Blind suppression of nonstationary diffuse noise based on spatial covariance matrix decomposition

Nobutaka Ito, Emmanuel Vincent, Tomohiro Nakatani, Nobutaka Ono, Shoko Araki, Shigeki Sagayama

Noname manuscript No.
(will be inserted by the editor)

Blind suppression of nonstationary diffuse noise based on spatial covariance matrix decomposition

Nobutaka Ito · Emmanuel Vincent · Tomohiro Nakatani · Nobutaka Ono · Shoko Araki · Shigeki Sagayama

Received: date / Accepted: date

Abstract We propose methods for blind suppression of nonstationary diffuse noise based on decomposition of the observed spatial covariance matrix into signal and noise parts. In modeling noise to regularize the ill-posed decomposition problem, we exploit spatial invariance (isotropy) instead of temporal invariance (stationarity). The isotropy assumption is that the spatial cross-spectrum of noise is dependent on the distance between microphones and independent of the direction between them. We propose methods for spatial covariance matrix decomposition based on least squares and maximum likelihood estimation. The methods are validated on real-world recordings.

1 Introduction

Noise suppression is the task of estimating a desired signal from its noisy observations by microphones. The difficulty of the noise suppression problem depends on the temporal and the spatial properties of noise. We can distinguish four categories of noise based on two independent axes: stationary or nonstationary noise, and point-source or diffuse noise. Here, diffuse noise is defined as noise from numerous directions caused by many point sources (*e.g.*, many interfering speakers) or by a continuous source (*e.g.*, the vibrating body of a vehicle).

Most of the conventional approaches to noise suppression have assumed either stationary or point-source noise. The spectral approach [1–3] can suppress stationary noise. Under the assumption of stationary noise, the signal power spectrum and/or the noise power spectrum are estimated and used to design spectral filters. On the other hand, the spatial approach [4,5] can suppress point-source noise by controlling the directivity, but cannot suppress diffuse noise sufficiently. This is because the number of spatial nulls is limited to the number of microphones minus one. These conventional approaches do not cover the remaining category: nonstationary, diffuse noise, which is omnipresent in the real world. Indeed, such noise is encountered in many environments, such as stations, airports, vehicles, factories, cafeterias, bars, streets, *etc.* This has significantly limited the application area of noise suppression techniques to the real world.

In principle, even nonstationary diffuse noise can be suppressed with a spectral filter, provided that the signal is sufficiently sparse in the time-frequency domain. Indeed, if we attenuate energy at the time-frequency points where the signal is inactive, the signal-to-noise ratio is expected to improve significantly. The question is how we can design such a filter without assuming noise stationarity. Since this is a highly ill-posed inverse problem, we need to restrict the search space by modeling the signal and noise appropriately.

To this end, we model noise based on spatial invariance (isotropy) instead of temporal invariance (stationarity). The isotropy assumption is that the spatial cross-spectrum of noise is dependent on the distance between microphones and independent of the direction between them. We propose methods for spatial covariance

N. Ito, S. Araki, and T. Nakatani
NTT Communication Science Laboratories
E-mail: ito.nobutaka@lab.ntt.co.jp

E. Vincent
Inria

N. Ono and S. Sagayama
National Institute of Informatics / The Graduate University for Advanced Studies

matrix decomposition based on least squares (LS) and maximum likelihood (ML) estimation. The estimated spatial covariance matrices of the signal and noise are used to design a time-varying multichannel Wiener filter for diffuse noise suppression. We discuss several variants of the isotropic noise model that correspond to additional assumptions about the noise properties or the array geometry. We also propose a general linear algebraic framework for treating these models in a unified manner. This allows us to derive general algorithms each to all noise models, instead of different algorithms for each noise model.

Both the LS and the ML methods have pros and cons. First, as shown later, ML is equivalent to the minimization of the Itakura-Saito divergence, which has proven to be effective for audio signal processing [6]. Second, based on a probabilistic generative model, the ML method can be combined 1) with other signal processing techniques based on a generative model [7–9] to deal with more general environments, and 2) with prior distributions to improve performance using our prior knowledge. Third, in terms of the computational time, the ML method is more expensive than the LS method, because the former employs matrix inversion at each iteration.

The rest of this paper is structured as follows. In Section 2, we describe our observation model and the multichannel Wiener filter for diffuse noise suppression. In Section 3, we describe our framework for modeling the signal and the diffuse noise. Section 4 proposes methods for spatial covariance matrix decomposition based on least squares estimation. Section 5 proposes a method for spatial covariance matrix decomposition based on maximum likelihood estimation. We evaluate the proposed methods experimentally in Section 6, and conclude in Section 7.

2 Background

In this section, we first describe our observation model. We then derive the time-varying multichannel Wiener filter for diffuse noise suppression, and point out that the design of the filter boils down to the estimation of the spatial covariance matrices of the signal and diffuse noise.

2.1 Observation model

In this paper, we focus on blind enhancement of a single desired signal in the presence of diffuse noise. We assume that the source location is stationary. Extension to the case of multiple and/or moving sources is part of future work.

In this paper, unless we note otherwise, we represent signals in the time-frequency domain (*e.g.*, in the short-time Fourier transform domain). We denote the number of frames by T, and the frame index by $t \in \{1, \dots, T\}$. We omit the frequency bin index for brevity. This should not cause confusion, since each frequency bin is processed independently in this paper. The observed signal can be modeled by

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{v}_t, \tag{1}$$

where the variables are defined as follows:

- $\boldsymbol{y}_t \in \mathbb{C}^M$: M-channel observed signal,
- $\boldsymbol{x}_t \in \mathbb{C}^M$: M-channel target signal,
- $\boldsymbol{v}_t \in \mathbb{C}^M$: M-channel diffuse noise.

The problem of diffuse noise suppression considered in this paper is formally defined as the estimation of $\mathcal{X} \triangleq \{\boldsymbol{x}_t\}_{1\le t\le T}$ from $\mathcal{Y} \triangleq \{\boldsymbol{y}_t\}_{1\le t\le T}$. Here, the notation $\{\boldsymbol{y}_t\}_{1\le t\le T}$ stands for $\{\boldsymbol{y}_t | 1 \le t \le T\}$, for example.

For simplicity, we make the following assumptions on $\boldsymbol{x}_t$ and $\boldsymbol{v}_t$:

- Temporal independence: $\{\boldsymbol{x}_t\}_{1\le t\le T}$ is an independent series. That is, for different t, u, $\boldsymbol{x}_t$ and $\boldsymbol{x}_u$ are independent. $\{\boldsymbol{v}_t\}_{1\le t\le T}$ is also an independent series.
- Mutual independence: $\{\boldsymbol{x}_t\}_{1\le t\le T}$ and $\{\boldsymbol{v}_t\}_{1\le t\le T}$ are mutually independent. That is, for any t, u, $\boldsymbol{x}_t$ and $\boldsymbol{v}_u$ are independent.
- Gaussianity: $\boldsymbol{x}_t$ and $\boldsymbol{v}_t$ are zero-mean complex-valued Gaussian variables with covariance matrices $\boldsymbol{\Phi}_t^{\boldsymbol{x}} \triangleq \mathcal{E}[\boldsymbol{x}_t \boldsymbol{x}_t^{\mathsf{H}}]$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \triangleq \mathcal{E}[\boldsymbol{v}_t \boldsymbol{v}_t^{\mathsf{H}}]$.

