General algorithms for estimating spectrogram and transfer functions of target signal for blind suppression of diffuse noise

Nobutaka Ito, Emmanuel Vincent, Nobutaka Ono, Shigeki Sagayama

GENERAL ALGORITHMS FOR ESTIMATING SPECTROGRAM AND TRANSFER FUNCTIONS OF TARGET SIGNAL FOR BLIND SUPPRESSION OF DIFFUSE NOISE

Nobutaka Ito[†]*, Emmanuel Vincent*[‡]*, Nobutaka Ono*[*]*, and Shigeki Sagayama*[†]

[†] The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[‡] Inria, 54600 Villers-lès-Nancy, France
[*] National Institute of Informatics / The Graduate University for Advanced Studies
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

ABSTRACT

We propose two algorithms for jointly estimating the power spectrogram and the room transfer functions of a target signal in diffuse noise. These estimates can be used to design a multichannel Wiener filter, and thereby separate a target signal *from an unknown direction* from diffuse noise. We express a diffuse noise model as a subspace of a matrix linear space, which consists of *Hermitian matrices* instead of Euclidean vectors. This general framework enables the design of new *general* algorithms applicable to all specific noise models, instead of multiple specific algorithms each applicable to a single model. The more general proposed algorithms resulted in superior noise suppression performance to our previous algorithms in terms of an output signal-to-noise ratio (SNR).

Index Terms— Diffuse noise, microphone arrays, multichannel Wiener filter, noise suppression, speech enhancement.

1. INTRODUCTION

This paper aims at microphone array signal processing for blind suppression of diffuse noise, specifically blind separation of a point-source target signal from diffuse noise. We distinguish two types of noise depending on its spatial propagation: *point-source noise* and *diffuse noise*. Diffuse noise is defined as noise due to many point sources (*e.g.*, many interfering speakers) or some continuous sources (*e.g.*, the vibrating body of a train car). While the suppression of point-source noise has been established [1, 2], suppression of diffuse noise still remains open and hinders successful application of noise suppression techniques in the real world.

To design optimal noise suppression filters, we need certain information on the signal and noise; blind suppression of noise requires estimating it from the noisy observation. Spatial filters (*e.g.*, null beamformers [1] and the minimum variance distortionless response (MVDR) beamformer [3]) require spatial information such as source locations, transfer functions, spatial covariance matrices, or the mixing matrix. On the other hand, time-frequency masks (*e.g.*, binary masks [4] and the Wiener mask [5]) require spectral information such as activation in the time-frequency domain and power spectrograms.

Typically, diffuse noise suppression is performed using the multichannel Wiener filter, which is decomposed into the MVDR beamformer and the subsequent Wiener mask [5, 6]. Designing the multichannel Wiener filter requires the room transfer functions and the power spectrogram of the target signal. Contrary to point-source noise, diffuse noise is modeled in the covariance matrix domain, instead of the linear time-frequency domain. Therefore, established methods such as independent component analysis [1] and clustering-based methods [2] cannot be applied in the presence of diffuse noise. Instead, a covariance matrix fitting approach is usually taken, in which the observed covariance matrix is fitted with the model covariance matrix. Zelinski [7], McCowan *et al.* [8], and Ito *et al.* [9, 10] proposed methods for estimating the power spectrogram of the target signal, based on different covariance matrix models of diffuse noise. These methods assumed prior knowledge of the transfer functions or the source location of the target signal. However, the MVDR beamformer is sensitive to estimation errors of the transfer functions [6], and inaccuracy of the given transfer functions or the assumption of planewave propagation can degrade the noise suppression performance significantly. Therefore, it is important to estimate the transfer functions of the target signal as well as its power spectrogram for effective suppression of diffuse noise.

In [9], we have proposed a first method for joint estimation of the power spectrogram and the transfer functions of a target signal. However, this method was limited to a diffuse noise model called a blind noise decorrelation model, which is only applicable to a specific class of array configuration.

In this paper, we propose *general* algorithms for jointly estimating the power spectrogram and the room transfer functions of the target signal, *applicable to existing noise models in the literature*. To this end, we propose a general linear algebraic framework for expressing diffuse noise models in a unified manner, where each model is specified as a subspace of a matrix linear space. The more general proposed algorithms resulted in superior noise suppression performance to our previous algorithms in terms of SNR.

The rest of this paper is structured as follows. In Section 2, we review the multichannel Wiener filter. In Section 3, we describe the proposed methods for joint estimation of the transfer functions and the power spectrogram. In Section 4, we experimentally evaluate the proposed methods, and we conclude in Section 5.

2. REVIEW: MULTICHANNEL WIENER FILTER FOR SUPPRESSING DIFFUSE NOISE

Throughout this paper, complex conjugation and Hermitian transposition are denoted by * and $^{\mathsf{H}}$, respectively. Signals are represented in the time-frequency domain as, *e.g.*, $\alpha(\tau, \omega)$, with τ and

This work was supported by Grant-in-Aid for JSPS Fellows 22-6927 from MEXT, Japan and by INRIA under the Associate Team Program VERSAMUS (http://versamus.inria.fr). The authors thank Yasuhisa-KOKI Co.,Ltd, Japan for the fabrication of a microphone array.

