Kawamura *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:35
https://doi.org/10.1186/s13636-023-00300-y

EURASIP Journal on Audio,
Speech, and Music Processing

METHODOLOGY **Open Access**

Acoustic object canceller: removing a known signal from monaural recording using blind synchronization

Takao Kawamura[1*], Kouei Yamaoka[1], Yukoh Wakabayashi[2], Nobutaka Ono[1] and Ryoichi Miyazaki[3]

Abstract

In this paper, we propose a technique for removing a specific type of interference from a monaural recording. Nonstationary interferences are generally challenging to eliminate from such recordings. However, if the interference is a known sound like a cell phone ringtone, music from a CD or streaming service, or a radio or TV broadcast, its source signal can be easily obtained. In our method, we define such interference as an acoustic object. Even if the sampling frequencies of the recording and the acoustic object do not match, we compensate for the mismatch and use the maximum likelihood estimation technique with the auxiliary function to remove the interference from the recording. We compare several probabilistic models for representing the object-canceled signal. Experimental evaluations confirm the effectiveness of our proposed method.

Keywords Noise suppression, Acoustic object, Blind synchronization, Sampling frequency mismatch, Maximum likelihood estimation, Majorization–minimization algorithm

1 Introduction

Unlike multichannel recording, to which various array signal processing techniques can be applied, removing nonstationary noise from a monaural recording is generally challenging. Some algorithms [1–3] for noise suppression are based on estimating a noise power spectrum. Because these algorithms assume stationary noise, the accuracy of noise estimation is imperfect.

However, the situation is different when the sound source waveform of the interference sound is known in advance. For instance, it becomes feasible to obtain signal waveforms for specific sounds such as ringtones of mobile phones, commercially distributed music, television broadcasts, and similar sounds. We define these signals as acoustic objects. As with general noise removal, various applications can be considered for removing these acoustic objects. For example, one might wish to eliminate mobile phone ringtones or alarms that were inadvertently included in a recording, remove the music to circumvent copyright issues, or attenuate any interfering noise to enhance the precision of speech recognition and acoustic scene recognition. Additionally, it may be desirable to remove announcements that are specific to certain locations in order to anonymize the location of the recording. This study aims to achieve high-precision removal of the acoustic object from monaural recordings by utilizing it.

We treat the obtained acoustic object as a new channel and apply array signal processing. Note that the recording contains an acoustic object regardless of when or where it was acquired. However, the sampling frequencies of the recording and the available acoustic object can be mismatched even when the nominal

*Correspondence:
Takao Kawamura
kawamura-takao@ed.tmu.ac.jp
[1] Department of Computer Science, Tokyo Metropolitan University, Tokyo, Japan
[2] Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan
[3] Department of Computer Science and Electronic Engineering, National Institute of Technology, Tokuyama College, Yamaguchi, Japan

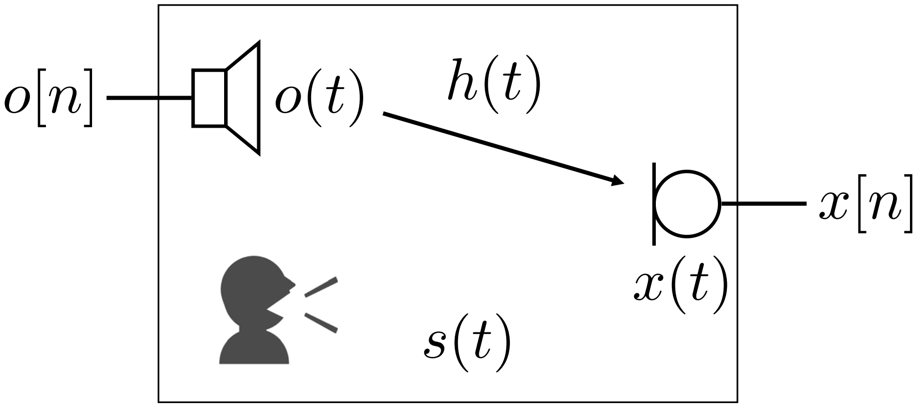

Fig. 1 Problem setting. The acoustic object $o(t)$ radiated through an acoustic path interferes with the monaural recording. Here, the loudspeaker has a D/A converter that converts $o[n]$ to $o(t)$, and the microphone has an A/D converter that converts $x(t)$ to $x[n]$

sampling frequencies are the same. The time drift due to the sampling frequency mismatch makes the frequency response time variant, which differs from the assumption in array signal processing.

An asynchronous microphone array [4–19], which consists of independent recording devices, also has a sampling frequency mismatch. Such a mismatch degrades the performance of signal processing [4–6]. To compensate the mismatch, blind synchronization methods that use only recordings without prior information have been proposed [7–16] and applied as preprocessing methods for array signal processing [17–19].

In this study, we propose an "acoustic object canceller," a framework to remove an acoustic object from a monaural recording. The monaural recording and the obtained acoustic object are treated as components of an asynchronous microphone array and we apply one of the blind synchronization methods [7] for compensating the sampling frequency mismatch. Then, the frequency response of the acoustic object is determined by the maximum likelihood estimation by the auxiliary function method, also known as the majorization–minimization (MM) algorithm [20], so the acoustic object is removed from the recording.

This paper is partially based on a conference paper [21] in which we proposed the framework of the acoustic object canceller. In summary, the main contributions of this paper are as follows.

- We consider three types of model for the object-canceled signal for the maximum likelihood estimation: generalized Gaussian distribution, multivariate Laplace distribution, and local Gaussian distribution.
- We experimentally investigate the dependence of the performance on the model parameters using the three types of desired sound and four types of acoustic objects.
- To confirm the effectiveness of the acoustic object canceller, we compare it with the amplitude-based noise suppression method [22].
- We also evaluated the sound quality of the proposed method using two speech quality metrics: Perceptual Evaluation of Speech Quality (PESQ) [23] and Short-Time Objective Intelligibility (STOI) [24].

The rest of this paper is organized as follows. In Section 2, we describe the problem setting. In Section 3, the acoustic object canceller is described. In Section 4, we carried out evaluations from the following four perspectives: (i) the effectiveness of the synchronization, (ii) the performance in the model for frequency response estimation, and (iii) comparison with the conventional method, and (iv) the evaluation of sound quality. We conclude the paper in Section 5.

2 Problem setting

We assume a situation where an acoustic object interferes with a monaural recording. Let $x(t)$ be a monaural recorded signal that is modeled by

$$x(t) = o(t - t_d) * h(t) + s(t), \tag{1}$$

where $o(t)$ and $h(t)$ are the acoustic object signal and the impulse response from the sound source of $o(t)$ to the microphone (see Fig. 1), and "$*$" is the convolution operator. $s(t)$ denotes signals other than the acoustic object signal, such as the desired signal to be recorded and background sound. The variable t_d is the time difference between $x(t)$ and $o(t)$. The objective of this study is to estimate $o(t - t_d) * h(t)$ including unknown variables $h(t)$ and t_d using $o(t)$ to obtain $s(t)$. Hereafter, we call $s(t)$ the target signal.

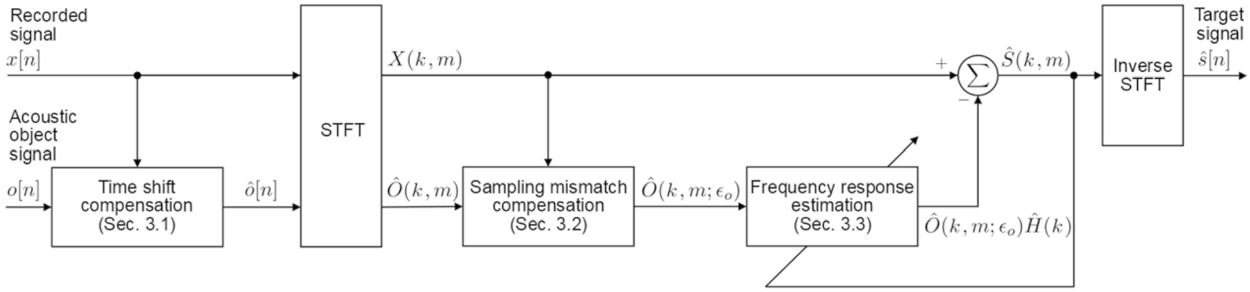
Recorded signal
$x[n]$
Acoustic object signal
$o[n]$
Time shift compensation (Sec. 3.1)
$\hat{o}[n]$
STFT
$X(k,m)$
$\hat{O}(k,m)$
Sampling mismatch compensation (Sec. 3.2)
$\hat{O}(k,m;\epsilon_o)$
Frequency response estimation (Sec. 3.3)
$\hat{O}(k,m;\epsilon_o)\hat{H}(k)$
+
−
Σ
$\hat{S}(k,m)$
Inverse STFT
Target signal
$\hat{s}[n]$

Fig. 2 Overview of procedures in proposed acoustic object canceller. First, we synchronize the recording and acoustic object (time shift compensation and sampling frequency mismatch compensation). Then the acoustic object is removed using an estimated frequency response

We formulate discrete signals of $x(t)$ and $o(t)$. Analog-to-digital (A/D) converters that sample the monaural recorded signal $x(t)$ and the acoustic object signal $o(t)$ are different. Therefore, even when the common nominal sampling frequencies are the same, the sampling frequency differs slightly, mainly because of differences in the clock generators. In this study, we assume that this slight sampling frequency mismatch can be expressed as an unknown time-invariant dimensionless quantity ϵ_o ($|\epsilon_o| \ll 1$) [7]. Let $x[n]$ and $o[n]$ denote the recorded signal and the acoustic object signal, expressed as

$$x[n] = x\left(\frac{n}{f_x}\right) = x\left(\frac{n}{(1+\epsilon_o)f_o}\right), \tag{2}$$

$$o[n] = o\left(\frac{n}{f_o}\right). \tag{3}$$

The $x[n]$ and $o[n]$ are the representations in the discrete-time domain, respectively. f_x and f_o denote the sampling frequency of the recording signal and acoustic object signal, respectively.