Here, the probability density function of the complex-valued Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{C}^M$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{C}^{M\times M}$ is given by

$$\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{\pi^M \det \boldsymbol{\Sigma}} \exp\left[-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]. \tag{2}$$

Modeling temporal correlation due to reverberation and more complex distributions are part of future work.

2.2 Time-variant multichannel Wiener filter for diffuse noise suppression

Here we review the time-varying multichannel Wiener filter for diffuse noise suppression. The filter is a time-varying spatiotemporal filter that is optimal in the sense of linear minimum mean square error (LMMSE). Designed properly, it can suppress diffuse noise effectively,

by attenuating energy at the time-frequency points at which the signal is inactive.

The LMMSE estimator is defined as the linear estimator of the form

$$\hat{\boldsymbol{x}}_t = \boldsymbol{W}_t^{\mathsf{H}} \boldsymbol{y}_t \tag{3}$$

that minimizes the mean square error

$$\mathcal{E}[\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|^2]. \tag{4}$$

In this paper, we denote the Frobenius/Euclidean norm of a matrix/vector by $\|\cdot\|$, and the Hermitian transpose by the superscript $^{\mathsf{H}}$. Partial differentiation of (4) w.r.t. $\boldsymbol{W}_t^*$ reveals that the optimum estimator is given by [7, 10, 11]

$$\hat{\boldsymbol{x}}_t = \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{y}_t, \tag{5}$$

where * denotes complex conjugation.

The time-varying multichannel Wiener filter (5) is also optimal in the senses of maximum *a posteriori* (MAP) and minimum mean square error (MMSE), when the signal and noise are modeled as zero-mean Gaussian random variables (see (37)).

2.3 Spatial covariance matrix decomposition

To design the time-varying multichannel Wiener filter (5), we need to know the spatial covariance matrices of the signal and the noise, $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$. However, we are not given these matrices, but only the spatial covariance matrix of the observed noisy data, $\boldsymbol{\Phi}_t^{\boldsymbol{y}} \triangleq \mathcal{E}[\boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{H}}]$. Therefore, to suppress diffuse noise effectively, it is crucial to estimate $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ accurately from $\boldsymbol{\Phi}_t^{\boldsymbol{y}}$. Under the assumption that the signal and the noise are mutually uncorrelated, these matrices are related as

$$\boldsymbol{\Phi}_t^{\boldsymbol{y}} = \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}. \tag{6}$$

Hence, we call the estimation of $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ *spatial covariance matrix decomposition*, which is the main focus of the paper.

Spatial covariance matrix decomposition is a highly ill-posed inverse problem. Indeed, we need to estimate both $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ given $\boldsymbol{\Phi}_t^{\boldsymbol{y}}$ only. Without any constraints on $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$, there would be infinitely many decompositions. Therefore, to obtain a reliable decomposition, it is important to restrict the search space by modeling $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ appropriately.

Regarding the target signal, the assumption that it is emitted from a stationary (*i.e.*, not moving) point source implies that its spatial characteristics do not change over time. Therefore, we can consider the following model of $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ [7]:

$$\boldsymbol{\Phi}_t^{\boldsymbol{x}} = \phi_t^x \boldsymbol{B}^{\boldsymbol{x}}. \tag{7}$$

Here, $\phi_t^x \in \mathbb{R}$ is a time-varying parameter representing the power spectrum, and $\boldsymbol{B}^{\boldsymbol{x}} \in \mathbb{C}^{M \times M}$ is a constant parameter called the *coherence matrix* representing the spatial characteristics. Especially, for low reverberation, $\boldsymbol{B}^{\boldsymbol{x}}$ can be approximated as a rank-one matrix

$$\boldsymbol{B}^{\boldsymbol{x}} = \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}, \tag{8}$$

where $\boldsymbol{h}$ is called the steering vector. More generally, we can model the eigenvalues of $\boldsymbol{B}^{\boldsymbol{x}}$ to be sparse.

3 Matrix linear subspace for unified treatment of diffuse noise models

In contrast to the target signal, diffuse noise originates from many sources or from a continuous source, and therefore its spatial characteristics vary over time. For the case of many sources, for example, even if they are not moving, the spatial characteristics of the whole noise vary over time, since different sources are active at different time-frequency points. Furthermore, diffuse noise is also spectrally nonstationary in general (*e.g.*, consider many concurrent speakers at a cocktail party). To this issue, in [12, 13], we have proposed diffuse noise modeling based on isotropy. The isotropy assumption is that the spatial cross-spectrum of noise is dependent on the distance between microphones and independent of the direction between them. Under this assumption, we can show that $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ belongs to a low-dimensional subspace $\mathcal{V}$ of the vector space $\mathcal{H}$ over $\mathbb{R}$ spanned by $M \times M$ Hermitian matrices [12, 13]. We call $\mathcal{V}$ *a matrix linear subspace* because it is a subspace of the space of matrices, $\mathcal{H}$.

There are several choices of $\mathcal{V}$ depending on additional assumptions regarding the array geometry or the noise field. Firstly, the most general choice with no additional assumptions is the *real-valued noise covariance model*. The isotropy assumption implies that $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ is symmetric, and covariance matrices are Hermitian by definition. Therefore, it follows that $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ is real-valued:

$$\mathcal{V}_{\text{real}} = \left\{ \boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{H} \middle| \boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathbb{R}^{M \times M} \right\}. \tag{9}$$

Secondly, for spatial class of array geometries called *crystal arrays*, the *blind noise decorrelation (BND) model* can be applied [12, 14]:

$$\mathcal{V}_{\text{BND}} = \left\{ \boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{H} \middle| \boldsymbol{U}^{\mathsf{H}} \boldsymbol{\Phi}_t^{\boldsymbol{v}} \boldsymbol{U} \text{ is diagonal} \right\}. \tag{10}$$

Here, $\boldsymbol{U}$ is a known constant unitary matrix. Thirdly, if we assume that diffuse noise is uncorrelated between

two points in the limit of a large distance between them, the following *spatially uncorrelated noise model* holds for large arrays [15]:

$$\mathcal{V}_{\mathrm{uncor}} = \left\{ \boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{H} \middle| \boldsymbol{\Phi}_t^{\boldsymbol{v}} \text{ is diagonal} \right\}. \tag{11}$$

Finally, assuming that the microphone coordinates are known and that noise planewaves with the same power spectrum arrive from all directions, the following *fixed noise coherence model* holds [16]:

$$\mathcal{V}_{\mathrm{coh}} = \left\{ \phi_t^v \boldsymbol{B}^{\boldsymbol{v}} \middle| \phi_t^v \in \mathbb{R} \right\}. \tag{12}$$

Here, $\boldsymbol{B}^{\boldsymbol{v}}$ is called a noise coherence matrix, and its (m,n)th entry is given by

$$b_{mn}^v = \mathrm{sinc}\left(\frac{2\pi f L_{mn}}{c} \right). \tag{13}$$

Here, $\mathrm{sinc}(\cdot)$ denotes the sine cardinal function, L_{mn} the distance between the mth and the nth microphones, and c the sound velocity.

