Jointly optimal denoising, dereverberation, and source separation

Tomohiro Nakatani, *Senior Member, IEEE,* Christoph Boeddeker, *Student Member, IEEE,* Keisuke Kinoshita, *Senior Member, IEEE,* Rintaro Ikeshita, *Member, IEEE,* Marc Delcroix, *Senior Member, IEEE,* Reinhold Haeb-Umbach, *Fellow, IEEE*

***Abstract*—This paper proposes methods that can optimize a Convolutional BeamFormer (CBF) for jointly performing denoising, dereverberation, and source separation (DN+DR+SS) in a computationally efficient way. Conventionally, cascade configuration composed of a Weighted Prediction Error minimization (WPE) dereverberation filter followed by a Minimum Variance Distortionless Response (MVDR) beamformer has been used as the state-of-the-art frontend of far-field speech recognition, however, overall optimality of this approach is not guaranteed. In the blind signal processing area, an approach for jointly optimizing dereverberation and source separation (DR+SS) has been proposed, however, this approach requires huge computing cost, and has not been extended for application to DN+DR+SS. To overcome the above limitations, this paper develops new approaches for jointly optimizing DN+DR+SS in a computationally much more efficient way. To this end, we first present an objective function to optimize a CBF for performing DN+DR+SS based on the maximum likelihood estimation, on an assumption that the steering vectors of the target signals are given or can be estimated, e.g., using a neural network. This paper refers to a CBF optimized by this objective function as a weighted Minimum-Power Distortionless Response (wMPDR) CBF. Then, we derive two algorithms for optimizing a wMPDR CBF based on two different ways of factorizing a CBF into WPE filters and beamformers: one based on extension of the conventional joint optimization approach proposed for DR+SS and the other based on a novel technique. Experiments using noisy reverberant sound mixtures show that the proposed optimization approaches greatly improve the performance of the speech enhancement in comparison with the conventional cascade configuration in terms of the signal distortion measures and ASR performance. It is also shown that the proposed approaches can greatly reduce the computing cost with improved estimation accuracy in comparison with the conventional joint optimization approach.**

***Index Terms*—Beamforming, dereverberation, source separation, microphone array, automatic speech recognition, maximum likelihood estimation**

I. Introduction

When a speech signal is captured by distant microphones, e.g., in a conference room, it often contains reverberation, diffuse noise, and extraneous speakers' voices. These components are detrimental to the intelligibility of the captured speech and often cause serious degradation in many applications such as hands-free teleconferencing and Automatic Speech Recognition (ASR).

Microphone array speech enhancement has been extensively studied to minimize the aforementioned detrimental effects in the acquired signal. For performing denoising (DN), beamforming techniques have been investigated for decades [1], [2], [3], [4], and the Minimum Variance Distortionless Response (MVDR) beamformer and the Minimum Power Distortionless Response (MPDR) beamformer, are now widely used as the state-of-the-art techniques. For source separation (SS), a number of blind signal processing techniques have been developed, including independent component analysis [5], independent vector analysis [6], and spatial clustering-based beamforming [7]. For dereverberation (DR), a Weighted Prediction Error minimization (WPE) based linear prediction technique [8], [9] and its variants [10] have been actively studied as an effective approach. With these techniques, for determining the coefficients of the filtering, it is crucial to accurately estimate statistics of the speech signals and the noise, such as their spatial covariances and time-varying variances. However, the estimation often becomes inaccurate when the signals mixed under reverberant and noisy conditions, which seriously degrades the performance of these techniques.

To enhance the robustness of the above techniques, recently, neural network-supported microphone array speech enhancement has been actively studied, and showed its effectiveness for denoising [11], dereverberation [12], and source separation [13], [14]. With this approach, estimation of the statistics of the signals and noise, such as Time-Frequency (TF) masks and time-varying variances, is conducted by neural networks [13], [15], [16], [17], while speech enhancement is performed by the microphone array signal processing. This combination is particularly effective because neural networks can very well capture spectral patterns of signals over wide TF ranges, and can reliably estimate such statistics of the signals, which was not well handled by the conventional signal processing. On the other hand, neural networks often introduce nonlinear distortions into the processed signal, which are harmful to perceived speech quality and ASR, while it can be well avoided by microphone array techniques. A number of articles have reported the usefulness of this combination, particularly for far-field ASR, e.g., at the REVERB challenge [18] and the CHiME-3/4/5 challenges [19], [20].

Despite the success of the neural network-supported microphone array speech enhancement, it is still not yet well investigated how to optimally combine individual microphone array techniques for performing denoising, dereverberation, and source separation (DN+DR+SS) at the same time in a computationally efficient way. For example, for denoising and dereverberation (DN+DR), cascade configuration of a WPE

T. Nakatani, K. Kinoshita, R. Ikeshita, and M. Delcroix are with NTT Corporation. C. Boeddeker and R. Haeb-Umbach are with Paderborn Univ.

Manuscript received January 1, 2020; revised XXXX XX, 2020.

filter followed by a MVDR/MPDR beamformer has been widely used as the state-of-the-art frontend, e.g., at the far-field ASR challenges [18], [19], [20], [21]. However, the WPE filter and the beamformer are separately optimized, and the overall optimality of this approach is not guaranteed. In order to perform DN+DR in an optimal way, several techniques have been proposed using a Kalman filter [22], [23], [24]. A technique, called Integrated Sidelobe Cancellation and Linear Prediction (ISCLP) [24], optimizes an integrated filter that can cancel noise and reverberation from the observed signal using a sidelobe cancellation framework. With this technique, however, a steering vector of the target signal needs to be estimated in advance directly from noisy reverberant speech, which is a challenging problem, and thus limits the overall estimation accuracy. In the blind signal processing area, on the other hand, a technique to jointly optimize a pair of a WPE filter followed by a beamformer has been proposed for dereverberation and source separation (DR+SS) under noiseless conditions [25], [26], [27]. One advantage of this approach is that we can access multichannel dereverberated signals obtained as the output of the WPE filter during the optimization, and utilize them to reliably estimate the beamformer. However, this approach requires 1) huge computing cost for the optimization, and 2) has not been extended for application to DN+DR+SS.

To overcome the above limitations, this paper develops algorithms for optimizing a Convolutional BeamFormer (CBF) that can perform DN+DR+SS in a computationally much more efficient way. A CBF is a filter that is applied to a multichannel observed signal to yield the desired output signals. For the optimization of a CBF, this paper first presents a common objective function based on the Maximum Likelihood (ML) criterion, on an assumption that the steering vectors of the desired signals are given, or can be estimated. This paper refers to a CBF optimized by this objective function as a weighted MPDR (wMPDR) CBF. Then, showing that a CBF can be factorized into WPE filter(s) and beamformer(s) in two different ways, we derive two different algorithms for optimizing the wMPDR CBF, based on the ways of the CBF factorization. The first approach, referred to as a source-packed factorization, is an extension of the conventional joint optimization technique proposed for DR+SS [25], [26], [27]. In this paper, we first show that the direct application of this approach to DN+DR+SS has serious problems in terms of the computational efficiency and the estimation accuracy, and then present its extension for solving the problems. The second approach, referred to as a source-wise factorization, is based on a novel factorization technique. It factorizes a CBF into a set of sub-filter pairs, each of which is composed of a WPE filter and a beamformer, and aimed at estimating each source independently. For both approaches, we also present a method for robustly estimating the steering vectors of the desired signals during the optimization of the wMPDR CBF using the output of the WPE filters. A neural network supported TF-mask estimation technique is also incorporated[1] for the estimation of the steering vectors. While both approaches work comparably well in terms of the estimation accuracy, the source-wise factorization has advantages in terms of computational efficiency. An additional benefit of the source-wise factorization is that it can be used, without loss of optimality, for extraction of a single target source from a sound mixture, which is now an important application area of speech enhancement [13], [29].

Experiments based on noisy reverberant sound mixtures created using the REVERB Challenge dataset [18] show that the proposed optimization approaches substantially improve the performance of DN+DR+SS in comparison to the conventional cascade configuration in terms of ASR performance and reduction of signal distortion. It is also shown that the two proposed approaches can greatly reduce the computing cost with improved estimation accuracy in comparison with the conventional joint optimization approach.

Certain parts of this paper have already been presented in our recent conference papers. In [30], the ML formulation for optimizing a CBF was derived for DN+DR. In [31], it was shown that a CBF for DN+DR can be factorized into a WPE filter and a wMPDR (non-convolutional) beamformer, and jointly optimized without loss of optimality. Furthermore, [32] presented ways to reliably estimate TF masks for DN+DR+SS. This paper integrates these techniques to perform DN+DR+SS in a computationally efficient way.

In the remainder of this paper, the model of the observed signal and that of the CBF are defined in Section II. Then, Section III presents the proposed optimization methods. Section IV summarizes characteristics and advantages of the proposed methods. In Sections V and VI, experimental results and concluding remarks are described, respectively.

II. Models of signal and beamformer

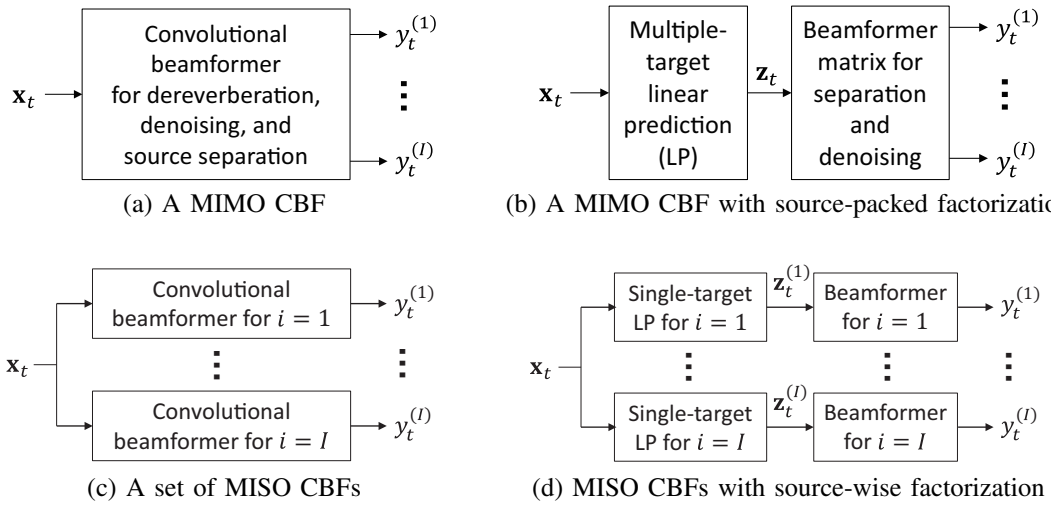
This paper assumes that I source signals are captured by $M(\quad I)$ microphones in a noisy reverberant environment. The captured signal at each TF point in the short-time Fourier transform (STFT) domain is modeled by

$$\mathbf{x}_{t;f} = \sum_{i=1}^{I} \mathbf{x}_{t;f}^{(i)} + \mathbf{n}_{t;f}; \tag{1}$$

$$\mathbf{x}_{t;f}^{(i)} = \mathbf{d}_{t;f}^{(i)} + \mathbf{r}_{t;f}^{(i)}; \tag{2}$$

where t and f are time and frequency indices, respectively, $\mathbf{x}_{t;f} = [x_{1;t;f}; \cdots; x_{M;t;f}]^{>} \ 2 \ \mathbb{C}^{M \ 1}$ is a column vector containing all microphone signals at a TF point. Here, $(\,)^{>}$ denotes the non-conjugate transpose. $\mathbf{x}_{t;f}^{(i)} = [x_{1;t;f}^{(i)}; \cdots; x_{M;t;f}^{(i)}]^{>}$ is a (noiseless) reverberant signal corresponding to the ith source, and $\mathbf{n}_{t;f} = [n_{1;t;f}; \cdots; n_{M;t;f}]^{>}$ is the additive diffuse noise. $\mathbf{x}_{t;f}^{(i)}$ for each source in Eq. (1) is further decomposed into two parts in Eq. (2), one consisting of the direct signal and early reflections, referred to as a desired signal $\mathbf{d}_{t;f}^{(i)}$, and the other corresponding to the late reverberation $\mathbf{r}_{t;f}^{(i)}$. Hereafter, the frequency indices of the symbols are omitted for brevity, on an assumption that each frequency bin is processed independently in the same way.