3 Acoustic object canceller

We propose a framework "acoustic object canceller" that removes the acoustic object signal from the monaural recorded signal. Figure 2 shows an overview of the signal processing blocks that make up the acoustic object canceller. The acoustic object canceller has two inputs and an output. The two inputs are the monaural recorded signal $x[n]$ and the obtained acoustic object signal $o[n]$. The output is the estimated target signal $\hat{s}[n]$ from which the acoustic object signal is removed.

The acoustic object canceller consists of three major processes: time shift compensation, sampling frequency mismatch compensation, and frequency response estimation. In time shift compensation, we synchronize the recorded and acoustic object signals by a rough time shift (detailed in Section 3.1). In sampling frequency mismatch compensation, we compensate for the sampling frequency mismatch using blind synchronization techniques proposed for ad-hoc microphone arrays [7] (detailed in Section 3.2). In frequency response estimation, the frequency response of the acoustic object signal is obtained by maximum likelihood estimation, assuming the model of the target signal (detailed in Section 3.3).

3.1 Time shift compensation

Time shift compensation is achieved by estimating the time difference t_d between the recorded signal $x[n]$ and the acoustic object signal $o[n]$ and time-shifting the acoustic object signal. However, accurate estimation of the continuous time difference t_d in Eq. (1) is challenging when a sampling frequency mismatch occurs, and $h(t)$ includes the time delay from the loudspeaker to the microphone. On the other hand, the estimation accuracy of t_d need not be perfect since slight estimation errors in t_d can be compensated by the frequency response estimation (described in Section 3.3). Therefore, the discrete-time difference τ is estimated instead of the continuous time difference t_d. Since we assume that the sampling frequency mismatch ϵ_o is $|\epsilon_o| \ll 1$, there is a sufficiently high correlation between $x[n]$ and $o[n]$ even without sampling frequency mismatch compensation. Thus, the estimated time difference $\hat{\tau}$ is calculated by finding the time shift τ that maximizes the cross-correlation function between $x[n]$ and $o[n]$:

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}}\left\{\sum_n o[n-\tau]x[n]\right\}. \tag{4}$$

Hereafter, the time-shifted version of $o[n]$ using the estimated time difference $\hat{\tau}$ is denoted by $\hat{o}[n] = o[n-\hat{\tau}]$.

3.2 Sampling frequency mismatch compensation

We compensate for the sampling frequency mismatch between the monaural recorded signal $x[n]$ and the time-shifted acoustic object signal $\hat{o}[n]$. Although resampling with a sinc function is necessary to compensate for sampling frequency mismatch, since most array signal processing is performed in the short-time Fourier transform (STFT) domain, it is efficient to obtain a compensated STFT domain representation. Therefore, some compensation techniques for sampling frequency mismatch based on the linear-phase drift (LPD) model have been proposed [7, 9, 14–16]. The LPD model applies a linear phase shift in the STFT domain. Hereafter, we denote the STFT domain representations of the monaural recorded signal $x[n]$ as $X(k, m)$ and the STFT domain representations of the time-shifted acoustic object signal $\hat{o}[n]$ as $\hat{O}(k, m)$. k and m are frequency and time frame indices, respectively.

In this study, we use the sampling frequency mismatch compensation technique [7]. We adopt the same assumptions and approximations as those in the application of sampling frequency mismatch compensation [17, 18]. We assume that the sources have stationary amplitudes and are motionless and approximate that the phase difference between channels caused by sampling frequency mismatch is constant in the time frame m. Then, the sampling frequency mismatch ϵ_o is compensated by a linear phase shift in the STFT domain. The signals compensated using accurate ϵ_o are expressed as

$$\hat{\mathbf{X}}(k, m; \epsilon_o) = \left[X(k, m), \hat{O}(k, m; \epsilon_o)\right]^{\top} \tag{5}$$

and are stationary at each discrete frequency k. Here, $\hat{O}(k, m; \epsilon_o)$ is the acoustic object signal with sampling frequency mismatch compensation by linear phase shift and is expressed as

$$\hat{O}(k, m; \epsilon_o) = \hat{O}(k, m) \exp\left(\frac{-2\pi jk\epsilon_o N_{\text{shift}} m}{N_{\text{FFT}}}\right), \tag{6}$$

where N_{FFT} and N_{shift} are the frame length and shift length of STFT, respectively. We assume that the monaural recorded and acoustic object signals $\hat{\mathbf{X}}(k, m; \epsilon_o)$ follow a multivariate Gaussian distribution with covariance matrix $\mathbf{V}(k)$, and accurate compensation of ϵ_o recovers the stationary $\hat{\mathbf{X}}(k, m; \epsilon_o)$. Then the log-likelihood function is expressed as

$$J(\mathbf{V}(k), \epsilon_o) = \sum_k \sum_m (-\hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}} \mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k, m; \epsilon_o) - \log \pi^2 - \log \det \mathbf{V}(k)), \tag{7}$$

where $\{\cdot\}^H$ denotes the conjugate transpose, and the covariance matrix $\mathbf{V}(k)$ is the parameter of the log-likelihood function. The $\mathbf{V}(k)$ is obtained by sample estimation using $\hat{\mathbf{X}}(k, m; \epsilon_o)$. The sample estimation for $\mathbf{V}(k)$ is described as

$$\mathbf{V}(k) \leftarrow \frac{1}{M} \sum_m \hat{\mathbf{X}}(k, m; \epsilon_o) \hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}}. \tag{8}$$

Here, M is the total number of time frames. Substituting Eq. (8) into Eq. (7), the first term in Eq. (7) is constant, as derived by the following equation:

$$\begin{aligned}
&\sum_k \sum_m -\hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}} \mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k, m; \epsilon_o) \\
&= \sum_k \sum_m -\mathrm{Tr}\left(\hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}} \mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k, m; \epsilon_o)\right) \\
&= \sum_k \sum_m -\mathrm{Tr}\left(\mathbf{V}(k)^{-1} \hat{\mathbf{X}}(k, m; \epsilon_o) \hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}}\right) \\
&= \sum_k -\mathrm{Tr}\left(\mathbf{V}(k)^{-1} \sum_m \hat{\mathbf{X}}(k, m; \epsilon_o) \hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}}\right) \\
&= \sum_k -\mathrm{Tr}(M \cdot \mathbf{I}) \\
&= -2MK.
\end{aligned} \tag{9}$$

where $\mathbf{I}$ and K indicate a 2×2 identity matrix and the total number of frequency bins, respectively. $\mathrm{Tr}(\cdot)$ denotes the trace of matrix. In this derivation, we use a matrix formula $\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{BCA})$ for any matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ such that $\mathbf{ABC}$ is a square matrix. The log-likelihood function simplifies to the following equation where the constant term is excluded.

$$J(\epsilon_o) = -\sum_k \log \det \sum_m \hat{\mathbf{X}}(k, m; \epsilon_o) \hat{\mathbf{X}}(k, m; \epsilon_o)^{\mathrm{H}} \tag{10}$$

When the sampling frequency mismatch ϵ_o is not compensated accurately, the log-likelihood function $J(\epsilon_o)$ will be small owing to the reduced stationary caused by drift. Therefore, we can estimate ϵ_o by maximizing $J(\epsilon_o)$. Unfortunately, an estimate of ϵ_o that maximizes the likelihood $J(\epsilon_o)$ cannot be obtained analytically. We perform a rough full search of ϵ_o and then a golden section search [7].

3.3 Frequency response estimation

From Eq. (1), when the length of the impulse response $h(t)$ is sufficiently smaller than the window length of STFT, the recorded signal in the STFT domain is described as

$$X(k, m) = \hat{O}(k, m; \epsilon_o) H(k) + S(k, m), \tag{11}$$

where $H(k)$ is the frequency response of the acoustic object signal. Note that $H(k)$ may not be the frequency response of the actual impulse response $h(t)$ due to the effect of $\hat{\tau}$ described in Eq. (4). We assume that $S(k, m)$ and $\hat{O}(k, m; \epsilon_o)$ are uncorrelated.

Table 1 Model of $S(k, m)$ and parameters. In Local Gaussian distribution based on NMF, we assume that the variance $r(k, m)$ can be expressed by NMF: $r(k,m) = \sum_c a(c,m)b(c,k)$

Model	p(S(k, m))	Parameters
Generalized Gaussian distribution	$p(S(k,m)) \propto \exp\left(-\left(\frac{\lvert S(k,m)\rvert}{\alpha}\right)^{\beta}\right)$	$H(k)$
Multivariate Laplace distribution	$p(\mathbf{S}(m)) \propto \exp\left(-\sqrt{\sum_k \left\lvert \frac{S(k,m)}{\sigma} \right\rvert^2}\right)$	$H(k)$
Local Gaussian distribution based on NMF	$p(S(k,m)) \propto \frac{1}{r(k,m)} \exp\left(-\frac{\lvert S(k,m)\rvert^2}{r(k,m)}\right)$	$H(k)$, $a(c, m)$, $b(c, k)$

The target signal $S(k, m)$ can be obtained by rewriting Eq. (11) as:

$$S(k,m) = X(k,m) - \hat{O}(k,m;\epsilon_o)H(k). \tag{12}$$

Since time-invariant $H(k)$ is the only unknown factor in Eq. (12), we focus on how to estimate it.

