






# Causal and Relaxed-Distortionless Response Beamforming for Online Target Source Extraction

Yoshiki Masuyama , *Graduate Student Member, IEEE*, Kouei Yamaoka , *Graduate Student Member, IEEE*, Yuma Kinoshita , *Member, IEEE*, Taishi Nakashima , *Student Member, IEEE*, and Nobutaka Ono , *Senior Member, IEEE*

**Abstract**—In this paper, we propose a low-latency beamforming method for target source extraction. Beamforming has been performed in the time-frequency domain and achieved promising results in offline applications. Meanwhile, it causes a long algorithmic delay due to the frame analysis. Such a delay is unacceptable in various low-latency real-time applications, including hearing aids. To reduce this delay, we propose a causal variant of the minimum power distortionless response (MPDR) beamformer. The proposed method constrains the non-causal components of the spatial filter to be zero in the optimization of the MPDR beamformer. The algorithmic delay is reduced to zero by applying the causal spatial filter in the time domain. We further propose to relax the distortionless constraint regarding the gain, which allows us to improve the extraction performance without a phase delay. The Douglas–Rachford splitting method and its online extension are adopted to solve the optimization problems of the proposed methods. In our experiment, the relaxed method outperformed various low-latency beamforming methods in terms of extraction performance.

**Index Terms**—Low-latency, hearing aids, beamforming, adaptive filtering, online optimization.

## I. INTRODUCTION

**T**ARGET source extraction is essential to assisting human-to-human communication in a noisy environment since interference signals degrade speech quality and intelligibility. When multiple microphones are available, beamforming has been widely used to extract the target signal by leveraging spatial information. Beamforming enhances a target signal arriving from a specific direction and suppresses interference signals [1], [2], [3]. For instance, the multichannel Wiener filter [4] and the generalized eigenvalue beamformer [5] intensively suppress noise with a slight distortion of the target signal. To avoid distorting the target signal, distortionless beamformers, including the minimum power distortionless response (MPDR) and minimum variance distortionless response (MVDR) beamformers,

have been investigated [6], [7]. The distortionless beamformers have achieved promising results in many applications, including assistive hearing devices [8], [9], [10], [11], [12].

The convolutive mixing process in the time domain can be well approximated by the instantaneous mixing process in the time-frequency (T-F) domain when a window for the short-time Fourier transform (STFT) is sufficiently long. This significantly reduces the computational complexity of constructing beamformers. A window longer than 100 ms has been widely used to handle reverberation of a popular length.

Beamforming has been typically performed in the T-F domain to assume the instantaneous mixing process. One drawback of beamforming in the T-F domain is its algorithmic delay induced by STFT and its inverse (iSTFT). For instance, when the window length is 100 ms, an algorithmic delay of at least 100 ms is inevitable on the time-domain extracted signal.

In real-time applications including hearing aids [13], [14] and in-car communication systems [15], [16], [17], we should reduce a delay of the processed signal. As users of assistive hearing devices hear both unprocessed and processed signals simultaneously, a delay of the processed signal deteriorates the quality of the heard signal. For instance, listeners without hearing loss can notice a delay of 3 ms and feel uncomfortable with a delay longer than 10 ms [18]. Although listeners with hearing loss tend to tolerate a longer delay, a delay longer than 6 ms is noticeable when they use open-fitting hearing aids [19]. It is thus important to avoid the long algorithmic delay caused by STFT and iSTFT with a popular window length.

To prevent the long algorithmic delay, various beamforming approaches have been developed. One approach uses asymmetric windows [20], [21]. This approach uses a long window in STFT and a short window in iSTFT. The long window allows us to assume the instantaneous mixing process, and the short window reduces the algorithmic delay. Another approach considers the convolutive mixing process in the T-F domain and uses short windows in both STFT and iSTFT [16], [17]. Although various T-F domain approaches have been proposed [16], [17], [20], [21], [22], they cannot reduce the algorithmic delay to zero owing to the use of STFT.

A time-domain approach has also been presented [14]. In detail, a spatial filter constructed in the T-F domain is converted to the time domain. Then, the filter is convolved with the observed time-domain signal, where we can reduce the

Manuscript received 29 November 2022; revised 25 June 2023 and 26 August 2023; accepted 13 October 2023. Date of publication 1 November 2023; date of current version 16 November 2023. This work was supported in part by JSPS KAKENHI under Grants JP20H00613 and JP21J21371 and in part by JST CREST under Grant JPMJCR19A3, Japan. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hakan Erdogan. (Corresponding author: Yoshiki Masuyama.)

The authors are with Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University, Hino, Tokyo 191-0065, Japan (e-mail: yoshiki.masuyama@ieee.org; yamaoka-kouei@ed.tmu.ac.jp; ykinoshita@tsc.u-tokai.ac.jp; nakashima-taishi@ed.tmu.ac.jp; onono@tmu.ac.jp).

Digital Object Identifier 10.1109/TASLP.2023.3329377

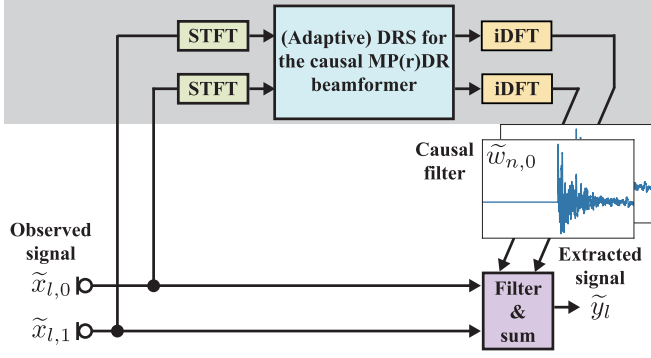


Fig. 1. Overview of the proposed method. The causal filter is optimized by the Douglas–Rachford splitting (DRS) method or its adaptive version based on the STFT of the observed signal (blue block). The target signal is extracted by using the obtained causal filter in the time domain (purple block), which is performed per sample of the observed signal.

algorithmic delay by truncating the non-causal components of the filter. The time-domain approach achieved excellent performance with independent vector analysis under non-reverberant conditions [14]. This approach, however, cannot maintain the directivity and frequency response of the original filter under general conditions. Furthermore, in the case of distortionless beamforming, the truncated filter does not satisfy the distortionless constraint.

