Constant Separating Vector-based Blind Source Extraction and Dereverberation for a Moving Speaker

Tetsuya Ueda* , Shoji Makino*
*Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kita-Kyushu, Fukuoka 808-0135, Japan
Email: t.ueda@akane.waseda.jp, s.makino@waseda.jp

***Abstract*—This paper proposes a multi-channel speech extraction method for moving sound sources in a long reverberant environment. Constant Separating Vector (CSV) mixing model has been devised for batch processing speech extraction to extract a moving target speech stably. Also, based on this mixture model, an update algorithm using auxiliary function technology has been proposed as a fast and stable source extraction. However, source extraction performance will be limited when the reverberation time is long. In recent years, joint optimization technique has been researched to achieve effective dereverberation and source extraction simultaneously under highly reverberant environments. However, the extension to the CSV mixing model is yet to be discovered. To realize moving source extraction under a highly reverberant environment, we derive the update algorithm when the dereverberation mechanism is installed in the conventional method. In our proposed method, we estimate a dereverberation system focusing only on the extracted target sound, which achieves effective source extraction with a small additional computational cost. Our experiment shows that the proposed algorithm achieves sufficient blind dereverberation and source extraction.**

***Index Terms*—Blind Source Extraction, Constant Separating Vector, Auxiliary Function, Blind Dereverberation, Moving Source Extraction.**

I. Introduction

In the development of speech applications, there is an increasing need to extract specific sounds from microphone signals recorded with a mixture of different sounds. Because we cannot always use spatial information about the sounds to extract from the microphones, Blind Source Extraction (BSE) has been researched recently. BSE extracts speech sounds using only the observed microphone signals without any prior information.

As a basis for BSE, Independent Component Analysis (ICA) [1], [2] based blind source separation has been studied to achieve source separation by maximizing the independence between separated signals. Recently, several algorithms based on frequency-domain ICA have been developed [3]–[7], which provide flexibility in utilizing various models for the time-frequency representations of source signals and array responses. For example, Independent Vector Analysis (IVA) simultaneously solves source separation and frequency-domain permutation problems by assuming that the magnitudes of the frequency components originating from the same source tend to vary coherently over time [3], [4]. IVA has been extended to auxiliary function-based IVA (AuxIVA) [5], [6] as a fast approach with rapid convergence and a stable calculation. Moreover, the above methods are extended to the BSE scenario to extract N sources using $M(> N)$ microphones assuming the noise signals derive from the $M - N$ outputs [8]–[10]. In particular, auxiliary function-based IVE (AuxIVE) [9], [10] can optimally extract sources much fewer in number than the number of microphones and skip most of the calculations for estimating the noise statistics.

In order to perform source separation/extraction in a more realistic environment, it is necessary to consider the situation where the sound sources are moving. One approach for time-varying mixtures is to perform time-varying source separation/extraction at each time interval by updating the separation system using forgetting coefficients [7]. Although this approach is useful, it suffers from the following issues: 1) the performance of the source extraction degrades with respect to the movement of the sound source, 2) the sound objects to be extracted change with time (discontinuity problem), and 3) the forgetting coefficients depend on the situation. For the above issues, a source extraction method based on the Constant Separating Vector (CSV) mixing model has been studied in recent years [11], [12]. The CSV-based methods estimate a time-invariant separation filter, while the mixing parameters are time-varying. This model enables us to avoid the discontinuity problem and eliminate the need to set forgetting coefficients. Its extension using the auxiliary function-based algorithm (CSV-AuxIVE) has been proposed recently, which holds the advantages of the CSV mixing model and the rapid and stable convergence [13].

