INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer

Tomohiro Nakatani and Keisuke Kinoshita

NTT Communication Science Laboratories, NTT Corporation
tnak@ieee.org, keisuke.kinoshita@ieee.org

Abstract

This article presents frame-by-frame online processing algorithms for a Weighted Power minimization Distortionless response convolutional beamformer (WPD). The WPD unifies widely-used multichannel dereverberation and denoising methods, namely 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 based on maximum likelihood estimation. We derive two different online algorithms, one based on frame-by-frame recursive updating of the spatio-temporal covariance matrix of the captured signal, and the other on recursive least square estimation of the convolutional beamformer. In addition, for both algorithms, the desired signal's relative transfer function (RTF) is estimated by online processing using a neural network based online mask estimation. Experiments using the REVERB challenge dataset show the effectiveness of both algorithms in terms of objective speech enhancement measures and automatic speech recognition (ASR) performance.

Index Terms: dereverberation, denoising, online processing, speech recognition

1. 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 automatic speech recognition (ASR).

Microphone array signal processing techniques have been proposed to minimize the above detrimental effects in the acquired signal. For example, a minimum variance distortionless response beamformer (MVDR) [1–4] and a weighted prediction error based dereverberation method (WPE) [5–8] are widely-used approaches to denoising and dereverberation, respectively. Frame-by-frame online processing has also been developed for these techniques [9–13]. Furthermore, the cascade integration of these techniques has been investigated to perform both denoising and dereverberation [14–17]. Their usefulness, particularly for ASR, has been extensively studied, e.g., at the REVERB challenge [18] and the CHiME-3/4/5 challenges [19–21]. Advances in these techniques have led to recent progress on commercially available devices, such as smart speakers [22–24].

Recently, a promising convolutional beamforming technique, known as a Weighted Power minimization Distortionless response beamformer (WPD), has been proposed [25, 26] that unifies the WPE and a variant of MVDR, namely a minimum power distortionless response beamformer (MPDR) [27–29], into a single convolutional beamformer, and achieves both denoising and dereverberation simultaneously based on maximum likelihood estimation [26] with weighted power minimization with a distortionless constraint. It was experimentally confirmed that this beamformer can substantially outperform the conventional cascade integration of the WPE and the MPDR.

This paper presents frame-by-frame online processing algorithms for the WPD. Two different algorithms, referred to as recursive WPD and recursive least square WPD (RLS-WPD), are derived. While the recursive WPD estimates the spatio-temporal covariance matrix used for the WPD by a recursive update based on the Woodbury matrix identity [30], the RLS-WPD employs the RLS estimation [31] for a frame-by-frame update of the beamformer coefficients adopting a generalized sidelobe canceler (GSC)-like beamformer structure [32]. The latter algorithm can be viewed also as a variant of a Kalman filter based denoising and dereverberation method based on integrated sidelobe canceler and linear prediction (ISCLP) [33]. For both recursive WPD and RLS-WPD, a neural network based online mask estimation is introduced for online estimation of the desired signal's RTF.

Experiments using the REVERB challenge dataset [34] show that both online algorithms for the WPD can greatly improve the quality of the enhanced speech in terms of objective speech enhancement measures and ASR performance without any prior knowledge on the recording conditions, such as a room impulse response, by frame-by-frame online processing, and substantially outperform conventional online beamforming approaches, including the cascade integration of an online WPE and an online MPDR. It may be worth noting that a key to the performance improvement by both algorithms was the use of a multi-input multi-output (MIMO) WPE [6] for the preprocessing of the RTF estimation, as with the conventional WPD [26].

As regards the remainder of this paper, we describe the conventional WPD in Section II. The two algorithms for the online WPD are presented in Section III and related work is summarized in Section IV. The experimental results and concluding remarks are given in Sections V and VI, respectively. Hereafter, we refer to the conventional WPD as batch WPD to distinguish it from the online WPD proposed in this paper. Also, we use the term "online" to represent "frame-by-frame online" for brevity.

