Generalization of Multi-Channel Linear Prediction Methods for Blind MIMO Impulse Response Shortening

Takuya Yoshioka, *Member, IEEE*, and Tomohiro Nakatani, *Senior Member, IEEE*

***Abstract*—The performance of many microphone array processing techniques deteriorates in the presence of reverberation. To provide a widely applicable solution to this longstanding problem, this paper generalizes existing dereverberation methods using subband-domain multi-channel linear prediction filters so that the resultant generalized algorithm can blindly shorten a multiple-input multiple-output (MIMO) room impulse response between a set of unknown number of sources and a microphone array. Unlike existing dereverberation methods, the presented algorithm is developed without assuming specific acoustic conditions, and provides a firm theoretical underpinning for the applicability of the subband-domain multi-channel linear prediction methods. The generalization is achieved by using a new cost function for estimating the prediction filter and an efficient optimization algorithm. The proposed generalized algorithm makes it easier to understand the common background underlying different dereverberation methods and future technical development. Indeed, this paper also derives two alternative dereverberation methods from the proposed algorithm, which are advantageous in terms of computational complexity. Experimental results are reported, showing that the proposed generalized algorithm effectively achieves blind MIMO impulse response shortening especially in a mid-to-high frequency range.**

***Index Terms*—Blind equalization, dereverberation, linear prediction MIMO.**

I. Introduction

THE efficacy of many microphone array processing techniques, such as sound source localization and beamforming, is degraded in reverberant environments because they rely on the assumption that there is little reverberation. Although several methods have been proposed to circumvent this problem [1], [2], they can be used only for specific applications.

To provide a widely applicable solution to this problem, this paper presents an algorithm for shortening a multiple-input multiple-output (MIMO) room impulse response between a set of unknown number of sources and the microphone array with a blind processing approach, which we call *blind MIMO impulse response shortening*. In other words, we aim at transforming reverberant speech signals observed by a microphone array into the same number of dereverberated signals without using knowledge of the room impulse response. The source localization and beamforming performance will be improved in reverberant environments by preprocessing microphone signals with blind MIMO impulse response shortening.

A. Related Work

Blind MIMO impulse response shortening is a dereverberation process that has the following four properties, which must be satisfied when this technique is employed as a preprocessor for other microphone array systems.

(c1) The process produces the same number of dereverberated signals as microphones. This property is required since the performance of many microphone array processing techniques increases with the number of microphones.

(c2) The process does not require knowledge of the number of sound sources. This property is required since it is often difficult to estimate the correct number of sources with blind processing.

(c3) The dereverberated signals are linear convolutive mixtures of the source signals using truncated versions of the original room impulse responses while the dereverberated signals are allowed to be superimposed by statistically independent noise. This property is required since almost all microphone array techniques assume that input signals are linear convolutive mixtures of source signals plus some additive noise [3].

(c4) The process conserves the time differences of arrival (TDOAs) at microphone positions. This property is required since many source localization algorithms are explicitly or implicitly based on TDOAs at microphone positions [1].

Let us review existing dereverberation methods with the above four properties in mind. The widely employed spectral subtraction-based approach does not have property (c3) since it corrupts the linear relationship between source and microphone signals [4], [5]. On the other hand, blind deconvolution, another conventional approach to dereverberation, fulfills (c3) [6]. However, many blind deconvolution algorithms are based on the assumption that source signals are produced by independent and identically distributed processes while speech signals are nonstationary and correlated in time, resulting in the "overwhitening" of speech signals. Although there are several methods for avoiding overwhitening, most of them produce only one dereverberated signal, thus failing to satisfy (c1) [7], [8].

Manuscript received April 10, 2012; revised July 10, 2012; accepted July 11, 2012. Date of publication July 31, 2012; date of current version October 09, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

The authors are with the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kyoto 619-0237, Japan (e-mail: yoshioka.takuya@lab.ntt.co.jp; nakatani.tomohiro@lab.ntt.co.jp).

Digital Object Identifier 10.1109/TASL.2012.2210879

The Weighted Prediction Error (WPE) method [9] can be extended to perform blind MIMO impulse response shortening, although it was originally proposed for single-source dereverberation. This method uses subband-domain multi-channel linear prediction filters, allowing us to conserve TDOAs at microphone positions (condition (c4)) as well as the linear relationship between sources and microphones (condition (c3)). In addition, it can yield as many dereverberated signals as microphones by predicting each microphone signal separately (condition (c1)). Since it does not require explicit knowledge of the number of sources (condition (c2)), the WPE method satisfies all the four properties. However, the cost function for estimating the prediction filters was derived with the maximum likelihood (ML) method, assuming that there is only one speaker in a room and that there is no background noise. Although [10] reports that the WPE method can perform dereverberation to a limited extent even when there are multiple speakers, no mathematical justification has ever been presented as to why the method works under the multi-speaker situation.

Joint dereverberation and source separation methods such as TRINICON [11] and the conditional separation and dereverberation (CSD) method [10] are also relevant to blind MIMO impulse response shortening. These methods aim at cancelling a MIMO room impulse response rather than shortening it, and therefore requirement (c4) is not necessarily satisfied. In addition, they need to know the correct number of sources, and thus fail to satisfy (c2). Let us focus on the CSD method since it is highly relevant to the algorithm presented in this paper. This method uses dereverberation and source separation filters connected in tandem, where the dereverberation filter takes the form of subband-domain multi-channel linear prediction. Both filters are estimated jointly by minimizing a predetermined cost function. As with the WPE method, the cost function was derived by using the ML method on the assumption that there is no background noise and that we know the correct number of sources. These assumptions mean that the CSD method does not comply with (c2). Even if the correct source number is known, the dereverberation filter converts reverberant signals into only the same number of dereverberated signals as sources. Therefore, the CSD method may be used to perform blind MIMO impulse response shortening only when we have the same number of microphones as sources, otherwise it does not satisfy (c1).

B. Contribution of This Paper

We can see from the above review that, although promising, the existing methods using subband-domain multi-channel linear prediction have limitations when viewed as blind MIMO impulse response shortening methods. The root of these limitations lies in the fact that the cost functions are derived with the ML method using certain assumptions about the number of sources and the presence/absence of background noise.

To extend the applicability of the subband-domain multi-channel linear prediction approach, we present a novel cost function for estimating the prediction filters without assuming specific acoustic conditions. The proposed cost function is developed to account for the following three observations. First, non-reverberant speech signals are almost uncorrelated in time in the subband domain. Secondly, we do not need to make output signals spatially uncorrelated because our goal is to shorten MIMO room impulse responses and not to separate source signals. Thirdly, exploiting the nonstationarity of speech signals plays a critical role in achieving dereverberation [12]. With these things in mind, the proposed cost function measures the degree of temporal correlation in non-stationary output signals (i.e., the correlation between output signal values at different points in time), ignoring spatial correlation (i.e., the correlation between different output channels). In addition, we propose an efficient algorithm for minimizing the proposed cost function. Interestingly, this algorithm includes the WPE and CSD methods as special cases, thus revealing the common mathematical background underlying these existing methods. For this reason, we call the proposed algorithm the Generalized WPE (GWPE) algorithm.

The remainder of this paper is organized as follows. Section II defines the problem addressed in this work and briefly reviews the WPE method. Section III proposes the cost function for prediction filter estimation and formulates the optimization problem to be solved. Section IV describes the proposed efficient optimization algorithm. This method requires a sequence of spatial correlation matrices of dereverberated signals, and Section V discusses several approaches for estimating this sequence. Section VI reports our experimental results, and Section VII concludes this paper.

II. Blind MIMO Impulse Response Shortening

We begin by formulating blind MIMO impulse response shortening. For notational simplicity, in the following, we use superscript $*$ to denote the conjugate transpose of a complex vector or matrix, the transpose of a real vector or matrix, and the conjugate of a complex number without making any distinction between them.

A. Problem Formulation

A general acoustic system in a room is described as follows. Let M and N be the numbers of speakers and microphones, respectively. Also, let $s^m[k]$ $(1 \leq m \leq M)$, $v^n[k]$, and $y^n[k]$ $(1 \leq n \leq N)$ denote the speech signal of the mth speaker, the noise at the nth microphone, and the speech signal observed by the nth microphone, respectively, where k is a fullband-domain time index. We assume that the noise is statistically independent of the clean speech signals. The microphone signals are generated according to the following equation:

$$\boldsymbol{y}[k] = \sum_{\kappa=0}^{J-1} \boldsymbol{H}^*[\kappa]\boldsymbol{s}[k-\kappa] + \boldsymbol{v}[k], \tag{1}$$

where $\boldsymbol{y}[k] = [y^1[k], \ldots, y^N[k]]^*$, $\boldsymbol{s}[k] = [s^1[k], \ldots, s^M[k]]^*$, $\boldsymbol{v}[k] = [v^1[k], \ldots, v^N[k]]^*$, and $\{\boldsymbol{H}[\kappa]\}_{0 \leq \kappa \leq J-1}$ is a MIMO room impulse response of order J between the speakers and microphones. $\boldsymbol{H}[\kappa]$ is an M-by-N matrix defined as

$$\boldsymbol{H}[\kappa] = \begin{bmatrix} h^{1,1}[\kappa] & \cdots & h^{N,1}[\kappa] \\ \vdots & \ddots & \vdots \\ h^{1,M}[\kappa] & \cdots & h^{N,M}[\kappa] \end{bmatrix}, \tag{2}$$

where $\{h^{n,m}[\kappa]\}_{0\leq\kappa\leq J-1}$ is the room impulse response from the mth speaker to the nth microphone. The signals and the room impulse response can take any real values.

To reduce the filter order, we employ a subband signal processing scheme by rewriting the acoustic system (1) as

$$\boldsymbol{y}_l(t) = \sum_{\tau=0}^{J_l-1} \boldsymbol{H}_l^*(\tau)\boldsymbol{s}_l(t-\tau) + \boldsymbol{v}_l(t), \tag{3}$$

where $\boldsymbol{y}_l(t)$, $\boldsymbol{s}_l(t)$, $\boldsymbol{v}_l(t)$, and $\boldsymbol{H}_l(\tau)$ are the complex-valued subband-domain counterparts of microphone signal vector $\boldsymbol{y}[k]$, clean speech signal vector $\boldsymbol{s}[k]$, noise vector $\boldsymbol{v}[k]$, and MIMO room impulse response $\boldsymbol{H}[\kappa]$, respectively, while J_l corresponds to room impulse response order J. Subscript l denotes a subband index and t is a subband-domain time index. Note that, in (3), the noise vector $\boldsymbol{v}_l(t)$ includes both the background noise and the effect of speech energy leakage over adjacent subbands.

We define blind MIMO impulse response shortening as finding a MIMO linear filter for each subband l that virtually truncates $\{\boldsymbol{H}_l(\tau)\}_{0\leq\tau\leq J_l-1}$ up to Δ taps, where Δ is an arbitrary integer. When we employ the multi-channel linear prediction approach, our goal can be specifically described as follows.

