the DOA-based ATFs contain only direct path components. Thus, such BFs may suppress the early reflections of the SOI, even if we use a distortionless constraint. Since early reflections amplify the SOI in the observations, largely reducing these reflections can lead to a substantial degradation in speech enhancement. To mitigate this degradation, a BF must be developed that effectively performs speech enhancement even with DOA-based ATFs. Hereafter, we deal with the lack of multipath components in DOA-based ATFs as the modeling errors of ATFs.

In contrast, it may be worth noting that a variation of MPDR, weighted MPDR (wMPDR) [11–13], can perform speech enhancement better than MPDR even with the modeling errors of ATFs. wMPDR estimates a spatial filter so that its outputs follow a time-varying Gaussian (TVG) with time-varying variances based on the Maximum Likelihood Estimation (MLE). The MLE used by wMPDR is shown equivalent to minimizing the power of the observed signal normalized by the time-varying variance of the SOI. Accordingly, wMPDR can estimate the spatial filter that mainly reduces the interferences appearing in the SOI-absent periods while retaining the SOI and its reflections.

On the other hand, we can create a DOA-informed speech enhancement method by introducing spatial regularization to BSS. BSS is a technique that separates individual source signals from microphone observations without prior information of the signals or the room acoustics. In contrast, spatial regularization guides BSS to separate the SOIs corresponding to specified DOAs.

Independent Component Analysis (ICA) [14] is a BSS algorithm that estimates spatial filters as those that maximize the independence between separated signals. To apply ICA to time-frequency domain audio signals, Independent Vector Analysis (IVA) [15–17] was proposed; it separates sources across all frequencies using frequency-independent probabilistic source models. Based on IVA, Independent Vector Extraction (IVE) [18–20] was developed to separate $N(<M)$ sources (i.e., solving a BSS problem in overdetermined situations) in a computationally efficient way.

Since BSS can only estimate a specified number of sources, additional techniques are necessary to determine which sources correspond to the SOIs. Spatial regularization [3] can be used for this purpose. For example, IVA with spatial regularization, called Spatially-Regularized IVA (SRIVA) [21–24], estimates the spatial filters by considering both the BSS objective and the spatial regularization to determine the SOIs. Spatial regularization has also been incorporated into IVE (SRIVE) [23]. SRIVE can extract only the SOIs by dealing with the other sounds collectively as a noise signal. Spatial regularization has been experimentally proven to be effective even with DOA-based ATFs [21–26]. Therefore, SRIVE can be a useful method for DOA-informed speech enhancement when the number of SOIs is limited to one.

Table 1 Comparison of DOA-informed speech enhancement methods

Methods	Robustness against modeling errors of ATFs (included in DOA-based ATFs)	Accuracy of speech enhancement in underdetermined situations
MPDR [1]	Low	Low
Sw-MPDR [6]	Low	High
wMPDR [11–13]	High	Low
SRIVE [23]	High	Low
Sw-wMPDR [proposed]	High	High
Sw-SRIVE [proposed]	High	High

One issue in SRIVE is that it requires more microphones than sources to achieve accurate speech enhancement. To overcome this issue, switching IVE (Sw-IVE) [10] was proposed by incorporating the switching mechanism that was first introduced to MPDR into IVE. However, previous research derived an optimization algorithm only for cases with $M \geq N$ (i.e., determined and overdetermined situations). Thus, it cannot be applied to underdetermined situations.

Table 1 summarizes the advantages and disadvantages of the conventional methods. In short, none of the conventional DOA-informed speech enhancement methods based on beamforming and BSS can simultaneously solve the following two problems[1]:

1) Robust speech enhancement against modeling errors of ATFs (included in DOA-based ATFs).
2) Accurate speech enhancement in underdetermined situations

As the main contributions of this paper, we propose two DOA-informed speech enhancement methods that feature the following aspects:

- We introduce a TVG source model into Sw-MPDR [6] to improve its speech enhancement performance when using DOA-based ATFs. The extended method is called Switching wMPDR (Sw-wMPDR).
- We introduce a switching mechanism to SRIVE [23] to improve its speech enhancement performance in underdetermined situations. The extended method is called Sw-SRIVE.

Table 1 also shows the advantages of the proposed methods over the conventional methods.

First, with Sw-wMPDR, we cluster the time frames and estimate spatial filters for individual clusters based on MLE, assuming that the SOI follows a TVG distribution. Here, similar to wMPDR [11–13], we expect that using the TVG source model increases the robustness of the switching BF's speech enhancement when there are modeling errors in ATFs, and thus we can solve the above-mentioned problem 1 of Sw-MPDR.

Next, with Sw-SRIVE, we cluster the time frames and estimate the spatial filters for individual clusters based on the BSS objective with spatial regularization. The switching mechanism enables Sw-SRIVE to suppress interferences more effectively than SRIVE [23] in underdetermined situations, and thus we can solve the above-mentioned problem 2 of SRIVE.

In addition, we propose two new techniques to improve the performance of Sw-SRIVE. The first is robust clustering of time frames in underdetermined situations. Conventionally, IVE and its extensions assume that all the noise signals (other than the SOIs) follow a stationary Gaussian distribution. However, those signals become non-stationary when the observed signal includes interference speech signals. Our preliminary experiments showed that this mismatch degrades the clustering of time frames by Sw-SRIVE. To avoid that problem, we introduce a clustering technique that is robust against the errors in the assumption.

The second new technique aims to stabilize the optimization of Sw-SRIVE. Instability in the optimization arises with Sw-SRIVE when a specific cluster is composed of only an insufficient number of time frames. To avoid this problem, we use a single-state IVE objective to regularize the optimization. We call this technique state regularization.

In experiments, we show how our proposed methods outperform the conventional methods in underdetermined situations using different numbers of speech signals. In addition, we show that the above robust clustering and state regularization improve the speech enhancement performance of Sw-SRIVE.

This paper is an extended version of our conference papers, which proposed spatially-regularized switching IVA (Sw-SRIVA) [27] and Sw-wMPDR [28]. The extension presented in this paper includes the following:

[1] We discuss DNN-based approaches in the next section.

1. Extension of Sw-SRIVA to Sw-SRIVE, including robust clustering for Sw-SRIVE to handle underdetermined situations in Section 5.3 and state regularization for Sw-SRIVE's stable parameter optimization in Section 5.4.
2. Thorough discussion of the effectiveness of utilizing the TVG source model for DOA-informed speech enhancement in Section 4.3 and its experimental validation in Section 7.2.

2 Related work

A super-directive beamformer can be applied to DOA-informed speech enhancement [29]. It assumes an isotropic noise field where the noise comes to microphones from arbitrary directions with equal power and determines the noise SCM based on this assumption. Although this method can perform DOA-based speech enhancement without prior knowledge of the noise statistics, the performance degrades in an underdetermined situation because it largely deviates from the isotropic noise field. In contrast, wMPDR and Sw-wMPDR can adapt to the characteristics of the noise in the observed signal and thus more effectively reduce it.

BSS enhances speech signals including an SOI without DOA by separating the microphone signals into individual signals [14–17, 30, 31]. However, BSS requires prior knowledge of the number of speech signals captured by microphones. Moreover, post-processing is necessary to determine the SOI from the separated signals. On the other hand, our proposed methods, Sw-wMPDR and Sw-SRIVE, do not require post-processing or information about the number of speech signals.

A method was proposed to enhance an SOI by estimating its mask with its DOA and leveraging the mask for spatial filter estimation [32]. The method uses a Neural Network (NN) to estimate the SOI's mask based on pairs of microphone observations and the DOA. However, this method does not fit the goal of this paper well in the following aspects. The method's effectiveness has only been experimentally confirmed in determined situations where the observed signals involve two speakers. In addition, the observed signals are generated using the image method [33]. The image method introduces mismatches compared to real-world environments. NNs tend to suffer from reduced estimation accuracy when there are mismatches between training and testing data. Moreover, the NNs come with high computational costs and specialized hardware requirements. In contrast, we propose methods for underdetermined situations, typically involving more than two speakers, and conduct experiments using impulse responses recorded in real environments. It may be worth noting that the performance of model-based approaches degrades with mismatches between the model and the data. To establish model-based approaches that are more adaptable to the data, this paper proposes methods using the time-varying source model and the switching mechanism.

3 Problem formulation

This paper considers a multi-input single-output speech enhancement that estimates a source image of the SOI at the first microphone. We assume that the reverberation time is not so large in this paper[2].

Suppose that an SOI $s(f, t)$ and interference speech signals $u_n(f,t)$ for $n = 1, \ldots, N-1$ are mixed and captured by M microphones. We represent observed signal $\boldsymbol{x}(f,t)$ at each time frame indexed by $t = 1, \ldots, T$ and frequency bin indexed by $f = 1, \ldots, F$ in the Short-Time Fourier Transform (STFT) domain as

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

where $(\cdot)^{\mathsf{T}}$ denotes the transpose. We model the observed signal $\boldsymbol{x}(f,t)$ by

$$\boldsymbol{x}(f,t) = \boldsymbol{h}(f)s(f,t) + \sum_{n=1}^{N-1} \boldsymbol{h}_n(f)u_n(f,t) + \boldsymbol{r}(f,t), \tag{2}$$

where $\boldsymbol{h}(f)$ is the ATF of the SOI from the source location to the microphones, $\boldsymbol{h}_n(f)$ is that of the nth interference, and $\boldsymbol{r}(f,t)$ is the ambient noise.

We also suppose that we have an estimate of the SOI's ATF (hereafter called a steering vector to be distinguished from a true ATF). Assuming that the SOI's DOA and the microphone locations are given or estimated, we obtain steering vector $\boldsymbol{a}(f)$ based on the plane-wave assumption and relative Time Delays of Arrival (TDOA) $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_M]^{\mathsf{T}} \in \mathbb{R}^M$ from the SOI to M microphones:

$$\begin{aligned}\boldsymbol{a}(f) &= [a_1(f), \ldots, a_M(f)]^{\mathsf{T}} \in \mathbb{C}^M \\ &= \frac{1}{\sqrt{M}} \exp\left(-2\pi \frac{(f-1)}{N_{\mathrm{F}}} \boldsymbol{\tau}\sqrt{-1}\right),\end{aligned} \tag{3}$$

where each element $a_m(f)$ is an estimate of an ATF from an SOI to the mth microphone and N_{F} is the number of points used for STFT. This paper defines the number of frequency bins F as $N_{\mathrm{F}}/2 + 1$. Equation (3) normalizes the norm of steering vector $\boldsymbol{a}(f)$ to a constant ($= 1$) because otherwise the value of the spatial regularization term (see (31) below) changes with the norm. Note that $\boldsymbol{a}(f)$ contains substantial errors from true ATFs $\boldsymbol{h}(f)$ in (2) because $\boldsymbol{a}(f)$ only contains a direct path component without any reflection paths.

[2] Although it might be possible to incorporate (e.g.,) dereverberation to adapt to long reverberation [7, 9], this paper relegates this task to future work.

Because this paper aims to obtain a source image of the SOI at the first microphone, we set $\tau_1 = 0$.

3.1 Estimation model

This section describes two important estimation models: a separation model with a switching mechanism and a source model. Both are used in our proposed methods in Sections 4 and 5.

3.1.1 Separation model with switching mechanism

To obtain an estimate of SOI $\hat{s}(f,t)$, we use a switching mechanism, which is also used in Sw-MPDR and Sw-IVE [6, 9, 10]. With this mechanism, J separation matrices, $\boldsymbol{W}_j(f)$ for $1 \le j \le J$, are applied to the observed signal to yield J different sets of SOI $\hat{s}_j(f,t) \in \mathbb{C}$ and noise signals $\hat{\boldsymbol{z}}_j(f,t) \in \mathbb{C}^{M-1}$:

$$\begin{bmatrix} \hat{s}_j(f,t) \\ \hat{\boldsymbol{z}}_j(f,t) \end{bmatrix} = \boldsymbol{W}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,t) \text{ for } 1 \le j \le J, \tag{4}$$

$$\begin{aligned} \boldsymbol{W}_j(f) &= [\boldsymbol{w}_{j,1}(f), \boldsymbol{w}_{j,2}(f), \ldots, \boldsymbol{w}_{j,M}(f)] \\ &= [\boldsymbol{w}_{j,1}(f), \boldsymbol{W}_{j,\mathrm{Z}}(f)] \in \mathbb{C}^{M\times M}. \end{aligned} \tag{5}$$

This paper refers to $\boldsymbol{w}_{j,1}(f)$ as a spatial filter and $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ as a noise filter. Note that although we do not aim to obtain an estimate of noise signal $\hat{\boldsymbol{z}}_j(f,t)$, both Sw-wMPDR and Sw-SRIVE use the model (4) to derive optimization criteria for spatial filter $\boldsymbol{w}_{j,1}(f)$. Then, one of the J separated sources is selected as the final output for each time frequency:

$$\underbrace{\begin{bmatrix} \hat{s}(f,t) \\ \hat{\boldsymbol{z}}(f,t) \end{bmatrix}}_{\hat{\boldsymbol{y}}(f,t)} = \sum_{j=1}^{J} \delta_j(f,t) \begin{bmatrix} \hat{s}_j(f,t) \\ \hat{\boldsymbol{z}}_j(f,t) \end{bmatrix}, \tag{6}$$

$$\sum_{j=1}^{J} \delta_j(f,t) = 1 \quad \text{and} \quad \delta_j(f,t) \in \{0,1\}. \tag{7}$$

The selection is implemented using a time-frequency dependent switching weight $\delta_j(f,t)$ in (6), which takes a binary value. Hereafter, j and J are referred to as an index of a switching state and the total number of states, respectively.

