Signal Processing 107 (2015) 185–196

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation

Shigeki Miyabe [a,*], Nobutaka Ono [b], Shoji Makino [a,1]

[a] Life Science Center of Tsukuba Advanced Research Alliance, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
[b] Principles of Informatics Research Division, National Institute of Informatics/Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form
10 September 2014
Accepted 11 September 2014
Available online 22 September 2014

Keywords:
Ad hoc microphone array
Drift
Sampling frequency
Maximum likelihood estimation
Blind source separation

ABSTRACT

In this paper, we propose a novel method for the blind compensation of drift for the asynchronous recording of an ad hoc microphone array. Digital signals simultaneously observed by different recording devices have drift of the time differences between the observation channels because of the sampling frequency mismatch among the devices. On the basis of a model in which the time difference is constant within each short time frame but varies in proportion to the central time of the frame, the effect of the sampling frequency mismatch can be compensated in the short-time Fourier transform (STFT) domain by a linear phase shift. By assuming that the sources are motionless and have stationary amplitudes, the observation is regarded as being stationary when drift does not occur. Thus, we formulate a likelihood to evaluate the stationarity in the STFT domain to evaluate the compensation of drift. The maximum likelihood estimation is obtained effectively by a golden section search. Using the estimated parameters, we compensate the drift by STFT analysis with a noninteger frame shift. The effectiveness of the proposed blind drift compensation method is evaluated in an experiment in which artificial drift is generated.

1. Introduction

Microphone array signal processing is a framework for analyzing spatial information of a sound field observed with multiple microphones to perform speech enhancement, source separation, source localization, and so forth, which are difficult by the processing of single-channel observations [1]. Microphone arrays are used in various applications, including teleconferencing, hands-free speech recognition, hearing aids, acoustic monitoring, spatial audio, and computer games. While conventional microphone array signal processing assumes that multichannel signals are observed by a unified analog-to-digital converter (ADC), recently increasing attention has been focused on an extension of the microphone array framework, the so-called ad hoc microphone array, where a combination of observations by independent recording devices is treated as multichannel recording [2]. The nonnecessity of wired channels to achieve synchronization enables the downsizing of the recording devices, which is an important attribute for hearing aids [3]. In addition, this framework is suitable for recording meetings because of the easy construction of the microphone array by the combination of widely available portable recording devices, such as cell phones, IC recorders, and video cameras, and the freedom of the microphone arrangement, which enables recording with a

* Corresponding author. Tel.: +81 29 853 6566; fax: +81 29 853 7387.
E-mail address: miyabe@tara.tsukuba.ac.jp (S. Miyabe).
[1] The following co-author is a member of EURASIP: Shoji Makino.

http://dx.doi.org/10.1016/j.sigpro.2014.09.015

high SNR by setting each device close to each speaker [2,4]. Also, considerable effort has been made to develop wireless acoustic sensor networks (WASNs), where the recording devices are connected by wireless networks [5].

However, the increased freedom of ad hoc microphone arrays raises various issues that do not arise in conventional array signal processing. For example, the array geometry is unknown [2,6–8], the recording devices have different unknown gains [2], each device starts recording independently [7,8], and the sampling frequencies are not common among the observation channels [9–15]. Also, in WASNs, the efficiencies of communication and distributed computation are important issues to achieve array signal processing with a limited bandwidth and array nodes with low computational power [5].

Among these issues of ad hoc microphone arrays, one of the most important is the mismatch of sampling frequencies. Since each ADC is not synchronized with the others, the individual variability of clocks results in a slight mismatch of the sampling frequencies, causing a change in the time difference between channels due to the constant skew, the so-called *drift*. Array signal processing generally assumes that the sources do not move and utilizes the phase differences inherent to the positions of the sources. However, drift causes the phase differences to constantly change as if the sources are moving, preventing the use of array signal processing to analyze phase differences assuming static sources [10,11]. Also, the asynchronous recording causes offsets of the recording start time. Estimation and compensation of the sampling frequency mismatch and recording start offset are indispensable as preprocesses in array signal processing.

While there are two types of methods for estimating the sampling frequency mismatch, i.e., supervised and unsupervised, our research focuses on unsupervised estimation. The former, supervised estimation, mainly estimates the relation between the clock time of the ADC and the absolute time by using the time stamp received from satellites or wireless networks [9,13–15]. While the advantage of this approach is that the recording start offset can be estimated in addition to the sampling frequency mismatch, the disadvantage is the constraint that the recording devices must have the ability to receive the time stamp. Another problem is that the accuracy of the time stamp is generally much lower than that required in array signal processing, and training involving the long observation of time stamps is required for accurate estimation. Unsupervised estimation cannot obtain a precise estimate of the recording start offset without prior information. Inaccurate compensation of the recording start offset is problematic with particular classes of supervised array signal processing which receive time differences of arrival (TDOAs) of positions of specific sources as prior information because the offset changes the relation between the TDOAs and the positions. However, in a blind scenario of array signal processing such as blind source separation (BSS) [16], where only the observation is given, a rough compensation of the recording start offset is sufficient. Therefore, in this paper we focus on the accurate compensation of the drift and the rough compensation of the recording start offset in an unsupervised manner.

To the best of our knowledge, there have been few works on the blind estimation of sampling frequency mismatch. Liu et al. proposed a method of estimation involving the iteration of independent component analysis (ICA) and evaluation of the correlation between the estimated independent components by utilizing the property that ICA can extract uncorrelated independent sources only when the drift is well compensated [10]. However, its applicability is limited to determined systems with equal numbers of sources and microphones so that the independent components can be extracted. Markovich-Golan et al. estimated the sampling frequency mismatch as the rate of change of the phase in the interchannel correlation of a noise observation [12]. The modeling of the sampling frequency mismatch using the phase is very similar to our proposed framework, which is discussed later. However, the scenario of this method, which assumes voice activity detection (VAD), is different from our fully blind estimation scenario.

In this paper, we propose a novel method for blind drift compensation by maximum likelihood estimation of the sampling frequency mismatch in the short-time Fourier transform (STFT) domain. The basic idea of the maximum likelihood estimation is published in a previous conference proceedings paper [17]. The optimization algorithm is followed by resampling as a modification of the STFT analysis with a noninteger frame shift, which we proposed in another conference proceedings paper [18]. In addition, we newly propose an iterative algorithm for estimating the recording start offset and sampling frequency mismatch, which enables the accurate compensation of drift even with a long observation. We model the drift in the STFT domain as a linear phase by ignoring the drift inside each short time frame. By assuming that the sources are unmoving and have stationary amplitudes, the sound wave to be observed is regarded as stationary regardless of the number of sources. Since the stationarity collapses with the pseudo-movement of the sources caused by drift, the stationarity can be a cue to estimate the drift. Thus, we derive a likelihood function in the STFT domain to measure the stationarity and evaluate the compensation of the sampling frequency mismatch. The maximum likelihood estimate is searched for efficiently by performing a golden section search. We also show that the likelihood function evaluates the coherence between the channels. To compensate the recording start offset, we shift the observed signal in the time domain to maximize the interchannel correlation of the signals with the sampling frequency mismatch compensated. Since the accuracies of the estimation of the offset and the compensations of the drift are mutually dependent, particularly when the observation is long, we iterate these procedures. We evaluate the effectiveness of the proposed method of blind drift compensation in an experiment to emulate the asynchronous recording of an ad hoc microphone array by giving an artificial sampling frequency mismatch.

The rest of the paper is organized as follows. In Section 2, we formulate the asynchronous observation of the ad hoc microphone array. In Section 3, we describe our modeling of the drift in the STFT domain. In Section 4, we derive the likelihood function used to estimate the

sampling frequency mismatch, carry out an efficient search for the maximum likelihood estimate, and analyze the properties of the likelihood function. In Section 5, we describe STFT analysis with noninteger frame shift for computationally simple resampling. In Section 6, we describe the whole algorithm used in the proposed method of blind drift compensation. In Section 7, we evaluate the effectiveness of the proposed method. Finally, the paper is concluded in Section 8.