One drawback of the above CSV-AuxIVE is that it needs to solve the degradation of source extraction due to long reverberation. We can solve the degradation by applying the Weighted Prediction Error (WPE)-based dereverberation methods [14]. Although there are researches [15], [16] to optimize the dereverberation filter based on the output of both target source and noise to realize effective source extraction, its extension to the CSV mixing model is unclear. To realize moving source extraction under a highly reverberant environment, we propose a CSV-AuxIVE-based source extraction algorithm with WPE-based dereverberation, which we call CSV-WPEIVE. In the proposed method, we update the dereverberation filter based

on the output of only the extracted signal. We will show the extraction performance of our proposed algorithm using a simulation experiment.

The following structure in this paper is described. First, the source separation model and cost function used in this study are described in Section II. Next, the update algorithm proposed in this paper based on the cost function is described in Section III. The performance of the proposed algorithm through evaluation experiment is described in Section IV, and conclusion is given in Section V.

II. Problem formulation

Let us consider the situation where we extract a source of interest (SOI) from M microphones. We represent the observed signals $\boldsymbol{x}_{f,n} = [x_{1,f,n}, \ldots, x_{M,f,n}]^{\mathsf{T}} \in \mathbb{C}^M$, the SOI $s_{f,n} \in \mathbb{C}$, and the background signals $\boldsymbol{z}_{f,n} \in \mathbb{C}^{M-1}$ at each time frame $n = 1, \ldots, N$ and frequency bin $f = 1, \ldots, F$ in the STFT domain.

Let the frames be divided into $T \geq 1$ time intervals called blocks, and each block includes N_{b} frames for simplicity, hence $N = TN_{\mathrm{b}}$. Hereafter, we treat the frame index $\{(t-1)N_{\mathrm{b}} + 1, ..., tN_{\mathrm{b}}\}$ as the same block index t for $t = 1, \ldots, T$. For example, we denote $\boldsymbol{x}_{f,(t-1)N_{\mathrm{b}}+n'}$ as $\boldsymbol{x}_{f,t,n'}$ for $t = 1, \ldots, T$ and $n' = 1, \ldots, N_b$.

In the CSV mixing model, the relation between $s_{f,t,n'}$, $\boldsymbol{z}_{f,t,n'}$, and $\boldsymbol{x}_{f,t,n'}$ can be written in a semi-time-varying model:

$$\boldsymbol{x}_{f,t,n'} = \boldsymbol{A}_{f,t} \begin{bmatrix} s_{f,t,n'} \\ \boldsymbol{z}_{f,t,n'} \end{bmatrix}, \tag{1}$$

where $\boldsymbol{A}_{f,t}$ is a mixing matrix parameterized as

$$\boldsymbol{A}_{f,t} = [\boldsymbol{a}_{f,t}\ \boldsymbol{Q}_{f,t}] = \begin{bmatrix} \gamma_{f,t} & \boldsymbol{h}_f^{\mathsf{H}} \\ \boldsymbol{g}_{f,t} & \frac{1}{\gamma_{f,t}}(\boldsymbol{g}_{f,t}\boldsymbol{h}_f^{\mathsf{H}} - \boldsymbol{I}_{M-1}) \end{bmatrix}. \tag{2}$$

Similarly, the separation model can be written as follows:

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

where $\boldsymbol{W}_{f,t} = \boldsymbol{A}_{f,t}^{-1}$ is a separation matrix parameterized as

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

The time-invariant separation filter $\boldsymbol{w}_f$ in (4) enables us to extract one source stably. Recently, source extraction algorithm using the CSV mixing model and the auxiliary function update has been proposed as fast and stable convergence [13]. However, in a highly reverberant condition where the length of the room impulse responses is longer than the STFT frame length, the ability of source extraction will be limited. In this paper, we extend the separation model in (3) by introducing a dereverberation filter $\boldsymbol{D}_f \in \mathbb{C}^{ML \times M}$:

$$\boldsymbol{y}_{f,t,n'} = \boldsymbol{x}_{f,t,n'} - \boldsymbol{D}_f^{\mathsf{H}} \overline{\boldsymbol{x}}_{f,t,n'}, \tag{5}$$

$$\begin{bmatrix} s_{f,t,n'} \\ \boldsymbol{z}_{f,t,n'} \end{bmatrix} = \boldsymbol{W}_{f,t}^{\mathsf{H}} \boldsymbol{y}_{f,t,n'}, \tag{6}$$

where $\overline{\boldsymbol{x}}_{f,t,n'} \in \mathbb{C}^{ML}$ is a vector containing a past observation sequence for L frames. Hereafter, we omit the index n' to reduce redundancy.