978-1-4799-1180-6/13/$31.00

ω denoting the frame index and the angular frequency. The covariance matrix of a zero-mean vector signal $\boldsymbol{\alpha}(\tau,\omega)$ is denoted by $\boldsymbol{\Phi}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\tau,\omega) \triangleq \mathcal{E}[\boldsymbol{\alpha}(\tau,\omega)\boldsymbol{\alpha}^{\mathsf{H}}(\tau,\omega)]$, where $\mathcal{E}[\cdot]$ is expectation. $\|\cdot\|$ denotes the Frobenius norm of a matrix or the Euclidean norm of a vector. $\mathrm{diag}(\alpha_1,\ldots,\alpha_i)$ denotes the diagonal matrix whose diagonal entries equal $\alpha_1,\ldots,\alpha_i$.

We model the observation by an array of M microphones as a mixture of a point-source target signal and diffuse noise. The target signal observed by the array can be modeled by a source-filter model as the target signal $s(\tau,\omega)$ filtered by transfer functions $\boldsymbol{h}(\omega) \in \mathbb{C}^M$. Here, we assume that the target source is static, and therefore $\boldsymbol{h}(\omega)$ is time-invariant. Precisely, choosing the first microphone as a reference, we denote by $s(\tau,\omega)$ the target signal observed by the first microphone, and therefore we set

$$h_1(\omega) = 1. \tag{1}$$

On the other hand, diffuse noise cannot be modeled by using transfer functions, and therefore, we simply denote it by $\boldsymbol{v}(\tau,\omega) \in \mathbb{C}^M$. Consequently, the observed signal $\boldsymbol{x}(\tau,\omega) \in \mathbb{C}^M$ is modeled as follows:

$$\boldsymbol{x}(\tau,\omega) = s(\tau,\omega)\boldsymbol{h}(\omega) + \boldsymbol{v}(\tau,\omega). \tag{2}$$

Diffuse noise suppression is formulated as the problem of estimating $s(\tau,\omega)$ given $\boldsymbol{x}(\tau,\omega)$.

The multichannel Wiener filter is often employed for suppressing diffuse noise. The filter is the linear estimator of $s(\tau,\omega)$ that minimizes the mean square error, and is given by $\hat{s}(\tau,\omega) = \mathcal{E}[s(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)]\boldsymbol{\Phi}^{-1}_{\boldsymbol{x}\boldsymbol{x}}(\tau,\omega)\boldsymbol{x}(\tau,\omega)$. Under the model (2) and the assumption that $s(\tau,\omega)$ and $\boldsymbol{v}(\tau,\omega)$ are mutually uncorrelated, the filter is decomposed into the MVDR beamformer and the Wiener mask $p(\tau,\omega)$ [5, 6]:

$$\hat{s}(\tau,\omega) = \underbrace{\frac{\phi_{ss}(\tau,\omega)}{\phi_{yy}(\tau,\omega)}}_{\triangleq\, p(\tau,\omega)} \cdot \underbrace{\frac{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}_{\boldsymbol{x}\boldsymbol{x}}(\tau,\omega)\boldsymbol{x}(\tau,\omega)}{\boldsymbol{h}^{\mathsf{H}}(\omega)\boldsymbol{\Phi}^{-1}_{\boldsymbol{x}\boldsymbol{x}}(\tau,\omega)\boldsymbol{h}(\omega)}}_{\triangleq\, y(\tau,\omega)}, \tag{3}$$

where $\phi_{ss}(\tau,\omega) \triangleq \mathcal{E}[|s(\tau,\omega)|^2]$ is the power spectrogram of the target signal, and $\phi_{yy}(\tau,\omega) \triangleq \mathcal{E}[|y(\tau,\omega)|^2]$ that of the output of the MVDR beamformer. For the blind design of (3), we need to estimate $\boldsymbol{h}(\omega)$ and $\phi_{ss}(\tau,\omega)$ from the observed noisy signals.

3. GENERAL ALGORITHMS FOR ESTIMATING POWER SPECTROGRAM AND ROOM TRANSFER FUNCTIONS

As we have seen in Section 2, designing the multichannel Wiener filter (3) requires the room transfer functions $\boldsymbol{h}(\omega)$ and the power spectrogram $\phi_{ss}(\tau,\omega)$ of the target signal. This section describes the proposed methods for jointly estimating $\boldsymbol{h}(\omega)$ and $\phi_{ss}(\tau,\omega)$ from the observed signals contaminated by diffuse noise.

The rest of this section is organized as follows. Section 3.1 briefly reviews existing models of diffuse noise. Section 3.2 describes the proposed linear algebraic framework for unifying these models. Section 3.3 formulates the estimation problem as covariance matrix fitting. In Sections 3.4 and 3.5, we propose two alternative algorithms for the optimization.

3.1. Existing diffuse noise models reviewed

We first overview existing parametric models of diffuse noise [7–11], which operate on spatial covariance matrices. In this subsection, $\alpha_i(\tau,\omega)$ denotes an unknown real-valued variable.