In this paper, we propose an optimization-based method to construct a causal MPDR beamformer that simultaneously satisfies both distortionless and causality constraints. The overview of the proposed method is illustrated in Fig. 1. It consists of two parts: optimizing the spatial filter in the T-F domain and convolving the filter with the observed signal in the time domain. Since the filter is constrained to be causal, the algorithmic delay for the second part is zero. When an application tolerates a small number of non-causal components, we can improve the extraction performance by allowing non-causal components for the spatial filter. Furthermore, we propose to relax the distortionless constraint and name the beamformer a causal minimum power relaxed-distortionless response (MPrDR) beamformer. This relaxation increases the degree of freedom (DoF) of the spatial filter without any additional delay of the target signal. To solve the optimization problems, we propose to use the Douglas–Rachford splitting (DRS) method [23]. As the DRS method separately handles the distortionless and causality constraints, we can efficiently update the spatial filter frequency-wise and microphone-wise. We also present an online extension of the optimization algorithm that can leverage observed signals given sequentially.

This paper is an extended version of our conference paper [24], in which an offline optimization algorithm for the causal MPDR and MPrDR beamformers was presented. The main contribution of this paper is developing the online optimization algorithm to adapt the causal MPDR and MPrDR beamformers to non-stationary environments. We also provide a closed-form derivation of the causal MPDR beamformer. The effectiveness of the causality constraint is discussed from the viewpoint of

reducing the mismatch between the linear and circular convolutions. Moreover, by extensive experimental evaluations, we investigate the relationship between hyperparameters and the performance of the proposed methods.

## II. TARGET SOURCE EXTRACTION BY BEAMFORMING

### A. Signal Model

Let the target and interference signals be observed by  $M$  microphones. The mixture observed by the  $m$ th microphone is denoted as  $\tilde{x}_{l,m}$ , where  $l = 0, \dots, L-1$  and  $m = 0, \dots, M-1$  are the sample and microphone indices, respectively. Let us denote STFT of  $\tilde{x}_{l,m}$  as

$$x_{t,f,m} = \sum_{n=0}^{N-1} \tilde{x}_{n+rt,m} g_n e^{-2\pi i f n / F}, \quad (1)$$

where  $i$  is the imaginary unit,  $r$  is the window shift,  $\mathbf{g}$  is the window of length  $N$ , and  $t = 0, \dots, T-1$  and  $f = 0, \dots, F-1$  are the time frame and frequency bin indices, respectively. By using a sufficiently long window, we assume the following instantaneous mixing process in the T-F domain:

$$\mathbf{x}_{t,f} = \mathbf{a}_f s_{t,f} + \mathbf{u}_{t,f}, \quad (2)$$

where  $s_{t,f} \in \mathbb{C}$  is the STFT of the target source image at the reference microphone. The  $M$ -channel STFTs of the observed mixture and the interference signal are respectively given by

$$\mathbf{x}_{t,f} = [x_{t,f,0}, \dots, x_{t,f,M-1}]^T, \quad (3)$$

$$\mathbf{u}_{t,f} = [u_{t,f,0}, \dots, u_{t,f,M-1}]^T, \quad (4)$$

where  $(\cdot)^T$  denotes the transpose. In (2),  $\mathbf{a}_f$  is the relative transfer function (RTF) [25], [26] of the target signal given by

$$\mathbf{a}_f = [1, a_{f,1}, \dots, a_{f,M-1}]^T, \quad (5)$$

where we set the 0th microphone to the reference one without loss of generality. In this paper, we assume that the RTF of the target signal is known or estimated in advance.

### B. MPDR Beamforming in T-F Domain

In the T-F domain, target source extraction by beamforming is formulated as

$$y_{t,f} = \mathbf{w}_f^H \mathbf{x}_{t,f}, \quad (6)$$

where  $\mathbf{w}_f \in \mathbb{C}^M$  is a spatial filter,  $(\cdot)^H$  denotes the Hermitian transpose, and  $y_{t,f} \in \mathbb{C}$  is the extracted target source image. The MPDR beamformer [6] is a popular beamformer that minimizes the power of the extracted signal under the distortionless constraint. The spatial filter is obtained by solving the following optimization problem:

$$\min_{\mathbf{w}_f} \mathbf{w}_f^H \mathbf{R}_f \mathbf{w}_f \quad (7a)$$

$$\text{s.t. } \mathbf{w}_f^H \mathbf{a}_f = 1, \quad (7b)$$

where  $\mathbf{R}_f \in \mathbb{C}^{M \times M}$  is the spatial covariance matrix (SCM):

$$\mathbf{R}_f = \frac{1}{T'} \sum_{t=T'_{\text{start}}}^{T'_{\text{end}}} \mathbf{x}_{t,f} \mathbf{x}_{t,f}^H. \quad (8)$$

Here,  $T'_{\text{start}}$  and  $T'_{\text{end}}$  indicate the duration for computing the SCM, and  $T' = T'_{\text{end}} - T'_{\text{start}} + 1$ . In the offline MPDR beamforming, the SCM is computed from the entire observation where  $T'_{\text{start}}$  and  $T'_{\text{end}}$  are 0 and  $T - 1$ , respectively. In online processing, we consider computing the SCM from the mixtures observed before the target frames, where  $t = T'_{\text{start}}, \dots, T'_{\text{end}}$  corresponds to the previous time frames. Another online extension is introduced in Section IV-A.