[1]It is worth noting that the proposed techniques can also be applied to the conventional blind signal processing for DR+SS as discussed in [28].



(a) A MIMO CBF (b) A MIMO CBF with source-packed factorization

(c) A set of MISO CBFs (d) MISO CBFs with source-wise factorization

Fig. 1. A Multi-Input Multi-Output (MIMO) CBF and its three different implementations. They are equivalent to each other in the sense that whatever values are set to coefficients of one implementation, certain coefficients of the other implementations can be determined such that they realize the same input-output relationship. Thus, the optimal solutions of all the implementations are identical as long as they are optimized based on the same objective function.

In this paper, the goal of DN+DR+SS is to estimate $\mathbf{d}_t^{(i)}$ for each source i from $\mathbf{x}_t$ in Eq. (1) by reducing $\mathbf{r}_t^{(i)}$ of the source i, $\mathbf{x}_t^{(i')}$ of all the other sources $i' \neq i$, and the diffuse noise $\mathbf{n}_t$. It is known that in noisy reverberant environments the early reflections enhance the intelligibility of speech for human perception [33] and improve the ASR performance by computer [34], and thus we include them in the desired signal. Hereafter, this paper uses $m = 1$ as the reference microphone, and describes a method for estimating the desired signal, $d_{1;t}^{(i)}$, at the microphone without loss of generality.

To achieve the above goal, we further model $\mathbf{d}_t^{(i)}$ as

$$\mathbf{d}_t^{(i)} = \mathbf{v}^{(i)} s_t^{(i)} = \boldsymbol{\nu}^{(i)} d_{1;t}^{(i)}; \tag{3}$$

where $s_t^{(i)}$ is the ith clean speech at a TF point. In Eq. (3), the desired signal of the ith source, $\mathbf{d}_t^{(i)}$, is modeled by $\mathbf{v}^{(i)} s_t^{(i)}$, i.e., a product in the STFT domain of the clean speech with a transfer function $\mathbf{v}^{(i)}$, hereafter referred to as a steering vector, assuming that the duration of the impulse response corresponding to the direct signal and early reflections in the time domain is sufficiently short in comparison with the analysis window [35]. Then, the desired signal is further rewritten as $\boldsymbol{\nu}^{(i)} d_{1;t}^{(i)}$, i.e., a product of the desired signal at the reference microphone $d_{1;t}^{(i)} = v_1^{(i)} s_t^{(i)}$ with a Relative Transfer Function (RTF) [36] which is defined as the steering vector divided by its reference microphone element, namely

$$\boldsymbol{\nu}^{(i)} = \mathbf{v}^{(i)} / v_1^{(i)}. \tag{4}$$

In contrast, assuming that the duration of the late reverberation in the time domain is larger than the analysis window, the late reverberation $\mathbf{r}_t^{(i)}$ is modeled by a convolution in the STFT domain [37] of the clean speech with a time series of acoustic transfer functions corresponding to the late reverberation as

$$\mathbf{r}_t^{(i)} = \sum_{\tau=b}^{L_a-1} \mathbf{a}^{(i)}(\tau) s_{t-\tau}^{(i)}; \tag{5}$$

where $\mathbf{a}^{(i)} = [a_{1;}^{(i)}(\tau); \ldots; a_{M;}^{(i)}(\tau)]^\top$ for $\tau = b, \ldots, L_a - 1$ are the convolutional acoustic transfer functions, and b is the mixing time, representing the relative frame delay of the late reverberation start time to the direct signal.

In this paper, we further assume that $\mathbf{d}_t^{(i)}$ is statistically independent[2] of the following variables:

- $s_{t'}^{(i)}$ for $t' < t - b$ (and thus $\mathbf{d}_t^{(i)}$ is statistically independent of $\mathbf{x}_{t'}^{(i)}$ for $t' \leq t - b$),
- $\mathbf{r}_{t''}^{(i)}$ for $t'' < t$,
- $\mathbf{x}_{t'}^{(i')}$ and $\mathbf{n}_{t'}$ for all t, t' and $i' \neq i$.

These assumptions are used to derive the optimization algorithms described in the following.

A. Definition of a CBF and its three different implementations

We now define a CBF, which will be later factorized into WPE filter(s) and beamformer(s), as

$$\mathbf{y}_t = \mathbf{W}_0^{\mathsf{H}} \mathbf{x}_t + \sum_{\tau=b}^{L-1} \mathbf{W}^{\mathsf{H}}(\tau) \mathbf{x}_{t-\tau}; \tag{6}$$

where $\mathbf{y}_t = [y_t^{(1)}; \ldots; y_t^{(I)}]^\top \in \mathbb{C}^{I \times 1}$ is the output of the CBF corresponding to estimates of I desired signals, $\mathbf{W}(\tau) \in \mathbb{C}^{M \times I}$ for each $\tau \in \{0, \ldots, b, \ldots, L-1\}$ is a matrix composed of the beamformer coefficients, $(\cdot)^{\mathsf{H}}$ denotes conjugate transpose, and b is the prediction delay of the CBF. In this paper, we set b equal to the mixing time introduced in Eq. (5), so that the desired signals are included only in the first term of Eq. (6), and that they are statistically independent of the second term according to the assumptions introduced in the signal model. Then, this paper performs DN+DR+SS by estimating beamformer coefficients that can estimate the desired signals included in the first term of Eq. (6).

[2] See [8] for more precise discussion on the statistical independence between $\mathbf{d}_t^{(i)}$ and $s_{t'}$ for $t' < t - b$.

For the sake of notational simplicity, we also introduce a matrix representation of a CBF as

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{w}_0 \\ \overline{\mathbf{W}} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{x}_t \\ \overline{\mathbf{x}}_t \end{bmatrix}; \tag{7}$$

where $\overline{\mathbf{W}}$ is a matrix containing $\mathbf{W}$ for $L-1$ and $\overline{\mathbf{x}}_t$ is a column vector containing past multichannel observed signals $\mathbf{x}_t$ for $L-1$, respectively, defined as

$$\overline{\mathbf{W}} = [\mathbf{W}^{\mathsf{H}}, \cdots, \mathbf{W}_{L-1}^{\mathsf{H}}]^{\mathsf{H}} \in \mathbb{C}^{M(L-1)\times I}; \tag{8}$$

$$\overline{\mathbf{x}}_t = [\mathbf{x}_t^{\mathsf{H}}, \cdots, \mathbf{x}_{t-L+1}^{\mathsf{H}}]^{\mathsf{H}} \in \mathbb{C}^{M(L-1)\times 1}. \tag{9}$$

Hereafter, we refer to the CBF defined by Eqs. (6) and (7) as a MIMO CBF.

In the following, we further present three different implementations of the CBF, including the two ways of factorizing the CBF. Figure 1 illustrates the MIMO CBF and its three different implementations.

1) Source-packed factorization: With the implementation shown in Fig. 1 (b), we directly factorize[3] the MIMO CBF in Eq. (7) as

$$\begin{bmatrix} \mathbf{w}_0 \\ \overline{\mathbf{W}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_M \\ \overline{\mathbf{G}} \end{bmatrix} \mathbf{Q}; \tag{10}$$

where $\mathbf{Q} \in \mathbb{C}^{M\times I}$, $\overline{\mathbf{G}} \in \mathbb{C}^{M(L-1)\times M}$, and $\mathbf{I}_M \in \mathbb{R}^{M\times M}$ is an identity matrix. Then, Eq. (6) can be rewritten as a pair of a (convolutional) linear prediction filter followed by a (non-convolutional) beamformer matrix, respectively, defined as

$$\mathbf{z}_t = \mathbf{x}_t - \overline{\mathbf{G}}^{\mathsf{H}}\overline{\mathbf{x}}_t; \tag{11}$$

$$\mathbf{y}_t = \mathbf{Q}^{\mathsf{H}}\mathbf{z}_t. \tag{12}$$

Here, $\mathbf{z}_t \in \mathbb{C}^{M\times 1}$ and $\overline{\mathbf{G}}$ are the output and the prediction matrix of the linear prediction, and $\mathbf{Q}$ is the coefficient matrix of the beamformer. Eq. (11) is supposed to dereverberate all the sources at the same time, and thus referred to as a multiple-target linear prediction, while Eq. (12) is supposed to perform denoising and source separation at the same time. Because the individual sources are not distinguished in the output of the WPE filter, this implementation is called source-packed factorization.

One example of the source-packed factorization is the cascade configuration composed of a WPE filter followed by a beamformer, which has been widely used for DN+DR+SS in the far-field speech recognition area [14], [20], [38], and the other example is one used in the joint optimization of a WPE filter and a beamformer, which has been investigated for DR+SS in the blind signal processing area [25], [26], [27].

2) Multi-Input Single-Output (MISO) CBF: Next, we define a set of MISO CBFs shown in Fig. 1 (c). They are obtained by decomposing the beamformer coefficients in Eq. (7) as

$$\begin{bmatrix} \mathbf{w}_0 \\ \overline{\mathbf{W}} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_0^{(1)} & \mathbf{w}_0^{(2)} & \cdots & \mathbf{w}_0^{(I)} \\ \overline{\mathbf{w}}^{(1)} & \overline{\mathbf{w}}^{(2)} & \cdots & \overline{\mathbf{w}}^{(I)} \end{bmatrix}; \tag{13}$$

where $\mathbf{w}_0^{(i)} \in \mathbb{C}^{M\times 1}$ and $\overline{\mathbf{w}}^{(i)} \in \mathbb{C}^{M(L-1)\times 1}$ are column vectors, respectively, containing the ith columns of $\mathbf{W}_0$ and $\overline{\mathbf{W}}$, and are used to extract the ith desired signal. Then, Eq. (7) can be rewritten for each source i as

$$y_t^{(i)} = \begin{bmatrix} \mathbf{w}_0^{(i)} \\ \overline{\mathbf{w}}^{(i)} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{x}_t \\ \overline{\mathbf{x}}_t \end{bmatrix}. \tag{14}$$

For example, MISO CBFs were used in [30], [39]. ISCLP proposed in [24] can also be viewed as a realization of a MISO CBF using a sidelobe cancellation framework [40].

3) Source-wise factorization: With the source-wise factorization shown in Fig. 1 (d), we further factorize each MISO CBF defined in Eq. (14) for a source i as

$$\begin{bmatrix} \mathbf{w}_0^{(i)} \\ \overline{\mathbf{w}}^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_M \\ \overline{\mathbf{G}}^{(i)} \end{bmatrix} \mathbf{q}^{(i)}; \tag{15}$$

where $\mathbf{q}^{(i)} \in \mathbb{C}^{M\times 1}$ and $\overline{\mathbf{G}}^{(i)} \in \mathbb{C}^{M(L-1)\times M}$. Then, Eq. (14) can be rewritten as a pair of a linear prediction filter followed by a beamformer, respectively, defined as

$$\mathbf{z}_t^{(i)} = \mathbf{x}_t - \overline{\mathbf{G}}^{(i)\mathsf{H}}\overline{\mathbf{x}}_t; \tag{16}$$

$$y_t^{(i)} = \mathbf{q}^{(i)\mathsf{H}}\mathbf{z}_t^{(i)}; \tag{17}$$

where $\mathbf{z}_t^{(i)} \in \mathbb{C}^{M\times 1}$ and $\overline{\mathbf{G}}^{(i)}$ are the output and the prediction matrix of the linear prediction, and $\mathbf{q}^{(i)}$ is the coefficient vector of the beamformer. Because Eq. (16) is performed only for estimation of the ith source, it is referred to as a single-target linear prediction.

4) Relationship between two factorization approaches: The difference between the two factorization approaches, namely Fig. 1 (b) and (d), is based only on the way to perform linear prediction, Eq. (11) or Eq. (16), and more specifically based on whether the prediction matrices, $\overline{\mathbf{G}}$ and $\overline{\mathbf{G}}^{(i)}$, are common to all the sources or different over different sources. According to this, different optimization algorithms with different characteristics are derived, as will be shown in Section III. In contrast, the beamformer parts, $\mathbf{Q}$ and $\mathbf{q}^{(i)}$ in Eqs. (12) and (17) are identical in the two approaches, viewing $\mathbf{q}^{(i)}$ as the ith column of $\mathbf{Q}$, because they satisfy $\mathbf{W}_0 = \mathbf{Q}$ in Eq. (10) and $\mathbf{w}_0^{(i)} = \mathbf{q}^{(i)}$ in Eq. (15).