2. Conventional method – batch WPD

Assume that a speech signal spoken by a speaker is captured by M microphones in a noisy reverberant environment. Then, with the WPD, the captured signal is modeled in the short time Fourier transform (STFT) domain (see [25] for more detail) as

$$\mathbf{x}_{f,t} = \mathbf{d}_{f,t} + \mathbf{r}_{f,t} + \mathbf{n}_{f,t}, \tag{1}$$

$$\mathbf{d}_{f,t} = \mathbf{v}_f s_{f,t}, \tag{2}$$

$$\mathbf{r}_{f,t} = \sum_{\tau=b}^{L_a+b-1} \mathbf{a}_{f,\tau} s_{f,t-\tau}, \tag{3}$$

where f is a frequency index, t and τ are time frame indices, and $\mathbf{x}_{f,t} = [x_{f,t}^{(1)}, x_{f,t}^{(2)}, \ldots, x_{f,t}^{(M)}]^\top \in \mathbb{C}^M$ contains

http://dx.doi.org/10.21437/Interspeech.2019-1286

the STFT coefficients of the captured signal[1] at all the microphones, where $\top$ denotes the non-conjugate transpose. In eq. (1), $\mathbf{d}_{f,t} + \mathbf{r}_{f,t}$ corresponds to reverberant speech, and $\mathbf{n}_{f,t}$ is additive noise. In $\mathbf{d}_{f,t}+\mathbf{r}_{f,t}$, $\mathbf{d}_{f,t}$ is the sum of the direct signal and the early reflections, and $\mathbf{r}_{f,t}$ is late reverberation [35]. Hereafter, $\mathbf{d}_{f,t}$ is referred to as the desired signal. In eq. (2), assuming that the duration of the early reflections in the time domain is shorter than the analysis window, $\mathbf{d}_{f,t}$ is modeled by a product of the clean speech, $s_{f,t} \in \mathbb{C}$, and a transfer function, $\mathbf{v}_f \in \mathbb{C}^M$, referred to as a steering vector. In contrast, in eq. (3) $\mathbf{r}_{f,t}$ is modeled by a frequency-domain convolution of $s_{f,t}$ with a convolutional transfer function [36], $\mathbf{a}_{f,t} \in \mathbb{C}^M$, assuming that the duration of the late reverberation is longer than the analysis window. b is the time frame index that divides the reverberant speech into the desired signal and the late reverberation, and L_a is the length of the convolutional transfer function.

The goal for the WPD is to reduce $\mathbf{r}_{f,t}$ and $\mathbf{n}_{f,t}$ from $\mathbf{x}_{f,t}$, while keeping, $\mathbf{d}_{f,t}$, unchanged. This paper sets $m = q$ as the reference microphone, and describes a method for estimating the desired signal, $d^{(q)}_{f,t} = v^{(q)}_f s_{f,t}$, at the microphone without loss of generality. The WPD estimates $d^{(q)}_{f,t}$ by applying a convolutional beamformer, $\mathbf{w}_{f,t} = [w^{(1)}_{f,t}, w^{(2)}_{f,t}, \ldots, w^{(M)}_{f,t}]^\top \in \mathbb{C}^M$ for $t = 0, b, b+1, \ldots, L+b-1$, to the captured signal as

$$\hat{d}^{(q)}_{f,t} = \mathbf{w}^{\mathsf{H}}_{f,0}\mathbf{x}_{f,t} + \sum_{\tau=b}^{L+b-1} \mathbf{w}^{\mathsf{H}}_{f,\tau}\mathbf{x}_{f,t-\tau}, \quad (4)$$

$$= \bar{\mathbf{w}}^{\mathsf{H}}_f \bar{\mathbf{x}}_{f,t}, \quad (5)$$

where H denotes the conjugate transpose, L is the length of the convolutional beamformer, b is a prediction delay, corresponding to b in eq. (3), $\bar{\mathbf{x}}_{f,t} = [\mathbf{x}^\top_{f,t}, \mathbf{x}^\top_{f,t-b}, \mathbf{x}^\top_{f,t-b-1}, \ldots, \mathbf{x}^\top_{f,t-L-b+1}]^\top \in \mathbb{C}^{M(L+1)}$, and $\bar{\mathbf{w}}_f = [\mathbf{w}^\top_{f,0}, \mathbf{w}^\top_{f,b}, \mathbf{w}^\top_{f,b+1}, \ldots, \mathbf{w}^\top_{f,L+b-1}]^\top \in \mathbb{C}^{M(L+1)}$. Maximum likelihood estimation of the convolutional beamformer, $\bar{\mathbf{w}}_f$, for the WPD [26] is achieved by the following constrained optimization criterion.