The solution of the optimization problem in (7) can be calculated in a closed-form:

$$\mathbf{w}_f = \frac{\mathbf{R}_f^{-1} \mathbf{a}_f}{\mathbf{a}_f^H \mathbf{R}_f^{-1} \mathbf{a}_f}. \quad (9)$$

The extracted time-domain signal  $\tilde{y}_l$  is computed by applying iSTFT to  $y_{t,f}$  in (6). We can obtain the MVDR beamformer [6] by replacing the SCM of the observed signal  $\mathbf{R}_f$  in (9) by that of the interference signal.

MPDR and MVDR beamformers have demonstrated their effectiveness in offline [27], [28], [29], [30], [31] and frame-level online [32], [33], [34] settings. One limitation of the T-F domain beamforming methods is their algorithmic delay. In detail, the delay usually becomes longer than 100 ms because a window longer than 100 ms is widely used in STFT to model reverberation by the instantaneous mixing process in (2). Such a long delay, however, is not acceptable in various real-time applications, including hearing aids [18], [19].

### C. Low-Latency Spatial Filtering in Time Domain

To avoid the algorithmic delay due to STFT and iSTFT, a time-domain method has been developed [14]. It comprises the estimation of the spatial filter in the T-F domain and the extraction of the target signal in the time domain. In detail, the estimated spatial filter  $\mathbf{w}_f$  is converted to the time domain by the inverse discrete Fourier transform (iDFT) as follows:

$$\tilde{\mathbf{w}}_m = \mathbf{D}^{-1} [w_{0,m}, \dots, w_{F-1,m}]^H, \quad (10)$$

where  $\mathbf{D} \in \mathbb{C}^{F \times F}$  is the DFT matrix. In the time domain, we assign the indices of the filter as

$$\tilde{\mathbf{w}}_m = [\tilde{w}_{0,m}, \dots, \tilde{w}_{F/2-1,m}, \tilde{w}_{-F/2,m}, \dots, \tilde{w}_{-1,m}], \quad (11)$$

where the number of DFT points  $F$  is assumed to be even. As illustrated in Fig. 2(a), the positive and negative indices in (11) correspond to the causal and non-causal components of the filters, respectively. The target signal is then extracted by filter-and-sum as follows:

$$\tilde{y}_l = \sum_{m=0}^{M-1} \sum_{n=-F/2}^{F/2-1} \tilde{w}_{n,m} \tilde{x}_{l-n,m}. \quad (12)$$

In this time-domain implementation, the algorithmic delay is reduced to  $F/2$  samples.

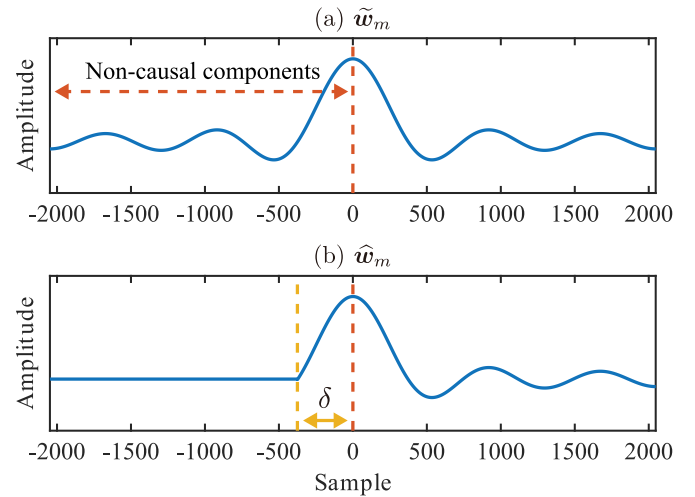


Fig. 2. Illustration of (a) the original filter  $\tilde{\mathbf{w}}_m$  and (b) the truncated one  $\hat{\mathbf{w}}_m$ . Here, we circularly shift the filters so that the zero index is centered.

To further shorten the algorithmic delay, non-causal components of the spatial filter  $\tilde{\mathbf{w}}_{n,m}$  are truncated as:

$$\hat{\mathbf{w}}_{n,m} = \begin{cases} \tilde{\mathbf{w}}_{n,m} & (n \in \mathcal{N}_\delta) \\ 0 & (n \notin \mathcal{N}_\delta) \end{cases}, \quad (13)$$

where  $\delta$  is the number of allowable non-causal components depending on applications, and  $\mathcal{N}_\delta = [-\delta, F/2 - 1]$ . The truncated filter has only a few non-causal components as depicted in Fig. 2(b), and we call such a filter a quasi-causal filter. By using the quasi-causal filter, the algorithmic delay for (12) becomes only  $\delta$  samples. Although the truncation reduces the algorithmic delay to arbitrary samples, it does not retain the directivity or frequency response of the filter optimized in the T-F domain. When it is applied to the MPDR beamformer, the distortionless constraint in (7b) is also no longer satisfied.

## III. CAUSAL MPDR BEAMFORMER

Our low-latency target source extraction method, named the causal MPDR beamforming, is explained in Section III-A. Its optimization problem and closed-form solution are presented in Sections III-B and III-C, respectively. We further propose to relax the distortionless constraint and adopt the DRS method to its optimization in Sections III-D and III-E, respectively. An advantage of the causality constraint is discussed in Section III-F.

### A. Overview of Proposed System

An overview of our low-latency target source extraction method is illustrated in Fig. 1. It consists of two parts: optimization of the spatial filter in the T-F domain and filter-and-sum in the time domain. In the first part, we construct a quasi-causal spatial filter that satisfies the distortionless constraint. The number of non-causal components is adjustable depending on applications.

In the second part, the target signal is extracted sample-by-sample with the quasi-causal spatial filter, which is performed in parallel with the first part. That is, by setting  $T'_{\text{start}}$  and  $T'_{\text{end}}$  in

(8) to the previous time frames, we can reduce the algorithmic delay for target source extraction to  $\delta$  samples. The delay can even be zero samples.

