Joint Optimization of Sampling Rate Offsets and Demixing Filters Using Auxiliary Function Method

Hayato Takeuchi*, Takao Kawamura*, Nobutaka Ono* and Shoko Araki†
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: takeuchi-hayato1@ed.tmu.ac.jp, kawamura-takao@ed.tmu.ac.jp, onono@tmu.ac.jp
† NTT, Inc. Japan
E-mail: shoko.araki@ntt.com

***Abstract*—In this paper, we propose a blind source separation (BSS) method for signals affected by sampling rate offset (SRO), which occasionally occurs in distributed recording scenarios. Conventional approaches typically follow a two-step procedure: estimating and compensating for the SRO, followed by applying BSS. In contrast, our proposed method defines a single objective function through which both the SROs and the demixing filters are jointly optimized, ensuring mutual consistency between synchronization and separation. The method is motivated by the idea that successful source separation implicitly requires proper synchronization, allowing us to guide synchronization using the separation objective itself. We believe that this unified formulation also provides a principled framework that can be extended to incorporate prior knowledge, regularization, or machine learning techniques. We evaluate the performance of the proposed method through experiments. The results demonstrate that our method achieves performance comparable to existing approaches, showing its promise for practical applications.**

I. Introduction

Blind source separation (BSS) aims to extract individual sound sources from a mixture of signals recorded by multiple microphones, without requiring prior information about the sources or the recording environment. BSS plays a crucial role in various applications such as hearing aids, speech recognition systems, and immersive audio systems. Numerous BSS methods have been developed, including independent component analysis (ICA) [1]–[3], independent vector analysis (IVA) [4], [5], and independent low-rank matrix analysis (ILRMA) [6].

A distributed microphone array consists of multiple independent recording devices such as smartphones, voice recorders, or laptops placed arbitrarily in space. This configuration allows for flexibility in deployment without the need for wired connections [7]. When equipped with communication capabilities, such arrays also enhance usability [8]. Distributed microphone arrays have been utilized in a wide range of tasks including sound source separation [9], [10], localization [11], voice activity detection [12], and acoustic scene classification [13], [14].

In distributed arrays, each device digitizes the acoustic signal independently, which results in synchronization mismatches such as sampling time offsets (STO) and sampling rate offsets (SRO). These asynchronies significantly degrade the performance of BSS systems [15], [16]. Since conventional BSS methods assume that the observed signals are time-synchronized, they are not directly applicable to such asynchronous settings. To handle this issue, a typical approach adopts a two-step procedure [17], [18]: synchronization techniques [19]–[23] are adapted to the recorded signals to estimate and compensate for the SRO, followed by BSS processing.

In this paper, we propose a unified approach that simultaneously estimates the sampling rate offsets and the demixing filters using the auxiliary function method [24]. The method is motivated by the idea that successful source separation inherently requires proper synchronization. We hypothesize that synchronization can therefore be guided solely by maximizing separation performance, enabling both components to be optimized jointly within a single objective function. This unified formulation not only ensures consistency between synchronization and separation, but also facilitates the incorporation of prior knowledge or regularization. We believe that it could further enable extensions such as end-to-end training with machine learning or online adaptation. We then derive update rules based on this objective function to efficiently optimize both the SROs and the demixing filters.

Experimental evaluations demonstrate that our method achieves performance comparable to conventional two-step methods.

II. Conventional Methods

To perform BSS from signals obtained by an asynchronous distributed microphone array, conventional methods have proposed a two-step approach [17], [18] where source separation is performed after compensating for SRO.

