Switching Convolutional Beamformer

Tomohiro Nakatani, Rintaro Ikeshita, Naoyuki Kamo, Keisuke Kinoshita, Shoko Araki, and Hiroshi Sawada
NTT Corporation, Japan

***Abstract*—This paper proposes a time-varying Convolutional BeamFormer (CBF), called a switching CBF, which can capture time-varying characteristics of an observed signal to perform beamforming and dereverberation accurately and simultaneously. With a switching CBF, time frames of a time-varying observed signal are grouped into several clusters, each of which can be taken as time-invariant, and individual clusters are separately processed by different time-invariant CBFs. Conventionally, a switching BeamFormer (BF) and a switching Weighted Prediction Error (WPE) dereverberation filter have been shown effective for the respective problems. This paper presents a method to integrate and jointly optimize them based on Maximum Likelihood (ML) estimation and extends it to work with a Neural Network (NN)-based spectral prior based on Maximum a Posteriori (MAP) estimation. Experiments show that a switching CBF largely outperforms a conventional time-invariant CBF in terms of improved ASR scores.**

***Index Terms*—Beamforming, dereverberation, microphone array, switching system, maximum likelihood estimation**

I. Introduction

When a speech signal is captured by distant microphones, e.g., in a conference room, it often contains such interference signals as reverberation, diffuse noise, and voices from extraneous speakers. They all reduce the intelligibility of the captured speech and often cause serious degradation in many speech applications, such as hands-free teleconferencing and Automatic Speech Recognition (ASR).

Mask-based BeamFormers (BFs) [1]–[3] have been actively studied to minimize the aforementioned interference signals in acquired signals. Masks indicate the time-frequency (TF) regions that are dominated by target speakers' voices and are used to estimate acoustic transfer functions (ATFs) from the speakers to microphones. Many useful techniques have been proposed to estimate masks, e.g., neural networks (NNs) [3], [4] and clustering of microphone array signals [5], [6]. The mask-based BF approach effectively optimizes BFs and Convolutional BFs (CBFs) that can jointly perform denoising, dereverberation, and source separation [7]–[9].

On the other hand, considering that the interference signals to be reduced are time-varying, e.g., due to their time-varying power and presence, a time-varying BF largely outperforms a time-invariant BF in terms of estimation accuracy [10], [11]. An efficient way to implement this idea is to use a switching BF [12]. A switching BF is modeled by a weighed sum of a set of time-invariant BFs and achieves time-varying beamforming making the weight time-varying. The weight and BF coefficients are jointly optimized by minimizing the noise power in the observed signal [13], [14]. Also, a time-varying Weighted Prediction Error (WPE) dereverberation filter with a switching mechanism, called a switching WPE filter [15], outperformed a conventional time-invariant WPE filter [16], [17].

To establish a comprehensive framework of a mask-based CBF [8], [18], this paper presents a new formulation called a switching CBF that incorporates the switching mechanism. It consists of a switching weighted Minimum-Power Distortionless Response (wMPDR) BF and a switching WPE filter [15] and jointly optimizes them assuming that the sources are time-varying Gaussians with time-varying variances, similar to a conventional mask-based CBF [8]. The optimization algorithm is derived based on Maximum Likelihood (ML) estimation and extended to work with an NN-based spectral prior based on Maximum a Posteriori (MAP) estimation. Experiments with noisy reverberant speech mixtures show that the proposed switching CBF largely outperforms the conventional state-of-the-art, a mask-based CBF [8], in terms of improved ASR performance.

The following are the contribution of this paper: 1) the formulation of a switching BF/CBF based on a time-varying Gaussian source model; 2) derivation of an algorithm for jointly optimizing switching BFs and a switching WPE filter; and 3) experimental evaluation of the new framework. In the remainder of this paper, a probabilistic model of a time varying CBF and its implementation using a switching mechanism are described in Sections 2 and 3. Section 4 derives the optimization algorithm. Experiments and concluding remarks are given in Sections 5 and 6.