In addition, it should be noted that all the above implementations of a CBF are equivalent to each other in the sense that whatever values are set to coefficients of one implementation, certain coefficients of the other implementations can be determined such that they realize the same input-output relationship. Thus, the optimal solutions of all the implementations are identical as long as they are based on the same objective function.

III. ML ESTIMATION OF CBF

In this section, two different optimization algorithms are derived, respectively, using (b) source-packed factorization and (d) source-wise factorization. For the derivation, we assume that the RTFs $\boldsymbol{\nu}^{(i)}$ and the time-varying variances of the output signals yielded by the optimal CBF, denoted by $\lambda_t^{(i)}$, are given. Later in Section III-E, we describe a way for estimating $\lambda_t^{(i)}$ jointly with the CBF coefficients based on the ML criterion,

[3]The existence of $\overline{\mathbf{G}}$ that satisfies $\overline{\mathbf{W}} = \overline{\mathbf{G}}\mathbf{Q}$ is guaranteed for any $\overline{\mathbf{W}}$ when $M \geq I$ and $\mathrm{rank}\{\mathbf{Q}\} = I$.

and a way for estimating $\boldsymbol{\nu}^{(i)}$ based on the output of the WPE filter obtained at a step of the optimization.

A. Probabilistic model

First, we formulate the objective function for DN+DR+SS by reinterpreting the objective function proposed for DN+DR in [30]. For this formulation, we interpret DN+DR+SS to be composed of a set of separate processing steps, each of which applies DN+DR to enhance a source i by reducing the late reverberation of the source (DR) and by reducing the additive noise including the other sources and the diffuse noise (DN). With this interpretation, we introduce the following assumptions similar to [30].

> The output of the optimal CBF for each i, namely $y_t^{(i)}$, follows zero-mean complex Gaussian distribution with a time-varying variance $\lambda_t^{(i)} = \mathrm{E}\left[|y_t^{(i)}|^2 \right]$ [8].
>
> The beamformer satisfies a distortionless constraint for each source i defined using the RTF $\boldsymbol{\nu}^{(i)}$ in Eq. (4) as

$$\mathbf{w}_0^{(i)\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1 \quad \text{or} \quad \mathbf{q}^{(i)\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1 \ : \tag{18}$$

Then, according to the discussion in [30], we can approximately derive the objective function to minimize for estimating the CBF coefficients for source i, e.g., $\boldsymbol{\theta}^{(i)} = \{\mathbf{w}_0^{(i)}, \overline{\mathbf{w}}^{(i)}\}$, based on the ML estimation as

$$L_i(\boldsymbol{\theta}^{(i)}) = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{|y_t^{(i)}|^2}{\lambda_t^{(i)}} + \log \lambda_t^{(i)} \right) \text{ s.t. } \mathbf{w}_0^{(i)\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1. \tag{19}$$

The objective function for estimating all the sources can then be obtained by summing Eq. (19) over all the sources as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{I} L_i(\boldsymbol{\theta}^{(i)}), \text{ s.t. } \mathbf{w}_0^{(i)\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1 \text{ for all } i, \tag{20}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(I)}\}$. This objective function is used commonly for all the implementations of a CBF. In this paper, we call a CBF optimized by the above objective function as weighted MPDR (wMPDR) CBF because it minimizes the average power of the output $y_t^{(i)}$ weighted by the time-varying variance, $\lambda_t^{(i)}$, of the signal.

Here, let us briefly explain how DN+DR+SS is performed by Eqs. (19) and (20). Substituting Eqs. (1) and (2) in Eq. (14) and using the model of the desired signal in Eq. (3) and the distortionless constraint in Eq. (18), we obtain

$$y_t^{(i)} = d_{1;t}^{(i)} + \hat{r}_t^{(i)} + \sum_{i' \neq i} \hat{x}_t^{(i')} + \hat{n}_t, \tag{21}$$

where $\hat{r}_t^{(i)}$, $\hat{x}_t^{(i')}$ for $i' \neq i$, and $\hat{n}_t$ are late reverberation of the ith source, all the other sources, and the additive diffuse noise remaining in the output of the CBF, respectively, written using the MISO CBF form as

$$\hat{r}_t^{(i)} = \begin{bmatrix} \mathbf{w}_0^{(i)} \\ \overline{\mathbf{w}}^{(i)} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{r}_t^{(i)} \\ \overline{\mathbf{x}}_t^{(i)} \end{bmatrix}; \tag{22}$$

$$\hat{x}_t^{(i')} = \begin{bmatrix} \mathbf{w}_0^{(i)} \\ \overline{\mathbf{w}}^{(i)} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{x}_t^{(i')} \\ \overline{\mathbf{x}}_t^{(i')} \end{bmatrix}; \tag{23}$$

$$\hat{n}_t = \begin{bmatrix} \mathbf{w}_0^{(i)} \\ \overline{\mathbf{w}}^{(i)} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \mathbf{n}_t \\ \overline{\mathbf{n}}_t \end{bmatrix}; \tag{24}$$

where $\overline{\mathbf{n}}_t = [\mathbf{n}_t^{\mathsf{T}}, \ldots, \mathbf{n}_{t-L+1}^{\mathsf{T}}]^{\mathsf{T}}$. According to the statistical independence assumptions introduced in Section II, $d_{1;t}^{(i)}$ is statistically independent of $\hat{r}_t^{(i)}$, $\hat{x}_t^{(i')}$, and $\hat{n}_t$. Then, substituting Eq. (21) in Eq. (19) and omitting constant terms, we obtain in the expectation sense

$$\mathrm{E}\left\{ L_i(\boldsymbol{\theta}^{(i)}) \right\} = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathrm{E}\left[|\hat{r}_t^{(i)} + \sum_{i' \neq i} \hat{x}_t^{(i')} + \hat{n}_t|^2 \right]}{\lambda_t^{(i)}}. \tag{25}$$

The above equation indicates that minimization of the objective function indeed minimizes the sum of $\hat{r}_t^{(i)}$, $\hat{x}_t^{(i')}$ for $i' \neq i$, and $\hat{n}_t$ in Eq. (21).

Before deriving the optimization algorithms, we define a matrix that is frequently used in the derivation, referred to as a variance-normalized spatio-temporal covariance matrix. Letting $\underline{\mathbf{x}}_t$ be a column vector composed of the current and past observed signals at all the microphones, defined as

$$\underline{\mathbf{x}}_t = [\mathbf{x}_t^{\mathsf{T}}, \overline{\mathbf{x}}_t^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M(L+1) \times 1}; \tag{26}$$

the matrix is defined as

$$\underline{\mathbf{R}}_{\mathbf{x}}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \frac{\underline{\mathbf{x}}_t \underline{\mathbf{x}}_t^{\mathsf{H}}}{\lambda_t^{(i)}} \in \mathbb{C}^{M(L+1) \times M(L+1)}. \tag{27}$$

A factorized form of the matrix is also defined as

$$\underline{\mathbf{R}}_{\mathbf{x}}^{(i)} = \begin{bmatrix} \mathbf{R}_{\mathbf{x}}^{(i)} & \mathbf{P}_{\mathbf{x}}^{(i)\mathsf{H}} \\ \mathbf{P}_{\mathbf{x}}^{(i)} & \overline{\mathbf{R}}_{\mathbf{x}}^{(i)} \end{bmatrix}; \tag{28}$$

where

$$\mathbf{R}_{\mathbf{x}}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbf{x}_t \mathbf{x}_t^{\mathsf{H}}}{\lambda_t^{(i)}} \in \mathbb{C}^{M \times M}; \tag{29}$$

$$\mathbf{P}_{\mathbf{x}}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \mathbf{x}_t^{\mathsf{H}}}{\lambda_t^{(i)}} \in \mathbb{C}^{M(L) \times M}; \tag{30}$$

$$\overline{\mathbf{R}}_{\mathbf{x}}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \overline{\mathbf{x}}_t^{\mathsf{H}}}{\lambda_t^{(i)}} \in \mathbb{C}^{M(L) \times M(L)}. \tag{31}$$

B. Optimization based on source-packed factorization

This subsection discusses methods for optimizing a CBF with the source-packed factorization. In the following, after describing a method for directly applying the conventional joint optimization technique used for DR+SS to DN+DR+SS, we summarize the problems in it, and present the solutions to the problems.

1) Direct application of conventional technique: With the source-packed factorization in Eqs. (11) and (12), it is difficult to estimate both $\mathbf{Q}$ and $\overline{\mathbf{G}}$ at the same time in closed form even when $\overset{(i)}{t}$ and $\boldsymbol{\nu}^{(i)}$ are all given. Instead, we use an iterative and alternate estimation scheme, following the idea from the blind signal processing technique [25], [26], [27], where at each estimation step, one of $\mathbf{Q}$ and $\overline{\mathbf{G}}$ is updated while fixing the other.

For updating $\overline{\mathbf{G}}$, we fix $\mathbf{Q}$ at its previously estimated value. For the derivation of the algorithm, the representation of linear prediction in Eq. (11) is slightly modified as

$$\mathbf{z}_t = \mathbf{x}_t \quad \overline{\mathbf{X}}_t \overline{\mathbf{g}}; \tag{32}$$

where $\overline{\mathbf{X}}_t$ and $\overline{\mathbf{g}}$ are equivalent to $\overline{\mathbf{x}}_t$ and $\overline{\mathbf{G}}$ with modified matrix structure defined as

$$\overline{\mathbf{X}}_t = \mathbf{I}_M \quad \overline{\mathbf{x}}_t^{>} \; 2 \; \mathbb{C}^{M \quad M^2(L \quad)}; \tag{33}$$

$$\overline{\mathbf{g}} = \overline{\mathbf{g}}_1^{>}; : : : ; \overline{\mathbf{g}}_M^{>} \; {}^{\mathsf{H}} \; 2 \; \mathbb{C}^{M^2(L \quad) \quad 1}; \tag{34}$$

where $\quad$ is Kronecker product, and $\overline{\mathbf{g}}_m$ is the mth column of $\overline{\mathbf{G}}$. Then, considering that the CBF in Eqs. (11) and (12) can be written as $y_t^{(i)} = \mathbf{q}^{(i)\;\mathsf{H}} \mathbf{x}_t \quad \overline{\mathbf{X}}_t \overline{\mathbf{g}}$ and omitting normalization terms, the objective function in Eq. (20) becomes

$$L_{\overline{\mathbf{g}}}(\overline{\mathbf{g}}) = \frac{1}{T} \sum_{t=1}^{\overline{K}} \mathbf{x}_t \quad \overline{\mathbf{X}}_t \overline{\mathbf{g}} \; {}^{2}_{\mathbf{q};t}; \tag{35}$$

where $k\mathbf{x}k_{\mathbf{R}}^2 = \mathbf{x}^{\mathsf{H}}\mathbf{R}\mathbf{x}$, and $\quad_{\mathbf{q};t}$ is a semi-definite Hermitian matrix defined as

$$\quad_{\mathbf{q};t} = \sum_{i=1}^{K} \frac{\mathbf{q}^{(i)} \; \mathbf{q}^{(i)\;\mathsf{H}}}{\overset{(i)}{t}} \; 2 \; \mathbb{C}^{M \quad M}. \tag{36}$$

Because Eq. (35) is a quadratic form with a lower bound, $\overline{\mathbf{g}}$ that minimizes it can be obtained as

$$\overline{\mathbf{g}} = \quad^{+}; \tag{37}$$

$$= \frac{1}{T} \sum_{t} \overline{\mathbf{X}}_t^{\mathsf{H}} \quad_{\mathbf{q};t} \overline{\mathbf{X}}_t \; 2 \; \mathbb{C}^{M^2(L \quad) \quad M^2(L \quad)}; \tag{38}$$

$$= \frac{1}{T} \sum_{t} \overline{\mathbf{X}}_t^{\mathsf{H}} \quad_{\mathbf{q};t} \mathbf{x}_t \; 2 \; \mathbb{C}^{M^2(L \quad) \quad 1}; \tag{39}$$

where $(\;)^{+}$ is the Moore-Penrose pseudo-inverse. Because the rank of $\quad$ is equal to or smaller than $MI(L \quad)$ as will be shown in Section III-B2, $\quad$ is rank deficient for over-determined cases, namely when $M > I$, and thus the use of the pseudo-inverse is indispensable. Eqs. (37) to (39) are equivalent to those used in the dereverberation step for DR+SS [25], [26], [27] except that in this paper denoising is additionally included in the objective and that over-determined cases are also considered. This filter is referred to as the multiple-target WPE filter in this paper.