Definition 1: Let us define N-dimensional vector $\tilde{\boldsymbol{y}}_l(t) = \left[\tilde{y}_l^{1*}(t),\ldots,\tilde{y}_l^{N*}(t)\right]^*$ and $\boldsymbol{x}_l(t) = \left[x_l^{1*}(t),\ldots,x_l^{N*}(t)\right]^*$ as a linear prediction of $\boldsymbol{y}_l(t)$ and the corresponding prediction error vector, respectively, i.e.,

$$\tilde{\boldsymbol{y}}_l(t) = \sum_{\tau=\Delta}^{\Delta+K_l-1} \boldsymbol{G}_l^*(\tau)\boldsymbol{y}_l(t-\tau) \tag{4}$$

$$\boldsymbol{x}_l(t) = \boldsymbol{y}_l(t) - \tilde{\boldsymbol{y}}_l(t), \tag{5}$$

where $\{\boldsymbol{G}_l(\tau)\}_{\Delta\leq\tau\leq\Delta+K_l-1}$ is the N-by-N (i.e., MIMO) prediction filter and K_l is the prediction order. The problem addressed in this paper is to adjust $\mathcal{G}_l = \{\boldsymbol{G}_l(\tau)\}_{\Delta\leq\tau\leq\Delta+K_l-1}$, given T samples of the subband microphone signals $(\boldsymbol{y}_l(t))_{t\in\mathcal{T}}$, where $\mathcal{T} = \{t\}_{1\leq t\leq T}$, so that $\boldsymbol{x}_l(t)$ can be represented as

$$\boldsymbol{x}_l(t) = \sum_{\tau=0}^{\Delta-1} \boldsymbol{H}_l^*(\tau)\boldsymbol{s}_l(t-\tau) + \text{noise}, \tag{6}$$

where the noise in each output channel is independent of the clean speech signal vector $\boldsymbol{s}_l(t)$. □

The existence of a desired prediction filter, shortening the MIMO room impulse response up to Δ taps, is easily confirmed based on the theory of multi-channel linear prediction. Specifically, under a few conditions, there exists $\mathcal{G}_l$ value that produces $\boldsymbol{x}_l(t)$ satisfying the following relationship:

$$\boldsymbol{x}_l(t) = \sum_{\tau=0}^{\Delta-1} \boldsymbol{H}_l^*(\tau)\boldsymbol{s}_l(t-\tau) + \tilde{\boldsymbol{v}}_l(t), \tag{7}$$

where $\tilde{\boldsymbol{v}}_l(t)$ is given by

$$\tilde{\boldsymbol{v}}_l(t) = \boldsymbol{v}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \boldsymbol{G}_l^*(\tau)\boldsymbol{v}_l(t-\tau) \tag{8}$$

and thus statistically independent of the clean speech signal vector $\boldsymbol{s}_l(t)$ [13].

B. Weighted Prediction Error Method and its Limitations

The question is how to estimate such a desired prediction filter. We review the multi-channel linear prediction and WPE methods, which partly solve this problem, and point out their limitations in order to make our motivation clear.

The basic multi-channel linear prediction method solves the problem when $\boldsymbol{s}_l(t)$ is stationary and uncorrelated over time and $\boldsymbol{v}_l(t) = 0$. Under these conditions, the desired prediction filter can be estimated by minimizing the sum of the squared prediction errors. In other words, the cost function employed by the multi-channel linear prediction method is given by

$$F_{\mathrm{PE}}(\mathcal{G}_l) = \sum_{t\in\mathcal{T}} \left\| \boldsymbol{y}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \boldsymbol{G}_l^*(\tau)\boldsymbol{y}_l(t-\tau) \right\|^2, \tag{9}$$

where $||\cdot||^2$ represents the vector norm. Thanks to the decimation performed in the course of the subband decomposition, it may be reasonable to assume that $\boldsymbol{s}_l(t)$ is temporally uncorrelated. However, minimization of the sum of the squared prediction errors does not yield a good $\boldsymbol{G}_l(\tau)$ estimate since speech signals are nonstationary and do not meet the stationarity assumption.

To account for the nonstationarity, the WPE method makes two assumptions. The first assumption is that, for each m, the mth clean speech signal $s_l^m(t)$ is sampled from a complex normal distribution with mean 0 and *time-varying* variance $\lambda_l^m(t)$, which corresponds to the expected value of $|s_l^m(t)|^2$. The second assumption is that there is one speaker (i.e., $M = 1$) and no background noise in a room. The second assumption causes (7) to degenerate to the following equation:

$$x_l^n(t) = \sum_{\tau=0}^{\Delta-1} h_l^{n,1*}(\tau)s_l^1(t-\tau), \quad 1\leq n\leq N, \tag{10}$$

which indicates that the prediction error for each microphone, $x_l^n(t)$, is the convolution of $s_l^1(t)$ and the initial Δ-tap part of $h_l^{n,1}(\tau)$. Therefore, a dereverberated signal can be obtained by calculating only one of the M-channel prediction errors. For example, we may use the first-channel prediction errors as

$$x_l^1(t) = y_l^1(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \boldsymbol{g}_l^{1*}(\tau)\boldsymbol{y}_l(t-\tau), \tag{11}$$

where $\boldsymbol{g}_l^1(\tau)$ is the first column of $\boldsymbol{G}_l(\tau)$. Hence, we only need to estimate $\mathcal{G}_l^1 = \left\{\boldsymbol{g}_l^1(\tau)\right\}_{\Delta\leq\tau\leq\Delta+K_l-1}$ instead of the whole MIMO filter $\mathcal{G}_l$.

By using the ML method with these assumptions, the cost function of the WPE method is obtained as

$$F_{\mathrm{WPE}}\left(\mathcal{G}_l^1\right) = \sum_{t\in\mathcal{T}} \frac{\left\| y_l^1(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \boldsymbol{g}_l^{1*}(\tau)\boldsymbol{y}_l(t-\tau) \right\|^2}{\lambda_l^1(t)}. \tag{12}$$

F_{WPE} is different from F_{PE} in that each prediction error is weighted by the reciprocal of the time-varying clean speech signal power $\lambda_l^1(t)$. This weighting operation equalizes the degree to which each prediction error contributes to the cost function and has been shown to play a critical role in dereverberation [14]. In practice, each $\lambda_l^1(t)$ value is estimated along with $\mathcal{G}_l^1$ by iterating two optimization processes: one optimizes $\mathcal{G}_l^1$ for a fixed $\left\{\lambda_l^1(t)\right\}_{t\in\mathcal{T}}$ estimate; the other optimizes $\left\{\lambda_l^1(t)\right\}_{t\in\mathcal{T}}$ for a fixed $\mathcal{G}_l^1$ estimate. Finally, the whole prediction filter $\mathcal{G}_l$ is constructed as

$$\boldsymbol{G}_l(\tau) = \left[\boldsymbol{g}_l^1(\tau),\ldots,\boldsymbol{g}_l^N(\tau)\right], \tag{13}$$

where $\boldsymbol{g}_l^2(\tau),\ldots,\boldsymbol{g}_l^N(\tau)$ are obtained by repeating the same procedure for channels $2,\ldots,M$ separately.

As such, the WPE method lacks a theoretical background when it is applied to blind MIMO impulse response shortening under noisy and/or multi-speaker situations since the cost function F_{WPE} assumes a single speaker and the absence of noise. This motivates development of a new cost function for estimating the desired prediction filter without assuming specific acoustic conditions.

Before proceeding, let us make a preliminary assumption that will form the basis of the subsequent derivation. We assume that a speech signal is generated by a random process. So, let $(\boldsymbol{Y}_l(t))_{t\in\mathcal{T}}$ denote a sequence of random variables and let us assume that $(\boldsymbol{y}_l(t))_{t\in\mathcal{T}}$ is its realization. Then, $(\boldsymbol{x}_l(t))_{t\in\mathcal{T}}$ may be considered to be a realization of the random variable sequence $(\boldsymbol{X}_l(t))_{t\in\mathcal{T}}$ given by

$$\boldsymbol{X}_l(t) = \sum_{\tau=\Delta}^{\Delta+K_l-1} \boldsymbol{G}_l^*(\tau)\boldsymbol{Y}_l(t-\tau). \tag{14}$$

We allow $(\boldsymbol{X}_l(t))_{t\in\mathcal{T}}$ and $(\boldsymbol{Y}_l(t))_{t\in\mathcal{T}}$ to be nonstationary to account for the nonstationarity of speech signals.

III. New Cost Function for Prediction Filter Estimation

The goal in this section is to find a cost function

$$F(\mathcal{G}_l) = C(\boldsymbol{X}_l(1),\ldots,\boldsymbol{X}_l(T)), \tag{15}$$

where $C(\cdot)$ measures the degree to which a realization of $(\boldsymbol{X}_l(t))_{t\in\mathcal{T}}$ is reverberant. Once such a cost function is given, we will obtain an estimate of $\mathcal{G}_l = \{\boldsymbol{G}_l(\tau)\}_{\Delta\leq\tau\leq\Delta+K_l-1}$ by minimizing the cost function.

We derive such a cost function by leveraging the temporal correlation characteristics of speech signals. A clean speech signal has autocorrelation coefficients of nearly zero for time lags greater than tens of milliseconds while a reverberant signal has large autocorrelation coefficients for those large time lags. Accordingly, we can reasonably assume that a set of dereverberated speech signals, which may be still corrupted by additive noise, has zero temporal correlation coefficients except for the zeroth lag in each subband regardless of the number of sources (as long as we choose an appropriate subband width and decimation factor). On the other hand, these dereverberated speech signals may be correlated in space because blind impulse response shortening is not intended to separate individual speech signals. On the basis of these observations, we conjecture that a prediction filter that shortens MIMO room impulse responses can be obtained by making the output random vector sequence $(\boldsymbol{X}_l(t))_{t\in\mathcal{T}}$ as temporally uncorrelated as possible without enforcing spatial uncorrelatedness. In other words, we minimize the correlation between $\boldsymbol{X}_l(1),\ldots,\boldsymbol{X}_l(T)$ without uncorrelating the vector components $X_l^1(t),\ldots,X_l^N(t)$ for each t.

A. Correlation Measure for Multivariate Random Variables

To give shape to the idea sketched above, we need a measure of the correlation between multivariate random variables $\boldsymbol{X}_l(1),\ldots,\boldsymbol{X}_l(T)$. We propose a new correlation measure, which is called the *Hadamard-Fischer (HF) mutual correlation*[1] and applicable to multivariate random variables while correlation measures used for source separation (e.g., [16]) assume univariate variables.