II. Probabilistic model of time-varying CBF

This section presents a probabilistic model of a time-varying CBF for developing a switching CBF in the next section. Suppose that N source signals are captured by M distant microphones with reverberation and background noise. Let $x_{m,t,f}$ be the captured signal at the mth microphone and a time-frequency point (t, f) in the short-time Fourier transform (STFT) domain, and let $(\cdot)^\top$ denote a non-conjugate transpose, and the captured signal at all the microphones, $\mathbf{x}_{t,f} = [x_{1,t,f}, \ldots, x_{M,t,f}]^\top \in \mathbb{C}^M$, is modeled:

$$\mathbf{x}_{t,f} = \sum_{n=1}^{N} \mathbf{d}_{n,t,f} + \sum_{n=1}^{N} \mathbf{l}_{n,t,f} + \mathbf{v}_{t,f}, \tag{1}$$

$$\mathbf{d}_{n,t,f} = \mathbf{h}_{n,f} s_{n,t,f} \quad \text{for all } n, \tag{2}$$

where $\mathbf{d}_{n,t,f} = [d_{n,1,t,f}, \ldots, d_{n,M,t,f}]^\top \in \mathbb{C}^M$ is the direct signal plus the early reflections of the nth source [19], [20], $\mathbf{l}_{n,t,f}$ is the source's late reverberation, and $\mathbf{v}_{t,f}$ is the diffuse noise. This paper deals with $\mathbf{d}_{n,t,f}$ for each n as a signal to be estimated, called a desired signal, and models it by a

product of a time-invariant Acoustic Transfer Function (ATF) $\mathbf{h}_{n,f} \in \mathbb{C}^M$ and nth clean source signal $s_{n,t,f} \in \mathbb{C}$ in Eq. (2). Note that for estimating each desired signal, $\mathbf{d}_{n,t,f}$, the other components in Eq. (1) are time-varying interference signals to be reduced. Hereafter, we omit frequency index f in all the symbols assuming that the same processing is independently applied to each frequency.

We estimate the desired signals by defining a time-varying CBF as

$$\begin{bmatrix} \mathbf{y}_t \\ \tilde{\mathbf{v}}_t \end{bmatrix} = \begin{bmatrix} \mathbf{W}_t & \mathbf{B}_t \\ \bar{\mathbf{W}}_t & \bar{\mathbf{B}}_t \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{x}_t \\ \bar{\mathbf{x}}_t \end{bmatrix} \in \mathbb{C}^M, \tag{3}$$
$$\bar{\mathbf{x}}_t = [\mathbf{x}_{t-D}^\top, \ldots, \mathbf{x}_{t-L+1}^\top]^\top \in \mathbb{C}^{M(L-D)},$$

where $\mathbf{y}_t = [y_{1,t}, \ldots, y_{N,t}]^\top \in \mathbb{C}^N$ is an estimate of the desired signals at the reference channel [21], i.e., the estimate of $\mathbf{d}_{r,t} = [d_{1,r,t}, \ldots, d_{N,r,t}]^\top$, letting r be the index of the reference channel, $\mathbf{W}_t \in \mathbb{C}^{M\times N}$ and $\bar{\mathbf{W}}_t \in \mathbb{C}^{M(L-D)\times N}$ are the CBF's time-varying coefficient matrices applied to the current captured signal $\mathbf{x}_t$ and the past captured signal sequence $\bar{\mathbf{x}}_t$, and $(\cdot)^{\mathsf{H}}$ is an Hermitian transpose. Here prediction delay D (≥ 1) is introduced to set the dereverberation goal to reduce only the late reverberation and preserve the desired signals [16]. In contrast, $\tilde{\mathbf{v}}_t \in \mathbb{C}^{M-N}$ is an auxiliary output corresponding to a noise estimate, and $\mathbf{B}_t \in \mathbb{C}^{M\times(M-N)}$ and $\bar{\mathbf{B}}_t \in \mathbb{C}^{M(L-D)\times(M-N)}$ are auxiliary coefficient matrices to generate $\tilde{\mathbf{v}}_t$. They are just introduced for deriving a ML objective in the following.