Each of these noise models was applied to *non-blind* diffuse noise suppression with a known target steering vector [14–17], and to noise-robust multiple source localization [18]. In [14], we proposed the first method for *blind* diffuse noise suppression, which was limited to the blind noise decorrelation model. In this paper, we propose general blind diffuse noise suppression methods applicable to all the above noise models.

4 Spatial covariance matrix decomposition based on least squares estimation

4.1 Cost function: square error

In this section, we propose two methods for spatial covariance matrix decomposition based on the minimization of the following Euclidean error subject to $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{V}$:

$$J_{\mathrm{LS}}(\Theta) \triangleq \sum_t \|\boldsymbol{\Phi}_t^{\boldsymbol{y}} - (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})\|^2 \tag{14}$$

$$= \sum_t \|\boldsymbol{\Phi}_t^{\boldsymbol{y}} - (\phi_t^x \boldsymbol{B}^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})\|^2. \tag{15}$$

Θ denotes the parameter set given by

$$\Theta \triangleq \{\{\boldsymbol{\Phi}_t^{\boldsymbol{x}}\}_{1\le t\le T}, \{\boldsymbol{\Phi}_t^{\boldsymbol{v}}\}_{1\le t\le T}\} \tag{16}$$

$$= \{\{\phi_t^x\}_{1\le t\le T}, \boldsymbol{B}^{\boldsymbol{x}}, \{\boldsymbol{\Phi}_t^{\boldsymbol{v}}\}_{1\le t\le T}\}. \tag{17}$$

The minimizers of (15) are not unique, and so an additional constraint is necessary. To see this, note that (15) is decomposed into $\mathcal{V}$ and $\mathcal{V}^\perp$ components as follows:

$$J_{\mathrm{LS}}(\Theta) = \sum_{t=1}^T \|\mathcal{P}[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}[\boldsymbol{B}^{\boldsymbol{x}}] - \boldsymbol{\Phi}_t^{\boldsymbol{v}}\|^2 + \sum_{t=1}^T \|\mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]\|^2. \tag{18}$$

Here, $\mathcal{V}^\perp$ denotes the orthogonal complement of $\mathcal{V}$, and $\mathcal{P}$ and $\mathcal{P}^\perp$ the orthogonal projection operators onto $\mathcal{V}$ and $\mathcal{V}^\perp$, respectively. We can eliminate $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ from (18) by replacing it with the maximizer $\boldsymbol{\Phi}_t^{\boldsymbol{v}} = \mathcal{P}[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}[\boldsymbol{B}^{\boldsymbol{x}}]$ of (18). Therefore, the minimization of $J_{\mathrm{LS}}(\Theta)$ w.r.t. Θ is equivalent to the minimization of

$$\sum_{t=1}^T \|\mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]\|^2 \tag{19}$$

w.r.t. $\{\phi_t^x\}_{1\le t\le T}$ and $\boldsymbol{B}^{\boldsymbol{x}}$. Since (19) does not depend on $\mathcal{P}[\boldsymbol{B}^{\boldsymbol{x}}]$, it has infinitely many optimal solutions. This indeterminacy is resolved by exploiting the sparseness of the eigenvalues of $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ [12]. In Section 4.2, we propose first estimating the $\mathcal{V}^\perp$-component, and then completing the missing $\mathcal{V}$-component through matrix completion techniques. In Section 4.3, on the other hand, we propose minimizing $J_{\mathrm{LS}}(\Theta)$ directly subject to a rank-one constraint on $\boldsymbol{B}^{\boldsymbol{x}}$.

4.2 Optimization algorithm based on low-rank matrix completion

Based on the observation in Section 4.1, the first algorithm for least squares estimation follows the following procedure:

1. estimate $\boldsymbol{Z} \triangleq \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]$ and $\{\phi_t^x\}_{1\le t\le T}$ by minimizing $\sum_{t=1}^T \|\mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]\|^2$,
2. estimate $\boldsymbol{B}^{\boldsymbol{x}}$ through low-rank matrix completion [19] of $\boldsymbol{Z}$,
3. reestimate $\{\phi_t^x\}_{1\le t\le T}$ using $\boldsymbol{B}^{\boldsymbol{x}}$.

The algorithm eliminates $\{\boldsymbol{\Phi}_t^{\boldsymbol{v}}\}_{1\le t\le T}$ as a nuisance parameters, and it estimates the rest. We denote the parameter set in this method by $\Omega \triangleq \{\{\phi_t^x\}_{1\le t\le T}, \boldsymbol{h}\}$.

In the first step, if we eliminate the nuisance parameter $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ by replacing it with its optimal value

$$\boldsymbol{\Phi}_t^{\boldsymbol{v}} \leftarrow \mathcal{P}[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \mathcal{P}[\boldsymbol{\Phi}_t^{\boldsymbol{x}}], \tag{20}$$

the cost function reduces to

$$\sum_{t=1}^T \|\mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]\|^2 = \sum_{t=1}^T \|\mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \boldsymbol{Z}\|^2. \tag{21}$$

This can be minimized w.r.t. ϕ_t^x and $\boldsymbol{Z}$ by alternately applying the following update rules

$$\phi_t^x \leftarrow \frac{\langle \mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}], \boldsymbol{Z}\rangle}{\|\boldsymbol{Z}\|^2}, \tag{22}$$

$$\boldsymbol{Z} \leftarrow \frac{\sum_t \phi_t^x \mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}]}{\sum_t (\phi_t^x)^2}, \tag{23}$$

which is based on coordinate descent. Prior to the iterations, $\boldsymbol{Z}$ can be initialized roughly, *e.g.*, by eigenvalue truncation of $\boldsymbol{\Phi}_t^{\boldsymbol{y}}$.