In this study, we adopt maximum likelihood estimation to estimate $H(k)$ instead of using power minimization, which has been commonly used in conventional echo and noise cancellers. It is known that assuming a suitable distribution $p(x)$, which represents the statistical characteristics of the desired sound, is effective in various applications, such as in echo canceller [25] and blind source separation [26–28]. We assume three distributions often assumed in the blind source separation: the generalized Gaussian distribution [26], multivariate Laplace distribution [27], and local Gaussian distribution with variance represented by nonnegative matrix factorization (NMF) [28] (see Table 1).

In maximum likelihood estimation, we estimate the frequency response as

$$\hat{H}(k) = \underset{H(k)}{\operatorname{argmin}}\, \mathcal{C}(H(k)), \tag{13}$$

where $\mathcal{C}(H(k))$ is a negative log-likelihood function and is described as

$$\mathcal{C}(H(k)) = \sum_k \sum_m -\log p(S(k,m)). \tag{14}$$

In the following section, we derive update formulae to estimate frequency response.

3.3.1 Generalized Gaussian distribution

The probability density function of the generalized Gaussian distribution is given as

$$p(S(k,m)) \propto \exp\left(-\left(\frac{\lvert S(k,m)\rvert}{\alpha}\right)^{\beta}\right), \tag{15}$$

where α and β are the scaling and shape parameters, respectively. It includes a Gaussian distribution when $\beta = 2$ and a Laplace distribution when $\beta = 1$. Hereafter, we consider $0 < \beta \leq 2$ that corresponds to a super-Gaussian distribution.

Under the above assumptions, the objective function to be minimized, that is, the negative log-likelihood function, is given by

$$\mathcal{C}(H(k)) = \frac{1}{\alpha^{\beta}} \sum_k \sum_m \lvert S(k,m)\rvert^{\beta}, \tag{16}$$

where parameter-independent terms are omitted. Note that, in the case of $\beta = 2$, minimizing Eq. (16) is equivalent to minimizing the power of the target signal $S(k, m)$, as has been commonly used in the conventional echo canceller and noise canceller.

The optimization problem to minimize Eq. (16) in terms of $H(k)$ has no closed-form solutions in the case of $\beta \neq 2$. We apply the auxiliary function method, also known as the majorization–minimization (MM) algorithm [20]. In the auxiliary function method, we define the auxiliary function, which is an upper bound of an objective function and is easier to optimize. Given the auxiliary function, we can derive an efficient algorithm that minimizes the objective function by iteratively minimizing the auxiliary function instead of the objective function.

An auxiliary function for Eq. (16) is obtained by the theorem described in [29]. According to the theorem, for the continuous and differentiable even function $G(x)$ of x, if $G'(x)/x$ is continuous, $x > 0$, positive, and monotonically decreasing,

$$G(x) \leq \frac{G'(x_0)}{2x_0}x^2 + \left(G(x_0) - \frac{x_0 G'(x_0)}{2}\right) \tag{17}$$

holds for any x, and the equality condition is $x = \pm x_0$.

From Eq. (17), the auxiliary function of Eq. (16) is calculated as

$$\begin{aligned}&\frac{1}{\alpha^{\beta}}\sum_k\sum_m|S(k,m)|^{\beta}\\&\leq\frac{1}{\alpha^{\beta}}\sum_k\sum_m\frac{\beta|S_0(k,m)|^{\beta-1}}{2|S_0(k,m)|}|S(k,m)|^2\\&=\frac{1}{\alpha^{\beta}}\sum_k\sum_m\frac{\beta|S_0(k,m)|^{\beta-2}}{2}|S(k,m)|^2,\end{aligned}\tag{18}$$

$$S_0(k,m)=X(k,m)-\hat{O}(k,m;\epsilon_o)H_0(k).\tag{19}$$

Here, $H_0(k)$ is an auxiliary variable. The auxiliary function can be written as follows. Note that terms that do not depend on $H(k)$ are omitted.

$$\mathcal{Q}(H(k),H_0(k))=\frac{1}{\alpha^{\beta}}\sum_k\sum_m\frac{\beta|S_0(k,m)|^{\beta-2}}{2}|S(k,m)|^2\tag{20}$$

Equation (20) has a closed-form solution of $H(k)$ because it is quadratic in form with respect to $H(k)$.

The following update equation is obtained by differentiating Eq. (20) with respect to $H(k)$ and setting it to 0 and then substituting the frequency response before the update for $H_0(k)$.

$$\hat{S}(k,m)\leftarrow X(k,m)-\hat{O}(k,m;\epsilon_o)\hat{H}(k)\tag{21}$$

$$\hat{H}(k)\leftarrow\frac{\sum_m\hat{O}^*(k,m;\epsilon_o)X(k,m)|\hat{S}(k,m)|^{\beta-2}}{\sum_m|\hat{O}(k,m;\epsilon_o)|^2|\hat{S}(k,m)|^{\beta-2}}\tag{22}$$

$\{\cdot\}^*$ denotes the complex conjugate operator. The estimated target signal $\hat{S}(k,m)$ is obtained by applying these updates sufficiently.

3.3.2 Multivariate Laplace distribution

The probability density function of the multivariate Laplace distribution is shown as

$$p(\mathbf{S}(m))\propto\exp\left(-\sqrt{\sum_k\left|\frac{S(k,m)}{\sigma}\right|^2}\right),\tag{23}$$

where $\mathbf{S}(m)=[S(0,m),S(1,m),\ldots,S(K,m)]^{\top}$ and σ is the scaling parameter. Equation (23) depends on the norm of the vector $\mathbf{S}(m)$ that assembles all frequency components of the target signal into one vector.

Using this probability density function, the objective function to be minimized, the negative log-likelihood function, can be obtained as

$$\mathcal{C}(H(k))=\frac{1}{\sigma}\sum_m\sqrt{\sum_k|S(k,m)|^2},\tag{24}$$

where terms that do not depend on $H(k)$ are omitted. In Eq. (24), $\sum_k|S(k,m)|^2$ is included in the square root and has no closed-form solution for $H(k)$.

Therefore, we apply the auxiliary function method to Eq. (24) to obtain the solution (see Appendix). We obtain the update rules shown as

$$\hat{S}(k,m)\leftarrow X(k,m)-\hat{O}(k,m;\epsilon_o)\hat{H}(k),\tag{25}$$

$$\hat{H}(k)\leftarrow\frac{\sum_m\frac{X(k,m)\hat{O}^*(k,m;\epsilon_o)}{\sqrt{\sum_k|\hat{S}(k,m)|^2}}}{\sum_m\frac{|\hat{O}(k,m;\epsilon_o)|^2}{\sqrt{\sum_k|\hat{S}(k,m)|^2}}}.\tag{26}$$

By sufficiently updating Eq. (25) and Eq. (26), the target signal is obtained as $\hat{S}(k,m)$.

3.3.3 Local Gaussian distribution based on NMF

The probability density function of the local Gaussian distribution is shown as

$$p(S(k,m))\propto\frac{1}{r(k,m)}\exp\left(-\frac{|S(k,m)|^2}{r(k,m)}\right),\tag{27}$$

where $r(k,\ m)$ is the variance of the local Gaussian distribution. We assume that the variance $r(k,\ m)$ can be expressed by NMF,

$$r(k,m)=\sum_c a(c,m)b(c,k),\tag{28}$$

where $a(c,m)$ and $b(c,k)$ denote the activation and the basis in NMF, respectively. c denotes the index of the basis.

The objective function, the negative log-likelihood function, is given by

$$\begin{aligned}&\mathcal{C}(H(k),a(c,m),b(c,k))\\&\quad=\sum_k\sum_m\frac{|S(k,m)|^2}{r(k,m)}+\log|r(k,m)|,\end{aligned}\tag{29}$$

where parameter-independent terms are omitted. Equation (29) is a quadratic form for $H(k)$ and has closed-form solutions. On the other hand, Eq. (29) has no closed-form solutions for $a(c,\ m)$ and $b(c,\ k)$ because the first and second terms are an inverse function and a logarithmic function of them, respectively.

Therefore, we apply the auxiliary function method to obtain the solutions for $a(c,\ m)$ and $b(c,\ k)$ (see Appendix). We obtain the following update formulae:

$$\hat{a}(c,m)\leftarrow\hat{a}(c,m)\sqrt{\frac{\sum_k\frac{\hat{b}(c,k)|X(k,m)-\hat{O}(k,m;\epsilon_o)\hat{H}(k)|^2}{(\sum_{c'}\hat{a}(c',m)\hat{b}(c',k))^2}}{\sum_k\frac{\hat{b}(c,k)}{\sum_{c'}\hat{a}(c',m)\hat{b}(c',k)}}},\tag{30}$$

$$\hat{b}(c,k) \leftarrow \hat{b}(c,k)\sqrt{\frac{\sum_m \frac{\hat{a}(c,m)|X(k,m)-\hat{O}(k,m;\epsilon_o)\hat{H}(k)|^2}{(\sum_{c'}\hat{a}(c',m)\hat{b}(c',k))^2}}{\sum_m \frac{\hat{a}(c,m)}{\sum_{c'}\hat{a}(c',m)\hat{b}(c',k)}}}, \tag{31}$$

$$\hat{H}(k) \leftarrow \frac{\sum_m \frac{X(k,m)\hat{O}^*(k,m;\epsilon_o)}{\sum_c \hat{a}(c,m)\hat{b}(c,k)}}{\sum_m \frac{|\hat{O}(k,m;\epsilon_o)|^2}{\sum_c \hat{a}(c,m)\hat{b}(c,k)}}. \tag{32}$$

The estimated target signal is obtained by sufficiently updating Eqs. (30), (31), and (32).

4 Experimental evaluations

4.1 Experimental conditions

In this experiment, we created a dataset by simulation. Initially, we generated $s[n]$ and $o[n] * h[n]$. To make $s[n]$ and $o[n] * h[n]$, we used Pyroomacoustics [30]. A $4.1 \times 3.8 \times 2.8$ m^3 virtual room where T_{60} is 0.40 s was considered. The speech and acoustic object sources and a microphone were randomly positioned in the virtual room and impulse responses were made.