Definition 2: Let $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_N$ be complex-valued multivariate random vectors and $\boldsymbol{U}$ be the vector wherein these random vectors are stacked as $\boldsymbol{U} = [\boldsymbol{U}_1^*,\ldots,\boldsymbol{U}_N^*]^*$. The HF mutual correlation between these random vectors is given by

$$C_{\mathrm{HF}}(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_N) = \frac{1}{N}(\sum_{n=1}^{N}\log\left(\det E\left(\boldsymbol{U}_n\boldsymbol{U}_n^*\right)\right) - \log\left(\det E\left(\boldsymbol{U}\boldsymbol{U}^*\right)\right)), \tag{16}$$

where operator $\det$ calculates a matrix determinant while $E(\cdot)$ denotes an expectation operator.

The following theorem holds with respect to the HF mutual correlation. (This theorem can be readily proven by using the Hadamard-Fischer inequality [17], [18] and that is why we call $C_{\mathrm{HF}}(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_N)$ the HF mutual correlation. See the Appendix for the proof.)

Theorem 1: The HF mutual correlation is nonnegative and zero if and only if all of $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_N$ are mutually uncorrelated. In other words, we have

$$C_{\mathrm{HF}}(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_N) \geq 0, \tag{17}$$

where the equality holds if and only if $E\left(\boldsymbol{U}_m\boldsymbol{U}_n^*\right) = \boldsymbol{O}$, where $\boldsymbol{O}$ denotes a zero matrix, for all combinations of m and n values satisfying $1 \leq m \neq n \leq N$.

1) Note: The equality condition allows the elements of vector $\boldsymbol{U}_n$ to be mutually correlated. □

The above theorem suggests that $C_{\mathrm{HF}}(\boldsymbol{X}_l(1),\ldots,\boldsymbol{X}_l(T))$ measures the degree of temporal correlation of $(\boldsymbol{X}_l(t))_{t\in\mathcal{T}}$. Hence, we will use the HF mutual correlation in (15). It is worthwhile noting that the HF mutual correlation is equivalent to mutual information if all of $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_N$ are normally distributed. The fact that the HF mutual correlation does not require the normality condition is of theoretical importance because a speech signal has a super-Gaussian distribution [19].

[1] The HF mutual correlation is identical to the correlation measure used for source separation in [15]. However, the cited paper states that the non-negativity of the cost function can be derived from Oppenheim's inequality and provides no proof. Thus, for completeness, we define the HF mutual correlation and shows its appropriateness as a measure of correlation.

B. Optimization Problem Statement

Now, let us turn our attention back to the blind MIMO impulse response shortening problem. Here, we present an optimization problem for prediction filter $\mathcal{G}_l = \{\boldsymbol{G}_l(\tau)\}_{\Delta \le \tau \le \Delta + K_l - 1}$.

Using the HF mutual correlation in (15), our goal is defined as to minimize the following cost function with respect to $\mathcal{G}_l$:

$$\begin{aligned} F(\mathcal{G}_l) &= C_{\text{HF}}(\boldsymbol{X}_l(1), \ldots, \boldsymbol{X}_l(T)) \\ &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \log \left(\det E \left(\boldsymbol{X}_l(t) \boldsymbol{X}_l^*(t) \right) \right) \\ &\quad - \log \left(\det E \left(\boldsymbol{X}_l \boldsymbol{X}_l^* \right) \right), \end{aligned} \tag{18}$$

where $\boldsymbol{X}_l$ is the vector wherein $\boldsymbol{X}_l(t)$'s are stacked as $\boldsymbol{X}_l = [\boldsymbol{X}_l^*(T), \ldots, \boldsymbol{X}_l^*(1)]^*$. This cost function takes its minimum when random vector sequence $(\boldsymbol{X}_l(t))_{t \in \mathcal{T}}$ is temporally uncorrelated. As described in the note on Theorem 1, the elements of $\boldsymbol{X}_l(t)$ may be (spatially) correlated with each other at the minimum point.

This cost function reduces to

$$F(\mathcal{G}_l) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \log \left(\det E \left(\boldsymbol{X}_l(t) \boldsymbol{X}_l^*(t) \right) \right). \tag{19}$$

This can be confirmed as follows. $\boldsymbol{X}_l$ is given by $\boldsymbol{X}_l = \boldsymbol{G}_l^* \boldsymbol{Y}_l$, where $\boldsymbol{Y}_l = [\boldsymbol{Y}_l^*(T), \ldots, \boldsymbol{Y}_l^*(1)]^*$ and $\boldsymbol{G}_l$ is a convolution matrix consisting of the impulse response $\{\boldsymbol{I}, \boldsymbol{O}, \ldots, \boldsymbol{O}, -\boldsymbol{G}_l(\Delta), \ldots, -\boldsymbol{G}_l(\Delta + K_l - 1)\}$. Since the determinant of a block triangular matrix is given by the product of the determinants of the diagonal submatrices, we have

$$\begin{aligned} \det E(\boldsymbol{X}_l \boldsymbol{X}_l^*) &= |\det \boldsymbol{G}_l|^2 \det E \left(\boldsymbol{Y}_l \boldsymbol{Y}_l^* \right) \\ &= \det E \left(\boldsymbol{Y}_l \boldsymbol{Y}_l^* \right) \\ &= \text{constant}. \end{aligned} \tag{20}$$

Equations (18) and (20) combine to lead to (19).

Unfortunately, the minimization of the cost function given by (19) has no analytical solution. To estimate the prediction filter $\mathcal{G}_l$ according to (19), we may replace each $E\left(\boldsymbol{X}_l(t)\boldsymbol{X}_l^*(t)\right)$ by a temporally local average of $\boldsymbol{x}_l(t)\boldsymbol{x}_l^*(t)$ and find the $\hat{\mathcal{G}}_l$ that makes the first order derivative of $F(\mathcal{G}_l)$ zero. However, due to the logarithmic function in (19), we cannot find such a $\mathcal{G}_l$ analytically. Since the prediction filter order is usually large, it is impractical to use numerical optimization algorithms such as gradient-based algorithms. Therefore, we need to go one step further to invent a feasible optimization algorithm.

IV. Optimization Algorithm

We employ the auxiliary function approach [20] to derive the proposed algorithm, i.e., the GWPE algorithm, to minimize $F(\mathcal{G}_l)$. The auxiliary function approach enables us to split the optimization problem into two subproblems: one optimizes a prediction filter and the other optimizes a set of newly introduced auxiliary variables. The key to success is that the prediction filter optimization is performed analytically, and thus the computational complexity is low enough to be used in practical scenarios. Section IV-A summarizes the auxiliary function approach. Sections IV-B–IV-D elaborate on the derivation of the proposed algorithm.

A. Auxiliary Function Approach

The basic concept of the auxiliary function approach is to iteratively minimize the upper bound of the cost function, which is easier to minimize than the original cost function. The upper bound is called an auxiliary function and is defined as follows.

Definition 3: $\tilde{f}(\boldsymbol{x}, \boldsymbol{y})$ is an auxiliary function for $f(\boldsymbol{x})$ if

$$f(\boldsymbol{x}) = \min_{\boldsymbol{y}} \tilde{f}(\boldsymbol{x}, \boldsymbol{y}) \text{ for all } \boldsymbol{x}. \tag{21}$$

By using the auxiliary function, we can obtain an optimization algorithm as follows. Suppose that we want to iteratively minimize $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$, starting from the initial guess $\hat{\boldsymbol{x}}$. Let $\tilde{f}(\boldsymbol{x}, \boldsymbol{y})$ be an auxiliary function for $f(\boldsymbol{x})$. With the auxiliary function approach, we iteratively perform a set of update procedures consisting of $\hat{\boldsymbol{y}} = \text{argmin}_{\boldsymbol{y}} \tilde{f}(\hat{\boldsymbol{x}}, \boldsymbol{y})$ and $\hat{\boldsymbol{x}} = \text{argmin}_{\boldsymbol{x}} \tilde{f}(\boldsymbol{x}, \hat{\boldsymbol{y}})$. Based on the definition of the auxiliary function, this update rule never increases the true cost function value. Therefore, if the cost function is lower-bounded, the sequence of the updated estimates converges although global optimality is not guaranteed. One advantage of this approach is that we can construct a simple algorithm that requires no control parameters such as a learning rate if we can find an appropriate auxiliary function.

B. Algorithm Overview

Now, we derive the proposed GWPE algorithm for obtaining the prediction filter $\hat{\mathcal{G}}_l$ that minimizes $F(\mathcal{G}_l)$, given by (19). We find an auxiliary function for $F(\mathcal{G}_l)$ by exploiting the following lemma. (The proof is given in the Appendix.)

Lemma 2: Let $\boldsymbol{U}$ be an N-dimensional multivariate random vector. Then, for all positive definite Hermitian matrices $\boldsymbol{\Lambda}$, we have the following inequality:

$$\log |\det E(\boldsymbol{U}\boldsymbol{U}^*)| \le E(\boldsymbol{U}^* \boldsymbol{\Lambda}^{-1} \boldsymbol{U}) - N + \log(\det \boldsymbol{\Lambda}) \tag{22}$$

with equality if and only if $\boldsymbol{\Lambda}$ is the covariance matrix of $\boldsymbol{U}$ as $\boldsymbol{\Lambda} = E(\boldsymbol{U}\boldsymbol{U}^*)$. □

Based on this lemma, we can deduce the following auxiliary function for $F(\mathcal{G}_l)$.

Theorem 3: $\tilde{F}(\mathcal{G}_l, \mathcal{L}_l)$ defined as follows is an auxiliary function for $F(\mathcal{G}_l)$:

$$\begin{aligned} \tilde{F}(\mathcal{G}_l, \mathcal{L}_l) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \Big(& E\left(\boldsymbol{X}_l^*(t) \boldsymbol{\Lambda}_l(t)^{-1} \boldsymbol{X}_l(t) \right) \\ & -N + \log(\det \boldsymbol{\Lambda}_l(t)) \Big), \end{aligned} \tag{23}$$

where $\boldsymbol{\Lambda}_l(t)$ is a positive definite Hermitian matrix and $\mathcal{L}_l = \{\boldsymbol{\Lambda}_l(t)\}_{t \in \mathcal{T}}$. □

By using this auxiliary function, we can sketch the proposed optimization algorithm as shown in Table I. We found empirically that starting with $\hat{\boldsymbol{G}}_l(\tau) = \boldsymbol{O}$ for all τ values worked very well in most cases. The algorithms for performing steps 2 and 3 in the table are described in subsequent subsections.

TABLE I
SKETCH OF PROPOSED OPTIMIZATION ALGORITHM.

1) Initialize $\hat{\mathfrak{G}}_l$.
2) Compute $\hat{\mathcal{L}}_l = \mathrm{argmin}_{\mathcal{L}_l} \tilde{F}(\hat{\mathfrak{G}}_l, \mathcal{L}_l)$.
3) Compute $\hat{\mathfrak{G}}_l = \mathrm{argmin}_{\mathfrak{G}_l} \tilde{F}(\mathfrak{G}_l, \hat{\mathcal{L}}_l)$.
4) Return to step 2 unless convergence is reached.