2. Effect of sampling frequency mismatch on discrete signals

We suppose that sound pressures $x_1(t)$ and $x_2(t)$ on two microphones are sampled by different ADCs, where t denotes the continuous time. The sound pressures are observed as the discretized signals denoted as $x_1[n]$ for $n=0,\dots,N_1-1$ and $x_2[n]$ for $n=0,\dots,N_2-1$, where n denote the discrete time, and N_i for $i=1,2$ is the number of observed samples of the ith channel, and N_1 and N_2 are not necessarily the same. Throughout this paper, the notations $(\cdot)$ and $[\cdot]$ are used for denoting continuous and discrete time signals, respectively. Although the signals are available as observation only at the integer-valued discrete times, we accept the noninteger values of the discrete times. We also suppose that the sampling frequency of $x_1[n]$ is f_s and that of $x_2[n]$ is $(1+\epsilon)f_s$ for a dimensionless number ϵ used to define the sampling frequency mismatch, without loss of generality. In this paper, we assume that the ADCs have common nominal sampling frequencies and that $|\epsilon| \ll 1$. Note that we focus on the compensation of drift between two channels in this paper, but the extension to an arbitrary number of channels can be carried out easily. The relations between $x_i[n]$ and $x_i(t)$ for $i=1,2$ are given by

$$x_1[n] = x_1\left(\frac{n}{f_s}\right), \tag{1}$$

$$x_2[n] = x_2\left(\frac{n}{(1+\epsilon)f_s} + T_{21}\right), \tag{2}$$

where the origin of the continuous time $t=0$ is defined as the time when the sampling of $x_1[n]$ starts, and T_{21} is the continuous time when the sampling of $x_2[n]$ starts. The discrete times of these two channels have independent correspondence to the continuous time as described in (1) and (2), and the phase difference between the channels linearly changes according to the time, hereafter we refer this behavior as drift. Fig. 1 shows a conceptual diagram of the asynchronous recording in the case that $x_1(t)=x_2(t)$. We denote the pair of discrete times corresponding to the identical continuous time t as n_1 for the first channel and n_2 for the second channel. Then the discrete time pair n_1 and n_2 satisfies the following condition:

$$n_2 = (1+\epsilon)(n_1 - D_{21}), \tag{3}$$

$$D_{21} = f_s T_{21}, \tag{4}$$

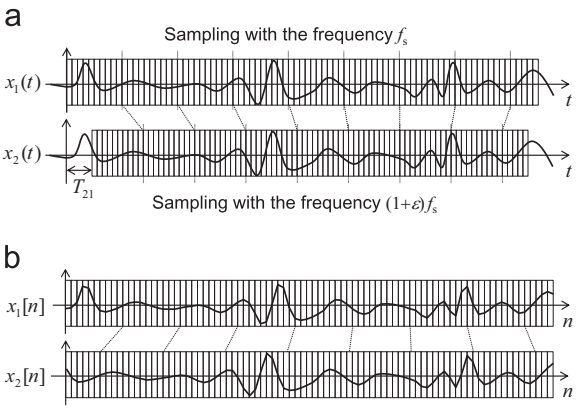

Fig. 1. Conceptual diagram of the asynchronous recording in the case when the analog waveforms of the two channels are the same, i.e., $x_1(t)=x_2(t)$. The sampling frequency mismatch with $\epsilon>0$ expands the digital waveform of $x_2[n]$. (a) Analogue waveform and sampling and (b) Resultant digital waveform with drift.

where D_{21} denotes the discrete time of the first channel when the recording of the second channel starts. Note that hereafter we use the notation of the pair n_1 and n_2 when we consider the correspondence to the identical continuous time.

According to the above modeling of the sampling of the asynchronous recording, the discretized signal $\hat{x}_2[n]$ of the second channel with precise synchronization to the first channel is given as

$$\begin{aligned}\hat{x}_2[n] &= x_2\left(\frac{n}{f_s}\right) \\ &= x_2[(1+\epsilon)(n-D_{21})].\end{aligned} \tag{5}$$

Thus, to achieve the synchronization, we have to obtain the amplitude of the second channel at the noninteger-valued discrete times $(1+\epsilon)(n-D_{21})$, and the problem reduces to resampling. The precise resampling is given by the following infinite convolution of the sinc function:

$$\hat{x}_2[n] = \sum_{n'=-\infty}^{\infty} \mathrm{sinc}((1+\epsilon)(n-f_s T_{21}) - n')x_2[n']. \tag{6}$$

Since an infinitely long observation is unavailable in practice, the infinite convolution has to be approximated, for example, by a truncated convolution. Additionally, a simplified model is necessary to formulate an effective estimate of the unknown mismatch parameters.

3. Modeling of sampling frequency mismatch in STFT domain

3.1. Statement of problem

Here we discuss the modeling and compensation of the sampling frequency mismatch, which is the basis of the proposed method. Since parameter estimation in array signal processing is generally formulated as statistical optimization in the STFT domain, it is sufficient to compensate for the effect of drift appearing in the STFT domain. Thus, the analysis in the STFT domain is appropriate for the estimation and compensation of drift as a preprocessing step in array signal processing.

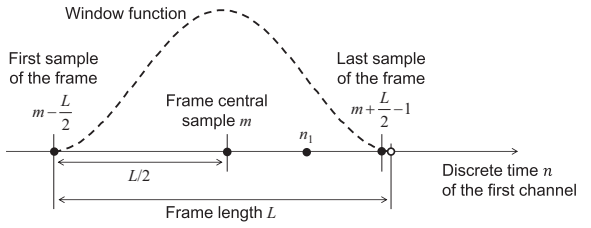

Fig. 2. Discrete time n_1 of the first channel in the frame centered at sample m.

We define the STFT signal $X_i(k,m)$, $i=1,2$, $k=-L/2,\ldots,L/2-1$ obtained by the frame analysis of $x_i[n]$ centered at the mth sample with a window of length L samples as

$$X_i(k,m)=\sum_{l=0}^{L-1} w(l)x_i\left[l+m-\frac{L}{2}\right]\exp\left(-\frac{2\pi jkl}{L}\right), \tag{7}$$

where $w(l)$ is an appropriate window function, k is the discrete frequency index, and $j=\sqrt{-1}$. We assume that L is even. Note that the discrete Fourier transform is substituted by the fast Fourier transform in practical processing.

We discuss the correspondence between n_1 and n_2 inside the frame centered at sample m in the first channel, i.e., $m-L/2\le n_1\le m+L/2-1$ as shown in Fig. 2. From (3), the following correspondence is obtained:

$$(n_2-m)-(n_1-m)=\epsilon m+\epsilon(n_1-m)-(1+\epsilon)D_{21}. \tag{8}$$

Considering that (n_1-m) is the time inside the frame, the terms on the right side of (8) have the following properties:

- The first term ϵm increases or decreases in proportion to the frame central time m.
- The second term $\epsilon(n_1-m)$ is proportional to the time inside the frame.
- The third term $(1+\epsilon)D_{21}$ originates from the difference in the recording start times.

3.2. Constant-drift model within frame

Let us start from the second term on the right of (8). Suppose that both ϵ and L are small and their product is much smaller than one sampling interval,

$$|\epsilon L|\ll 1. \tag{9}$$

Then we obtain the following approximation because $-L/2\le n_1-m\le L/2-1$:

$$(n_2-m)-(n_1-m)\approx\epsilon m-(1+\epsilon)D_{21}, \tag{10}$$

which indicates that the drift inside each frame is sufficiently small to be ignored. In addition, if the frame of the second channel is centered at

$$\phi_{21}(m)=(1+\epsilon)(m-D_{21}), \tag{11}$$

then the frame is synchronous with the frame of the first channel centered at the mth sample because of the correspondence between $n_1=m$ and $n_2=\phi_{21}(m)$, indicating an identical continuous time.