A. Blind Source Separation

In BSS, only the observed signals mixed from multiple sources are used to estimate the individual source signals before mixing. Let M and K denote the number of microphones and sources, respectively. The short-time Fourier transform (STFT) coefficients of the M-channel signal are denoted as $\boldsymbol{x}[t,f] = [x_0[t,f],\dots,x_{M-1}[t,f]]^{\mathsf{T}}$, and the time-invariant demixing filter is represented as $\boldsymbol{W}[f] = [\boldsymbol{w}_0[f],\dots,\boldsymbol{w}_{K-1}[f]]^{\mathsf{H}}$. t and f denote the time frame index and the frequency bin index, respectively. The estimated separated signal $\boldsymbol{y}[t,f] = [y_0[t,f],\dots,y_{K-1}[t,f]]^{\mathsf{T}}$ is expressed as:

$$\boldsymbol{y}[t,f] = \boldsymbol{W}[f]\boldsymbol{x}[t,f]. \tag{1}$$

979-8-3315-7206-8/25/$31.00

Representative methods for estimating demixing filter include auxiliary-function-based frequency domain ICA [3], and auxiliary-function-based independent vector analysis (AuxIVA) [5], ILRMA [6]. Here, it is assumed that the observed signals from each microphone are synchronized. It is well known that when the microphones are asynchronous, the performance degrades.

B. SRO Compensation Using the linear phase drift (LPD) Model

In an asynchronous distributed microphone array, signal synchronization is essential. Assuming that the 0-th microphone is used as the reference, the sampling rate r_m of the m-th microphone is expressed as follows:

$$r_m = (1 + \varepsilon_m) r_0, \tag{2}$$

where ε_m denotes the SRO of the m-th microphone, and it is assumed that $\varepsilon_0 = 0$. Let the continuous-time signal at each microphone be $\chi_m(t)$. The discrete-time signal $\chi_m[\tau]$ is then:

$$\chi_m[\tau] = \chi_m \left(\frac{\tau}{(1 + \varepsilon_m) r_0} + \Delta_m \right), \tag{3}$$

where τ is the discrete index of the sample, and Δ_m is the STO between the reference microphone and the m-th microphone. In this paper, we assume that the STO is compensated for in advance based on existing methods.

The STFT coefficient $x_m[t, f]$ of the signal $\chi_m[\tau]$ is given by:

$$x_m[t, f] = \sum_{l=-\lfloor L/2 \rfloor}^{\lfloor (L-1)/2 \rfloor} g[l] \chi_m[l + St] e^{-2\pi j f l / N}, \tag{4}$$

where j is the imaginary unit, $g[l]$ is the window function, L is the frame length, S is the shift size, and N is the number of discrete Fourier transform points. Here, F denotes the number of frequency bins up to the Nyquist frequency. In the LPD model, assuming that the SRO is sufficiently small, the impact of the SRO on the STFT coefficients can be compensated as follows:

$$\widehat{\boldsymbol{x}}[t, f] = [\widehat{x}_0[t, f], \ldots, \widehat{x}_{M-1}[t, f]]^{\mathsf{T}}, \tag{5}$$

$$\widehat{x}_m[t, f] = x_m[t, f] e^{j t f \kappa \varepsilon_m}, \tag{6}$$

where $\kappa = 2\pi S / N$. $\boldsymbol{\varepsilon} = [\varepsilon_0, \ldots, \varepsilon_{M-1}]$ denotes the SROs, and $\varepsilon_0 = 0$. Representative methods for estimating the SRO include those based on coherence drift (CD) [23], correlation maximization (CM) [20], maximum likelihood estimation (ML) [19], [25], double-cross-correlation processor (DXCP) [21].

III. Proposed method

In this study, we propose a method for jointly optimizing the SROs and the demixing filters using the auxiliary function approach. Here, we assume that the number of microphones and the number of sources are both equal, that is $M = K$.

We consider that successful source separation inherently requires proper synchronization. Based on this idea, we hypothesize that both the SROs and the demixing filters can be jointly optimized by using only the objective function of source separation, without explicitly defining a separate synchronization criterion. To realize this, we adopt the standard IVA framework, in which the SRO-compensated signal $\widehat{\boldsymbol{x}}[t, f]$ is used in place of the original observation.