C. $\mathcal{L}_l$ Update

Invoking Lemma 2, we see that the minimization of auxiliary function $\tilde{F}(\mathcal{G}_l, \mathcal{L}_l)$ with respect to $\mathcal{L}_l$ for a fixed $\mathcal{G}_l$ estimate is achieved by

$$\hat{\mathbf{\Lambda}}_l(t) = E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right), \tag{24}$$

where $(\hat{\boldsymbol{X}}_l(t))_{t\in\mathcal{T}}$ is the convolution of the prediction error filter estimate and the input $(\boldsymbol{Y}_l(t))_{t\in\mathcal{T}}$ as

$$\hat{\boldsymbol{X}}_l(t) = \boldsymbol{Y}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \hat{\boldsymbol{G}}_l^*(\tau)\boldsymbol{Y}_l(t-\tau). \tag{25}$$

In the following, we call $\hat{\mathbf{\Lambda}}_l(t)$ a spatial correlation marix to highlight the fact that $E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right)$ consists of the (zeroth-lag) auto-and cross-correlations between the dereverberated signals, which characterize the coupling of (spatially distributed) output channels.

In practice, we have to estimate the spatial correlation matrix sequence $\left(E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right)\right)_{t\in\mathcal{T}}$ on the basis of its realization, $(\hat{\boldsymbol{x}}_l(t))_{t\in\mathcal{T}}$, given by

$$\hat{\boldsymbol{x}}_l(t) = \boldsymbol{y}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \hat{\boldsymbol{G}}_l^*(\tau)\boldsymbol{y}_l(t-\tau). \tag{26}$$

We will discuss several approaches for doing this in Section V.

D. $\mathcal{G}_l$ Update

The formula for minimizing $\tilde{F}(\mathcal{G}_l, \mathcal{L}_l)$ while fixing $\mathcal{L}_l$ at $\tilde{\mathcal{L}}_l$ can be easily derived because (23) is quadratic with respect to $\mathcal{G}_l$. Here, we describe only the resultant formula since it is derived in a similar way to the CSD method [10].

Let $\boldsymbol{g}_l$ be a vector in which the columns of all prediction matrices $\{\boldsymbol{G}_l(\tau)\}_{\Delta\le\tau\le\Delta+K_l-1}$ are stacked as

$$\boldsymbol{g}_l = \left[\begin{array}{c} \boldsymbol{g}_l^1(\Delta) \\ \vdots \\ \boldsymbol{g}_l^N(\Delta) \\ \hline \vdots \\ \hline \boldsymbol{g}_l^1(\Delta+K_l-1) \\ \vdots \\ \boldsymbol{g}_l^N(\Delta+K_l-1) \end{array}\right], \tag{27}$$

where $\boldsymbol{g}_l^m(\tau)$ represents the mth column of $\boldsymbol{G}_l(\tau)$ as shown in (13). Then, representing the elementwise complex conjugation of matrix $\boldsymbol{A}$ by $\overline{\boldsymbol{A}}$, the optimal prediction filter that minimizes (23) is obtained by

$$\hat{\overline{\boldsymbol{g}_l}} = \boldsymbol{R}_l^{-1}\boldsymbol{r}_l, \tag{28}$$

TABLE II
DETAILS OF THE PROPOSED OPTIMIZATION ALGORITHM.

1) (*Initialization*) Initialize $\hat{\mathfrak{G}}_l$ as, for instance, $\hat{\boldsymbol{G}}_l(\tau) = \boldsymbol{O}$ for all τ values with $\Delta \le \tau \le \Delta + K_l - 1$.
2) (*Dereverberation*) Compute
$$\hat{\boldsymbol{x}}_l(t) = \boldsymbol{y}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1} \hat{\boldsymbol{G}}_l^*(\tau)\boldsymbol{y}_l(t-\tau) \quad \forall t \in \mathfrak{T}.$$
3) (*Spatial correlation matrix estimation*) Estimate
$$\hat{\mathbf{\Lambda}}_l(t) = E(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)) \quad \forall t \in \mathfrak{T}$$
by using $(\boldsymbol{x}_l(t))_{t\in\mathfrak{T}}$.
4) (*Computation of weighted sample correlation matrix/vector*) Compute
$$\hat{\boldsymbol{R}}_l = \sum_{t\in\mathfrak{T}} \overline{\boldsymbol{\psi}_l(t-\Delta)}\hat{\mathbf{\Lambda}}_l(t)^{-1}\overline{\boldsymbol{\psi}_l^*(t-\Delta)}$$
$$\hat{\boldsymbol{r}}_l = \sum_{t\in\mathfrak{T}} \overline{\boldsymbol{\psi}_l(t-\Delta)}\hat{\mathbf{\Lambda}}_l(t)^{-1}\boldsymbol{y}_l(t).$$
5) (*Prediction filter update*) Compute
$$\hat{\overline{\boldsymbol{g}_l}} = \hat{\boldsymbol{R}}_l^{-1}\hat{\boldsymbol{r}}_l$$
The updated prediction matrices are obtained by rearranging the $\hat{\overline{\boldsymbol{g}_l}}$ entries.
6) (*Convergence check*) Return to step2 unless convergence is reached.

where matrix $\boldsymbol{R}_l$ and vector $\boldsymbol{r}_l$ are given by

$$\boldsymbol{R}_l = \sum_{t\in\mathcal{T}} E\Big(\overline{\mathbf{\Psi}_l(t-\Delta)}\hat{\mathbf{\Lambda}}_l(t)^{-1}\overline{\mathbf{\Psi}_l^*(t-\Delta)}\Big) \tag{29}$$

$$\boldsymbol{r}_l = \sum_{t\in\mathcal{T}} E\Big(\overline{\mathbf{\Psi}_l(t-\Delta)}\hat{\mathbf{\Lambda}}_l(t)^{-1}\boldsymbol{Y}_l(t)\Big), \tag{30}$$

respectively. Here, $\mathbf{\Psi}_l(t)$ is a matrix comprising $\boldsymbol{Y}_l(t), \dots, \boldsymbol{Y}_l(t-K_l+1)$ as

$$\mathbf{\Psi}_l(t) = \left[\tilde{\boldsymbol{Y}}_l^*(t), \dots, \tilde{\boldsymbol{Y}}_l^*(t-K_l+1)\right]^*, \tag{31}$$

where

$$\tilde{\boldsymbol{Y}}_l(t) = \underbrace{\begin{bmatrix} \boldsymbol{Y}_l(t) & & \boldsymbol{O} \\ & \ddots & \\ \boldsymbol{O} & & \boldsymbol{Y}_l(t) \end{bmatrix}}_{N}. \tag{32}$$

In practice, we need to estimate $\boldsymbol{R}_l$ and $\boldsymbol{r}_l$ by using the realization $(\boldsymbol{y}_l(t))_{t\in\mathcal{T}}$ of $(\boldsymbol{Y}_l(t))_{t\in\mathcal{T}}$. We use the sample averages to estimate $\boldsymbol{R}_l$ and $\boldsymbol{r}_l$ as follows:

$$\hat{\boldsymbol{R}}_l = \sum_{t\in\mathcal{T}} \overline{\boldsymbol{\psi}_l(t-\Delta)}\hat{\mathbf{\Lambda}}_l(t)^{-1}\overline{\boldsymbol{\psi}_l^*(t-\Delta)} \tag{33}$$

$$\hat{\boldsymbol{r}}_l = \sum_{t\in\mathcal{T}} \overline{\boldsymbol{\psi}_l(t-\Delta)}\hat{\mathbf{\Lambda}}_l(t)^{-1}\boldsymbol{y}_l(t), \tag{34}$$

where $\boldsymbol{\psi}_l(t)$ is a realization of $\mathbf{\Psi}_l(t)$ and is computed by replacing $\boldsymbol{Y}_l(t), \dots, \boldsymbol{Y}_l(t-K_l+1)$ by their respective realizations $\boldsymbol{y}_l(t), \dots, \boldsymbol{y}_l(t-K_l+1)$ in (31). It may be useful to note that (28) reduces to the covariance method for the standard linear prediction when $M = N = 1$ and $\hat{\mathbf{\Lambda}}_l(t) = 1$ for all t. In light of this, we call $\hat{\boldsymbol{R}}_l$ and $\hat{\boldsymbol{r}}_l$ a weighted sample correlation matrix and a weighted sample correlation vector, respectively.

In summary, the proposed GWPE algorithm can be described as shown in Table II. It should be kept in mind that, in order to implement the GWPE method, we need to define a way of estimating the spatial correlation matrix sequence $(E(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)))_{t\in\mathcal{T}}$ on the basis of the available data $(\hat{\boldsymbol{x}}_l(t))_{t\in\mathcal{T}}$.

V. Spatial Correlation Matrix Estimation

Finally, we address the problem of estimating spatial correlation matrices. A straightforward method using $\hat{\boldsymbol{\Lambda}}_l(t) = E(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)) \approx \sum_{\tau=t-\delta}^{t+\delta}\hat{\boldsymbol{x}}_l(\tau)\hat{\boldsymbol{x}}_l^*(\tau)/(2\delta+1)$, where δ is a small positive integer, makes the computation of weighted sample correlation matrices/vectors (step 4 in Table II) very time consuming. Thus, we propose four alternative methods that use structured approximations, two of which clarify the relationship between the WPE, CSD, and our proposed GWPE methods.

A. Diagonal Matrix Method

The first method assumes each spatial correlation matrix to be diagonal as follows:

$$\hat{\boldsymbol{\Lambda}}_l(t) = E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right) \approx \mathrm{diag}\left(\lambda_l^m(t)\right)_{1\le m\le N}, \tag{35}$$

where $\lambda_l^m(t)$ is the power of $\hat{X}_l^m(t)$, i.e., $\lambda_l^m(t) = E(|\hat{X}_l^m(t)|^2)$. Such a $\lambda_l^m(t)$ value may be estimated as $\hat{\lambda}_l^{m}(t) = |\hat{x}_l^m(t)|^2$. The diagonal matrix method is a rather crude approximation because it ignores the spatial coupling between the output channels. However, as far as we have tested, this method does have some dereverberation efficacy in most cases.

The advantage of the diagonal matrix method is its relatively low computational complexity. Indeed, under the diagonal matrix assumption, $\hat{\boldsymbol{R}}_l$ becomes a block diagonal matrix (after permuting some columns), and therefore $\mathcal{G}_l^1 = \{\boldsymbol{g}_l^1(t)\}_{\Delta\le t\le\Delta+K_l-1}, \ldots, \mathcal{G}_l^N = \{\boldsymbol{g}_l^N(t)\}_{\Delta\le t\le\Delta+K_l-1}$ can be updated separately.