$$\bar{\mathbf{w}} = \arg\min_{\bar{\mathbf{w}}} \sum_t \frac{|\bar{\mathbf{w}}^{\mathsf{H}}_f \bar{\mathbf{x}}_{f,t}|^2}{\sigma^2_{f,t}} \quad \text{s.t.} \quad \mathbf{w}^{\mathsf{H}}_{f,0}\mathbf{v}_f = v^{(q)}_f, \quad (6)$$

where $\sigma^2_{f,t} = |d^{(q)}_{f,t}|^2$ is the power of the desired signal at t. By this estimation, based on the constraint, $\bar{\mathbf{w}}^{\mathsf{H}}_{f,0}\mathbf{v}_f = v^{(q)}_f$, $\bar{\mathbf{w}}_{f,0}$ does not distort the desired signal included in the first term in eq. (4), and based on the use of b and an assumption that a clean speech signal does not have long-term correlation, the second term in eq. (4), or $\mathbf{w}_{f,t}$ for $t > 0$, cannot predict and thus distort the desired signal included in the first term. As a consequence, the WPD reduces the weighted power of the captured signal without distorting the desired signal, resulting in a reduction of the noise and the reverberation. The scale ambiguity in the steering vector estimation means that in practice an RTF [37, 38] is estimated, which is defined as $\tilde{\mathbf{v}}_f = \mathbf{v}_f / v^{(q)}_f$, instead of estimating the steering vector. With an RTF, the constraint in eq. (6) is rewritten simply as $\mathbf{w}^{\mathsf{H}}_{f,0}\tilde{\mathbf{v}}_f = 1$.

Figure 1 illustrates the processing flow of the conventional WPD, which performs simultaneous denoising and dereverberation by batch processing. With the conventional WPD, after the MIMO WPE [6] has been applied to the captured signal to obtain a multichannel dereverberated signal, $\mathbf{z}_{f,t}$, the RTF is estimated based on generalized eigenvalue decomposition with noise covariance whitening [39, 40]. Letting Ψ^{Z}_f and Ψ^{n}_f be time-invariant spatial covariance matrices of the dereverberated signal, $\mathbf{z}_{f,t}$, and that of the noise in $\mathbf{z}_{f,t}$, the WPD estimates the RTF as follows:

$$\dot{\mathbf{v}}_f = \text{MaxEig}((\Psi^{\mathsf{n}}_f)^{-1}\Psi^{\mathsf{Z}}_f) \quad (7)$$

$$\mathbf{v}_f = \Psi^{\mathsf{n}}_f \dot{\mathbf{v}}_f, \quad (8)$$

$$\tilde{\mathbf{v}}_f = \mathbf{v}_f / v^{(q)}_f, \quad (9)$$

where $\text{MaxEig}(\cdot)$ is a function that extracts the eigenvector corresponding to the maximum eigenvalue. This method with the MIMO WPE can reduce the effect of the reverberation and the noise on the RTF estimation.

A power normalized spatio-temporal covariance matrix used in the conventional WPD is defined and estimated as

$$R_f = \sum_t \frac{\bar{\mathbf{x}}_{f,t}\bar{\mathbf{x}}^{\mathsf{H}}_{f,t}}{\sigma^2_{f,t}}, \quad (10)$$

where $\sigma^2_{f,t}$ is estimated from the power of the captured signal as $\sigma^2_{f,t} = \mathbf{x}^{\mathsf{H}}_{f,t}\mathbf{x}_{f,t}/M$ for the sake of simplicity in our experiments. Once $R_{f,t}$ and $\mathbf{v}_{f,t}$ are estimated, the convolutional beamformer is estimated as follows

$$\bar{\mathbf{w}}_f = \frac{R^{-1}_f \bar{\mathbf{v}}_f}{\bar{\mathbf{v}}^{\mathsf{H}}_f R^{-1}_f \bar{\mathbf{v}}_f}, \quad (11)$$

where $\bar{\mathbf{v}}_f = [\tilde{\mathbf{v}}^\top_f, 0, 0, \ldots, 0]^\top \in \mathbb{C}^{M(L+1)}$.

