Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation

Tomohiro Nakatani and Keisuke Kinoshita

NTT Communication Science Laboratories, NTT Corporation, Japan

Abstract—**This article describes a probabilistic formulation of a Weighted Power minimization Distortionless response convolutional beamformer (WPD). The WPD unifies a weighted prediction error based dereverberation method (WPE) and a minimum power distortionless response beamformer (MPDR) into a single convolutional beamformer, and achieves simultaneous dereverberation and denoising in an optimal way. However, the optimization criterion is obtained simply by combining existing criteria without any clear theoretical justification. This article presents a generative model and a probabilistic formulation of a WPD, and derives an optimization algorithm based on a maximum likelihood estimation. We also describe a method for estimating the steering vector of the desired signal by utilizing WPE within the WPD framework to provide an effective and efficient beamformer for denoising and dereverberation.**

Index Terms—**Denoising, dereverberation, microphone array, speech enhancement, maximum likelihood estimation**

I. INTRODUCTION

When a speech signal is captured by distant microphones, e.g., in a conference room, it will inevitably contain additive noise and reverberation components. These components are detrimental to the perceived quality of the observed speech signal and often cause serious degradation in many applications such as hands-free teleconferencing and ASR.

Microphone array signal processing has been investigated to minimize the aforementioned detrimental effects on the acquired signal. A filter-and-sum beamformer [1], a minimum-variance distortionless response beamformer (MVDR) and MPDR [2]–[6], and a maximum signal-to-noise ratio beamformer [7]–[9] are widely-used techniques for denoising, while WPE and its variants [10]–[14] are emerging dereverberation techniques. Techniques for reducing both noise and reverberation have also been investigated, for example, using both MVDR and WPE by cascade integration [15]–[18]. The usefulness of these techniques, particularly for ASR, has been extensively studied, e.g., at the REVERB challenge [19] and the CHiME-3/4/5 challenges [20]–[22]. Moreover, advances in these techniques have led to recent progress on commercial devices, such as smart speakers [23]–[25].

Recently, a new unified beamforming approach has been proposed for achieving denoising and dereverberation both simultaneously and optimally [26], [27]. In [26], researchers introduced a convolutional beamformer that unifies WPE and MPDR, and presented a method for optimizing the beamformer by using a single optimization criterion. It is referred to as a Weighted Power minimization Distortionless response beamformer (WPD). We showed experimentally that this beamformer provides substantially better simultaneous denoising and dereverberation performance than conventional

cascade integration approaches[1]. However, the reason for the effectiveness of the unified approach remains somewhat unclear because the optimization criterion was constructed simply by combining existing criteria without any clear theoretical justification.

To clarify the mechanism of the unified approach, this article presents a generative model and a probabilistic formulation of a WPD, and derives an optimization algorithm based on a maximum likelihood (ML) estimation. We also present an iterative estimation method based on WPD for the steering vector of the desired signal, which was conventionally assumed to be given or estimated in advance. A key to successful estimation is to use multi-input multi-output (MIMO) dereverberation by WPE [11] within the WPD framework. Experiments using the REVERB challenge dataset show the importance of using the MIMO dereverberation to achieve superior denoising and dereverberation in terms of objective speech enhancement measures and ASR performance.

In the remainder of this paper, we define the model of the signal and the beamformer in Section II, and present the probabilistic formulation of WPD in Section III. Section IV derives the optimization algorithm and describes a processing flow that utilizes WPE. The experimental results and concluding remarks are given in Sections V and VI, respectively.

II. MODEL OF SIGNAL AND BEAMFORMER

We assume that a single speech signal is captured by M microphones in a noisy reverberant environment. The captured signal in the short-time Fourier transform (STFT) domain is approximately modeled at each frequency bin [28] by

$$\mathbf{x}_t = \sum_{\tau=0}^{L_a} \mathbf{a}_\tau s_{t-\tau} + \mathbf{n}_t, \qquad (1)$$

where t and τ are frame indices, $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(M)}]^\top$ is, letting $\top$ denotes a non-conjugate transpose, a column vector containing STFT coefficients of all the microphone signals at a time frame t, s_t is the STFT coefficient of the clean speech signal, $\mathbf{a}_t = [a_t^{(1)}, a_t^{(2)}, \ldots, a_t^{(M)}]^\top$ for $t = 0, 1, \ldots, L_a$ is a set of column vectors containing convolutional acoustic transfer functions from the speaker location to all the microphones, and $\mathbf{n}_t = [n_t^{(1)}, n_t^{(2)}, \ldots, n_t^{(M)}]^\top$ is the additive noise. In this paper, the frequency indices of the