There were three types of target signals $s[n]$: (i) speech signal convolved with the impulse response, (ii) environmental sound signal, and (iii) a mixture of speech signal convolved with the impulse response and environmental sound signal at 5 dB. As the speech signal, we used utterances in the Japanese Newspaper Article Sentences (JNAS) corpus [31]. This corpus includes utterance signals of a sentence in Japanese. To obtain an utterance signal longer than the object signal, we concatenated the utterance signals of the same speaker. As the environmental sound, we used the TUT Acoustic scenes 2016, Evaluation dataset [32]. This dataset includes a 10-s environmental sound signal. To obtain an environmental sound signal longer than the object signal, we concatenated the environmental sound signals of the sequential scene "Grocery store." In target (i), $S(k, m)$ is known to follow a super-Gaussian distribution [33]. For target (ii), $S(k, m)$ may follow a Gaussian distribution due to the presence of various sounds. In target (iii), $S(k, m)$ is considered to follow a distribution closer to Gaussian than in target (i).

The acoustic object signals $o[n]$ were the following four types of sound: Electronic Alarm, BGM, Broadcast, and Announce. We used a windows notification sound signal in the Electronic Alarm case. In the BGM case, we used the mixture signal of "ANiMAL-ClinicA" in DSD100 [34]. In the Broadcast case, we used an audio signal from a YouTube video [1]. In the Announce case, we used a train announcement signal of a JR East Yamanote Line in-train automatic announcement [35]. The signal length of the Electronic Alarm was 6 s. We clipped other signals (BGM, Broadcast, Announce) to 30 s. Electronic Alarm, BGM, Broadcast, and Announcement assumed a short duration of music, a long duration of music, combine a long duration of speech and music, and a long duration of speech, respectively. The sampling frequency of all signals was unified at 16,000 Hz.

To to make a recorded signal $x[n]$, we randomly determined a time difference t_d and mixed $s[n]$ and $o[n] * h[n]$ at an input signal-to-noise ratio (SNR).

$$\mathrm{SNR}_{\mathrm{input}} = 10\log_{10}\frac{\sum_n s[n]^2}{\sum_n (o[n] * h[n])^2} \tag{33}$$

Here, the sum of n is taken for the period when either the target signal or the acoustic object signal is not silent. The sampling frequency mismatch was simulated by resampling the recorded signals. For each input SNR and mismatch combination, we generated ten recorded signals with random time differences and source placements.

For evaluation, we used the SNR improvement is the difference between the input SNR $\mathrm{SNR}_{\mathrm{input}}$ and output SNR $\mathrm{SNR}_{\mathrm{output}}$ in Section 4.2, Section 4.3 and Section 4.4. We define the output SNR:

$$\mathrm{SNR}_{\mathrm{output}} = 10\log_{10}\frac{\sum_n s[n]^2}{\sum_n (\hat{s}[n] - s[n])^2}, \tag{34}$$

where $\hat{s}[n]$ is the estimated target signal.

For STFT, the fast Fourier transform was performed at 8192 points with a 4096-length Hamming window, and a shift length was half the window length. The number of update iterations was 20 for the generalized Gaussian and multivariate Laplace distributions to attain sufficient enhancement. However, for the local Gaussian distribution based on NMF, 20 was insufficient, and we iterated the updates 200 times to obtain sufficient enhancement. We set the initial frequency response $H_{\mathrm{init}}(k) = 1$.

4.2 Effectiveness of mismatch compensation

In this experiment, we evaluated the effectiveness of sampling frequency mismatch compensation from the following two perspectives: the difference in the acoustic object type and the difference in sampling frequency mismatch. We compared "w/o sync." and "w/ sync.," which indicate SNR improvement without and with blind synchronization, respectively. We used recorded signals $x[n]$ where the target signal $s[n]$ was target iii), and the acoustic object signals $o[n]$ were Alarm, BGM, Broadcast, and Announce. We set $\mathrm{SNR}_{\mathrm{input}}$ as 0 dB. The sampling frequency mismatches ϵ_o were ±31.25 and ±62.5 ppm. We assumed the multivariate Laplace distribution,

[1] https://www.youtube.com/watch?v=FISSRWuC7Y0&list=PL8oGxvf_g9snQdfDPrMQb0kzEmDRO4oUi&index=13

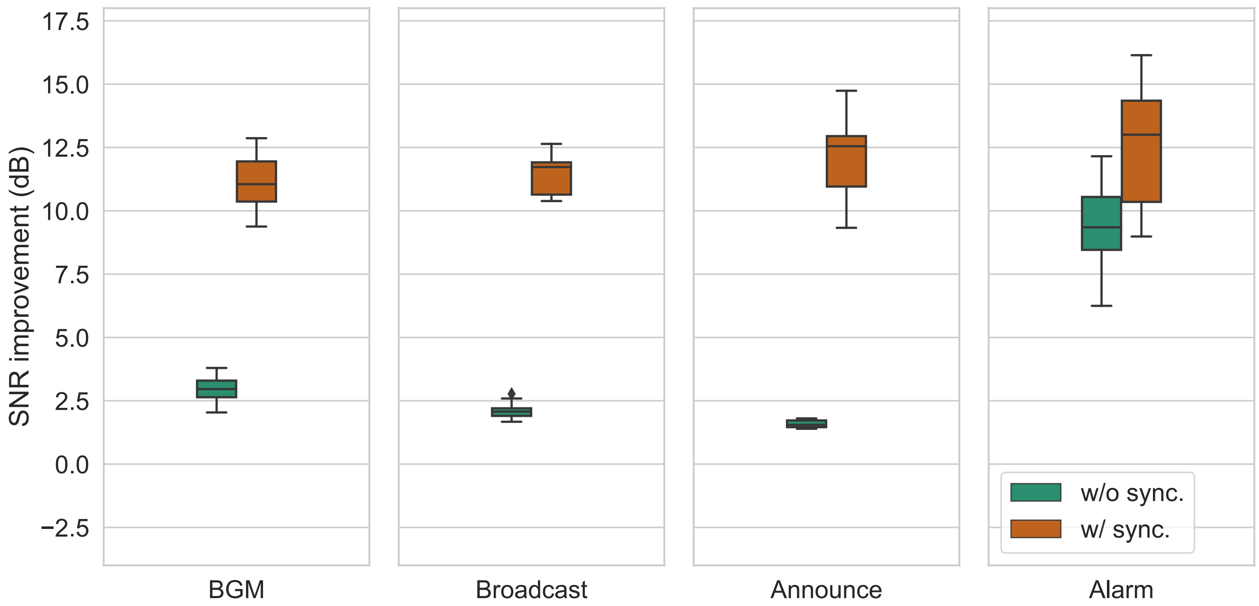

Fig. 3 SNR improvement between four types of acoustic objects (BGM, Broadcast, Announce, and Alarm) where target signal was target (iii), input SNR was 0 dB and sampling frequency mismatch was 62.5 ppm

which has no parameter dependence in the frequency response estimation because we investigated the parameter dependence of models for three types of target signals in Section 4.3.

Figure 3 shows the SNR improvement for different acoustic object types (BGM, Broadcast, Announce, and Alarm). We focus on the results of recorded signals, where the sampling frequency mismatch was 62.5 ppm. In the the box-plots, the box extends from the first to the third quartile of the SNR improvements, with a line at the median. The whiskers extend from the box by 1.5× the interquartile range. Outliers are those past the end of the whiskers. From Fig. 3, we have demonstrated that the performance was significantly improved by applying the blind synchronization technique. On the other hand, the difference in SNR improvement with and without blind synchronization was almost insignificant when the acoustic object was Alarm. It would be reasonable to infer that the shorter the signal length of the acoustic object signal, the less affected the performance and the less susceptible the sampling frequency mismatch. Therefore, it suggests that when the signal length of the acoustic object signal was short, it had less impact on performance by time drift due to sampling frequency mismatch.

Figure 4 shows the removal performance for different sampling frequency mismatches (±31.25, ±62.5 ppm), where positive and negative ppm correspond to upsampling and downsampling, respectively. We focused on the results where acoustic object signals were BGM, Broadcast, and Announce because these acoustic objects were significantly affected by sampling frequency mismatch. Figure 4 demonstrates the performance improvement upon applying the blind synchronization technique. In addition, we confirmed that the difference in SNR improvement between the absolute values of 62.5 ppm and 31.25 ppm of sampling frequency mismatch is about 2 dB when the synchronization method is not applied.

We have demonstrated that in environments where mismatches occur, the removal performance is affected by the signal length of the acoustic object (see Fig. 3) and the amount of sampling frequency mismatch (see Fig. 4) and that the blind synchronization technique could reduce these effects. We also confirmed these findings where the target signal was targets (i) and (ii) in the preliminary experiments.

Figure 5 shows examples of spectrograms. The upper left shows the recorded signal where we used target (iii), BGM, and set input SNR at 0 dB and sampling frequency mismatch at 62.5 ppm. The upper right shows the target (iii). The lower left shows the estimated target signal without blind compensation for sampling frequency mismatch. The lower right shows the estimated target signal with blind compensation for sampling frequency mismatch.

From Fig. 5, we confirm that the acoustic object signal was almost completely removed by the proposed method with blind synchronization (lower right) compared with that without synchronization (lower left).