Finally, the beamforming is performed as in eq. (5).

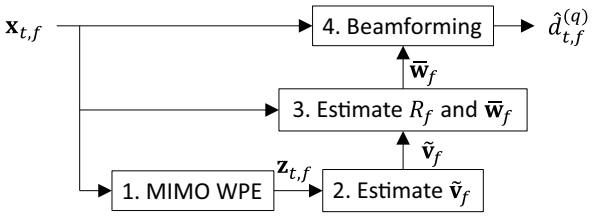

Figure 1: *Processing flow of the WPD. It receives a multichannel captured signal, $\mathbf{x}_{f,t}$, and outputs a single channel denoised and dereverberated signal $\hat{d}^{(q)}_{f,t}$. $\mathbf{z}_{f,t}$, $\tilde{\mathbf{v}}_f$, R_f, and $\bar{\mathbf{w}}_f$ denote a multichannel dereverberated signal, an RTF, a power-normalized spatio-temporal covariance matrix, and a convolutional beamformer, respectively.*

3. Proposed method – online WPD

To achieve online processing for the WPD, it is necessary to make the following three blocks work in an online processing manner.

1. MIMO WPE
2. Estimation of the RTF, $\tilde{\mathbf{v}}_f$
3. Estimation of the spatio-temporal covariance matrix, R_f, and the convolutional beamformer, $\bar{\mathbf{w}}_f$

Because an online algorithm has already been proposed for the MIMO WPE [11, 12], we describe online algorithms of the remaining two blocks in the following. Specifically, we describe two different algorithms for the estimation of R_f and $\bar{\mathbf{w}}_f$, namely a recursive WPD and an RLS-WPD, while presenting a common procedure for the estimation of $\tilde{\mathbf{v}}_f$.

Hereafter, time frame indices are attached to variables to be updated at each time frame, e.g., as $R_{f,t}$, $\mathbf{v}_{f,t}$, $\bar{\mathbf{w}}_{f,t}$, $\Psi^{\mathsf{Z}}_{f,t}$, and $\Psi^{\mathsf{n}}_{f,t}$ for R_f, $\mathbf{v}_f$, $\bar{\mathbf{w}}_f$, Ψ^{Z}_f, and Ψ^{n}_f.

[1] In this paper, a time series of STFT coefficients of a signal are referred to simply as a signal.

3.1. Online RTF estimation

Here, we describe how we can make the RTF estimation method used by the WPD perform online processing. First, $\Psi^{\mathsf{Z}}_{f,t}$ can be estimated recursively at each t as

$$\Psi^{\mathsf{Z}}_{f,t} = \alpha_z \Psi^{\mathsf{Z}}_{f,t-1} + \mathbf{z}_{f,t}\mathbf{z}^{\mathsf{H}}_{f,t}, \tag{12}$$

where α_z is a forgetting factor. $\Psi^{\mathsf{n}}_{f,t}$ and its inverse, on the other hand, are estimated by online processing assuming that masks, $\gamma_{f,t}$, are available. A mask at each time-frequency (TF) point takes a value between 0 and 1, and indicates whether speech or noise dominates the TF point. $\gamma_{f,t} = 1$ means that noise dominates the TF point. With the masks, $\Psi^{\mathsf{n}}_{f,t}$ is recursively updated as in eq. (13) [9], and its inverse, $(\Psi^{\mathsf{n}}_{f,t})^{-1}$, is recursively updated as in eqs. (14) and (15) based on the Woodbury matrix identity [30].