The switching mechanism assumes that time frames of the observed signal can be clustered into several groups, each containing fewer sources than microphones, even when the entire observation contains more sources than microphones, i.e., in an underdetermined situation. Accordingly, we can achieve more effective interference suppression by clustering the time frames effectively and applying different spatial filter to each cluster, rather than applying a single time-invariant filter to the entire observation. Although the separation model of the switching mechanism can be viewed as a time-varying separation model defined as $\boldsymbol{W}(f,t) = \sum_j \delta_j(f,t)\boldsymbol{W}_j(f)$, its goal is completely different from that of conventional online processing such as [26, 34]. Conventional online processing updates the statistics of the signals at each time frame with a certain forgetting factor, assuming that the signal is stationary for a certain period prior to the frame. However, such methods cannot handle the problem discussed in this paper because active sources (or signal statistics) change quickly from frame to frame in the underdetermined situation. In contrast, the switching mechanism can handle such situations simply by switching the filter.

In this paper, steering vector $\boldsymbol{a}(f)$ in (3) will be used to define distortionless and blocking constraints in Sw-wMPDR (Section 4) and to specify spatial regularization in Sw-SRIVE (Section 5), for estimating spatial filter $\boldsymbol{w}_{j,1}(f)$ and noise filter $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ in (5). The switching mechanism (6), on the other hand, is used to cluster time-frequency bins into a fixed number of clusters and to estimate $\boldsymbol{w}_{j,1}(f)$ and $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ separately for each cluster.

3.1.2 Source model

To estimate $\boldsymbol{W}_j(f)$ and $\delta_j(f,t)$, we use the following model and assumption, both of which are also used in wMPDR and SRIVE [13, 23].

First, we use the TVG source model where the extracted SOI $\hat{s}(f,t)$ adheres to the following distribution:

$$\begin{aligned} p(\hat{s}(f,t)) &= \mathcal{N}_{\mathbb{C}}(0, v(f,t)) \\ &= \frac{1}{\pi v(f,t)} \exp\left(-\frac{|\hat{s}(f,t)|^2}{v(f,t)}\right), \end{aligned} \tag{8}$$

where $v(f,t)$ is a time-varying source variance of $\hat{s}(f,t)$.

Next, we assume that SOI $\hat{s}(f,t)$ and noise signals $\hat{\boldsymbol{z}}(f,t)$ are mutually independent over all times and frequencies:

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

Based on the above TVG source model and Appendix, we can derive a negative log-likelihood function for a given observed signal $\mathcal{X} = \{x_m(f,t)\}_{m,f,t}$:

$$\begin{aligned} &\mathcal{L}_{\mathrm{NL}}(\mathcal{X}; \mathcal{W}, \mathcal{V}, \mathcal{D}) \\ &\overset{c}{=} \sum_{j,f,t} \delta_j(f,t) \Bigg(-2\log|\det \boldsymbol{W}_j(f)| \\ &\quad + \log v(f,t) + \frac{|\hat{s}_j(f,t)|^2}{v(f,t)} - \log p(\hat{\boldsymbol{z}}_j(f,t)) \Bigg), \end{aligned} \tag{10}$$

where $\mathcal{W} = \{\boldsymbol{W}_j(f)\}_{j,f}$, $\mathcal{V} = \{v(f,t)\}_{f,t}$, $\mathcal{D} = \{\delta_j(f,t)\}_{j,f,t}$, and $\overset{c}{=}$ denotes the equality up to the constant terms. Although the right hand side of (10) does not explicitly include $x_m(f,t)$, $x_m(f,t)$ is implicitly included via $\hat{s}_j(f,t)$ and $\hat{\boldsymbol{z}}_j(f,t)$ according to (4). Hereafter, this likelihood

function is utilized for estimating $\mathcal{W}$ and $\mathcal{D}$ and obtaining a source image of the SOI.

In the next section, we present two proposed methods by introducing additional conditions (e.g., constraints and regularization) to the likelihood function (10). Specifically, the new conditions use the DOA-based ATFs to determine the SOI from the mixture.

4 Proposed method: Switching weighted MPDR (Sw-wMPDR)

In this section, we extend the conventional Sw-MPDR [6] to our proposed method, Sw-wMPDR. Based on the switching mechanism, Sw-wMPDR can handle underdetermined situations well. In addition, the TVG source model makes Sw-wMPDR perform more effectively than Sw-MPDR for DOA-informed speech enhancement. After defining the cost function in Section 4.1, we derive a parameter optimization algorithm of Sw-wMPDR based on MLE in Section 4.2. We describe how the TVG source model supports DOA-informed speech enhancement in Section 4.3.

4.1 Cost function of Sw-wMPDR

In addition to the estimation models introduced in Section 3.1, we introduce the following constraints for $\forall j$ and f to spatial filter $\boldsymbol{w}_{j,1}(f)$ and noise filter $\boldsymbol{W}_{j,\mathrm{Z}}(f)$:

$$\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{a}(f) = 1, \tag{11}$$

$$\boldsymbol{W}_{j,\mathrm{Z}}^{\mathsf{H}}(f)\boldsymbol{a}(f) = \boldsymbol{0}_{M-1}, \tag{12}$$

$$\boldsymbol{W}_{j,\mathrm{Z}}(f) = \boldsymbol{W}_{j',\mathrm{Z}}(f) \text{ for } \forall j, j', \tag{13}$$

where $\boldsymbol{0}_M \in \mathbb{R}^M$ is a zero vector. Equation (11) is a distortionless constraint that makes spatial filter $\boldsymbol{w}_{j,1}(f)$ respond with value 1 to steering vector $\boldsymbol{a}(f)$. Based on (12), $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ becomes a blocking matrix that cancels the signal space spanned by $\boldsymbol{a}(f)$. Also, (13) assumes that noise filter $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ has the same value in each state j[3].

Using the constraints and disregarding constant terms unrelated with $\mathcal{W}_{\mathrm{S}} = \{\boldsymbol{w}_{j,1}(f)\}_{j,f}$, $\mathcal{V}$, and $\mathcal{D}$, we can rewrite the negative log-likelihood function in (10) as Sw-wMPDR's cost function $\mathcal{C}_{\text{Sw-wMPDR}}$ with distortionless constraint [28, 35]:

$$\begin{aligned}&\mathcal{C}_{\text{Sw-wMPDR}}(\mathcal{X}; \mathcal{W}_{\mathrm{S}}, \mathcal{V}, \mathcal{D})\\ &\overset{c}{=} \sum_{j,f,t} \delta_j(f,t)\left(\log \nu(f,t) + \frac{|\hat{s}_j(f,t)|^2}{\nu(f,t)}\right)\\ &\text{s.t. } \boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{a}(f) = 1 \text{ for } \forall f, j.\end{aligned} \tag{14}$$

Note that $\det \boldsymbol{W}_j(f) = \det \boldsymbol{W}_{j'}(f)$ and $\hat{\boldsymbol{z}}_j(f,t) = \hat{\boldsymbol{z}}_{j'}(f,t)$ hold for $\forall j$ and j' according to previous papers [28, 35] and (11)–(13), thus we can disregard the first and last terms (i.e., $-2\log|\det \boldsymbol{W}_j(f)|$ and $-\log p(\hat{\boldsymbol{z}}_j(f,t))$) in the parentheses of (10).

4.2 Optimization algorithm of Sw-wMPDR

Based on the above cost function, we can find a local optimum of the parameters by alternately updating one of $\mathcal{W}_{\mathrm{S}}$, $\mathcal{D}$, and $\mathcal{V}$ while fixing the other parameters. After initializing $\mathcal{D}$ and $\mathcal{V}$ at certain values for all j, f, and t, we iterate the following updates until a convergence is obtained:

$$\boldsymbol{\Sigma}_j(f) = \sum_{t=1}^{T} \frac{\delta_j(f,t)}{\nu(f,t)} \boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t), \tag{15}$$

$$\boldsymbol{w}_{j,1}(f) \leftarrow \frac{\boldsymbol{\Sigma}_j^{-1}(f)\boldsymbol{a}(f)}{\boldsymbol{a}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_j^{-1}(f)\boldsymbol{a}(f)}, \tag{16}$$

$$\delta_j(f,t) \leftarrow \begin{cases} 1 & \text{if } j = \underset{j'}{\operatorname{argmin}} |\boldsymbol{w}_{j',1}^{\mathsf{H}}(f)\boldsymbol{x}(f,t)|^2 \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

$$\nu(f,t) \leftarrow \sum_{j=1}^{J} \delta_j(f,t) |\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{x}(f,t)|^2, \tag{18}$$

where $\boldsymbol{\Sigma}_j(f)$ is an SCM of the observed signal for each state j normalized by time-varying source variance $\nu(f,t)$. The normalization is derived from the MLE based on the TVG source model.

4.3 Effectiveness of TVG source model for SOI enhancement

We now explain why the TVG source model is useful for DOA-informed speech enhancement. The key characteristic is that Sw-wMPDR (similar to wMPDR [11–13]) minimizes the power of the observed signal normalized by the power estimated by the TVG source model.

Before discussing Sw-wMPDR, let us discuss the tendencies of MPDR and MVDR. MPDR minimizes the power of the observed signal under the distortionless constraint to optimize the spatial filter $\boldsymbol{w}_{j,1}(f)$. Here, an SOI is included in the signal to be minimized. Thus, parts of the SOI that are not protected by the DOA-based ATF are significantly suppressed. Typically, early reflections are among the parts suppressed by MPDR. As we explained in the introduction,

[3] We might realize better speech enhancement by formulating the noise filter to be state-dependent. However, to the best of our knowledge, it is difficult to derive a computationally efficient optimization algorithm with the formulation and based on the estimation model introduced in this paper. Thus, we used an assumption (13) to derive the computationally efficient Sw-wMPDR.

excessive reduction of early reflections leads to a substantial degradation in speech enhancement. In contrast, MVDR minimizes the power of the interferences. Here, the SOI is not included in the signal to be minimized. Thus, MVDR can be relatively robust against the ATF modeling errors caused by using a DOA-based ATF.

Sw-wMPDR has robustness against ATF modeling errors similar to MVDR. Sw-wMPDR minimizes (14) to optimize $\boldsymbol{w}_{j,1}(f)$. During the iterative optimization, $\boldsymbol{w}_{j,1}(f)$ for a state j and at a frequency f is updated while fixing $\delta_j(f,t)$ and $v(f,t)$. Specifically, this is done by minimizing the following function under the distortionless constraint ($\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{a}(f)=1$):

$$\mathcal{C}_{\text{Sw-wMPDR}}(\boldsymbol{w}_{j,1}(f)) \stackrel{c}{=} \sum_t \delta_j(f,t)\frac{|\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{x}(f,t)|^2}{v(f,t)}. \tag{19}$$

In other words, Sw-wMPDR minimizes the power of the observed signal normalized by its estimated power $v(f,t)$. Due to this weighting mechanism, Sw-wMPDR tends not to suppress signals in frames with larger $v(f,t)$. Because the direct signal is protected by the distortionless constraint with the DOA-based ATF, the estimated power $v(f,t)$ becomes relatively larger during SOI-present frames than during SOI-absent frames. In addition, most of the early reflections of an SOI are included in the same time frames as those of their direct signal. As a result, Sw-wMPDR tends not to suppress the early reflections even when we use a DOA-based ATF as the distortionless constraint.

During the optimization using the TVG source model, $v(f,t)$ can become zero or an extremely small value at a frame. This makes the SCM $\boldsymbol{\Sigma}_j(f)$ in (16) unstable. Thus, setting an appropriate floor of $v(f,t)$ is essential for Sw-wMPDR to work effectively with DOA-based ATFs. In this paper, we apply flooring after (18):

$$v(f,t) \leftarrow \begin{cases} v(f,t) & \text{if } v(f,t) > \epsilon(f) \\ \epsilon(f) & \text{otherwise} \end{cases}, \tag{20}$$

$$\epsilon(f) = \frac{\epsilon_{\text{floor}}}{TM}\sum_{t=1}^{T}\|\boldsymbol{x}(f,t)\|_2^2, \tag{21}$$

where ϵ_{floor} is a flooring coefficient and $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}$.

In Section 7, we experimentally confirm that Sw-wMPDR ourperforms Sw-MPDR [6] using the TVG source model with flooring.