### B. Proposed Formulation

We extend the optimization problem in (7) to ensure the causality of the spatial filter. The proposed method is formulated as the following constrained optimization problem:

$$\min_{(\mathbf{w}_f)_{f=0}^{F-1}} \sum_{f=0}^{F-1} \mathbf{w}_f^H \mathbf{R}_f \mathbf{w}_f, \quad (14a)$$

$$\text{s.t. } \mathbf{w}_f^H \mathbf{a}_f = 1 \quad \forall f, \quad (14b)$$

$$(\mathbf{D}^{-1}[\mathbf{w}_{0,m}, \dots, \mathbf{w}_{F-1,m}]^H)_n = 0 \quad \forall m, n \notin \mathcal{N}_\delta, \quad (14c)$$

where the constraint in (14c) means that the filters have non-causal components of at most  $\delta$  samples. The proposed formulation aims to find a distortionless quasi-causal spatial filter that minimizes the cost function in (14a). When  $\delta = F/2$ , i.e., all entries are not constrained to be zero, the optimization problem in (14) coincides with that of the original MPDR beamformer in (7).

Before solving the optimization problem in (14), we show the existence of its solution. The cost function in (14a) is strongly convex by assuming that the SCM  $\mathbf{R}_f$  is positive-definite at all frequencies.<sup>1</sup> In addition, the constraints in (14b)–(14c) are affine, and there exists a spatial filter that satisfies both constraints even when  $\delta = 0$ :

$$w_{f,m} = \begin{cases} 1 & (m = 0) \\ 0 & (m \neq 0) \end{cases}. \quad (15)$$

Hence, the optimization problem in (14) has a solution [23]. Many convex optimization algorithms can globally solve it regardless of initialization.

### C. Closed-Form Solution of Causal MPDR Beamformer

The optimization problem of the causal MPDR beamformer in (14) is an equality constrained least-squares problem. It can be solved in a closed-form by considering the Karush–Kuhn–Tucker (KKT) condition [35]. Let us concatenate the spatial filters at all frequencies as

$$\underline{\mathbf{w}} = [\mathbf{w}_0^T, \dots, \mathbf{w}_{F-1}^T]^T. \quad (16)$$

On the basis of this notation, we reformulate the optimization problem in (14) as

$$\min_{\underline{\mathbf{w}}} \underline{\mathbf{w}}^H \underline{\mathbf{R}} \underline{\mathbf{w}} \quad (17a)$$

$$\text{s.t. } \mathbf{A}^H \underline{\mathbf{w}} = \mathbf{1}, \quad \mathbf{B}^H \underline{\mathbf{w}} = \mathbf{0}, \quad (17b)$$

where  $\mathbf{1} \in \mathbb{C}^F$  and  $\mathbf{0} \in \mathbb{C}^{(F/2-\delta)M}$  are vectors of ones and zeros, respectively. Here,  $\underline{\mathbf{R}} \in \mathbb{C}^{FM \times FM}$  and  $\mathbf{A} \in \mathbb{C}^{FM \times F}$  are respectively given by

$$\underline{\mathbf{R}} = \text{blkdiag}(\mathbf{R}_0, \dots, \mathbf{R}_{F-1}), \quad (18)$$

<sup>1</sup>The positive-definiteness of the SCM is usually assumed in the MPDR beamforming to compute the inverse of the SCM in (9).

$$\mathbf{A} = \text{blkdiag}(\mathbf{a}_0, \dots, \mathbf{a}_{F-1}), \quad (19)$$

where  $\text{blkdiag}(\cdot)$  constructs a block-diagonal matrix by concatenating inputted matrices diagonally. In (17b),  $\mathbf{B} \in \mathbb{C}^{FM \times (F/2-\delta)M}$  is a matrix for converting spatial filters to the time domain and extracting unacceptable non-causal components as follows:

$$\mathbf{B}^H \underline{\mathbf{w}} = \underline{\mathbf{E}} \underline{\mathbf{D}}^{-1} \mathbf{P} \underline{\mathbf{w}} \\ = [\mathbf{E} \tilde{\mathbf{w}}_0, \dots, \mathbf{E} \tilde{\mathbf{w}}_{M-1}]^T, \quad (20)$$

$$\underline{\mathbf{D}} = \text{blkdiag}(\mathbf{D}, \dots, \mathbf{D}) \in \mathbb{C}^{FM \times FM}, \quad (21)$$

$$\underline{\mathbf{E}} = \text{blkdiag}(\mathbf{E}, \dots, \mathbf{E}) \in \mathbb{R}^{(F/2-\delta)M \times FM}, \quad (22)$$

where  $\mathbf{P} \in \mathbb{R}^{FM \times FM}$  is a permutation matrix that changes the order of  $\underline{\mathbf{w}}$  from the frequency-wise one to the microphone-wise one. Its  $(f + Fm, m + fM)$ th entry is one for all  $f = 0, \dots, F-1$  and  $m = 0, \dots, M-1$ , and the other entries are zero. In (20),  $\mathbf{E} \in \mathbb{R}^{(F/2-\delta) \times F}$  extracts the non-causal components of a filter except for the allowed  $\delta$  samples.<sup>2</sup>

Let  $\mathbf{C} \in \mathbb{C}^{FM \times (F+(F/2-\delta)M)}$  be the horizontal concatenation of  $\mathbf{A}$  and  $\mathbf{B}$ , and  $\mathbf{v} \in \mathbb{C}^{F+(F/2-\delta)M}$  be the vertical concatenation of  $\mathbf{1}$  and  $\mathbf{0}$ . By considering the KKT conditions for the reformulated optimization problem in (17), we can obtain its solution by solving the following linear system [36]:

$$\begin{pmatrix} \underline{\mathbf{R}} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{w}}^* \\ \underline{\boldsymbol{\kappa}}^* \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{v} \end{pmatrix}, \quad (23)$$

where  $\underline{\mathbf{w}}^* \in \mathbb{C}^{FM}$  is the solution of the optimization problem in (17), and  $\underline{\boldsymbol{\kappa}}^* \in \mathbb{C}^{F+(F/2-\delta)M}$  is the KKT multiplier. In particular,  $\underline{\mathbf{w}}^*$  is obtained by

$$\underline{\mathbf{w}}^* = \underline{\mathbf{R}}^{-1} \mathbf{C} [\mathbf{C}^H \underline{\mathbf{R}}^{-1} \mathbf{C}]^{-1} \mathbf{v}. \quad (24)$$

As the main difference from the original MPDR beamformer in (9), the causal MPDR beamformer should be jointly optimized at all frequencies because the relationship between frequencies affects the causality.