In the second step, we estimate $\boldsymbol{B}^{\boldsymbol{x}}$ using $\boldsymbol{Z} = \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]$ obtained in the first step, by low-rank matrix completion techniques [12,19,20]. The low-rank matrix completion techniques are originally techniques for completing missing entries of a low-rank matrix, and we have extended them to the completion of a missing subspace [12]. Here, we apply this technique to the completion of the missing matrix subspace $\mathcal{V}$. As an example, we consider here the following optimization problem under a low-rank constraint:

$$\min_{\boldsymbol{B}^{\boldsymbol{x}}} \|\mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}] - \boldsymbol{Z}\|^2 \tag{24}$$

$$\text{s.t. } \boldsymbol{B}^{\boldsymbol{x}}: \text{ Hermitian positive semidefinite, } \mathrm{rank}(\boldsymbol{B}^{\boldsymbol{x}}) \le R.$$

That is, we seek for the matrix $\boldsymbol{B}^{\boldsymbol{x}}$ with a rank no larger than R (a predetermined upper bound of the rank), for which $\mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]$ is closest to $\boldsymbol{Z}$ obtained in the first step. Other techniques for low-rank matrix completion such as trace norm minimization [12,20] can also be applied.

The criterion (24) can be optimized by alternately iterating the following update rules [12,19]:

1. $\boldsymbol{Y} \leftarrow \mathcal{P}[\boldsymbol{B}^{\boldsymbol{x}}] + \boldsymbol{Z}$
2. $(\lambda_r, \boldsymbol{\mu}_r) \leftarrow$the rth largest eigenvalue and the corresponding eigenvector of $\boldsymbol{Y}$
3. $\boldsymbol{B}^{\boldsymbol{x}} \leftarrow \sum_{r=1}^R \max\{\lambda_r, 0\}\boldsymbol{\mu}_r\boldsymbol{\mu}_r^{\mathsf{H}}$

It is guaranteed that the above procedure decreases the cost function in (24) monotonically. $\boldsymbol{B}^{\boldsymbol{x}}$ can be initialized by

$$\boldsymbol{B}^{\boldsymbol{x}} \leftarrow \frac{\sum_t \phi_t^x \boldsymbol{\Phi}_t^{\boldsymbol{y}}}{\sum_t (\phi_t^x)^2}, \tag{25}$$

which can be computed using ϕ_t^x estimated in the first step. (25) is derived by the minimization of $\sum_t \|\boldsymbol{\Phi}_t^{\boldsymbol{y}} - \phi_t^x \boldsymbol{B}^{\boldsymbol{x}}\|^2$ obtained by neglecting the noise contribution in the cost function (15).

In the final step, $\boldsymbol{Z} = \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]$ is updated using $\boldsymbol{B}^{\boldsymbol{x}}$ estimated in the second step, and ϕ_t^x is reestimated by (22).

The algorithm is summarized as follows, with `iter_num` denoting the number of iterations:

Algorithm 1

Initialize $\boldsymbol{Z}$ *by rank-one approximation of* $\boldsymbol{\Phi}_t^{\boldsymbol{y}}$
for *cnt* $= 1$ *to* `iter_num` ***do***
  for $t = 1$ *to* T ***do***
    $\phi_t^x \leftarrow \dfrac{\langle \mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}], \boldsymbol{Z}\rangle}{\|\boldsymbol{Z}\|^2}$
  end for
  $\boldsymbol{Z} \leftarrow \dfrac{\sum_t \phi_t^x \mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}]}{\sum_t (\phi_t^x)^2}$
end for
Initialize $\boldsymbol{B}^{\boldsymbol{x}}$ *by* $\boldsymbol{B}^{\boldsymbol{x}} \leftarrow \dfrac{\sum_t \phi_t^x \boldsymbol{\Phi}_t^{\boldsymbol{y}}}{\sum_t (\phi_t^x)^2}$
for *cnt* $= 1$ *to* `iter_num` ***do***
  $\boldsymbol{Y} \leftarrow \mathcal{P}[\boldsymbol{B}^{\boldsymbol{x}}] + \boldsymbol{Z}$
  $(\lambda_r, \boldsymbol{\mu}_r) \leftarrow$*the* r*th largest eigenvalue and the corresponding eigenvector of* $\boldsymbol{Y}$
  $\boldsymbol{B}^{\boldsymbol{x}} \leftarrow \sum_{r=1}^R \max\{\lambda_r, 0\}\boldsymbol{\mu}_r\boldsymbol{\mu}_r^{\mathsf{H}}$
end for
for $t = 1$ *to* T ***do***
  $\phi_t^x \leftarrow \dfrac{\langle \mathcal{P}^\perp[\boldsymbol{\Phi}_t^{\boldsymbol{y}}], \mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]\rangle}{\|\mathcal{P}^\perp[\boldsymbol{B}^{\boldsymbol{x}}]\|^2}$
  $\phi_t^x \leftarrow \max\{\phi_t^x, 0\}$
end for

4.3 Optimization algorithm based on a rank-one constraint

The second method minimizes the cost function (15), subject to the rank-one constraint

$$\boldsymbol{B}^{\boldsymbol{x}} = \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}. \tag{26}$$

Furthermore, we pose a unit norm constraint $\|\boldsymbol{h}\| = 1$ on $\boldsymbol{h}$, which simplifies the cost function and leads to a closed-form update rule of $\boldsymbol{h}$. Based on coordinate descent, we can can derive the following algorithm, which alternately minimizes (15) w.r.t. $\{\phi_t^x\}_{1\le t\le T}$, $\boldsymbol{h}$, and $\{\boldsymbol{\Phi}_t^{\boldsymbol{v}}\}_{1\le t\le T}$. Our preliminary experiments have shown that the algorithm in Section 4.2 gives good initial values of $\{\phi_t^x\}_{1\le t\le T}$ and $\boldsymbol{h}$.

Algorithm 2

Initialize $\{\phi_t^x\}_{1\le t\le T}$ *and* $\boldsymbol{h}$ *by Algorithm 1*
for *cnt* $= 1$ *to* `iter_num` ***do***
  for $t = 1$ *to* T ***do***
    $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \leftarrow \mathcal{P}[\boldsymbol{\Phi}_t^{\boldsymbol{y}}] - \phi_t^x \mathcal{P}[\boldsymbol{h}\boldsymbol{h}^H]$
    $\phi_t^x \leftarrow \max\{\boldsymbol{h}^H(\boldsymbol{\Phi}_t^{\boldsymbol{y}} - \boldsymbol{\Phi}_t^{\boldsymbol{v}})\boldsymbol{h}, 0\}$
  end for
  $\boldsymbol{h} \leftarrow$ *unit principal eigenvector of* $\sum_t \phi_t^x(\boldsymbol{\Phi}_t^{\boldsymbol{y}} - \boldsymbol{\Phi}_t^{\boldsymbol{v}})$
end for
for $t = 1$ *to* T ***do***
  $\phi_t^x \leftarrow |h_1|^2 \phi_t^x$
end for
$\boldsymbol{h} \leftarrow \boldsymbol{h}/h_1$

5 Spatial covariance matrix decomposition based on maximum likelihood

In this section, we propose a method for spatial covariance matrix decomposition based on maximum likelihood estimation.