Then, the objective function of IVA is given by [26]:

$$\begin{aligned} \mathcal{J}(\mathcal{W}, \boldsymbol{\varepsilon}) = & \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \sqrt{\sum_{f=0}^{F-1} |\boldsymbol{w}_k^{\mathsf{H}}[f] \widehat{\boldsymbol{x}}[t, f]|^2} \\ & - \sum_{f=0}^{F-1} 2 \log |\det \boldsymbol{W}[f]|, \end{aligned} \tag{7}$$

where $\mathcal{W} = \{\boldsymbol{W}[f]\}_{f=0}^{F-1}$. On the other hand, since both $\boldsymbol{w}_k[f]$ and $\boldsymbol{\varepsilon}$ appear inside the square root, this function does not admit a closed-form solution with respect to $\boldsymbol{w}_k[f]$ and $\boldsymbol{\varepsilon}$. To address this issue, we apply the auxiliary function method [24]. In this approach, an auxiliary function that upper-bounds the objective function is constructed, and optimization is performed by alternately updating the auxiliary variables and the target variables. The auxiliary function with respect to the demixing filter in Eq. (7) can be derived based on [5]:

$$\begin{aligned} \mathcal{J}^{+}(\mathcal{W}, \boldsymbol{\varepsilon} | \boldsymbol{v}) = & \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{K-1} \left(\frac{1}{2 v_k[t]} \sum_{f=0}^{F-1} \left| \boldsymbol{w}_k^{\mathsf{H}}[f] \widehat{\boldsymbol{x}}[t, f] \right|^2 + v_k[t] \right) \\ & - \sum_{f=0}^{F-1} 2 \log \left| \det \boldsymbol{W}[f] \right|, \end{aligned} \tag{8}$$

where, $v_k[t]$ is an auxiliary variable and $[\boldsymbol{v}]_{k,t} = v_k[t]$. The equality $\mathcal{J}^{+}(\mathcal{W}, \boldsymbol{\varepsilon} | \boldsymbol{v}) = \mathcal{J}(\mathcal{W}, \boldsymbol{\varepsilon})$ is satisfied when $v_k[t] = \sqrt{\sum_f |y_k[t, f]|^2}$ for all k and t. In this study, optimization is performed with respect to $\mathcal{W}$ and $\boldsymbol{\varepsilon}$.

A. Update rules of demixing filter

Eq. (8) has a closed-form solution with respect to $\boldsymbol{w}_k[f]$. By differentiating with respect to $\boldsymbol{w}_k^{\mathsf{H}}[f]$ and setting the result to zero, the following update rule can be derived.

$$v_k[t] \leftarrow \sqrt{\sum_{f=0}^{F-1} \left| \boldsymbol{w}_k^{\mathsf{H}}[f] \widehat{\boldsymbol{x}}[t, f] \right|^2}, \tag{9}$$

$$\boldsymbol{V}_k[f] = \frac{1}{T} \sum_{t=0}^{T-1} \frac{\widehat{\boldsymbol{x}}[t, f] \widehat{\boldsymbol{x}}^{\mathsf{H}}[t, f]}{2 v_k[t]}, \tag{10}$$

$$\boldsymbol{w}_k[f] \leftarrow (\boldsymbol{W}[f] \boldsymbol{V}_k[f])^{-1} \boldsymbol{e}_k, \tag{11}$$

$$\boldsymbol{w}_k[f] \leftarrow \frac{\boldsymbol{w}_k[f]}{\sqrt{\boldsymbol{w}_k^{\mathsf{H}}[f] \boldsymbol{V}_k[f] \boldsymbol{w}_k[f]}}. \tag{12}$$

Here, the auxiliary variable $v_k[t]$ is first updated using the demixing filter from the previous iteration. Then, the weighted

covariance matrix $\boldsymbol{V}_k[f]$ is computed based on the updated $v_k[t]$. Subsequently, the demixing vector $\boldsymbol{w}_k^{\mathsf{H}}[f]$ is updated using both $v_k[t]$ and $\boldsymbol{V}_k[f]$. $\boldsymbol{e}_k$ denotes a unit vector whose k-th element is 1. Note that the SRO $\boldsymbol{\varepsilon}$ remains fixed while updating the auxiliary variable, the weighted covariance matrix, and the demixing filter.