---

[1]A recent study has revealed that the WPD yields the same outputs even when it is implemented with a certain cascade configuration consisting of a WPE and an MPDR. The details will be discussed in future publications. The analysis presented in this paper is based on a form of convolutional beamformer that unifies a WPE and an MPDR into a single filter.

symbols are omitted for brevity, and on the assumption that each frequency bin is processed independently in the same way. Hereafter, we refer to the STFT coefficients of signals simply as signals.

The first term in Eq. (1) can be further decomposed into two parts, one consisting of the direct signal and early reflections, and referred to as a desired signal $\mathbf{d}_t$, and the other corresponding to the late reverberation $\mathbf{r}_t$ [29]. With this decomposition, Eq. (1) is rewritten as

$$\mathbf{x}_t = \mathbf{d}_t + \mathbf{r}_t + \mathbf{n}_t, \tag{2}$$

$$\mathbf{d}_t = \sum_{\tau=0}^{b-1} \mathbf{a}_\tau s_{t-\tau}, \tag{3}$$

$$\mathbf{r}_t = \sum_{\tau=b}^{L_a} \mathbf{a}_\tau s_{t-\tau}, \tag{4}$$

where b is a frame index that divides the reverberation into the two parts. The goal of the beamforming is to preserve $\mathbf{d}_t$, while reducing $\mathbf{r}_t$ and $\mathbf{n}_t$ from $\mathbf{x}_t$. In this paper, $\mathbf{d}_t$, $\mathbf{r}_t$, and $\mathbf{n}_t$ are assumed to be statistically independent of each other. (See [10] for a more precise discussion of the statistical independence of $\mathbf{d}_t$ and $\mathbf{r}_t$.)

With WPD, we further assume that the transfer function corresponding to the desired signal can be approximated by a product of a vector $\mathbf{v}$ with a clean speech signal, i.e., $\mathbf{d}_t = \mathbf{v}s_t$, in the STFT domain. Then Eq. (2) becomes

$$\mathbf{x}_t = \mathbf{v}s_t + \mathbf{r}_t + \mathbf{n}_t, \tag{5}$$

Here, $\mathbf{v}$ is also termed a steering vector. This paper sets $m = 1$ as the reference microphone, and describes a method for estimating the desired signal at the microphone without loss of generality. The desired signal at the reference microphone is represented as

$$d_t^{(1)} = v^{(1)} s_t, \tag{6}$$

where $v^{(1)}$ is the element of $\mathbf{v}$ at the reference microphone.

A. Model of convolutional beamformer

We now define a MIMO convolutional beamformer as

$$\mathbf{y}_t = W_0^{\mathsf{H}}\mathbf{x}_t + \sum_{\tau=b}^{L_w} W_\tau^{\mathsf{H}}\mathbf{x}_{t-\tau}, \tag{7}$$

where $\mathbf{y}_t$ is the output of the beamformer, W_t for each t ($= 0, b, b+1, \ldots, L_w$) is an $M \times M$ dimensional matrix, which is composed of the beamformer coefficients, H denotes conjugate transpose, and b is the prediction delay that corresponds to b in Eq. (3) and is introduced to prevent the desired signal from being distorted by the convolutional beamforming [10]. We further decompose W_t for each t as follows

$$W_t = [\mathbf{w}_t, B_t], \tag{8}$$

where $\mathbf{w}_t = [w_t^{(1)}, w_t^{(2)}, \ldots, w_t^{(M)}]^\top$ is an M-dimensional column vector that denoises and dereverberates the captured signal, and B_t is an $M \times (M-1)$ dimensional matrix that extracts the noise from the captured signal. Note that our interest is in the estimation of $\mathbf{w}_t$, but B_t is introduced because it is necessary for the probabilistic formulation of a WPD.

