PROBABILISTIC INTEGRATION OF DIFFUSE NOISE SUPPRESSION AND DEREVERBERATION

Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani

NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4, Hikaridai, Seika-cho, "Keihanna Science City" Kyoto 619-0237 Japan

ABSTRACT

This paper deals with joint suppression of diffuse noise and reverberation, to enhance perceived speech quality and speech recognition performance. Although diffuse noise and reverberation are both omnipresent in the real world, conventional methods have modeled only one while neglecting the other. In contrast, we propose a novel joint suppression method that employs a unified probabilistic model of observed signals affected by both diffuse noise and reverberation. Through likelihood maximization, this unified model enables proper parameter estimation that takes into account both diffuse noise and reverberation. As a byproduct, we also propose a novel method for diffuse noise suppression. Experimental results demonstrate the effectiveness of the proposed joint suppression method in terms of dereverberation and denoising.

Index Terms— Denoising, dereverberation, diffuse noise, expectation-maximization, speech enhancement.

1. INTRODUCTION

Blind signal processing is a highly underdetermined task in general. Blind dereverberation, for instance, aims to reconstruct an unknown desired signal using its convolved versions with unknown room impulse responses. It would be impossible to perform such blind signal processing without any prior knowledge about signals and the environment. Therefore, it is crucial to construct an appropriate model that describe our knowledge about the general *structures* of signals and the environment.

Linear prediction based dereverberation [1–5] is among the most promising dereverberation techniques [6–8]. It exploits quasi-stationarity of the desired speech and predictability of reverberation as the above-mentioned structures. On the one hand, the desired speech is correlated only within several tens of milliseconds, and approximately uncorrelated for a longer time lag [1]. This structure is reflected in speech modeling as a time-varying random variable that is uncorrelated across time frames. On the other hand, reverberation is predictable using past samples. This structure is modeled as an autoregressive model [1], which excludes a few most recent samples from prediction to avoid whiting out the above short-term correlation of speech. These models enable distinction between the desired speech and reverberation and estimation of prediction gains for dereverberation through likelihood maximization.

Similarly, our recently developed method for diffuse noise suppression [9,10] exploits locality of the desired speech source and diffuseness of noise. The desired speech is emitted from a small source (*i.e.*, mouth) that is approximately spatially stationary within a few seconds. This structure is modeled using a time-invariant, rank-one spatial coherence matrix (*i.e.*, normalized spatial covariance matrix). On the other hand, diffuse noise is caused by many sources (*e.g.*, many people talking concurrently), and thus spatially diffuse. This structure is modeled as a low-dimensional subspace in a linear space spanned by spatial covariance matrices [9, 10]. These models enable distinction between the desired signal and diffuse noise and estimation of the parameters for denoising such as the signal power.

For our purpose of jointly suppressing diffuse noise and reverberation in this paper, we need to distinguish the desired signal from diffuse noise and reverberation. To this end, we construct a unified probabilistic model by integrating probabilistic models for diffuse noise suppression and dereverberation. This unified model enables appropriate parameter estimation that takes into account both diffuse noise and reverberation, which have been treated separately in the literature. The primary focus of this paper resides in the theoretical aspect and a preliminary experiment, with thorough experimental evaluation being future work.

Yoshioka *et al.* [2, 3] have also dealt with joint denoising and dereverberation, but assumed that noise statistics or speech absence periods are known. In contrast, our methods estimate the noise spatial covariance matrix in a fully blind manner. Togami *et al.* [4] have proposed a method involving blind estimation of noise statistics, but assumed noise stationarity. Our methods can deal with even nonstationary noise, owing to noise modeling based on the spatial structure, instead of the spatiotemporal structure.

The rest of this paper is structured as follows. Section 2 proposes a method for diffuse noise suppression. This method is novel by itself, and also probabilistically integrated into a joint suppression method in Section 3. Section 4 verifies the proposed joint suppression method through an experiment, and Section 5 concludes this paper.

2. PROBABILISTIC DIFFUSE NOISE SUPPRESSION

In this paper, we consider a sound field that contains 1) a spatially stationary point source of a desired signal and 2) diffuse noise. Since we focus on the integration of diffuse noise suppression and dereverberation, other issues such as source movement, multiple sources, and an unknown and/or time-varying source number are beyond the scope of this paper.