5 Proposed method: Switching SRIVE (Sw-SRIVE)

In this section, we extend the conventional SRIVE [23] to our proposed method, Sw-SRIVE. Based on the switching mechanism, Sw-SRIVE can improve its speech enhancement performance to surpass SRIVE. In addition, Sw-SRIVE uses spatial regularization instead of a distortionless constraint. While a distortionless constraint is a hard constraint that makes a spatial filter have a specific response regardless of the accuracy of the steering vector, spatial regularization is a soft constraint in which we can control the strength of the constraint by tuning its weight. This soft constraint makes Sw-IVE [10] robust even when the DOA-based steering vector deviates from the true ATF. After defining the cost function in Section 5.1, we derive the parameter optimization algorithm of Sw-SRIVE in Section 5.2. Then, we propose a robust clustering technique to perform Sw-SRIVE effectively in Section 5.3. Finally, we introduce state regularization for SRIVE's stable parameter optimization in Section 5.4.

5.1 Cost function of Sw-SRIVE

In addition to the estimation models introduced in Section 3.1, we assume that the extracted noise signal $\hat{\boldsymbol{z}}_j(f,t)$ adheres to a multivariate stationary complex Gaussian distribution at each j [10]:

$$\begin{aligned} p(\hat{\boldsymbol{z}}_j(f,t)) &= \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_{M-1}, \boldsymbol{\Omega}_j(f)) \\ &= \frac{\pi^{-(M-1)}}{\det \boldsymbol{\Omega}_j(f)} \exp\left(-\hat{\boldsymbol{z}}_j^{\mathsf{H}}(f,t)\boldsymbol{\Omega}_j^{-1}(f)\hat{\boldsymbol{z}}_j(f,t)\right), \end{aligned} \tag{22}$$

where $\boldsymbol{\Omega}_j(f) \in \mathbb{C}^{(M-1)\times(M-1)}$ is a covariance matrix of $\hat{\boldsymbol{z}}_j(f,t)$. Under the above assumption, negative log-likelihood function $\mathcal{L}_{\text{NL}}$ without spatial regularization in (10) can be rewritten as $\mathcal{L}_{\text{Sw-IVE}}$:

$$\mathcal{L}_{\text{Sw-IVE}}(\mathcal{X};\Theta) \stackrel{c}{=} \sum_{j,f,t}\delta_j(f,t)\mathcal{L}(\theta_j(f,t)), \tag{23}$$

$$\begin{aligned} \mathcal{L}(\theta_j(f,t)) &= -2\log\left|\det \boldsymbol{W}_j(f)\right| \\ &+ \log v(f,t) + \frac{|\hat{s}_j(f,t)|^2}{v(f,t)} + \log\det\boldsymbol{\Omega}_j(f) \\ &+ \hat{\boldsymbol{z}}_j^{\mathsf{H}}(f,t)\boldsymbol{\Omega}_j^{-1}(f)\hat{\boldsymbol{z}}_j(f,t), \end{aligned} \tag{24}$$

where $\Theta = \{\mathcal{V},\mathcal{D},\mathcal{W},\mathcal{O}\}$, $\mathcal{O} = \{\boldsymbol{\Omega}_j(f)\}_{j,f}$, and $\theta_j(f,t) = (\boldsymbol{W}_j(f), \boldsymbol{\Omega}_j(f), v(f,t))$. It may be worth noting that $\mathcal{L}(\theta_j(f,t))$ corresponds to the negative log-likelihood function of IVE (without switching or spatial regularization). Thus, (23) can be viewed to switch the IVE's likelihood functions with different state parameters $\theta_j(f,t)$ using switching weight $\delta_j(f,t)$.

For Sw-IVE [10], it is difficult to determine which source is the SOI to be extracted when more than one source is included in the observed signal. To identify the SOI, Sw-SRIVE leverages the spatial regularization for the spatial filter estimation.

As discussed in [20], IVE and Sw-IVE are unable to uniquely determine noise signals $\hat{\boldsymbol{z}}_j(f,t)$ and noise filter $\boldsymbol{W}_{j,\mathrm{Z}}(f)$. This ambiguity could potentially have a detrimental impact on the regularization process [26]. To address this problem, we define a cost function that merges the regularization term with the likelihood function in a transformed domain. In this domain, the estimated noise signal adheres to a complex Gaussian distribution with the mean of zero vector, $\boldsymbol{0}_{M-1}$ and a constant covariance of identity matrix $\boldsymbol{I}_{M-1} \in \mathbb{R}^{(M-1)\times(M-1)}$:

$$p(\tilde{\boldsymbol{z}}_j(f,t)) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_{M-1}, \boldsymbol{I}_{M-1}), \tag{25}$$

where $\tilde{\boldsymbol{z}}_j(f,t)$ is the transformed noise signal. The transformed domain does not have scale ambiguity for noise filters, which avoids the detrimental impact on the regularization process explained in [26]. Moreover, we can use a computationally efficient update rule [19] for noise filters (see (42) below). Once we obtain the cost function in the transformed domain, we transform it back to the original domain. Note that once we determine the term in the original domain, we do not need to transform the signal domain as preprocessing of Sw-SRIVE.

With this transformation, (23) and (24) can be rewritten without modifying its value [20]:

$$\mathcal{L}^{\mathrm{trans}}_{\text{Sw-IVE}}(\mathcal{X}; \tilde{\Theta}) = \sum_{j,f,t} \delta_j(f,t) \mathcal{L}^{\mathrm{trans}}(\tilde{\theta}_j(f,t)), \tag{26}$$

$$\begin{aligned}\mathcal{L}^{\mathrm{trans}}(\tilde{\theta}_j(f,t)) &= -2\log\left|\det \tilde{\boldsymbol{W}}_j(f)\right| \\ &+ \log v(f,t) + \frac{|\hat{s}_j(f,t)|^2}{v(f,t)} + \tilde{\boldsymbol{z}}_j^{\mathsf{H}}(f,t)\tilde{\boldsymbol{z}}_j(f,t),\end{aligned} \tag{27}$$

where $\tilde{\theta}_j(f,t) = (\tilde{\boldsymbol{W}}_j(f), v(f,t))$,

$$\tilde{\boldsymbol{z}}_j(f,t) = \boldsymbol{L}_j^{\mathsf{H}}(f)\hat{\boldsymbol{z}}_j(f,t), \tag{28}$$

$$\begin{aligned}\tilde{\boldsymbol{W}}_j(f) &= \left[\boldsymbol{w}_{j,1}(f), \tilde{\boldsymbol{W}}_{j,\mathrm{Z}}(f)\right] \\ &= \left[\boldsymbol{w}_{j,1}(f), \boldsymbol{W}_{j,\mathrm{Z}}(f)\boldsymbol{L}_j(f)\right],\end{aligned} \tag{29}$$

and $\boldsymbol{L}_j(f) \in \mathbb{C}^{(M-1)\times(M-1)}$ is a matrix satisfying $\boldsymbol{L}_j(f)\boldsymbol{L}_j^{\mathsf{H}}(f) = \boldsymbol{\Omega}_j^{-1}(f)$. For example, $\boldsymbol{L}_j(f)$ can be determined as the Cholesky decomposition of $\boldsymbol{\Omega}_j^{-1}(f)$. The equality between (23) and (26) can be shown by substituting (28) and (29) into (27) followed by certain mathematical manipulations.

In the above-transformed domain, we define the cost function for Sw-SRIVE that combines the negative log-likelihood function with a spatial regularization term $\mathcal{J}^{\mathrm{trans}}_{\mathrm{SR}}(\tilde{\mathcal{W}}_{\mathrm{Z}}, \mathcal{D})$:

$$\mathcal{C}^{\mathrm{trans}}_{\mathrm{cost}}(\tilde{\Theta}) = \mathcal{L}^{\mathrm{trans}}_{\text{Sw-IVE}}(\mathcal{X}; \tilde{\Theta}) + \mathcal{J}^{\mathrm{trans}}_{\mathrm{SR}}(\tilde{\mathcal{W}}_{\mathrm{Z}}, \mathcal{D}), \tag{30}$$

where $\tilde{\Theta} = \{\mathcal{V}, \mathcal{D}, \tilde{\mathcal{W}}\}$, $\tilde{\mathcal{W}} = \{\tilde{\boldsymbol{W}}_j(f)\}_{j,f}$, and $\tilde{\mathcal{W}}_{\mathrm{Z}} = \{\tilde{\boldsymbol{W}}_{j,\mathrm{Z}}(f)\}_{j,f}$. This paper adopts a regularization term used for SRIVE [23] with an extension for the switching mechanism:

$$\mathcal{J}^{\mathrm{trans}}_{\mathrm{SR}}(\tilde{\mathcal{W}}_{\mathrm{Z}}, \mathcal{D}) = \lambda_{\mathrm{SR}} \sum_{j,f,t} \delta_j(f,t) \|\tilde{\boldsymbol{W}}_{j,\mathrm{Z}}^{\mathsf{H}}(f)\boldsymbol{a}(f)\|_2^2, \tag{31}$$

where λ_{SR} is a regularization weight. The regularization term in (31) forces noise filter $\tilde{\boldsymbol{W}}_{j,\mathrm{Z}}(f)$ to direct a spatial null to steering vector $\boldsymbol{a}(f)$. Equation (31) is one of the novelties in our paper since no paper has introduced spatial regularization into Sw-IVE. A major difference from the regularization term for SRIVE [23] is that we regularize J noise filters using the same term with switching weight $\delta_j(f,t)$ in (31), rather than applying the term to only one noise filter with no switching weight.

Then, we obtain the cost function in the original domain by transforming (30) back to the original domain without modifying its value:

$$\mathcal{C}_{\mathrm{cost}}(\Theta) = \mathcal{L}_{\text{Sw-IVE}}(\mathcal{X}; \Theta) + \mathcal{J}_{\mathrm{SR}}(\mathcal{W}_{\mathrm{Z}}, \mathcal{O}, \mathcal{D}), \tag{32}$$

where

$$\mathcal{J}_{\mathrm{SR}}(\mathcal{W}_{\mathrm{Z}}, \mathcal{O}, \mathcal{D}) = \lambda_{\mathrm{SR}} \sum_{j,f,t} \delta_j(f,t) \mathcal{J}(\boldsymbol{W}_j(f), \boldsymbol{\Omega}_j(f)), \tag{33}$$

$$\mathcal{J}(\boldsymbol{W}_j(f), \boldsymbol{\Omega}_j(f)) = \|\boldsymbol{W}_{j,\mathrm{Z}}^{\mathsf{H}}(f)\boldsymbol{a}(f)\|^2_{\boldsymbol{\Omega}_j^{-1}(f)}, \tag{34}$$

$\|\boldsymbol{x}\|^2_{\Sigma} = \boldsymbol{x}^{\mathsf{H}}\Sigma\boldsymbol{x}$, and $\mathcal{W}_{\mathrm{Z}} = \{\boldsymbol{W}_{j,\mathrm{Z}}(f)\}_{j,f}$.

We describe the relationship between the spatial regularization term used here and the assumptions (11) and (12) used in Sw-wMPDR. The regularization to noise filter $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ used in (34) can be considered as introducing a blocking matrix assumption (12) not as a constraint but as a regularization term. Similarly, it is also possible to introduce a distortionless constraint in (11) as a regularization term for spatial filter $\boldsymbol{w}_{j,1}(f)$ [22, 24, 26]. For the sake of simplicity, however, we do not adopt this type of regularization term in this paper.

5.2 Optimization algorithm of Sw-SRIVE

We optimize parameters Θ for Sw-SRIVE by minimizing the cost function in (32). Because no closed-form solution was obtained for the optimization, we alternately update one among $\mathcal{W}$, $\mathcal{V}$, $\mathcal{O}$, and $\mathcal{D}$ by fixing the other parameters. We iterate the update until a convergence is obtained. After initializing $\mathcal{W}$, $\mathcal{D}$, and $\mathcal{V}$ at certain values, we iterate the updates described in the following.

The update rule for $\boldsymbol{W}_j(f)$ at each state j can be derived in almost the same way as that for conventional IVE [19, 20]. Extracting the terms related with $\boldsymbol{W}_j(f)$ from (32) yields

$$\begin{aligned}\mathcal{C}_{\boldsymbol{W}_j(f)} \propto & -2\log\left|\det \boldsymbol{W}_j(f)\right| \\ & + \mathrm{tr}(\boldsymbol{W}_{j,\mathrm{Z}}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,\mathrm{Z}}(f)\boldsymbol{W}_{j,\mathrm{Z}}(f)\boldsymbol{\Omega}_j^{-1}(f)) \\ & + \boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_{j,\mathrm{S}}(f)\boldsymbol{w}_{j,1}(f),\end{aligned} \tag{35}$$

where

$$\boldsymbol{\Pi}_{j,\mathrm{Z}}(f) = \boldsymbol{\Sigma}_{j,\mathrm{Z}}(f) + \lambda_{\mathrm{SR}}\boldsymbol{a}(f)\boldsymbol{a}^{\mathsf{H}}(f), \tag{36}$$

$$\boldsymbol{\Sigma}_{j,\mathrm{S}}(f) = \frac{1}{T_j(f)}\sum_{t=1}^{T}\frac{\delta_j(f,t)}{\nu(t)}\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t), \tag{37}$$

$$\boldsymbol{\Sigma}_{j,\mathrm{Z}}(f) = \frac{1}{T_j(f)}\sum_{t=1}^{T}\delta_j(f,t)\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t), \tag{38}$$

$$T_j(f) = \sum_{t=1}^{T}\delta_j(f,t), \tag{39}$$

and $\nu(t) \leftarrow \frac{1}{F}\sum_{f=1}^{F}\nu(f,t)$. Note that this paper adopts a frequency-independent source model only when updating separation matrices $\mathcal{W}$, following a previously proposed practical technique [9]. The frequency-independent source model was originally utilized to IVA [15, 16] to separate sources across all frequencies.