To characterize $\mathbf{w}_t$ and B_t, we introduce the following constraints.

$$\mathbf{w}_0^{\mathsf{H}}\mathbf{v} = v^{(1)}, \tag{9}$$

$$B_0^{\mathsf{H}}\mathbf{v} = 0. \tag{10}$$

Eq. (9) specifies that $\mathbf{w}_0$ extracts the desired signal at the reference channel with no distortion. Eq. (10) specifies that B_0 blocks the signal subspace spanned by $\mathbf{v}$ that includes the desired signal. With the constraints, and based on Eqs. (2) and (5), Eq. (7) can be rewritten as

$$\mathbf{y}_t = \begin{bmatrix} y_t^{(1)} \\ \mathbf{y}_t^{(2:M)} \end{bmatrix} = \begin{bmatrix} d_t^{(1)} + \tilde{r}_t + \tilde{n}_t \\ \dot{\mathbf{r}}_t + \dot{\mathbf{n}}_t \end{bmatrix}, \tag{11}$$

where $y_t^{(1)}$ is the first element of $\mathbf{y}_t$, $\mathbf{y}_t^{(2:M)}$ is a column vector containing the other elements of $\mathbf{y}_t$, and

$$\tilde{r}_t = \mathbf{w}_0^{\mathsf{H}}\mathbf{r}_t + \sum_{\tau=b}^{L_w} \mathbf{w}_\tau^{\mathsf{H}}(\mathbf{d}_{t-\tau} + \mathbf{r}_{t-\tau}), \tag{12}$$

$$\tilde{n}_t = \mathbf{w}_0^{\mathsf{H}}\mathbf{n}_t + \sum_{\tau=b}^{L_w} \mathbf{w}_\tau^{\mathsf{H}}\mathbf{n}_{t-\tau}, \tag{13}$$

$$\dot{\mathbf{r}}_t = B_0^{\mathsf{H}}\mathbf{r}_t + \sum_{\tau=b}^{L_w} B_\tau^{\mathsf{H}}(\mathbf{d}_{t-\tau} + \mathbf{r}_{t-\tau}), \tag{14}$$

$$\dot{\mathbf{n}}_t = B_0^{\mathsf{H}}\mathbf{n}_t + \sum_{\tau=b}^{L_w} B_\tau^{\mathsf{H}}\mathbf{n}_{t-\tau}, \tag{15}$$

In Eq. (11), $\tilde{r}_t$ and $\dot{\mathbf{r}}_t$ are the reverberation that remains after the beamforming, and $\tilde{n}_t$ and $\dot{\mathbf{n}}_t$ represent the noise that remains after the beamforming.

III. PROBABILISTIC FORMULATION

Let $\theta_w = \{\mathbf{w}_0, \mathbf{w}_b, \ldots, \mathbf{w}_{L_w}\}$, $\theta_B = \{B_0, B_b, \ldots, B_{L_w}\}$, $\theta_\sigma = \{\sigma_t^2 \mid t = 1, 2, \ldots, T\}$, and $\theta = \{\theta_w, \theta_B, \theta_\sigma, \mathbf{v}\}$ be model parameter sets, where σ_t^2 is the time-varying power of the desired signal, and T is the number of available time frames. Then, assuming based on Eq. (7) that the probabilistic uncertainty of $\mathbf{x}_t$ is derived only from $\mathbf{y}_t$ when $\mathbf{x}_{t-\tau}$ for $\tau = b, b+1, \ldots, L_w$ are given, the likelihood function can be defined and rewritten as

$$\mathcal{L}(\theta) = \log p(\{\mathbf{x}_t\}; \theta), \tag{16}$$

$$= \sum_t \log p(\mathbf{x}_t \mid \mathbf{x}_{t-b}, \mathbf{x}_{t-b-1}, \ldots, \mathbf{x}_{t-L_w}; \theta) \tag{17}$$

$$= \sum_t \log p\left(W_0^{\mathsf{H}}\mathbf{x}_t \mid \sum_{\tau=b}^{L_w} W_\tau^{\mathsf{H}}\mathbf{x}_{t-\tau}; \theta\right)$$
$$+ 2T \log |\det(W_0)|, \tag{18}$$