The *spatially uncorrelated noise model* [7] states that the diffuse noise components at different microphones are uncorrelated with each other. This model corresponds to a parametric spatial covariance matrix

$$\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega) = \mathrm{diag}\big(\alpha_1(\tau,\omega), \alpha_2(\tau,\omega), \cdots, \alpha_M(\tau,\omega)\big). \tag{4}$$

This model is valid for a microphone array with aperture much larger than the wavelength. The *fixed noise coherence model* [8, 11] states that $\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega)$ equals a known time-invariant coherence matrix $\boldsymbol{\Gamma}(\omega)$ scaled by a time-varying unknown factor $\alpha_1(\tau,\omega)$ modeling the noise power spectrogram as follows:

$$\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega) = \alpha_1(\tau,\omega)\boldsymbol{\Gamma}(\omega). \tag{5}$$

This model is valid for a microphone array in the free field and perfectly diffuse noise. The *blind noise decorrelation (BND) model* [9] states that $\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega)$ is diagonalized by a known unitary matrix $\boldsymbol{P}$, and is valid for a certain class of symmetric arrays called crystal arrays [9]. This model is written as follows:

$$\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega) = \boldsymbol{P}\mathrm{diag}\big(\alpha_1(\tau,\omega), \alpha_2(\tau,\omega), \cdots, \alpha_M(\tau,\omega)\big)\boldsymbol{P}^{\mathsf{H}}. \tag{6}$$

The *real-valued noise covariance model* [10] states that $\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega)$ is real-valued symmetric. Therefore, $\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega)$ is modeled parametrically, *e.g.*, for $M = 3$ as follows:

$$\boldsymbol{\Phi}_{\boldsymbol{v}\boldsymbol{v}}(\tau,\omega) = \begin{bmatrix} \alpha_1(\tau,\omega) & \alpha_2(\tau,\omega) & \alpha_3(\tau,\omega) \\ \alpha_2(\tau,\omega) & \alpha_4(\tau,\omega) & \alpha_5(\tau,\omega) \\ \alpha_3(\tau,\omega) & \alpha_5(\tau,\omega) & \alpha_6(\tau,\omega) \end{bmatrix}. \tag{7}$$

This model is valid for arbitrary array configurations, but it tends to overfit the data due to the high dimensionality.

The above models have been applied to the estimation of $\phi_{ss}(\tau,\omega)$ with known $\boldsymbol{h}(\omega)$ [7–10]. The blind noise decorrelation (BND) model has also been applied to joint estimation of $\phi_{ss}(\tau,\omega)$ and $\boldsymbol{h}(\omega)$ [9]. We unify these models in a linear algebraic framework in Section 3.2, and apply the general noise model to joint estimation of $\phi_{ss}(\tau,\omega)$ and $\boldsymbol{h}(\omega)$ in Sections 3.3, 3.4, and 3.5. We evaluate the noise suppression performance of the proposed methods with each noise model in Section 4.

3.2. Proposed linear algebraic framework for unifying existing diffuse noise models

In this subsection, we propose a linear algebraic framework (see Fig. 1) for unifying existing models in Section 3.1. This general formulation has several advantages. First, it highlights the theoretical connections between the previous models [7–11] by describing them in a unified framework. Second, as shown in Sections 3.3, 3.4, and 3.5, it enables the design of new general algorithms applicable to all specific noise models, instead of multiple specific algorithms each applicable to a single model. Third, it facilitates the design of new noise models in the future by restricting the search space for these models; instead of searching for arbitrary, *e.g.*, nonlinear models, we shall restrict ourselves to linear subspace models.

Since the models in Section 3.1 are linear w.r.t. $\alpha_i(\tau,\omega)$, the resulting spatial covariance matrices belong to a subspace $\mathcal{V}(\omega)$ of the $\mathbb{R}$-linear space of the $M \times M$ Hermitian matrices

$$\mathcal{H} \triangleq \{\boldsymbol{A} \in \mathbb{C}^{M\times M} | \boldsymbol{A}^{\mathsf{H}} = \boldsymbol{A}\}. \tag{8}$$

$\mathcal{H}$ is endowed with the inner product

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle \triangleq \sum_{m=1}^{M}\sum_{n=1}^{M} a_{mn}b^*_{mn} = \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^{\mathsf{H}}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) \tag{9}$$

Table 1. Diffuse noise models in the literature expressed as a matrix linear subspace $\mathcal{V}(\omega)$. A basis of $\mathcal{V}(\omega)$ and the orthogonal projection operator $\mathcal{P}_\omega$ onto $\mathcal{V}(\omega)$ are shown for each specific model. $\boldsymbol{E}_{mn} \in \mathcal{H}$ denotes the $M \times M$ matrix, whose entries are all zeros except the (m,n) and (n,m) entries equal to one. $\mathcal{D}$ denotes the operation of replacing the off-diagonal entries of a matrix by zeros. $\boldsymbol{P}$ denotes a unitary matrix for blind noise decorrelation [9]. $\boldsymbol{p}_m$ denotes the mth column of $\boldsymbol{P}$. $\Re$ denotes the operation of taking the real part.