For the update of $\mathbf{Q}$, fixing $\overline{\mathbf{g}}$ at its previously estimated value, the objective in Eq. (20) can be rewritten as

$$L_{\mathbf{Q}}(\mathbf{Q}) = \sum_{i=1}^{K} \mathbf{q}^{(i)} \; {}^{2}_{\mathbf{R}_{\mathbf{z}}^{(i)}} \quad \text{s.t.} \quad \mathbf{q}^{(i)\;\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1; \tag{40}$$

where $\mathbf{R}_{\mathbf{z}}^{(i)}$ is a variance-normalized spatial covariance matrix of the output of the multiple-target WPE filter, calculated as

$$\mathbf{R}_{\mathbf{z}}^{(i)} = \frac{1}{T} \sum_{t=1}^{\overline{K}} \frac{\mathbf{z}_t \mathbf{z}_t^{\mathsf{H}}}{\overset{(i)}{t}}. \tag{41}$$

Then, $\mathbf{q}^{(i)}$ that minimizes Eq. (40) under the distortionless constraint $\mathbf{q}^{(i)\;\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1$ can be obtained as

$$\mathbf{q}^{(i)} = \frac{\mathbf{R}_{\mathbf{z}}^{(i)} \quad^{1} \boldsymbol{\nu}^{(i)}}{\boldsymbol{\nu}^{(i)\;\mathsf{H}} \; \mathbf{R}_{\mathbf{z}}^{(i)} \quad^{1} \boldsymbol{\nu}^{(i)}}. \tag{42}$$

Because the above beamformer minimizes the average power of $\mathbf{z}_t$ weighted by the time-varying variance, we call this as weighted MPDR (wMPDR) beamformer[4]. As will be shown in Section III-C, a wMPDR beamformer is a special case of a wMPDR CBF. A wMPDR CBF is reduced to a wMPDR beamformer when setting the length of the CBF $L = 1$, i.e., by just making it a non-convolutional beamformer.

The above algorithm, however, has two serious problems. Firstly, the size of the covariance matrix in Eq. (38) is very large, requiring huge computing cost for calculating it and its inverse. Secondly, as will be shown in the experiments, the iterative and alternate estimation of $\mathbf{Q}$ and $\overline{\mathbf{G}}$ tends to converge to a sub-optimal point. This is probably because the update of $\overline{\mathbf{G}}$ is performed based only on the output of the fixed beamformer in the iterative and alternate estimation, as in Eq. (19), while the signal dimension of the beamformer output, i.e., I, is reduced from that of the original signal space, i.e., M, with the over-determined case, i.e., $I < M$. As a consequence, signal components that are relevant for the update of $\overline{\mathbf{G}}$ may be reduced in the beamformer output, especially when the estimation of $\mathbf{Q}$ is not accurate at the early stage of the optimization. This can seriously degrade the update of $\overline{\mathbf{G}}$.

2) Proposed extension: Here, we present two techniques to mitigate the above problems within the source-packed factorization approach. The first one is used to reduce the computing cost. As shown in Appendix A, Eqs. (38) and (39) can be rewritten, using Eq. (28), as

$$= \sum_{i=1}^{K} \mathbf{q}^{(i)} \; \mathbf{q}^{(i)\;\mathsf{H}} \quad \overline{\mathbf{R}}_{\mathbf{x}}^{(i)\;>}; \tag{43}$$

$$= \sum_{i=1}^{K} \mathbf{q}^{(i)} \quad \mathbf{P}_{\mathbf{x}}^{(i)} \mathbf{q}^{(i)}; \tag{44}$$

where $(\;)$ denotes complex conjugate. In the above equations, the majority of the calculation is coming from that of $\overline{\mathbf{R}}_{\mathbf{x}}^{(i)}$. Because the size of the matrix is much smaller than that of , we can greatly reduce the computing cost with this modification[5] in comparison with direct calculation of Eqs. (38)

[4]A wMPDR beamformer is also called a Maximum-Likelihood Distortionless Response (MLDR) beamformer in [41].

[5]In general, computational complexity of a matrix multiplication is more than $O(n^2)$. Because the size of is M-times larger than $\overline{\mathbf{R}}_{\mathbf{x}}$, it is suggested that the computational complexity for calculating is at least M^2 times larger than that for calculating $\overline{\mathbf{R}}_{\mathbf{x}}$.

and (39). Although we still need to calculate the inverse of the huge matrix even with this modification, the cost is relatively small in comparison with the direct calculation of . (Note that Eq. (43) also shows the rank of to be equal to or smaller than $MI(L$).)

The second technique introduces a heuristic to improve the update of the WPE filter. In order to use a whole M-dimensional signal space to be considered for the update, we modify the CBF to output not only I desired signals, but also $M - I$ auxiliary signals included in the orthogonal complement $\mathbf{Q}^?$ of $\mathbf{Q}$, and model the auxiliary signals as zero-mean time-varying complex Gaussians. With this modification, the optimization is performed by calculating the summation in Eqs. (43) and (44) over not only $1 \le i \le I$ but also $I < i \le M$, letting $\mathbf{q}^{(I+1)}, \ldots, \mathbf{q}^{(M)}$ be the orthonormal bases for the orthogonal complement $\mathbf{Q}^?$. Because it is not important to distinguish the variances $\lambda_t^{(i)}$ of the auxiliary signals, we use the same value for them calculated as

$$\lambda_t^? = \frac{1}{M - I} \sum_{i=I+1}^{M} \left| \mathbf{q}^{(i)H} \mathbf{z}_t \right|^2, \tag{45}$$

and calculate $\mathbf{P}_{\mathbf{x}}^?$ and $\overline{\mathbf{R}}_{\mathbf{x}}^?$ based on Eqs. (30) and (31) accordingly. In summary, we can implement this modification by adding the following terms, respectively, to and in Eqs. (43) and (44).

$$^? = \sum_{i=I+1}^{M} \mathbf{q}^{(i)} \left(\mathbf{q}^{(i)H} \overline{\mathbf{R}}_{\mathbf{x}}^? \right)^{>}, \tag{46}$$

$$^? = \sum_{i=i+1}^{M} \mathbf{q}^{(i)} \mathbf{P}_{\mathbf{x}}^? \mathbf{q}^{(i)} \quad . \tag{47}$$

C. Direct optimization of MISO CBFs

Before deriving the optimization with the source-wise factorization, we show that we can directly optimize the MISO CBFs in Eq. (14), and summarize its characteristics. With this setting, the CBFs and the objective function are both defined separately for each source in Eqs. (14) and (19), and thus, the optimization can be performed separately for each source. The resultant algorithm is, therefore, identical to that proposed for DN+DR in [42], where this type of CBF is also referred to as a Weighted Power minimization Distortionless response (WPD) CBF.

For presenting the solution, we introduce the following vector representation of Eq. (14).

$$y_t^{(i)} = \underline{\mathbf{w}}^{(i)H} \underline{\mathbf{x}}_t, \tag{48}$$

where $\underline{\mathbf{w}}^{(i)}$ is defined as

$$\underline{\mathbf{w}}^{(i)} = \begin{bmatrix} w_0^{(i)} \\ \overline{\mathbf{w}}^{(i)} \end{bmatrix}; \tag{49}$$

Then, when $\lambda_t^{(i)}$ and $\mathbf{v}^{(i)}$ are given, Eq. (19) becomes a simple constraint quadratic form as

$$L_{\mathbf{w}}(\underline{\mathbf{w}}^{(i)}) = \left\| \underline{\mathbf{w}}^{(i)} \right\|^2_{\underline{\mathbf{R}}_{\mathbf{x}}^{(i)}} \text{ s.t. } \underline{\mathbf{w}}^{(i)H} \underline{\mathbf{v}}^{(i)} = 1, \tag{50}$$

where $\underline{\mathbf{R}}_{\mathbf{x}}^{(i)}$ is the covariance matrix defined in Eq. (28), and $\underline{\mathbf{v}}^{(i)} = \left[\mathbf{v}^{(i)\top}, 0, \ldots, 0 \right]^\top \in \mathbb{C}^{M(L+1)}$ corresponds to the RTF $\mathbf{v}^{(i)}$ with zero padding. Finally, the solution is given as

$$\underline{\mathbf{w}}^{(i)} = \frac{\underline{\mathbf{R}}_{\mathbf{x}}^{(i)-1} \underline{\mathbf{v}}^{(i)}}{\underline{\mathbf{v}}^{(i)H} \underline{\mathbf{R}}_{\mathbf{x}}^{(i)-1} \underline{\mathbf{v}}^{(i)}}. \tag{51}$$

The above equation gives the simplest form of the solution to a wMPDR CBF. It clearly shows that a wMPDR CBF is a general case of a wMPDR beamformer. By setting $L = 1$ in the above solution, namely by letting it be a non-convolutional beamformer, it reduces to the solution of a wMPDR beamformer in Eq. (42).

An advantage of the solution using the MISO CBFs is that it can be obtained by a closed form equation provided the RTFs and the time-varying variances of the desired signals are given, and we do not need to consider the interaction between DN and DR. With this approach, however, it is necessary to estimate the RTFs directly from a reverberant observation similar to [24]. A solution to this problem is to use dereverberation preprocessing based on a WPE filter for the RTF estimation. It was shown in [30] that the output of a WPE filter can be obtained in a computationally efficient way within the framework of this approach. However, the source-wise factorization approach described in the following can more naturally solve this problem. So, this paper adopts it as the solution.

D. Optimization based on source-wise factorization

With the source-wise factorization, similar to the case with the direct optimization of the MISO CBFs, the optimization can be performed separately for each source, and the resultant algorithm is identical to that proposed for DN+DR in [31].

Considering that a CBF can be written based on Eqs. (16) and (17) as $y_t^{(i)} = \mathbf{q}^{(i)H} \mathbf{x}_t - \overline{\mathbf{G}}^{(i)H} \overline{\mathbf{x}}_t$ and using the factorized form of $\underline{\mathbf{R}}_{\mathbf{x}}^{(i)}$ in Eq. (28), the objective function in Eq. (19) can be rewritten as

$$L_i\left(\overline{\mathbf{G}}^{(i)}, \mathbf{q}^{(i)}\right) = \left\| \overline{\mathbf{G}}^{(i)} - \overline{\mathbf{R}}_{\mathbf{x}}^{(i)-1} \mathbf{P}_{\mathbf{x}}^{(i)} \mathbf{q}^{(i)} \right\|^2_{\overline{\mathbf{R}}_{\mathbf{x}}^{(i)}} + \left\| \mathbf{q}^{(i)} \right\|^2_{\mathbf{R}_{\mathbf{x}}^{(i)} - \mathbf{P}_{\mathbf{x}}^{(i)H} \overline{\mathbf{R}}_{\mathbf{x}}^{(i)-1} \mathbf{P}_{\mathbf{x}}^{(i)}}. \tag{52}$$

In the above objective function, $\overline{\mathbf{G}}^{(i)}$ is contained only in the first term, and the term can be minimized, not depending on the value of $\mathbf{q}^{(i)}$, when $\overline{\mathbf{G}}^{(i)}$ takes the following value.

$$\overline{\mathbf{G}}^{(i)} = \overline{\mathbf{R}}_{\mathbf{x}}^{(i)-1} \mathbf{P}_{\mathbf{x}}^{(i)}. \tag{53}$$

So, this is a solution[6] of $\overline{\mathbf{G}}^{(i)}$ that globally minimizes the objective function given the time-varing variance $\lambda_t^{(i)}$. Interestingly, this solution is identical to that of the conventional

[6]This is not a unique solution. The first term is minimized even when an arbitrary matrix, of which null space includes $\mathbf{q}^{(i)}$, is added to Eq. (53).