$$= \sum_t \log p(\mathbf{y}_t; \theta) + 2T \log |\det(W_0)|. \tag{19}$$

Now, we assume that the optimal beamformer θ_w can reduce the level of the reverberation $\tilde{r}_t$ and that of the noise $\tilde{n}_t$ in Eq. (11) to be negligibly small, i.e., $\tilde{r}_t + \tilde{n}_t \simeq 0$. Then, the first

row and the remaining rows in Eqs. (11) can be considered statistically independent of each other, and thus $p(\mathbf{y}_t; \theta)$ can be decomposed into $p(y_t^{(1)}; \theta_w)$ and $p(\mathbf{y}_t^{(2:M)}; \theta_B)$. Then, the likelihood function can be rewritten as

$$\mathcal{L}(\theta) = \sum_t \log p(y_t^{(1)}; \theta_w, \theta_\sigma) + \sum_t \log p(\mathbf{y}_t^{(2:M)}; \theta_B)$$
$$+ 2T \log \frac{|v^{(1)}|}{||\mathbf{v}||_2} + T \log \det(B_0^{\mathsf{H}} B_0), \quad (20)$$

where the last two terms in Eq. (20) are derived from the last term in Eq. (19) based on Eqs. (9) and (10) (See Appendix), and $|| \cdot ||_2$ denotes the Euclidean norm of a vector.

Note that because it is difficult to optimize $\mathbf{v}$ based on the ML estimation, we estimate it separately from the ML estimation as described in Section IV-B. Then, for the estimation of θ_w, we need only to estimate $\Theta = \{\theta_w, \theta_\sigma\}$, based solely on the first term in Eq. (20).

Finally, we introduce a model of the probability density function (pdf) of $y_t^{(1)}$ as

$$p(y_t^{(1)}; \theta_w, \theta_\sigma) = \mathcal{N}_C(y_t^{(1)} = \bar{\mathbf{w}}^{\mathsf{H}} \bar{\mathbf{x}}_t; 0, \sigma_t^2), \quad (21)$$

where $\mathcal{N}_C(x; \mu, \sigma^2)$ is a pdf of a complex Gaussian distribution with a mean μ and a covariance σ^2, defined as

$$\mathcal{N}_C(x; \mu, \sigma^2) = \frac{1}{\pi \sigma^2} \exp \left(-\frac{|x - \mu|^2}{\sigma^2} \right). \quad (22)$$

Then, the likelihood function to be maximized becomes

$$\mathcal{L}(\Theta) = -\sum_t \frac{|y_t^{(1)}|^2}{\sigma_t^2} - \sum_t \log \sigma_t^2, \quad (23)$$

$$= -\sum_t \frac{|\bar{\mathbf{w}}^{\mathsf{H}} \bar{\mathbf{x}}_t|^2}{\sigma_t^2} - \sum_t \log \sigma_t^2, \quad (24)$$

where we set $\bar{\mathbf{w}} = [\mathbf{w}_0^{\top}, \mathbf{w}_b^{\top}, \mathbf{w}_{b+1}^{\top}, \ldots, \mathbf{w}_{L_w}^{\top}]^{\top}$ and $\bar{\mathbf{x}}_t = [\mathbf{x}_t^{\top}, \mathbf{x}_{t-b}^{\top}, \mathbf{x}_{t-b-1}^{\top}, \ldots, \mathbf{x}_{t-L_w+1}^{\top}]^{\top}$. Note that $\bar{\mathbf{w}}$ and $\bar{\mathbf{x}}_t$ have a time gap between their first and the second elements, corresponding to the prediction delay b.

IV. SOLUTION TO ML ESTIMATION

Because it is difficult to obtain a closed form solution that maximizes Eq. (24), we adopt an iterative estimation scheme, by which Eq. (24) is maximized to a stationary point by alternately updating θ_w and θ_σ from certain initial values.