### D. Relaxed Distortionless Constraint

The causality constraint in (14c) limits the DoF of a spatial filter. In particular, when  $\delta = 0$ , the constraint restricts the  $FM/2$  entries of the spatial filter. To increase the DoF and improve the extraction performance, we propose the causal MPrDR beamformer that relaxes the distortionless constraint regarding the gain of the spatial filter as follows:

$$\min_{(\mathbf{w}_f, z_f)_{f=0}^{F-1}} \sum_{f=0}^{F-1} \mathbf{w}_f^H \mathbf{R}_f \mathbf{w}_f \quad (25a)$$

$$\text{s.t. } \mathbf{w}_f^H \mathbf{a}_f = 1 + z_f \quad \forall f, \quad (25b)$$

<sup>2</sup>Different from (14c), the complex conjugation is not applied to the filters in (20) before the inverse DFT. To constrain the non-causal components of the complex conjugate of the filters,  $[w_{0,m}, \dots, w_{F-1,m}]^H$ , we assign the indices of the filter as  $[\tilde{w}_{0,m}, \tilde{w}_{-1,m}, \dots, \tilde{w}_{-F/2,m}, \tilde{w}_{F/2-1,m}, \dots, \tilde{w}_{1,m}]$  to  $\mathbf{D}^{-1}[w_{0,m}, \dots, w_{F-1,m}]^T$  in (20). Then, the multiplication of  $\mathbf{E}$  extracts the components whose index  $n$  is not in  $\mathcal{N}_\delta$ .

$$z_f \geq 0 \quad \forall f, \quad (25c)$$

$$(\mathbf{D}^{-1}[w_{0,m}, \dots, w_{F-1,m}]^H)_n = 0 \quad \forall m, n \notin \mathcal{N}_\delta, \quad (25d)$$

where  $z_f$  is a frequency-wise gain parameter. Since it is restricted to be a non-negative value, this relaxation of the distortionless constraint does not cause a delay in the target signal. The frequency-wise gain parameter can also be interpreted as an additional zero-phase filter. In (25), the gain of the spatial filter for the target direction varies at each frequency. Although the relaxed distortionless constraint potentially distorts the extracted signal, the gain parameter might not be extremely large. This is because  $z_f$  is optimized to be small to minimize the power of the extracted signal. In the case of  $\delta = F/2$ , we can omit the causality constraint in (25d) and solve the optimization problem in (25) separately for each frequency. When we obtain  $w_f$  that minimizes (25a) under the constraint in (25b) with a positive  $z_f$ , it is always possible to minimize (25a) by dividing  $w_f$  by  $1 + z_f$  that satisfies (25b) with  $z_f = 0$ . That is, the optimal gain parameter  $z_f$  becomes zero at all frequencies in the case without the causality constraint. Then, the solution of (25) coincides with the original MPDR beamformer.

The relaxed optimization problem is still a convex optimization problem because the first constraint in (25b) is still affine, and the non-negativity constraint in (25c) is convex. The spatial filter defined by (15) satisfies all the constraints in (25), and thus there exists a solution.

### E. DRS Method for Proposed Beamforming

The optimization problem of the causal MPrDR beamformer in (25) is difficult to solve in a closed-form. Hence, we propose to adopt the DRS method [23] that can handle multiple constraints and be easily extended to an online setting [37].

To use the DRS method, we first recast the variables and define two constraint sets. Let us concatenate the spatial filter  $w_f$  and the gain parameter  $z_f$ :

$$\zeta_f = [w_{f,0}, \dots, w_{f,M-1}, z_f]^T. \quad (26)$$

We extend the STFT of the observed signal and the RTF of the target signal as follows:

$$\mathbf{X}_{t,f} = [x_{t,f,0}, \dots, x_{t,f,M-1}, 0]^T, \quad (27)$$

$$\boldsymbol{\alpha}_f = [1, a_{f,1}, \dots, a_{f,M-1}, -1]^T. \quad (28)$$

The set of spatial filters that satisfy the relaxed distortionless constraint in (25b) is represented by

$$\mathcal{A}_f = \{\boldsymbol{\eta} \in \mathbb{C}^{M+1} \mid \boldsymbol{\eta}^H \boldsymbol{\alpha}_f = 1\}. \quad (29)$$

The set of filters that satisfy the causality constraint in (25d) is given by

$$\mathcal{C} = \{\boldsymbol{\xi} \in \mathbb{C}^F \mid (\mathbf{D}^{-1}[\xi_0, \dots, \xi_{F-1}]^H)_n = 0 \quad \forall n \notin \mathcal{N}_\delta\}. \quad (30)$$

On the basis of these notations, the optimization problem in (25) is reformulated as a minimization of the following sum of two functions with respect to  $\zeta_f$  defined in (26):

$$\mathcal{G}(\zeta_0, \dots, \zeta_{F-1}) + \mathcal{H}(\zeta_0, \dots, \zeta_{F-1}), \quad (31)$$

where the cost functions are defined as

$$\begin{aligned} \mathcal{G}(\zeta_0, \dots, \zeta_{F-1}) &= \sum_{f=0}^{F-1} \mathcal{G}_f(\zeta_f) \\ &= \sum_{f=0}^{F-1} \zeta_f^H \boldsymbol{\Gamma}_f \zeta_f + \iota_{\mathcal{A}_f}(\zeta_f), \end{aligned} \quad (32)$$

$$\begin{aligned} \mathcal{H}(\zeta_0, \dots, \zeta_{F-1}) &= \sum_{f=0}^{F-1} \iota_{\mathbb{R}_+}(z_f) + \sum_{m=0}^{M-1} \iota_{\mathcal{C}}([w_{0,m}, \dots, w_{F-1,m}]^T). \end{aligned} \quad (33)$$