Algorithm 1: Source-packed factorization-based optimization for estimation of all the sources.

Data: Observed signal $\mathbf{x}_t$ for all t
TF masks $\lambda_t^{(i)}$ for all t and $1 \le i \le I$
Result: Estimated sources $y_t^{(i)}$ for all t and $1 \le i \le I$
Initialize $\lambda_t^{(i)}$ as $\|\mathbf{x}_t\|^2_{\mathbf{I}_M}/M$ for all t and $1 \le i \le I$
Initialize $\mathbf{q}^{(i)}$ as the ith column of $\mathbf{I}_M$ for $1 \le i \le I$
Initialize $\mathbf{z}_t$ as $\mathbf{x}_t$ for all t
repeat
  $\overline{\mathbf{R}}_{\mathbf{x}}^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \overline{\mathbf{x}}_t^{\mathsf{H}}}{\lambda_t^{(i)}}$ for $1 \le i \le I$
  $\mathbf{P}_{\mathbf{x}}^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \mathbf{x}_t^{\mathsf{H}}}{\lambda_t^{(i)}}$ for $1 \le i \le I$
  $\quad + \sum_{i=1}^{I} \mathbf{q}^{(i)} \mathbf{q}^{(i)\mathsf{H}} \overline{\mathbf{R}}_{\mathbf{x}}^{(i)\top}$
  $\quad + \sum_{i=1}^{I} \mathbf{q}^{(i)} \mathbf{P}_{\mathbf{x}}^{(i)} \mathbf{q}^{(i)}$
  Begin Add orthogonal complement beamformer
    Set $\mathbf{q}^{(I+1)}, \ldots, \mathbf{q}^{(M)}$ be orthonormal bases for the orthogonal complement $\mathbf{Q}^{?}$ of $\mathbf{Q}$
    $\lambda_t^{?} \leftarrow \frac{1}{M-I}\sum_{i=I+1}^{M} |\mathbf{q}^{(i)\mathsf{H}} \mathbf{z}_t|^2$
    $\overline{\mathbf{R}}_{\mathbf{x}}^{?} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \overline{\mathbf{x}}_t^{\mathsf{H}}}{\lambda_t^{?}}$
    $\mathbf{P}_{\mathbf{x}}^{?} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \mathbf{x}_t^{\mathsf{H}}}{\lambda_t^{?}}$
    $\quad + \sum_{i=I+1}^{M} \mathbf{q}^{(i)} \mathbf{q}^{(i)\mathsf{H}} \overline{\mathbf{R}}_{\mathbf{x}}^{?\top}$
    $\quad + \sum_{i=I+1}^{M} \mathbf{q}^{(i)} \mathbf{P}_{\mathbf{x}}^{?} \mathbf{q}^{(i)}$
  End
  $\overline{\mathbf{g}} \leftarrow$ [ ]
  $\mathbf{z}_t \leftarrow \mathbf{x}_t - \overline{\mathbf{X}}_t \overline{\mathbf{g}}$
  Estimate $\boldsymbol{\nu}^{(i)}$ based on $\mathbf{z}_t$ and $\lambda_t^{(i)}$ for $1 \le i \le I$
  $\mathbf{R}_{\mathbf{z}}^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{z}_t (\mathbf{z}_t)^{\mathsf{H}}}{\lambda_t^{(i)}}$ for $1 \le i \le I$
  $\mathbf{q}^{(i)} \leftarrow \frac{(\mathbf{R}_{\mathbf{z}}^{(i)})^{-1}\boldsymbol{\nu}^{(i)}}{(\boldsymbol{\nu}^{(i)})^{\mathsf{H}} \mathbf{R}_{\mathbf{z}}^{(i)-1}\boldsymbol{\nu}^{(i)}}$ for $1 \le i \le I$
  $y_t^{(i)} \leftarrow \mathbf{q}^{(i)\mathsf{H}} \mathbf{z}_t$ for $1 \le i \le I$
  $\lambda_t^{(i)} \leftarrow |y_t^{(i)}|^2$ for $1 \le i \le I$
until *convergence*

WPE dereverberation. This means that the WPE filter optimized solely for dereverberation can perform the optimal dereverberation for the joint optimization not depending on the subsequent beamforming, provided the time-varying variance of the desired source is given for the optimization. In addition, unlike the source-packed factorization approach, this approach does not need to compensate for the dimensionality reduction of the beamformer output for the update of $\overline{\mathbf{G}}^{(i)}$ because it considers a whole signal space without adding any modification. We refer to this filter $\overline{\mathbf{G}}^{(i)}$ as a single-target WPE filter in this paper.

Once $\overline{\mathbf{G}}^{(i)}$ is obtained as the above solution, the objective function in Eq. (19) can be rewritten as

$$L_i(\mathbf{q}^{(i)}) = \mathbf{q}^{(i)\mathsf{H}} \mathbf{R}_{\mathbf{z}}^{(i)} \mathbf{q}^{(i)} \quad \text{s.t.} \quad \mathbf{q}^{(i)\mathsf{H}} \boldsymbol{\nu}^{(i)} = 1, \tag{54}$$

Algorithm 2: Source-wise factorization-based optimization for estimation of the ith source

Data: Observed signal $\mathbf{x}_t$ for all t
TF masks $\lambda_t^{(i)}$ for all t
Result: Estimated ith source $y_t^{(i)}$ for all t
Initialize $\lambda_t^{(i)}$ as $\|\mathbf{x}_t\|^2_{\mathbf{I}_M}/M$ for all t
repeat
  $\overline{\mathbf{R}}_{\mathbf{x}}^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \overline{\mathbf{x}}_t^{\mathsf{H}}}{\lambda_t^{(i)}}$
  $\mathbf{P}_{\mathbf{x}}^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\overline{\mathbf{x}}_t \mathbf{x}_t^{\mathsf{H}}}{\lambda_t^{(i)}}$
  $\overline{\mathbf{G}}^{(i)} \leftarrow \overline{\mathbf{R}}_{\mathbf{x}}^{(i)-1} \mathbf{P}_{\mathbf{x}}^{(i)}$
  $\mathbf{z}_t^{(i)} \leftarrow \mathbf{x}_t - \overline{\mathbf{G}}^{(i)\mathsf{H}} \overline{\mathbf{x}}_t$
  Estimate $\boldsymbol{\nu}^{(i)}$ based on $\mathbf{z}_t^{(i)}$ and $\lambda_t^{(i)}$
  $\mathbf{R}_{\mathbf{z}}^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{z}_t^{(i)} \mathbf{z}_t^{(i)\mathsf{H}}}{\lambda_t^{(i)}}$
  $\mathbf{q}^{(i)} \leftarrow \frac{(\mathbf{R}_{\mathbf{z}}^{(i)})^{-1}\boldsymbol{\nu}^{(i)}}{(\boldsymbol{\nu}^{(i)})^{\mathsf{H}} \mathbf{R}_{\mathbf{z}}^{(i)-1}\boldsymbol{\nu}^{(i)}}$
  $y_t^{(i)} \leftarrow \mathbf{q}^{(i)\mathsf{H}} \mathbf{z}_t^{(i)}$
  $\lambda_t^{(i)} \leftarrow |y_t^{(i)}|^2$
until *convergence*

where $\mathbf{R}_{\mathbf{z}}^{(i)}$ is a variance-normalized covariance matrix of the output of the single-target WPE filter, calculated as

$$\mathbf{R}_{\mathbf{z}}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbf{z}_t^{(i)} \mathbf{z}_t^{(i)\mathsf{H}}}{\lambda_t^{(i)}} \in \mathbb{C}^{M \times M}. \tag{55}$$

Then, the solution can be obtained, under the distortionless constraint, as a wMPDR beamformer defined by

$$\mathbf{q}^{(i)} = \frac{\mathbf{R}_{\mathbf{z}}^{(i)-1}\boldsymbol{\nu}^{(i)}}{\boldsymbol{\nu}^{(i)\mathsf{H}} \mathbf{R}_{\mathbf{z}}^{(i)-1}\boldsymbol{\nu}^{(i)}}. \tag{56}$$

Eqs. (54) to (56) are very similar to Eqs. (40) to (42), and the difference is whether the dereverberation is performed by multiple-target WPE filter or single-target WPE filter.

With the source-wise factorization, the solution can be obtained in closed form when $\lambda_t^{(i)}$ and $\boldsymbol{\nu}^{(i)}$ are given, similar to the case with the direct optimization of the MISO CBFs. In addition, the output of the WPE filter is obtained as $\mathbf{z}_t^{(i)}$ in Eq. (16), and can be efficiently used for estimation of the RTFs. Furthermore, the size of the temporal-spatial covariance matrix in Eq. (31) is much smaller than that in Eq. (38) of the source-packed factorization, and thus the computational cost can be reduced. (See Section IV for more detailed discussion on the computing cost.)

E. Processing flow with estimation of $\lambda_t^{(i)}$ and $\boldsymbol{\nu}^{(i)}$

This subsection describes examples of processing flows, shown in Algorithms 1 and 2, for optimizing a CBF based on source-packed factorization and source-wise factorization, including the estimation of the time-varying variances, $\lambda_t^{(i)}$,

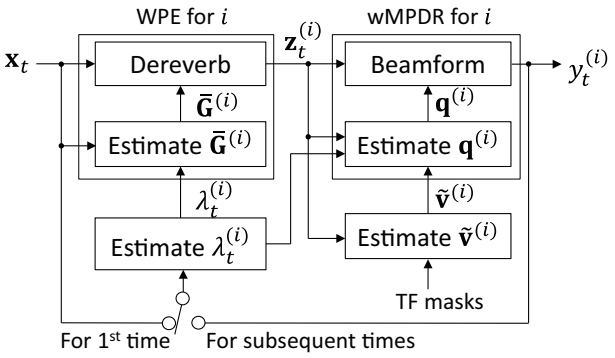


Fig. 2. Processing flow of source-wise factorization-based CBF for estimating a source i.

and the RTFs, $\mathbf{v}^{(i)}$. Hereafter, we refer to both algorithms as A-1 and A-2 for brevity. While A-1 estimates all the sources, $y_t^{(i)}$ for all i, at the same time from the observed signal $\mathbf{x}_t$, A-2 estimates only one of the sources, $y_t^{(i)}$ for a certain i, and (if necessary) is repeatedly applied to the observed signal to estimate all the sources one after another. As auxiliary inputs, TF masks are provided for both algorithms. A TF mask $_t^{(i)}$ is associated with a source and a TF point, takes a value between 0 and 1, and indicates whether the desired signal of the source dominates the TF point ($_t^{(i)} = 1$) or not ($_t^{(i)} = 0$). The TF masks over all the TF points are used to estimate the RTF(s) of the desired signal(s) in line 19 of A-1 and line 7 of A-2. (See Section III-E1 for the detail of estimation of the TF masks and the RTFs.)

Both algorithms estimate the time-varying variances $_t^{(i)}$ based on the same objective as that for the CBF, defined in Eq. (19). Because a closed form solution to the estimation of the CBF and the time-varying variances is not known, an iterative and alternate optimization scheme is introduced to both algorithms. In each iteration, the time-varying variances, $_t^{(i)}$, are updated in line 23 of A-1 and line 11 of A-2 as the power of the previously estimated values of the desired signal $y_t^{(i)}$, and then the CBF and the desired signal $y_t^{(i)}$ are updated while fixing the time-varying variances. The iteration is repeated until convergence is obtained.

The optimization methods described in Sections III-B and III-D are used in respective algorithms for update of the CBF and the desired signal(s). The WPE filter is first estimated in lines 5 to 17 of A-1 and lines 3 to 5 of A-2, and applied in line 18 of A-1 and line 6 of A-2. After the RTF(s) is updated using the dereverberated signals, the wMPDR beamformer is estimated in lines 20 and 21 of A-1 and lines 8 and 9 of A-2, and applied in line 22 of A-1 and line 10 of A-2.

Figure 2 also illustrates the processing flow of a CBF with the source-wise factorization for estimating a source i.