To derive the ML objective, we assume that a certain desired CBF satisfies the following conditions:

1) Let $\mathbf{w}_{n,t}$ be the nth column of $\mathbf{W}_t$ and assume that the ATF $\mathbf{h}_n$ of the nth source is given (or can be estimated). Then $\mathbf{w}_{n,t}$ and $\mathbf{B}_t$ for all n and t satisfy
$$\mathbf{w}_{n,t}^{\mathsf{H}}\mathbf{h}_n = h_{n,r} \quad \text{and} \quad \mathbf{B}_t^{\mathsf{H}}\mathbf{h}_n = \mathbf{0}, \tag{4}$$
i.e., $\mathbf{w}_{n,t}$ does not modify the nth desired signal at reference channel r in $\mathbf{x}_t$ (distortionless constraint), and $\mathbf{B}_t$ blocks all the desired signals in the estimated noise (blocking constraint).
2) The CBF outputs, $y_{n,t}$ and $\tilde{\mathbf{v}}_t$ for all n and t, are mutually independent, and satisfy
$$p(\{y_{n,t}\}_{n,t}, \{\tilde{\mathbf{v}}_{t'}\}_{t'}) = \prod_{n,t} p(y_{n,t}) \prod_{t'} p(\tilde{\mathbf{v}}_{t'}), \tag{5}$$
3) Each $y_{n,t}$ can be modeled by a time-varying Gaussian with a mean zero and time-varying variance $\lambda_{n,t}$ as
$$p(y_{n,t}; \lambda_{n,t}) = \frac{1}{\pi\lambda_{n,t}} \exp\left(-\frac{|y_{n,t}|^2}{\lambda_{n,t}}\right). \tag{6}$$

Then we obtain the ML objective to be maximized to estimate the CBF by following the discussions in our previous papers [18], [22, proposition 7] and disregarding the terms unrelated with $\mathbf{y}_t$:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}\left(\frac{|y_{n,t}|^2}{\lambda_{n,t}} + \log\lambda_{n,t}\right) \tag{7}$$
$$\text{s.t.} \quad \mathbf{w}_{n,t}^{\mathsf{H}}\mathbf{h}_n = h_{n,r} \quad \text{for all } n \text{ and } t,$$

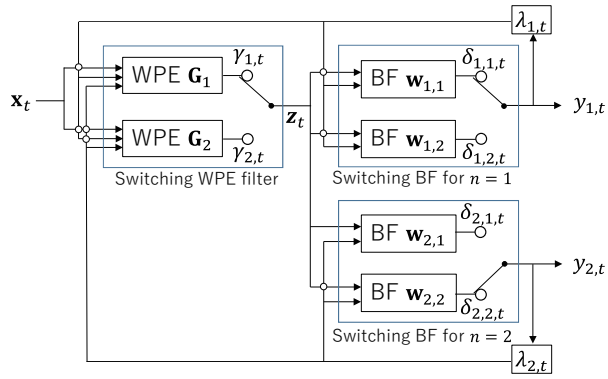


Fig. 1. Processing flow of a switching CBF

where $\theta = \{\{\lambda_{n,t}\}_{n,t}, \{\mathbf{W}_t\}_t, \{\bar{\mathbf{W}}_t\}_t\}$. Although the discussions in previous papers [18], [22] were given for a time-invariant system, extending them for the time-varying CBF is straightforward.

Finally, similar to a conventional mask-based CBF [7], [8], we introduce a factorized form of the time-varying CBF as

$$\begin{bmatrix} \mathbf{W}_t \\ \bar{\mathbf{W}}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I}_M \\ -\mathbf{G}_t \end{bmatrix} \mathbf{W}_t, \tag{8}$$

where $\mathbf{G}_t \in \mathbb{C}^{M(L-D)\times M}$ is a coefficient matrix that satisfies $\bar{\mathbf{W}}_t = -\mathbf{G}_t\mathbf{W}_t$, and $\mathbf{I}_M \in \mathbb{R}^{M\times M}$ is an identity matrix. Using this factorization, $\mathbf{y}_t$ in Eq. (3) is obtained as