Since the above cost function has the same form as that of IVE, we can use the same optimization technique [19, 20, 23], which updates $\boldsymbol{W}_j(f)$ with a sequence of $\boldsymbol{w}_{j,1}(f) \rightarrow \boldsymbol{W}_{j,\mathrm{Z}}(f)$ for $j = 1, \ldots, J$. We update $\boldsymbol{w}_{j,1}(f)$ using

$$\boldsymbol{w}_{j,1}(f) = (\boldsymbol{W}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_{j,\mathrm{S}}(f))^{-1}\boldsymbol{e}_1, \tag{40}$$

$$\boldsymbol{w}_{j,1}(f) = \frac{\boldsymbol{w}_{j,1}(f)}{\sqrt{\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{\Sigma}_{j,\mathrm{S}}(f)\boldsymbol{w}_{j,1}(f)}}, \tag{41}$$

where $\boldsymbol{e}_1$ denotes the first column of $\boldsymbol{I}_M$. Then, we update $\boldsymbol{W}_{j,\mathrm{Z}}(f)$ using

$$\boldsymbol{W}_{j,\mathrm{Z}}(f) \leftarrow \begin{bmatrix} -\dfrac{\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,\mathrm{Z}}(f)\boldsymbol{E}_{\mathrm{Z}}}{\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,\mathrm{Z}}(f)\boldsymbol{e}_1} \\ \boldsymbol{I}_{M-1} \end{bmatrix}, \tag{42}$$

where $\boldsymbol{E}_{\mathrm{Z}}$ is the last $M-1$ columns of $\boldsymbol{I}_M$.

After updating $\mathcal{W}$ and obtaining signals according to (4) and (6), we update $\mathcal{V}$ and $\mathcal{O}$ based on the cost function in (32):

$$\nu(f,t) \leftarrow |\hat{s}(f,t)|^2, \tag{43}$$

$$\boldsymbol{\Omega}_j(f) \leftarrow \boldsymbol{W}_{j,\mathrm{Z}}^{\mathsf{H}}(f)\boldsymbol{\Pi}_{j,\mathrm{Z}}(f)\boldsymbol{W}_{j,\mathrm{Z}}(f). \tag{44}$$

Finally, we update switching weights $\mathcal{D}$ by setting $\delta_j(f,t) = 1$ for a state j that gives the minimum cost function among all states at each time frequency:

$$\delta_j(f,t) \leftarrow \begin{cases} 1 \text{ if } j = \underset{j'}{\mathrm{argmin}}\,\mathcal{C}(\theta_{j'}(f,t)) \\ 0 \text{ otherwise} \end{cases}, \tag{45}$$

$$\mathcal{C}(\theta_{j'}(f,t)) = \mathcal{L}(\theta_{j'}(f,t)) + \mathcal{J}(\boldsymbol{W}_{j'}(f), \boldsymbol{\Omega}_{j'}(f)). \tag{46}$$

Let us briefly summarize the relationship among IVE [19, 20], Sw-IVE [10], SRIVE [23], and Sw-SRIVE:

1. Sw-SRIVE is equivalent to SRIVE when we set the number of switching states $J = 1$ (accordingly skipping the update of $\delta_j(f,t)$ in (45)).
2. Sw-SRIVE becomes Sw-IVE when setting spatial regularization weight $\lambda_{\mathrm{SR}} = 0$.
3. Sw-SRIVE is reduced to IVE when dropping both the spatial regularization and the switching mechanism.

5.3 Robust clustering technique for Sw-SRIVE

Our preliminary experiments found that Sw-SRIVE sometimes fails to greatly outperform SRIVE when estimating an SOI in underdetermined situations. One possible cause could be the erroneous assumption introduced in (22) that noise signals $\hat{\boldsymbol{z}}_j(f,t)$ are stationary Gaussian at each switching state. In underdetermined situations, $\hat{\boldsymbol{z}}_j(f,t)$ tends to be non-stationary since they include interference speech signals. This model mismatch degrades the SOI estimation when we update switching weights $\delta_j(f,t)$ depending on noise covariance matrix $\boldsymbol{\Omega}_j(f)$ in (22). Thus, to mitigate the dependency, we introduce a technique that updates $\delta_j(f,t)$ by dropping the terms related with $\boldsymbol{\Omega}_j(f)$ from the cost function in (32):

$$\delta_j(f,t) \leftarrow \begin{cases} 1 \text{ if } j = \underset{j'}{\mathrm{argmin}}\,\mathcal{L}^{+}(\boldsymbol{W}_{j'}(f), \nu(f,t)) \\ 0 \text{ otherwise} \end{cases}, \tag{47}$$

where

$$\begin{aligned}\mathcal{L}^{+}(\boldsymbol{W}_j(f), \nu(f,t)) = & -2\log\left|\det \boldsymbol{W}_j(f)\right| \\ & + \log\nu(f,t) + \frac{|\hat{s}_j(f,t)|^2}{\nu(f,t)}.\end{aligned} \tag{48}$$

Equation (47) updates switching weights $\delta_j(f,t)$ without the influence from the erroneous estimation of

$\boldsymbol{\Omega}_j(f)$, which can happen in underdetermined situations as denoted in this section. As a consequence, $\delta_j(f,t)$ is updated based mainly on terms related to the SOI. Thus, (47) is more robust against the erroneous estimation of $\boldsymbol{\Omega}_j(f)$ than (45). We refer to the clustering performed by (47) as robust clustering and Sw-SRIVE adopting it as Sw-SRIVE+.

5.4 State regularization for Sw-SRIVE

Although Sw-SRIVE and Sw-SRIVE+ estimate the SOI more accurately than SRIVE, they sometimes suffer from unstable optimization. After updating $\delta_j(f,t)$ using (45) or (47), the number of frames $T_j(f)$ assigned to a certain state j can be very small. This makes the parameter optimization unstable.

To avoid the above problem, this paper introduces state regularization. In concrete, we use the following cost function for the parameter optimization, in which the last term is newly added to (32):

$$\mathcal{C}^{+}_{\mathrm{cost}}(\Theta) = \mathcal{L}_{\text{Sw-IVE}}(\mathcal{X};\Theta) + \mathcal{J}_{\mathrm{SR}}(\mathcal{W}_{\mathrm{Z}},\mathcal{O},\mathcal{D}) + \lambda_{\mathrm{state}} \sum_{j,f,t} \mathcal{L}(\theta_j(f,t)). \tag{49}$$

The last term in (49) is a state regularization term. This term is defined as the sum of the negative log-likelihood function of IVE, $\mathcal{L}(\theta_j(f,t))$, for all states j, multiplied by weight λ_{state}. Since this term does not consider any switching mechanism, the likelihood value for the parameters of each state j is evaluated across all time frames. If we assign a relatively small weight to λ_{state}, the regularization term becomes insignificant in the cost function for a state including a substantial number of time frames. Conversely, for a state that includes a minimal number of time frames, the term $\mathcal{L}_{\text{Sw-IVE}}(\mathcal{X};\Theta)$ becomes insignificant, and the regularization term primarily influences the cost function. By considering both terms, we can ensure the stability of Sw-SRIVE's optimization, regardless of the number of time frames contained in each state.

We can derive a parameter optimization algorithm of minimizing the cost function (49) as the algorithm described in Section 5.2 with the following modifications:

- Set $\delta'_j(f,t) = \delta_j(f,t) + \lambda_{\mathrm{state}}$ and substitute $\delta_j(f,t)$ in (37), (38), and (39) with $\delta'_j(f,t)$.
- Modify (43) as

$$v(f,t) \leftarrow \frac{\sum_{j=1}^{J} \delta'_j(f,t)|\hat{s}_j(f,t)|^2}{1 + J\lambda_{\mathrm{state}}}. \tag{50}$$

6 Computational complexity of each method

We evaluated the computational complexity of each method for each iterative update assuming $T \gg MJ$ and showed it in Table 2. The table includes the complexity of MPDR, although it does not use iterative updates.

In each complexity in the table, the complexity of $\mathcal{O}(M^2TF)$ is for calculating $\boldsymbol{\Sigma}_j(f)$, $\boldsymbol{\Sigma}_{j,\mathrm{S}}(f)$, $\boldsymbol{\Sigma}_{j,\mathrm{Z}}(f)$, and all methods include some of these. Note that the calculation is independent of the number of states J under an assumption that $\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)$ is calculated only once in each iterative update, even if we introduce state regularization. Next, the complexity of $\mathcal{O}(MJTF)$ included in Sw-MPDR, Sw-wMPDR, and Sw-SRIVE+ is for calculating $\hat{s}(f,t)$ in (4) and (6). $\mathcal{O}(M^2JTF)$ included in the complexity of Sw-IVE and Sw-SRIVE is for further calculating $\hat{\boldsymbol{z}}_j(f,t)$ because $\hat{\boldsymbol{z}}_j(f,t)$ is used for updating $\delta_j(f,t)$. It should be noted that Sw-SRIVE+ does not require $\hat{\boldsymbol{z}}_j(f,t)$ since it changes the update rule of $\delta_j(f,t)$ from (45) to (47).

In summary, the increase in the computational complexity of Sw-wMPDR and Sw-SRIVE+ is only $\mathcal{O}(MJTF)$ required for calculating $\hat{s}(f,t)$ in comparison with that of wMPDR, IVE, and SRIVE. In contrast, Sw-SRIVE requires further additional complexity, $\mathcal{O}(M^2JTF)$, for calculating $\hat{\boldsymbol{z}}_j(f,t)$.

7 Experiment

In this section, we experimentally evaluate the effectiveness of our proposed methods in underdetermined situations, focusing on the following two aspects:

- Effectiveness of introducing the TVG source model to Sw-MPDR when using DOA-based steering vectors.
- Effectiveness of introducing the switching mechanism to SRIVE in underdetermined situations.

Table 2 Computational complexity of each method for updating parameters in each iterative update under assumption $T \gg MJ$

	Complexity
MPDR	$\mathcal{O}(M^2TF)$
wMPDR [12, 13]	$\mathcal{O}(M^2TF)$
Sw-MPDR [5, 6]	$\mathcal{O}(M^2TF + MJTF)$
Sw-wMPDR [proposed]	$\mathcal{O}(M^2TF + MJTF)$
IVE [19, 20]	$\mathcal{O}(M^2TF)$
SRIVE [23]	$\mathcal{O}(M^2TF)$
Sw-IVE [10]	$\mathcal{O}(M^2JTF)$
Sw-SRIVE [proposed]	$\mathcal{O}(M^2JTF)$
Sw-SRIVE+ [proposed]	$\mathcal{O}(M^2TF + MJTF)$

7.1 Experimental condition

We conducted experiments using TIMIT-ConvMix, which is composed of simulated noisy reverberant mixtures. To generate the mixtures, we first concatenated utterances extracted from the TIMIT corpus [36] to obtain a set of single-speaker clean utterance sequences, each of which is 15 s long. Then, we mixed N-utterance sequences and five[4] different additive noise signals extracted from the CHiME-3 dataset [38] after individually reverberating them. We convolved the utterances and the noise signals with Room Impulse Responses (RIRs) labeled "E2A," obtained from the RWCP database [37]. In the database, we used three microphones (numbered 22, 23, and 24) attached to the linear array and randomly assigned nine different source locations corresponding to the azimuths of $10°, 30°, 50°, \ldots, 170°$ to the utterances and the noises. RT_{60} of the E2A's RIRs was 0.3 s. We used all four types of noise signals: bus, cafe, pedestrian area, and street junction, 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 compared four BF-based methods, MPDR [1], wMPDR [11–13], Sw-MPDR [6], and the proposed Sw-wMPDR, and compared two Spatial-Regularization-based methods (SR-based methods), SRIVE [23] and the proposed Sw-SRIVE. For all the methods, we updated spatial filter $\boldsymbol{w}_{j,1}(f)$ or separation matrix $\boldsymbol{W}_j(f)$ 50 times, except for MPDR, which does not use iterative updates. For the methods using a TVG source model, we initialized $v(f,t) = |x_1(f,t)|^2$. For the methods using a switching mechanism (i.e., $J > 1$), we randomly initialized switching weights $\delta_j(f,t)$ at a positive real value in a range of 1 ± 10^{-3} and normalized it to satisfy $\sum_{j=1}^{J} \delta_j(f,t) = 1$. After initialization, we updated $\delta_j(f,t)$ to binary values during parameter optimization. For SR-based methods, we initialized $\boldsymbol{W}_j(f) = \boldsymbol{I}_M$. We applied projection back [39] post-processing to solve the scale ambiguity. For Sw-SRIVE and Sw-SRIVE+, we adopted Spatially-Regularized Single-State (SRSS) initialization [27], which uses the first 25 updates for the initialization using a single-state separation matrix $\boldsymbol{W}_1(f)$, and then estimates $\mathcal{W}$ using the remaining 25 updates after duplicating $\boldsymbol{W}_1(f)$ to $\boldsymbol{W}_2(f), \ldots, \boldsymbol{W}_J(f)$. This technique is useful for avoiding the inter-state permutation problem [9], which occurs in IVA-related methods with a switching mechanism. We used a Hann window for a short-time analysis, where the frame length and the shift size were set to 2048 points (128 ms) and 1024 points (64 ms). We set sampling frequency f_s to 16 kHz. Because we used linear arrays, we set relative TDOA $\boldsymbol{\tau}$ in (3) for the DOA-based steering vectors:

$$\tau_m = f_s \frac{d(m-1)}{c} \cos\left(2\pi \frac{\phi}{360°}\right), \tag{51}$$

where $c = 343$ m/s is the speed of sound and $d = 0.0281$ meter is the distance between adjacent microphones. We used angle labels for source locations in the RWCP database as SOI's direction ϕ. Note that even using the angle labels provided for the database, the steering vector in (3) should contain substantial errors from the true ATF because it does not include the effects of reflection paths.