In the step employed to update θ_w, while fixing $\hat{\sigma}_t^2$ for each t, θ_w is updated as one that maximizes the likelihood function shown below with the distortionless constant in Eq. (9).

$$\mathcal{L}(\theta_w) = -\sum_t \frac{|\bar{\mathbf{w}}^{\mathsf{H}} \bar{\mathbf{x}}_t|^2}{\hat{\sigma}_t^2} \quad \text{s.t.} \quad \mathbf{w}_0^{\mathsf{H}} \mathbf{v} = v^{(1)}, \quad (25)$$

The solution that maximizes Eq. (25) can be derived based on the Lagrange multiplier method as

$$\hat{\bar{\mathbf{w}}} = \frac{R^{-1} \bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}} R^{-1} \bar{\mathbf{v}}}, \quad (26)$$

where $\bar{\mathbf{v}} = [\mathbf{v}^{\top}/v^{(1)}, 0, 0, \ldots, 0]^{\top}$ is a column vector containing $\mathbf{v}/v^{(1)}$ followed by $M(L_w - b + 1)$ zeros, and R is
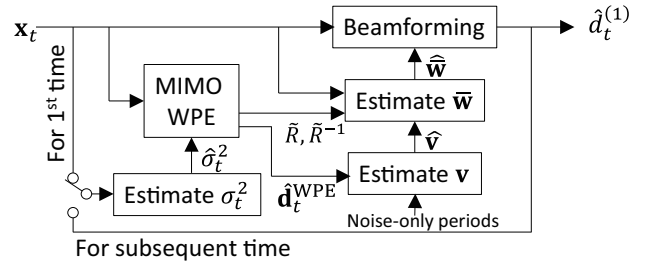

Fig. 1. Iterative beamformer estimation by WPD performed with WPE.

a power-normalized temporal-spatial covariance matrix with a prediction delay calculated as

$$R = \sum_t \frac{\bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^{\mathsf{H}}}{\hat{\sigma}_t^2}. \quad (27)$$

Then, the estimate of the desired signal is obtained as

$$\hat{d}_t^{(1)} = \hat{\bar{\mathbf{w}}}^{\mathsf{H}} \bar{\mathbf{x}}_t. \quad (28)$$

In the other step used to update θ_σ, σ_t^2 can be updated as the power of the estimated desired signal, i.e., $\hat{\sigma}_t^2 = |\hat{d}_t^{(1)}|^2$.

It may be worth noting that the ML formulation presented here is reduced to that for MPDR (and its realization based on a generalized sidelobe canceller) if we set $\mathbf{r}_t = 0$ for all t and $W_\tau = 0$ for $\tau \geq b$ in the formulation and assume that σ_t^2 is time invariant.

A. Analysis of the solution

Because $d_t^{(1)}$, $\tilde{r}_t$, and $\tilde{n}_t$ in Eq. (11) are mutually independent, Eq. (25) can be decomposed, under the distortionless constraint, as

$$\mathcal{L}(\theta_w) = -\sum_t \frac{|d_t^{(1)}|^2}{\hat{\sigma}_t^2} - \sum_t \frac{|\tilde{r}_t|^2}{\hat{\sigma}_t^2} - \sum_t \frac{|\tilde{n}_t|^2}{\hat{\sigma}_t^2}. \quad (29)$$

The first term in Eq. (29) does not depend on θ_w. Thus if we obtain a beamformer θ_w that maximizes Eq. (25) for fixed $\hat{\sigma}_t^2$, it surely maximizes the sum of the second and third terms in Eq. (29). As a consequence, the beamformer can perform denoising and dereverberation simultaneously.

B. Overall processing flow with estimation of $\mathbf{v}$ using WPE

For the accurate estimation of $\mathbf{v}$ from the captured signal, it is crucial to exclude any influence of noise and reverberation. For this purpose, we first reduce the effect of reverberation by performing MIMO dereverberation using WPE [11] within the WPD framework (see Fig. 1). Because WPE and WPD share most of the calculation that requires the majority of their computing cost, namely the calculation of R in Eq. (27) and its inverse, WPE can be performed very efficiently within the WPD framework to obtain the dereverberated multichannel signal, $\hat{\mathbf{d}}_t^{\text{WPE}}$. In concrete terms, let $\tilde{R}$ be an $M(L_w - b + 1) \times M(L_w - b + 1)$ dimensional submatrix of R, obtained by excluding the first M rows and M columns from R. For WPE, we need to calculate $\tilde{R}$ and its inverse, which accounts for the majority of the computing cost of WPE. Then, once we have $\tilde{R}$ and its inverse, we can calculate R and its inverse

TABLE I
CD (DB), FWSSNR (DB), AND WER (%) OF ENHANCED SPEECH
OBTAINED AFTER 1ST ITERATION USING REVERB CHALLENGE EVAL
SET. NO ENH MEANS NO SPEECH ENHANCEMENT. BOLDFACE INDICATES
THE BEST SCORE FOR EACH METRIC.