2.1. Probabilistic observation model

We define diffuse noise suppression as the problem of estimating the microphone image of the desired signal from signals observed with M microphones in the above sound field. We operate in the short-time Fourier transform (STFT) domain, and denote the time

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

and frequency indices by $t \in \{1, \ldots, T\}$ and $f \in \{1, \ldots, F\}$, respectively. Let $\boldsymbol{y}_{tf}, \boldsymbol{x}_{tf}, \boldsymbol{v}_{tf} \in \mathbb{C}^M$ be the observed signals, the microphone image of the desired signal, and that of diffuse noise, respectively. The observed signals $\boldsymbol{y}_{tf}$ are modeled by

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

As we see in (1), in this paper, we omit the subscript f for brevity, which should not cause confusion, because the proposed methods process each frequency bin independently. For simplicity, we assume that $\boldsymbol{x}_t$ and $\boldsymbol{v}_t$ are Gaussian random variables with zero mean and time-varying covariance matrices $\boldsymbol{\Phi}_t^{\boldsymbol{x}}, \boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathbb{C}^{M \times M}$. We also assume that $\{\boldsymbol{x}_t\}_{t=1}^T$ and $\{\boldsymbol{v}_t\}_{t=1}^T$ are mutually and temporally independent. We shall relax the temporal independence assumption on $\{\boldsymbol{x}_t\}_{t=1}^T$ in Section 3 by modeling its temporal correlation due to reverberation. Studies of more appropriate distributions than the Gaussian distribution and temporal noise correlation across frames are included in the future work.

Looking at (1), we see a fundamental underdetermined nature of our inverse problem. Indeed, we know only $\boldsymbol{y}_t$ in the left-hand side, but neither $\boldsymbol{x}_t$ nor $\boldsymbol{v}_t$ in the right-hand side. Therefore, to make the problem solvable, we need to construct appropriate models of $\boldsymbol{x}_t$ and $\boldsymbol{v}_t$. Because of the above assumptions of zero-mean Gaussian distributions and independence, modeling of $\boldsymbol{x}_t$ and $\boldsymbol{v}_t$ boils down to modeling of $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$.

The desired speech is emitted from a small source (*i.e.*, mouth) that is approximately spatially stationary within a short observation period. Therefore, $\boldsymbol{x}_t$ is often modeled using a source signal $x_t \in \mathbb{C}$ and time-invariant room transfer functions $\boldsymbol{h} \in \mathbb{C}^M$ as $\boldsymbol{x}_t = x_t \boldsymbol{h}$. The corresponding model in the spatial covariance domain is the following rank-one covariance model as employed in [9, 10]: $\boldsymbol{\Phi}_t^{\boldsymbol{x}} = \phi_t^x \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}$. Here, $\phi_t^x \in \mathbb{R}$ denotes the time-varying power spectrum of the desired signal, and $(\cdot)^{\mathsf{H}}$ Hermitian transposition. In this paper, we replace the above rank-one matrix $\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}$ by a full-rank matrix $\boldsymbol{B} \in \mathbb{C}^{M \times M}$, because of two reasons. First, the full-rank model has proven more effective than the rank-one model in the context of source separation [11]. Second, the Gaussian distribution for the desired signal is not well-defined for a rank-deficient covariance matrix. Consequently, our signal model is expressed as

$$\boldsymbol{\Phi}_t^{\boldsymbol{x}} = \phi_t^x \boldsymbol{B}. \tag{2}$$

Note that the time-invariant spatial coherence matrix $\boldsymbol{B}$ reflects the spatial stationarity of the desired speech source.

On the other hand, diffuse noise is caused by many, possibly nonstationary sources (*e.g.*, many people talking concurrently), where different noise sources are active at different time-frequency points. Consequently, diffuse noise $\boldsymbol{v}_t$ is generally nonstationary both temporally and spatially, and therefore it is difficult to model its structure in the linear time-frequency domain. To model diffuse noise properly, in [9, 10], we have proposed modeling based on a matrix linear space. It is expressed as

$$\boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{V}, \tag{3}$$

where $\mathcal{V}$ denotes a matrix subspace corresponding to diffuse noise, which is a low-dimensional subspace of the vector space $\mathcal{H}$ spanned by the $M \times M$ Hermitian matrices. By this matrix subspace, we can appropriately model the structure of diffuse noise. For more details of this matrix subspace model, we refer the readers to [9, 10]. Equations (1)–(3) constitute our probabilistic model for diffuse noise suppression.