Next, we assume the same probabilistic model proposed by the previous research [8], [12], [13]. Let $p(\boldsymbol{s}_t)$ denote the joint pdf of the SOI vector component $\boldsymbol{s}_t = [s_{1,t}, \ldots, s_{F,t}] \in \mathbb{C}^F$ and $p(\boldsymbol{z}_{f,t})$ denote the pdf of $\boldsymbol{z}_{f,t}$, respectively. Here, we assume $\boldsymbol{s}_t$ and $\boldsymbol{z}_{f,t}$ are mutually independent over all times and frequencies:

$$p(\{\boldsymbol{s}_t, \boldsymbol{z}_{f,t}\}_{f,t}) = \prod_t p(\boldsymbol{s}_t) \prod_{f,t} p(\boldsymbol{z}_{f,t}). \tag{7}$$

We set $p(\boldsymbol{s}_t)$ as the following pdf to reflect the block-dependent variance:

$$p(\boldsymbol{s}_t) = g\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right)\left(\prod_{f=1}^F \hat{\sigma}_{f,t}\right)^{-2}, \tag{8}$$

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

$$g\left(\left\{\frac{s_{f,t}}{\hat{\sigma}_{f,t}}\right\}_f\right) = C\exp(-G_R(r_t)), \tag{9}$$

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

$$p(\boldsymbol{z}_{f,t}) = \mathcal{N}(\boldsymbol{0}_{M-1}, \boldsymbol{C}_{\boldsymbol{z}_{f,t}}), \tag{10}$$

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

From the above assumptions, we can obtain the negative log-likelihood of the given signal $\mathcal{X} = \{x_{m,f,t}\}_{m,f,t}$:

$$\mathcal{L}(\mathcal{X}) \overset{c}{=} \frac{1}{T}\sum_t^T \Bigg\{ \mathbb{E}\left[G_R\left(r_t\right)\right] + \sum_f \Bigg(\log \hat{\sigma}_{f,t}^2 + \mathbb{E}\left[\boldsymbol{z}_{f,t}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{z}_{f,t_{\mathrm{b}}}}^{-1} \boldsymbol{z}_{f,t}\right] - \log|\gamma_{f,t}|^{2(M-2)} \Bigg) \Bigg\}, \tag{11}$$

where $\overset{c}{=}$ denotes equality up to the constant terms. By applying auxiliary function techniques [6], we can obtain the following auxiliary function to be minimized:

$$\mathcal{L}_{\mathrm{aux}}(\mathcal{X}) \overset{c}{=} \frac{1}{T}\sum_{t=1}^T \sum_{f=1}^F \Bigg\{ \frac{1}{2}\frac{\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{V}_{f,t} \boldsymbol{w}_f}{\hat{\sigma}_{f,t}^2} + \log \hat{\sigma}_{f,t}^2 + \mathbb{E}\left[\boldsymbol{z}_{f,t}^{\mathsf{H}} \boldsymbol{C}_{\boldsymbol{z}_{f,t_{\mathrm{b}}}}^{-1} \boldsymbol{z}_{f,t}\right] - (M-2)\log|\gamma_{f,t}|^2 \Bigg\}, \tag{12}$$

where

$$\boldsymbol{V}_{f,t} = \mathbb{E}\left[\psi(r_t)\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathsf{H}}\right]. \tag{13}$$

III. Optimization process

We use a coordinate descent method to reduce the cost function in (12) by repeatedly updating each $\mathcal{W} = \{\boldsymbol{w}_f\}_f$, $\mathcal{D} = \{\boldsymbol{D}_f\}_f$, and $\mathcal{R} = \{r_t\}_t$ one by one. The following describes each update step.