5.1 Probabilistic generative model of the observed signals

Based on the assumptions in Section 2.1, we can derive a probabilistic generative model of the oberved signals. The joint distribution $p(\boldsymbol{x}_t, \boldsymbol{y}_t; \Theta)$ of the observed signals and the target signal is calculated as follows:

$$\begin{aligned}
&\log p(\boldsymbol{x}_t, \boldsymbol{y}_t; \Theta) \\
&= \log p(\boldsymbol{x}_t; \Theta) + \log p(\boldsymbol{y}_t|\boldsymbol{x}_t; \Theta) && (27)\\
&= -2M \log \pi - \log\det \boldsymbol{\Phi}_t^{\boldsymbol{x}} - \boldsymbol{x}_t^{\mathsf{H}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}})^{-1} \boldsymbol{x}_t && (28)\\
&\quad - \log\det \boldsymbol{\Phi}_t^{\boldsymbol{v}} - (\boldsymbol{y}_t - \boldsymbol{x}_t)^{\mathsf{H}} (\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} (\boldsymbol{y}_t - \boldsymbol{x}_t).
\end{aligned}$$

Marginalizing this w.r.t. $\boldsymbol{x}_t$, we have the following marginal distribution of the observed signals $\boldsymbol{y}_t$:

$$p(\boldsymbol{y}_t; \Theta) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_t; 0, \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}). \tag{29}$$

5.2 Objective function: likelihood

We estimate the parameters by maximizing the following log-likelihood function subject to $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{V}$:

$$\begin{aligned}
J_{\mathrm{ML}}(\Theta) &= \sum_{t=1}^{T} \log p(\boldsymbol{y}_t; \Theta) && (30)\\
&= \sum_{t=1}^{T} \log \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_t; 0, \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}). && (31)
\end{aligned}$$

Here, Θ is the parameter set defined by

$$\Theta \triangleq \big\{ \{\phi_t^x\}_{1\le t\le T}, \boldsymbol{B}^{\boldsymbol{x}}, \{\boldsymbol{\Phi}_t^{\boldsymbol{v}}\}_{1\le t\le T} \big\}. \tag{32}$$

The maximization of the likelihood can also be viewed as the minimization of a matrix Itakura-Saito divergence [21,22]. Indeed, $J_{\mathrm{ML}}(\Theta)$ can be rewritten as

$$J_{\mathrm{ML}}(\Theta) = -\sum_{t=1}^{T} D_{\mathrm{IS}}\big(\boldsymbol{\Phi}_t^{\boldsymbol{y}}; \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}\big) + \text{const.}, \tag{33}$$

where D_{IS} denotes the following matrix Itakura-Saito divergence:

$$\begin{aligned}
&D_{\mathrm{IS}}\big(\boldsymbol{\Phi}_t^{\boldsymbol{y}}; \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}\big) \triangleq \mathrm{Tr}\big\{\boldsymbol{\Phi}_t^{\boldsymbol{y}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\big\} && (34)\\
&- \log\det\big\{\boldsymbol{\Phi}_t^{\boldsymbol{y}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\big\} - M.
\end{aligned}$$

Since $D_{\mathrm{IS}}\big(k\boldsymbol{A}, k\boldsymbol{B}\big) = D_{\mathrm{IS}}\big(\boldsymbol{A}, \boldsymbol{B}\big)$ holds for any positive scalar k, we see that D_{IS} is suitable for audio signals, which have logarithmic characteristics. The divergence is a generalization of the Itakura-Saito divergence [6], which has proven to be effective for speech processing.

5.3 Optimization algorithm based on expectation-maximization

Since $J(\Theta)$ contains the term $\log\det(\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})$ and $(\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}$, it is difficult to obtain a close-form expression of the optimal solution. On the other hand, regarding $\{\boldsymbol{x}_t\}_{1\le t\le T}$ as hidden variable, and based on the expectation-maximization (EM) algorithm [26], we can derive an efficient optimization algorithm. The EM algorithm iterates the following E-step and M-step alternately, and it is guaranteed to converge to a local optimum.

- E-step: Update the posterior distribution $p(\boldsymbol{x}_t|\boldsymbol{y}_t; \Theta')$ of $\boldsymbol{x}_t$, using the current estimate of Θ'.
- M-step: Update Θ by maximizing the Q-function

$$Q(\Theta; \Theta') = \sum_{t=1}^{T} \big\langle \log p(\boldsymbol{x}_t, \boldsymbol{y}_t; \Theta) \big\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t; \Theta')}. \tag{35}$$

  Here, $\langle\cdot\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')}$ denotes the expectation w.r.t. to the current posterior probability $p(\boldsymbol{x}_t|\boldsymbol{y}_t; \Theta')$ of the hidden variable $\boldsymbol{x}_t$.

In the E-step, we update the posterior distribution, which can be derived as follows. From (28) and the Bayes rule,

$$\begin{aligned}
&\log p(\boldsymbol{x}_t|\boldsymbol{y}_t; \Theta) \overset{\mathrm{c}}{=} \log p(\boldsymbol{x}_t, \boldsymbol{y}_t; \Theta) && (36)\\
&\overset{\mathrm{c}}{=} -\Big(\boldsymbol{x}_t - \boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\Big)^{\mathsf{H}} \Big(\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\Big)^{-1} \Big(\boldsymbol{x}_t - \boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\Big). && (37)
\end{aligned}$$

Here,

$$\begin{aligned}
\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}} &\triangleq \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{y}_t, && (38)\\
\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}} &\triangleq \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{\Phi}_t^{\boldsymbol{v}}, && (39)
\end{aligned}$$

and $\overset{\mathrm{c}}{=}$ means the equality up to a difference of a constant independent of $\boldsymbol{x}_t$. Therefore, the posterior probability is given by

$$p(\boldsymbol{x}_t|\boldsymbol{y}_t; \Theta) = \mathcal{N}_{\mathbb{C}}\big(\boldsymbol{x}_t; \boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}, \boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\big). \tag{40}$$

Because the Gaussian distribution (40) is completely determined by its mean $\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}$ and covariance matrix $\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}$, in the E-step it suffices to update them.