B. Update rules of SRO

In Eq. (8), since $\widehat{x}$ contains elements of $\boldsymbol{\varepsilon}$ in the exponent as Eq. (6), there is no closed-form solution with respect to the SRO $\boldsymbol{\varepsilon}$. Therefore, we optimize Eq. (8) by an auxiliary function method [25], [27]. First, Eq. (5) is rewritten as follows:

$$\widehat{\boldsymbol{x}}[t,f] = \mathrm{diag}(\boldsymbol{p}[t,f])\boldsymbol{x}[t,f], \tag{13}$$

$$\boldsymbol{p}[t,f] = (e^{jtf\kappa\varepsilon_0}, \ldots, e^{jtf\kappa\varepsilon_{M-1}})^{\mathsf{T}}. \tag{14}$$

By excluding terms that do not depend on the SRO, the objective function of Eq. (8) can be rewritten as follows.

$$\begin{aligned}\mathcal{Q}(\mathcal{W},\boldsymbol{\varepsilon}|\boldsymbol{v}) &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{K-1}\left(\frac{1}{2v_k[t]}\sum_{f=0}^{F-1}\left|\boldsymbol{w}_k^{\mathsf{H}}[f]\widehat{\boldsymbol{x}}[t,f]\right|^2\right)\\ &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\boldsymbol{p}[t,f]^{\mathsf{H}}\boldsymbol{\Upsilon}[t,f]\boldsymbol{p}[t,f],\end{aligned} \tag{15}$$

$$\begin{aligned}&\boldsymbol{\Upsilon}[t,f]\\ &= \mathrm{diag}(\boldsymbol{x}[t,f])^{\mathsf{H}}\left(\sum_{k=0}^{K-1}\frac{1}{2v_k[t]}\boldsymbol{w}_k[f]\boldsymbol{w}_k^{\mathsf{H}}[f]\right)\mathrm{diag}(\boldsymbol{x}[t,f]).\end{aligned} \tag{16}$$

Interestingly, although Eq. (16) is derived from the objective function of IVA for blind source separation, it has a similar mathematical form to Eq. (24) in [25], which was proposed for synchronization in distributed recording scenarios. Specifically, [25] assumes that the multichannel observed signals follow a time-invariant multivariate Gaussian distribution and formulates an objective function for jointly estimating all pairwise SROs, not just relative to a reference microphone.

Despite the fundamental difference in modeling assumptions, source separation based IVA with synchronization in this paper versus pure synchronization based on time-invariant Gaussian modeling in [25], the resulting cost functions for SRO optimization share the same structure, differing only in the definition of $\boldsymbol{\Upsilon}[t,f]$. In [25], $\boldsymbol{\Upsilon}[t,f]$ is defined as follows (same as Eq. (24)):

$$\boldsymbol{\Upsilon}[t,f] = \mathrm{diag}(\boldsymbol{x}[t,f])^{\mathsf{H}}\boldsymbol{V}[f]^{-1}\,\mathrm{diag}(\boldsymbol{x}[t,f]), \tag{17}$$

where $\boldsymbol{V}[f]$ is the spatial covariance matrix, which is computed by fixing the SROs using Eq. (5). Therefore, following [25], an auxiliary function can be designed and the update rule can be derived. The derived update rules are shown as:

$$\xi_{m,n}[t,f] = ft\kappa(\widetilde{\varepsilon}_n - \widetilde{\varepsilon}_m), \tag{18}$$

$$\lambda_{m,n}[t,f] = \frac{|\boldsymbol{\Upsilon}_{m,n}[t,f]|}{2}\,\mathrm{sinc}\left(\xi_{m,n}[t,f] - \mu_{m,n}[t,f]\right), \tag{19}$$

$$\begin{aligned}\mu_{m,n}[t,f] =& 2\pi\left\lfloor\frac{\xi_{m,n}[t,f] + \angle\boldsymbol{\Upsilon}_{m,n}[t,f]}{2\pi}\right\rfloor\\ &+ \pi - \angle\boldsymbol{\Upsilon}_{m,n}[t,f],\end{aligned} \tag{20}$$