1) Methods for estimating TF masks and RTFs: In experiments, for estimating TF masks, $_t^{(i)}$, for all i and t at each frequency, we used a Convolutional Neural Network that works in the TF domain and is trained using utterance-level Permutation Invariant Training criterion (CNN-uPIT) [43]. According to our preliminary experiments [32], we set the network structure as a CNN with a large receptive field similar to one used by a fully-Convolutional Time-domain Audio Separation Network (Conv-TasNet) [44]. The network was trained so that it receives the output of the WPE filter that is obtained at the first iteration in the iterative optimization of the CBF, and estimates the TF masks of the desired signals. The input of the network was set as concatenation of the real and imaginary parts of STFT coefficients, and the loss function was set as the (scale-dependent) signal-to-distortion ratio (SDR) of the enhanced signal obtained by multiplying the estimated masks to the observed signal. For the training and validation data, we synthesized mixtures using two utterances randomly extracted from the WSJ-CAM0 corpus [45] and two room impulse responses and background noise extracted from the REVERB Challenge training set [18].

For the estimation of the RTFs, $\mathbf{v}^{(i)}$, we adopt a method based on eigenvalue decomposition with noise covariance whitening [46], [47]. With this technique, the steering vector $\mathbf{v}^{(i)}$ is first estimated as

$$\mathbf{v}^{(i)} = R_{ni}\mathrm{MaxEig}\left(R_{ni}^{-1}R_i\right); \tag{57}$$

where MaxEig( ) is a function that calculates the eigenvector corresponding to the maximum eigenvalue, R_i and R_{ni} are a spatial covariance matrix of the i-th desired signal and that of the other signals, respectively, estimated as

$$R_i = \frac{\sum_t {}_t^{(i)}\mathbf{z}_t^{(i)}\mathbf{z}_t^{(i)H}}{\sum_t {}_t^{(i)}}; \tag{58}$$

$$R_{ni} = \frac{\sum_t (1 - {}_t^{(i)})\mathbf{z}_t^{(i)}\mathbf{z}_t^{(i)H}}{\sum_t (1 - {}_t^{(i)})}. \tag{59}$$

Then, the RTF is obtained by Eq. (4).

IV. Discussion

In summary, the proposed techniques can optimize a CBF for jointly performing DN+DR+SS with greatly reduced computing cost in comparison with the direct application of the conventional joint optimization technique proposed for DR+SS to DN+DR+SS. With the conventional technique, it is necessary to calculate the huge covariance matrix in order to take into account the dependency of $\bar{\mathbf{G}}$ on $\mathbf{Q}$ inherently introduced in the source-packed factorization. This makes the computing cost of the conventional technique extremely high. In contrast, the proposed extension of the source-packed factorization approach reduces the size of the matrix to be calculated substantively from $M^2(L$) for to $M(L$) for $\bar{\mathbf{R}}_{\mathbf{x}}$, and thus can effectively reduce the computing cost.

On the other hand, with the source-wise factorization, $\bar{\mathbf{G}}^{(i)}$ can be optimized independently of $\mathbf{q}^{(i)}$, which also allows us to reduce the size of the matrix to be calculated to the same as that of the proposed extension of the source-packed factorization approach. In addition, we can skip the calculation of an additional matrix, $\bar{\mathbf{R}}_{\mathbf{x}}^{?}$, and that of the inverse of the huge matrix, ^{1}, which are required for the proposed extension of the source-packed factorization approach. This makes the source-wise factorization approach computationally

TABLE I
CBFS COMPARED IN EXPERIMENTS. (1) AND (2) ARE CONVENTIONAL CASCADE CONFIGURATION APPROACHES, (5) IS A CONVENTIONAL JOINT OPTIMIZATION APPROACH, (6) AND (7) ARE PROPOSED JOINT OPTIMIZATION APPROACHES, AND (3) AND (4) ARE TEST CONDITIONS USED JUST FOR COMPARISON. (5), (6), AND (7) ARE CATEGORIZED AS "JOINTLY OPTIMAL" BECAUSE THEY ARE COMPOSED OF WPE AND WMPDR AND OPTIMIZED BASED ON INTEGRATED VARIANCE ESTIMATION (SEE FIG. 3 FOR THE DIFFERENCE BETWEEN SEPARATE AND INTEGRATED VARIANCE ESTIMATION).

Name of method	Jointly optimal	WPE	BF	Variance estimation	Category
(1) WPE+MPDR (separate)		Multiple-target	MPDR	Separate	Cascade (conventional)
(2) WPE+MVDR (separate)		Multiple-target	MVDR	Separate	Cascade (conventional)
(3) WPE+wMPDR (separate)		Multiple-target	wMPDR	Separate	Test condition
(4) WPE+MPDR (integrated)		Single-target	MPDR	Integrated	Test condition
(5) Source-packed factorization (conventional)	✓	Multiple-target	wMPDR	Integrated	Jointly optimal (conventional)
(6) Source-packed factorization (extended)	✓	Multiple-target	wMPDR	Integrated	Jointly optimal (proposed)
(7) Source-wise factorization	✓	Single-target	wMPDR	Integrated	Jointly optimal (proposed)

further efficient. A drawback of the source-wise factorization is that it has to handle I-times larger number of dereverberated signals than the source-packed factorization.

The source-wise factorization approach has additional benefits w.r.t. computational efficiency when it is used in specific scenarios listed below:

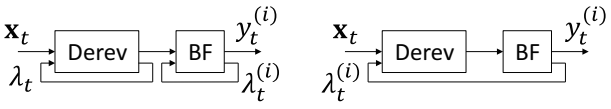
The source-wise factorization approach can estimate the CBF by a closed-form equation when time-varying source variances are given, or estimated, e.g., using neural networks [15], [12]. In such a case, we can skip the iterative optimization. In contrast, the source-packed factorization approach needs to maintain iterations to estimate Q and $\bar{\mathsf{g}}$ alternately due to their mutual dependency.

The source-wise factorization approach is advantageous when it is combined with neural network-based single target speaker extraction that has been actively studied recently [13]. With this combination, we can skip the estimation of sources other than the target source, allowing us to further reduce the computing cost.

V. EXPERIMENTS

This section experimentally confirms the effectiveness of the proposed joint optimization approaches. Table I summarizes optimization methods to be compared in the experiments (See Sections V-C and V-D for the detail of the methods). We compare them in the following three aspects.

1) Effectiveness of joint optimization
   We compare a CBF with and without joint optimization in terms of estimation accuracy. The source-wise factorization approach (Table I (7)) is compared with the conventional cascade configuration (Table I (1) and (2)), and two additional test conditions (Table I (3) and (4)).
2) Comparison among joint optimization approaches
   We compare three joint optimization approaches, i.e., the source-packed factorization approach with its conventional setting (Table I (5)), its proposed extension (Table I (6)), and the source-wise factorization approach (Table I (7)), described, respectively, in Sections III-B1, III-B2, and III-D, in terms of computational efficiency and estimation accuracy.
3) Evaluation using oracle masks
   We use oracle masks instead of estimated masks for evaluating a CBF, in order to test the performance of a CBF using a different type of masks and also to obtain the top-line performance of a CBF.



(a) Separate optimization (b) Integrated optimization

Fig. 3. Separate and integrated variance optimization schemes. While the separate variance optimization updates $_t$ for Derev as the variance of the Derev output, the integrated variance optimization updates it as the variance of the beamformer output. As a consequence, $_t$ for Derev is common to all the sources with separate variance optimization.

A. Dataset and evaluation metrics

For the evaluation, we prepared for a set of noisy reverberant speech mixtures (REVERB-2MIX) using the REVERB Challenge dataset (REVERB) [18]. Each utterance in REVERB contains a single reverberant speech with moderate stationary diffuse noise. For generating a set of test data, we mixed two utterances extracted from REVERB, one from its development set (Dev set) and the other from its evaluation set (Eval set), so that each pair of mixed utterances were recorded in the same room, by the same microphone array, and under the same condition (near or far, RealData or SimData). We categorize the test data according to the original categories of the data in REVERB (e.g., SimData or RealData). We created the same number of mixtures in the test data as in the REVERB Eval set, such that each utterance in the REVERB Eval set is contained in one of the mixtures in the test data. Furthermore, the length of each mixture in the test data was set at the same as that of the corresponding utterance in the REVERB Eval set, while the utterance from Dev set was trimmed or zero-padded at its end to have the same length as that of Eval set.

For experiments in Section V-E, we also prepared for a set of noisy reverberant speech mixtures, each of which is composed of three speaker utterances (REVERB-3MIX). We created REVERB-3MIX by adding one utterance extracted from REVERB Dev set to each mixture in REVERB-2MIX. Only RealData (i.e., real recordings of reverberant data) was created for REVERB-3MIX.

In the experiments, we estimated two and three speech signals from each mixture, respectively, for REVERB-2MIX

TABLE II
BEAMFORMER CONFIGURATIONS USED IN EXPERIMENTS

	M	L at each freq. range (kHz)			#iterations
		0.0-0.8	0.8-1.5	1.5-8.0	
Config-1	8	20	16	8	10
Config-2	4	20	16	8	10

and REVERB-3MIX, and evaluated only one of them corresponding to the REVERB Eval set, using baseline evaluation tools prepared for it. We selected the signal to be evaluated from all the estimated speech signals based on the correlation between the separated signals and the original signal in the REVERB Eval set. As objective measures for speech enhancement [48], we used the Cepstrum Distance (CD), the Frequency-Weighted Segmental SNR (FWSSNR), the Perceptual Evaluation of Speech Quality (PESQ), and the Short-Time Objective Intelligibility measure (STOI) [49]. To evaluate the ASR performance, we used a baseline ASR system for REVERB that was recently developed using Kaldi [50]. This system is composed of a Time-Delay Neural Network (TDNN) acoustic model trained using lattice-free maximum mutual information (LF-MMI) and online i-vector extraction, and a trigram language model. They are trained on the REVERB training set.

B. Configurations of CBF

Table I summarizes two configurations of the CBF examined in experiments including the number of microphones M, the filter length L, and the number of optimization iterations. The sampling frequency was 16 kHz. A Hann window was used for a short-time analysis with the frame length and shift being set at 32 ms and 8 ms, respectively. The prediction delay was set at = 4 for the WPE filter.

In the iterative optimization, the time-varying variances of sources were initialized as those of the observed signal for the WPE filter and as 1 for the wMPDR beamformer for all the methods.

C. Experiment-1: effectiveness of joint optimization

In this experiment, we evaluated the effectiveness of the joint optimization focusing on its two characteristics. First, we compared three different filter combinations, a WPE filter followed by a wMPDR beamformer (WPE+wMPDR), a WPE filter followed by an MPDR beamformer (WPE+MPDR), and a WPE filter followed by an MVDR beamformer (WPE+MVDR). The first combination is required for the jointly optimal processing, and the others have been used for the conventional cascade configuration. Second, we compared two different variance optimization schemes shown in Fig. 3, namely "separate" and "integrated" variance optimization schemes. With the separate variance optimization, the iterative estimation of the time-varying variance was performed separately for the WPE filter and for the beamformer. This is the scheme used by the conventional cascade configuration. In contrast, with the integrated variance optimization, the iterative estimation was performed jointly for the WPE filter and the beamformer. A significant difference between the two schemes is whether the WPE filter uses the same variances for all the sources or different variances dependant on the sources estimated by the beamformer.

TABLE III
WER (%) FOR REALDATA AND CD (DB), FWSSNR (DB), PESQ AND STOI FOR SIMDATA IN REVERB-2MIX OBTAINED USING DIFFERENT BEAMFORMERS AFTER FIVE ESTIMATION ITERATIONS WITH CONFIG-1. SCORES FOR REVERB-2MIX AND REVERB (I.E., SINGLE SPEAKER) WITH NO ENHANCEMENT (NO ENH), ARE ALSO SHOWN.