3.3. Compensation model of sampling frequency mismatch in STFT domain

Next, we discuss the third term $-(1+\epsilon)D_{21}$ on the right of (8), which is caused by the difference in the recording start times. Although accurate estimation of the difference in the recording start times is difficult, a small error is acceptable for the class of array signal processing without the TDOA prior such as BSS, as described in Section 1. Thus, the satisfactory compensation of the recording start offset with the error much smaller than the frame size is obtained by the time shift with the maximum correlation. We now redefine the discrete time of the second channel so that the recording start offset is compensated as $D_{21}\approx 0$, with $n_2=0$ corresponding to $n_1\approx 0$. Then (10) is approximated as

$$(n_2-m)-(n_1-m)\approx\epsilon m. \tag{12}$$

If, for the central samples m of all the frames, the condition

$$|\epsilon m|\ll L \tag{13}$$

is satisfied and the shift of the frame is much smaller than the frame size L, the time difference ϵm in (12) can be regarded as a small shift of the whole signal inside each frame. Such a small shift can be compensated by a linear phase in the STFT domain as

$$\hat{X}_2(k,m;\epsilon)=X_2(k,m)\exp\left(\frac{2\pi jk\epsilon m}{L}\right) \tag{14}$$

to obtain the STFT signal $\hat{X}_2(k,m;\epsilon)$ with the sampling frequency mismatch compensated.

3.4. Trade-off in frame size

To obtain the compensation model of (12), we introduced two assumptions in (9) and (13) related to the frame size L. These assumptions contradict each other when the sampling frequency mismatch ϵ is large and the signal lengths N_1 and N_2 are large, as summarized in the following:

1. If L is large, (9) cannot be satisfied and the drift inside each frame becomes too large to ignore.
2. If L is small, (13) cannot be satisfied, and the drift cannot be compensated by a linear phase in the STFT domain as in (14), particularly when the observation is long and $|m|$ takes a large value.

Thus, the frame size has to be chosen carefully. We show a conceptual diagram of the trade-off in Fig. 3.

The drift inside each frame is unignorable when the sampling frequency mismatch ϵ and the frame length L are large and the condition in (9) cannot be satisfied. As can be seen in Fig. 3, the largest effect of the drift inside the frame appears in both ends of the frame with the error of the discrete time of $|\epsilon L/2|$ samples. Such that the time error of $|\epsilon L/2|$ samples corresponds to the error of the phase $|\pi jk\epsilon|$

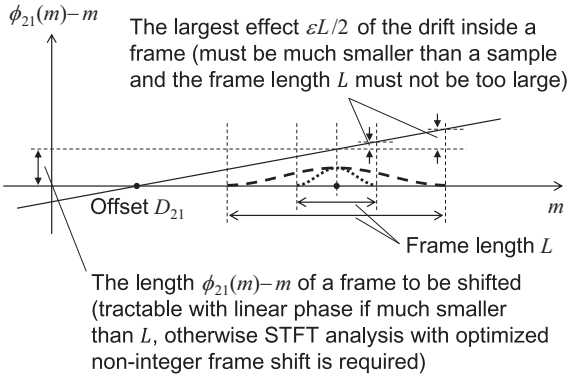

Fig. 3. Trade-off in the frame length L.

at the discrete frequency k, and the effect of the drift inside the frame is large at high frequencies. The typical quantity of the error is in the order of 10 μs with the sampling frequency normally in the order of 10^{-5} and the frame analysis with the length of the order of 0.1 s. Note that the effect of the errors is reduced with a typical choice of the window function $w(l)$ to suppress the amplitude near both ends.

In contrast to the condition so that we can ignore the drift inside frames, the condition in (13) to approximate the shift of the signal with the linear phase is that the frame length L is much larger than the signal shift $|\phi_{21}(m)-m| \approx |\epsilon m|$. It is hard to satisfy this condition when the observation is long and the sampling frequency mismatch ϵ is large. We show a typical example. Supposing we observe signals of the length 10 s with a nominal sampling frequency of 16 kHz, and $D_{21} \approx 0$. With a sampling frequency mismatch of $\epsilon = 5 \times 10^{-5}$ and a window length of $L=2048$, the maximum shift of the signal in the last frame is about 8 samples, which is sufficiently small to be approximated by a linear phase. However, for a longer observation, the time shift becomes closer to the frame length L, and exceeds the frame length in the case of observation for longer than 2560 s.

To solve the trade-off, we employ the optimization of the STFT analysis with the noninteger frame shift as described in Section 5. With the optimized STFT analysis, we do not need to consider the condition in (13), and the frame length L should be decided considering the condition in (9) and the suitable frame length for the array signal processing.

4. Maximum likelihood estimation of sampling frequency mismatch assuming spatial stationarity

4.1. Probabilistic model in STFT domain

The drift caused by the sampling frequency mismatch causes the TDOAs of each sound source to change slowly with time as if the source is moving. Thus, if the movements of sources are not large, the compensation of the sampling frequency mismatch can be evaluated by considering how static the TDOAs are. Also, by assuming that the sources are stationary, spatial stationarity can be used as a measure of the sampling frequency mismatch. With these assumptions, we derive a likelihood for the sampling frequency mismatch using the compensation model given by (14). Note that in this section we assume that the condition in (13) is satisfied, and the model mismatch is fixed in the algorithm described in Section 6.

We assume that all the sources are stationary and that their amplitudes are stationary on a long-term basis. Then the compensated observed signal $\hat{\boldsymbol{X}}(k,m;\epsilon)$ expressed in vector notation, given by

$$\hat{\boldsymbol{X}}(k,m;\epsilon) = [X_1(k,m), \hat{X}_2(k,m;\epsilon)]^{\mathsf{T}}, \tag{15}$$

where $\{\cdot\}^{\mathsf{T}}$ denotes matrix transposition, is regarded as a stationary random variable if the sampling frequency mismatch ϵ is estimated accurately. We also assume that the compensated observation $\hat{\boldsymbol{X}}(k,m;\epsilon)$ with the stationarity recovered by the precise estimation of ϵ has a zero-mean bivariate complex normal distribution for each frequency bin k, whose density of each frequency bin is given by

$$p\left(\hat{\boldsymbol{X}}(k,m;\epsilon);\mathbf{V}(k)\right) = \frac{\exp(-\hat{\boldsymbol{X}}(k,m;\epsilon)^{\mathsf{H}}\mathbf{V}(k)^{-1}\hat{\boldsymbol{X}}(k,m;\epsilon))}{\pi^2 \det \mathbf{V}(k)}, \tag{16}$$

where $\mathbf{V}(k)$ denotes the covariance matrix. Thus, accurate estimation of ϵ recovers the stationarity of $\hat{\boldsymbol{X}}(k,m;\epsilon)$ and maximizes the following log likelihood function $J(\mathbf{V},\epsilon)$, which evaluates the fit with the zero-mean bivariate normal distribution:

$$\begin{aligned} J(\mathbf{V},\epsilon) &= \sum_{k,m} \log p(\hat{\boldsymbol{X}}(k,m;\epsilon);\mathbf{V}(k)) \\ &= \sum_{k,m} (-\log \pi^2 - \log\det \mathbf{V}(k) \\ &\quad -\hat{\boldsymbol{X}}(k,m;\epsilon)^{\mathsf{H}}\mathbf{V}(k)^{-1}\hat{\boldsymbol{X}}(k,m;\epsilon)), \end{aligned} \tag{17}$$

where $\{\cdot\}^{\mathsf{H}}$ denotes conjugate transposition. $\mathbf{V}$ denotes the group of all covariance matrices $\mathbf{V}(k)$ for $k=-L/2,\ldots,L/2-1$, which can be substituted by the following sample estimate:

$$\mathbf{V}(k) \leftarrow \frac{1}{|\forall m|}\sum_m \hat{\boldsymbol{X}}(k,m;\epsilon)\hat{\boldsymbol{X}}(k,m;\epsilon)^{\mathsf{H}}, \tag{18}$$

where $|\forall m|$ denotes the number of frames. By substituting (18) and omitting the constants, a simplified version $J(\epsilon)$ of the log likelihood function $J(\mathbf{V},\epsilon)$ is given by

$$J(\epsilon) = -\sum_k \log\det \sum_m \hat{\boldsymbol{X}}(k,m;\epsilon)\hat{\boldsymbol{X}}(k,m;\epsilon)^{\mathsf{H}}. \tag{19}$$