Interestingly, the GWPE algorithm using the diagonal matrix approximation is equivalent to performing the WPE method separately for all channels. This can be confirmed as follows. First, we rewrite (23) as

$$\tilde{F}(\mathcal{G}_l,\mathcal{L}_l) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} E\Bigg(\left(\boldsymbol{Y}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1}\boldsymbol{G}_l^*(\tau)\boldsymbol{Y}_l(t-\tau)\right)^* \times\hat{\boldsymbol{\Lambda}}_l(t)^{-1}\left(\boldsymbol{Y}_l(t) - \sum_{\tau=\Delta}^{\Delta+K_l-1}\boldsymbol{G}_l^*(\tau)\boldsymbol{Y}_l(t-\tau)\right)\Bigg), \tag{36}$$

and we denote this function as $F_{\mathrm{GWPE}}(\mathcal{G}_l)$. Substituting (35) into (36) and replacing each expectation by the corresponding realization, we obtain

$$F_{\mathrm{GWPE}}(\mathcal{G}_l) = \frac{1}{|\mathcal{T}|}\sum_{m=1}^{N} F_{\mathrm{WPE}}\left(\mathcal{G}_l^m\right), \tag{37}$$

where $F_{\mathrm{WPE}}\left(\mathcal{G}_l^m\right)$ is given by (12). Therefore, the WPE method can be derived as a special case of the GWPE algorithm. This observation shows that the WPE method can be reasonably applied to blind MIMO impulse response shortening, and that we can modify it to take account of the spatial coupling between the output channels.

B. Scaled Identity Matrix Method

The second method further simplifies the diagonal matrix method, assuming that the powers of different output channels are equal at each time instant t. Hence, we have the following approximation:

$$\hat{\boldsymbol{\Lambda}}_l(t) = E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right) \approx \lambda_l(t)\boldsymbol{I}. \tag{38}$$

$\lambda_l(t)$ is a "spatially averaged output power", which changes with time t and may be estimated as $\hat{\lambda}_l(t) = \sum_{m=1}^{N}\left|\hat{x}_l^m(t)\right|^2/N$. This approximation is reasonable when a microphone array is so small that the powers of different output channels tend to change in synchronization.

The advantage of the scaled identity matrix method over the diagonal matrix method lies in the fact that the weighted sample correlation matrix $\hat{\boldsymbol{R}}_l$ and vector $\hat{\boldsymbol{r}}_l$ can be computed more efficiently because the multiplication of matrix $\hat{\boldsymbol{\Lambda}}_l(t)^{-1}$ in (33) and (34) reduces to a scalar multiplication. The computation saving offered by this approach becomes significant as the number of microphones increases.

C. Scaled Full-Correlation Matrix Method

The third method uses a full-correlation matrix in place of the identity matrix used in the scaled identity matrix method. To be more precise, we use the following approximation:

$$\hat{\boldsymbol{\Lambda}}_l(t) = E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right) \approx \lambda_l(t)\boldsymbol{\Phi}_l. \tag{39}$$

$\boldsymbol{\Phi}_l$ is a time-invariant full-correlation matrix that describes the temporally averaged coupling characteristics between output channels.

Unlike the first two methods, the scaled full-correlation matrix method takes account of the spatial correlation between different output channels. At the same time, this method still enables us to compute $\hat{\boldsymbol{R}}_l$ and $\hat{\boldsymbol{r}}_l$ as efficiently as the scaled identity matrix approach. Note however that when $\hat{\boldsymbol{\Lambda}}_l(t)$ has non-zero non-diagonal elements, all the elements of $\mathcal{G}_l$ must be updated jointly according to step 5 in Table II. This increases the computational complexity especially when both the number of microphones and the prediction order are large.

$\boldsymbol{\Phi}_l$ and $\{\lambda_l(t)\}_{t\in\mathcal{T}}$ may be estimated by iterating a set of estimation procedures consisting of

$$\hat{\boldsymbol{\Phi}}_l = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\frac{1}{\hat{\lambda}_l(t)}\hat{\boldsymbol{x}}_l(t)\hat{\boldsymbol{x}}_l^*(t) \tag{40}$$

and

$$\hat{\lambda}_l(t) = \frac{1}{N}\hat{\boldsymbol{x}}_l^*(t)\hat{\boldsymbol{\Phi}}_l^{-1}\hat{\boldsymbol{x}}_l(t). \tag{41}$$

Our implementation sets the iteration number at two.

D. Independent Component Analysis Method

The fourth approach takes account of the spatial correlation between different output channels by using independent component analysis (ICA). Specifically, we assume that $\hat{\boldsymbol{X}}_l(t)$ can be approximated as

$$\hat{\boldsymbol{X}}_l(t) \approx \boldsymbol{A}_l\boldsymbol{U}_l(t), \tag{42}$$

where $\boldsymbol{A}_l$ is a time-invariant matrix and $\boldsymbol{U}_l(t)$ is a random vector whose elements are statistically independent of each other. $\boldsymbol{A}_l$ is estimated by applying ICA to the data $(\boldsymbol{x}_l(t))_{t\in\mathcal{T}}$. Then, we have

$$\hat{\boldsymbol{\Lambda}}_l(t) = E\left(\hat{\boldsymbol{X}}_l(t)\hat{\boldsymbol{X}}_l^*(t)\right) \approx \boldsymbol{A}_l \operatorname{diag}\left(\lambda_l^m(t)\right)_{1\leq m\leq N} \boldsymbol{A}_l^*, \tag{43}$$

where $\lambda_l^m(t)$ is the power of the mth element of $\boldsymbol{U}_l(t)$, which may be estimated as $\hat{\lambda}_l^m(t) = |u_l^m(t)|^2$, where $u_l^m(t)$ is the mth element of $\boldsymbol{A}_l^{-1}\boldsymbol{x}_l(t)$. Note that if we assume that all $\lambda_l^1(t), \ldots, \lambda_l^N(t)$ are equal, this method degenerates to the scaled full-correlation matrix method.

The GWPE algorithm with ICA-based spatial correlation estimation is essentially equivalent to the dereverberation subsystem of the CSD method when the numbers of sources and microphones are equal. The difference between the two algorithms lies only in that, for each m, the power spectrum $\{\lambda_l^m(t)\}_{l\in\mathcal{F}}$, where $\mathcal{F}$ is a set of all subband indices, is modeled with an all-pole model in the CSD method while the GWPE algorithm does not use such constraints. This observation shows that the CSD method is a special case of the GWPE algorithm, and that applying it to blind MIMO impulse response shortening is theoretically sound. Furthermore, if we wish to do so, it is possible to modify the spatial correlation matrix estimator to take account of the background noise.

VI. Experimental Results

In this section, we demonstrate that the proposed GWPE algorithm achieves MIMO impulse response shortening in noisy reverberant environments. We conducted two sets of experiments. In the first set, we closely examined the impulse response shortening effect by comparing sound decay curves and room impulse responses obtained before and after applying the GWPE algorithm. In addition, we estimated the degree of improvement in the direct-to-reverberation ratio (DRR) for quantitative evaluation. The second set of experiments was undertaken to see how sound source localization accuracy and noise reduction beamformer performance changed when we preprocessed microphone signals with the GWPE algorithm. The goal was to demonstrate that the proposed algorithm can make microphone array systems robust against reverberation. (In addition to these experiments, we confirmed that the GWPE algorithm improves the performance of meeting speech recognition as detailed in [21].)

We recorded reverberant speech and acoustic noise separately and mixed them on a computer to simulate noisy reverberant speech in order to make the experiments as realistic as possible while keeping the experimental conditions controllable. The precise recording conditions are as follows. We recorded eight-channel speech signals in a varechoic chamber. The chamber was 4.45 m wide and 3.35 m long with a 2.5 m high ceiling. We employed two reverberation times (T_{60}): 0.39 s and 0.65 s. An eight-element equidistant linear microphone array with 3 cm intervals was placed against the wall. A loudspeaker was placed in front of the microphone array at a distance of 2 m. For each T_{60}, we played clean speech recordings through the loudspeaker and captured them with the microphone array to obtain reverberant speech signals. We also played white Gaussian noise through another loudspeaker and recorded the sound with the same microphone array. The noise loudspeaker was placed 2 m to the left of the microphone array at an angle of 60 degrees. The recorded speech and noise were mixed at SNRs of 10 and 15 dB. Thus, we obtained four sets of noisy reverberant speech signals that differed in terms of T_{60} or SNR.

We used utterances from the TIDIGITS test set [22], resampled at 8 kHz, as the clean speech signals. TIDIGITS consists of connected digit strings. Although TIDIGITS contains very short utterances consisting of a single digit, the GWPE method requires microphone signals of a certain length to reliably estimate prediction filters. Hence, we concatenated a few utterances of the same speaker so that the duration of each audio file became approximately 5 s. Thus, a total of 1213 audio files were obtained and used for each of the four environmental conditions (i.e., T_{60} and SNR combinations).

Subband decomposition was performed with an oversampled uniform discrete Fourier transform filterbank using a fast Fourier transform [23], where a prototype lowpass filter of order 256 was designed with the classical window-based method. The number of subbands and the decimation factor were set at 128 and 64, respectively. The prediction order K_l for each subband was determined according to the subband center frequency. Specifically, we used the following K_l values: $K_l = 18$ for $f_l < 800$, $K_l = 15$ for $800 \leq f_l < 1500$, and $K_l = 12$ for $f_l \geq 1500$, where f_l is the center frequency in Hz of the lth frequency band to take advantage of the fact that reverberation diminishes faster in higher frequency bands. The Δ value was set at 2. These parameter values were commonly used for all the four environmental conditions.

A. Direct Performance Evaluation

In the first set of experiments, we evaluated the impulse response shortening effect of the GWPE algorithm. To do this, we first looked at the results for a T_{60} of 0.39 s and an input SNR of 15 dB. We used the scaled identity matrix approximation method as the spatial correlation matrix estimator.

Fig. 1 shows how the sound decay curve of the room impulse response between the loudspeaker and the first microphone was changed by the GWPE algorithm. The dashed and solid lines are the sound decay curves obtained before and after applying the GWPE algorithm, respectively. These curves were calculated as follows. First, we estimated the impulse response between the speaker and the first microphone positions for each of the 1213 files with a T_{60} of 0.39 s and an SNR of 15 dB based on the clean speech signal and the corresponding noiseless reverberant signal (i.e., the reverberant signal before being mixed with the noise). The impulse response estimation method used in this step is described in the next paragraph. Then, the sound decay curve for each estimated impulse response was calculated by using Schroeder's method [24]. Finally, the 1213 sound decay curves were averaged to obtain the dashed line in Fig. 1. Prior to the averaging computation, the delays in each of the 1213 sound decay curves were normalized. The solid line in Fig. 1 was also obtained based on the clean speech signals and the dereverberated signals of the first microphone.