2.2. Estimator of desired signal

Based on the probabilistic model in Section 2.1, we can derive the maximum *a posteriori* (MAP) estimator of $\boldsymbol{x}_t$, which can be calculated given the parameters $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$. The estimation of the parameters will be described later. The MAP estimator is defined as the $\boldsymbol{x}_t$ that maximizes the posterior probability $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$. Using the Bayes' theorem, we have $p(\boldsymbol{x}_t|\boldsymbol{y}_t) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with

$$\boldsymbol{\mu}_t \triangleq \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{y}_t, \tag{4}$$

$$\boldsymbol{\Sigma}_t \triangleq \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{\Phi}_t^{\boldsymbol{v}}. \tag{5}$$

Here, $\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the complex-valued Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Therefore, the MAP estimator is given by (4), which is the well-known multichannel Wiener filter [11–14]. Note that the MAP and the minimum mean square error (MMSE) estimators coincide under our Gaussian assumptions.

2.3. Parameter estimation algorithm

Now we move on to estimation of the parameters $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$. We can derive a simple algorithm for maximizing the likelihood function $p(\{\boldsymbol{y}_t\}_{t=1}^T; \{\boldsymbol{\Phi}_t^{\boldsymbol{x}}\}_{t=1}^T, \{\boldsymbol{\Phi}_t^{\boldsymbol{v}}\}_{t=1}^T)$ based on the expectation-maximization (EM) [15] by regarding $\{\boldsymbol{x}_t\}_{t=1}^T$ as hidden variables. Because of space limitation, we omit detailed derivation. One iteration of the algorithm is shown below, where $\mathcal{P}[\cdot]$ denotes the orthogonal projection onto the subspace $\mathcal{V}$.

[E-step]

1. $\boldsymbol{\mu}_t \leftarrow \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{y}_t$.
2. $\boldsymbol{\Sigma}_t \leftarrow \boldsymbol{\Phi}_t^{\boldsymbol{x}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1} \boldsymbol{\Phi}_t^{\boldsymbol{v}}$.
3. $\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}} \leftarrow \boldsymbol{\mu}_t \boldsymbol{\mu}_t^{\mathsf{H}} + \boldsymbol{\Sigma}_t$.
4. $\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}} \leftarrow (\boldsymbol{y}_t - \boldsymbol{\mu}_t)(\boldsymbol{y}_t - \boldsymbol{\mu}_t)^{\mathsf{H}} + \boldsymbol{\Sigma}_t$.

[M-step]

1. $\phi_t^x \leftarrow \dfrac{1}{M} \mathrm{Tr}\big[\boldsymbol{B}^{-1} \hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}\big]$.
2. $\boldsymbol{B} \leftarrow \dfrac{1}{T} \displaystyle\sum_{t=1}^T \dfrac{1}{\phi_t^x} \hat{\boldsymbol{\Phi}}_t^{\boldsymbol{x}}$.
3. $\boldsymbol{\Phi}_t^{\boldsymbol{x}} \leftarrow \phi_t^x \boldsymbol{B}$.
4. $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \leftarrow \mathcal{P}\big[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\big]$.

2.4. Discussion

Like the algorithm in Section 2.3, our previous methods in [9] estimate $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$. These previous methods employ the Euclidean distance between the observed covariance matrix $\boldsymbol{\Phi}_t^{\boldsymbol{y}} \triangleq \mathcal{E}[\boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{H}}]$ and the model covariance matrix $\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}$ as a cost function:

$$d_{\mathrm{E}}(\boldsymbol{\Phi}_t^{\boldsymbol{y}}, \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}) = \big\| \boldsymbol{\Phi}_t^{\boldsymbol{y}} - \boldsymbol{\Phi}_t^{\boldsymbol{x}} - \boldsymbol{\Phi}_t^{\boldsymbol{v}} \big\|_{\mathrm{F}}. \tag{6}$$