Note that the sum of the quadratic forms in the last term of (17) is constant with the sample estimate of the covariance matrix given by (18), as

$$\begin{aligned} &\sum_m \hat{\boldsymbol{X}}(k,m;\epsilon)^{\mathsf{H}}\mathbf{V}(k)\hat{\boldsymbol{X}}(k,m;\epsilon) \\ &= \sum_m \hat{\boldsymbol{X}}(k,m;\epsilon)^{\mathsf{H}} \left(\frac{\sum_{m'}\hat{\boldsymbol{X}}(k,m';\epsilon)\hat{\boldsymbol{X}}(k,m';\epsilon)^{\mathsf{H}}}{|\forall m|}\right)^{-1} \hat{\boldsymbol{X}}(k,m;\epsilon) \\ &= 2|\forall m|. \end{aligned} \tag{20}$$

Unfortunately, an estimate of ϵ that maximizes the likelihood $J(\epsilon)$ cannot be obtained analytically. In addition, as plotted in Fig. 4, $J(\epsilon)$ has not only a clear global maximum but also several local maxima. We discuss an effective

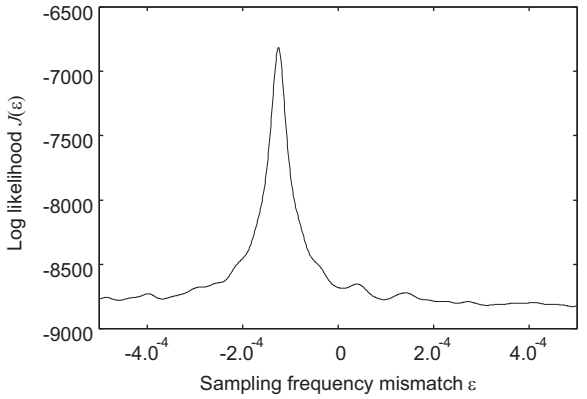

Fig. 4. Example of the log likelihood function $J(\epsilon)$ for the observation of a two-speech mixture of 10 s with the sampling frequency mismatch $\epsilon = -1.25 \times 10^{-4}$.

search method to obtain the global maximum in the following.

4.2. Efficient optimization of maximum likelihood estimate by golden section search

Since the only parameter to be estimated in the maximization of the likelihood is the sampling frequency mismatch ϵ, the problem reduces to a line search. If we can assume that the log likelihood function $J(\epsilon)$ is unimodal, efficient optimization can be obtained by performing a golden section search. In the example shown in Fig. 4, $J(\epsilon)$ given by (19) is usually locally unimodal around the global maximum. Thus, after specifying the unimodal range including the global maximum, we perform a golden section search.

To specify the unimodal range around the maximum, we discretize ϵ roughly and select the discretized value that maximizes $J(\epsilon)$. We generate D uniform samples in the range $[-E, E]$ $(E > 0)$ and define the discretized values ϵ_d, $d = 0, \ldots, D-1$ as

$$\epsilon_d = -E + \frac{2dE}{D-1}. \tag{21}$$

Then we compare all the values of $J(\epsilon_d)$ and find the optimal index d^* that maximizes $J(\epsilon_d)$ as

$$d^* = \mathrm{argmax}_d J(\epsilon_d), \ d = 0, \ldots, D-1. \tag{22}$$

The range parameter E can be defined easily by considering the possible range of ϵ. Since the sampling frequency mismatch normally takes a value on the order of 10^{-5}, E can be set to approximately 10^{-4} or larger. The appropriate choice of D depends on the setting of E and the signal characteristics. At least for the observation of speech, the shape tends to be similar to the example in Fig. 4, and the appropriate value of D should satisfy the condition $2E/(D-1) < 10^{-4}$.

After the search of the discretized values to obtain the coarse estimate ϵ_{d^*}, the golden section search refines the estimation by narrowing the search range in an iterative manner. We show the algorithm in Table 1. The initial search range is set to $[\epsilon_{d^*-1}, \epsilon_{d^*+1}]$, and the iteration continues until the range shrinks to the desired resolution $\rho(>0)$. Finally, the estimate ϵ is given as the center of the narrowed search range.

Table 1
Algorithm of the golden section search used to search for the maximum of $J(\epsilon)$.

Definition and initialization:
Set $\varphi \stackrel{\mathrm{def}}{=} (\sqrt{5}-1)/2$, $a \leftarrow \epsilon_{d^*-1}$, $b \leftarrow \epsilon_{d^*+1}$.
Step 1:
Set $p \leftarrow b - \varphi(b-a)$, $q \leftarrow a + \varphi(b-a)$.
Calculate $J(p)$ and $J(q)$.
Step 2:
If $J(p) \le J(q)$
  Set $a \leftarrow p$, $p \leftarrow q$, $q \leftarrow a + \varphi(b-a)$.
Else
  Set $b \leftarrow q$, $q \leftarrow p$, $p \leftarrow b - \varphi(b-a)$.
End if
Step 3:
If $b - a > \rho$
  Go back to Step 1.
Else
  Obtain the result $\epsilon \leftarrow (a+b)/2$.
  Terminate the algorithm.
End if

4.3. Robustness of proposed likelihood evaluation

Here we discuss the robustness of the evaluation of the log likelihood given by (19) against noise and the mismatch of the Gaussian assumption. Before the discussion of robustness, we simplify the log likelihood function given in (19) by discarding the residual constants. By using the coherence $\gamma_{12}^2(k;\epsilon)$ between channels after compensating the sampling frequency mismatch ϵ, defined as

$$\hat{\gamma}_{12}^2(k;\epsilon) = \frac{|\sum_m X_1(k,m)^* \hat{X}_2(k,m;\epsilon)|^2}{(\sum_m |X_1(k,m)|^2)(\sum_m |\hat{X}_2(k,m;\epsilon)|^2)}, \tag{23}$$

the log likelihood $J(\epsilon)$ is further simplified by separating the constants as follows:

$$\begin{aligned} J(\epsilon) &= -\sum_k \log \left(\left(\sum_m |X_1(k,m)|^2 \right) \left(\sum_m ||\hat{X}_2(k,m;\epsilon)|^2 \right) \right. \\ &\quad \left. - \left| \sum_m X_1(k,m)^* X_2(k,m;\epsilon) \right|^2 \right) \\ &= -\sum_k \log \left((1 - \hat{\gamma}_{12}^2(k;\epsilon)) \left(\sum_m |X_1(k,m)|^2 \right) \left(\sum_m |X_2(k,m)|^2 \right) \right) \\ &= -\sum_k \log(1 - \hat{\gamma}_{12}^2(k;\epsilon)) + \mathrm{const} \quad ., \end{aligned} \tag{24}$$

where $\{\cdot\}^*$ denotes the complex conjugate. Note that in (24) we used the equality $|\hat{X}_2(k,m;\epsilon)| = |X_2(k,m)|$, which is trivial from (14).