In the M-step, we update the parameters so that the Q-function is maximized. The explicit form of Q-function is given by the following expression:

$$\begin{aligned}
Q(\Theta; \Theta') &= -\sum_{t=1}^{T} \log\det \boldsymbol{\Phi}_t^{\boldsymbol{x}} && (41)\\
&- \sum_{t=1}^{T} \mathrm{Tr}\Big[(\boldsymbol{\Phi}_t^{\boldsymbol{x}})^{-1} \big\langle \boldsymbol{x}_t \boldsymbol{x}_t^{\mathsf{H}} \big\rangle \Big] - \sum_{t=1}^{T} \log\det \boldsymbol{\Phi}_t^{\boldsymbol{v}} \\
&- \sum_{t=1}^{T} \mathrm{Tr}\Big[(\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \big\langle (\boldsymbol{y}_t - \boldsymbol{x}_t)(\boldsymbol{y}_t - \boldsymbol{x}_t)^{\mathsf{H}} \big\rangle \Big].
\end{aligned}$$

Here, we omitted a constant, and abbreviated $\langle\cdot\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')}$ as $\langle\cdot\rangle$. The expectations $\langle \boldsymbol{x}_t\boldsymbol{x}_t^{\mathsf{H}}\rangle$ and $\langle(\boldsymbol{y}_t-\boldsymbol{x}_t)(\boldsymbol{y}_t-\boldsymbol{x}_t)^{\mathsf{H}}\rangle$ in (41) can be computed using the current mean and covariance matrix $\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'$ and $\big(\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'$ as follows:

$$\begin{aligned}\big\langle \boldsymbol{x}_t\boldsymbol{x}_t^{\mathsf{H}}\big\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')} &= \Big\langle \big\{\boldsymbol{x}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big\} \qquad (42)\\ &\quad\times\big\{\boldsymbol{x}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big\}^{\mathsf{H}}\Big\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')}\\ &= \big(\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'^{\mathsf{H}}, \qquad (43)\end{aligned}$$

$$\begin{aligned}&\big\langle(\boldsymbol{y}_t-\boldsymbol{x}_t)(\boldsymbol{y}_t-\boldsymbol{x}_t)^{\mathsf{H}}\big\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')}\\ &= \Big\langle \big\{\boldsymbol{y}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'-\boldsymbol{x}_t\big\} \qquad (44)\\ &\quad\times\big\{\boldsymbol{y}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'-\boldsymbol{x}_t\big\}^{\mathsf{H}}\Big\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')}\\ &= \big(\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big\{\boldsymbol{y}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big\}\big\{\boldsymbol{y}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big\}^{\mathsf{H}}. \qquad (45)\end{aligned}$$

The update rules for the M-step can be derived by maximizing the Q-function (41) w.r.t. the parameters (see Appendix A). The following summarizes one iteration of the algorithm derived in the above:

Algorithm 3
for $t=1$ *to* T ***do***
$\quad\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}} \leftarrow \boldsymbol{\Phi}_t^{\boldsymbol{x}}(\boldsymbol{\Phi}_t^{\boldsymbol{x}}+\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\boldsymbol{y}_t$
$\quad\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}} \leftarrow \boldsymbol{\Phi}_t^{\boldsymbol{x}}(\boldsymbol{\Phi}_t^{\boldsymbol{x}}+\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\boldsymbol{\Phi}_t^{\boldsymbol{v}}$
$\quad\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}} \leftarrow \boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}+\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)^{H}$
$\quad\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}} \leftarrow \boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}+\big(\boldsymbol{y}_t-\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)\big(\boldsymbol{y}_t-\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)^{H}$
end for
$\boldsymbol{B}^{\boldsymbol{x}} \leftarrow \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\dfrac{1}{\phi_t^x}\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}$
for $t=1$ *to* T ***do***
$\quad\phi_t^x \leftarrow \dfrac{1}{M}Tr\big[(\boldsymbol{B}^{\boldsymbol{x}})^{-1}\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}\big]$
$\quad\boldsymbol{\Phi}_t^{\boldsymbol{x}} \leftarrow \phi_t^x\boldsymbol{B}^{\boldsymbol{x}}$
$\quad\boldsymbol{\Phi}_t^{\boldsymbol{v}} \leftarrow \begin{cases}\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big], & \mathcal{V}=\mathcal{V}_{uncor},\mathcal{V}_{BND},\mathcal{V}_{real}\\ \dfrac{1}{M}Tr\big[(\boldsymbol{B}^{\boldsymbol{v}})^{-1}\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big]\boldsymbol{B}^{\boldsymbol{v}}, & \mathcal{V}=\mathcal{V}_{coh}\end{cases}$
end for

5.4 Comparison between least squares and maximum likelihood estimation

Compared to the least squares estimation, the Itakura-Saito divergence in the maximum likelihood estimation has the advantage of scale invariance, and it is expected to be more suitable for audio signals, which have a logarithmic nature. Furthermore, maximum likelihood estimation is based on a generative model of the observed signal, and therefore it can be integrated with other speech enhancement techniques (*e.g.*, source separation [7,23,9], dereverberation [8]) based on a generative model. This is a great advantage in extending the proposed methods to a versatile method that can be applied to various real-world environments.

Fig. 1 Fabricated 12-element spherical microphone array of diameter 15 cm. The microphones are mounted on a rigid spherical shell.

6 Experimental performance evaluation on real-world data

We experimentally validated the proposed methods on real-world data.

6.1 Experimental conditions

We fabricated a 12-channel spherical microphone array with microphones at the vertices of an icosahedron of diameter 15 cm (see Fig. 1). The microphones were mounted on a rigid spherical shell. With the array, we recorded the signal and noise images in a room at the University of Tokyo (Fig. 2). The signal image was recorded while the loudspeaker played female speech [27], and the noise image while the windows were open. They were mixed to generate the observed signals.

We compared the following four methods for spatial covariance matrix decomposition:

- `conv-LS`: conventional least squares method proposed in [14],
- `prop-LS1`: proposed least squares method with low-rank matrix completion in Section 4.2,
- `prop-LS2`: proposed least squares method with a rank-1 constraint in Section 4.3,

Table 1 Detailed experimental conditions

– D/A board: M-AUDIO Fast track pro (4-channel)
– Loudspeaker: BOSE 101MM
– Loudspeaker amplifier: BOSE 1705II
– Microphones: SONY ECM-C10 (electret-type; omnidirectional)
– A/D board: Tokyo Electron Device TD-BD-16ADUSB (16-channel; with microphone amplifiers)
– Data length: 2 s
– Sampling frequency: 16 kHz
– Frame length: 2048 samples
– Frame shift: 64 samples
– Window: Hamming window
– Number of iterations: `iter_num` = 20

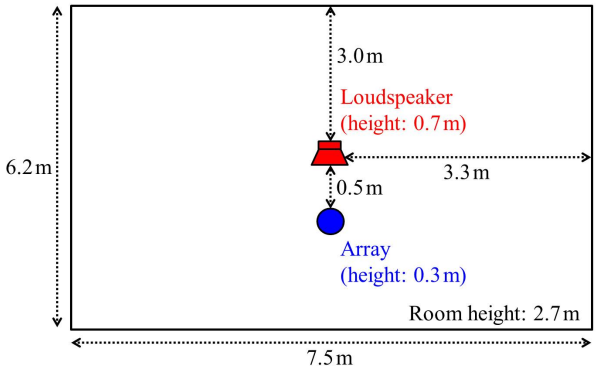

Fig. 2 The room layout in the experiment.

– `prop-ML`: proposed maximum-likelihood method in Section 5.

Each method was combined with the following four noise models (see Section 3):

– `coh`: fixed noise coherence model,
– `uncor`: spatially uncorrelated noise model,
– `BND`: blind noise decorrelation model,
– `real`: real-valued noise covariance model.