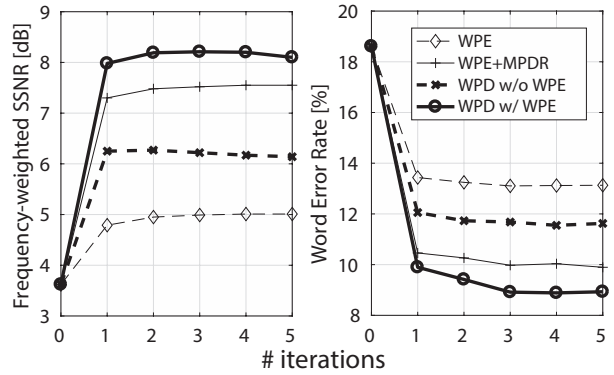|	SimData			RealData
	CD	FWSSNR	WER	WER
No Enh	3.97	3.62	4.35	18.61
MPDR	3.43	5.97	5.56	14.68
WPE	3.74	4.79	4.37	13.44
WPE+MPDR	3.02	7.30	4.42	10.46
WPD w/o WPE	3.23	6.25	4.82	12.06
WPD w/ WPE	**2.65**	**7.98**	**3.83**	**9.90**

Fig. 2. Performance curve of FWSSNRs (dB) and WERs (%) with increases in # of estimation iterations. FWSSNRs are evaluated using SimData of Eval set and WERs are evaluated using RealData of Eval set.

with little additional computing cost [30]. Finally, we estimate $\mathbf{v}$ from $\mathbf{d}_t^{\mathrm{WPE}}$ by reducing the effect of the noise based on generalized eigenvalue decomposition with noise covariance whitening [31], [32], assuming that the noise-only periods are given.

Figure 1 shows the overall processing flow of WPD when utilizing WPE to estimate $\mathbf{v}$. Adopting the power of the captured signal as the initial value of σ_t^2, we iterate WPD jointly with WPE, and update $\mathbf{v}$ and σ_t^2 using the outputs of WPE and WPD, respectively. In each iteration, the beamformer $\hat{\bar{\mathbf{w}}}$ and the desired signal $\hat{d}_t^{(1)}$ are updated based on Eqs. (26), (27), and (28) using the estimated $\hat{\mathbf{v}}$ and $\hat{\sigma}_t^2$.

V. EXPERIMENTS

A. Dataset and evaluation metrics

We evaluated the performance of the proposed method using the REVERB Challenge dataset [19]. The evaluation set (Eval set) of the dataset is composed of simulated data (SimData) and real recordings (RealData). Each utterance in the dataset contains reverberant speech uttered by a speaker and stationary additive noise. The distance between the speaker and the microphone array is varied from 0.5 m to 2.5 m. For SimData, the reverberation time is varied from about 0.25 s to 0.7 s, and the signal-to-noise ratio (SNR) is set at about 20 dB.

Evaluation metrics prepared for the challenge were used in the experiments. As objective measures for evaluating speech enhancement performance [33], we used the cepstrum distance (CD), and the frequency-weighted segmental SNR (FWSSNR). To evaluate the ASR performance, we used a baseline ASR system recently developed using Kaldi [34]. This is a fairly competitive system composed of a TDNN acoustic model trained using lattice-free MMI and online i-vector extraction, and a tri-gram language model.

B. Methods to be compared and analysis conditions

WPD (Proposed) was compared with WPE, MPDR, and the integration of WPE followed by MPDR in a cascade configuration (WPE+MPDR). To confirm the importance of utilizing WPE within the WPD framework, we examined the performance of WPD with and without WPE. The two configurations are respectively referred to as WPD w/ and w/o WPE. Without WPE, WPD estimates $\mathbf{v}$ from the captured signal, and does not update it during the iterative estimation.