First, we discuss the effect of the non-Gaussianity of the observed signal. The complex amplitudes of the acoustic signal in the STFT domain can be regarded as having zero mean, but it is well known that many sound sources including speech tend to have super-Gaussian complex amplitudes in the STFT domain. As can be seen in (24), only the interchannel coherence is evaluated in the log likelihood function $J(\epsilon)$, it can be expected that non-Gaussianity does not affect the estimation of the sampling

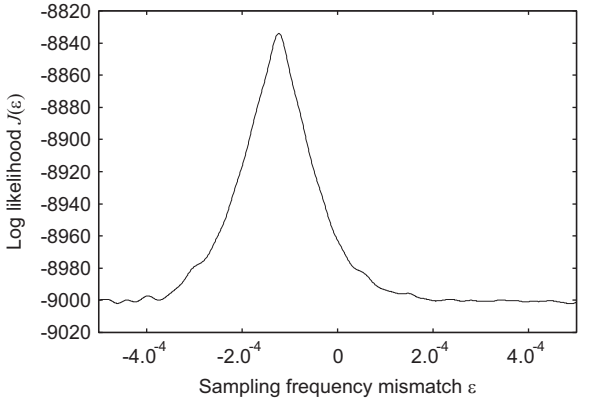

Fig. 5. Example of the log likelihood function $J(\epsilon)$ for the noisy observation of a two-speech mixture with the sampling frequency mismatch $\epsilon = -1.25 \times 10^{-4}$. The noise is uncorrelated white Gaussian with the SNR of 0 dB.

frequency mismatch ϵ so much. We will investigate this by experiments using speech.

Next, we discuss the robustness against noise. The estimation of ϵ is possible as long as the amplitudes at the two microphones are correlated because the log likelihood function $J(\epsilon)$ evaluates only the interchannel coherence $\gamma_{12}^2(k;\epsilon)$ by effectively weighting the significant frequency bins. In the frequency bins where only a few unmoving sources are observed, high stationarity is maintained only when drift is compensated properly. Particularly in the frequency where only one source is dominant and the coherence $\hat{\gamma}_{12}^2(k;\epsilon)$ is close to one, the kth frequency significantly affects the log likelihood $J(\epsilon)$ with infinitely high value of $-\log(1-\hat{\gamma}_{12}^2(k;\epsilon))$. In contrast, $J(\epsilon)$ is not strongly affected by noisy frequency bins with low coherence because of the low value of $-\log(1-\hat{\gamma}_{12}^2(k;\epsilon)) \approx 0$. Therefore, the log likelihood $J(\epsilon)$ automatically takes only the significant frequency bins into account, and the maximum likelihood function estimation of the sampling frequency mismatch is expected to have high robustness against noise. Fig. 5 shows an example of the log likelihood of an asynchronous observation of asynchronous speech mixture used in Fig. 4 with the uncorrelated white Gaussian noise imposed. We can see that the log likelihood function $J(\epsilon)$ is slightly flattened by the noise but the peak is still clear under this severe condition with the very low SNR of 0 dB.

Note that we have confirmed that the stationarity can be evaluated satisfactorily using $J(\epsilon)$ even when there are many sources in the observation. However, the proposed method is weak against the large movement of the sources because the movement degrades the stationarity and causes the time variation of the interchannel correlation.

5. Estimation of drift-compensated signal by STFT analysis with noninteger frame shift

In this section, we describe an efficient resampling method based on modified STFT analysis to estimate the STFT signal with the drift compensated under the condition that the sampling frequency mismatch ϵ and the recording start offset D_{21} are given. Here we assume that the drift inside each frame can be ignored, as discussed in Section 3.2. This signal estimation procedure is used in both the parameter estimation and the drift compensation as discussed in Section 6. Although it is possible to use various resampling methods to compensate for the drift with the parameters ϵ and D_{21} given, many of the accurate resampling methods are computationally complex. Our proposed method is a simple modification of STFT analysis and its computational cost is slightly larger than that of the original analysis. Also, the time-domain signal is obtained by inverse STFT (ISTFT) analysis.

As we showed in Section 3.2, the drift inside each frame can be ignored under the condition given by (9) that $|\epsilon L|$ is small, and the drift can be compensated by shifting the frame central time. According to (11), the time of the second channel synchronous with the time m at the center of the frame of the first channel is given by $\phi_{21}(m)$. Thus, we estimate the STFT signal $X_2(k,\phi_{21}(m))$ whose frame center is positioned at the discrete time $\phi_{21}(m)$.

In the beginning, we perform the STFT analysis of $x_1[n]$ as in (7):

$$X_1(k,m) = \sum_{l=0}^{L-1} w(l) x_1\left[l+m-\frac{L}{2}\right] \exp\left(-\frac{2\pi j k l}{L}\right), \tag{25}$$

with a uniform integer frame shift of R samples, e.g., $m = 0, R, 2R, \ldots$. With such integer values of m, the corresponding frame central times $\phi_{21}(m)$ of the second channel are generally noninteger-valued. Since the precise estimation of the amplitude at a noninteger discrete time is too complicated as shown in (6), we approximate the frame analysis with the frame center rounded to the nearest integer in the time domain, and compensate the round-off error of the frame center by a linear phase in the STFT domain. The rounded central time is written as $\lfloor \phi_{21}(m) \rceil$, where $\lfloor \cdot \rceil$ denotes the rounding operation to the nearest integer, i.e.,

$$\lfloor \phi_{21}(m) \rceil = \arg\min_n |\phi_{21}(m) - n|, \quad n \in \mathbb{Z}. \tag{26}$$

Since the error appears as the delay of $(\phi_{21}(m) - \lfloor \phi_{21}(m) \rceil)$ samples, the linear phase compensation gives $X_2(k,\phi_{21}(m))$ as

$$X_2(k,\phi_{21}(m)) = \sum_{l=0}^{L-1} w(l) x_2\left[l+\lfloor \phi_{21}(m) \rceil - \frac{L}{2}\right] \cdot \exp\left(\frac{2\phi j k(\phi_{21}(m) - \lfloor \phi_{21}(m) \rceil)}{L}\right). \tag{27}$$

Now the obtained STFT signal $X_2(k,\phi_{21}(m))$ is synchronous with $X_1(k,m)$, which is analyzed using the frame shift R. Thus, when we need to obtain the time domain signal $\hat{x}_2[n]$ with the drift compensation, a good approximation is obtained by the ISTFT analysis of $X_2(k,\phi_{21}(m))$ using the original integer frame shift R.

6. Algorithm of blind drift compensation

Here we describe an algorithm for blind compensation of the drift combining the maximum likelihood estimation of ϵ and the optimized STFT analysis with the noninteger frame shift. The algorithm iteratively estimates the offset D_{21} in the time domain and the sampling frequency

mismatch ϵ in the STFT domain. Although an error of the estimation of D_{21} much smaller than the frame length in STFT analysis is accepted in array signal processing without the use of TDOA priors, as described in Section 1, satisfactory estimation cannot be guaranteed under the existence of drift. Also, the unignorable error of D_{21} degrades the estimation accuracy of ϵ owing to the insufficient STFT analysis. Thus, the estimations of D_{21} and ϵ are mutually dependent, and the estimation procedures are iterated to resolve the dependence.

6.1. Initial frame analysis

First we obtain the initial STFT signal with the appropriately compensated offset to evaluate the sampling frequency mismatch in the STFT domain.