A. Update of $\mathcal{W}$

We first use orthogonal constraint [8] that the SOI $s_{f,t}$ has zero sample correlation with the noise signal $\boldsymbol{z}_{f,t}$, i.e., $\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{C}_{f,t} \boldsymbol{B}_{f,t} = \boldsymbol{0}_{1\times(M-1)}$. Under the distortionless constraint and the orthogonal constraint, we can estimate the t-th mixing vector $\boldsymbol{a}_{f,t}$:

$$\boldsymbol{a}_{f,t} = \frac{\boldsymbol{C}_{f,t}\boldsymbol{w}_f}{\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{C}_{f,t} \boldsymbol{w}_f}. \tag{14}$$

Next, we find an update rule of the separation filter $\boldsymbol{w}_f$. By putting the derivatives from (12), we obtain

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

where we used the same technique as the previous researches [8], [13] to replace the derivative from the second to fourth terms in (12) as $\boldsymbol{0}_M$. Using conventional technique [13], we take the linearized solution of $\boldsymbol{w}_f$ by fixing $\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{V}_{f,t} \boldsymbol{w}_f$ and $\hat{\sigma}_{f,t}^2$ as constant terms:

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

B. Update of $\mathcal{D}$

When updating the dereverberation filter $\boldsymbol{D}_f$, we need to treat the noise signal $\boldsymbol{z}_{f,t}$ that depends on $\boldsymbol{D}_f$. In previous research, we can consider a minimum solution of updating $\boldsymbol{D}_f$ using Kronecker product [17]–[19]. However, we need to decompose $\boldsymbol{W}_f$ to M separation filters, which results in a huge computation. In another approach, source-wise factorization [19] decomposes the dereverberation filter $\boldsymbol{D}_f$ into $\boldsymbol{D}_{f,\text{SOI}}$ and $\boldsymbol{D}_{f,\text{Noise}}$ and optimize each filter. However, it conflicts with the orthogonal constraint, and we cannot apply the same discussion in Section III-A. Also, we cannot use a technique [16, Algorithm 2] to update $\boldsymbol{D}_f$ because we use block-variant filter $\boldsymbol{B}_{f,t}$. We can consider each block-variant dereverberation filter $\boldsymbol{D}_{f,1}, ..., \boldsymbol{D}_{f,T}$. But we did not conduct the decomposition because the source extraction performance degraded in our preliminary experiments. In this paper, to realize a simple and effective update, we update the dereverberation filter $\boldsymbol{D}_f$ based on only the first term in (12).

By ignoring the third term in (12) and dropping the constant terms with respect to $\mathcal{D}$, we obtain

$$\mathcal{L}(\mathcal{D}) \overset{c}{=} \sum_{f=1}^{F} \left\| \left(\boldsymbol{D}_f - \boldsymbol{R}_f^{-1} \boldsymbol{P}_f \right) \boldsymbol{w}_f \right\|_{\boldsymbol{R}_f}^2, \tag{17}$$

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

$$\boldsymbol{R}_f = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[\psi(r_t) \frac{\overline{\boldsymbol{x}}_{f,t} \overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}}{\hat{\sigma}_{f,t}^2} \right], \tag{18}$$

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

Because $\boldsymbol{R}_f$ is positive definitive, we can minimize the cost function in (17) by solving:

$$\boldsymbol{D}_f \leftarrow \boldsymbol{R}_f^{-1} \boldsymbol{P}_f. \tag{20}$$

C. Update of $\mathcal{R}$

After updating $\boldsymbol{y}_{f,t}$ and $s_{f,t}$ using (5) and (6), we update the variance r_t by calculating the square root sum of the weighted signals:

$$r_t \leftarrow \sqrt{\sum_{f=1}^{F} \left| \frac{s_{f,t}}{\hat{\sigma}_{f,t}} \right|^2}. \tag{21}$$