$$(\boldsymbol{D}\boldsymbol{\varepsilon}_{\setminus 0})_{mM+n} = \begin{cases} 0 & m=0, n=0\\ \varepsilon_n & m=0, n\neq 0\\ -\varepsilon_m & m\neq 0, n=0\\ \varepsilon_n - \varepsilon_m & m\neq 0, n\neq 0\end{cases}, \tag{21}$$

$$\boldsymbol{A} = \sum_{t=0}^{T-1}\sum_{f=0}^{F-1}(ft\kappa)^2\boldsymbol{\Lambda}[t,f], \tag{22}$$

$$\boldsymbol{b} = \sum_{t=0}^{T-1}\sum_{f=0}^{F-1}ft\kappa\boldsymbol{\Lambda}[t,f]\boldsymbol{\mu}[t,f], \tag{23}$$

$$\boldsymbol{\varepsilon}_{\setminus 0} \leftarrow (\boldsymbol{D}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathsf{T}}\boldsymbol{b}. \tag{24}$$

Here, $\boldsymbol{\varepsilon}_{\setminus 0} = [\epsilon_1, \ldots, \epsilon_{M-1}]^{\mathsf{T}}$ and $\boldsymbol{D} \in \mathbb{R}^{M^2\times(M-1)}$ is a matrix used to compute the differences in $\boldsymbol{\varepsilon}_{\setminus 0}$. $\boldsymbol{\Lambda}[t,f]$ is a diagonal matrix whose $(mM+n, mM+n)$-th entry is $\lambda_{m,n}[t,f]$. $\boldsymbol{\mu}[t,f]$ is a vector whose $(mM+n)$-th entry is $\mu_{m,n}[t,f]$. Note that the auxiliary variable $v_k[t]$ and the demixing filter $\boldsymbol{W}[f]$ remain fixed while updating the SRO $\boldsymbol{\varepsilon}$.

C. Algorithm

The algorithm of the proposed method for the case of $K=2$ is shown in Algorithm 1, based on the update rules derived in Sections III-A and III-B. The index $i = 0, \ldots, I-1$ denotes the iteration number, where I is the total number of iterations. First, $\boldsymbol{W}[f]$ is updated given $\boldsymbol{\varepsilon}$. Next, $\boldsymbol{\varepsilon}$ is updated based on the updated $\boldsymbol{W}[f]$. Note that $v_k[t]$ is updated again before updating $\boldsymbol{\varepsilon}$. By repeating this process I times, joint optimization is achieved.

IV. Experiments

A. Experimental condition

In this experiment, we evaluated the effectiveness of the proposed method through simulations in two settings: (i) two sources and two microphones, and (ii) three sources and three microphones, using the simulation framework in [28]. The speech signals used in the experiment were five utterances selected from the Japanese Newspaper Article Sentences (JNAS) corpus provided by the Acoustical Society of Japan (ASJ) [29]. Ten types of mixtures were generated by combining these utterances. Each mixture signal had a total duration of 30 sec. The layouts of the simulated room for (i) and (ii) are illustrated in Fig. 1 and Fig. 2. The reverberation time was set to 200 ms. The observation signal at mic1 was used as the reference signal, and the sampling rate was 16000 Hz. To simulate SRO, the observation signals were resampled as follows: (i) the signal at mic2 was resampled to 16001.6 Hz, corresponding to an SRO of 100 ppm (parts per million), (ii) the signals at mic2 and mic3 were resampled to 16001.6 Hz and 16000.96 Hz, corresponding to SROs of 100 ppm and 60 ppm, respectively.