In (32),  $\boldsymbol{\Gamma}_f$  is the extended SCM given by

$$\boldsymbol{\Gamma}_f = \frac{1}{T'} \sum_{t=T'_{\text{start}}}^{T'_{\text{end}}} \mathbf{X}_{t,f} \mathbf{X}_{t,f}^H, \quad (34)$$

and  $\iota_{\mathcal{S}}(\cdot)$  is the indicator function with respect to a set  $\mathcal{S}$ :

$$\iota_{\mathcal{S}}(x) = \begin{cases} 0 & (x \in \mathcal{S}) \\ \infty & (x \notin \mathcal{S}) \end{cases}. \quad (35)$$

The function  $\mathcal{G}(\cdot)$  is a sum of frequency-wise functions  $\mathcal{G}_f(\cdot)$  and covers the cost function in (25a) and the relaxed distortionless constraint in (25b). The other function  $\mathcal{H}(\cdot)$  consists of the indicator functions related to the constraints in (25c)–(25d).

The sum of two functions in (31) can be minimized by the following iterative procedure of the DRS method [23]:

$$(\zeta_0, \dots, \zeta_{F-1}) \leftarrow \text{prox}_{\mathcal{H}/\rho}(\phi_0, \dots, \phi_{F-1}), \quad (36)$$

$$\varphi_f \leftarrow 2\zeta_f - \phi_f \quad \forall f, \quad (37)$$

$$(\psi_0, \dots, \psi_{F-1}) \leftarrow \text{prox}_{\mathcal{G}/\rho}(\varphi_0, \dots, \varphi_{F-1}), \quad (38)$$

$$\phi_f \leftarrow \phi_f + \lambda(\psi_f - \zeta_f) \quad \forall f, \quad (39)$$

where  $\rho \in \mathbb{R}_+$  and  $\lambda \in (0, 2)$  are hyperparameters. Here,  $\phi_f$ ,  $\varphi_f$ , and  $\psi_f$  are auxiliary variables whose sizes are the same as  $\zeta_f$ . In (36) and (38), the functions  $\mathcal{H}(\cdot)$  and  $\mathcal{G}(\cdot)$  are separately handled via the proximity operators [38]:

$$\text{prox}_{\mathcal{G}/\rho}(\mathbf{g}) = \arg \min_{\mathbf{h}} (\mathcal{G}(\mathbf{h}) + \rho \|\mathbf{h} - \mathbf{g}\|_2^2), \quad (40)$$

$$\text{prox}_{\mathcal{H}/\rho}(\mathbf{g}) = \arg \min_{\mathbf{h}} (\mathcal{H}(\mathbf{h}) + \rho \|\mathbf{h} - \mathbf{g}\|_2^2), \quad (41)$$

where  $\mathbf{g} \in \mathbb{C}^{F(M+1)}$  is the concatenation of the inputs of each function,  $\mathbf{h}$  is a dummy variable whose size is the same as  $\mathbf{g}$ , and  $\|\cdot\|_2$  is the  $\ell_2$  norm. The proximity operators of  $\mathcal{G}(\cdot)$  and  $\mathcal{H}(\cdot)$  are explained in Sections III-E1 and III-E2, respectively.

1) *Proximity Operator of  $\mathcal{G}(\cdot)$* : The proximity operator of  $\mathcal{G}(\cdot)$  can be computed frequency-wise:

$$\begin{aligned} &\text{prox}_{\mathcal{G}/\rho}(\varphi_0, \dots, \varphi_{F-1}) \\ &= \left[ \text{prox}_{\mathcal{G}_0/\rho}(\varphi_0), \dots, \text{prox}_{\mathcal{G}_{F-1}/\rho}(\varphi_{F-1}) \right]^T, \end{aligned} \quad (42)$$

because  $\mathcal{G}(\cdot)$  is separable for each frequency [38]. As  $\mathcal{A}_f$  corresponds to the relaxed distortionless constraint in (25b),

$\text{prox}_{\mathcal{G}_f/\rho}(\varphi_f)$  is the solution of the following equality constrained least squares problem:

$$\min_{\boldsymbol{\eta}} \boldsymbol{\eta}^H \boldsymbol{\Gamma}_f \boldsymbol{\eta} + \rho \|\boldsymbol{\eta} - \varphi_f\|_2^2 \quad (43a)$$

$$\text{s.t. } \boldsymbol{\eta}^H \boldsymbol{\alpha}_f = 1. \quad (43b)$$

Considering the KKT conditions, we obtain the following linear system:

$$\begin{pmatrix} \boldsymbol{\Gamma}_f + \rho \mathbf{I} & \boldsymbol{\alpha}_f \\ \boldsymbol{\alpha}_f^H & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\eta}^* \\ \sigma^* \end{pmatrix} = \begin{pmatrix} \rho \varphi_f \\ 1 \end{pmatrix}, \quad (44)$$

where  $\mathbf{I} \in \mathbb{C}^{(M+1) \times (M+1)}$  is the identity matrix,  $\boldsymbol{\eta}^* \in \mathbb{C}^{M+1}$  is the solution of the optimization problem in (43), and  $\sigma^* \in \mathbb{C}$  is the KKT multiplier. As the KKT system is nonsingular, the solution can be computed in a closed-form.