$$\mathbf{z}_t = \mathbf{x}_t - \mathbf{G}_t^{\mathsf{H}}\bar{\mathbf{x}}_t, \tag{9}$$
$$\mathbf{y}_t = \mathbf{W}_t^{\mathsf{H}}\mathbf{z}_t, \tag{10}$$

where Eq. (9) is a WPE filter yielding a dereverberated signal $\mathbf{z}_t$ from $\mathbf{x}_t$ using a prediction matrix $\mathbf{G}_t$ and Eq. (10) is a (non-convolutional) BF $\mathbf{W}_t$ that extracts $\mathbf{y}_t$ from $\mathbf{z}_t$.

III. Configurations of Switching CBF

Because the above time-varying CBF is so flexible that overfitting to the observed signal can easily happen, we need to introduce certain constraints to avoid it. For this purpose, we introduce switching mechanisms into the WPE filter and the BFs (Fig. 1). They are composed of a set of time-invariant WPE filters and time-invariant BFs, which are controlled by separate time-varying switches, and mathematically modeled by the sums of the time-invariant filters with time-varying switching weights as

$$\mathbf{G}_t = \sum_{i=1}^{I}\gamma_{i,t}\mathbf{G}_i \quad \text{and} \quad \mathbf{w}_{n,t} = \sum_{j=1}^{J}\delta_{n,j,t}\mathbf{w}_{n,j}, \tag{11}$$

where I and J are the numbers of the switching states of the WPE filter and the BFs, $\mathbf{G}_i$ for $1 \leq i \leq I$ is a prediction matrix of the ith time-invariant WPE filter, $\mathbf{w}_{n,j}$ for $1 \leq j \leq J$ is the jth time-invariant BF for the nth source, and $\gamma_{i,t} \in \mathbb{R}$ for $0 \leq \gamma_{i,t} \leq 1$ and $\delta_{n,j,t} \in \mathbb{R}$ for $0 \leq \delta_{n,j,t} \leq 1$ are their time-varying switching weights satisfying $\sum_{i=1}^{I}\gamma_{i,t} = 1$ and $\sum_{j=1}^{J}\delta_{n,j,t} = 1$. In this paper, for brevity, we only consider

hard switches, and allow $\gamma_{i,t}$ and $\delta_{n,j,t}$ to take only binary values, 0 or 1.

In the above configuration, we introduced separate switches into the WPE filter and BFs, considering that the interference signals to be reduced by the respective filters have different time-varying characteristics. Also, a time-varying WPE filter was shared by all the BFs. Different configurations could be considered, for example, by making certain parts of the switches work synchronously, and/or by introducing different WPE filters separately for individual BFs. Such configurations might be investigated in future work.

IV. Optimization algorithm

This section describes the algorithm that optimizes the switching CBF based on the ML objective in Eq. (7). We next present a few variations, including one that works with an NN-based spectral prior based on MAP estimation.

A. Joint optimization algorithm

Based on Eqs. (9), (10), and (11), the output of switching CBF $y_{n,t}$ can be written as

$$\mathbf{z}_{i,t} = \mathbf{x}_t - \mathbf{G}_i^{\mathsf{H}} \bar{\mathbf{x}}_t, \tag{12}$$

$$y_{n,i,j,t} = \mathbf{w}_{n,j}^{\mathsf{H}} \mathbf{z}_{i,t}, \tag{13}$$

$$y_{n,t} = \sum_{i,j} \delta_{n,j,t} \gamma_{i,t} y_{n,i,j,t}. \tag{14}$$

Here, the parameters to be optimized consist of the following parameter subsets: $\Theta_{\mathbf{G}} = \{\mathbf{G}_i\}_i$, $\Theta_{\mathbf{w}} = \{\mathbf{w}_{n,j}\}_{n,j}$, $\Theta_\gamma = \{\gamma_{i,t}\}_{i,t}$, $\Theta_\delta = \{\delta_{n,j,t}\}_{n,j,t}$, and $\Theta_\lambda = \{\lambda_{n,t}\}_{n,t}$.