The STFT was computed using a 2048-point Hann window with a hop size of 1024 samples and a 4096-point DFT. The

Algorithm 1 Algorithm to perform BSS and estimate SRO

Input: Initial SRO ε and demixing filter $\boldsymbol{W}[f]$, D, κ

Output: Final estimate of SRO ε and demixing filter $\boldsymbol{W}[f]$

for $i = 0, \ldots, I-1$ **do**

$\widehat{\boldsymbol{x}}[f,t] = \mathrm{diag}(\boldsymbol{p}[t,f])\boldsymbol{x}[f,t]$

$\boldsymbol{y}[f,t] = \boldsymbol{W}[f]\widehat{\boldsymbol{x}}[f,t]$

$v_k[t] \leftarrow \sqrt{\sum_f |y_k[f,t]|^2}$

$\boldsymbol{V}_k[f] = (1/T)\sum_t \widehat{\boldsymbol{x}}[f,t]\widehat{\boldsymbol{x}}^{\mathsf{H}}[f,t]/2v_k[t]$

$\boldsymbol{w}_k[f] \leftarrow \mathrm{Update}(\boldsymbol{V}_k[f], \boldsymbol{W}[f])$ with Eqs. (11) and (12)

$\boldsymbol{y}[f,t] = \boldsymbol{W}[f]\widehat{\boldsymbol{x}}[f,t]$

$v_k[t] \leftarrow \sqrt{\sum_f |y_k[f,t]|^2}$

$\boldsymbol{\Upsilon}[t,f] = \mathrm{diag}(\boldsymbol{x}[t,f])^{\mathsf{H}} \left(\sum_k \frac{\boldsymbol{w}_k[f]\boldsymbol{w}_k[f]^{\mathsf{H}}}{2v_k[t]}\right) \mathrm{diag}(\boldsymbol{x}[t,f])$

$\widetilde{\boldsymbol{\varepsilon}} \leftarrow \boldsymbol{\varepsilon}$

$\xi_{m,n}[t,f] = ft\kappa(\widetilde{\varepsilon}_n - \widetilde{\varepsilon}_m)$

$\lambda_{m,n}[t,f] = \frac{|\boldsymbol{\Upsilon}_{m,n}[t,f]|}{2} \mathrm{sinc}\left(\xi_{m,n}[t,f] - \mu_{m,n}[t,f]\right)$

$\mu_{m,n}[t,f] = 2\pi \left\lfloor \frac{\xi_{m,n}[t,f] + \angle \boldsymbol{\Upsilon}_{m,n}[t,f]}{2\pi} \right\rfloor + \pi - \angle \boldsymbol{\Upsilon}_{m,n}[t,f]$

$\boldsymbol{A} = \sum_t \sum_f (ft\kappa)^2 \boldsymbol{\Lambda}[t,f]$

$\boldsymbol{b} = \sum_t \sum_f ft\kappa \boldsymbol{\Lambda}[t,f]\boldsymbol{\mu}[t,f]$

$\boldsymbol{\varepsilon}_{\setminus 0} \leftarrow (\boldsymbol{D}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathsf{T}}\boldsymbol{b}$

end for

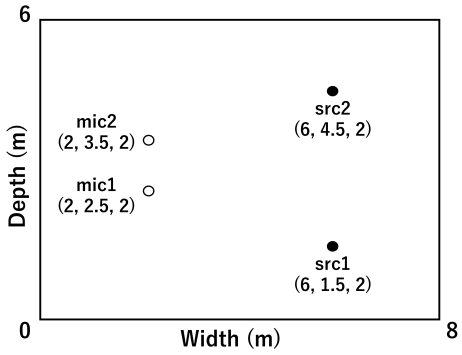

Fig. 1. Arrangement of sound sources and microphones in simulation experiments (i). The room size is $(8 \times 6 \times 4)$ m^3.

number of iterations was set to 200, which is sufficiently large to ensure convergence. Following the approach in [25], to avoid convergence to local optima during SRO updates, the frequency range used for the updates was restricted based on the iteration index. In this experiment, all frequency bins were used when updating the demixing filter, while the frequency range restriction was applied only when updating the SRO. As shown in Table I, the incremental bandwidth expansion was applied in the same way as in [25], where the number of iterations were divided into three equal parts and the same frequency ranges as in [25] were used in each stage. The initial value of the demixing filter $\boldsymbol{W}[f]$ was set to the identity matrix. We tested four initial values for the SRO (0, 75, 150, and 225 ppm) and selected the one that yielded the minimum objective function value after iteration.