In the evaluation, we adopted the following as the metrics of the speech enhancement performance [40, 41]: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Noise Ratio (SNR), and Perceptual Evaluation of Speech Quality (PESQ). We calculated SDR and SIR using the MUSEVAL V4 toolkit [42] with its bss_eval_images configuration. We calculated SNR as the power ratio of outputs when we applied the filter to the SOI and noise signals in the mixture. Since both BF-based and SR-based methods aim to estimate the source image, we generated an SOI's reference signal for each mixture by convolving the SOI's clean utterance sequence with the RIR used for generating the mixture after truncating the RIR at 128 ms. Note that the SDR obtained in this paper is different from that obtained using bss_eval_sources, which forgives channel errors that can be accounted for by a time-invariant 512-tap filter. In contrast, bss_eval_images does not allow any distortion including gain errors when calculating SDR. Also, we did not use scale-invariant SDR [43] because both BF-based and SR-based methods aim to estimate the source image, including its scale.

7.2 Evaluation results in underdetermined situations using SDR improvement

7.2.1 Evaluation of BF-based methods

In this section, after examining the effect of the flooring of Sw-wMPDR and wMPDR in underdetermined situations, we compare their performance with the other BF-based methods.

First, we applied Sw-wMPDR and wMPDR [11–13] to the mixture by varying flooring coefficient $\epsilon_{\mathrm{floor}}$ from 10^{-12} to 10^{3}. We calculated the SDR improvement (iSDR) of each method by changing the combination of (N, M) to (3, 2), (4, 2), (4, 3), (5, 2), and (5, 3). We then averaged the iSDRs for all the combinations and showed it in Fig. 1 (hereafter, we refer to the iSDR calculated in this way as "average iSDR" in this paper). In

[4] Because only nine source locations are available for RIRs in the RWCP database [37], we used four noise signals instead of five when $N = 5$.

the figure, the average iSDR of Sw-wMPDR tends to decrease when we set flooring coefficient $\epsilon_{\text{floor}} < 10^{-10}$, denoting that the flooring is essential for Sw-wMPDR to work effectively. The average iSDRs of Sw-wMPDR and wMPDR also tend to decrease as we increase $\epsilon_{\text{floor}} > 10^{-4}$. The above results mean that Sw-wMPDR and wMPDR work effectively by setting appropriate flooring coefficient ϵ_{floor} between 10^{-10} and 10^{-4} when using DOA-based steering vectors in underdetermined situations. In the following experiments, we adopted $\epsilon_{\text{floor}} = 10^{-12}$, 10^{-10}, and 10^{-10} for wMPDR, Sw-wMPDR ($J = 2$), and ($J = 3$) as they achieved the best average iSDRs in Fig. 1.

Next, we compared the BF-based methods for each (N, M) combination in underdetermined situations. The iSDRs are shown in the upper section of Table 3. When we focus on the BF-based methods without the TVG source model, Sw-MPDR decreased the iSDR for combinations $(N, M) \in \{(3, 2), (4, 3), (5, 3)\}$, as we increased the number of switching states J. On the other hand, Sw-wMPDR increased the iSDR for all underdetermined situations. Considering the average iSDR shown in the table, we conclude that Sw-wMPDR improved the DOA-informed speech enhancement performance the best among the BF-based methods.

We analyzed the reason for the above results using the directional responses of estimated spatial filters $\boldsymbol{w}_{j,1}(f)$. Figure 2 shows the directional responses of $\boldsymbol{w}_{j,1}(f)$ to the given ATFs in case $(N, M) = (3, 2)$. On the y-axis, $\boldsymbol{h}(f)$ and $\boldsymbol{h}_n(f)$ represent the true ATFs of the SOI and the nth interference. We determined the true ATF of each source as the primary eigenvector of the SCM of the noiseless source images. The preferred result is that $\boldsymbol{w}_{j,1}(f)$ suppresses $\boldsymbol{h}_n(f)$ for $\forall n$ and f while preserving $\boldsymbol{h}(f)$. The directional response of $\boldsymbol{w}_{j,1}(f)$ to $\boldsymbol{h}(f)$ is calculated as $\min_j |\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{h}(f)|^2$. To enhance clarity in the figure, we normalized each response by dividing it by the maximum response at each frequency. Consequently, the maximum normalized response corresponds to 0 dB (or is represented by white in the figure). Note that we estimated spatial filter $\boldsymbol{w}_{j,1}(f)$ without using the true ATFs but DOA-based steering vectors $\boldsymbol{a}(f)$ calculated by (3). In the figure, when comparing Fig. 2a and b, $\boldsymbol{w}_{j,1}(f)$ of Sw-MPDR [6] reduced the response to true ATF $\boldsymbol{h}(f)$ when the number of switching states J increased from one (i.e., MPDR) to three, corresponding to the iSDR degradation in Table 3. On the other hand, $\boldsymbol{w}_{j,1}(f)$ of Sw-wMPDR preserves $\boldsymbol{h}(f)$ while suppressing $\boldsymbol{h}_n(f)$ better than Sw-MPDR (Fig. 2b and d). As discussed in Section 4.3, Sw-wMPDR thus reduces the interferences more effectively than Sw-MPDR when we use the DOA-based steering vectors.

7.2.2 Evaluation of SR-based methods

We compared the iSDRs of Sw-SRIVE, Sw-SRIVE+, and SRIVE [23] in each (N, M) combination using Fig. 3. Each

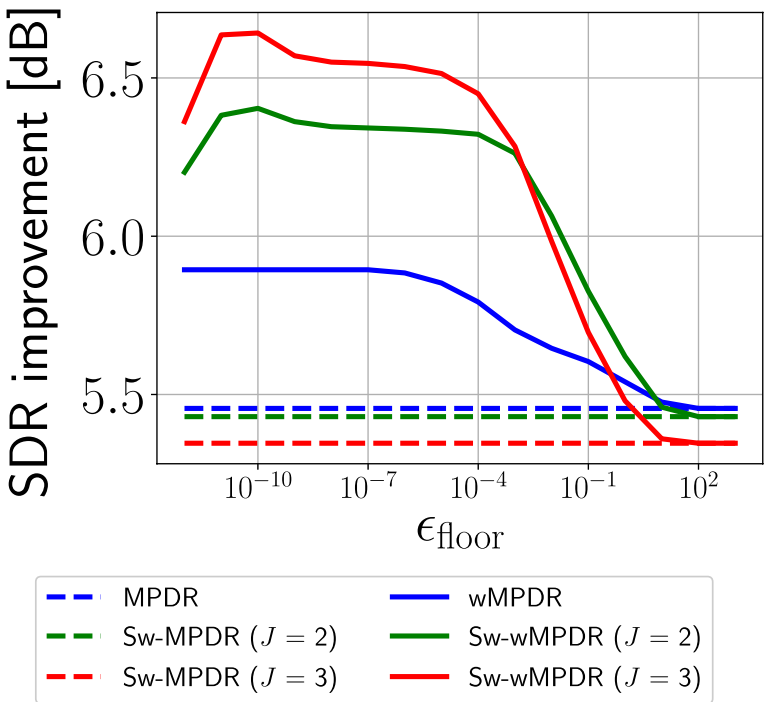

Fig. 1 Average iSDR of each BF-based method when varying flooring coefficient ϵ_{floor}

Table 3 SDR improvement [dB] obtained when changing the number of sources N and the number of microphones M. In each condition (N, M), fonts with **bold** and ***italic bold*** faces show the best scores among BF-based and SR-based methods, respectively

(*N*, *M*)		**(3, 2)**	**(4, 2)**	**(4, 3)**	**(5, 2)**	**(5, 3)**	**Average**
BF-based	MPDR [1]	4.69	5.22	5.43	5.50	6.44	5.46
	Sw-MPDR ($J = 2$) [6]	4.11	5.44	5.00	6.28	6.32	5.43
	Sw-MPDR ($J = 3$) [6]	3.81	5.32	4.93	6.41	6.26	5.35
	wMPDR [12, 13]	5.26	5.36	6.29	5.46	7.10	5.89
	Sw-wMPDR ($J = 2$)	5.88	6.09	6.50	6.21	7.34	6.40
	Sw-wMPDR ($J = 3$)	**6.16**	**6.45**	**6.55**	**6.60**	**7.45**	**6.64**
SR-based	SRIVE [23]	5.75	5.31	6.21	5.45	6.60	5.86
	Sw-SRIVE ($J = 2$)	5.84	5.47	6.21	5.54	6.61	5.93
	Sw-SRIVE ($J = 3$)	6.01	5.63	6.26	5.70	6.63	6.05
	Sw-SRIVE+ ($J = 2$)	6.49	6.20	6.15	***6.34***	6.74	6.38
	Sw-SRIVE+ ($J = 3$)	***6.81***	***6.21***	***6.46***	6.29	***6.84***	***6.52***

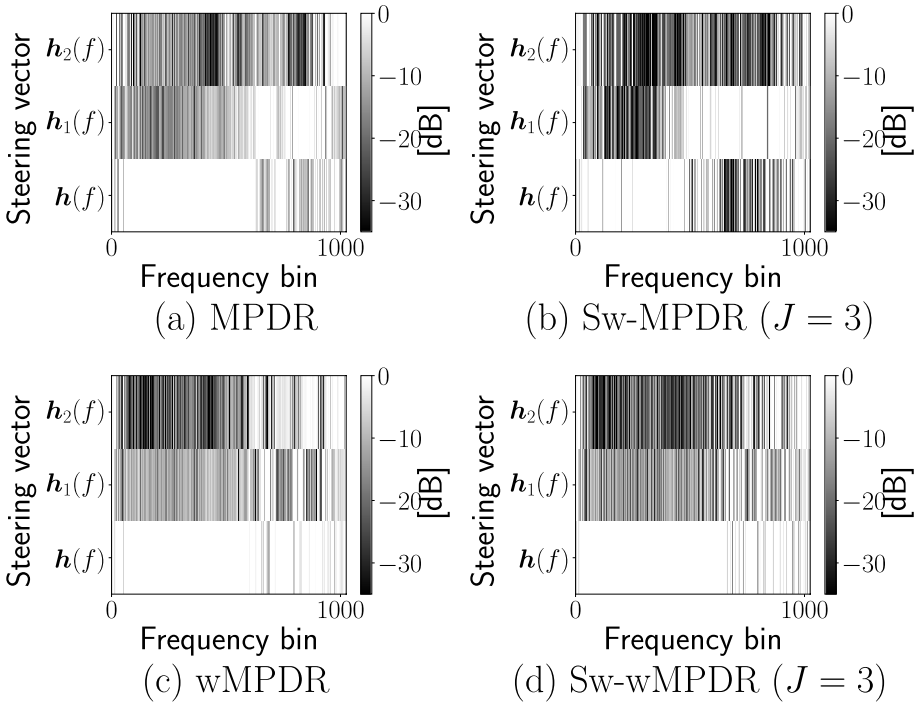

Fig. 2 Directional response in each BF-based method in case $(N, M) = (3, 2)$. The directional response of $\boldsymbol{w}_{j,1}(f)$ to $\boldsymbol{h}(f)$ is calculated as $\min_j |\boldsymbol{w}_{j,1}^{\mathsf{H}}(f)\boldsymbol{h}(f)|^2$. $\boldsymbol{h}(f)$ and $\boldsymbol{h}_n(f)$ represent the true ATFs of the SOI and that of the nth interference. In these subfigures, the SOI, 1st interference, and 2nd interference are located to the azimuths of 10°, 110°, 150°, respectively

subfigure shows the iSDR obtained when varying the weight of spatial regularization λ_{SR} from 10^{-3} to 10^{1}. Note that setting $\lambda_{\mathrm{SR}} = 10^{-3}$ almost corresponds to a case without using spatial regularization (i.e., Sw-IVE [10]). In this experiment, we set $\lambda_{\mathrm{state}} = 0.1$ for Sw-SRIVE and Sw-SRIVE+.