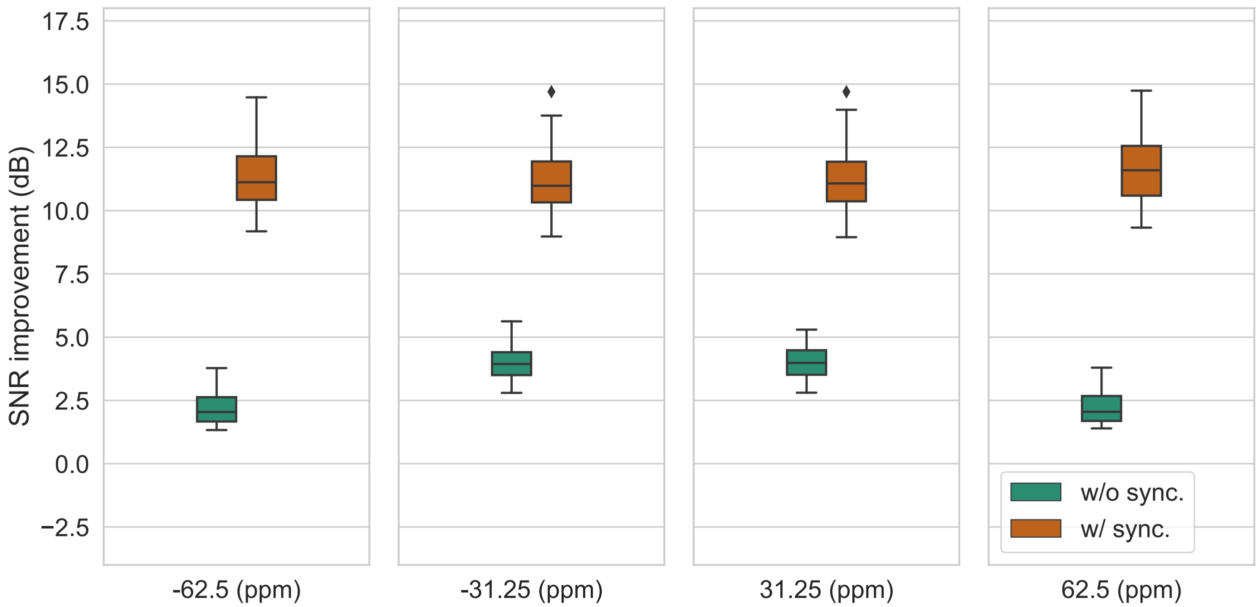

Fig. 4 SNR improvement between four types of sampling frequency mismatches (±31.25, ±62.5 ppm) where target signal is target (iii), input SNR was 0 dB and acoustic object signals were BGM, Broadcast, and Announce

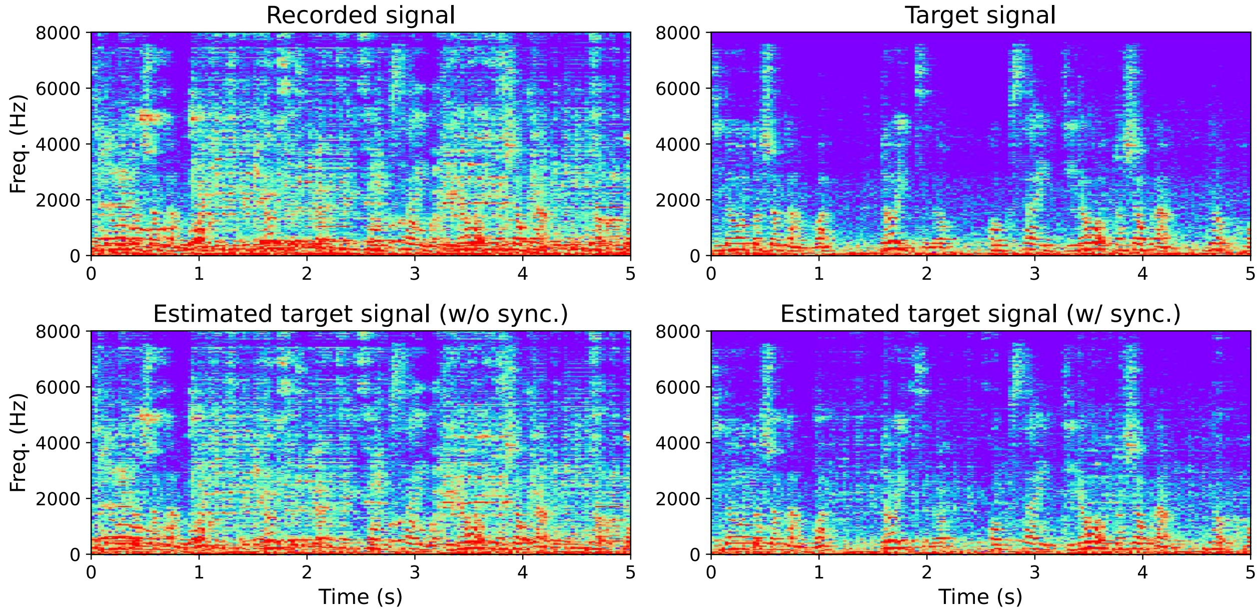

Fig. 5 Examples of spectrograms. Upper left shows the recorded signal where we used target (iii), BGM, and set input SNR at 0 dB and sampling frequency mismatch at 62.5 ppm. Upper right shows the target (iii). Lower left shows the estimated target signal without blind compensation for sampling frequency mismatch. Lower right shows the estimated target signal with blind compensation for sampling frequency mismatch

4.3 Performance with change in the model

We compared the SNR improvement of the models with various parameters for frequency response estimation. We used recorded signals $x[n]$ where the target signals $s[n]$ were target (i), target (ii) and target (iii), and 30-s acoustic object signals $o[n]$ (BGM, Broadcast, and Announce) that signal lengths are the same. We set $\text{SNR}_{\text{input}}$ as $-5, 0, 5, 10$ dB. The sampling frequency mismatches ϵ_o were ±31.25 and ±62.5 ppm. Since we confirmed the effectiveness of sampling frequency mismatch

Table 2 Average SNR improvement of different target models. The generalized Gaussian distribution is denoted as "L-GG," the multivariate Laplace distribution as "M-Laplace," and the local Gaussian distribution based on NMF as "L-G-NMF." Parameters are the shape parameter β for L-GG and the basis number c for L-G-NMF

Model	Parameter	Average SNR improvemnt (dB)		
		Target (i)	Target (ii)	Target (iii)
L-GG	0.4	10.66	9.75	10.36
	0.8	11.00	10.29	10.80
	1.2	11.19	10.61	11.00
	1.6	11.23	10.82	11.07
	2	11.01	10.94	10.96
M-Laplace	N/A	11.28	10.88	11.06
L-G-NMF	1	11.03	10.81	10.90
	2	11.02	10.79	10.94
	5	11.03	10.73	10.96
	10	10.97	10.67	10.91

compensation in the previous experiment, this experiment was focused only on the results with sampling frequency mismatch compensation. We set the shape parameter β in the generalized Gaussian distribution from 0.4 to 2.0 in 0.4 increments, and the basis numbers c in the local Gaussian distribution based on NMF to 1, 2, 5, and 10.

Table 2 shows the SNR improvement averaged by each target for each model. Here, the generalized Gaussian distribution is denoted as "L-GG," the multivariate Laplace distribution as "M-Laplace," and the local Gaussian distribution based on NMF as "L-G-NMF." According to Table 2, the averaged SNR improvement differed depending on the model and parameters.

First, we focused on the results for L-GG. For (i) speech signal, the highest performance was attained when $\beta < 2$, which corresponds to a super-Gaussian distribution. For (ii) environmental sound signal, the peak of performance was obtained with $\beta = 2$, which corresponds to a Gaussian distribution. We have demonstrated that the parameters that maximized the SNR improvement changed depending on the target signal type.

Second, we focused on the results for L-G-NMF. In L-G-NMF, there was no significant difference in SNR improvement for a change in the number of bases c. In L-G-NMF, when the number of bases is small, the model might have insufficient capability to represent the target signal. On the other hand, when the number of bases is large, the model might represent the target signal but also represent the acoustic object signal than when the number of bases is small. It would be reasonable to infer that results were not significantly changed by this tradeoff relationship.

Finally, we focused on the results for M-Laplace. We have demonstrated that the SNR improvement with M-Laplace was greater than with other models when the target signal was target (i). It suggests that M-Laplace represented the co-occurrence relationship of spectra and that the speech signal fit the model.

4.4 Comparison with amplitude-based method

In this experiment, we compared the proposed method with the conventional amplitude-based method [22] (see Appendix) from the following two perspectives: the difference in input SNR and the difference in the target type. We compared the SNR improvement among three approaches: "w/o sync." (without blind synchronization), "w/ sync." (with blind synchronization), and "Amp." (amplitude-based method). We utilized the recorded signals that were previously employed in Section 4.3. We assumed the multivariate Laplace distribution as the target model in the proposed method since Table 2 shows no significant performance differences, and multivariate Laplace distribution is parameter independent.

Figure 6 shows the SNR improvement for the four different types of input SNR ($-5, 0, 5, 10$ dB). Figure 6 demonstrates that the performance of the proposed method ("w/ sync.") was the best. In the high input SNR case, $o(t)$ is smaller than $s(t)$. This may reduce the estimation accuracy of τ and $h(t)$ and lead to a decrease in SNR improvement. We also confirmed that the conventional amplitude-based method showed little performance improvement when the input SNR was 10 dB.

Figure 7 shows the SNR improvement for each target type. According to Fig. 7, the SNR improvement of the conventional method is greater than that without synchronization. On the other hand, the results of the method with blind synchronization (proposed method) is higher than that of the conventional amplitude-based method. We have demonstrated the effectiveness of the proposed method with blind synchronization.

Figure 8 shows examples of spectrograms. The upper left shows the recorded signal where we used target (iii), BGM, and set input SNR at 0 dB and sampling frequency mismatch at 62.5 ppm. The upper right shows the target (iii). The lower left shows the estimated target signal of the conventional amplitude-based method. The lower right shows the estimated target signal of the proposed method.