Noise models	References	Basis of $\mathcal{V}(\omega)$	$\mathcal{P}_\omega[\boldsymbol{A}]$
Spatially uncorrelated noise model	[7]	$\{\boldsymbol{E}_{mm} \mid 1 \le m \le M\}$	$\mathcal{D}[\boldsymbol{A}]$
Fixed noise coherence model	[8, 11]	$\boldsymbol{\Gamma}(\omega)$	$\frac{\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Gamma}(\omega))}{\|\boldsymbol{\Gamma}(\omega)\|^2}\boldsymbol{\Gamma}(\omega)$
Blind noise decorrelation (BND) model	[9]	$\{\boldsymbol{p}_m \boldsymbol{p}_m^{\mathsf{H}} \mid 1 \le m \le M\}$	$\boldsymbol{P}\mathcal{D}[\boldsymbol{P}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{P}]\boldsymbol{P}^{\mathsf{H}}$
Real-valued noise covariance model	[10]	$\{\boldsymbol{E}_{mn} \mid 1 \le m \le n \le M\}$	$\Re[\boldsymbol{A}]$

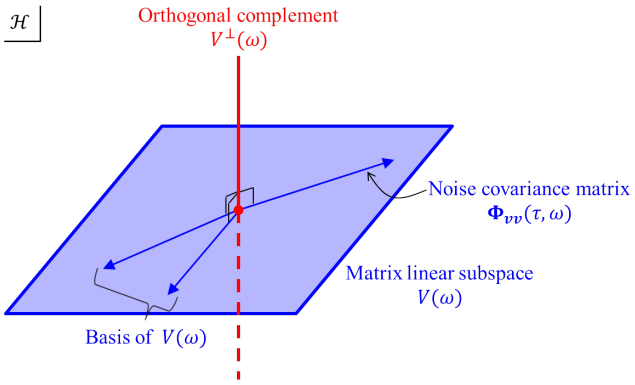

Fig. 1. We express a diffuse noise model in terms of a matrix linear space, which consists of *matrices* instead of Euclidean vectors. Specifically, we assume that the diffuse noise covariance matrix $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)$ belongs to a certain subspace $\mathcal{V}(\omega)$ of the linear space $\mathcal{H}$ spanned by all $M \times M$ Hermitian matrices (M: number of microphones). Existing diffuse noise models can be expressed as $\mathcal{V}(\omega)$, where the difference of the model boils down to that of the basis (see Table 1). This framework enables derivation of general algorithms applicable to all these noise models (see Sections 3.4 and 3.5).

and with the Frobenius norm $\|\boldsymbol{A}\| \triangleq \sqrt{\langle \boldsymbol{A}, \boldsymbol{A}\rangle}$. $\mathcal{V}(\omega)$ is specified either by a set of basis vectors or by the orthogonal projection operator $\mathcal{P}_\omega$:

$$\mathcal{P}_\omega[\boldsymbol{A}] \triangleq \sum_{i=1}^{P} \langle \boldsymbol{A}, \boldsymbol{Q}_i(\omega)\rangle \boldsymbol{Q}_i(\omega), \tag{10}$$

where $\{\boldsymbol{Q}_i(\omega)\}_{i=1}^P$ denotes an orthonormal basis of $\mathcal{V}(\omega)$ with $P \triangleq \dim\mathcal{V}(\omega)$. Table 1 shows a set of basis vectors of $\mathcal{V}(\omega)$ and the expression of the projection $\mathcal{P}_\omega$ for the models in Section 3.1.

3.3. Formulation of the optimization problem

While the point-source target signal is modeled by the source-filter model in the time-frequency domain, diffuse noise is modeled by a parametric covariance matrix, as we have seen in Section 3.2. In this subsection, we formulate the estimation of $\boldsymbol{h}(\omega)$ and $\phi_{ss}(\tau,\omega)$ as a covariance matrix fitting problem, where the observed spatial covariance matrix is fitted with a model spatial covariance matrix.

First, we model the observed spatial covariance matrix. Assuming that $s(\tau,\omega)$ and $\boldsymbol{v}(\tau,\omega)$ are uncorrelated with each other, the observed covariance matrix is modeled as the sum of the target and the noise covariance matrices. The noise spatial covariance matrix is modeled using the matrix linear space as $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega) \in \mathcal{V}(\omega)$. On the other hand, since the target signal is modeled using time-invariant transfer functions in the time-frequency domain, the corresponding model in the covariance matrix domain is the rank-one matrix constant up to a time-varying scale factor:

$$\mathcal{E}[|s(\tau,\omega)|^2 \boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega)] = \phi_{ss}(\tau,\omega)\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega). \tag{11}$$

Therefore, the observed spatial covariance matrix is modeled as follows:

$$\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau,\omega) = \phi_{ss}(\tau,\omega)\boldsymbol{h}(\omega)\boldsymbol{h}^{\mathsf{H}}(\omega) + \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega). \tag{12}$$