$$\Psi^{\mathsf{n}}_{f,t} = \alpha_n \Psi^{\mathsf{n}}_{f,t-1} + \gamma_{f,t}\mathbf{z}_{f,t}\mathbf{z}^{\mathsf{H}}_{f,t}, \tag{13}$$

$$\mathbf{k}_{f,t} = \frac{\gamma_{f,t}(\Psi^{\mathsf{n}}_{f,t-1})^{-1}\mathbf{z}_{f,t}}{\alpha_n + \gamma_{f,t}\mathbf{z}^{\mathsf{H}}_{f,t}(\Psi^{\mathsf{n}}_{f,t-1})^{-1}\mathbf{z}_{f,t}}, \tag{14}$$

$$(\Psi^{\mathsf{n}}_{f,t})^{-1} = \frac{1}{\alpha_n}\left((\Psi^{\mathsf{n}}_{f,t-1})^{-1} - \mathbf{k}_{f,t}\mathbf{z}^{\mathsf{H}}_{f,t}(\Psi^{\mathsf{n}}_{f,t-1})^{-1}\right), \tag{15}$$

where α_n is a forgetting factor. In the experiments, we use a neural network for the online estimation of the masks.

Because MaxEig$(\cdot)$ in eq. (7) is computationally demanding and thus unsuitable for online processing, we approximate it with the power method. With it, eq. (7) can be calculated by online processing as

$$\dot{\mathbf{v}}_{f,t} = (\Psi^{\mathsf{n}}_{f,t})^{-1}\Psi^{\mathsf{Z}}_{f,t}\dot{\mathbf{v}}_{f,t-1}/\dot{v}^{(q)}_{f,t-1}, \tag{16}$$

Then, $\tilde{\mathbf{v}}_{f,t}$ is obtained at each t in the same way as in eqs. (8) and (9).

3.2. Recursive WPD

With a recursive WPD, the spatio-temporal covariance matrix, $R_{f,t}$, is defined by a recursive form at each time frame t as

$$R_{f,t} = \sum_{\tau=0}^{t} \alpha_R^{t-\tau} \frac{\bar{\mathbf{x}}_{f,\tau}\bar{\mathbf{x}}^{\mathsf{H}}_{f,\tau}}{\sigma^2_{f,\tau}}, \tag{17}$$

where α_R is a forgetting factor. Then, the inverse of $R_{f,t}$, namely $R^{-1}_{f,t}$, can be recursively updated again using Woodbury matrix identity by

$$\mathbf{h}_{f,t} = \frac{R^{-1}_{f,t-1}\bar{\mathbf{x}}_{f,t}}{\alpha_R\sigma^2_{f,t} + \bar{\mathbf{x}}^{\mathsf{H}}_{f,t}R^{-1}_{f,t-1}\bar{\mathbf{x}}_{f,t}}, \tag{18}$$

$$R^{-1}_{f,t} = \frac{1}{\alpha_R}(R^{-1}_{f,t-1} - \mathbf{h}_{f,t}\bar{\mathbf{x}}^{\mathsf{H}}_{f,t}R^{-1}_{f,t-1}), \tag{19}$$

The estimation of $\bar{\mathbf{w}}_{f,t}$ and $d^{(q)}_{f,t}$ can be performed at each time frame in the same way as in eqs. (11) and (5).

3.3. RLS-WPD

An RLS-WPD can be derived by adopting a beamformer structure that is similar to one used for ISCLP [33], namely by introducing a GSC-like beamformer structure [32] to the WPD instead of using the distortionless constraint, $\mathbf{w}^{\mathsf{H}}_{f,0}\tilde{\mathbf{v}}_f = 1$. With this algorithm, $\mathbf{w}_{f,0}$ in eq. (4) is parameterized as

$$\mathbf{w}_{f,0} = (\tilde{\mathbf{v}}^{\mathsf{H}}_f\tilde{\mathbf{v}}_f)^{-1}\tilde{\mathbf{v}}_f + B_f\mathbf{g}_f, \tag{20}$$