In this initial frame analysis, we assume that sampling frequency mismatch does not occur, i.e.,

$$\epsilon \leftarrow 0. \tag{28}$$

Under this assumption, an integer value of D_{21} with an acceptable error much smaller than the frame length is obtained to maximize the correlation between channels:

$$D_{21} \leftarrow \arg\max_{\delta} \sum_{n=\max(0,\delta)}^{\min(N_1,N_2+\delta)-1} x_1[n]x_2[n-\delta], \quad -N_2 < \delta < N_1. \tag{29}$$

Note that the error of this estimation of D_{21} cannot be made sufficiently small when the assumption $\epsilon \approx 0$ in (28) has a large error and the observation length is large. Thus, the estimation of D_{21} is renewed after the estimation of ϵ in the STFT domain. Using this estimate of D_{21}, we obtain the STFT signal $x'_i(n)$, $i=1,2$ with the redefined discrete time that cancels the offset as

$$x'_1[n] \leftarrow x_1[n], \tag{30}$$

$$x'_2[n] \leftarrow x_2[n-D_{21}]. \tag{31}$$

Finally, we perform the initial STFT analysis to obtain $X'_i(k,m)$, $i=1,2$ using the uniform frame shift R, e.g., $m=0,R,2R,\ldots$, for both channels, as

$$X'_i(k,m) = \sum_{l=0}^{L-1} w(l)x'_i\left[l+m-\frac{L}{2}\right]\exp\left(-\frac{2\pi jkl}{L}\right). \tag{32}$$

6.2. Estimation of sampling frequency mismatch

Here we describe the estimation of the sampling frequency mismatch ϵ. The estimation is given as the update for the estimate of ϵ from the current one. We formulate this problem to obtain a parameter ϵ' which updates the estimation of the sampling frequency $(1+\epsilon)f_s$ as follows:

$$(1+\epsilon)f_s \leftarrow (1+\epsilon')(1+\epsilon)f_s. \tag{33}$$

Although in Section 4 we assumed that the third term $-(1+\epsilon)D_{21}$ on the right of (8) is zero, this term should be considered in practical processing even though we do not have accurate estimates of ϵ and D_{21} here. Since the time-domain signal $x'_i[n], i=1,2$ has been given a time shift to maximize the correlation, the gap between the corresponding discrete times $|n_2-n_1|$ is expected to be minimum around the central sample M of the samples overlapping between the channels, where M is given by

$$M \leftarrow \left\lfloor \frac{\min(N_1-D_{21},N_2)-\max(0,D_{21})-1}{2} \right\rfloor. \tag{34}$$

Thus, the drift-compensated STFT signal $\hat{X}'(k,m;\epsilon)$ with the minimal effect of the third term $-(1+\epsilon)D_{21}$ in (8) is obtained by a shift of the $\epsilon(m-M)$ samples to $X'_2(k,m)$ as

$$\hat{X}'_2(k,m;\epsilon) = X'_2(k,m)\exp\left(\frac{2\pi jk\epsilon(m-M)}{L}\right). \tag{35}$$

According to (19), the log likelihood function $J'(\epsilon')$ used to evaluate the compensated STFT signal $\hat{X}'_2(k,m;\epsilon')$ is given as

$$J'(\epsilon') = -\sum_k \log\det \sum_m \hat{\boldsymbol{X}}'(k,m;\epsilon')\hat{\boldsymbol{X}}'(k,m;\epsilon')^{\mathrm{H}}, \tag{36}$$

$$\hat{\boldsymbol{X}}'(k,m;\epsilon') = [X'_1(k,m),\hat{X}'_2(k,m;\epsilon')]^{\mathrm{T}}. \tag{37}$$

The value of ϵ' that maximizes $J'(\epsilon')$ is estimated by a simple modification of the procedure in Section 4.2, where ϵ is substituted with ϵ' and $J(\epsilon)$ is substituted with $J'(\epsilon')$ in which the current estimate ϵ is treated as a constant.

Finally after the estimation of ϵ' by the golden section search, we update ϵ using ϵ' as

$$\epsilon \leftarrow (1+\epsilon)(1+\epsilon')-1, \tag{38}$$

which gives the update of the estimated sampling frequency $(1+\epsilon)f_s$ in (33).

6.3. Estimation of offset

Now we describe the update of the estimation of the offset D_{21}. First, we obtain an estimate of the drift-compensated time-domain signal $\hat{x}_2[n]$ by the procedure in Section 5 with the substitution of the estimated ϵ and the assumption $D_{21}=0$. Then the offset D_{21} is estimated by maximizing the correlation between $x_1[n]$ and $\hat{x}_2[n]$:

$$D_{21} \leftarrow \arg\max_{\delta} \sum_{n=\max(0,\delta)}^{\min(N_1,N_2+\delta)-1} x_1[n]\hat{x}_2[n-\delta], \quad -N_2 < \delta < N_1. \tag{39}$$

Finally, we estimate the drift-compensated STFT signal $X_2(m,\phi_{21}(m))$ by substituting the estimated parameters ϵ and D_{21} again into the procedure in Section 5.

6.4. Iterative update

The estimation procedure for ϵ in Section 6.2 and that for D_{21} in Section 6.3 are conducted separately. However, the estimates of these parameters are mutually dependent and the error of one parameter propagates to the other. In particular, when the observation is long and ϵ is large, the initial STFT analysis with the assumption $\epsilon=0$ is strongly affected by the mismatch of the assumption in (12) and degrades the estimation accuracy of ϵ in Section 6.2; thus, the estimation accuracy of D_{21} in Section 6.3 is also subsequently degraded. To achieve accurate parameter

Table 2
Iterative algorithm of blind synchronization.

Initialization:
Set the iteration number $j \leftarrow 1$.
Set $\epsilon \leftarrow 0$.
Initialize $X_i'(k,m),\ i=1,2$ according to Section 6.1.
Step 1:
Update ϵ according to Section 6.2.
Step 2:
Update D_{21} and obtain $X_1(k,m)$ and $X_2(k,\phi_{21}(m))$ according to Section 6.3.
Step 3:
Set $j \leftarrow j+1$.
If j reaches the maximum iteration number
Terminate the algorithm.
Else
Update $X_2'(k,m)$ as in (40).
Go back to Step 1.
End if

estimation in such a long observation, we propose the iteration of these procedures to reduce the propagation of the error in each stage to the next stage.

Table 2 summarizes the iterative optimization algorithm. After the processing described in Sections 6.1–6.3, we return to the procedure in Section 6.2 after the following update of the STFT signal in the second channel $X_2'(k,m)$ for the update of ϵ:

$$X_2'(k,m) \leftarrow X_2(k,\phi_{21}(m)). \tag{40}$$

The procedures in Sections 6.2 and 6.3 after the update given by (40) are repeated until the iteration reaches the maximum number of iterations. The iterative algorithm is summarized in Table 2. Note that such iteration is required only for particularly long observations such as those for 10 min.

7. Experimental evaluation

In this section we evaluate the effectiveness of the proposed method of blind drift compensation. First, we evaluate the estimation accuracy of the sampling frequency mismatch by the proposed method. Second, we evaluate the performance of BSS with the drift compensation to assess the suitability of the proposed drift compensation as a preprocessing step in array signal processing.

7.1. Experimental setup

The observed signals are two-channel recordings of two-speaker mixtures made by the convolution of measured impulse responses and speech signals. The speech signals are made by the concatenation of word utterances chosen from the utterances in the ATR Japanese speech database [19]. We evaluated all 12 combinations of two speakers selected from two male and two female speakers. The original sampling frequency of the observation was 16,000 kHz, and we modified the sampling frequency of one channel by ± 0.5, ± 1, and ± 1.5 Hz. These modifications correspond to sampling frequency mismatches of ± 31.25, ± 62.5, and ± 93.75 ppm (parts per minute, 10^{-6}), respectively, which are realistic values for the practical bias of clock generators in real ADCs. To generate an artificial sampling frequency mismatch, we performed resampling with a polyphase filter, which is implemented as a standard command in MATLAB. We used auxiliary-function-based independent vector analysis [20] to conduct two-channel BSS of the two speakers. The placements of microphones and speakers are shown in Fig. 6. The sources are placed 1.5 m apart from the middle of the two microphones. We evaluated two different source positions, and the combinations of the horizontal angles of the two sources are $[-50^\circ, 30^\circ]$ and $[-60^\circ, -10^\circ]$. Other conditions are listed in Table 3.