We formulate the estimation of $\boldsymbol{h}(\omega)$ and $\phi_{ss}(\tau,\omega)$ as the problem of fitting model (12) to the observed covariance matrix. That is, for a measure of discrepancy $D(\cdot,\cdot)$ between two matrices, we solve the following optimization problem:

$$\min_{\Theta} \sum_{\tau} D(\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau), \phi_{ss}(\tau)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}} + \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)), \tag{13}$$
$$\text{s.t. } \phi_{ss}(\tau) \ge 0, \|\boldsymbol{h}\| = 1, \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau) \in \mathcal{V}.$$

Here, we have omitted ω for the optimization problem is formulated frequency bin-wise, and $\Theta \triangleq \{\{\phi_{ss}(\tau)\}_\tau, \boldsymbol{h}, \{\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\}_\tau\}$ denotes the set of parameters to be estimated. Choosing the Eulidean distance as the discrepancy measure, (13) becomes

$$\min_{\Theta} \sum_{\tau} \|\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau) - \phi_{ss}(\tau)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}} - \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\|^2, \tag{14}$$
$$\text{s.t. } \phi_{ss}(\tau) \ge 0, \|\boldsymbol{h}\| = 1, \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau) \in \mathcal{V}.$$

We plan to study more sophisticated measures than the Eulidean distance in the future.

Note that, in (14), we impose the constraint $\|\boldsymbol{h}\| = 1$ instead of (1), to derive simple algorithms. Once the solutions $\phi_{ss}(\tau)$ and $\boldsymbol{h}$ of (14) are obtained, they are post-processed to satisfy (1) as follows: $\phi_{ss}(\tau) \leftarrow |h_1|^2\phi_{ss}(\tau)$, $\boldsymbol{h} \leftarrow \boldsymbol{h}/h_1$.

3.4. Method 1: estimation via sequential procedure

The first algorithm for solving the optimization problem in (14) is derived by eliminating the nuisance parameters $\{\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\}_\tau$, and solving the resulting optimization problem through a sequential procedure.

Noting that $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau) \in \mathcal{V}$, we can decompose the error in (14) into $\mathcal{V}$ and $\mathcal{V}^\perp$ components as follows:

$$\begin{aligned}&\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau) - \phi_{ss}(\tau)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}} - \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\\ &= \underbrace{\mathcal{P}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau) - \phi_{ss}(\tau)\mathcal{P}[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}] - \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)}_{\in\mathcal{V}}\\ &+ \underbrace{\mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau) - \phi_{ss}(\tau)\mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}]}_{\in\mathcal{V}^\perp}.\end{aligned} \tag{15}$$

Here, $\mathcal{V}^\perp$ denotes the orthogonal complement of $\mathcal{V}$. Therefore, applying the Pythagorean theorem, we can decompose the square error in (14) into $\mathcal{V}$ and $\mathcal{V}^\perp$ components as follows:

$$\begin{aligned}&\|\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau) - \phi_{ss}(\tau)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}} - \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\|^2 \\ &= \|\mathcal{P}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau) - \phi_{ss}(\tau)\mathcal{P}[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}] - \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\|^2 \\ &+ \|\mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau) - \phi_{ss}(\tau)\mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}]\|^2.\end{aligned} \tag{16}$$

Here, $\mathcal{P}^\perp$ denotes the orthogonal projection operator onto $\mathcal{V}^\perp$:

$$\mathcal{P}^\perp[\boldsymbol{A}] \triangleq \boldsymbol{A} - \mathcal{P}[\boldsymbol{A}]. \tag{17}$$

Therefore, by removing the nuisance parameter $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)$ by replacing it with its optimal value

$$\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau) \leftarrow \mathcal{P}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau) - \phi_{ss}(\tau)\mathcal{P}[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}], \tag{18}$$

the optimization problem reduces to

$$\begin{aligned}&\min_{\Omega} J(\Omega) \triangleq \sum_{\tau} \|\mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau) - \phi_{ss}(\tau)\mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}]\|^2 \\ &\text{s.t. } \phi_{ss}(\tau) \geq 0, \|\boldsymbol{h}\| = 1,\end{aligned} \tag{19}$$

where $\Omega \triangleq \{\{\phi_{ss}(\tau)\}_\tau, \boldsymbol{h}\}$ is the set of desired parameters.

$J(\Omega)$ in (19) depends on $\boldsymbol{h}$ in a rather complex way. Therefore, instead of directly minimizing $J(\Omega)$ w.r.t. Ω, we follow the following sequential procedure:

1. Estimate $\{\phi_{ss}(\tau)\}_\tau$ and $\boldsymbol{Z} \triangleq \mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}]$ by minimizing $J(\Omega)$ w.r.t. $\{\phi_{ss}(\tau)\}_\tau$ and $\boldsymbol{Z}$.
2. Reconstruct $\boldsymbol{W} \triangleq \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}$ by low-rank matrix completion [13] of $\boldsymbol{Z}$.
3. Estimate $\boldsymbol{h}$ as a unit principal eigenvector of $\boldsymbol{W}$. Scale $\boldsymbol{h}$ by $\boldsymbol{h} \leftarrow \boldsymbol{h}/h_1$.
4. Reestimate $\{\phi_{ss}(\tau)\}_\tau$ by minimizing $J(\Omega)$ w.r.t. $\{\phi_{ss}(\tau)\}_\tau$ given $\boldsymbol{h}$.