Because no closed form solution is known for the optimization, we use iterative estimation based on a coordinate ascent method [23]. It updates each parameter subset alternately by fixing the other parameter subsets, and iterates the update until convergence is obtained. The following describes each update step in the iteration.

1) Updates of $\Theta_{\mathbf{G}}$ and Θ_γ: First, extracting the terms related with $\Theta_{\mathbf{G}}$ and Θ_γ from Eq. (7), we obtain

$$\mathcal{L}(\Theta_{\mathbf{G}}, \Theta_\gamma) = -\sum_{t,n,i,j} \frac{\delta_{n,j,t}\gamma_{i,t}}{\lambda_{n,t}} |\mathbf{w}_{n,j}^{\mathsf{H}} \left(\mathbf{x}_t - \mathbf{G}_i^{\mathsf{H}} \bar{\mathbf{x}}_t\right)|^2. \tag{15}$$

Since the above equation is a simple quadratic form in terms of $\mathbf{G}_i$, we can obtain a closed form solution for it when fixing the other parameters. Let $\mathbf{g}_i = \mathrm{vec}(\mathbf{G}_i)$, where $\mathrm{vec}(\mathbf{A})$ is an operation to reshape a matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ to a vector $\mathbf{a} = [\mathbf{a}_1^\top, \ldots, \mathbf{a}_M^\top]^\top$. Then the solution is given by

$$\mathbf{g}_i \leftarrow \Psi_i^+ \mathrm{vec}(\Phi_i) \in \mathbb{C}^{M^2(L-D)}, \tag{16}$$

where $(\cdot)^+$ denotes a pseudo-inverse,[1]

$$\Psi_i = \sum_{j,n} \left(\mathbf{w}_{n,j}\mathbf{w}_{n,j}^{\mathsf{H}}\right)^* \otimes \mathbf{R}_{n,i,j} \in \mathbb{C}^{M^2(L-D)\times M^2(L-D)}, \tag{17}$$

$$\Phi_i = \sum_{j,n} \mathbf{P}_{n,i,j} \left(\mathbf{w}_{n,j}\mathbf{w}_{n,j}^{\mathsf{H}}\right) \in \mathbb{C}^{M(L-D)\times M}, \tag{18}$$

[1]We used diagonal loading for calculating a pseudo-inverse in our experiments.

$$\mathbf{R}_{n,i,j} = \sum_t \frac{\delta_{n,j,t}\gamma_{t,i}}{\lambda_{n,t}} \bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^{\mathsf{H}} \in \mathbb{C}^{M(L-D)\times M(L-D)}, \tag{19}$$

$$\mathbf{P}_{n,i,j} = \sum_t \frac{\delta_{n,j,t}\gamma_{i,t}}{\lambda_{n,t}} \bar{\mathbf{x}}_t \mathbf{x}_t^{\mathsf{H}} \in \mathbb{C}^{M(L-D)\times M}, \tag{20}$$

$(\cdot)^*$ is a complex conjugate, and $\otimes$ is a Kronecker product. After updating $\mathbf{z}_{i,t}$ by Eq. (12), $\gamma_{i,t}$ is updated by

$$\gamma_{i,t} \leftarrow \begin{cases} 1 & \text{for } \; i = \mathrm{argmin}_{i'} \sum_n \frac{|\sum_j \delta_{n,j,t}\mathbf{w}_{n,j}^{\mathsf{H}}\mathbf{z}_{i',t}|^2}{\lambda_{n,t}} \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