The observed signals were analyzed by the short-time Fourier transform (STFT). The lowest 14 frequency bins (below 100 Hz) were discarded, which contained only noise. The observed spatial covariance matrix for the least squares methods was computed locally by averaging $\boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{H}}$ over 48 consecutive frames. The other conditions are summarized in Table 1.

6.2 Experimental results

Fig. 3 shows the spectrograms and the output SNR of the observed, enhanced, and reference signals (noise model: BND). The maximum-likelihood method ("`prop-ML`") has resulted in a larger SNR than the least squares methods ("`prop-LS1`" and "`prop-LS2`"), mainly because of less signal distortion. Furthermore, plotted in the logarithmic scale, the spectrogram of the ML method resembles the reference signal more, which can be explained by the logarithmic nature of the Itakura-Saito divergence. Especially, the LS methods attenuated some frequency components, while the ML method avoided this issue because the logarithmic measure strongly penalizes zeros in the estimated spectra.

Table 2 shows the output SNRs (dB) [14] of the observed and the enhanced signals. We show results obtained with oracle initialization (calculated with the reference signal) to examine the sensitivity of the noise suppression performance to initialization. The maximum-likelihood method with the BND model yielded the highest output SNR of 8.6 dB. Comparing `prop-LS1`, `prop-LS2`, and `prop-ML`, `prop-ML` gave the highest SNR for all noise models except `coh`. For `coh`, the SNR of `prop-LS2` was the highest (5.2 dB), while that of `prop-ML` was −0.9 dB. The algorithm `prop-ML` did not work well for `coh`. This is because the sine cardinal noise coherence matrix approaches a rank-one (hence, singular) matrix at low frequencies, and hence the update of $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$, which includes inversion of the matrix, diverges.

7 Conclusion

In this paper, we have proposed a blind method for suppressing nonstationary diffuse noise. Based on the isotropic noise models, we proposed methods for spatial covariance matrix decomposition. The decomposition algorithms are based on least squares or maximum-likelihood estimation. In the experimental evaluation, the maximum likelihood estimation has resulted in a superior performance to the least squares estimation.

References

1. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
2. R. Martin, "Spectral subtraction based on minimum statistics," in *Proc.EUSIPCO*, 1994, pp. 1982–1185.
3. Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
4. D.E. Dudgeon D.H. Johnson, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ, 1993.
5. M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, Heidelberg, 2001.
6. F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Rep. 6th International Congress on Acoustics*, Aug. 1968, pp. C–17–C–20.

Table 2 Objective evaluation of the proposed methods in terms of the SNR (dB). For the blind initialization, the highest SNR is shown in **boldface**, and the highest SNR within each noise model is shown in *Italic*. The input SNR was −1.0 dB.

method	conv-LS	prop-LS1				prop-LS2				prop-ML			
noise model	BND	coh	uncor	BND	real	coh	uncor	BND	real	coh	uncor	BND	real
blind initialization	7.7	4.7	7.5	4.5	−6.5	*5.2*	7.5	4.9	−6.5	−0.9	*8.4*	**8.6**	*5.8*
oracle initialization	9.2	4.8	7.5	4.5	−6.5	5.2	7.5	4.9	−6.5	−0.3	8.3	11.6	17.6

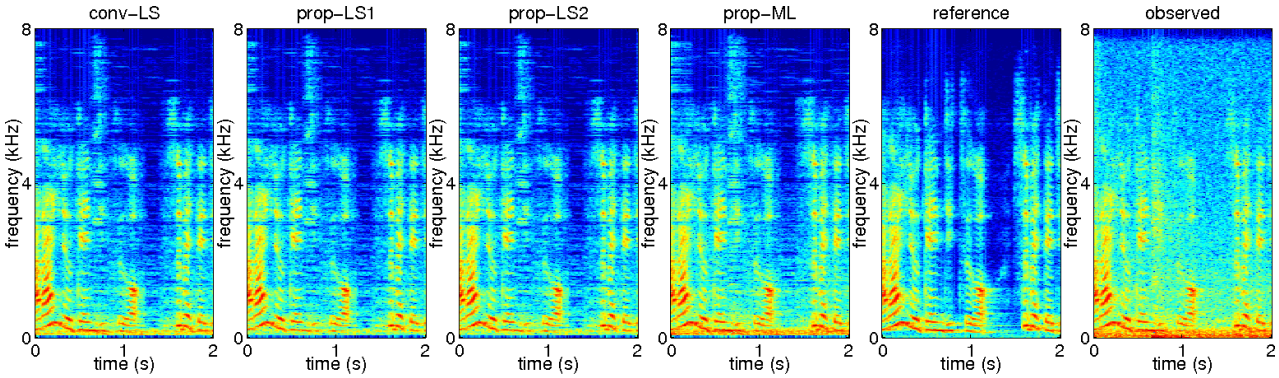

Fig. 3 Spectrograms of the observed, enhanced, and reference signals. The SNRs of these signals were as follows: conv-LS: 7.7 dB; prop-LS1: 4.5 dB; prop-LS2: 4.9 dB; prop-ML: 8.6 dB; observed: −0.2 dB. The plot range of the power has been equalized for all spectrograms

7. N.Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
8. T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.
9. E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Sig. Proc. Mag.*, vol. 31, no. 3, May 2014.
10. K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 39–60. Springer, Berlin Heidelberg, 2001.
11. S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. SP*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.
12. N. Ito, *Robust Microphone Array Signal Processing against Diffuse Noise*, Ph.D. thesis, the University of Tokyo, 2012.
13. N. Ito, E. Vincent, N. Ono, and S. Sagayama, "General algorithms for estimating spectrogram and transfer functions of target signal for blind suppression of diffuse noise," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2013.
14. N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *"IEEE Trans. ASLP"*, vol. 19, no. 7, pp. 2101–2110, Sep. 2011.
15. R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP*, Apr. 1988, pp. 2578–2581.
16. I.A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. SAP*, vol. 11, no. 6, pp. 709–716, Nov. 2003.
17. N. Ito, N. Ono, and S. Sagayama, "Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra," in *Proc. ICASSP*, Mar. 2010, pp. 2818 – 2821.
18. N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Crystal-MUSIC: Accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays," in *Proc. of LVA/ICA, Lecture Notes in Computer Science*, Sep. 2010, vol. 6365, pp. 81–88.
19. N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. International Conference on Machine Learning (ICML)*. AAAI Press, 2003, pp. 720–727.
20. K. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 3, pp. 615–640, Sept. 2010.
21. D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. SP*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.
22. A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, Mar. 2010.
23. H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, May 2013.
24. K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," in *Proc. International Conference on Machine Learning (ICML)*, Jun. 2013, pp. 576–584.
25. B. Kulis, M. Sustik, and I. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *Journal of Machine Learning Research*, vol. 10, pp. 341–376, Feb. 2009.
26. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algo-

rithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.
27. A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communications*, vol. 9, no. 4, pp. 357–363, Aug. 1990.