2) *Proximity Operator of  $\mathcal{H}(\cdot)$* : The second cost function  $\mathcal{H}(\cdot)$  is separable across the gain parameter for each frequency  $z_f$  and the spatial filter for each microphone  $[w_{0,m}, \dots, w_{F-1,m}]^T$ . Hence, we can evaluate the proximity operator of  $\mathcal{H}(\cdot)$  by the proximity operator for each indicator function [38]. The proximity operator for  $\iota_{\mathbb{R}_+}(\cdot)$  is given by

$$\text{prox}_{\iota_{\mathbb{R}_+}}(\varsigma) = \max(\text{Re}[\varsigma], 0), \quad (45)$$

where  $\varsigma \in \mathbb{C}$  is a dummy variable, and  $\text{Re}[\cdot]$  returns the real part of its input. Then, as  $\mathbf{D}$  is unitary,  $\text{prox}_{\iota_{\mathbb{C}}}(\cdot)$  can be evaluated as [38]

$$\overline{\text{prox}_{\iota_{\mathbb{C}}}(\boldsymbol{\varrho})} = \mathbf{D} \text{prox}_{\iota_{\mathbb{C}}}(D^{-1} \overline{\boldsymbol{\varrho}}), \quad (46)$$

where  $\boldsymbol{\varrho} \in \mathbb{C}^F$  is a dummy variable,  $\overline{(\cdot)}$  denotes the complex conjugation, and  $\text{prox}_{\iota_{\mathbb{C}}}(\cdot)$  is the projection onto the set of quasi-causal filters in the time domain:

$$\text{prox}_{\iota_{\mathbb{C}}}(\tilde{\boldsymbol{w}})_n = \begin{cases} \tilde{w}_n & (n \in \mathcal{N}_\delta) \\ 0 & (n \notin \mathcal{N}_\delta) \end{cases}. \quad (47)$$

This procedure is the same as the truncation of the unallowable non-causal components in (13). In [14], the truncation is presented heuristically and performed as a postprocessing. On the other hand, the proposed method iteratively applies the truncation in the optimization algorithm.

3) *Summary of Proposed Algorithm*: The DRS method for the causal MPrDR beamformer in (25) is summarized in Algorithm 1, where  $k = 0, \dots, K-1$  is the iteration index. Here,  $\theta_{f,m}$  and  $\vartheta_f$  are auxiliary variables related to  $w_{f,m}$  and  $z_f$ , respectively. This algorithm is designed for (25), but we can adopt it to the causal MPDR beamformer in (14) by applying  $\vartheta_f \leftarrow 0$  instead of  $\vartheta_f \leftarrow \text{prox}_{\iota_{\mathbb{R}_+}}(z_f)$ . Through our experiments, we set the initial value of  $\phi_f$  to zero.

In Algorithm 1, we update  $\vartheta_f$ ,  $\varphi_f$ ,  $\psi_f$ , and  $\phi_f$  in parallel for each frequency. In the update of  $\psi_f$ ,  $\text{prox}_{\mathcal{G}_f/\rho}(\cdot)$  costs  $\mathcal{O}(M^2)$  where we assume that the inverse of the KKT matrix in (44) is computed in advance. On the other hand,  $[\theta_{0,m}, \dots, \theta_{F-1,m}]^T$  is updated by  $\text{prox}_{\iota_{\mathbb{C}}}(\cdot)$  in parallel for each microphone. Its computational cost can be reduced to  $\mathcal{O}(F \log F)$  by using the fast Fourier transform. Hence, the computational cost of Algorithm 1 per iteration is the maximum of  $\mathcal{O}(FM^2)$  and  $\mathcal{O}(FM \log F)$ . On the other hand, the closed-form method (24)

---

**Algorithm 1: DRS Method for Causal MPrDR Beamformer.**


---

**Input:**  $w_f, z_f, \mathbf{R}_f, \mathbf{a}_f, \phi_f, \rho, \lambda$

**Output:**  $D^{-1}[\theta_{0,m}, \dots, \theta_{F-1,m}]^H \forall m$

**for**  $k = 0, \dots, K-1$  **do**

**for**  $m = 0, \dots, M-1$  **do**

$[\theta_{0,m}, \dots, \theta_{F-1,m}]^T \leftarrow \text{prox}_{\iota_{\mathbb{C}}}([w_{0,m}, \dots, w_{F-1,m}]^T)$

**end for**

**for**  $f = 0, \dots, F-1$  **do**

$\vartheta_f \leftarrow \text{prox}_{\iota_{\mathbb{R}_+}}(z_f)$

**end for**

  Concatenate  $\theta_{f,0}, \dots, \theta_{f,M-1}$  and  $\vartheta_f$  as  $\zeta_f$

**for**  $f = 0, \dots, F-1$  **do**

$\varphi_f \leftarrow 2\zeta_f - \phi_f$

$\psi_f = \text{prox}_{\mathcal{G}_f/\rho}(\varphi_f)$

$\phi_f \leftarrow \phi_f + \lambda(\psi_f - \zeta_f)$

**end for**

  Split  $\phi_f$  into  $w_{f,m}$  and  $z_f$

**end for**

---

costs  $\mathcal{O}(F^3 M^3)$ . We thus expect that Algorithm 1 is faster than the closed-form method even with a large number of iterations.<sup>3</sup>

#### F. Advantage of Causality Constraint

We extract the target signal by filter-and-sum in (12) by using the quasi-causal spatial filter obtained through Algorithm 1. The algorithmic delay for target source extraction is reduced to only  $\delta$  samples owing to the causality constraint, and the filter-and-sum does not cause a phase delay of the target signal due to the exact or relaxed distortionless constraints. Meanwhile, the interference signal has a phase delay and a frequency-dependent group delay because the distortionless constraint holds on only the target signal.

In the original MPDR beamformer, the cost function in (7a) is the power of the signal extracted by the T-F-domain beamforming in (6). It uses the circular convolution and does not coincide with the filter-and-sum using the linear convolution in (12). Hence, the original MPDR beamformer in (9) is not optimal for the time-domain filter-and-sum. Intriguingly, the causality constraint reduces this mismatch and completely removes it when  $\delta = 0$ . In detail, we assume that the window length  $N$  is equal to or less than half of the number of DFT points  $F$ . In addition, the pair of STFT and iSTFT is assumed to perfectly reconstruct a signal without windowing in iSTFT. Under these assumptions, the iSTFT of the result of the T-F-domain beamforming coincides with the output of the time-domain filter-and-sum. Proof of this property is presented in Appendix A.

For instance, the above assumptions are satisfied by using the Hann window, and setting the window shift and the number of DFT points to half and twice the window size, respectively.