Here, $\| \cdot \|_{\mathrm{F}}$ denotes the Frobenius norm, and $\mathcal{E}[\boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{H}}]$ is practically replaced by an empirical estimate using short-term data around time t. On the other hand, the likelihood maximization in Section 2.3 is equivalent to minimization of a multichannel Itakura-Saito divergence [16–18]:

$$\begin{aligned} &d_{\mathrm{IS}}(\boldsymbol{\Phi}_t^{\boldsymbol{y}}, \boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}}) \\ &= -\log\det\big[\boldsymbol{\Phi}_t^{\boldsymbol{y}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\big] + \mathrm{Tr}\big[\boldsymbol{\Phi}_t^{\boldsymbol{y}} (\boldsymbol{\Phi}_t^{\boldsymbol{x}} + \boldsymbol{\Phi}_t^{\boldsymbol{v}})^{-1}\big] \end{aligned} \tag{7}$$

with $\boldsymbol{\Phi}_t^{\boldsymbol{y}} \triangleq \boldsymbol{y}_t \boldsymbol{y}_t^{\mathsf{H}}$. Owing to the scale invariance, the Itakura-Saito divergence can more properly measure the discrepancy between audio spectra, which have a logarithmic nature [16].

Note that we have defined $\boldsymbol{\Phi}_t^{\boldsymbol{y}}$ differently in (6) and (7) with and without the expectation operation. While we use here only one sample to calculate $\boldsymbol{\Phi}_t^{\boldsymbol{y}}$ for simplicity, we could also use several samples as in (6). This corresponds to maximum likelihood estimation with $\boldsymbol{\Phi}_t^{\boldsymbol{x}}$ and $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$ assumed to be constant for a few samples.

3. PROBABILISTIC INTEGRATION OF DIFFUSE NOISE SUPPRESSION AND DEREVERBERATION

3.1. Probabilistic observation model

In this section, we aim to suppress diffuse noise and reverberation jointly. That is, we aim to estimate the dereverberated desired signal from noisy, reverberant observations. To this end, we extend the probabilistic model for denoising in Section 2.1 by combining it with a probabilistic model for dereverberation [1, 4].

In Section 2.1, we modeled $\boldsymbol{x}_t$ as temporally independent. Here, we model it more precisely with a multichannel delayed autoregressive model as in linear prediction based dereverberation [1]:

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{G}_k^{\mathsf{H}} \boldsymbol{x}_{t-\Delta-k} + \boldsymbol{s}_t. \tag{8}$$

Here, $\boldsymbol{s}_t \in \mathbb{C}^M$ denotes the dereverberated desired signal, $\boldsymbol{G}_k \in \mathbb{C}^{M\times M}$ a prediction gain matrix, K a preset prediction order, and $\Delta \in \mathbb{N}$ a preset prediction delay. Equation (8) means that the present sample $\boldsymbol{x}_t$ is predicted using K past samples $\{\boldsymbol{x}_{t-\Delta-k}\}_{k=1}^{K}$. Note that the Δ most recent samples $\{\boldsymbol{x}_{t-i}\}_{i=1}^{\Delta}$ are excluded from the prediction to avoid whiting out the inherent correlation of the desired signal. The delay Δ is set equivalent to the duration of the inherent correlation, namely several tens of milliseconds. The case $\Delta = 0$ corresponds to the classical autoregressive model with no delay. We define $\boldsymbol{x}_t = 0$ for $t \leq 0$.

As in [4, 11], we model $\boldsymbol{s}_t$ by $\boldsymbol{s}_t \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_{M\times 1}, \boldsymbol{\Phi}_t^{\boldsymbol{s}})$ with $\boldsymbol{\Phi}_t^{\boldsymbol{s}} = \phi_t^s \boldsymbol{B}$. Here, $\boldsymbol{0}_{k\times l}$ denotes the $k \times l$ zero matrix, and $\perp\!\!\!\perp$ statistical independence. The above temporal independence assumption results in such $\{\boldsymbol{G}_k\}_{k=1}^{K}$ that temporally decorrelate $\boldsymbol{s}_t$ as much as possible, except for the inherent speech correlation corresponding to the delay Δ. Furthermore, the time-varying covariance matrix makes the autoregressive model applicable to time-varying speech signals as in [1]. Moreover, as compare to [1], the introduction of the spatial coherence matrix $\boldsymbol{B}$ makes it possible to utilize the spatial information about the desired signal. On the other hand, modeling of diffuse noise is the same as in Section 2.1. That is, $\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{v}_t$ and $\boldsymbol{v}_t \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_{M\times 1}, \boldsymbol{\Phi}_t^{\boldsymbol{v}})$, with $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \in \mathcal{V}$.