where $B_f \in \mathbb{C}^{M\times(M-1)}$ is a blocking matrix that corresponds to the orthogonal complement of $\tilde{\mathbf{v}}_f$, satisfying $B^{\mathsf{H}}_f\tilde{\mathbf{v}}_f = 0$, and $\mathbf{g}_f \in \mathbb{C}^{M-1}$ contains filter coefficients to be optimized. With this structure, $\mathbf{w}_{f,0}$ always satisfies $\mathbf{w}^{\mathsf{H}}_{f,0}\tilde{\mathbf{v}}_f = 1$ for any $\mathbf{g}_f$, and thus the cost function in eq. (6) can be a simple quadratic form with no constraint. Accordingly, we can use the RLS estimation [31] for online beamformer estimation.

The resultant online algorithm at each t becomes

$$\hat{d}^{(q)}_{f,t} = (\tilde{\mathbf{v}}^{\mathsf{H}}_{f,t}\tilde{\mathbf{v}}_{f,t})^{-1}\tilde{\mathbf{v}}^{\mathsf{H}}_{f,t}\mathbf{x}_{f,t} + \tilde{\mathbf{w}}^{\mathsf{H}}_{f,t-1}\tilde{\mathbf{x}}_{f,t}, \tag{21}$$

$$\mathbf{u}_{f,t} = \frac{\tilde{R}^{-1}_{f,t-1}\tilde{\mathbf{x}}_{f,t}}{\alpha_{\tilde{R}}\sigma^2_{f,t} + \tilde{\mathbf{x}}^{\mathsf{H}}_{f,t}\tilde{R}^{-1}_{f,t-1}\tilde{\mathbf{x}}_{f,t}}, \tag{22}$$

$$\tilde{R}^{-1}_{f,t} = \frac{1}{\alpha_{\tilde{R}}}(\tilde{R}_{f,t-1} - \mathbf{u}_{f,t}\tilde{\mathbf{x}}^{\mathsf{H}}_{f,t}\tilde{R}^{-1}_{f,t-1}), \tag{23}$$

$$\tilde{\mathbf{w}}_{f,t} = \tilde{\mathbf{w}}_{f,t} - \mathbf{u}_{f,t}(\bar{d}^{(q)}_{f,t})^{\mathsf{H}}, \tag{24}$$

where $\hat{d}^{(q)}_{f,t}$ is the estimated desired signal, $\tilde{\mathbf{w}}_{f,t} \in \mathbb{C}^{ML+M-1}$ contains estimated convolutional beamformer coefficients, $\tilde{\mathbf{x}}_{f,t} = [(B^{\mathsf{H}}_{f,t}\mathbf{x}_{f,t})^{\top}, \mathbf{x}^{\top}_{f,t-b}, \mathbf{x}^{\top}_{f,t-b-1}, \ldots, \mathbf{x}^{\top}_{f,t-L-b+1}]^{\top} \in \mathbb{C}^{ML+M-1}$, $\alpha_{\tilde{R}}$ is a forgetting factor, and $\tilde{R}_{f,t}$ is defined as

$$\tilde{R}_{f,t} = \sum_{\tau=0}^{t} \alpha_{\tilde{R}}^{t-\tau} \frac{\tilde{\mathbf{x}}_{f,\tau}\tilde{\mathbf{x}}^{\mathsf{H}}_{f,\tau}}{\sigma^2_{f,\tau}}, \tag{25}$$

3.4. Difference between recursive WPD and RLS-WPD

While the recursive WPD and RLS-WPD have a lot in common, they are different in the way to use the estimated RTF. The Recursive WPD estimates the beamformer at t depending only on the RTF estimated at t via eq. (11). In contrast, the RLS-WPD estimates the beamformer at t depending also on the RTF estimated at $\tau < t$, e.g., via the definition of $\tilde{R}_{f,t}$ in eq. (25). This difference makes the behavior of the two algorithms somewhat different with each other.