This paper uses the coarse-fine source variance model proposed in the joint optimization of WPE and IVA [15]. While avoiding the frequency-domain permutation problem by using the frequency-invariant variance r_t for updating the separation filter $\boldsymbol{w}_f$, we replace the variance r_t in (18) and (19) as a frequency-variant one $r_{f,t}$:

$$r_t \leftarrow r_{f,t} = \left| \frac{s_{f,t}}{\hat{\sigma}_{f,t}} \right|. \tag{22}$$

Finally, we summarized our algorithm, CSV-WPEIVE in Algorithm 1.

IV. Experiment

We conducted an experiment to evaluate the source extraction performance of the proposed CSV-WPEIVE.

A. Experimental condition

We considered a situation where we had $M(=6)$ microphones, one target source, and $M-1(=5)$ point noises were mixed and observed in the microphones. We generated ten mixtures for the experiment. We obtained one point-source speech signal from the test set of the TIMIT corpus [21] and concatenated them so that the length of each signal becomes 20 seconds. We obtained point-source noise signals recorded in a cafe (CAF) from the third 'CHiME' Speech Separation and Recognition Challenge [22]. We obtained room impulse response (RIR) data using the image method [23]. We used the `signal generator`[1] to simulate RIRs. Figure 1 illustrates the mixing condition. We let SOI move at a uniform angular velocity of 9 [degree/s] in $67.5^\circ \leq \theta \leq 112.5^\circ$. We let other interfering noises be fixed. Before mixing SOI and noise, we adjusted the SNR to inputSNR $= 10 \log \frac{\lambda_1}{\lambda_2 + ..., + \lambda_M}$ as specified value, where λ_1 corresponds to the sample variance

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

Algorithm 1: Processing flow of CSV-WPEIVE

Input : observed signal $\boldsymbol{x}_{f,t}$ for $\forall f,t$
Output: source signals $s_{f,t}$ for $\forall f,t$

$\boldsymbol{w}_f \leftarrow \boldsymbol{e}_1$
$\boldsymbol{D}_f \leftarrow \boldsymbol{0}_{ML\times M}$
$\boldsymbol{y}_{f,t} \leftarrow \boldsymbol{x}_{f,t} - \boldsymbol{D}_f^{\mathsf{H}}\overline{\boldsymbol{x}}_{f,t}$ for $\forall f,t$
$s_{f,t} \leftarrow \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{y}_{f,t}$ for $\forall f,t$
$\boldsymbol{C}_{f,t} \leftarrow \mathbb{E}[\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathsf{H}}]$ for $\forall f,t$
$\hat{\sigma}_{f,t}^2 \leftarrow \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{w}_f$ for $\forall f,t$
for Iter $= 1$ to N_{Iter} **do**
  $\boldsymbol{a}_{f,t} \leftarrow (\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{w})^{-1}\boldsymbol{C}_{f,t}\boldsymbol{w}_f$ for $\forall f,t$
  $\boldsymbol{V}_{f,t} \leftarrow \mathbb{E}\left[\psi(r_t)\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathsf{H}}\right]$ for $\forall f,t$
  Compute $\boldsymbol{w}_f$ according to (16) for $\forall f$
  $\boldsymbol{w}_f \leftarrow \boldsymbol{w}_f/\sqrt{\sum_{t=1}^{T}\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{V}_{f,t}\boldsymbol{w}_f}$ for $\forall f$
  $\boldsymbol{R}_f \leftarrow \sum_{t=1}^{T}\mathbb{E}\left[\psi(r_{f,t})\overline{\boldsymbol{x}}_{f,t}\overline{\boldsymbol{x}}_{f,t}^{\mathsf{H}}/\hat{\sigma}_{f,t}^2\right]/T$ for $\forall f$
  $\boldsymbol{P}_f \leftarrow \sum_{t=1}^{T}\mathbb{E}\left[\psi(r_{f,t})\overline{\boldsymbol{x}}_{f,t}\boldsymbol{x}_{f,t}^{\mathsf{H}}/\hat{\sigma}_{f,t}^2\right]/T$ for $\forall f$
  $\boldsymbol{D}_f \leftarrow \boldsymbol{R}_f^{-1}\boldsymbol{P}_f$ for $\forall f$
  $\boldsymbol{y}_{f,t} \leftarrow \boldsymbol{x}_{f,t} - \boldsymbol{D}_f^{\mathsf{H}}\overline{\boldsymbol{x}}_{f,t}$ for $\forall f,t$
  $s_{f,t} \leftarrow \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{y}_{f,t}$ for $\forall f,t$
  $\boldsymbol{C}_{f,t} \leftarrow \mathbb{E}[\boldsymbol{y}_{f,t}\boldsymbol{y}_{f,t}^{\mathsf{H}}]$ for $\forall f,t$
  $\hat{\sigma}_{f,t}^2 \leftarrow \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{C}_{f,t}\boldsymbol{w}_f$ for $\forall f,t$
  $r_{f,t} \leftarrow |s_{f,t}/\hat{\sigma}_{f,t}|$ for $\forall f,t$
  $r_t \leftarrow \sqrt{\sum_f r_{f,t}^2}$ for $\forall t$
end
Projection Back [20] to solve the scale ambiguity of $s_{f,t}$.