For later use, we transform the probabilistic model described in the above into the following augmented form:

$$\boldsymbol{F}^{\mathsf{H}} \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{s}}, \tag{9}$$

$$\tilde{\boldsymbol{s}} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_{MT\times 1}, \tilde{\boldsymbol{\Phi}}^{\boldsymbol{s}}), \tag{10}$$

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{v}}, \tag{11}$$

$$\tilde{\boldsymbol{v}} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}_{MT\times 1}, \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}). \tag{12}$$

Here, the vector $\tilde{\boldsymbol{y}} \in \mathbb{C}^{MT}$ contains the time sequence $(\boldsymbol{y}_t)_{t=1}^{T}$ stacked in the reverse order as $\tilde{\boldsymbol{y}} \triangleq \begin{bmatrix}\boldsymbol{y}_T^{\mathsf{T}} & \boldsymbol{y}_{T-1}^{\mathsf{T}} & \cdots & \boldsymbol{y}_1^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$ with $(\cdot)^{\mathsf{T}}$ denoting transposition. The vectors $\tilde{\boldsymbol{x}}$, $\tilde{\boldsymbol{v}}$, and $\tilde{\boldsymbol{s}}$ are defined similarly. The matrix $\boldsymbol{F} \in \mathbb{C}^{MT\times MT}$ is defined as the block Toeplitz matrix with T^2 blocks of size $M \times M$, where the (i, j)th block equals

$$\begin{cases} \boldsymbol{0}_{M\times M}, & \text{if } i-j \in \{-(T-1), \ldots, -1\} \cup \\ & \{1, \ldots, \Delta\} \cup \{\Delta+K+1, \ldots, T-1\}, \\ \boldsymbol{I}_M, & \text{if } i-j = 0, \\ -\boldsymbol{G}_{i-j-\Delta}, & \text{if } i-j \in \{\Delta+1, \ldots, \Delta+K\}. \end{cases} \tag{13}$$

Here, $\boldsymbol{I}_M$ denotes the $M \times M$ unit matrix. The matrix $\tilde{\boldsymbol{\Phi}}^{\boldsymbol{s}} \in \mathbb{C}^{MT\times MT}$ is defined as the block diagonal matrix whose diagonals are $(\boldsymbol{\Phi}_t^{\boldsymbol{s}})_{t=1}^{T}$ in the reverse temporal order, and $\tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}$ is defined similarly. From (9), we can interpret $\boldsymbol{F}^{\mathsf{H}}$ as a dereverberation operator, and $\boldsymbol{F}^{-\mathsf{H}} \triangleq (\boldsymbol{F}^{\mathsf{H}})^{-1}$ as a reverberation operator. Note here that $\boldsymbol{F}^{\mathsf{H}}$ is invertible due to $\det(\boldsymbol{F}^{\mathsf{H}}) = 1$, which follows from the fact that $\boldsymbol{F}^{\mathsf{H}}$ is upper triangular with diagonal entries all equal to one.

3.2. Estimator of desired signal

In a similar fashion as in Section 2.2, we can derive the MAP estimator of $\tilde{\boldsymbol{s}}$. First, we have $p(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}}) = \mathcal{N}_{\mathbb{C}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ with

$$\tilde{\boldsymbol{\mu}} \triangleq \tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} \big(\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} + \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}\big)^{-1} \tilde{\boldsymbol{y}}, \tag{14}$$

$$\tilde{\boldsymbol{\Sigma}} \triangleq \tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} \big(\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} + \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}\big)^{-1} \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}. \tag{15}$$

Here, $\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} \triangleq \boldsymbol{F}^{-\mathsf{H}} \tilde{\boldsymbol{\Phi}}^{\boldsymbol{s}} \boldsymbol{F}^{-1}$ is the spatial covariance matrix of $\boldsymbol{x}_t$. Due to the relation (9), $p(\tilde{\boldsymbol{s}}|\tilde{\boldsymbol{y}}) = \mathcal{N}_{\mathbb{C}}(\boldsymbol{F}^{\mathsf{H}}\tilde{\boldsymbol{\mu}}, \boldsymbol{F}^{\mathsf{H}}\tilde{\boldsymbol{\Sigma}}\boldsymbol{F})$. Therefore, the MAP estimator is given by

$$\underbrace{\boldsymbol{F}^{\mathsf{H}}}_{\text{dereverberation}} \underbrace{\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} \big(\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} + \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}\big)^{-1}}_{\text{denoising}} \tilde{\boldsymbol{y}}. \tag{16}$$