The update rules for the first step are derived through the differentiation of (19) w.r.t. $\phi_{ss}(\tau)$ and $\boldsymbol{Z}^*$ as follows:

$$\phi_{ss}(\tau) \leftarrow \frac{\langle \mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau), \boldsymbol{Z}\rangle}{\|\boldsymbol{Z}\|^2}, \tag{20}$$

$$\boldsymbol{Z} \leftarrow \frac{\sum_\tau \phi_{ss}(\tau)\mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau)}{\sum_\tau \phi_{ss}^2(\tau)}. \tag{21}$$

Since the updates of $\phi_{ss}(\tau)$ and $\boldsymbol{Z}$ are interdependent, we iterate (20) and (21) alternately. $\boldsymbol{Z}$ is initialized by $\boldsymbol{Z} \leftarrow \mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}]$ with $\boldsymbol{h}$ initialized by a conventional technique. In the following experiment, $\boldsymbol{h}$ is initialized by independent vector analysis (IVA) [14].

The second step is based on low-rank matrix completion [13]. We search for a Hermitian positive semidefinite matrix $\boldsymbol{W}$ of rank no more than 1, whose projection $\mathcal{P}^\perp[\boldsymbol{W}]$ is as close to $\boldsymbol{Z}$ as possible. This can be formulated as follows:

$$\begin{aligned}&\min_{\boldsymbol{W}} \|\mathcal{P}^\perp[\boldsymbol{W}] - \boldsymbol{Z}\|^2 \\ &\text{s.t. } \boldsymbol{W}: \text{ Hermitian positive semidefinite, } \mathrm{rank}(\boldsymbol{W}) \leq 1.\end{aligned} \tag{22}$$

We can decrease the cost function in (22) monotonically by iterating the following updates alternately. This procedure has been inspired by the algorithm in [13].

1. $\boldsymbol{Y} \leftarrow \mathcal{P}[\boldsymbol{W}] + \boldsymbol{Z}$.
2. Derive the eigendecomposition of $\boldsymbol{Y}$: $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathsf{H}}$, where $\boldsymbol{U}$ is unitary, and the diagonal entries (real-valued) of $\boldsymbol{\Sigma}$ are arranged in decreasing order.
3. $\boldsymbol{W} \leftarrow \max\{\sigma_{11}, 0\}\boldsymbol{u}_1\boldsymbol{u}_1^{\mathsf{H}}$, where σ_{11} denotes the (1,1) entry of $\boldsymbol{\Sigma}$, and $\boldsymbol{u}_1$ the first column of $\boldsymbol{U}$.

$\boldsymbol{W}$ is initialized by a rough estimation obtained by ignoring noise term in (14). By minimizing $\sum_\tau \|\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau) - \phi_{ss}(\tau)\boldsymbol{W}\|^2$, we have

$$\boldsymbol{W} \leftarrow \frac{\sum_\tau \phi_{ss}(\tau)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau)}{\sum_\tau \phi_{ss}^2(\tau)}. \tag{23}$$

The algorithm is summarized as follows, where T denotes the number of frames, and iter_num the number of iterations:

Algorithm 1 (Method 1: estimation via sequential procedure)

Initialize $\boldsymbol{Z}$ *by* $\boldsymbol{Z} \leftarrow \mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^H]$ *with* $\boldsymbol{h}$ *initialized by a conventional technique.*
for $cnt = 1$ *to* $iter\_num$ ***do***
  for $\tau = 1$ *to* T ***do***
    $\phi_{ss}(\tau) \leftarrow \frac{\langle \mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau), \boldsymbol{Z}\rangle}{\|\boldsymbol{Z}\|^2}$.
  end for
  $\boldsymbol{Z} \leftarrow \frac{\sum_\tau \phi_{ss}(\tau)\mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau)}{\sum_\tau \phi_{ss}^2(\tau)}$.
end for
Initialize $\boldsymbol{W}$ *by* $\boldsymbol{W} \leftarrow \frac{\sum_\tau \phi_{ss}(\tau)\boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau)}{\sum_\tau \phi_{ss}^2(\tau)}$.
for $cnt = 1$ *to* $iter\_num$ ***do***
  $\boldsymbol{Y} \leftarrow \mathcal{P}[\boldsymbol{W}] + \boldsymbol{Z}$.
  Derive the eigendecomposition of $\boldsymbol{Y}$: $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^H$, *where* $\boldsymbol{U}$ *is unitary, and the diagonal entries of* $\boldsymbol{\Sigma}$ *are arranged in decreasing order.*
  $\boldsymbol{W} \leftarrow \max\{\sigma_{11}, 0\}\boldsymbol{u}_1\boldsymbol{u}_1^H$, *where* σ_{11} *denotes the (1,1) entry of* $\boldsymbol{\Sigma}$, *and* $\boldsymbol{u}_1$ *the first column of* $\boldsymbol{U}$.
end for
$\boldsymbol{h} \leftarrow \boldsymbol{u}_1/u_{11}$, *where* u_{11} *denotes the first element of* $\boldsymbol{u}_1$.
for $\tau = 1$ *to* T ***do***
  $\phi_{ss}(\tau) \leftarrow \frac{\langle \mathcal{P}^\perp[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau), \mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^H]\rangle}{\|\mathcal{P}^\perp[\boldsymbol{h}\boldsymbol{h}^H]\|^2}$.
end for