2) Updates of $\Theta_{\mathbf{w}}$, Θ_δ, and Θ_λ: Extracting terms related with $\Theta_{\mathbf{w}}$ and Θ_δ from Eq. (7) yields

$$\mathcal{L}(\Theta_{\mathbf{w}}, \Theta_\delta) = -\sum_{n,j} \mathbf{w}_{n,j}^{\mathsf{H}} \Sigma_{n,j} \mathbf{w}_{n,j}, \tag{22}$$

$$\Sigma_{n,j} = \sum_t \frac{\delta_{n,j,t}}{\lambda_{n,t}} \mathbf{z}_t \mathbf{z}_t^{\mathsf{H}}, \tag{23}$$

$$\text{s.t.} \quad \mathbf{w}_{n,j}^{\mathsf{H}} \mathbf{h}_n = h_{n,r} \quad \text{for all } n \text{ and } j,$$

where $\mathbf{z}_t = \sum_i \gamma_{i,t}\mathbf{z}_{i,t}$ is the output of the switching WPE filter. Because the above objective is identical to that of a wMPDR BF [7], [24] except that it includes a time-varying weight $\delta_{n,j,t}$, we call the BF a switching wMPDR BF. $\mathbf{w}_{n,j}$, which maximizes Eq. (22) when fixing the other parameters, is obtained as

$$\mathbf{w}_{n,j} \leftarrow \frac{h_{n,r}^* \Sigma_{n,j}^{-1} \mathbf{h}_n}{\mathbf{h}_n^{\mathsf{H}} \Sigma_{n,j}^{-1} \mathbf{h}_n}. \tag{24}$$

Then $\delta_{n,j,t}$ is updated as

$$\delta_{n,j,t} \leftarrow \begin{cases} 1 & \text{if } \; j = \mathrm{argmin}_{j'} |\mathbf{w}_{n,j'}^{\mathsf{H}} \mathbf{z}_t|^2, \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

Finally, $\lambda_{n,t}$ is updated as

$$y_{n,t} = \sum_j \delta_{n,j,t} \mathbf{w}_{n,j}^{\mathsf{H}} \mathbf{z}_t, \tag{26}$$

$$\lambda_{n,t} \leftarrow |y_{n,t}|^2. \tag{27}$$

B. Variations

Here we introduce three variations of the above proposed algorithm for an ablation evaluation of the proposed switching CBF in experiments.

1) Switching wMPDR BF: The first is just composed of a switching BF and optimized based on the objective in Eq. (7). The solution is obtained by Eqs. (23)-(27) and by replacing $\mathbf{z}_t$ in the equations with $\mathbf{x}_t$. This is a switching wMPDR BF. It is different from a conventional switching BF [12] because it is based not on the minimization of the noise power but on the ML objective, which considers the time-varying characteristics of the source variances.

2) Separate optimization of switching CBF: The second variation uses the switching CBF structure in Fig. 1 although it separately optimizes the WPE filter and the BFs. The WPE part is identical to the conventional switching WPE filter [15], and the BF part is identical to the switching wMPDR BF; they are connected in a cascade configuration. Overall optimality is not guaranteed with this variation.

TABLE I

WERs (%) obtained with various #states of a WPE filter and BFs after five iterations of separate and joint optimizations w/ and w/o an NN spectral prior using $M = 2$ and 3 microphones. WER of captured signals with no speech enhancement was 62.5%.

	$M = 2$												$M = 3$											
	Separate optimization						Joint optimization						Separate optimization						Joint optimization					
I	w/o NN prior			w/ NN prior			w/o NN prior			w/ NN prior			w/o NN prior			w/ NN prior			w/o NN prior			w/ NN prior		
(#States	J (#States of BFs)												J (#States of BFs)											
of WPE)	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
No WPE	48.0	45.2	*44.0*	48.3	45.8	*44.0*	48.0	45.2	*44.0*	48.3	45.8	*44.0*	39.6	36.4	*35.2*	40.2	36.4	*36.1*	39.6	36.4	*35.2*	40.2	36.4	*36.1*
1	**43.5**	**40.2**	***38.8***	**44.5**	**41.7**	***39.5***	42.5	39.5	***38.3***	42.8	39.6	*38.3*	**33.8**	**31.1**	***29.8***	**34.5**	**30.8**	***29.8***	32.7	29.4	***29.0***	33.4	***29.1***	29.4
2	44.4	40.9	*40.2*	45.5	**41.7**	*40.8*	**41.3**	**39.3**	*38.4*	**40.4**	**38.2**	***37.8***	35.2	32.5	*31.4*	36.6	32.3	*30.7*	**32.1**	***29.2***	29.6	**32.2**	29.7	***28.8***