The part labeled "denoising" in (16) is a multichannel Wiener filter generalized to the reverberant case. Indeed, let $\boldsymbol{F} = \boldsymbol{I}_{MT}$, and (16) reduces to the standard multichannel Wiener filter (4). Hence, (16) is the cascade of denoising with the generalized multichannel Wiener filter, and dereverberation with the dereverberation operator $\boldsymbol{F}^{\mathsf{H}}$.

3.3. Parameter estimation algorithm

As in Section 2.3, we can derive a parameter estimation algorithm based on the EM by regarding $\tilde{\boldsymbol{x}}$ as a hidden variable. One iteration of the resulting algorithm is described below, where $\tilde{\boldsymbol{G}} \triangleq \begin{bmatrix}\boldsymbol{G}_1^{\mathsf{T}} & \boldsymbol{G}_2^{\mathsf{T}} & \cdots & \boldsymbol{G}_K^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$. The algorithm is applied to each frequency bin.

[E-step]

1. $\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} \big(\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} + \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}\big)^{-1} \tilde{\boldsymbol{y}}$.
2. $\tilde{\boldsymbol{\Sigma}} \leftarrow \tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} \big(\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}} + \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}\big)^{-1} \tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}$.
3. Update $\boldsymbol{\mu}_t$ so that $\tilde{\boldsymbol{\mu}} = \begin{bmatrix}\boldsymbol{\mu}_T^{\mathsf{T}} & \boldsymbol{\mu}_{T-1}^{\mathsf{T}} & \cdots & \boldsymbol{\mu}_1^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$.
4. Update $\boldsymbol{\Sigma}_{tu}$ so that

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma}_{TT} & \boldsymbol{\Sigma}_{T,T-1} & \cdots & \boldsymbol{\Sigma}_{T1} \\ \boldsymbol{\Sigma}_{T-1,T} & \boldsymbol{\Sigma}_{T-1,T-1} & \cdots & \boldsymbol{\Sigma}_{T-1,1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{1T} & \boldsymbol{\Sigma}_{1,T-1} & \cdots & \boldsymbol{\Sigma}_{11} \end{bmatrix}.$$

5. $\boldsymbol{\Psi}_{tu} \leftarrow \boldsymbol{\mu}_t \boldsymbol{\mu}_u^{\mathsf{H}} + \boldsymbol{\Sigma}_{tu}$.

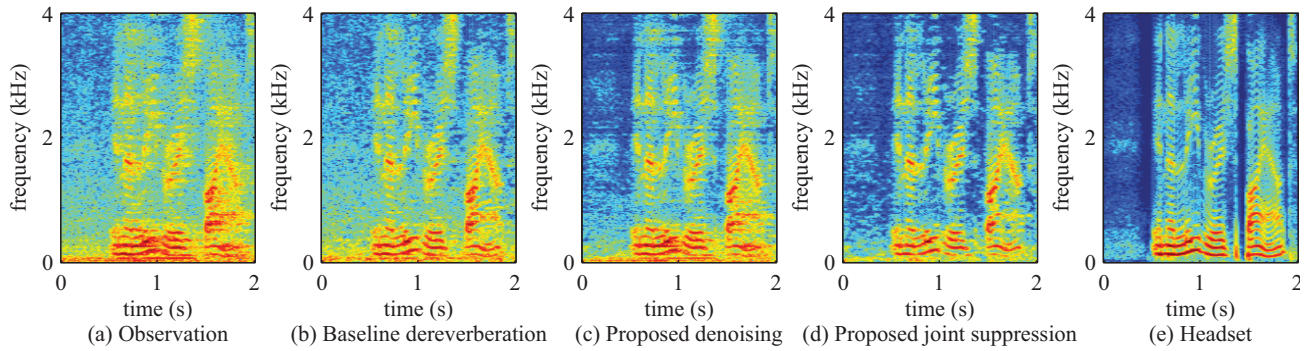

Fig. 1. Spectrograms of the experimental results. (a) Observation at the first microphone; (b) baseline dereverberation; (c) proposed denoising based on the integrated probabilistic model in Section 3; (d) proposed joint suppression based on the integrated probabilistic model; (e) headset recording as reference.