As a general trend, all the methods improved iSDRs by increasing λ_{SR} from 10^{-3} to around 10^{0}. This means that spatial regularization using DOA-based steering vectors effectively improved IVE and Sw-IVE. When comparing Sw-SRIVE and SRIVE at their respective best iSDRs around $\lambda_{\mathrm{SR}} = 10^{0}$, Sw-SRIVE achieved a higher iSDR than SRIVE for all cases except $(N, M) = (5, 3)$. For case $(N, M) = (5, 3)$, Sw-SRIVE and SRIVE were comparable. In addition, Sw-SRIVE+ achieved the highest iSDR of all the methods for almost all the cases. These results suggest the effectiveness of the switching mechanism with robust clustering in underdetermined situations.

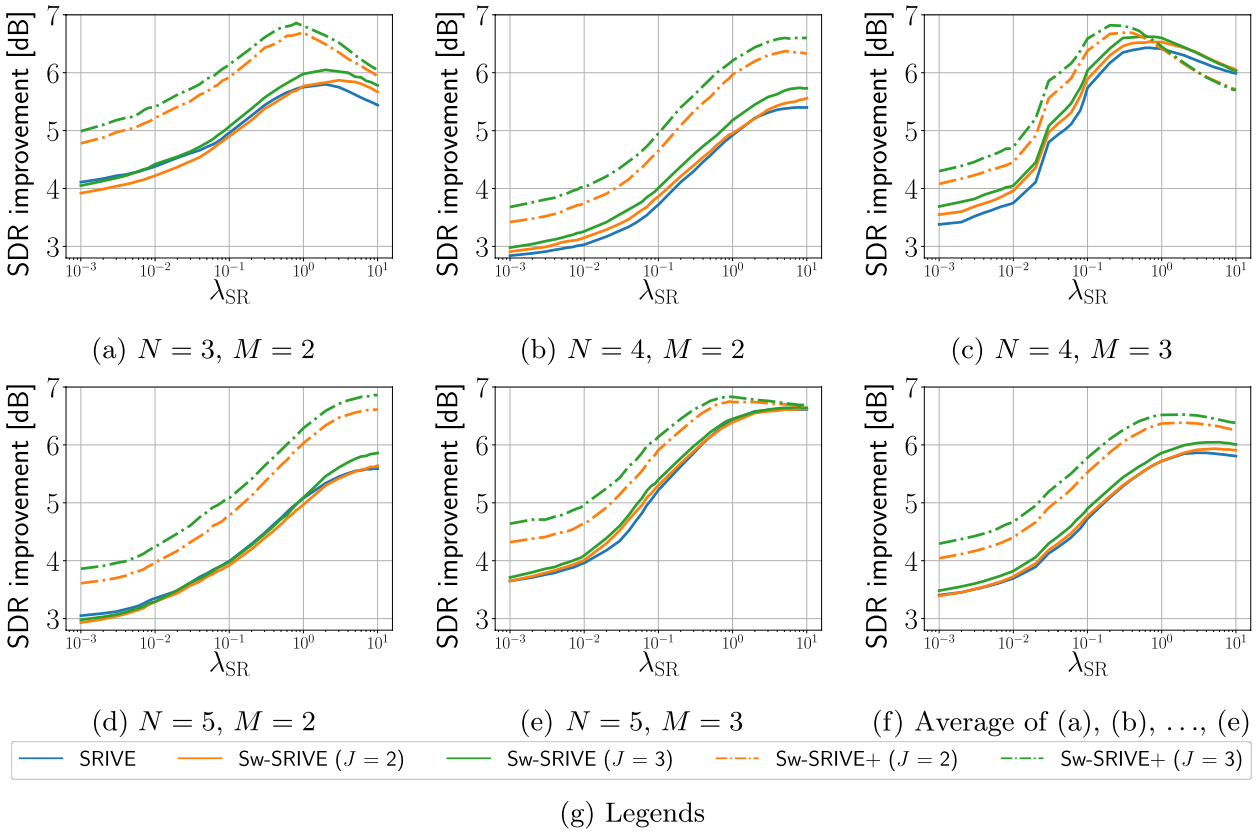

Fig. 3 SDR improvement [dB] by SR-based methods

We posit that disregarding noise covariance matrix $\boldsymbol{\Omega}_j(f)$ for robust clustering results in high consistency in the improvement of iSDR of Sw-SRIVE+ compared to Sw-SRIVE. In Sw-SRIVE, noise signal $\hat{\boldsymbol{z}}_j(f,t)$ is assumed to be stationary within each switching state. However, when we have more than one speech signal, $\hat{\boldsymbol{z}}_j(f,t)$ becomes non-stationary, leading to less effective estimation by Sw-SRIVE. In contrast, Sw-SRIVE+ can mitigate this issue by incorporating the robust clustering, leading to the consistent improvement compared with Sw-SRIVE.

In addition, we compared SR-based methods when varying the weight of state regularization λ_{state}. We calculated and showed the average iSDR of each method in Fig. 4. Note that setting $\lambda_{\mathrm{state}} = 10^{-3}$ almost corresponds to a case without using state regularization. Also, state regularization does not affect SRIVE as it does not use λ_{state}. As seen in the results, each SR-based method (except SRIVE) achieved the highest average iSDR at around $\lambda_{\mathrm{state}} = 10^{-1}$. In particular, Sw-SRIVE+ ($J = 3$) improved average iSDR by about 0.6 dB when increasing λ_{state} from 10^{-3} to 10^{-1}. The above results clearly demonstrate the high effectiveness of introducing state regularization for Sw-SRIVE.

7.2.3 Comparison between proposed methods

We compared proposed methods, Sw-wMPDR and Sw-SRIVE+, using Table 3. For each SR-based method, we set λ_{SR} to a value that achieved the highest average iSDR over all underdetermined settings. As the results show, Sw-SRIVE+ achieved a higher iSDR for case $(N,M) = (3,2)$. For case $(N,M) = (4,3)$, the iSDR of Sw-wMPDR became almost equal to Sw-SRIVE+. Sw-wMPDR achieved a higher iSDR for cases $(N,M) \in \{(4,2),(5,2),(5,3)\}$. The results suggest that Sw-wMPDR tends to show a higher iSDR than Sw-SRIVE+ when the microphone observation includes a relatively large number of sources; Sw-SRIVE+ has the opposite tendency.

7.3 Evaluation in underdetermined situations using other metrics and under various conditions

This subsection evaluates all the methods under comparison using different evaluation metrics other than iSDR and different recording conditions.

7.3.1 Under DOA estimation errors

We investigated how the average iSDR of each method was affected when the SOI's DOA included certain

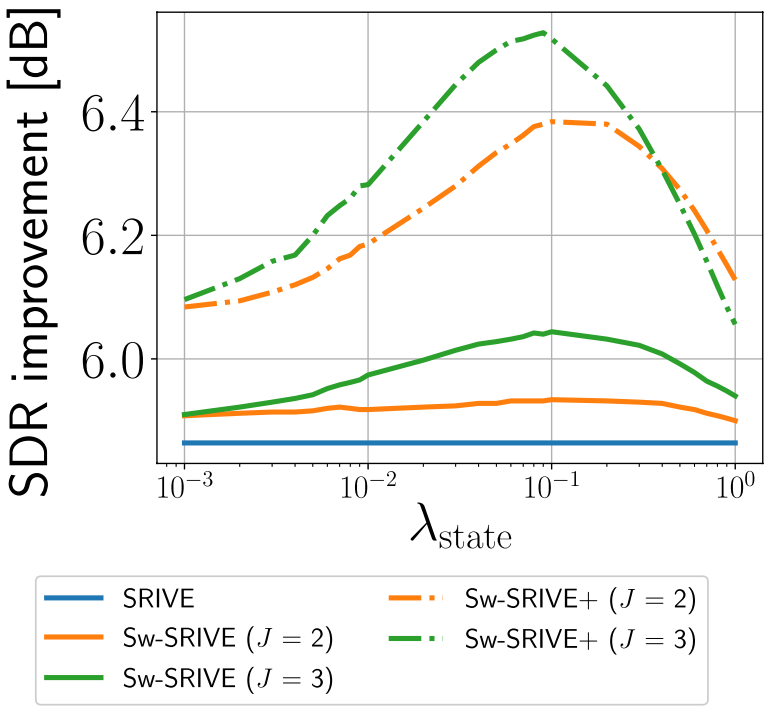

Fig. 4 Average iSDR of each SR-based method when varying state regularization weight λ_{state}

Table 4 Average iSDR obtained in underdetermined situations when setting DOAs with observation error calculated as $\phi + \phi_{\text{err}}$. ϕ_{err} is randomly generated with a uniform distribution in [$-\phi_{\text{max}}$, ϕ_{max}]. Fonts with **bold** and ***italic bold*** faces show the best scores among BF-based and SR-based methods, respectively

	ϕ_{max}	0°	5°	10°	15°	20°
BF-based	MPDR [1]	5.46	5.44	5.42	5.40	5.25
	Sw-MPDR ($J = 2$) [6]	5.43	5.42	5.41	5.42	5.10
	Sw-MPDR ($J = 3$) [6]	5.35	5.34	5.33	5.34	5.00
	wMPDR [12, 13]	5.89	5.88	5.87	5.88	5.81
	Sw-wMPDR ($J = 2$)	6.40	6.39	6.37	6.39	6.26
	Sw-wMPDR ($J = 3$)	**6.64**	**6.63**	**6.61**	**6.63**	**6.47**
SR-based	SRIVE [23]	5.86	5.85	5.83	5.80	5.85
	Sw-SRIVE ($J = 2$)	5.93	5.92	5.89	5.86	5.88
	Sw-SRIVE ($J = 3$)	6.05	6.03	6.02	5.98	6.00
	Sw-SRIVE+ ($J = 2$)	6.38	6.38	6.35	6.31	6.29
	Sw-SRIVE+ ($J = 3$)	***6.52***	***6.51***	***6.48***	***6.44***	***6.49***

estimation errors. The results are shown in Table 4. In this experiment, we added a DOA error, ϕ_{err}, to the actual SOI's DOA, ϕ, as $\phi + \phi_{\text{err}}$. Error ϕ_{err} was randomly generated with a uniform distribution in [$-\phi_{\text{max}}$, ϕ_{max}] for each utterance, where maximum angle ϕ_{max} was varied over $\phi_{\text{max}} = 0°, 5°, 10°, 15°$, and 20°. In the table, as ϕ_{max} increases, the iSDRs of MPDR [1] and Sw-MPDR [6] tend to decrease. In comparison, the decrease in the iSDRs of wMPDR [12, 13] and Sw-wMPDR is smaller than that of the former. This result again confirms the effectiveness of utilizing the TVG source model for DOA-informed speech enhancement. In addition, the table shows that the SR-based method is less affected by the DOA estimation error than the BF-based method. These results confirm the robustness of spatial regularization when using DOA-based steering vectors.

7.3.2 Using SIR, SNR, and PESQ

We evaluated the methods using average SIR improvement (iSIR), average SNR improvement (iSNR), and

Table 5 Average iSDR, iSIR, iSNR, and PESQ obtained in underdetermined situations. Fonts with **bold** and ***italic bold*** faces show the best scores among BF-based and SR-based methods, respectively

		iSDR	iSIR	iSNR	PESQ
BF-based	MPDR [1]	5.46	7.24	1.85	1.14
	Sw-MPDR ($J = 2$) [6]	5.43	8.98	-2.17	1.12
	Sw-MPDR ($J = 3$) [6]	5.35	**9.21**	-3.39	1.12
	wMPDR [12, 13]	5.89	6.38	**4.89**	1.18
	Sw-wMPDR ($J = 2$)	6.40	6.66	4.30	1.19
	Sw-wMPDR ($J = 3$)	**6.64**	6.91	3.77	**1.21**
SR-based	SRIVE [23]	5.86	8.04	3.41	1.19
	Sw-SRIVE ($J = 2$)	5.93	8.30	3.59	1.19
	Sw-SRIVE ($J = 3$)	6.05	8.64	***3.68***	1.20
	Sw-SRIVE+ ($J = 2$)	6.38	***10.23***	2.45	1.22
	Sw-SRIVE+ ($J = 3$)	***6.52***	10.07	2.94	***1.23***

average PESQ. Each score was obtained by averaging each metric over all (N, M) combinations. The results are shown in Table 5. First, focusing on the BF-based methods, Sw-MPDR ($J = 3$) [6] achieved the highest iSIR but the lowest iSNR. On the other hand, wMPDR [11, 13] achieved the highest iSNR. In contrast, Sw-wMPDR ($J = 3$) had the best iSDR and PESQ while its iSIR and iSNR scores are marginal. Next, focusing on the SR-based methods, Sw-SRIVE improved both iSIR and iSNR from SRIVE. In contrast, Sw-SRIVE+ further improved iSIR but yielded degraded iSNR from SRIVE. This might be caused by the introduction of the robust clustering in Section 5.3, which updates switching weight $\delta_j(f,t)$ disregarding the influence of the noise covariance matrices. When comparing our proposed methods, Sw-wMPDR is superior to Sw-SRIVE+ in terms of iSDR and iSNR, and Sw-SRIVE+ is superior in terms of iSIR and PESQ.