3) Using an NN-based spectral prior: The third variation introduces a power spectral prior to each source model using an NN. We adopt inverse-Gamma distribution $IG(\lambda; \alpha, \beta) = \beta^{\alpha}\Gamma(\alpha)^{-1}\lambda^{-(\alpha+1)}\exp(-\beta/\lambda)$ as the conjugate prior of the Gaussian source model and set a source power spectrum, $\eta_{n,t}$, estimated by an NN to β. Then instead of using Eq. (27), $\lambda_{n,t}$ is updated based on MAP estimation by

$$\lambda_{n,t} \leftarrow (|y_{n,t}|^2 + \eta_{n,t})/(\alpha + 2). \tag{28}$$

In experiments, we set $\alpha = 1$ and adopted an NN used in our previous paper [8].

V. Experiments

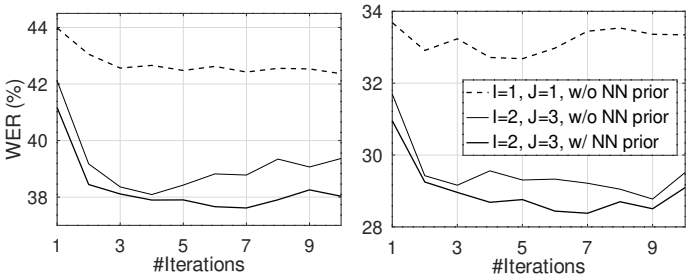
This section experimentally evaluates the performance of the switching CBF with a particular focus on the effect of the switching mechanism in combination with joint optimization and the NN prior (Section IV-B3).

A. Dataset, methods compared, and evaluation metrics

To evaluate the estimated source signals, we used the REVERB-2MIX dataset [25], which is composed of noisy reverberant speech mixtures. Each mixture was created by mixing two utterances (i.e., $N = 2$), extracted from the REVERB Challenge dataset (REVERB) [26]. Following the REVERB-2MIX guideline, evaluation was performed using separated signals that correspond to the evaluation set in REVERB.

To examine the effect of the switching mechanism, the number of switching states was varied over $I = 1, 2$ for a WPE filter and $J = 1, 2, 3$ for each wMPDR BF. We also examined the wMPDR BF without combining it with a WPE filter. ATF $\mathbf{h}_n$ of each speech source was estimated from the input of each wMPDR BF based on a method from our previous paper [8], i.e., based on time-frequency masks obtained using an NN and eigenvalue decomposition with noise covariance whitening [27], [28]. We set the frame length and the shift to 32 and 8 ms and used a Hann window for the short-time analysis. The sampling frequency was 16 kHz. For a WPE filter, the prediction delay was set at $D = 2$ and the prediction filter length was set at $L = 10$. The total number of optimization iterations was set at 10, and we updated the BFs twice in each iteration as it improved the performance.

We evaluated the ASR scores of the separated utterances using RealData in REVERB-2MIX. We used a baseline ASR system developed for REVERB with Kaldi [29] that was composed of a trigram language model, and a TDNN acoustic model trained using a lattice-free MMI and online i-vector extraction. They were trained on the REVERB training set.



(a) $M = 2$ (b) $M = 3$

Fig. 2. WERs (%) obtained after each iteration step of joint optimization using a conventional CBF ($I = 1$ and $J = 1$) and switching CBFs ($I = 2$ and $J = 3$) with/without a NN prior.

B. Evaluation results

Table I shows the WERs of the enhanced signals obtained after five iterations of separate and joint optimizations with various combinations of I and J with/without the NN prior. "No WPE" with $J = 1$ represents a conventional wMPDR BF [7] and $(I, J) = (1, 1)$ represents a conventional CBF composed of a conventional WPE filter followed by a conventional wMPDR BF [8]. "No WPE" with $J \geq 2$ represents a switching wMPDR BF, and the others are switching CBFs.