For all the methods, a Hann window was used for a short-time analysis with the frame length and the shift set at 32 ms and 8 ms, respectively. The sampling frequency was 16 kHz and $M = 8$ microphones were used for all the experiments. For WPE, WPE+MPDR, and WPD, the prediction delay was set at $b = 4$, and the length of the prediction filter was set at $L_w = 12, 10$, and 6, respectively, for frequency ranges of 0 to 0.8 kHz, 0.8 to 1.5 kHz, 1.5 to 8 kHz. For the estimation of $\mathbf{v}$, we assumed that each utterance had noise-only periods of 225 ms and 75 ms, respectively, at its beginning and ending parts.

C. Evaluation results

Table I summarizes the CDs, FWSSNRs, and WERs of the captured signals and the enhanced signals obtained after the first estimation iteration. In the table, all the methods improved the captured signal with all the measures except for the WERs on SimData. While WPD w/ WPE performed the best of all, WPD w/o WPE did not perform very well. This indicates that the reliable estimation of $\mathbf{v}$ is very important for successful beamforming by WPD, and it can be achieved by utilizing WPE for the estimation of $\mathbf{v}$ within the WPD framework.

Figure 2 shows the performance curve of the methods in terms of FWSSNRs and WERs when we performed the iterative estimation. We confirmed that WPD with WPE again greatly outperformed all the other methods for all the iteration times, and the iterative estimation was effective at least in the first few steps for WPD.

VI. CONCLUDING REMARKS

We presented a probabilistic formulation of WPD that achieves dereverberation and denoising both simultaneously and optimally, and derived an optimization algorithm based on the ML estimation. Furthermore, we proposed a method for effectively estimating the steering vector of the desired signal by incorporating WPE into the WPD framework. The experiments showed that WPD with WPE greatly outperformed conventional approaches, including one that utilizes WPE followed by MPDR in a cascade configuration, and that the incorporation of WPE into the WPD framework is very

important if we are to achieve a reliable estimation of the steering vector for successful beamforming.

APPENDIX
DECOMPOSITION OF $|\det(W_0)|$

Let us decompose the first column of W_0 in Eq. (8) as $\mathbf{w}_0 = \mathbf{e} + \mathbf{b}$, where $\mathbf{e}$ is a projection of $\mathbf{w}_0$ to $\mathbf{v}$, which is determined based on Eq. (9) as

$$\mathbf{e} = \frac{(v^{(1)})^{\mathsf{H}}}{||\mathbf{v}||_2^2}\mathbf{v}, \tag{30}$$

and $\mathbf{b}$ is a component that is orthogonal to $\mathbf{v}$. Note that $\mathbf{b}$ is linearly dependent on the subspace spanned by B_0.

Then, $|\det(W_0)|$ can be expanded as

$$|\det(W_0)| = |\det([\mathbf{e} + \mathbf{b}, B_0])|, \tag{31}$$

$$= |\det([\mathbf{e}, B_0]) + \det([\mathbf{b}, B_0])|, \tag{32}$$

$$= |\det([\mathbf{e}, B_0])|, \tag{33}$$

$$= \det\left(\begin{bmatrix} \mathbf{e}^{\mathsf{H}} \\ B_0^{\mathsf{H}} \end{bmatrix}[\mathbf{e}, B_0]\right)^{1/2}, \tag{34}$$

$$= \det\left(\begin{bmatrix} \mathbf{e}^{\mathsf{H}}\mathbf{e} & 0 \\ 0 & B_0^{\mathsf{H}}B_0 \end{bmatrix}\right)^{1/2}, \tag{35}$$

$$= \det(\mathbf{e}^{\mathsf{H}}\mathbf{e})^{1/2}\det(B_0^{\mathsf{H}}B_0)^{1/2}, \tag{36}$$

$$= \frac{|v^{(1)}|}{||\mathbf{v}||_2}\det(B_0^{\mathsf{H}}B_0)^{1/2}. \tag{37}$$

As a consequence, $|\det(W_0)|$ is decomposed into one based only on $\mathbf{v}$ and another based only on B_0. Note that $\det(B_0^{\mathsf{H}}B_0)^{1/2}$ is equal to the absolute value of the product of all the singular values of B_0.