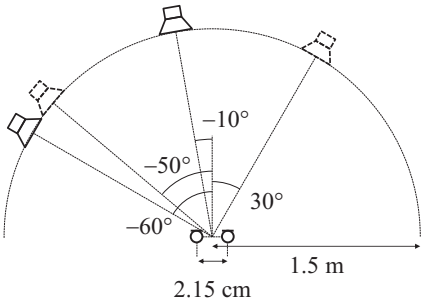

Fig. 6. Placement of microphones and sources. We evaluated two combinations of source positions with the horizontal angles of $[-50^\circ, 30^\circ]$ and $[-60^\circ, -10^\circ]$.

7.2. Accuracy of sampling frequency mismatch estimation

Fig. 7 shows the root mean squared errors (RMSEs) of the estimates of the sampling frequency mismatches for different signal lengths. We plotted the results of one and two iterations of the iterative estimation algorithm described in Section 6.4. We can see that the proposed estimation algorithm works appropriately even with short observed signals, and the accuracy improves with increasing length of the observed signals. Also we can see that the iterative procedure performs better when the observation is long, and the improved performance becomes clear for observations longer than 120 s. Thus, we confirmed that the estimation algorithm without iteration is satisfactory for short observations, but the iterative estimation algorithm is effective in the case of long observations.

7.3. Robustness

We analyzed the behavior of the proposed method of sampling frequency mismatch estimation to evaluate robustness. First, to evaluate the effect of the non-Gaussianity of the observed signal, we analyzed the kurtosis, which is a popular measure of Gaussianity. The kurtosis of complex variables [21] was evaluated in each frequency bin for each observation of 30 min, and the results were averaged over all the observations. Note that for the stability of estimation, the fourth-order moments were not obtained from the sample average but from the parameters of the gamma distributions fitted to the absolute squared amplitudes by using maximum likelihood estimators. The estimated kurtosis is shown in Fig. 8.

Table 3
Experimental conditions.

Length of observation	3, 5, 10, 20, 30, 60, 180, 300 and 1800 s
Reverberation time	T_{60} of 130 ms
Frame length L	4096 samples
Frame shift R	2048 samples
Microphone spacing	2.15 cm
Discretization search range E	5×10^{-4}
Discretization division D	20
Resolution ρ of golden section search	10^{-13}

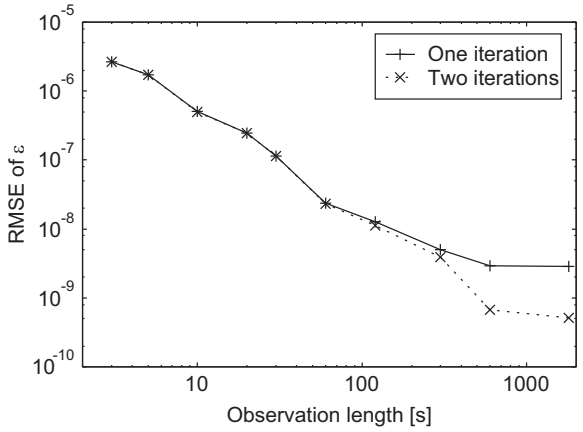

Fig. 7. Root mean squared errors (RMSEs) of estimates of sampling frequency mismatch.

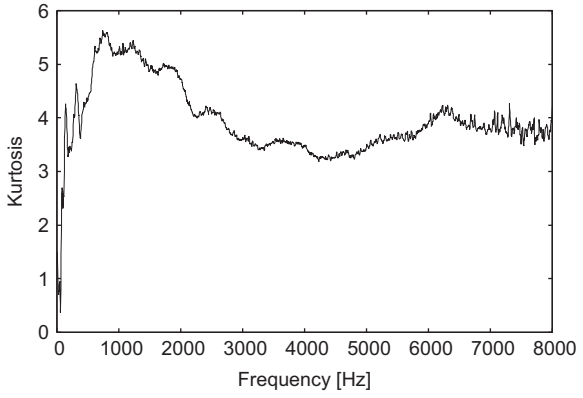

Fig. 8. Plot of kurtosis in each frequency bin.

The kurtosis of complex variables with a complex normal distribution is zero, and the kurtosis is larger than one when the variables are super-Gaussian. Fig. 8 shows that in most of the frequency bins the kurtosis is highly super-Gaussian, which is a characteristic of speech. Since accurate estimation was obtained for the super-Gaussian observation, it is confirmed that the non-Gaussianity of the observation does not affect the estimation.

To evaluate the robustness against noise, we superimposed uncorrelated white Gaussian noise on the observation and evaluated the RMSEs of the sampling frequency mismatch estimation. We compared the accuracy with a clean observation and noisy observations with SNRs of 20, 10, and 0 dB. The result is shown in Fig. 9. The RMSE increases with increasing noise, but the RMSE is only several times larger under a severe condition with an SNR of 0 dB than that for the clean observation. Therefore, it is confirmed that the proposed maximum likelihood estimation method is robust against noise as discussed in Section 4.3.

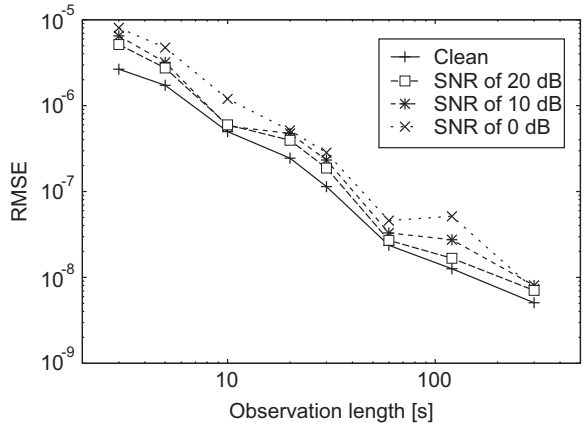

Fig. 9. Root mean squared errors (RMSEs) of estimated sampling frequency mismatch under the existence of uncorrelated white noise.

Since the proposed maximum likelihood estimation method assumes that the sources do not move so that the stationarity of the observation is maintained, the movement of sources degrades the estimation performance. To evaluate the effect of the movement of the sources, we changed the positions of the speakers from the directions $[-50^\circ, 30^\circ]$ to $[-60^\circ, -10^\circ]$ in the middle of the observation. Fig. 10 shows the RMSEs of the sampling frequency mismatch estimation when the speakers changed the positions. Although the estimation does not fail under this condition, the RMSEs increased approximately ten-fold. Thus, the proposed maximum likelihood estimation is weak against position shifts of the sources because of the degradation of stationarity.

7.4. Contribution to BSS

To assess the contribution of the proposed method of drift compensation to BSS, we evaluated the source separation performance in terms of the signal-to-distortion ratio (SDR) [22] at the first microphone as shown in Fig. 11. The SDRs were averaged for the two sources. We evaluated the following conditions:

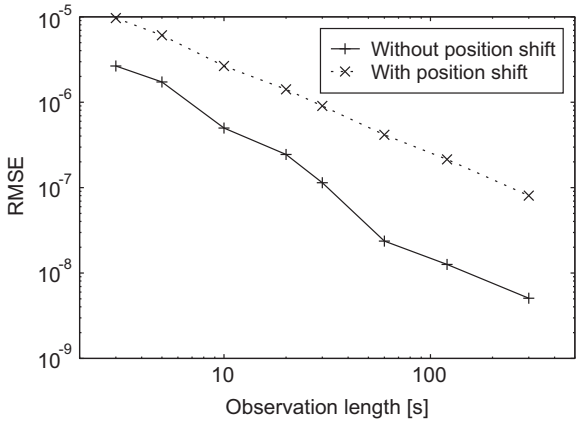

Fig. 10. Root mean squared errors (RMSEs) of estimated sampling frequency mismatch when the positions of the speakers changed in the middle of the observation.