of SOI and $\{\lambda_2 ..., \lambda_M\}$ correspond to that of noises. We set the sampling frequency to 16 kHz and the reverberation time $RT_{60} = 600$ ms.

We set the STFT frame length and shift as 512 and 256 samples, respectively. We set $N_{\mathrm{b}} = 100$ frames and the dereverberation filter length $L = 16$. For computational efficiency, we updated dereverberation filters $\mathcal{D}$ once every five updates in $\mathcal{W}$. In total, we updated $\mathcal{W}$ 100 times and $\mathcal{D}$ 20 times.

We used the average of the source-to-distortion ratios (SDR), the source-to-interference ratios (SIR), and the sources-to-artifact ratios (SAR) as the source separation accuracy [24]. We used the MUSEVAL V4 toolkit [25] with its bss_eval_images configuration and set the length of the bss_eval filter at 1 tap. We set the clean utterance sequence that convolved with the initial 32 ms part of the RIRs used for generating the corresponding mixtures as the reference. To evaluate the accuracy of both source extraction and dereverberation, we set the source images of the noise signals as interference reference of bss_eval. So SIR corresponds to how much we suppressed the images of the noise signal, and SAR corresponds to how much we suppressed the other components, e.g., reverberation.

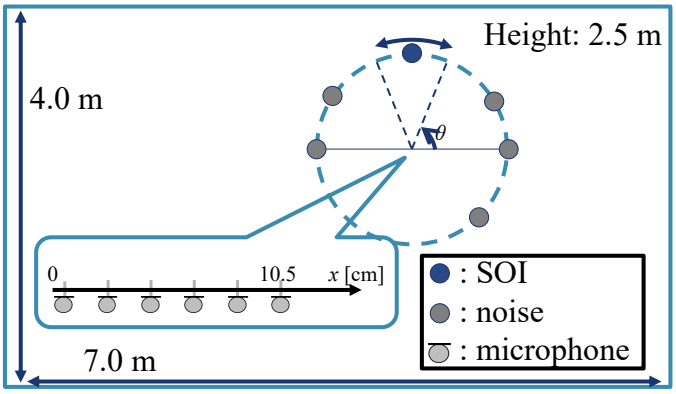

Fig. 1: Experimental sound source and microphone layout

TABLE I: SDR improvement (SDRi), SIR improvement (SIRi), and SAR [dB]. The bold font shows the top scores in each inputSNR.