To evaluate the effectiveness of the proposed method, we

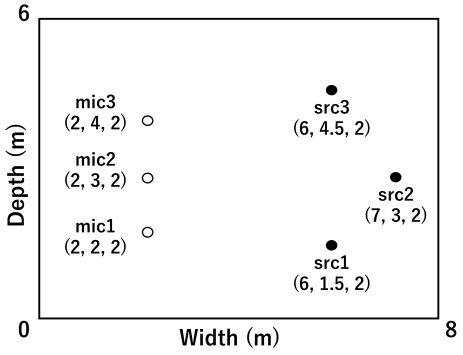

Fig. 2. Arrangement of sound sources and microphones in simulation experiments (ii). The room size is $(8 \times 6 \times 4)$ m^3.

TABLE I
NUMBER OF FREQUENCY BINS USED FOR UPDATING SRO PER ITERATION

Iteration index i	Number of frequency bins
$i \in [0, 65]$	512
$i \in [66, 132]$	1024
$i \in [133, 199]$	2049

compared it with the following cases: (1) BSS [5] performed after compensation using the ground-truth SRO, (2) BSS [5] performed after SRO compensation using a conventional SRO estimation method [25], (3) BSS [5] applied to asynchronous signals without any SRO compensation, and (4) the case where no source separation is performed. Note that in comparison method (2), both the SRO estimation and the demixing filter estimation were carried out for 200 iterations each. For the evaluation of the separation performance, we used the scale-invariant signal-to-distortion ratio (SI-SDR) [30] between the original and separated signals for each source (src1 and src2). Here, the scale of the separated signal was restored by projection back onto mic1. According to the evaluation of SRO estimation performance, we used the root mean square error (RMSE) between the ground-truth SRO and estimated SRO.

B. Results

The average SI-SDRs over 10 simulation trials for (i) and (ii) are shown in Table II and Table III. "Unprocessed" corresponds to the case where no source separation is performed, "BSS w/o sync." refers to BSS applied without any synchronization, "BSS w/o sync. of two-step" indicates the conventional two-step approach where synchronization is performed using a conventional SRO estimation method followed by BSS, "Prop." denotes the proposed method that jointly optimizes the SRO and the demixing filter, and "BSS w/ sync.*" represents the case where BSS is performed after synchronization using the ground-truth SRO (100 ppm in case (i), 100 ppm and 60 ppm in case (ii)).

From the experimental results, in case (i), "BSS w/o sync." improves the SI-SDR by more than 4 dB compared to "Unprocessed", but its performance is still inferior to that of "BSS w/ sync.*". On the other hand, in both case (i) and case (ii),

TABLE II
AVERAGE SI-SDR [DB] OF EXPERIMENT (I)

	src1	src2
Unprocessed	-20.1833	-6.7207
BSS w/o sync.	1.7749	-1.8627
BSS w/ sync. (two-step)	13.9694	11.7246
Prop.	14.3397	12.0866
BSS w/ sync.*	14.3078	12.0541

TABLE III
AVERAGE SI-SDR [DB] OF EXPERIMENT (II)

	src1	src2	src3
Unprocessed	-19.4616	-22.6113	-6.4915
BSS w/o sync.	-4.7033	-7.3012	-7.3858
BSS w/ sync. (two-step)	13.5481	13.1153	12.1462
Prop.	13.5483	13.1161	12.1460
BSS w/ sync.*	13.5335	13.0889	12.1300

"Prop." improved the SI-SDR by at least 12 dB compared to "BSS w/o sync.". Furthermore, it achieves comparable performance to "BSS w/ sync.*" and "BSS w/ sync. (two-step)". These results indicate that the single-step approach achieved performance comparable to the two-step approach. We also observed that the optimization behavior depended on the initial SRO values. Initializations with differences greater than 50 ppm from the true SRO (i.e., 0, 150, and 225 ppm) occasionally resulted in convergence to local optima. However, by selecting the result that achieved the minimum value of the objective function, both the proposed method and the two-step method were able to obtain the correct estimates. Further investigation under a broader range of conditions would be beneficial in future work.