3.5. Method 2: estimation in a single stage

The second algorithm for solving the optimization problem in (14) minimizes iteratively and alternately the cost function in (14) w.r.t. $\{\phi_{ss}(\tau)\}_\tau$, $\boldsymbol{h}$, and $\{\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau)\}_\tau$.

We define

$$\boldsymbol{Y}(\tau) \triangleq \boldsymbol{\Phi}_{\boldsymbol{xx}}(\tau) - \boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau). \tag{24}$$

Applying (24) and the constraint $\|\boldsymbol{h}\| = 1$, we can expand the cost function in (14) as follows:

$$\begin{aligned}&\sum_\tau \|\boldsymbol{Y}(\tau) - \phi_{ss}(\tau)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\|^2 \\ &= \sum_\tau \phi_{ss}^2(\tau) - 2\boldsymbol{h}^{\mathsf{H}}\left[\sum_\tau \phi_{ss}(\tau)\boldsymbol{Y}(\tau)\right]\boldsymbol{h} + \sum_\tau \|\boldsymbol{Y}(\tau)\|^2.\end{aligned} \tag{25}$$

Noting that (25) is quadratic w.r.t. $\phi_{ss}(\tau)$, the minimizer $\phi_{ss}(\tau)$ subject to the constraint $\phi_{ss}(\tau) \geq 0$ is straightforwardly obtained as follows:

$$\phi_{ss}(\tau) \leftarrow \max\{\boldsymbol{h}^{\mathsf{H}}\boldsymbol{Y}(\tau)\boldsymbol{h}, 0\}. \tag{26}$$

On the other hand, by applying the Courant-Fischer theorem, we have the following update rule for $\boldsymbol{h}$:

$$\boldsymbol{h} \leftarrow \text{unit principal vector of } \sum_\tau \phi_{ss}(\tau)\boldsymbol{Y}(\tau). \tag{27}$$

Fig. 2. Fabricated 12-element spherical microphone array of diameter 15 cm. The microphones are mounted on a rigid spherical shell.

Eliminating the nuisance parameter $\boldsymbol{\Phi}_{\boldsymbol{vv}}(\tau,\omega)$ as in the first method, we obtain the following algorithm:

Algorithm 2 (Method 2: estimation in a single stage)

Initialize $\boldsymbol{h}$ *so that* $\|\boldsymbol{h}\| = 1$ *by a conventional technique, and initialize* $\phi_{ss}(\tau)$ *by* $\phi_{ss}(\tau) \leftarrow \frac{\langle \mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}], \mathcal{P}^{\perp}[\boldsymbol{h}\boldsymbol{h}^{H}] \rangle}{\|\mathcal{P}^{\perp}[\boldsymbol{h}\boldsymbol{h}^{H}]\|^2}$.
for $cnt = 1$ *to* $iter\_num$ ***do***
  for $\tau = 1$ *to* T ***do***
    $\boldsymbol{Y}(\tau) \leftarrow \phi_{ss}(\tau)\mathcal{P}[\boldsymbol{h}\boldsymbol{h}^{H}] + \mathcal{P}^{\perp}[\boldsymbol{\Phi}_{\boldsymbol{xx}}](\tau)$.
    $\phi_{ss}(\tau) \leftarrow \max\{\boldsymbol{h}^{H}\boldsymbol{Y}(\tau)\boldsymbol{h}, 0\}$.
  end for
  $\boldsymbol{h} \leftarrow$ *unit principal eigenvector of* $\sum_{\tau} \phi_{ss}(\tau)\boldsymbol{Y}(\tau)$.
end for
for $\tau = 1$ *to* T ***do***
  $\phi_{ss}(\tau) \leftarrow |h_1|^2 \phi_{ss}(\tau)$.
end for
$\boldsymbol{h} \leftarrow \boldsymbol{h}/h_1$.

4. PERFORMANCE EVALUATION ON REAL-WORLD DATA

4.1. Experimental conditions

To record real-world data, we fabricated a 12-channel spherical microphone array with microphones at the vertices of an icosahedron of diameter 15 cm (see Fig. 2). The microphones were mounted on a rigid spherical shell.

With the array, we recorded the multichannel source and noise images in an experimental room at the University of Tokyo. The configuration in the experiment is shown in Fig. 3. The source image was recorded while the loudspeaker played female speech [15], and the noise image was recorded with the windows open. They were mixed to generate the observed signals.