7.3.3 Under long reverberation

We evaluated the performances of the proposed methods under long reverberation using RIRs labeled "JR1" from the RWCP database [37]. The RT_{60} of JR1's RIRs was 0.6 s. When creating a mixture using JR1's RIRs, we used the same microphones as with E2A's RIRs. Then, we randomly assigned nine different source locations corresponding to the azimuths of $50°, 60°, 70°, \ldots, 130°$ to the utterances and the noises. We set the minimum angle difference between each utterance to 20°.

We first evaluated the average iSDR of each method in an environment simulated using JR1's RIRs (JR1 environment) by setting various values to the hyperparameters, ϵ_{floor}, λ_{SR}, and λ_{state}, and showed the results in Fig. 5. Each figure shows that our proposed methods outperforms the conventional methods in the JR1 environment when we set appropriate hyperparameters. In addition, we can confirm that a certain range of hyperparameters is effective for both E2A and JR1 environments when using the proposed methods. For example, Sw-wMPDR achieved a higher iSDR than Sw-MPDR when we set ϵ_{floor} as $10^{-10} < \epsilon_{\text{floor}} < 10^{-4}$ (see Figs. 1 and 5a). Sw-SRIVE+ achieved higher iSDR than SRIVE when we set λ_{SR} as $10^{-3} < \lambda_{\text{SR}} < 10^{1}$ (see Figs. 3f and 5b) and $\lambda_{\text{state}} = 10^{-1}$ (see Figs. 4 and 5c). From the above results, we can say that these hyperparameters are not very sensitive to acoustic scenarios. Of course, to obtain the best results for each environment, these hyperparameters need to be tuned. Our future work might investigate in more detail the behavior of the proposed methods depending on the hyperparameters in various environments.

Next, we evaluated the average iSDR for each method using the optimal hyperparameters and showed the results in each environment. Table 6 shows the results. In the table, both Sw-wMPDR and Sw-SRIVE+ achieved higher iSDRs than their conventional methods in both E2A and JR1 environments.

7.3.4 Using convergence speed

We showed the convergence speed of BF-based and SR-based methods in Fig. 6. First, when comparing the convergence speed of IVE, SRIVE, wMPDR, and Sw-wMPDR at the first 25 iterations, wMPDR and Sw-wMPDR are almost the same and faster than SRIVE. This means that the convergence speed does not change significantly between wMPDR and Sw-wMPDR, and the convergence speed of BF-based methods is faster than that of SR-based methods. Also, the convergence speed of SRIVE is faster than that of IVE. It means that spatial regularization accelerates the convergence speed of IVE. Next, at the latter 25 iterations after conducting SRSS initialization, the iSDR of Sw-SRIVE converged from 25 to 30 iterations. On the other hand, the iSDR of Sw-SRIVE+ is improved dramatically from 25 to 30 iterations, followed by gradual improvement until the convergence at around 40 iterations. This result suggests that we should set different iteration numbers for Sw-SRIVE or Sw-SRIVE+ after SRSS initialization.

7.4 Evaluation in determined and overdetermined situations using SDR improvement

For references, we evaluated the speech enhancement methods in determined and overdetermined situations ($N \leq M$). The results are shown in Table 7. In

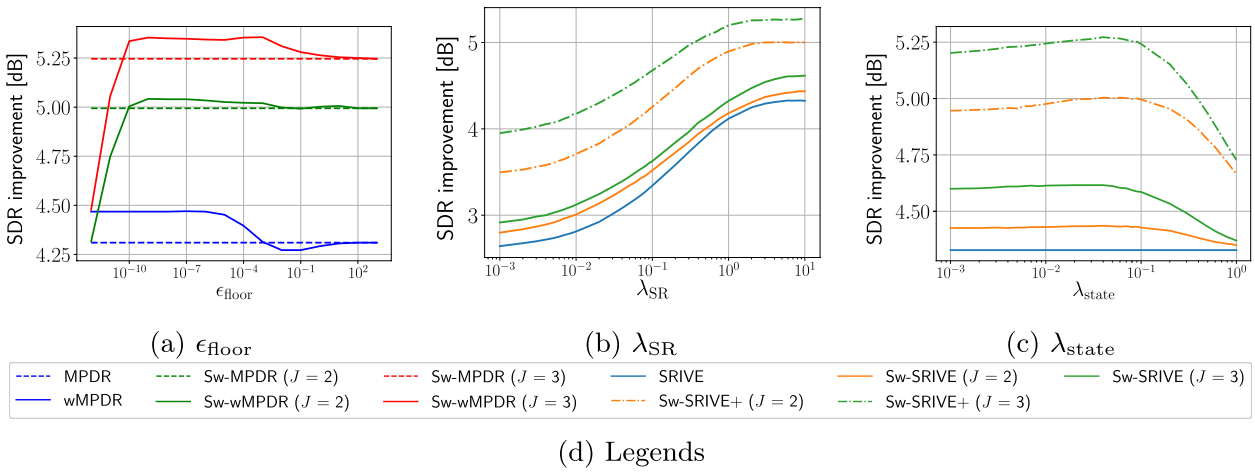

Fig. 5 Average iSDR of BF-based and SR-based methods in JR1 environment ($\text{RT}_{60} = 600$ ms) when varying flooring coefficient ϵ_{floor}, spatial regularization weight λ_{SR}, and state regularization weight λ_{state}

Table 6 iSDR [dB] obtained in both E2A and JR1 environments. Fonts with **bold** and ***italic bold*** faces show the best scores among BF-based and SR-based methods, respectively

		E2A (300 ms)	JR1 (600 ms)
BF-based	MPDR [1]	5.46	4.31
	Sw-MPDR ($J = 2$) [6]	5.43	4.99
	Sw-MPDR ($J = 3$) [6]	5.35	5.25
	wMPDR [12, 13]	5.89	4.47
	Sw-wMPDR ($J = 2$)	6.40	5.04
	Sw-wMPDR ($J = 3$)	**6.64**	**5.36**
SR-based	SRIVE [23]	5.86	4.33
	Sw-SRIVE ($J = 2$)	5.93	4.44
	Sw-SRIVE ($J = 3$)	6.05	4.62
	Sw-SRIVE+ ($J = 2$)	6.38	5.00
	Sw-SRIVE+ ($J = 3$)	***6.52***	***5.27***

this experiment, we adjusted weight λ_{SR} to a value that achieved the highest iSDR for each (N, M) combination.

As seen in the results, wMPDR achieved the highest iSDRs for BF-based methods. For SR-based methods, Sw-SRIVE+ ($J = 3$) achieved the highest iSDR for cases $(N, M) \in \{(2, 2), (3, 3)\}$. On the other hand, SRIVE achieved the highest iSDR for $(N, M) = (2, 3)$. These results mean that the switching mechanism used in our proposed methods becomes less effective in determined and overdetermined situations, i.e., when we have more microphones than necessary. The performance degradation of the proposed methods in determined and overdetermined situations may result from employing the switching mechanism with DOA-based steering vectors. Generally, the switching mechanism enables more effective noise reduction when we have fewer microphones than necessary. However, it may start introducing distortion to the SOI when we have excess microphones. Notably, this issue does not arise when relatively accurate steering vectors are available, as demonstrated in previous articles [9, 10].

It may be worth noting that when comparing the results of BF-based and SR-based methods, the latter show much higher iSDRs for all cases (note that there is no significant difference between these methods in underdetermined situations, as shown in Table 3). This difference may be caused by how they use the DOA-based ATFs. The BF-based methods use them in the distortionless constraint as a hard constraint, while the SR-based methods use them in the spatial regularization as a soft constraint. Using the soft constraint might make the SR-based methods relatively robust to the modeling errors in the DOA-based ATF, even in determined and overdetermined situations.

Table 7 also shows for reference the average iSDR obtained when using IVA [17], Sw-IVA [9], IVE [19, 20], and Sw-IVE [10]. When evaluating the iSDR of IVA and Sw-IVA, we used oracle information to select the SOI from the separated signals. We calculated the iSDRs of IVE and Sw-wIVE as those of SR-based methods with $\lambda_{\text{SR}} = 0$. Compared with the SR-based methods, IVA and Sw-IVA achieved much higher iSDRs, and IVE and Sw-IVE yielded much lower iSDRs. The higher iSDRs obtained by IVA and Sw-IVA may be caused partially by using the oracle information for the SOI selection. In contrast, the lower iSDRs obtained by IVE and Sw-IVE show the effectiveness of the SR-based methods.

To summarize our experimental results, we confirmed that our proposed methods, Sw-wMPDR and Sw-SRIVE,

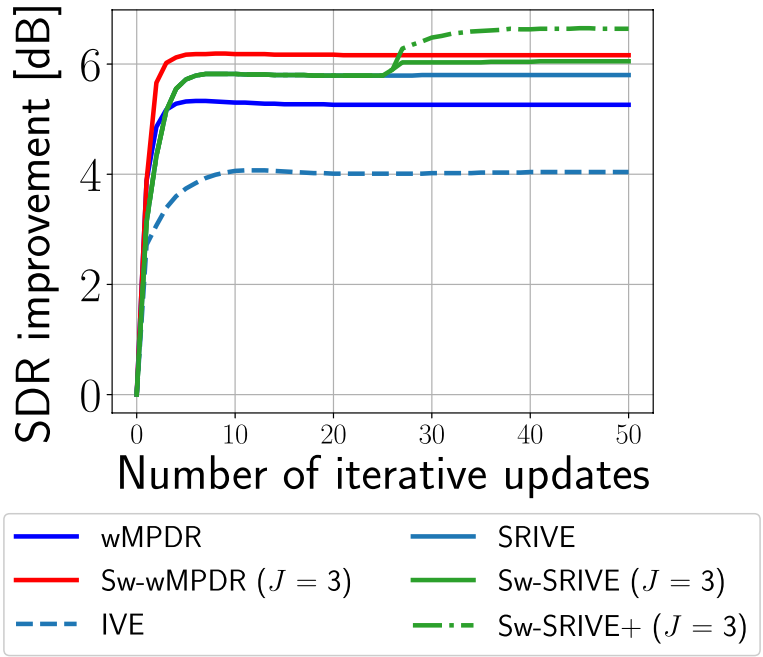

Fig. 6 iSDR when varying the number of iterative updates in case $(N, M) = (3, 2)$

Table 7 iSDR [dB] obtained in determined and overdetermined situations ($N \leq M$). Fonts with **bold** and ***italic bold*** faces show the best scores among BF-based and SR-based methods, respectively

		(2, 2)	(2, 3)	(3, 3)
BF-based	MPDR [1]	3.82	1.19	3.83
	Sw-MPDR ($J = 2$) [6]	1.70	0.99	3.32
	Sw-MPDR ($J = 3$) [6]	1.52	0.96	3.27
	wMPDR [12, 13]	**5.53**	**3.49**	5.21
	Sw-wMPDR ($J = 2$)	5.51	3.42	**5.22**
	Sw-wMPDR ($J = 3$)	5.29	3.26	5.19
SR-based	SRIVE [23]	7.38	***8.27***	5.81
	Sw-SRIVE ($J = 2$)	7.61	8.06	5.87
	Sw-SRIVE ($J = 3$)	7.64	7.82	6.05
	Sw-SRIVE+ ($J = 2$)	7.76	8.01	6.70
	Sw-SRIVE+ ($J = 3$)	***7.95***	8.17	***6.85***
BSS-based	IVA [17]	8.60	9.89	7.80
	Sw-IVA ($J = 2$) [9]	8.75	9.88	8.43
	Sw-IVA ($J = 3$) [9]	8.90	10.02	8.63
	IVE [19, 20]	4.64	4.04	3.08
	Sw-IVE ($J = 2$) [10]	4.77	4.12	3.08
	Sw-IVE ($J = 3$) [10]	4.73	3.99	3.15

achieved more accurate speech enhancement than Sw-MPDR [6] and SRIVE [23] in underdetermined situations ($N > M$), respectively. However, under general scenarios where overdetermined and determined situations can arise, a method must be devised that selects the optimal number of switches with or without prior knowledge of the situations. Future work will tackle this situation.