ACKNOWLEDGMENT

REFERENCES

[1] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[2] H. L. V. Trees, *Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory*. New York: Wiley-Interscience, 2002.

[3] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *The Journal of the Acoustical Society of America*, vol. 54, pp. 771–785, 1973.

[4] T. Higuchi, N. Ito *et al.*, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[5] H. Erdogan, J. R. Hershey *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," *Proc. Interspeech*, pp. 1981–1985, 2016.

[6] S. Emura, S. Araki *et al.*, "Distortionless beamforming optimized with l_1-norm minimization," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 936–940, 2018.

[7] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, 2007.

[8] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformer," *Proc. IEEE ICASSP*, pp. 41–44, 2007.

[9] J. Heymann, L. Drude *et al.*, "Beamnet: end-to-end training of a beamformer-supported multichannel ASR system," *Proc. IEEE ICASSP*, pp. 5235–5329, 2017.

[10] T. Nakatani, T. Yoshioka *et al.*, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[11] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[12] T. Yoshioka, H. Tachibana *et al.*, "Adaptive dereverberation of speech signals with speaker-position change detection," *Proc. IEEE ICASSP*, pp. 3733–3736, 2009.

[13] A. Jukić, T. van Waterschoot *et al.*, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[14] D. Giacobello and T. L. Jensen, "Speech dereverberation based on convex optimization algorithms for group sparse linear prediction," *Proc. IEEE ICASSP*, pp. 446–450, 2018.

[15] M. Delcroix, T. Yoshioka *et al.*, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process*, vol. Article ID 2015:60, doi:10.1186/s13634-015-0245-7, 2015.

[16] W. Yang, G. Huang *et al.*, "Dereverberation with differential microphone arrays and the weighted-prediction-error method," *Proc. IWAENC*, 2018.

[17] M. Togami, "Multichannel online speech dereverberation under noisy environments," *Proc. EUSIPCO*, pp. 1078–1082, 2015.

[18] L. Drude, C. Boeddeker *et al.*, "Integrating neural network based beamforming and weighted prediction error dereverberation," *Proc. Interspeech*, pp. pp. 3043–3047, 2018.

[19] K. Kinoshita, M. Delcroix *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. doi:10.1186/s13634-016-0306-6, 2016.

[20] J. Barker, R. Marxer *et al.*, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," *Proc. IEEE ASRU-2015*, pp. 504–511, 2015.

[21] E. Vincent, S. Watanabe *et al.*, "CHiME4 Challenge," http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/.

[22] J. Barker, S. Watanabe, and E. Vincent, "CHiME5 Challenge," http://spandh.dcs.shef.ac.uk/chime_challenge/.

[23] B. Li, T. N. Sainath *et al.*, "Acoustic modeling for Google Home," *Proc. Interspeech*, 2017.

[24] Audio Software Engineering and Siri Speech Team, "Optimizing Siri on HomePod in far-field settings," *Apple Machine Learning Journal*, vol. 1, no. 12, 2018.

[25] R. Haeb-Umbach, S. Watanabe *et al.*, "Speech processing for digital home assistants," *IEEE Signal Processing Magazine*, 2019.

[26] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 903–907, 2019.

[27] T. Dietzen, S. Doclo *et al.*, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," *Proc. IWAENC*, 2018.

[28] T. Nakatani, T. Yoshioka *et al.*, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," *Proc. IEEE ICASSP*, pp. 85–88, 2008.

[29] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustic Sociaty of America*, vol. 113, pp. 3233–3244, 2003.

[30] T.-T. Lu and S.-H. Shiou, "Inverses of 2 x 2 block matrices," *Computers & Mathematics with Applications*, vol. 43, pp. 119–129, 2002.

[31] N. Ito, S. Araki *et al.*, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," *Proc. IEEE ICASSP*, pp. 681–685, 2017.

[32] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," pp. 544–548, 2015.

[33] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE T-ASLP*, vol. 16, no. 1, pp. 229–238, 2008.

[34] D. Povey, A. Ghoshal *et al.*, "The Kaldi speech recognition toolkit," *Proc. IEEE ASRU*, 2011.