We compared the following four algorithms for estimating the power spectrogram and the room transfer functions:

- *conv1*: conventional power spectrogram estimation for known room transfer functions in [9], combined with independent vector analysis (IVA) [14] for room transfer function estimation
- *conv2*: conventional joint estimation in [9]

Table 2. 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	8 s
sampling frequency	16 kHz
frame length	2048 samples
frame shift	64 samples
window	Hamming window
number of iterations	100

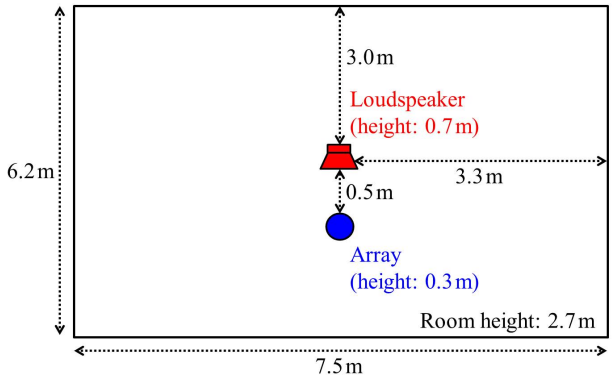

Fig. 3. Geometrical configuration of the experimental room, where the multichannel source and noise images were recorded. The source image was recorded while the loudspeaker played female speech, and the noise image was recorded with the windows open. They were mixed to generate the observed signals.

- *prop1*: proposed joint estimation in Section 3.4
- *prop2*: proposed joint estimation in Section 3.5

The estimates by conv1 were used to initialize the other three algorithms.

The observed signals were first analyzed by the short-time Fourier transform (STFT). The lower 14 frequency bins of the observed signals were discarded, which contain only noise. The observed covariance matrix for estimating $\boldsymbol{h}(\omega)$ and $\phi_{ss}(\tau,\omega)$ was computed locally by averaging $\boldsymbol{x}(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)$ over 48 consecutive frames. On the other hand, the observed covariance matrix for the MVDR beamformer was calculated as the long-term average of $\boldsymbol{x}(\tau,\omega)\boldsymbol{x}^{\mathsf{H}}(\tau,\omega)$ over the whole data. The other conditions are summarized in Table 2.

4.2. Experimental results

Fig. 4 shows the output SNR [9] of the multichannel Wiener filter designed with each of the above algorithms. The labels "uncor", "coh", "BND", and "real" refer to the diffuse noise models in Section 3.1. The input SNR was -0.2 dB. Proposed prop1 and prop2 resulted in output SNRs slightly better than conv2 when combined with the BND model. The poor performance of prop1 combined with the real-valued noise covariance model is likely due to local minima of the cost function resulting from the high dimensionality of the model. The other proposed methods improved the SNR by 5.3 to 10.7 dB compared to the observation.

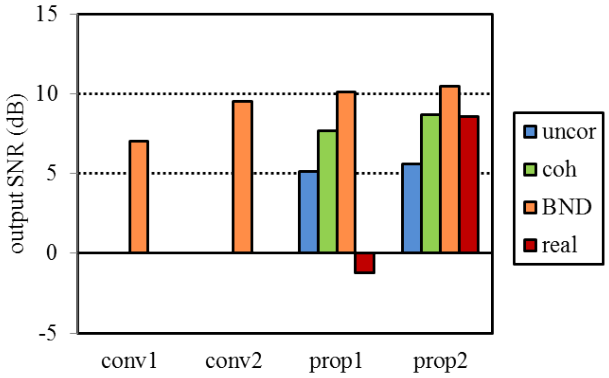

Fig. 4. Comparison of different algorithms for estimating the power spectrogram and the room transfer functions of the target signal. They are compared in terms of the output SNR of the multichannel Wiener filter. The proposed algorithms resulted in output SNRs slightly better than conventional conv2 when combined with the blind noise decorrelation (BND) model. Furthermore, the proposed algorithms are more general than the conventional algorithms in that the former apply to other noise models as well.

5. CONCLUSION

We proposed two algorithms for joint estimation of the power spectrogram and the room transfer functions of the target signal for blind suppression of diffuse noise. The algorithms are general in that they apply to existing diffuse noise models. Furthermore, the proposed algorithms resulted in higher output SNRs than our previous algorithms when combined with the blind noise decorrelation (BND) model.

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., New York, 2001.

[2] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[3] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 241–244.

[4] Ö. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[5] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 3, pp. 39–60. Springer-Verlag, Berlin, 2001.

[6] H. L. V. Trees, *Optimum Array Processing*, John Wiley & Sons, New York, 2002.

[7] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, New York, Apr. 1988, pp. 2578–2581.

[8] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[9] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2101–2110, Sept. 2011.

[10] 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. IEEE Int'l Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 2818 – 2821.

[11] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr., "Measurement of correlation coefficients in reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, Nov. 1955.

[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, 1990.

[13] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. International Conference on Machine Learning (ICML)*. AAAI Press, 2003, pp. 720–727.

[14] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2011, pp. 189–192.

[15] 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 Commun.*, vol. 9, no. 4, pp. 357–363, Aug. 1990.