In Fig. 8, we can confirm that the conventional and proposed methods almost completely removed the

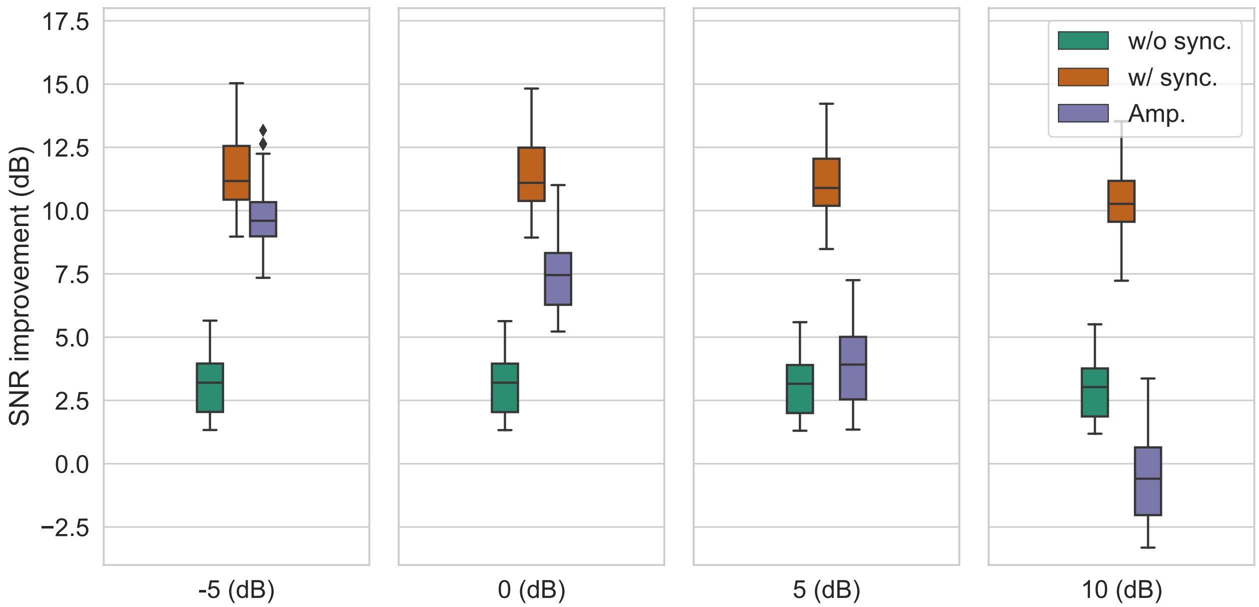

Fig. 6 SNR improvement between four types of input SNR (−5, 0, 5, 10 dB) where we use three types of target signals (targets i, ii, and iii), three types of acoustic object signals (BGM, Broadcast, and Announce), and four types of sampling frequency mismatches (±31.25, ±62.5 ppm)

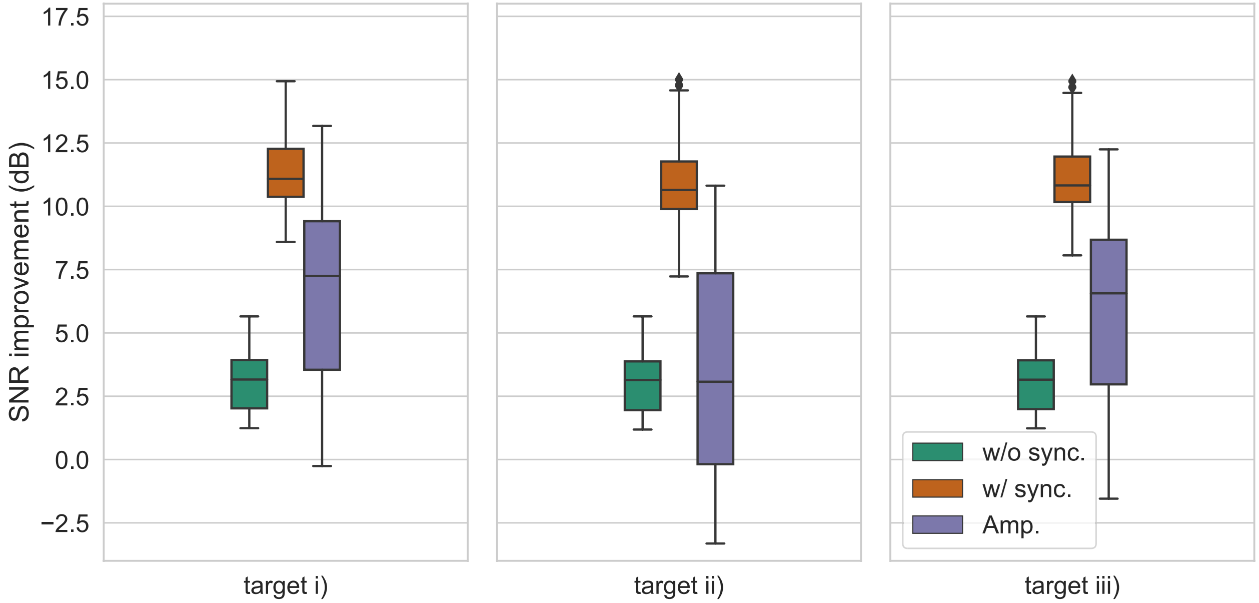

Fig. 7 SNR improvement between three types of target signal where we use three types of target signals (targets i, ii, and iii), three types of acoustic object signals (BGM, Broadcast, and Announce), four types of input SNR (−5, 0, 5, 10 dB), and four types of sampling frequency mismatches (±31.25, ±62.5 ppm)

acoustic object signal. A small target signal component was also removed in the conventional method, which may have contributed to the performance difference.

4.5 Evaluation of sound quality

In this experiment, we evaluated the sound quality of the proposed method by employing two speech quality metrics: Perceptual Evaluation of Speech Quality

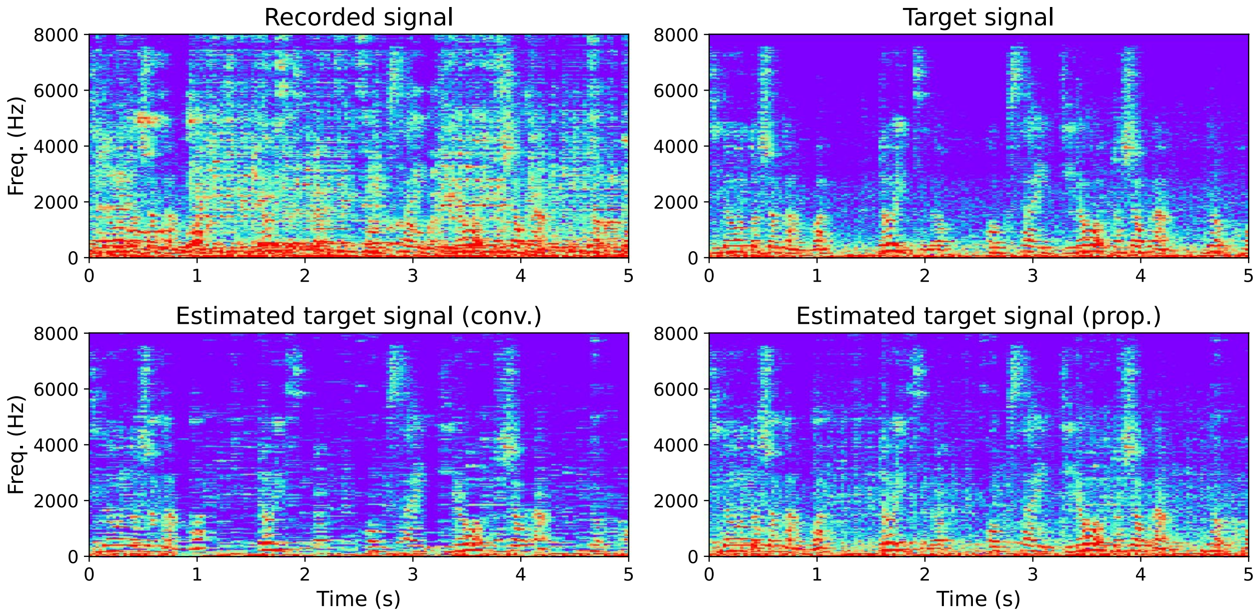

Fig. 8 Examples of spectrograms. Upper left shows the recorded signal where we used target (iii), BGM, and set input SNR at 0 dB and sampling frequency mismatch at 62.5 ppm. Upper right shows the target (iii). Lower left shows the target signal estimated by the conventional amplitude-based method. Lower right shows the target signal estimated by the proposed method

(PESQ) [23] and Short-Time Objective Intelligibility (STOI) [24]. The evaluation was conducted across different input SNRs to assess the robustness and performance of the proposed method. We conducted a comparison of the PESQ and STOI metrics among the following five types of signals: "Obs." (the recorded signal), "Amp." (amplitude-based method), "L-GG," "M-Laplace," and "L-G-NMF." We used recorded signals $x[n]$ where the target signals $s[n]$ was target (i), and 30-s acoustic object signals $o[n]$ (BGM, Broadcast, and Announce) that signal lengths are the same. We set $\mathrm{SNR}_{\mathrm{input}}$ as $-5, 0, 5, 10$ dB. The sampling frequency mismatches ϵ_o were ±31.25 and ±62.5 ppm. We set the shape parameter β in the generalized Gaussian distribution to 1.6, and the number of basis elements c in the local Gaussian distribution based on NMF to 5, as this parameter showed the highest SNR improvement in Table 2.