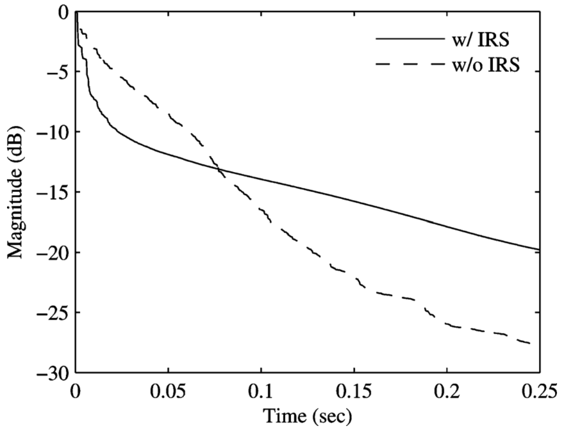


Fig. 1. Sound decay curves with and without impulse response shortening (IRS).

To estimate the impulse response between the speaker and the first microphone for each file, we assumed that the noiseless reverberant speech signal of the first microphone, which we denote by $r[k]$, could be represented as a linearly filtered version of the clean speech signal, $s[k]$, as $r[k] \approx \sum_{\kappa=0}^{I-1} f[\kappa]s[k-\kappa]$, where k is the full-band time index and $\{f[\kappa]\}_{0\leq\kappa\leq I-1}$ is the impulse response we want to obtain. The impulse response was estimated by minimizing the sum of the squared errors between $r[k]$ and $\sum_{\kappa=0}^{I-1} f[\kappa]s[k-\kappa]$, where the impulse response order, I, was set at 0.75 s. The blindly shortened impulse response (i.e., the impulse response obtained after applying the GWPE algorithm), required for drawing the solid line in Fig. 1, was estimated in the same way by using the noiseless components of the dereverberated signals instead of the noiseless reverberant signals. Here, the noiseless component of each dereverberated signal was estimated by processing the corresponding *noiseless* reverberant signal with the dereverberation filter estimated from the corresponding *noisy* reverberant signal.

In Fig. 1, we can see that the energy of the initial portion of the reverberation was effectively reduced. Recalling that the initial decay rate is a primary factor determining the perception of the decay [25], this result indicates that the GWPE algorithm successfully made the microphone signals less reverberant, as also implied by the improved DRRs presented later. On the other hand, in Fig. 1, we can also see an increase in the energy of the late reverberation. Indeed, when we listened to the noiseless components of the dereverberated signals (note that the dereverberated signals contained noise), we could hear both strong direct sounds and late reverberation components. However, the late reverberation components were hardly heard when we listened to the noisy versions of the dereverberated signals, i.e., they were masked by the noise. It should be noted that the decreased decay rate of the late reverberation is known to occur with many linear filter-based dereverberation techniques and effective remedies have already been proposed [26], [27].

To confirm the impulse response shortening effect visually, in Fig. 2, we plot the average impulse responses obtained before and after applying the GWPE algorithm. The right panels in Fig. 2 are magnified views of the left panels. (Note that the sound decay curves in Fig. 1 do not match these impulse responses exactly. This is because the decay curves in Fig. 1 were calculated by averaging the decay curves of the impulse responses for individual files while the impulse responses in Fig. 2 were obtained by averaging the impulse responses for individual files and thus the impulse response fluctuation affecting the decay curve shape was smoothed out in Fig. 2.) We can see that, in the first 10-ms-long portion starting at the arrival of the direct sound, the impulse response was almost unchanged. In the subsequent 10-ms-long portion, the shape of the impulse response was retained although the amplitude was attenuated. This result clearly demonstrates the impulse response shortening effect provided by the GWPE algorithm. Fig. 3 shows the frequency responses of the unprocessed and processed impulse responses. Numerous spectral peaks and dips were removed, indicating the successful cancellation of a large part of the reverberant distortion. However, several sharp peaks, dips, and tilts can be observed in the frequency response even after applying the GWPE algorithm, partly because the initial portions of the room impulse responses remain almost unchanged. Other techniques tailored to distortion caused by short impulse responses may be employed to compensate for the remaining distortion.

Finally, to evaluate the impulse response shortening performance quantitatively, we compared the DRRs obtained before and after applying the GWPE algorithm. The DRR is the ratio of the direct sound power to the reverberation power and was calculated as follows:

$$\mathrm{DRR} = 10\log_{10}\frac{\sum_k r^D[k]^2}{\sum_k r^R[k]^2}, \tag{44}$$

where the direct sound component, $r^D[k]$, and the reverberation component, $r^R[k]$, were calculated as $r^D[k] = \sum_{\kappa=0}^{\Phi-1} f[\kappa]s[k-\kappa]$ and $r^R[k] = \sum_{\kappa=\Phi}^{I-1} f[\kappa]s[k-\kappa]$, respectively. The Φ value in these equations determines the boundary between the direct and reverberation portions of the impulse response $\{f[\kappa]\}_{0\leq\kappa\leq I-1}$ and was set at 30 ms. Here, unlike the above investigations, we considered not only the impulse response for the first microphone but also those for the remaining seven microphones.

Table III lists the DRRs for all the four environmental conditions. Each of the DRRs in the table is the average over the corresponding 1213 results. We can draw the following conclusions from these results.

1) The GWPE algorithm consistently improved the DRRs for all the microphones.
2) A higher SNR resulted in greater DRR improvement although the improvement obtained when the SNR was 10 dB was still significant.

These results confirm that the GWPE algorithm can enhance the direct sound component for all the microphones in an array simultaneously, which implies the success of MIMO impulse response shortening.

B. Evaluation as a Microphone Array Preprocessor

In the second set of experiments, we evaluated the degree to which the sound source localization accuracy and the beamformer-based noise reduction performance were improved by preprocessing microphone signals with the proposed GWPE algorithm. Here, we tested both the scaled identity matrix approximation and scaled full-correlation matrix approximation methods for spatial correlation matrix estimation.

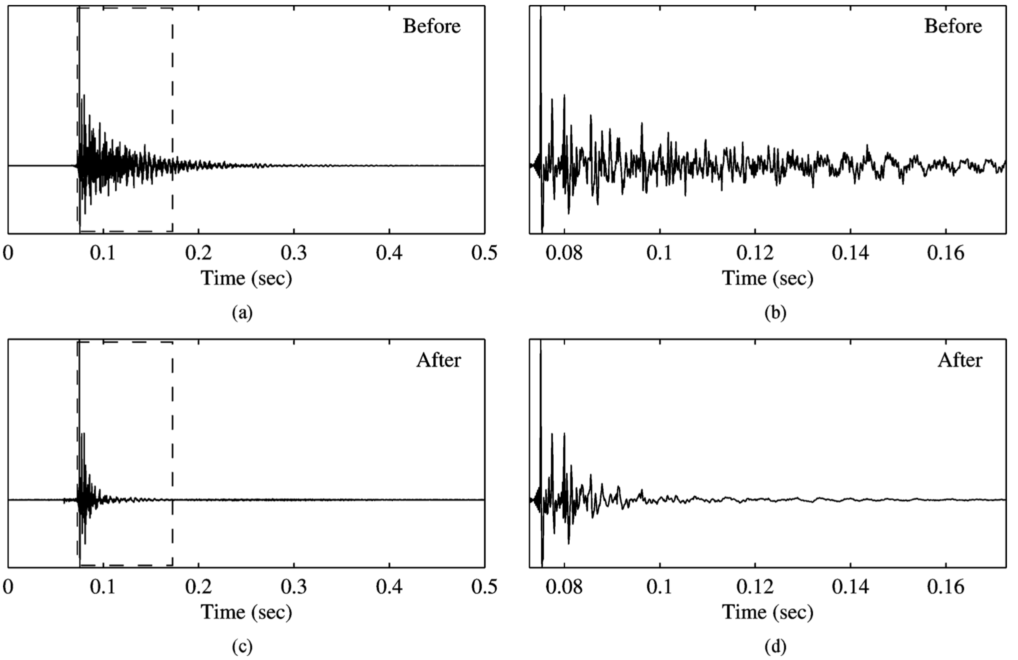


Fig. 2. Top left: Impulse response between first microphone and speaker position before impulse response shortening. Top right: Magnified view of the rectangular area in the top left panel. Bottom left: Impulse response between first microphone and speaker position after impulse response shortening. Bottom right: Magnified view of the rectangular area in the bottom left panel.

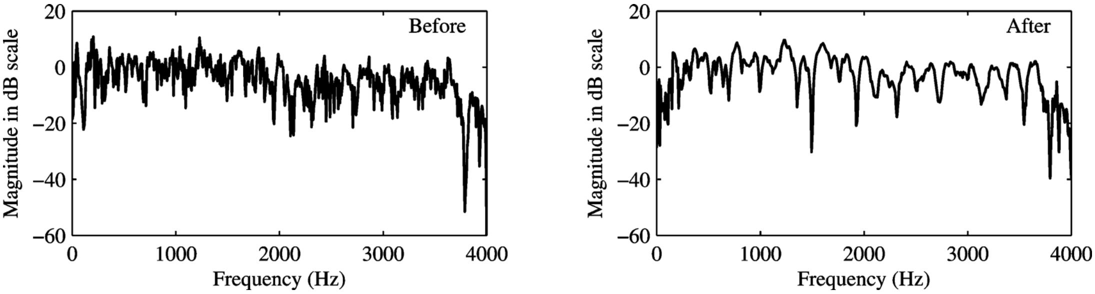


Fig. 3. Frequency responses corresponding to the impulse responses of Fig. 2.

TABLE III
DIRECT-TO-REVERBERATION RATIOS FOR EACH ENVIRONMENTAL CONDITION WITH AND WITHOUT IMPULSE RESPONSE SHORTENING (IRS).

T_{60}	0.39 s				0.65 s			
SNR	15 dB		10 dB		15 dB		10 dB	
	w/ IRS	w/o IRS	w/ IRS	w/o IRS	w/ IRS	w/o IRS	w/ IRS	w/o IRS
Microphone 1	15.37 dB	5.09 dB	13.94 dB	5.09 dB	11.57 dB	0.15 dB	10.39 dB	0.15 dB
Microphone 2	15.33 dB	5.24 dB	13.91 dB	5.24 dB	11.52 dB	0.05 dB	10.36 dB	0.05 dB
Microphone 3	15.27 dB	5.47 dB	13.86 dB	5.47 dB	11.48 dB	0.09 dB	10.32 dB	0.09 dB
Microphone 4	15.22 dB	5.74 dB	13.82 dB	5.74 dB	11.45 dB	0.15 dB	10.28 dB	0.15 dB
Microphone 5	15.16 dB	5.94 dB	13.75 dB	5.94 dB	11.45 dB	0.16 dB	10.30 dB	0.16 dB
Microphone 6	15.14 dB	6.12 dB	13.73 dB	6.12 dB	11.39 dB	0.13 dB	10.23 dB	0.13 dB
Microphone 7	15.15 dB	6.31 dB	13.76 dB	6.31 dB	11.37 dB	0.03 dB	10.24 dB	0.03 dB
Microphone 8	15.08 dB	6.51 dB	13.66 dB	6.51 dB	11.30 dB	-0.01 dB	10.11 dB	-0.01 dB

1) Source Localization: We used the steered response power-phase transform (SRP-PHAT) method, which itself is known to be robust against reverberant distortion [1], to perform the source localization. The SRP-PHAT method was applied to each 25-ms-long frame, shifted by 10 ms. The accuracy of the source localization was evaluated in terms of the localization correctness rate calculated as $T_{\mathrm{correct}}/T_{\mathrm{total}}$, where T_{correct} is the number of time frames in which a source direction is correctly identified, and T_{total} is the total number of frames. Unvoiced and silent frames were excluded from the localization correctness rate calculation. The source direction was regarded as being correctly identified if the direction estimation error was smaller than 10 degrees.