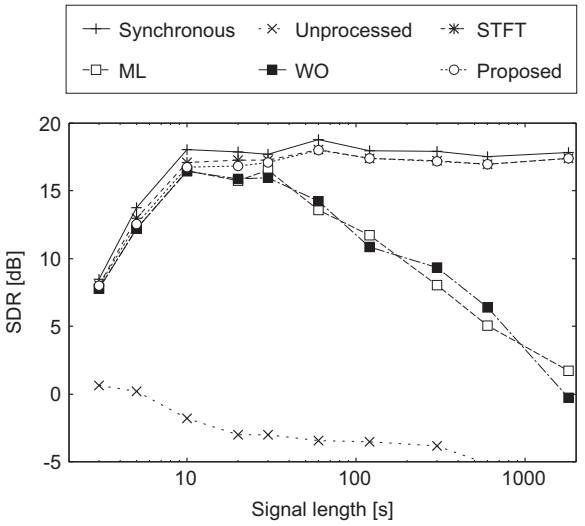

Fig. 11. Signal-to-distortion ratios (SDRs) of BSS performances for different conditions of drift compensation.

- *Synchronous*: BSS without artificial sampling frequency mismatch to obtain an upper limit for this task.
- *Unprocessed*: BSS without compensation of the asynchronous recording.
- *STFT-based resampling with true parameters (STFT)*: BSS of the STFT-based resampling with the true value of ϵ given and D_{21} estimated by the procedure in Section 6.3.
- *Maximum likelihood linear phase compensation (ML)*: BSS of the STFT signal obtained by the procedure in Section 4.2 without STFT-based resampling.
- *Without offset reestimation (WO)*: BSS of the STFT-based resampling without the reestimation of D_{21} obtained by the procedure in Section 6.3.
- *Proposed*: BSS of STFT-based resampling with drift compensation by the procedures in Section 6 without iteration of the procedures. The number of iterations in Section 6.4 was two.

Fig. 11 shows a comparison of the BSS performances. Since the SDRs of the unprocessed signal are very low, we can see that BSS is difficult without drift compensation under these conditions. The SDRs in the case of resampling with the true parameters are only 1 dB lower than those in the case of synchronous recording. Thus, we can conclude that the error of the signal expression in the STFT-based resampling is insignificant in array signal processing. The degradation in the SDRs of the maximum likelihood linear phase compensation is considerable for observations longer than 20 s compared with the other methods. Thus, resampling is essential and the proposed STFT-based resampling is satisfactory. The degradation in the SDRs of the BSS without offset reestimation is also considerable for observations longer than 30 s, but the proposed method does not exhibit such degradation. Thus, the effectiveness of the reestimation procedure for the offset described in Section 6.3 is confirmed. The SDRs of the proposed method exhibit insignificant degradation compared with STFT-based resampling with the true parameters, and the degradation from the synchronous recording was only about 1 dB. Thus, it is confirmed that the parameter estimation is sufficiently accurate. Therefore, the validity of the proposed compensation algorithm is verified, and its suitability for array signal processing as a process is also confirmed.

8. Conclusion

In this paper, we proposed a novel method for the blind compensation of drift arising in the asynchronous recording of an ad hoc microphone array. We modeled the drift as a shift of short time frames while ignoring the drift inside each frame. Thus, the compensation of drift was given by a linear phase in the STFT domain. By assuming that the sources are unmoving and have stationary amplitudes, we used stationarity as the cue to evaluate the drift compensation. A likelihood function to measure stationarity was formulated, and an efficient maximum likelihood search of the sampling frequency mismatch by a golden section search was described. We also proposed an efficient resampling method involving the modification of STFT analysis with a noninteger frame shift, and described the algorithm of the introduced blind drift compensation. In an experiment on the compensation of drift introduced artificially, we confirmed the accurate estimation and sufficient compensation of drift by the proposed method. The future work includes the online adaptation of the drift compensation.

References

[1] M. Brandstein, D. Ward (Eds.), Microphone Arrays: Signal Processing Techniques And Applications, Springer-Verlag, New York, Heidelberg, Berlin, 2001.
[2] Z. Liu, Z. Zhang, L.-W. He, P. Chou, Energy-based sound source localization and gain normalization for ad hoc microphone arrays, in: Proceedings of the ICASSP 2007, vol. II, 2007, pp. 761–764.
[3] O. Roy, M. Vetterli, Rate-constrained collaborative noise reduction for wireless hearing aids, IEEE Trans. Signal Process 57 (2) (2009) 645–657.

[4] M. Souden, K. Kinoshita, M. Delcroix, T. Nakatani, Distributed microphone array processing for speech source separation with classifier fusion, in: Proceedings of the MLSP 2012, 2012.
[5] A. Bertrand, Applications and trends in wireless acoustic sensor networks: a signal processing perspective, in: Proceedings of the SCVT 2011, 2011.
[6] V.C. Raykar, I.V. Kozintsev, R. Lienhart, Position calibration of microphones and loudspeakers in distributed computing platforms, IEEE Trans. Speech Audio Process. 13 (1) (2005) 70–83.
[7] N. Ono, H. Kohno, N. Ito, S. Sagayama, Blind alignment of asynchronously recorded signals for distributed microphone array, in: Proceedings of the WASPAA 2009, 2009, pp. 161–164.
[8] K. Hasegawa, N. Ono, S. Miyabe, S. Sagayama, Blind estimation of locations and time offsets for distributed recording devices, in: Proceedings of the LVA/ICA 2010, 2010, pp. 57–64.
[9] R. Lienhart, I. Kozintsev, S. Wehr, M. Yeung, On the importance of exact synchronization for distributed audio processing, in: Proceedings of the ICASSP, 2003, pp. 840–843.
[10] Z. Liu, Sound source separation with distributed microphone arrays in the presence of clock synchronization errors, in: Proceedings of the IWAENC 2007, 2007.
[11] E. Robledo-Arnuncio, T.S. Wada, B.-H. Juang, On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation, in: Proceedings of the WASPAA, 2007, pp. 21–24.
[12] S. Markovich-Golan, S. Gannot, I. Cohen, Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming, in: Proceedings of the IWAENC, 2012.
[13] F, Hoflinger, R. Zhang, J. Hoppe, A. Bannoura, L.M. Reindl, J. Wendeberg, M. Buhrer, C. Schindelhauer, Acoustic self-calibrating system for indoor smartphone tracking (ASSIST), in: Proceedings of the IPIN 2012, 2012.
[14] T. Janson, C. Schindelhauer, J. Wendeberg, Self-localization application for iPhone using only ambient sound signals, in: Proceedings of the IPIN 2012, 2012.
[15] J. Schmalenstroeer, R. Haeb-Umbach, Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model, in: Proceedings of the EUSIPCO 2013, 2013.
[16] S. Makino, T.-W. Lee, H. Sawada (Eds.), Blind Speech Separation, Springer-Verlag, New York, Heidelberg, Berlin, 2007.
[17] S. Miyabe, N. Ono, S. Makino, Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain, in: Proceedings of the ICASSP 2013, 2013, pp. 674–678.
[18] S. Miyabe, N. Ono, S. Makino, Optimizing frame analysis with non-integer shift for sampling mismatch compensation of long recording, in: Proceedings of the WASPAA 2013, 2013.
[19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, ATR Japanese speech database as a tool of speech recognition and synthesis, Speech Commun. 9 (4) (1990) 357–363.
[20] N. Ono, Stable and fast update rules for independent vector analysis based on auxiliary function technique, in: Proceedings of the WASPAA 2011, 2011, pp. 189–192.
[21] S. Javidi, D.P. Mandic, C.C. Took, A. Cichocki, Kurtosis-based blind source extraction of complex non-circular signals with application in EEG artifact removal in real-time, Front. Neurosci. 5 (105) (2011) 25, http://dx.doi.org/10.3389/fnins.2011.00105.
[22] E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1462–1469.