6. $\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{s}} \leftarrow \boldsymbol{\Psi}_{tt} - \tilde{\boldsymbol{G}}^{\mathsf{H}} \begin{bmatrix} \boldsymbol{\Psi}_{t-\Delta-1,t} \\ \vdots \\ \boldsymbol{\Psi}_{t-\Delta-K,t} \end{bmatrix}$
$+\tilde{\boldsymbol{G}}^{\mathsf{H}} \begin{bmatrix} \boldsymbol{\Psi}_{t-\Delta-1,t-\Delta-1} & \cdots & \boldsymbol{\Psi}_{t-\Delta-1,t-\Delta-K} \\ \vdots & & \vdots \\ \boldsymbol{\Psi}_{t-\Delta-K,t-\Delta-1} & \cdots & \boldsymbol{\Psi}_{t-\Delta-K,t-\Delta-K} \end{bmatrix} \tilde{\boldsymbol{G}}$
$- \begin{bmatrix} \boldsymbol{\Psi}_{t-\Delta-1,t}^{\mathsf{H}} & \cdots & \boldsymbol{\Psi}_{t-\Delta-K,t}^{\mathsf{H}} \end{bmatrix} \tilde{\boldsymbol{G}}$.
7. $\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}} \leftarrow (\boldsymbol{y}_t - \boldsymbol{\mu}_t)(\boldsymbol{y}_t - \boldsymbol{\mu}_t)^{\mathsf{H}} + \boldsymbol{\Sigma}_{tt}$.

[M-step]

1. $\phi_t^s \leftarrow \frac{1}{M} \mathrm{Tr}\left[\boldsymbol{B}^{-1} \hat{\boldsymbol{\Phi}}_t^{\boldsymbol{s}}\right]$.
2. $\boldsymbol{B} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\phi_t^s} \hat{\boldsymbol{\Phi}}_t^{\boldsymbol{s}}$.
3. $\boldsymbol{\Phi}_t^{\boldsymbol{s}} \leftarrow \phi_t^s \boldsymbol{B}$.
4. $\boldsymbol{\Phi}_t^{\boldsymbol{v}} \leftarrow \mathcal{P}\left[\hat{\boldsymbol{\Phi}}_t^{\boldsymbol{v}}\right]$.
5. $\boldsymbol{D} \leftarrow \sum_{t=1}^{T} \frac{1}{\phi_t^s} \begin{bmatrix} \boldsymbol{\Psi}_{t-\Delta-1,t-\Delta-1} & \cdots & \boldsymbol{\Psi}_{t-\Delta-1,t-\Delta-K} \\ \vdots & & \vdots \\ \boldsymbol{\Psi}_{t-\Delta-K,t-\Delta-1} & \cdots & \boldsymbol{\Psi}_{t-\Delta-K,t-\Delta-K} \end{bmatrix}$.
6. $\boldsymbol{N} \leftarrow \sum_{t=1}^{T} \frac{1}{\phi_t^s} \begin{bmatrix} \boldsymbol{\Psi}_{t-\Delta-1,t} \\ \vdots \\ \boldsymbol{\Psi}_{t-\Delta-K,t} \end{bmatrix}$.
7. $\tilde{\boldsymbol{G}} \leftarrow \boldsymbol{D}^{-1} \boldsymbol{N}$.
8. Update $\tilde{\boldsymbol{\Phi}}^{\boldsymbol{x}}$ and $\tilde{\boldsymbol{\Phi}}^{\boldsymbol{v}}$ by the definitions using $\boldsymbol{\Phi}_t^{\boldsymbol{s}}$, $\boldsymbol{\Phi}_t^{\boldsymbol{v}}$, and $\tilde{\boldsymbol{G}}$.