Table IV lists the localization correctness rates for each environmental condition and processing condition. Here, we considered three processing conditions: one did not perform impulse response shortening and applied the SRP-PHAT method to the microphone signals directly, the second used the GWPE algorithm with the scaled identity matrix approximation method before performing sound source localization, and the

TABLE IV
LOCALIZATION CORRECTNESS RATES FOR EACH ENVIRONMENTAL CONDITION AND PROCESSING CONDITION. PROCESSING CONDITIONS ARE AS FOLLOWS. W/O IRS: IMPULSE RESPONSE SHORTENING WAS DISABLED. W/IRS (IDENTITY): GWPE ALGORITHM WAS APPLIED BY USING SPATIAL CORRELATION MATRIX ESTIMATOR BASED ON SCALED IDENTITY MATRIX APPROXIMATION. W/IRS (FULL): GWPE ALGORITHM WAS APPLIED BY USING SPATIAL CORRELATION MATRIX ESTIMATOR BASED ON SCALED FULL-CORRELATION MATRIX APPROXIMATION.

T_{60}	0.39 s		0.65 s	
SNR	15 dB	10 dB	15 dB	10 dB
w/o IRS	89.7 %	87.8 %	90.8 %	89.9 %
w/ IRS (identity)	94.9 %	87.1 %	95.1 %	86.2 %
w/ IRS (full)	96.9 %	88.9 %	95.9 %	89.9 %

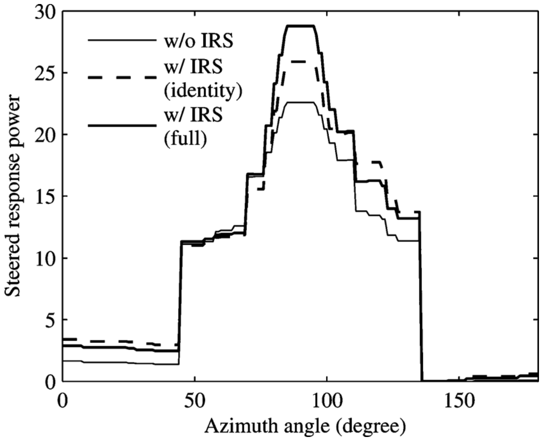


Fig. 4. SRP spectra for each processing condition for T_{60} of 0.65 s and SNR of 10 dB.

third used the scaled full-correlation matrix approximation method. When the background noise level was at an SNR of 15 dB, the GWPE algorithm improved the localization correctness rate considerably regardless of the spatial correlation matrix estimator. When the input SNR was 10 dB, the GWPE algorithm combined with the scaled identity matrix approximation method slightly degraded the localization correctness rate. The scaled full-correlation matrix approximation method did not exhibit such a negative effect.

To visualize the effect of the GWPE algorithm on sound source localization, the SRP spectra averaged over all frames and all of the 1213 files are shown in Fig. 4 for a T_{60} of 0.65 s and an SNR of 10 dB. We can clearly see that the GWPE algorithm sharpened the peak in the SRP spectrum, which indicates that the sound coming from the target direction was enhanced while the sounds coming from all the other directions were attenuated. This effect was prominent especially when we used the scaled full-correlation matrix approximation method for spatial correlation matrix estimation. This result may explain why the noise reduction performance was improved by the GWPE algorithm as described below.

2) Beamformer-Based Noise Reduction: We used a minimum variance distortionless response (MVDR) beamformer to perform noise reduction. The MVDR beamformer was applied to each of the frequency bins obtained by a short-time Fourier transform (STFT) with a 25-ms-long and 10-ms-shift hamming window. The noise reduction performance was evaluated by the improvement in the signal-to-noise ratio (SNR).

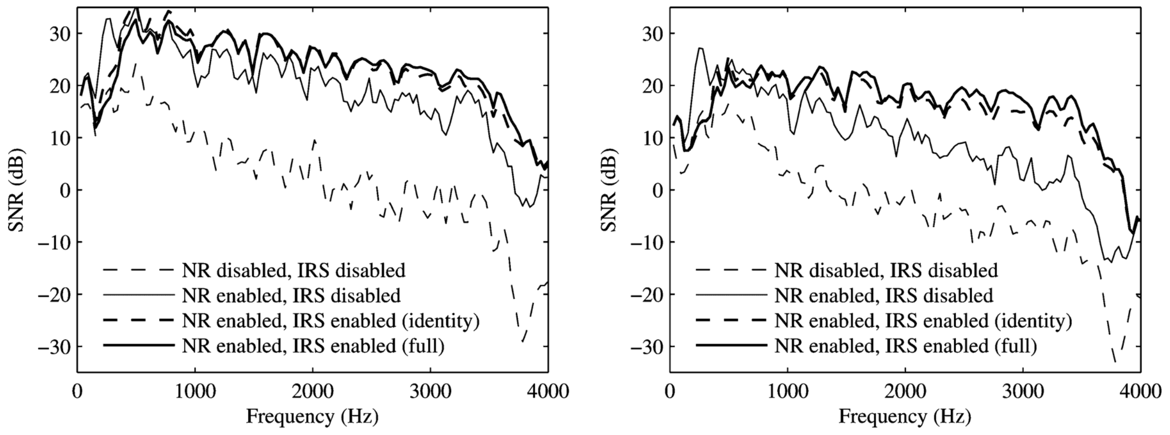
Fig. 5 shows the frequency-dependent (i.e., narrow band) SNRs on a dB scale for two different environmental conditions: one is with a T_{60} of 0.39 s and an input SNR of 15 dB; the other is with a T_{60} of 0.65 s and an input SNR of 10 dB. Comparing the thick (solid and dashed) lines with the thin solid line, we can see that the SNR was significantly improved in the mid-to-high frequency range regardless of the environmental conditions by performing impulse response shortening with the GWPE algorithm. With regard to spatial correlation matrix estimation, the scaled full-correlation matrix approximation method achieved slightly better SNRs than the scaled identify matrix approximation method. These results show that the GWPE algorithm successfully performed blind MIMO impulse response shortening, which boosted the effectiveness of the MVDR beamformer except in the low frequency region. Similar results were obtained for the other two environmental conditions.

The SNR degradation in the low frequency region was due to noise amplification caused by blind MIMO impulse response shortening. To confirm this, we show in Fig. 6 the dB-scale frequency-dependent SNRs obtained before MVDR beamforming for a T_{60} of 0.65 s and an input SNR of 10 dB. We can observe that the signals obtained with the GWPE algorithm had lower SNRs in the low frequency region than the microphone signals. This means that the GWPE algorithm amplified the background noise in the low frequency region, which explains why the combination of the GWPE algorithm and the MVDR beamformer underperformed the MVDR beamformer alone in that frequency region. Fortunately, this problem is not very serious because the GWPE method can be employed for each frequency band separately. Therefore, we have an option of performing blind MIMO impulse response shortening only in the mid-to-high frequency region.

VII. CONCLUSION

In this paper, we generalized the dereverberation approach based on subband-domain multi-channel prediction in order to derive a MIMO impulse response shortening algorithm without assuming specific acoustic conditions. This generalization was achieved by using a novel cost function to estimate the prediction filters and an efficient optimization algorithm based on the auxiliary function approach. In summary, the proposed GWPE algorithm provides a higher level framework from which the previously proposed WPE and CSD methods can be derived as special cases. This framework revealed that the difference between those two existing methods lies only in the way that we approximate a sequence of spatial correlation matrices of dereverberated signals. The proposed GWPE algorithm will help the future development of more advanced dereverberation techniques. In fact, we presented two new methods for approximating the spatial correlation matrix sequence, and we confirmed the effectiveness of these two approximation methods experimentally. Much more work is needed to assess the relative benefits of different approximation methods and to bring out the full potential of the GWPE algorithm.

As we noted at the beginning of this paper, MIMO impulse response shortening allows us to resolve the problem whereby typical microphone array processing techniques are seriously de-



(a)

(b)

Fig. 5. Frequency-dependent SNRs on dB scale. Left: Input SNR was 15 dB and T_{60} was 0.39 s. Right: Input SNR was 10 dB and T_{60} was 0.65 s. "NR" and "IRS" mean noise reduction and impulse response shortening, respectively, and "identity" and "full" mean scaled identity matrix approximation and scaled full-correlation matrix approximation, respectively.

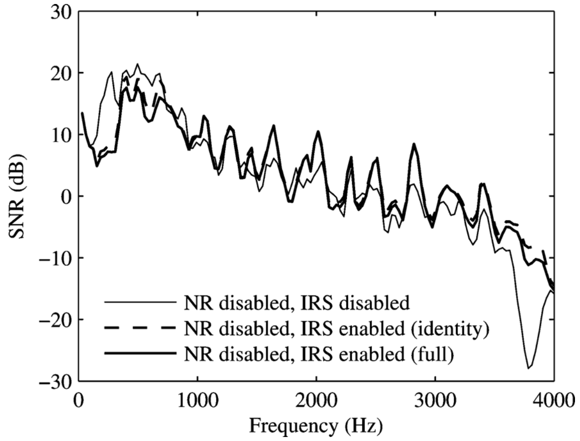


Fig. 6. Frequency-dependent SNRs on dB scale before MVDR beamforming for T_{60} of 0.65 s and SNR of 10 dB.

graded in reverberant environments. Thus, the proposed GWPE algorithm, achieving MIMO impulse response shortening, can be employed as a preprocessor for a variety of microphone array systems. For some applications, it will be better to combine the GWPE algorithm with other reverberation robustness techniques than to use the proposed algorithm alone. This is another direction that needs to be explored in the future.

Appendix A
Proof of Theorem 1

To prove Theorem 1, we use the Hadamard-Fischer inequality [17], [18]. This inequality states the following.

Theorem 4: Let $\boldsymbol{A} = [a_{i,j}]_{1 \le i,j \le N}$ be an N-by-N positive semidefinite Hermitian matrix. For any K with $1 \le K \le N$, let us use $\boldsymbol{A}^{[K]}$ and $\boldsymbol{A}^{(K)}$ to denote the K-by-K principal submatrix of $\boldsymbol{A}$ and the corresponding complementary principal submatrix, respectively, i.e., $\boldsymbol{A}^{[K]} = [a_{i,j}]_{1 \le i,j \le K}$ and $\boldsymbol{A}^{(K)} = [a_{i,j}]_{K+1 \le i,j \le N}$. If neither $\boldsymbol{A}^{[K]}$ nor $\boldsymbol{A}^{(K)}$ is singular, we have

$$\det \boldsymbol{A} \le \det \boldsymbol{A}^{[K]} \det \boldsymbol{A}^{(K)} \tag{45}$$

with equality if and only if $a_{i,j} = 0$ for all combinations of i and j values satisfying $1 \le i \le K$ and $K+1 \le j \le N$ or $K+1 \le i \le N$ and $1 \le j \le K$.