Figure 9 shows the average PESQ for the four different types of input SNR ($-5, 0, 5, 10$ dB). According to Fig. 9, the PESQ of the estimated target signals was higher than the recorded signal. We also confirmed that the PESQ of the estimated target signals of the amplitude-based method was lower than the proposed method. In particular, the larger the input SNR, the more significant the difference in PESQ between proposed and amplitude-based methods. It might be due to speech distortion in the amplitude-based method. We confirmed the effectiveness of the proposed method.

Figure 10 shows the average STOI for the four different types of input SNR ($-5, 0, 5, 10$ dB). According to Fig. 10, the STOI of the estimated target signals was higher than the recorded signal. We also confirmed that the STOI of the estimated target signals of the amplitude-based method was lower than the proposed method. In particular, the larger the input SNR, the more STOI differences between the proposed and amplitude-based methods. When the input SNR was 10 dB, the STOI of the estimated target signal by amplitude-based method showed little change from the STOI of the recorded signal. It might be due to speech distortion in the amplitude-based method. We confirmed the effectiveness of the proposed method.

5 Conclusion

In this study, we proposed the acoustic object canceller, a framework for removing the acoustic object signal from the monaural recorded signal. In the acoustic object canceller, first, we synchronized the monaural recorded signal and the available acoustic object signal. Second, we estimated the frequency response of the acoustic object by the maximum likelihood estimation assuming three model types: generalized Gaussian distribution, multivariate Laplace distribution, and local Gaussian distribution based on NMF. In the experiments, we have demonstrated the effectiveness of applying the synchronization technique and investigated the performance of the model types.

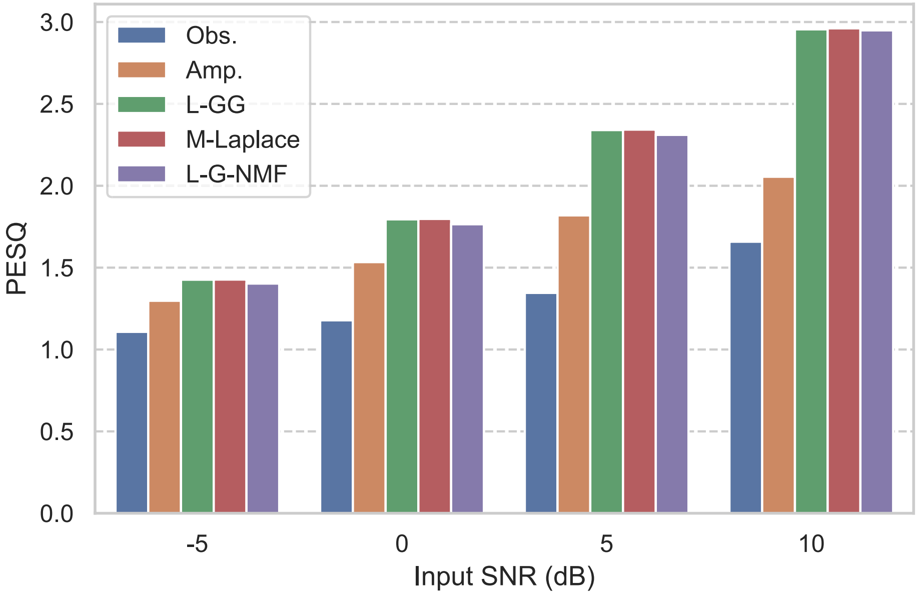

Fig. 9 Average PESQ between four types of input SNR (−5, 0, 5, 10 dB) where we use target (i) and three types of acoustic object signals (BGM, Broadcast, and Announce), and four types of sampling frequency mismatches (±31.25, ±62.5 ppm)

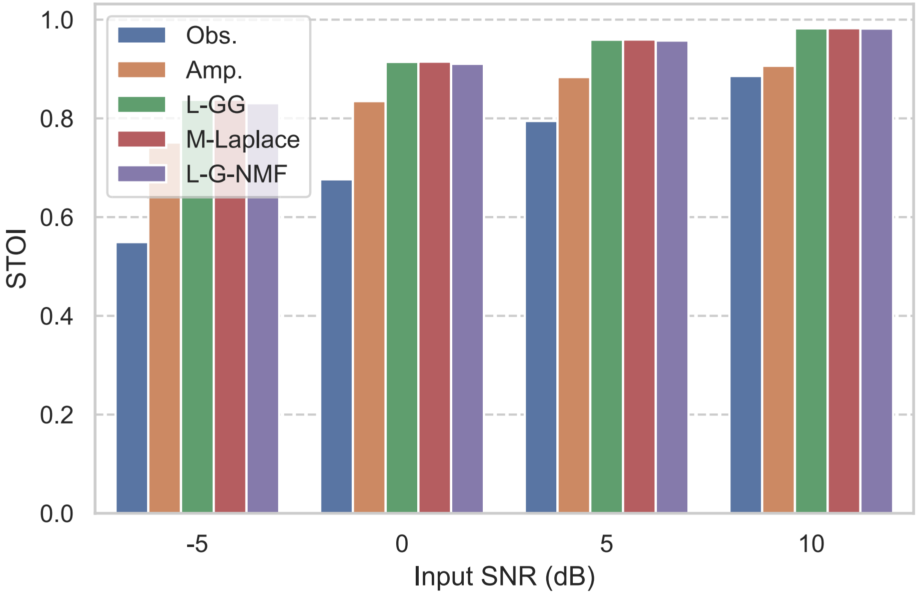

Fig. 10 Average STOI between four types of input SNR (−5, 0, 5, 10 dB) where we use target (i) and three types of acoustic object signals (BGM, Broadcast, and Announce), and four types of sampling frequency mismatches (±31.25, ±62.5 ppm)

Appendix

Frequency response estimation in multivariate Laplace distribution

In this appendix, we derive the frequency response $H(k)$ from Eq. (24) by the auxiliary function method. We use the properties of concave functions where the tangent line of concave functions always lies above the functions:

$$2\sqrt{x} \leq \frac{1}{\sqrt{x_1}}x + \sqrt{x_1}, \tag{35}$$

where the equality condition is $x = x_0$.

From Eqs. (24) and (35), the auxiliary function can be written as

$$\begin{aligned}&\mathcal{Q}(H(k), H_1(k))\\&\quad= \frac{1}{2\sigma}\sum_m \frac{1}{S_1(m)}\sum_k |S(k,m)|^2 + \frac{1}{2\sigma}\sum_m S_1(m),\end{aligned} \tag{36}$$

$$S_1(m) = \sqrt{\sum_k |X(k,m) - \hat{O}(k,m;\epsilon_o)H_1(k)|^2}, \tag{37}$$

where $H_1(k)$ is an auxiliary variable. Equation (36) has a closed-form solution because it is quadratic in form for $H(k)$. We differentiate Eq. (36) with respect to $H(k)$ and set it to 0, and then substitute the frequency response before the update for $H_1(k)$. Then, we obtain Eqs. (25) and (26).

Frequency response estimation in local Gaussian distribution based on NMF

Here, we derive the three parameters $H(k)$, $a(c, m)$, and $b(c, k)$ from Eq. (29) by the auxiliary function method. In the first term of Eq. (29), the inverse function is a convex function, so the auxiliary function is set from the following Jensen's inequality:

$$\frac{1}{\sum_i \lambda_i x_i} \leq \sum_i \lambda_i \frac{1}{x_i}, \tag{38}$$

where λ_i satisfies the condition

$$\sum_i \lambda_i = 1. \tag{39}$$

Therefore, from Eqs. (38) and (39), the auxiliary function of the first term in Eq. (29) is derived as

$$\begin{aligned}&\mathcal{Q}_1(H(k), a(c,m), b(c,k), \lambda(c,k,m))\\&\qquad= \sum_c \frac{\lambda(c,k,m)|S(k,m)|^2}{a(c,m)b(c,k)/\lambda(c,k,m)},\end{aligned} \tag{40}$$

$$\lambda(c,k,m) = \frac{a(c,m)b(c,k)}{\sum_{c'} a(c',m)b(c',k)}, \tag{41}$$

where $\sum_c \lambda(c,k,m) = 1$ and parameter-independent terms are omitted.

The second term in Eq. (29) cannot be solved in closed form for $a(c, m)$, $b(c, k)$ because the sum of the products of $a(c, m)$, $b(c, k)$ is in the logarithmic function. Therefore, to set the auxiliary function of the second term, we use inequalities characteristic of a concave function:

$$\log x \leq \frac{1}{x_0}(x - x_0) + \log x_0. \tag{42}$$

From Eq. (42), the auxiliary function of the second term in Eq. (29) is derived as

$$\begin{aligned}&\mathcal{Q}_2(H(k), a(c,m), b(c,k), \beta(k,m))\\&\quad= \frac{1}{\beta(k,m)}\left(\sum_c a(c,m)b(c,k) - \beta(k,m)\right) + \log\beta(k,m),\end{aligned} \tag{43}$$

where the equality condition for the inequality is

$$\beta(k,m) = \sum_c a(c,m)b(c,k). \tag{44}$$

From Eqs. (40) and (43), the auxiliary function of Eq. (29) is expressed as

$$\begin{aligned}&\mathcal{Q}(H(k), a(c,m), b(c,k), \lambda(c,k,m), \beta(k,m))\\&\quad= \sum_k\sum_m\sum_c \frac{\lambda(c,k,m)|S(k,m)|^2}{\frac{a(c,m)b(c,k)}{\lambda(c,k,m)}}\\&\quad+ \frac{1}{\beta(k,m)}\left(\sum_c a(c,m)b(c,k) - \beta(k,m)\right)\\&\quad+ \log\beta(k,m).\end{aligned} \tag{45}$$

Equation (45) has closed-form solutions for $a(c, m)$ and $b(c, k)$. We differentiate Eq. (45) with respect to $H(k)$, $a(c, m)$, and $b(c, k)$ and set it to 0 and then substitute the values before the update into the auxiliary variables $\lambda(c,k,m)$ and $\beta(k,m)$. Then, we obtain Eqs. (30), (31), and (32).