A Derivation of the update rules in the M-step of maximum likelihood estimation

By setting the partial derivative of the Q-function (41) to zero, we have

$$-UM\frac{1}{\phi_t^x}+\mathrm{Tr}\left[(\boldsymbol{B}^{\boldsymbol{x}})^{-1}\left\langle \boldsymbol{x}_t\boldsymbol{x}_t^{\mathsf{H}}\right\rangle\right]\frac{1}{(\phi_t^x)^2}=0. \tag{46}$$

By solving this w.r.t. ϕ_t^x, we get [7]

$$\phi_t^x=\frac{1}{M}\mathrm{Tr}\big[(\boldsymbol{B}^{\boldsymbol{x}})^{-1}\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}\big]. \tag{47}$$

Here, we defined

$$\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}\triangleq\left\langle \boldsymbol{x}_t\boldsymbol{x}_t^{\mathsf{H}}\right\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')} \tag{48}$$

$$=\big(\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'^{\mathsf{H}}. \tag{49}$$

Next, partial differentiation w.r.t. $\boldsymbol{B}^{\boldsymbol{x}}$ gives

$$-UT(\boldsymbol{B}^{\boldsymbol{x}})^{-1}+(\boldsymbol{B}^{\boldsymbol{x}})^{-1}\left(\sum_{t=1}^{T}\frac{1}{\phi_t^x}\left\langle \boldsymbol{x}_t\boldsymbol{x}_t^{\mathsf{H}}\right\rangle\right)(\boldsymbol{B}^{\boldsymbol{x}})^{-1}=0. \tag{50}$$

Solving this w.r.t. $\boldsymbol{B}^{\boldsymbol{x}}$, we have [7]

$$\boldsymbol{B}^{\boldsymbol{x}}=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\phi_t^x}\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}. \tag{51}$$

The update rule for $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ depends on the explicit form of the matrix subspace $\mathcal{V}$. In the following, we first show that for the class of $\mathcal{V}$ satisfying

$$\boldsymbol{\Phi}_t^{\boldsymbol{v}}\in\mathcal{V}\text{: positive definite}\Rightarrow(\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\in\mathcal{V}, \tag{52}$$

we can derive a unified update rule. Clearly, the subspaces $\mathcal{V}_{\mathrm{uncor}},\mathcal{V}_{\mathrm{BND}},\mathcal{V}_{\mathrm{real}}$ defined in Section 3 belong to the class. We then derive the update rule for $\mathcal{V}_{\mathrm{coh}}$, which does not belong to the class.

When $\mathcal{V}$ satisfies (52), the terms of (41) depending on $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ can be rewritten as

$$-U\log\det\boldsymbol{\Phi}_t^{\boldsymbol{v}}$$
$$-\mathrm{Tr}\left\{(\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\mathcal{P}\left[\left\langle(\boldsymbol{y}_t-\boldsymbol{x}_t)(\boldsymbol{y}_t-\boldsymbol{x}_t)^{\mathsf{H}}\right\rangle\right]\right\}. \tag{53}$$

Here, $\mathcal{P}[\cdot]$ denotes the orthogonal projection onto $\mathcal{V}$ defined using the standard inner product $\langle\boldsymbol{A},\boldsymbol{B}\rangle\triangleq\mathrm{Tr}[\boldsymbol{A}\boldsymbol{B}]$ of $\mathcal{H}$:

$$\mathcal{P}\big[\boldsymbol{A}\big]=\sum_{d=1}^{D}\mathrm{Tr}\big[\boldsymbol{A}\boldsymbol{Q}_d\big]\boldsymbol{Q}_d. \tag{54}$$

Here, $\{\boldsymbol{Q}_d\}_{1\le d\le D}$ is an orthonormal basis of $\mathcal{V}$, and D denotes the dimension of $\mathcal{V}$. The explicit form of $\boldsymbol{Q}_d$ depends on the choice of $\mathcal{V}$, for which the readers are referred to [12,13]. The term in $\mathcal{P}[\cdot]$ in (53) generally has both components parallel and orthogonal to $\mathcal{V}$. However, the latter vanishes owing to $(\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\in\mathcal{V}$, and hence (53). To derive $\boldsymbol{\Phi}_t^{\boldsymbol{v}}\in\mathcal{V}$ that maximizes (53), we forget the constraint $\boldsymbol{\Phi}_t^{\boldsymbol{v}}\in\mathcal{V}$ for the moment, and differentiate (53) w.r.t. $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$. We have

$$\boldsymbol{\Phi}_t^{\boldsymbol{v}}=\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big], \tag{55}$$

where

$$\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\triangleq\left\langle(\boldsymbol{y}_t-\boldsymbol{x}_t)(\boldsymbol{y}_t-\boldsymbol{x}_t)^{\mathsf{H}}\right\rangle_{p(\boldsymbol{x}_t|\boldsymbol{y}_t;\Theta')} \tag{56}$$

$$=\big(\boldsymbol{\Phi}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'+\big\{\boldsymbol{y}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big\}\big\{\boldsymbol{y}_t-\big(\boldsymbol{\mu}_t^{\boldsymbol{x}|\boldsymbol{y}}\big)'\big\}^{\mathsf{H}}. \tag{57}$$

As is clear from the definition of $\mathcal{P}[\cdot]$, (55) certainly satisfies $\boldsymbol{\Phi}_t^{\boldsymbol{v}}\in\mathcal{V}$.

Although we have derived (55) through partial differentiation, we can also derive it more intuitively as follows. Inverting the sign and ignoring a constant independent of $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$, (53) becomes the following matrix Itakura-Saito divergence:

$$D_{\mathrm{IS}}\big(\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big];\boldsymbol{\Phi}_t^{\boldsymbol{v}}\big)\triangleq\mathrm{Tr}\big\{\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big](\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\big\} \tag{58}$$
$$-\log\det\big\{\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big](\boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\big\}-M.$$

Therefore, the maximization of (53) is equivalent to the minimization of (58). $D_{\mathrm{IS}}(\cdot,\cdot)$ is nonnegative, and equal to zero if and only if the two arguments are equal. Since $\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big]$ belong to the feasible set $\mathcal{V}$ of $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$, (58) is minimized when $\boldsymbol{\Phi}_t^{\boldsymbol{v}}=\mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big]$.

Next we consider the case $\mathcal{V}=\mathcal{V}_{\mathrm{coh}}$. Substituting

$$\boldsymbol{\Phi}_t^{\boldsymbol{v}}=\phi_t^v\boldsymbol{B}^{\boldsymbol{v}} \tag{59}$$

into the Q-function (41), and differentiating it w.r.t. ϕ_t^v, we have, as in the derivation of (47),

$$\phi_t^v=\frac{1}{M}\mathrm{Tr}\big[(\boldsymbol{B}^{\boldsymbol{v}})^{-1}\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big]. \tag{60}$$