By applying the Hadamard-Fischer inequality recursively in $\det E(\boldsymbol{U}\boldsymbol{U}^*)$, we obtain

$$\det E(\boldsymbol{U}\boldsymbol{U}^*) \le \prod_{n=1}^{N} \det E\left(\boldsymbol{U}_n \boldsymbol{U}_n^*\right). \tag{46}$$

The condition for the equality is $E\left(\boldsymbol{U}_m \boldsymbol{U}_n^*\right) = \boldsymbol{O}$ for all $1 \le m \ne n \le N$. Theorem 1 is obtained by taking the log of (46).

Appendix B
Proof of Lemma 2

We prove that (22) holds for all natural numbers N by mathematical induction.

First, we show that this inequality holds when $N = 1$. Since the logarithmic function is concave, the following inequality holds:

$$\log(x) \le \frac{x}{\lambda} - 1 + \log(\lambda) \tag{47}$$

with equality if and only if $x = \lambda$. Thus, (22) is true when $N = 1$.

Let us hypothesize that (22) is true when $N = k - 1$. The next step is to show that this inequality also holds for $N = k$ under this hypothesis. To this end, we decompose k-dimensional random vector $\boldsymbol{U}$ and k-by-k positive definite Hermitian matrix $\boldsymbol{\Lambda}$ as

$$\boldsymbol{U} = \begin{bmatrix} \tilde{\boldsymbol{U}} \\ \tilde{U} \end{bmatrix} \text{ and } \boldsymbol{\Lambda} = \begin{bmatrix} \tilde{\boldsymbol{\Lambda}} & \tilde{\boldsymbol{\lambda}} \\ \tilde{\boldsymbol{\lambda}}^* & \tilde{\lambda} \end{bmatrix}, \tag{48}$$

respectively, where $\tilde{\boldsymbol{U}}$ is the $(k-1)$-dimensional vector that consists of the first $k-1$ entries of $\boldsymbol{U}$ while $\tilde{\boldsymbol{\Lambda}}$ is the $(k-1)$-by-$(k-1)$ principal submatrix of $\boldsymbol{\Lambda}$. We have an expression of $\boldsymbol{\Lambda}^{-1}$ as

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \tilde{\boldsymbol{\Lambda}}^{-1} + \frac{\phi\phi^*}{\varphi} & -\frac{\phi}{\varphi} \\ -\frac{\phi}{\varphi} & \frac{1}{\varphi} \end{bmatrix}, \tag{49}$$

where $\phi = \tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{\lambda}}$ and $\varphi = \tilde{\lambda} - \tilde{\boldsymbol{\lambda}}^{*}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{\lambda}}$. Moreover, $\det\boldsymbol{\Lambda}$ can be expressed by using φ and $\tilde{\boldsymbol{\Lambda}}$ as

$$\det\boldsymbol{\Lambda} = \varphi\det\tilde{\boldsymbol{\Lambda}}. \tag{50}$$

Plugging (49) and (50) into the right hand side of (22), we obtain

$$\begin{aligned}(\text{RHS of }(22)) &= \tilde{\boldsymbol{U}}^{*}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{U}} + \log(\det\tilde{\boldsymbol{\Lambda}}) \\ &+ \frac{1}{\varphi}\Big(\phi^{*}E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})\phi - E(\tilde{U}\tilde{\boldsymbol{U}}^{*})\phi \\ &\qquad - \phi^{*}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}) + E(|\tilde{U}|^{2})\Big) \\ &+ \log(\varphi) - k. \end{aligned} \tag{51}$$

Invoking the induction hypothesis and (47), the following inequality can be deduced:

$$\begin{aligned}(\text{RHS of }(22)) \geq \log(\det E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})) + \log\Big(\phi^{*}E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})\phi \\ -E(\tilde{U}\tilde{\boldsymbol{U}}^{*})\phi - \phi^{*}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}) + E(|\tilde{U}|^{2})\Big). \end{aligned} \tag{52}$$

The equality holds if and only if $\tilde{\boldsymbol{\Lambda}} = E(\boldsymbol{U}\boldsymbol{U}^{*})$ and $\varphi = \phi^{*}E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})\phi - E(\tilde{U}\tilde{\boldsymbol{U}}^{*})\phi - \phi^{*}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}) + E(|\tilde{U}|^{2})$. Concerning the second term of the right hand side of (52), we find that

$$\begin{aligned}&\phi^{*}E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})\phi - E(\tilde{U}\tilde{\boldsymbol{U}}^{*})\phi - \phi^{*}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}) \\ &\quad = (\phi - E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})^{-1}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}))^{*}E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*}) \\ &\qquad \times (\phi - E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})^{-1}E(\tilde{\boldsymbol{U}}\tilde{U}^{*})) \\ &\qquad - E(\tilde{U}\tilde{\boldsymbol{U}}^{*})E(\boldsymbol{U}\boldsymbol{U}^{*})^{-1}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}) \\ &\quad \geq -E(\tilde{U}\tilde{\boldsymbol{U}}^{*})E(\boldsymbol{U}\boldsymbol{U}^{*})^{-1}E(\tilde{\boldsymbol{U}}\tilde{U}^{*}) \end{aligned} \tag{53}$$

with equality if and only if $\phi = E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})^{-1}E(\tilde{\boldsymbol{U}}\tilde{U}^{*})$. Considering (52) and (53), we see that

$$\begin{aligned}(\text{RHS of }(22)) \geq \log(\det E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})) + \log\big(E(|\tilde{U}|^{2}) \\ -E(\tilde{U}\tilde{\boldsymbol{U}}^{*})E(\boldsymbol{U}\boldsymbol{U}^{*})^{-1}E(\tilde{\boldsymbol{U}}\tilde{U}^{*})\big). \end{aligned} \tag{54}$$

In particular, the equality holds if and only if $\tilde{\boldsymbol{\Lambda}} = E(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{*})$, $\tilde{\boldsymbol{\lambda}} = E(\tilde{\boldsymbol{U}}\tilde{U}^{*})$, and $\tilde{\lambda} = E(|\tilde{U}|^{2})$. It is easy to see that the right hand side of (54) is equal to that of (22), which completes the proof.

References

[1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001.

[2] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 624–639, Mar. 2011.

[3] *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001.

[4] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven Univ. of Technology, Eindhoven, The Netherlands, 2006.

[5] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009.

[6] *Unsupervised Adaptive Filtering Volume 2: Blind Deconvolution*, S. Haykin, Ed. New York: Wiley-Interscience, 2000.

[7] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. VI, pp. 3701–3704.

[8] M. Triki and D. T. M. Slock, "Iterated delay and predict equalization for blind speech dereverberation," in *Proc. Int. Workshop Acoust. Echo, Noise Contr.*, 2006, CD-ROM Proceedings.

[9] T. Yoshioka, T. Nakatani, K. Kinoshita, and M. Miyoshi, "Speech dereverberation and denoising based on time varying speech model and autoregressive reverberation model," in *Speech Processing in Modern Communication*, I. Cohen, J. Benesty, and S. Gannot, Eds. Berlin, Germany: Springer, 2010.

[10] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang, Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[11] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. New York: Springer, 2010.

[12] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007, Article ID 65698.

[13] Z. Ding, "Linear predictive algorithms for blind multichannel identification," in *Signal Processing Advances in Wireless and Mobile Communications*, G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, Eds. Upper Saddle River, NJ: Prentice-Hall, 2001, vol. 1, pp. 179–209.

[14] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[15] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 120–134, 2005.

[16] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Netw.*, vol. 8, no. 3, pp. 411–419, 1995.

[17] G. M. Engel and H. Schneider, "The Hadamard-Fischer inequality for a class of matrices defined by eigenvalue monotonicity," *Linear, Multilinear Algebra*, vol. 4, pp. 155–176, 1976.

[18] T. H. Pate, "Exterior products, elementary symmetric functions, and the Fischer determinant inequality," *Linear Algebra, Its Appl.*, vol. 255, pp. 203–242, 1997.

[19] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super Gaussian priors," *IEEE Trans. Speech, Audio, Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[20] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Syst. 13*, pp. 556–562, 2000.

[21] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 499–513, Feb. 2011.

[22] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1984, vol. 9, pp. 328–331.

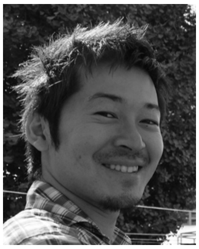
[23] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 243–248, Jun. 1976.

[24] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[25] B. Yegnanarayana and B. S. Ramakrishna, "Intelligibility of speech under nonexponential decay conditions," *J. Acoust. Soc. Amer.*, vol. 58, no. 4, pp. 853–857, 1975.

[26] M. Wu and D. Wang, "A two-stage algorithm for enhancement of reverberant speech," *IEEE Trans. Speech, Audio Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[27] K. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," (in Japanese) *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1579–1591, Jul. 2007.



Takuya Yoshioka (M'08) received the B.Eng., M.Inf., and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2004, 2006, and 2010, respectively.

He is a Researcher at NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. In 2005, he interned at NTT, where he conducted research of dereverberation. Since joining NTT in 2006, he has been working on the development of algorithms for speech enhancement and noise robust speech recognition.

Dr. Yoshioka received the Awaya Prize Young Researcher Award and the Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2010 and 2011, respectively, and the Young Researcher's Award in Speech Field from the Institute of Electronics, Information and Communication Engineers (IEICE) Information and Systems Society (ISS) in 2011.

Tomohiro Nakatani (SM'06) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 2002, respectively.

He is a Senior Research Scientist (Supervisor) of NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Since joining NTT Corporation as a Researcher in 1991, he has been investigating speech enhancement technologies for developing intelligent human-machine interfaces. Since 2005, he has visited the Georgia Institute of Technology as a Visiting Scholar for a year where he worked with Prof. B.-H. Juang. Since 2008, he has been a Visiting Assistant Professor in the Department of Media Science, Nagoya University.

Dr. Nakatani was honored to receive the 1997 JSAI Conference Best Paper Award, the 2002 ASJ Poster Award, the 2005 IEICE Best Paper Award, and the 2009 ASJ Technical Development Award. He has been a member of IEEE SP Society Audio and Acoustics Technical Committee since 2009, a member of IEEE CAS Society Blind Signal Processing Technical Committee since 2007, and a Chair of the IEEE Kansai Section Technical Program Committee since 2011. He served as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING for three years from 2008, and a Technical Program co-Chair of IEEE WASPAA-2007. He is a member of IEICE and ASJ.