4. Related Work

As we already described, ISCLP proposed in [33] is closely related with the RLS-WPD derived in this paper, in that both adopt a GSC-like beamformer structure. However, it has been reported in [33] that ISCLP performs no better than the conventional cascade integration of the online WPE and the online GSC. The difference between ISCLP and the RLS-WPD can be summarized as: 1) ISCLP employs a Kalman filter for online processing, while an RLS-WPD employs the RLS estimation [41], 2) ISCLP uses a fixed RTF estimated in advance using an oracle noise covariance matrix, while the RLS-WPD estimates it in an online manner using the neural network-based mask estimation, and 3) ISCLP employs a coherence matrix to model reverberation for the RTF estimation instead of using the MIMO WPE.

Several mask-based online beamformer techniques have been proposed for denoising [9, 10]. The online WPD proposed in this paper can be viewed as an extension of these techniques using a convolutional beamformer for simultaneous denoising and dereverberation.

5. Experiments

5.1. Dataset and evaluation metrics

We evaluated the performance of the proposed method using the REVERB Challenge dataset [34]. The evaluation set (Eval set) of the dataset is composed of simulated data (SimData) and real

Table 1: *CD (dB), FWSSNR (dB), and WER (%) of enhanced speech obtained using REVERB Challenge eval set. No Enh and Batch WPD indicate the performance with no speech enhancement and that obtained with batch WPD, respectively. Each ✓ in the "w/ MIMO WPE" column indicates that the RTF was estimated using the MIMO WPE. Boldface indicates the best score in each column.*

	w/ MIMO WPE	1st pass				2nd pass			
		SimData			RealData	SimData			RealData
		CD	FWSSNR	WER	WER	CD	FWSSNR	WER	WER
No Enh	n/a	3.97	3.62	4.35	18.61	-	-	-	-
Batch WPD	✓	2.65	7.98	3.83	9.90	-	-	-	-
MPDR		3.97	3.62	5.32	16.11	3.86	3.94	5.74	16.16
WPE	n/a	3.81	4.29	4.72	15.64	3.62	4.99	4.26	13.56
WPE + MPDR	✓	3.61	4.96	4.77	14.24	3.50	5.15	4.71	12.53
Recursive WPD		3.42	5.34	4.58	15.16	3.44	5.30	4.65	14.98
RLS-WPD		3.51	5.40	4.49	14.37	3.29	6.08	4.37	12.80
Recursive WPD (proposed)	✓	**3.37**	**6.57**	4.43	**12.99**	3.27	**6.29**	4.3	**11.86**
RLS-WPD (proposed)	✓	3.41	5.66	**4.30**	13.82	**3.21**	6.26	**4.14**	11.88

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 ranges 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 [42], 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 [43]. This is a fairly competitive system composed of a TDNN acoustic model trained using lattice-free MMI and online i-vector extraction, and a trigram language model.

5.2. Methods to be compared and analysis conditions

The recursive WPD and RLS-WPD were compared with WPE, MPDR, and the cascade integration of WPE followed by MPDR (WPE+MPDR), where WPE performs the online MIMO WPE based on the RLS estimation [11, 12], and MPDR performs the online MPDR based on the recursive update of the RTF and that of the spatial covariance matrix. To confirm the importance of the use of WPE for the RTF estimation, i.e., step 1 in Fig. 1, in the online processing, we examined the performance of the recursive WPD and RLS-WPD with and without the online MIMO WPE for the RTF estimation.

To separately evaluate the performance before and after the convergence of online processing, each utterance was passed to each method twice, where the parameters estimated at the end of the first pass were used as the initial values for the second pass. Then, the performances obtained at the 1st and the 2nd passes were taken as those before and after convergence, respectively.

With all the methods, a Hann window was used for a short-time analysis with the frame length and shift set at 64 ms and 16 ms, respectively. The sampling frequency was 16 kHz and $M = 8$ microphones were used for all the experiments. For WPE and WPD, the prediction delay was set at $b = 4$, and the prediction filter lengths were set at $L_w = 12, 10$, and 6, respectively, for frequency ranges of 0 to 0.8 kHz, 0.8 to 1.5 kHz, and 1.5 to 8 kHz. The masks, $\gamma_{f,t}$, which were used for the online estimation of $\dot{\mathbf{v}}_{f,t}$, were estimated using a long-short term memory (LSTM) network capable of performing online mask estimation [44, 45].