method (inputSNR = 10 dB)	SDRi	SIRi	SAR
CSV-AuxIVE [13]	2.18	7.10	-1.16
WPE → CSV-AuxIVE	2.58	8.71	-2.48
CSV-WPEIVE (proposed)	3.51	9.84	-0.23
CSV-WPEIVE (w/ coarse fine) (proposed)	**3.76**	**10.44**	**0.42**
method (inputSNR = 5 dB)	SDRi	SIRi	SAR
CSV-AuxIVE [13]	2.83	6.29	-1.77
WPE → CSV-AuxIVE	3.27	7.64	-3.94
CSV-WPEIVE (proposed)	4.18	9.06	-1.68
CSV-WPEIVE (w/ coarse fine) (proposed)	**4.36**	**10.21**	**-0.95**

B. Result

We show the source extraction performance in Table I. In the cascade configuration of WPE and CSV-AuxIVE (WPE → CSV-AuxIVE), the processing becomes unstable when we set long dereverberation filter length L. So we show the result with shorter $L = 4$ only for the cascade configuration. As the results show, WPE → CSV-AuxIVE could improve the SDR by about 0.4 dB than the conventional method, CSV-AuxIVE. Furthermore, the proposed joint optimization, CSV-WPEIVE, further improved the SDR by about 1.0 dB. In addition, we achieved an SDR improvement of 4.36 dB using the coarse-fine source variance model when inputSNR = 5 dB. Also, we confirmed that the proposed CSV-WPEIVE realized more accurate noise suppression and dereverberation in terms of SIR and SAR. The results show that the proposed CSV-WPEIVE achieves more precisely accurate source extraction.

We next calculated the computational time of each method in Table II. Compared to CSV-AuxIVE, CSV-WPEIVE improved the SDR from 2.83 dB to 4.36 dB at $L = 16$. However, it needs about 64 seconds of calculation for 20-second inputs. When we set a relatively short dereverberation length (for example, $L = 4$ frames), the CSV-WPEIVE improved by 1.16 dB by adding a 9.47 seconds calculation compared to CSV-AuxIVE.

TABLE II: SDR improvement (SDRi) [dB] and computational time [second]. We set inputSNR = 5 dB.

method	SDRi	computational time
CSV-AuxIVE [13]	2.83	6.51
CSV-WPEIVE (coarse fine) $L = 4$	3.99	3.44
CSV-WPEIVE (coarse fine) $L = 16$	4.36	63.91

V. Conclusion

In this paper, we proposed an extension of CSV-AuxIVE, CSV-WPEIVE, capable of simultaneously solving source extraction and dereverberation problems. Although CSV-AuxIVE has been proposed as a fast and stable source extraction for a moving target source, it does not consider the effect of reverberation. Although joint optimization of WPE and IVE has been researched, the extension to the CSV mixing model has yet to be discovered. In this paper, we proposed a joint optimization algorithm that updates WPE, focusing only on the extracted target sound. The experimental results showed that the proposed joint optimization algorithm of WPE and CSV-AuxIVE improved the source extraction performance in a noisy and highly reverberant environment.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19H04131.

References

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

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

[3] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

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

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

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

[7] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. HSCMA*, 2014, pp. 107–111.

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

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

[10] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *IEEE Trans. SP*, vol. 69, pp. 3252–3267, 2021.

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

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

[13] J. Janskỳ, Z. Koldovskỳ, J. Málek, T. Kounovskỳ, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–16, 2022.

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

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

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

[17] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. ASLP*, vol. 19, no. 1, pp. 69–84, 2010.

[18] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. ICASSP*, 2018, pp. 31–35.

[19] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.

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

[21] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus LDC93S1. web download. philadelphia: Linguistic data consortium," 1993.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*. IEEE, 2015, pp. 504–511.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

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