Enhancement method	WER	CD	FWSSNR	PESQ	STOI
No Enh (REVERB-2MIX)	62.49	5.44	1.12	1.12	0.55
No Enh (REVERB)	18.61	3.97	3.62	1.48	0.75
MPDR (w/o iteration)	30.79	4.40	3.07	1.45	0.73
MVDR (w/o iteration)	30.89	4.43	3.00	1.44	0.73
wMPDR	28.75	3.96	4.46	1.60	0.75
(1) WPE+MPDR (separate)	23.04	4.30	3.77	1.58	0.77
(2) WPE+MVDR (separate)	23.34	4.34	3.66	1.57	0.76
(3) WPE+wMPDR (separate)	21.53	3.74	5.42	1.77	**0.82**
(4) WPE+MPDR (integrated)	23.22	4.28	3.66	1.56	0.76
(7) Source-wise factorization	**20.03**	**3.67**	**5.57**	**1.80**	0.81

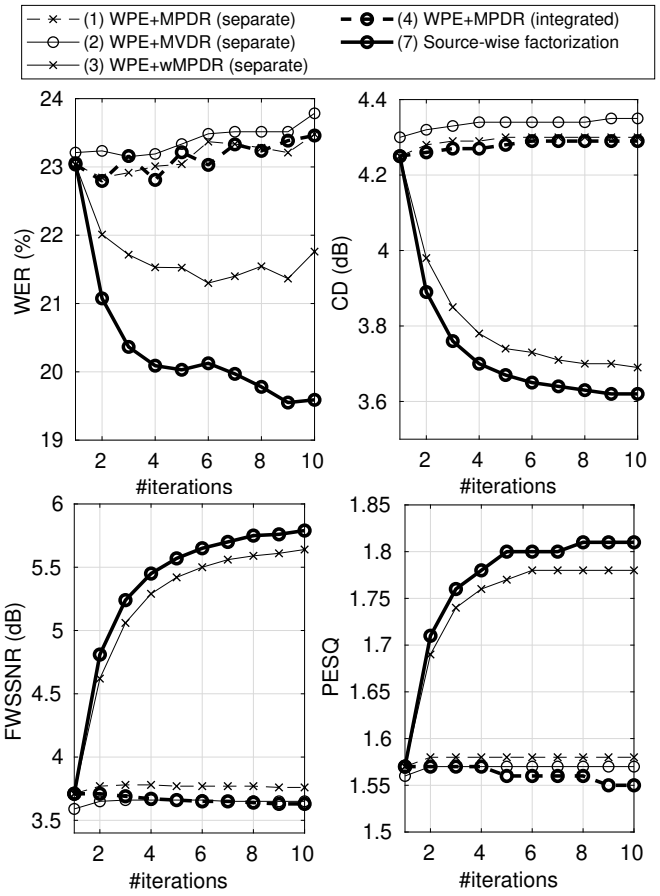
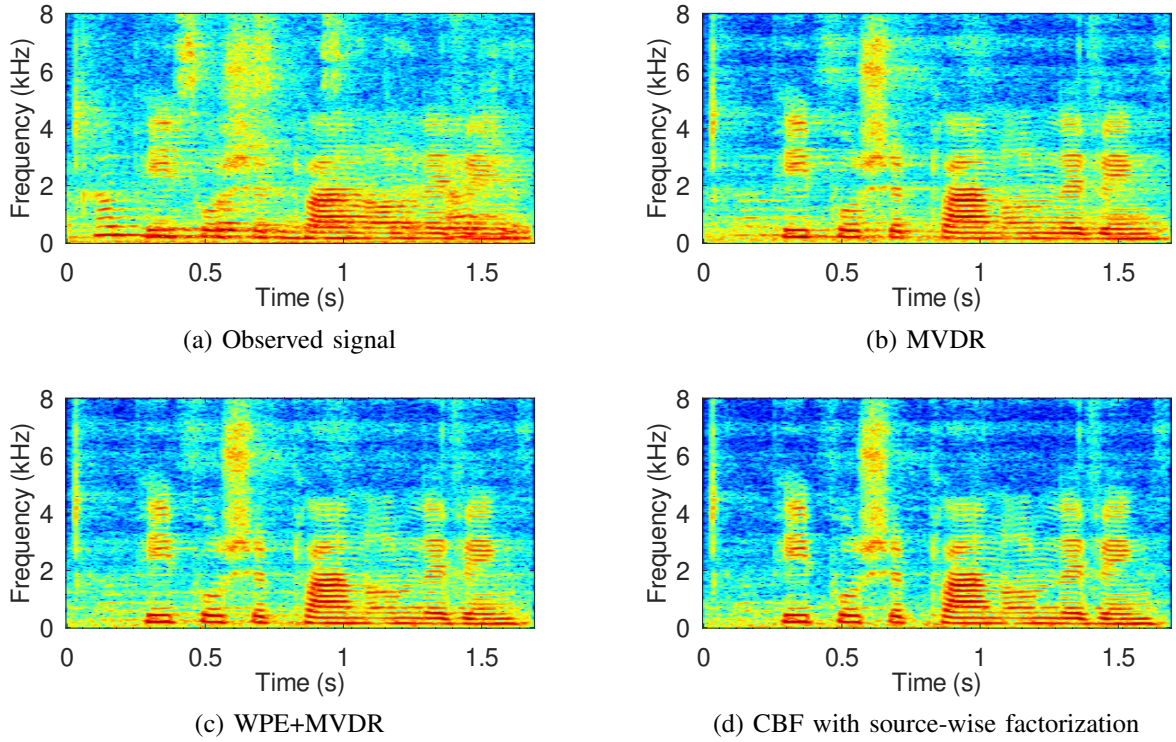


Fig. 4. Comparison among joint optimization and cascade configuration approaches when using WPE+MPDR and WPE+wMPDR with integrated and separate optimization schemed using Config-1 for REVERB-2MIX.

Table III compares WERs, CDs, FWSSNRs, PESQs, and STOIs obtained after five estimation iterations using three beamformers (MPDR, MVDR, and wMPDR), two conven-



(a) Observed signal (b) MVDR

(c) WPE+MVDR (d) CBF with source-wise factorization

Fig. 5. A spectrogram of (a) a noisy reverberant mixture in RealData of REVERB-2MIX and those of enhanced signals obtained by (b) MVDR, (c) WPE+MVDR and (d) CBF with source-wise factorization. The mixture is composed of two female speakers under far conditions.

tional cascade configuration approaches ((1) WPE+MPDR and (2) WPE+MVDR), two test conditions ((3) and (4)), and a proposed joint optimization approach ((7) source-wise factorization). All methods used configuration Config-1 in Table I. Table III shows that 1) WPE+MPDR, WPE+MVDR, and WPE+wMPDR greatly outperformed MPDR, MVDR, and wMPDR, respectively, with all the conditions, 2) the joint optimization approach, i.e., (7) source-wise factorization, substantially outperformed all the other methods in terms of all the measures except for a case in terms of STOI where WPE+wMPDR (separate) gave a slightly better score than (7) source-wise factorization. Furthermore, Fig. 4 shows convergence curves of the two cascade configuration approaches, two test conditions, and the joint optimization approach. The performance of (7) source-wise factorization was the best of all and improved as the number of iterations increased. The second best was (3) WPE+wMPDR (separate). In contrast, the other methods did not improve the scores after the first iteration with both integrated and separate variance optimization schemes.

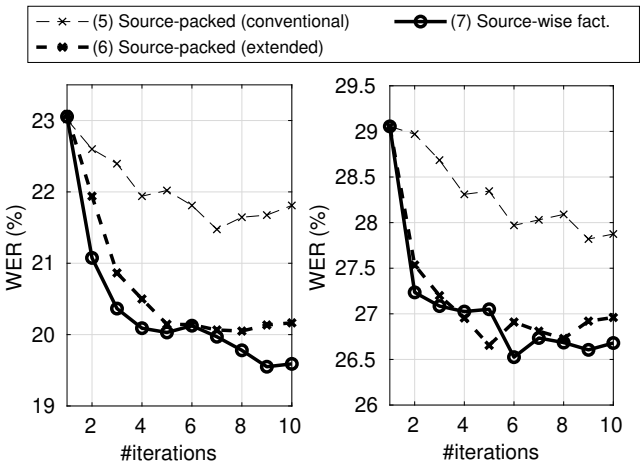
Figure 5 shows a spectrogram of a noisy reverberant mixture in RealData of REVERB-2MIX, and those of enhanced signals obtained using MVDR, WPE+MVDR, and CBF with source-wise factorization. The figure demonstrates that while all the enhancement methods were effective, the CBF with source-wise factorization was the best of all for achieving denoising, dereverberation, and source separation.

The above results clearly show that the two characteristics of the joint optimization approach, i.e., 1) the optimal combination of a WPE filter and a wMPDR beamformer, and 2) the integrated variance optimization, are both important for achieving the optimal performance.

D. Experiment-2: Comparison among joint optimization approaches

In this experiments, we compared three joint optimization approaches, namely two source-packed factorization approaches, respectively, described in Sections III-B1 and III-B2, and denoted as "(5) Source-packed factorization (conventional)" and "(6) Source-packed factorization (extended)," and the source-wise factorization approach, denoted as "(7) Source-wise factorization." (5) Source-packed factorization (conventional) corresponds to the conventional joint optimization technique, and (6) Source-packed factorization (extended) and (7) Source-wise factorization correspond to our proposed methods. Figure 6 compares the WERs obtained using the three joint optimization approaches with Config-1 and Config-2. It shows that the proposed methods, i.e., (6) Source-packed factorization (extended) and (7) Source-wise factorization, performed comparably well and both greatly outperformed (5) Source-packed factorization (conventional).

Table IV compares computing times required for the three approaches to estimate and apply the CBFs with ten estimation iterations for processing a mixture utterance with the length being 9.44 s. The computing time was measured by a Matlab interpreter as the elapsed time. The computing time for estimating masks, which was 0.63 s and 7.2 s, respectively, with and without a GPU (NVIDIA 2080ti), is not included in the table. As shown in the table, for both configurations, (6) Source-packed factorization (extended) greatly reduced



(a) Config-1 (b) Config-2

Fig. 6. WERs (%) obtained for REVERB-2MIX when jointly optimizing WPE+wMPDR based on the source-packed factorization (conventional/extended) and source-wise factorization approaches.

TABLE IV
COMPUTING TIME REQUIRED FOR PROCESSING A MIXTURE UTTERANCE WITH LENGTH OF 9.44 S IN REVERB-2MIX. THE COMPUTING TIME WAS MEASURED BY ELAPSED TIME ON A MATLAB INTERPRETER.

Method	Time (s)	
	Config-1	Config-2
(4) Source-packed factorization (conventional)	3467	688
(5) Source-packed factorization (extended)	209	33
(6) Source-wise factorization	40	23

the computing time in comparison with (5) Source-packed factorization (conventional), and (7) Source-wise factorization further reduced the computing time.

The above results clearly demonstrate the superiority of the two proposed approaches over the conventional joint optimization technique in terms of both computational efficiency and estimation accuracy. However, Table IV indicates that the proposed approaches still require relatively large computing cost, e.g., 40 s computing time for processing 9.44 s utterance with Config-1, to obtain high performance gain shown in Fig. 6 (a). Future work should include solving this problem. For example, it might be mitigated when we set the goal as extraction of a single target source. Then, thanks to the characteristics of the source-wise factorization, we can skip the estimation of the other sources, and skip the iterative estimation using source variances separately estimated, e.g., by a neural network. As a reference, the computing time, 40 s, in Table III required for the source-wise factorization with Config-1 is roughly reduced to 2:0 s for one iteration per source (namely 40 s=10=2), which results in the real-time factor being 0:21 (= 2:0 s=9:44 s).

E. Experiment-3: Evaluation using oracle masks

In this experiment, we examined the performance of CBFs using a different type of masks, i.e., oracle masks. An oracle mask is the power ratio of the desired signal to the observed signal at each TF point, and is calculated using reference signals. The oracle masks can be exactly calculated for SimData in REVERB-2MIX using signal components in the observed signals. In contrast, we can calculate the oracle masks only approximately for RealData because we cannot access the signal components. Thus, we first estimated the desired signals by applying dereverberation and denoising to utterances in REVERB, and then calculated the oracle masks using the estimated desired signals for REVERB-2MIX and REVERB-3MIX.