As for the initialization, all the elements of $\dot{\mathbf{v}}_{f,0}$ in eq. (16) were set at 1, and all the beamformer coefficients of $\tilde{\mathbf{w}}_{f,t}$ in eq. (24) were initialized to zero. $R_{f,t}^{-1}$, $\tilde{R}_{f,t}^{-1}$, $\Psi_{f,t}^{z}$, and $\Psi_{f,t}^{n}$ were all initialized as identity matrices. The forgetting factors were set at $\alpha_n = \alpha_R = \alpha_{\tilde{R}} = 0.9999$ and $\alpha_z = 0.66$.

5.3. Evaluation results

Table 1 summarizes the CDs, FWSSNRs, and WERs of the captured signals and enhanced signals obtained by the 1st and 2nd passes. In the table, all the methods improved the captured signal with all the measures for both passes except for the WERs for SimData. The recursive WPD and RLS-WPD worked comparably well and substantially outperformed the other methods including the cascade integration of WPE and MPDR. However, they did not perform so well without using the MIMO WPE for the RTF estimation. This indicates that the reliable estimation of the RTF is very important for successful beamforming by online WPD, and it can be achieved by utilizing the MIMO WPE for the estimation. The reason why the WERs for SimData were not well improved by beamforming is probably because the Kaldi ASR baseline was trained on SimData in the training set, and the mismatch between the training and testing was slight.

Note that the two proposed methods greatly improved the enhanced signal even in the 1st pass, and caused a consistent improvement in the 2nd pass. This demonstrates the desirable online processing behavior of the proposed methods.

6. Concluding remarks

This article presented frame-by-frame online processing algorithms for simultaneous denoising and dereverberation by a WPD. Two different algorithms were derived for the online processing: a recursive WPD based on the frame-by-frame recursive updating of the spatio-temporal covariance matrix of the captured signal, and an RLS-WPD based on the recursive least square estimation of a convolutional beamformer. With both algorithms, the desired signal's RTF was estimated by online processing using a neural network based online mask estimation. The experiments showed that the recursive WPD and RLS-WPD both greatly outperformed conventional approaches, including the one utilizing an online WPE followed by an online MPDR in a cascade configuration. It was also shown that the incorporation of the online WPE for the RTF estimation is very important for successful convolutional beamforming.

7. References

[1] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE,*, vol. 57, pp. 1408–1418, 1969.

[2] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 927–935, 1972.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] C. Böddeker, H. Erdogan *et al.*, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," *Proc. IEEE ICASSP*, pp. 6697–6701, 2018.

[10] T. Higuchi, K. Kinoshita *et al.*, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," *Proc. IEEE ICASSP*, pp. 531–535, 2018.

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

[12] J. Caroselli, I. Shafran *et al.*, "Adaptive multichannel dereverberation for automatic speech recognition," *Proc. Interspeech*, pp. 3877–3881, 2017.

[13] J. Heymann, L. Drude *et al.*, "Frame-online DNN-WPE dereverberation," *Proc. IWAENC*, pp. 466–470, 2018.

[14] 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.

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

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

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

[18] 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.

[19] 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.

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

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

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

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

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

[25] 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, 2018.

[26] ——, "Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation," *EUSIPCO*, 2019.

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

[28] 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.

[29] 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.

[30] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Review*, vol. 23, no. 1, pp. 53–60, 1981.

[31] S. Haykin, *Adaptive Filter Theory, fourth edition*. PrenticeHall, 2002.

[32] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly con-strained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

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

[34] K. Kinoshita, M. Delcroix *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," *Proc. IEEE WASPAA*, 2013.

[35] 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.

[36] 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.

[37] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[38] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. on Speech, and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[39] 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.

[40] 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.

[41] S. Haykin, *Adaptive filter theory*, 4th ed. NJ: Prentice Hall, 2002.

[42] 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.

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

[44] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," *Proc. IEEE ICASSP-2016*, pp. 196–200.

[45] Y. Matsui, T. Nakatani *et al.*, "Online integration of DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," *Proc. IWAENC*, pp. 71–75, 2018.