The average RMSEs of SRO for (i) and (ii) are summarized in Table IV and Table V, respectively. The methods referred to as "Prop." and "BSS w/ sync. (two-step)" are the same as those shown in Table II. In both cases, the difference in RMSE values between "Prop." and "BSS w/ sync. (two-step)" was less than 0.1 ppm. When considering the SI-SDR results in Table II, these differences suggest that this level of accuracy is sufficient for effective source separation.

V. CONCLUSIONS

In this paper, we proposed a blind source separation method for asynchronous signals based on the joint optimization of SROs and demixing filters using the auxiliary function approach. Unlike conventional two-step approaches, the proposed method defines a single objective function through which both the synchronization and separation processes are jointly optimized. This unified formulation ensures mutual consistency between SRO compensation and source separation, and provides a flexible framework for incorporating prior knowledge, regularization, or machine learning-based modules. Experimental evaluations using simulated data confirmed that the proposed method achieves separation performance comparable to conventional two-step methods. As future work, we plan to apply the proposed method to real recording scenarios to further validate its effectiveness and practical applicability.

TABLE IV
AVERAGE RMSE [PPM] OF EXPERIMENT (I)

	mic2
BSS w/ sync. (two-step)	0.4666
Prop.	0.4677

TABLE V
AVERAGE RMSE [PPM] OF EXPERIMENT (II)

	mic2	mic3
BSS w/ sync. (two-step)	0.3671	0.2232
Prop.	0.3561	0.2179

ACKNOWLEDGMENT

This work was supported by JST SICORP (JPMJSC2306).

REFERENCES

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

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998, ISSN: 0925-2312.

[3] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 165–172.

[4] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Independent Component Analysis and Blind Signal Separation (ICA)*, Springer, 2006, pp. 165–172.

[5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[7] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 161–164.

[8] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011.

[9] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5694–5698.

[10] D. Wang, T. Yoshioka, Z. Chen, X. Wang, T. Zhou, and Z. Meng, "Continuous speech separation with ad hoc microphone arrays," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1100–1104.

[11] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24,

[12] P. Giannoulis, A. Brutti, M. Matassoni, *et al.*, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1271–1275.

[13] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, 2017.

[14] T. Kawamura, Y. Kinoshita, N. Ono, and R. Scheibler, "Effectiveness of inter- and intra-subarray spatial features for acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] E. Robledo-Arnuncio, T. S. Wada, and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 34–37.

[16] Z. Liu, "Sound source seperation with distributed microphone arrays in the presence of clocks synchronization errors," in *Proc. International Workshop for Acoustic Echo and Noise Control (IWAENC)*, 2008.

[17] K. Ochi, N. Ono, S. Miyabe, and S. Makino, "Multi-talker speech recognition based on blind source separation with ad hoc microphone array using smartphones and cloud storage," in *Proc. Interspeech*, 2016, pp. 3369–3373.

[18] H. Nammoku, K. Yamaoka, T. Nakashima, Y. Wakabayashi, and N. Ono, "Analysis and source separation of overlapping speech using corpus of everyday Japanese conversation," in *Proc. International Congress on Acoustics (ICA)*, 2022.

[19] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015.

[20] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[21] A. Chinaev, P. Thüne, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021.

[22] J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, and R. Haeb-Umbach, "Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming," in *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2017.

[23] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[24] D. R. Hunter and K. L. and, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[25] Y. Masuyama, K. Yamaoka, T. Kawamura, and N. Ono, "Efficient joint optimization of sampling rate offsets using entire multichannel signal," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1816–1828, 2024.

[26] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, e12, 2019.

[27] K. Yamaoka, R. Scheibler, N. Ono, and Y. Wakabayashi, "Sub-sample time delay estimation via auxiliary-function-based iterative updates," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 130–134.

[28] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[29] K. Itou, M. Yamamoto, K. Takeda, *et al.*, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[30] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" In *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.