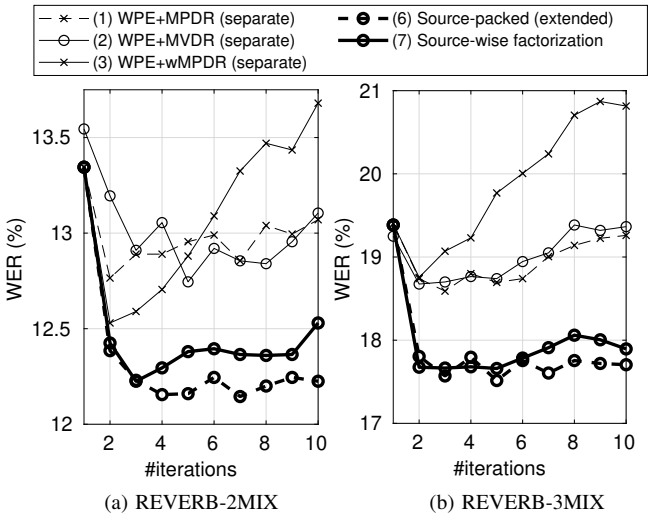
TABLE V
WER (%) FOR REALDATA AND CD (DB), FWSSNR (DB), PESQ, AND STOI FOR SIMDATA IN REVERB-2MIX OF ENHANCED SIGNALS OBTAINED BASED ON ORACLE MASKS USING DIFFERENT BEAMFORMERS AFTER THREE ESTIMATION ITERATIONS WITH CONFIG-1. SCORES FOR REVERB-2MIX WITH NO ENHANCEMENT (NO ENH) AND THOSE OBTAINED BY APPLYING AN OPTIMAL WMPDR CBF, WPD [30], TO REVERB (I.E., SINGLE SPEAKER), ARE ALSO SHOWN.

Enhancement method	WER	CD	FWSSNR	PESQ	STOI
No Enh (REVERB-2MIX)	62.49	5.44	1.12	1.12	0.55
WPD (REVERB) [30]	8.91	2.59	8.29	2.41	0.91
MPDR (w/o iteration)	20.16	3.53	5.49	1.86	0.84
MVDR (w/o iteration)	20.32	3.56	5.36	1.84	0.83
wMPDR	20.12	3.31	6.11	1.96	0.86
(1) WPE+MPDR (separate)	12.89	3.39	6.11	2.10	0.87
(2) WPE+MVDR (separate)	12.91	3.32	6.30	2.07	0.87
(3) WPE+wMPDR (separate)	12.59	3.12	6.84	2.21	0.89
(6) Source-packed fact.	**12.23**	3.02	7.15	**2.33**	**0.90**
(7) Source-wise fact.	**12.23**	**2.98**	**7.25**	2.32	**0.90**

Table V shows WERs, CDs, FWSSNRs, PESQs, and STOIs measured on enhanced signals obtained from REVERB-2MIX using various (non-convolutional) beamformers and CBFs after three estimation iterations. As a reference, the table also includes scores denoted by "WPD (REVERB)," which was reported in [30] and obtained by applying an optimal wMPDR CBF, referred to as WPD (see also Section III-C in this paper), to REVERB, i.e., noisy reverberant single speaker utterances. In addition, the convergence curves obtained using different CBFs in terms of WERs for REVERB-2MIX and REVERB-3MIX, and those obtained in terms of CDs, FWSSNRs, PESQs, and STOIs for REVERB-2MIX are shown, respectively, in Figs. 7 and 8. In all these results, the two joint optimization approaches, (6) source-packed factorization (extended) and (7) source-wise factorization, outperformed all the other methods in terms of all the measures. As a whole, almost the same tendency was observed as the cases using the estimated masks. One exception is that the WERs obtained using the source-wise factorization tended to increase after a few iterations although such tendency was not observed in terms of signal distortion measures. This means that improvement in the signal level distortion does not necessarily results in improvement in the WERs, and suggests the importance of optimization by ASR level criteria, similar to the conventional beamforming techniques [51], [52].

VI. CONCLUDING REMARKS

This paper presented methods for optimizing a CBF that performs DN+DR+SS based on the ML estimation. We introduced two different approaches for factorizing a CBF, i.e.,



(a) REVERB-2MIX (b) REVERB-3MIX

Fig. 7. Comparison of WERs among cascade configuration and joint optimization approaches using Config-1 for REVERB-2MIX and REVERB-3MIX.

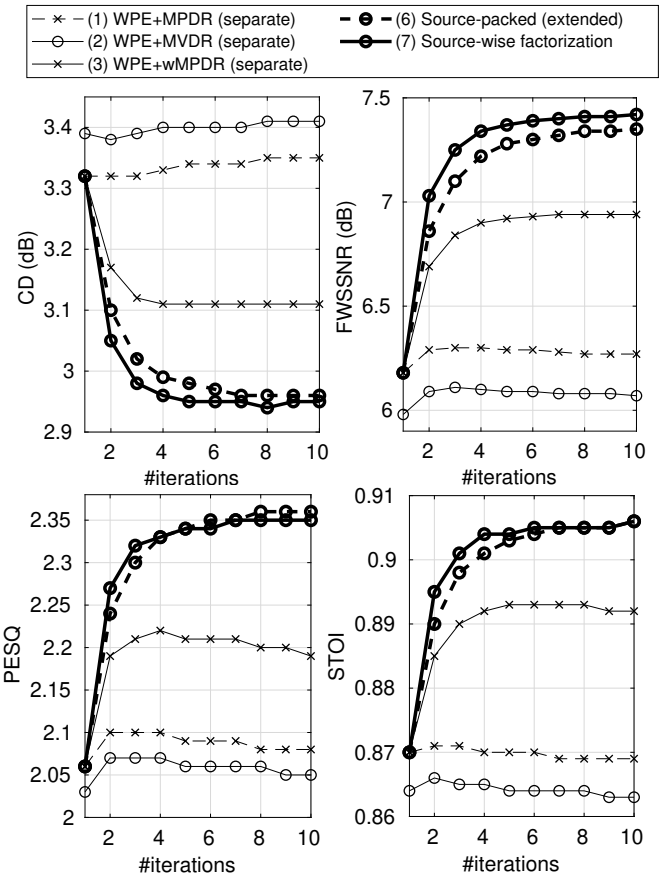


Fig. 8. Comparison of CDs, FWSSNRs, PESQs, and STOIS among cascade configuration and joint optimization approaches using Config-1 for REVERB-2MIX.

source-packed and source-wise factorization approaches, and derived the optimization algorithms for respective approaches. It was shown that a CBF can be factorized without loss of optimality into a multiple-target WPE filter followed by wMPDR beamformers using the source-packed factorization approach, and into a set of single-target WPE filters followed by wMPDR beamformers using the source-wise factorization approach. This paper also presented overall processing flows for both approaches on an assumption that TF masks are provided as auxiliary inputs. In the flows, the time varying source variances, which are required for the ML estimation, can be optimally estimated jointly with the CBF using iterative optimization, and steering vectors of the desired signals, which are required for beamformer optimization, can be reliably estimated based on the dereverberated multichannel signals obtained at a step of the optimization.

Experiments using noisy reverberant sound mixtures show that the proposed optimization approaches substantially improve the performance of the CBF in comparison with the conventional cascade configuration in terms of ASR performance and reduction of signal distortion. It is also shown that the proposed approaches can greatly reduce the computing cost with improved estimation accuracy in comparison with the conventional joint optimization technique. The proposed approaches, however, still result in relatively large computing costs to obtain high performance gain. The solution to this problem should be included in the future work.

APPENDIX A
DERIVATION OF EQS. (43) AND (44)

We can rewrite in Eq. (38) using Eq. (36) as

$$= \frac{1}{T} \sum_t \overline{\mathbf{X}}_t^{\mathsf{H}} \mathbf{q}_{;t} \overline{\mathbf{X}}_t; \tag{60}$$

$$= \frac{1}{T} \sum_t \sum_i \frac{1}{\lambda_t^{(i)}} \left(\mathbf{q}^{(i)\mathsf{H}} \overline{\mathbf{X}}_t \right)^{\mathsf{H}} \left(\mathbf{q}^{(i)\mathsf{H}} \overline{\mathbf{X}}_t \right)^{\mathsf{T}}: \tag{61}$$

Using Eq. (33), $\mathbf{q}^{(i)\mathsf{H}} \overline{\mathbf{X}}_t$ can further be rewritten as

$$\mathbf{q}^{(i)\mathsf{H}} \overline{\mathbf{X}}_t = \mathbf{q}^{(i)\mathsf{H}} \left(\mathbf{I}_M \otimes \overline{\mathbf{x}}_t^T \right); \tag{62}$$

$$= \mathbf{q}^{(i)\mathsf{H}} \left(\overline{\mathbf{x}}_t^T \right): \tag{63}$$

Substituting the above equation in Eq. (61) yields

$$= \frac{1}{T} \sum_t \sum_i \frac{1}{\lambda_t^{(i)}} \left(\mathbf{q}^{(i)} \otimes \overline{\mathbf{x}}_t^{\mathsf{H}} \right) \left(\mathbf{q}^{(i)\mathsf{H}} \otimes \overline{\mathbf{x}}_t^T \right); \tag{64}$$

$$= \frac{1}{T} \sum_t \sum_i \frac{1}{\lambda_t^{(i)}} \left(\mathbf{q}^{(i)} \mathbf{q}^{(i)\mathsf{H}} \right) \otimes \left(\overline{\mathbf{x}}_t \overline{\mathbf{x}}_t^{\mathsf{H}} \right); \tag{65}$$

$$= \sum_i \left(\mathbf{q}^{(i)} \mathbf{q}^{(i)\mathsf{H}} \right) \otimes \overline{\mathbf{R}}_{\mathbf{x}}^{(i)}: \tag{66}$$

Similarly, we can obtain

$$= \frac{1}{T} \sum_t \overline{\mathbf{X}}_t^{\mathsf{H}} \mathbf{q}_{\mathbf{x}} \mathbf{x}_t; \tag{67}$$

$$= \frac{1}{T} \sum_t \sum_i \frac{1}{\lambda_t^{(i)}} \left(\mathbf{q}^{(i)} \otimes \overline{\mathbf{x}}_t^{\mathsf{H}} \right) \left(\mathbf{q}^{(i)\mathsf{H}} \mathbf{x}_t \right); \tag{68}$$

$$= \frac{1}{T} \sum_t \sum_i \frac{1}{\lambda_t^{(i)}} \left(\mathbf{q}^{(i)} \otimes \mathbf{x}_t \right) \left(\mathbf{x}_t^{\mathsf{H}} \otimes \mathbf{q}^{(i)} \right); \tag{69}$$

$$= \sum_i \left(\mathbf{q}^{(i)} \otimes \mathbf{P}_{\mathbf{x}}^{(i)} \mathbf{q}^{(i)} \right): \tag{70}$$

REFERENCES

[1] 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.

[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] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2007.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. on Speech, and Audio Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[7] 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. on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2010.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

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

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

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

[12] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation," in *Proc. Interspeech*, 2017, pp. 384–388.

[13] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[14] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, , and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. Interspeech*, 2018.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM trans. on Audio, Speech, and Language Processing*, vol. 23, no. 1, 2015.

[16] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[17] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Trans. Audio, Speech, and Language Processing*, pp. 1901–1913, 2017.

[18] 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.

[19] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU-2015*, 2015, pp. 504–511.

[20] 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.

[21] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants," *IEEE Signal Processing Magazine*, 2019.

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

[23] S. Braun and E. A. P. Habets, "Linear prediction based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM trans. on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1119–1129, 2018.

[24] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. IWAENC*, 2018.

[25] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, January 2011.

[26] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. IWAENC*, 2014.

[27] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. IEEE ICASSP*, 2018, pp. 31–35.

[28] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Submitted to Interspeech*, 2020.

[29] Z. Koldovsky and P. Tichavský, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. on Signal Processing*, vol. 67, no. 4, pp. 1050––1064, 2019.

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

[31] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. ICASSP*, 2020, pp. 216–220.

[32] T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Declroix, and S. Araki, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *Proc. IEEE ICASSP*, 2020.

[33] 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.

[34] 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.

[35] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, pp. 337–340, 2007.

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

[37] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. IEEE ICASSP*, 2008, pp. 85–88.

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

[39] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "Independent low-rank matrix analysis with decorrelation learning," in *IEEE WASPAA*, 2019.

[40] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," in *Proc. Interspeech*, 2019.

[41] 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.

[42] 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, April 2019.

[43] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," in *Interspeech*, 2019.

[44] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.