4. EXPERIMENTAL VERIFICATION

We have conducted an experiment to examine the proposed joint suppression method in terms of denoising and dereverberation capability. We took observed signals from the REVERB challenge database [19], specifically the real-world eight-channel recording `AMI_WSJ20-Array1-*_T10c020c.wav` in `RealData`. The recording contains reverberation of $\mathrm{RT}_{60} \sim 0.7$ s and some noise, and was truncated at 2 s. The sampling frequency was 16 kHz. As a baseline method, we neglected noise in the proposed joint suppression method by replacing 1–3 of the E-step with trivial rules $\boldsymbol{\mu}_t \leftarrow \boldsymbol{y}_t$ and $\boldsymbol{\Sigma}_{tu} \leftarrow \boldsymbol{0}_{M \times M}$. The resulting algorithm is identical to the conventional frequency-domain variance-normalized delayed linear prediction (NDLP) [1], except that the former takes spatial signal correlation into account and estimates the signal variance using all channels. The frame length and hop for STFT were 1024 and 256 points (equivalent to 64 and 16 ms), respectively; the window type Hamming; the prediction order $K = 3$; the prediction delay $\Delta = 3$; the number of iterations 20. We evaluate the following measures as defined in [19]: the cepstrum distance (CD), the log-likelihood ratio (LLR), the frequency-weighted segmental signal-to-noise ratio (FWSegSNR), and the speech-to-reverberation modulation energy ratio (SRMR). These measures were evaluated by using the headset recording `AMI_WSJ20-Headset1_T10c020c.wav` as reference. For CD and LLR, a lower value indicates a better quality.

Table 1. Evaluation using objective measures.

Methods	CD (dB)	LLR	FWSegSNR (dB)	SRMR
(a)	4.2	0.8	−3.5	3.0
(b)	4.2	0.8	−2.6	*5.2*
(c)	*3.6*	**0.6**	*−1.4*	3.3
(d)	**3.4**	*0.7*	**0.0**	**8.3**

Fig. 1 shows the spectrograms of the results. We restrict the frequency range to show to 0–4 kHz, where the difference between the methods was clearest. Compared to the observed signal at the first microphone in Fig. 1(a), the baseline dereverberation method in Fig. 1(b) contains less reverberation, especially around 1 s and 1.7 s; but almost as much noise. Contrarily, the proposed denoising based on joint modeling in Fig. 1(c) contains less noise than Fig. 1(a) especially during the speech absence in the beginning; but almost as much reverberation. Finally, proposed joint suppression based on joint modeling in Fig. 1(d) suppressed both effectively.

Table 1 shows the evaluated objective measures. The labels (a)–(d) are defined as per Fig. 1. The boldface and the Italic figures represent the best and the second best scores for each criterion, respectively. We see that the proposed joint suppression gave the best score for CD, FWSegSNR, and SRMR, and the second best score for LLR. These results show the effectiveness of the proposed method.

5. CONCLUSION

We discussed probabilistic integration of diffuse noise suppression and dereverberation. The prediction gain matrices and the signal and noise covariance matrices are estimated jointly by maximizing the unified likelihood function through the EM algorithm. In the experiment, the proposed joint suppression method outperformed the baseline method in terms of CD, FWSegSNR, and SRMR. The future work includes such issues as source movement, multiple sources, and an unknown and/or time-varying source number.

6. REFERENCES

[1] 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.

[2] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. ASLP*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[3] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Enhancement of noisy reverberant speech by linear filtering followed by nonlinear noise suppression," in *Proc. IWAENC*, Sep. 2008.

[4] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with Kalman smoother," in *Proc. ICASSP*, May 2013, pp. 7447–7451.

[5] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment," in *Proc. Interspeech*, Aug. 2007, pp. 854–857.

[6] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Adv. Sig. Process.*, vol. 2003, pp. 1074–1090, Oct. 2003.

[7] E.A.P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. ICASSP*, Mar. 2005, vol. 4, pp. 173–176.

[8] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P.A. Naylor and N.D. Gaubitch, Eds., pp. 311–385. Springer, London, 2010.

[9] 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.

[10] N. Ito, *Robust Microphone Array Signal Processing against Diffuse Noise*, Ph.D. thesis, the University of Tokyo, 2012.

[11] 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.

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

[13] 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.

[14] 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.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[16] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multi-channel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, May 2013.

[17] 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.

[18] 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.

[19] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E.A.P. Habets, R. Häb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*, Oct. 2013.