Derivation for amplitude-based method

Noise reduction algorithms use two different SNRs, the a posteriori and the a priori SNRs [22]. In Eq. (11), the a posteriori SNR is defined as

$$\gamma(k,m) = \frac{|X(k,m)|^2}{|O(k,m)H(k)|^2} = \frac{|X(k,m)|^2}{|O(k,m)|^2 G(k)}, \tag{46}$$

where $G(k)$ is $|H(k)|^2$ and an unknown factor. The a priori SNR is defined as

$$\xi(k,m) = \frac{|S(k,m)|^2}{|O(k,m)H(k)|^2} = \frac{|S(k,m)|^2}{|O(k,m)|^2 G(k)}. \tag{47}$$

We consider the Wiener filter as a noise reduction algorithm described as

$$\hat{S}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)}X(k,m). \tag{48}$$

To calculate filter coefficients, we estimate $G(k)$ that minimizes the difference between the power of recorded signal $\sigma_x(k)$ and that of the premix acoustic object signal $\sigma_o(k)$. The $G(k)$ is derived as

$$\hat{G}(k) = \underset{G(k)}{\operatorname{argmin}} |\sigma_x(k) - \sigma_o(k)|^2 = \frac{\sum_m |X(k,m)|^2}{\sum_m |O(k,m)|^2}. \tag{49}$$

We estimate a priori SNR as

$$\hat{\xi}(k,m) = \max\{0, \gamma(k,m) - 1\}. \tag{50}$$

Abbreviations

MM	Majorization–minimization
A/D	Analog-to-digital
STFT	Short-time Fourier transform
LPD	Linear-phase drift
NMF	Nonnegative matrix factorization
SNR	Signal-to-noise ratio
L-GG	Generalized Gaussian distribution
M-Laplace	Multivariate Laplace distribution
L-G-NMF	Local Gaussian distribution based on NMF
PESQ	Perceptual evaluation of speech quality
STOI	Short-time objective intelligibility

Acknowledgements

This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP20H00613.

Authors' contributions

TK proposed the methodology, conducted the experiments, and wrote the manuscript. KY and YW supervised the design of the experiments and refined the manuscript. NO supervised the entire part of the research and refined the manuscript. RM supervised the design of methodology and experiments. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The JNAS corpus used in the experiments of this paper is not publicly available due it is provided after submission and review of the Pledge of Use to Speech Resources Consortium. The TUT Acoustic scenes 2016, Evaluation dataset used in the experiments of this article is available in zenodo, https://zenodo.org/record/165995. DSD100 dataset used in the experiments of this paper is available at https://sigsep.github.io/datasets/dsd100.html JR East Yamanote Line in-train automatic announcement is not publicly available due it is commercially available at https://www.teichiku.co.jp/JReast/cd25530.html.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 16 March 2023 Accepted: 17 August 2023
Published online: 11 September 2023

References

1. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
2. R.C. Hendriks, R. Heusdens, J. Jensen, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. MMSE based noise PSD tracking with low complexity (IEEE, Dallas, TX, USA, 2010), pp. 4266–4269. https://doi.org/10.1109/ICASSP.2010.5495680
3. T. Gerkmann, R.C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2011)
4. Z. Liu, in *Proc. International Workshop for Acoustic Echo and Noise Control (IWAENC)*. Sound source separation with distributed microphone arrays in the presence of clock synchronization errors (2008)
5. R. Lienhart, I. Kozintsev, S. Wehr, M. Yeung, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. On the importance of exact synchronization for distributed audio signal processing (IEEE, Hong Kong, 2003), pp. IV–840. https://doi.org/10.1109/ICASSP.2003.1202774
6. E. Robledo-Arnuncio, T.S. Wada, B.H. Juang, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation (IEEE, New Paltz, NY, USA, 2007), pp. 34–37. https://doi.org/10.1109/ASPAA.2007.4393044
7. S. Miyabe, N. Ono, S. Makino, Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. Signal Process. **107**, 185–196 (2015)
8. L. Wang, S. Doclo, Correlation maximization-based sampling rate offset estimation for distributed microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 571–582 (2016)
9. J. Schmalenstroeer, J. Heymann, L. Drude, C. Boeddecker, R. Haeb-Umbach, in *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*. Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming (IEEE, Luton, UK, 2017), pp. 1–6. https://doi.org/10.1109/MMSP.2017.8122278
10. D. Cherkassky, S. Gannot, Blind synchronization in wireless acoustic sensor networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(3), 651–661 (2017)
11. M.H. Bahari, A. Bertrand, M. Moonen, Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(3), 674–686 (2017)
12. K. Itoyama, K. Nakadai, in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Synchronization of microphones based on rank minimization of warped spectrum for asynchronous distributed recording (IEEE, Las Vegas, NV, USA, 2020), pp. 4842–4847. https://doi.org/10.1109/IROS45743.2020.9341584
13. A. Chinaev, P. Thüne, G. Enzner, Double-cross-correlation processing for blind sampling-rate and time-offset estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1881–1896 (2021)
14. Y. Masuyama, K. Yamaoka, N. Ono, in *Proc. INTERSPEECH*. Joint optimization of sampling rate offsets based on entire signal relationship among distributed microphones (ISCA, Incheon, Korea, 2022), pp. 704–708. https://doi.org/10.21437/Interspeech.2022-97
15. T. Gburrek, J. Schmalenstroeer, R. Haeb-Umbach, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes (IEEE, Singapore, Singapore, 2022), pp. 916–920. https://doi.org/10.1109/ICASSP43922.2022.9746284
16. P. Didier, T. Van Waterschoot, S. Doclo, M. Moonen, Sampling rate offset estimation and compensation for distributed adaptive node-specific signal estimation in wireless acoustic sensor networks. IEEE Open J. Signal Process. **4**, 71–79 (2023)
17. K. Ochi, N. Ono, S. Miyabe, S. Makino, in *Proc. INTERSPEECH*. Multi-talker speech recognition based on blind source separation with ad hoc microphone array using smartphones and cloud storage (ISCA, San Francisco, USA, 2016), pp. 3369–3373. https://doi.org/10.21437/Interspeech.2016-758
18. S. Araki, N. Ono, K. Kinoshita, M. Delcroix, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer (IEEE, Calgary, AB, Canada, 2018), pp. 5694–5698. https://doi.org/10.1109/ICASSP.2018.8462458
19. T. Gburrek, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, R. Haeb-Umbach, A meeting transcription system

for an ad-hoc acoustic sensor network. arXiv preprint arXiv:2205.00944 (2022)

20. D.R. Hunter, K. Lange, A tutorial on MM algorithms. Am. Stat. **58**(1), 30–37 (2004)
21. T. Kawamura, N. Ono, R. Scheibler, Y. Wakabayashi, R. Miyazaki, in *Proc. European Signal Processing Conference (EUSIPCO)*. Acoustic object canceller using blind compensation for sampling frequency mismatch (IEEE, Amsterdam, Netherlands, 2021), pp. 880–884. https://doi.org/10.23919/Eusipco47968.2020.9287658
22. C. Breithaupt, R. Martin, Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions. IEEE Trans. Audio Speech Lang. Process. **19**(2), 277–289 (2010)
23. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs (IEEE, Salt Lake City, UT, USA, 2001), pp. 749–752. https://doi.org/10.1109/ICASSP.2001.941023
24. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A short-time objective intelligibility measure for time-frequency weighted noisy speech (IEEE, Dallas, TX, USA, 2010), pp. 4214–4217. https://doi.org/10.1109/ICASSP.2010.5495701
25. B.J. Cho, H.M. Park, Stereo acoustic echo cancellation based on maximum likelihood estimation with inter-channel-correlated echo compensation. IEEE Trans. Signal Process. **68**, 5188–5203 (2020)
26. P. Smaragdis, Blind separation of convolved mixtures in the frequency domain. Neurocomputing **22**(1–3), 21–34 (1998)
27. T. Kim, H.T. Attias, S.Y. Lee, T.W. Lee, Blind source separation exploiting higher-order frequency dependencies. IEEE Trans. Audio Speech Lang. Process. **15**(1), 70–79 (2006)
28. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(9), 1626–1641 (2016)
29. N. Ono, S. Miyabe, in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Auxiliary-function-based independent component analysis for super-Gaussian sources (Springer, Berlin, Heidelberg, 2010), pp. 165–172. https://doi.org/10.1007/978-3-642-15995-4_21
30. R. Scheibler, E. Bezzam, I. Dokmanić, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Pyroomacoustics: A python package for audio room simulation and array processing algorithms (IEEE, Calgary, AB, Canada, 2018), pp. 351–355. https://doi.org/10.1109/ICASSP.2018.8461310
31. K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. J Acoust. Soc. Jpn (E) **20**(3), 199–206 (1999)
32. A. Mesaros, T. Heittola, T. Virtanen, in *Proc. European Signal Processing Conference (EUSIPCO)*. TUT database for acoustic scene classification and sound event detection (IEEE, Budapest, Hungary, 2016), pp. 1128–1132. https://doi.org/10.1109/EUSIPCO.2016.7760424
33. R. Martin, Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Process. **13**(5), 845–856 (2005)
34. A. Liutkus, F.R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, J. Fontecave, The 2016 signal separation evaluation campaign, in *Proc. Latent Variable Analysis and Signal Separation (LVA/ICA)*. (Springer International Publishing, Cham, 2017), pp.323–332
35. *JR Higashinihon Yamanote-sen Shanai & Eki Home Jido Hoso Kanzen Original Ongenshu (in Japanese)* (Teichiku Entertainment, Shibuya, 2006). ISBN: 4988004100840

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.