8 Conclusion

We proposed methods that enhance a single SOI using the prior knowledge of the SOI's DOA in underdetermined situations. We developed Sw-wMPDR by introducing a time-varying Gaussian source model to Sw-MPDR. Using the model and adopting appropriate source variance flooring, Sw-wMPDR effectively preserves the SOI while suppressing the interference speech signals. We also developed Sw-SRIVE by introducing a switching mechanism to SRIVE. We proposed two new techniques, robust clustering and state regularization, for Sw-SRIVE to make it work effectively in underdetermined situations. With robust clustering, we developed Sw-SRIVE+ based on the modified updating rule of switching weights. For state regularization, we introduced a new regularization term for stabilizing the optimization of Sw-SRIVE+. Experiments showed that both of our proposed methods, Sw-wMPDR and Sw-SRIVE+, achieved better DOA-informed SOI enhancement in terms of SDR improvement than the conventional methods, Sw-MPDR and SRIVE, in underdetermined situations.

Future work may develop online methods for DOA-informed speech enhancement that can cope with scenarios where speakers are moving.

Appendix

Here, we introduce how to derive (10). First, using (4) and (6), $\boldsymbol{x}(f,t)$ can be expressed by $\boldsymbol{W}_j(f)$, $\delta_j(f,t)$, $\hat{s}(f,t)$, and $\hat{\boldsymbol{z}}(f,t)$:

$$\begin{aligned}\boldsymbol{x}(f,t) &= \sum_j \delta_j(f,t)\boldsymbol{W}_j^{-\mathsf{H}}(f)\underbrace{\begin{bmatrix}\hat{s}(f,t)\\ \hat{\boldsymbol{z}}(f,t)\end{bmatrix}}_{\hat{\boldsymbol{y}}(f,t)} \\ &= \boldsymbol{W}^{-\mathsf{H}}(f,t)\hat{\boldsymbol{y}}(f,t),\end{aligned} \tag{52}$$

where $\boldsymbol{W}(f,t) = \sum_j \delta_j(f,t)\boldsymbol{W}_j(f)$. According to this relationship, we can write $p(\boldsymbol{x}(f,t)) = |\det \boldsymbol{W}(f,t)|^2 p(\hat{\boldsymbol{y}}(f,t))$, and thus can derive a negative log-likelihood function for a given observed signal $\mathcal{X}$:

$$\begin{aligned}&\mathcal{L}_{\mathrm{NL}}(\mathcal{X};\mathcal{W},\mathcal{V},\mathcal{D}) = -\sum_{f,t}\log p(\boldsymbol{x}(f,t)) \\ &= -\sum_{f,t}\log\left[|\det \boldsymbol{W}(f,t)|^2 p(\hat{\boldsymbol{y}}(f,t))\right] \\ &= -\sum_{f,t}\Bigg[2\log|\det\sum_j \delta_j(f,t)\boldsymbol{W}_j(f)| \\ &\qquad + \log p(\hat{\boldsymbol{y}}(f,t))\Bigg].\end{aligned} \tag{53}$$

Because $\delta_j(f,t)$ takes 1 for a state j and takes 0 for the other states,

$$\begin{aligned}&\log|\det\sum_j \delta_j(f,t)\boldsymbol{W}_j(f)| \\ &= \sum_j \delta_j(f,t)\log|\det \boldsymbol{W}_j(f)|.\end{aligned} \tag{54}$$

Based on (6) and (9) using the same property of $\delta_j(f,t)$,

$$\begin{aligned}&\log p(\hat{\boldsymbol{y}}(f,t)) = \log p(\hat{s}(f,t)) + \log p(\hat{\boldsymbol{z}}(f,t)) \\ &= \left(\log p\left(\sum_j \delta_j(f,t)\hat{s}_j(f,t)\right)\right. \\ &\qquad \left. + \log p\left(\sum_j \delta_j(f,t)\hat{\boldsymbol{z}}_j(f,t)\right)\right) \\ &= \sum_j \delta_j(f,t)\left(\log p(\hat{s}_j(f,t)) + \log p(\hat{\boldsymbol{z}}_j(f,t))\right).\end{aligned} \tag{55}$$

By substituting (54) and (55) to (53), we obtain

$$\begin{aligned}&\mathcal{L}_{\mathrm{NL}}(\mathcal{X};\mathcal{W},\mathcal{V},\mathcal{D}) = \sum_{j,f,t}\delta_j(f,t)\Big(-2\log|\det \boldsymbol{W}_j(f)| \\ &- \log p(\hat{s}_j(f,t)) - \log p(\hat{\boldsymbol{z}}_j(f,t))\Big).\end{aligned} \tag{56}$$

By inserting the TVG source model in (8) to the second term in the parentheses of the above function, we can derive (10).

Abbreviations

SOI	Speech signal of Interest
DOA	Direction of Arrival
BF	Beamformer
BSS	Blind Source Separation
TVG	Time-varying Gaussian
ATF	Acoustic Transfer Function
SCM	Spatial Covariance Matrix
MLE	Maximum Likelihood Estimation
MVDR	Minimum Variance Distortionless Response
MPDR	Minimum Power Distortionless Response
Sw-MPDR	Switching MPDR
wMPDR	Weighted MPDR
Sw-wMPDR	Switching wMPDR
ICA	Independent Component Analysis
IVA	Independent Vector Analysis
IVE	Independent Vector Extraction
SRIVA	Spatially-Regularized IVA
SRIVE	Spatially-Regularized IVE
Sw-IVE	Switching IVE
Sw-SRIVE	Switching SRIVE
NN	Neural Network
STFT	Short-Time Fourier Transform
TDOA	Time Delays of Arrival
RIR	Room Impulse Response
SDR	Signal-to-Distortion Ratio
SIR	Signal-to-Interference Ratio
SNR	Signal-to-Noise Ratio
PESQ	Perceptual Evaluation of Speech Quality

Acknowledgements

Not applicable.

Authors' contributions

TU designed the proposed methods, evaluated the experiments, and wrote the manuscript. TN supervised every part of the research, provided invaluable technical feedback, and refined the manuscript. RI provided invaluable technical feedback on the manuscript draft and refined the manuscript. SA and SM reviewed and revised the manuscript. All the authors read and approved the final manuscript.

Funding

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

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

SM is an editor of the collection "Advanced Signal Processing and Machine Learning for Acoustic Scene Analysis and Signal Enhancement," where this manuscript is submitted.

Received: 29 February 2024 Accepted: 19 September 2024
Published online: 08 October 2024

References

1. O.L. Frost, An algorithm for linearly constrained adaptive array processing. Proc. IEEE **60**(8), 926–935 (1972)
2. B.D. Van Veen, K.M. Buckley, Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag. **5**(2), 4–24 (1988)
3. L.C. Parra, C.V. Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming. IEEE Trans. Speech Audio Process. **10**(6), 352–362 (2002)

4. F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, H. Asoh, Detection and separation of speech event using audio and video information fusion and its application to robust speech interface. EURASIP J. Adv. Signal Process. **2004**, 1–12 (2004)
5. K. Yamaoka, A. Brendel, N. Ono, S. Makino, M. Buerger, T. Yamada, W. Kellermann, in *Proc. EUSIPCO*. Time-frequencybin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios (IEEE, Piscataway, 2018), pp. 1582–1586
6. K. Yamaoka, N. Ono, S. Makino, Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement. IEEE/ACM Trans. ASLP **29**, 3461–3475 (2021)
7. R. Ikeshita, N. Kamo, T. Nakatani, Blind signal dereverberation based on mixture of weighted prediction error models. IEEE Signal Process. Lett. **28**, 399–403 (2021)
8. R. Ikeshita, K. Kinoshita, N. Kamo, T. Nakatani, Online speech dereverberation using mixture of multichannel linear prediction models. IEEE Signal Process. Lett. **28**, 1580–1584 (2021)
9. T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, S. Araki, Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms. IEEE/ACM Trans. ASLP **30**, 1032–1047 (2022)
10. T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, S. Araki, in *Proc. ICA*. Switching independent vector extraction and its joint optimization with weighted prediction error dereverberation (Springer, Cham, 2022)
11. T. Nakatani, K. Kinoshita, A unified convolutional beamformer for simultaneous denoising and dereverberation. IEEE Signal Process. Lett. **26**(6), 903–907 (2019)
12. B.J. Cho, J.M. Lee, H.M. Park, A beamforming algorithm based on maximum likelihood of a complex Gaussian distribution with time-varying variances for robust speech recognition. IEEE Signal Process. Lett. **26**(9), 1398–1402 (2019)
13. C. Boeddeker, T. Nakatani, K. Kinoshita, R. Haeb-Umbach, in *Proc. ICASSP*. Jointly optimal dereverberation and beamforming (IEEE, Piscataway, 2020), pp. 216–220
14. A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications. Neural Netw. **13**(4–5), 411–430 (2000)
15. T. Kim, H.T. Attias, S.Y. Lee, T.W. Lee, Blind source separation exploiting higher-order frequency dependencies. IEEE Trans. ASLP **15**(1), 70–79 (2007)
16. A. Hiroe, in *Proc. ICA*. Solution of permutation problem in frequency domain ICA, using multivariate probability density functions (Springer, Cham, 2006), pp. 601–608
17. N. Ono, S. Miyabe, in *Proc. LVA/ICA*. Auxiliary-function-based independent component analysis for super-Gaussian sources (Springer, Cham, 2010), pp. 165–172
18. Z. Koldovsky, P. Tichavsky, Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence. IEEE Trans. Signal Process. **67**(4), 1050–1064 (2018)
19. R. Scheibler, N. Ono, in *Proc. WASPAA*. Independent vector analysis with more microphones than sources (IEEE, Piscataway, 2019), pp. 185–189
20. R. Ikeshita, T. Nakatani, S. Araki, Block coordinate descent algorithms for auxiliary-function-based independent vector extraction. IEEE Trans. Signal Process. **69**, 3252–3267 (2021)
21. A.H. Khan, M. Taseska, E.A. Habets, in *Proc. LVA/ICA*. A geometrically constrained independent vector analysis algorithm for online source extraction (Springer, Cham, 2015), pp. 396–403
22. L. Li, K. Koishida, in *Proc. ICASSP*. Geometrically constrained independent vector analysis for directional speech enhancement (IEEE, Piscataway, 2020), pp. 846–850
23. A. Brendel, T. Haubner, W. Kellermann, A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis. IEEE Trans. Signal Process. **68**, 3545–3558 (2020)
24. K. Goto, T. Ueda, L. Li, T. Yamada, S. Makino, in *Proc. EUSIPCO*. Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering (IEEE, Piscataway, 2022), pp. 757–761
25. Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, in *Proc. ICASSP*. Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis (IEEE, Piscataway, 2018), pp. 746–750
26. T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, S. Makino, Blind and spatially-regularized online joint optimization of source separation, dereverberation, and noise reduction. IEEE/ACM Trans. ASLP **32**, 1157–1172 (2024)
27. T. Ueda, T. Nakatani, R. Ikeshita, S. Araki, S. Makino, in *Proc. APSIPA*. Spatially-regularized switching independent vector analysis (IEEE, Piscataway, 2023), pp. 2040–2046
28. T. Nakatani, R. Ikeshita, N. Kamo, K. Kinoshita, S. Araki, H. Sawada, in *Proc. EUSIPCO*. Switching convolutional beamformer (IEEE, Piscataway, 2021), pp. 266–270
29. J. Bitzer, K.U. Simmer, in *Microphone Arrays: Signal Processing Techniques and Applications*. Superdirective microphone arrays (Springer, Cham, 2001), pp. 19–38
30. N.Q. Duong, E. Vincent, R. Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans. ASLP **18**(7), 1830–1840 (2010)
31. N. Ito, R. Ikeshita, H. Sawada, T. Nakatani, A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel wiener filter. IEEE/ACM Trans. ASLP **29**, 1950–1965 (2021)
32. F. Grondin, J.S. Lauzon, J. Vincent, F. Michaud, in *Proc. Interspeech*. GEV beamforming supported by DOA-based masks generated on pairs of microphones (IEEE, Piscataway, 2020), pp. 3341–3345
33. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
34. T. Taniguchi, N. Ono, A. Kawamura, S. Sagayama, in *Proc. HSCMA*. An auxiliary-function approach to online independent vector analysis for real-time blind source separation (IEEE, Piscataway, 2014), pp. 107–111
35. T. Nakatani, K. Kinoshita, in *Proc. EUSIPCO*. Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation (IEEE, Piscataway, 2019), pp. 1–5
36. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus, in Linguistic Data Consortium (Philadelphia), 1993
37. S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, in *Proc. LREC*. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition (ELCA, France, 2000), pp. 965–968. https://www.elra.eu/
38. J. Barker, R. Marxer, E. Vincent, S. Watanabe, in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines (IEEE, Piscataway, 2015), pp. 504–511
39. N. Murata, S. Ikeda, A. Ziehe, An approach to blind source separation based on temporal structure of speech signals. Neurocomputing **41**(1–4), 1–24 (2001)
40. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. ASLP **14**(4), 1462–1469 (2006)
41. A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, in *Proc. ICASSP*. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs (IEEE, Piscataway, 2001), vol. 2, pp. 749–752.
42. Museval. https://github.com/sigsep/sigsep-mus-eval. Accessed 15 Feb 2024
43. J. Le Roux, S. Wisdom, H. Erdogan, J.R. Hershey, in *Proc. ICASSP*. Sdr–half-baked or well done? (IEEE, 2019), pp. 626–630

Publisher's Note