Let us first look at the results with $M = 2$ microphones on the table's left half. Increasing the number of wMPDR BF states, J, almost always reduced the WERs under all the conditions. Although using a conventional WPE filter ($I = 1$) as pre-processing always reduced the WERs from "No WPE," using a switching WPE filter ($I = 2$) as pre-processing with the separate optimization increased the WERs with/without the NN-prior. This result may be because the time-varying filtering of the switching WPE filter unfavorably affected the performance of the following switching BFs. In contrast, using joint optimization almost consistently improved the WERs under identical conditions, especially when combined with the NN prior. This result suggests that joint optimization effectively mitigated the unfavorable effect on the BFs caused by the time-varying filtering of the switching WPE filter.

The right half of the table shows the results with $M = 3$ microphones. Although more exceptions are included in them,

they overall have the same tendency as those with $M = 2$ microphones.

Finally, Fig. 2 shows the convergence curves of the WERs obtained using the conventional CBF and switching CBFs with/without the NN prior. All the CBFs were estimated based on joint optimization. The figure clearly demonstrates the effectiveness of the switching mechanism and the NN prior introduced to the switching CBFs.

VI. Concluding remarks

This paper proposed a switching CBF that can capture the time-varying characteristics of observed signals to perform accurate and simultaneous dereverberation and beamforming. A switching CBF is composed of a switching WPE filter followed by switching wMPDR BFs, and jointly optimizes their time-varying switching weights and all the filter coefficients. Experiments showed that 1) the switching BFs with/without a WPE filter consistently improved the performance from conventional BFs in terms of ASR scores and 2) the switching WPE filter further improved the performance when it was jointly optimized with the switching BFs, especially in combination with an NN spectral prior. The proposed switching CBF (joint optimization with an NN prior) improved the WERs from 42.5% to 37.8% for $M = 2$ and from 32.7% to 28.8% for $M = 3$ when compared with the best scores obtained by a conventional CBF (joint optimization without an NN prior).

References

[1] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2010.

[2] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE ICASSP*, 2010, pp. 241–244.

[3] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multichannel ASR system," in *Proc. IEEE ICASSP*, 2017, pp. 5235–5329.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 708–712.

[5] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," *Proc. EUSIPCO*, pp. 1153–1157, 2016.

[6] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/Paderborn university joint investigation for dinner party ASR," in *Proc. Interspeech*, 2019, pp. 1248–1252.

[7] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. IEEE ICASSP*, 2020.

[8] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 2020.

[9] W. Zhang, A. S. Subramanian, X. Chang, S. Watanabe, and Y. Qian, "End-to-end far-field speech recognition with unified dereverberation and beamforming," in *Proc. Interspeech*, 2020.

[10] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *Proc. IEEE ICASSP*, 2019, pp. 6855–6859.

[11] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *Proc. IEEE ICASSP*, 2021.

[12] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *Proc. IEEE ICASSP*, 2019, pp. 7908–7912.

[13] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

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

[15] R. Ikeshita, N. Kamo, and T. Nakatani, "Blind signal dereverberation based on mixture of weighted prediction error models," *IEEE Signal Processing Letters*, 2021.

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "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.

[17] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multichannel 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.

[18] T. Nakatani and K. Kinoshita, "Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation," in *Proc. EUSIPCO*, 2019.

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

[20] T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama, "Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria," in *Proc. Interspeech*, 2007, pp. 1082–1085.

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

[22] R. Ikeshita, T. Nakatani, and S. Araki, "Block coordinate descent algorithms for auxiliary-function-based independent vector extraction," *arXiv:2010.08959*, 2021.

[23] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[24] B. J. Cho, J. Lee, and H. Park, "A beamforming algorithm based on maximum likelihood of a complex Gaussian distribution with time-varying variances for robust speech recognition," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1398–1402, August 2019.

[25] "REVERB-2MIX," https://github.com/nttcslab-sp/REVERB-2MIX/.

[26] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "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*, 2016.

[27] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. ASLP*, vol. 17, no. 6, pp. 1071–1086, 2009.

[28] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, pp. 37–51, May 2018.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.