<sup>3</sup>Before the iterative procedure of the DRS method, the inverse of the KKT matrix in (23) should be computed for all frequencies, which costs  $\mathcal{O}(FM^3)$ . It is, however, neglectable by assuming  $K \gg M$ .

As a result, the causality constraint reduces the discrepancy between beamforming in the time and T-F domains, which is advantageous for time-domain target source extraction. We will show this effectiveness of the causality constraint through an experiment in Section V-B.

#### IV. ONLINE CAUSAL MPRDR BEAMFORMING

In this section, we propose a frame-level online causal MPRDR beamforming. An online extension of Algorithm 1 is derived based on the adaptive DRS (ADRS) method [37].

##### A. Problem Formulation for Online Target Source Extraction

When the causal MPRDR beamformer is applied to low-latency real-time applications, the spatial filter should be updated in a block online manner. In detail, the spatial filter is updated by solving the optimization problem in (25) for every  $T$  frames where  $T'_{\text{start}}$  and  $T'_{\text{end}}$  in (8) corresponds to the start and end of the previous block. We extract the target signal from the subsequent observed signals by using the obtained spatial filter. This implementation, however, is not optimal for the situation with interference sources moving within  $T$  frames. To adapt to such situations, we derive a frame-by-frame online algorithm for the causal MPRDR beamformer.

In the online setting, the STFT coefficients  $\mathbf{x}_{t,f}$  are given sequentially. We recursively update the SCM at each time frame  $\mathbf{R}_{t,f}$  as follows [39], [40]:

$$\mathbf{R}_{t,f} = \beta \mathbf{R}_{t-1,f} + (1 - \beta) \mathbf{x}_{t,f} \mathbf{x}_{t,f}^H, \quad (48)$$

where  $\beta \in [0, 1)$  is a forgetting factor. Based on this time-varying SCM, the spatial filter for each time frame  $\mathbf{w}_{t,f}$  is updated to solve the following optimization problem:

$$\min_{(\mathbf{w}_f, z_f)_{f=0}^{F-1}} \sum_{f=0}^{F-1} \mathbf{w}_f^H \mathbf{R}_{t,f} \mathbf{w}_f \quad (49a)$$

$$\text{s.t.} \quad \mathbf{w}_f^H \mathbf{a}_f = 1 + z_f \quad \forall f, \quad (49b)$$

$$z_f \in \mathbb{R}_+ \quad \forall f, \quad (49c)$$

$$(\mathbf{D}^{-1} [w_{0,m}, \dots, w_{F-1,m}]^H)_n = 0 \quad \forall m, n \notin \mathcal{N}_\delta, \quad (49d)$$

where the constraints are the same as in the offline version in (25) because we assume that the target source does not move. The online formulation in (49) aims to minimize a time-varying cost function under the time-invariant constraints.

##### B. ADRS Method for Causal MPRDR Beamforming

To minimize the time-varying cost function in (49a), we propose to apply the ADRS method [37] that has achieved promising results in echo cancellation [37] and active noise control [41]. The ADRS method for (49) is summarized in Algorithm 2. In this online algorithm, we update the optimization variables  $K$  times for each time frame. Then,  $w_{f,m}$  and  $z_f$  at the last iteration for the  $t$ th time frame are used as the initial values for the  $(t + 1)$ th

---

#### Algorithm 2: ADRS Method for Causal MPRDR Beamformer.

---

**Input:**  $w_{f,m}, z_f, \mathbf{R}_{t,f}, \phi_f, \rho, \lambda$   
**Output:**  $\mathbf{D}^{-1} [\theta_{0,m}^{[t]}, \dots, \theta_{F-1,m}^{[t]}]^H \quad \forall t, m$   
**for**  $t = 0, 1, \dots$  **do**  
  **for**  $k = 0, \dots, K - 1$  **do**  
    **for**  $m = 1, \dots, M$  **do**  
       $[\theta_{0,m}, \dots, \theta_{F-1,m}]^T \leftarrow \text{prox}_{\ell_c}([w_{0,m}, \dots, w_{F-1,m}]^T)$   
    **end for**  
    **for**  $f = 1, \dots, F$  **do**  
       $\vartheta_f \leftarrow \text{prox}_{\ell_{\mathbb{R}_+}}(z_f)$   
    **end for**  
    Concatenate  $\theta_{f,0}, \dots, \theta_{f,M-1}$  and  $\vartheta_f$  as  $\zeta_f$   
    **for**  $f = 0, \dots, F - 1$  **do**  
       $\varphi_f \leftarrow 2\zeta_f - \phi_f$   
       $\psi_f \leftarrow \text{prox}_{\mathcal{G}_{t,f}/\rho}(\varphi_f)$   
       $\phi_f \leftarrow \phi_f + \lambda(\psi_f - \zeta_f)$   
    **end for**  
    Split  $\phi_f$  into  $w_{f,m}$  and  $z_f$   
  **end for**  
   $\theta_{f,m}^{[t]} \leftarrow \theta_{f,m}$   
**end for**

---

time frame. This warm starting strategy is effective when the SCM varies smoothly in successive time frames. The difference between Algorithms 1 and 2 is replacing  $\text{prox}_{\mathcal{G}_f/\rho}(\cdot)$  by the time-varying proximity operator  $\text{prox}_{\mathcal{G}_{t,f}/\rho}(\cdot)$ . It is computed in (44) by replacing  $\Gamma_f$  by  $\Gamma_{t,f}$ :

$$\Gamma_{t,f} = \beta \Gamma_{t-1,f} + (1 - \beta) \chi_{t,f} \chi_{t,f}^H. \quad (50)$$

The updated spatial filter  $\theta_{f,m}^{[t]}$  is converted to the time domain and used to extract the target signal from the subsequent  $r$  samples of the observed mixture.

In the online version of the original MPDR beamforming [32], the spatial filter has been computed by (9) on the basis of the time-varying SCM updated by (48). That is, the spatial filter is independent at each time frame. In Algorithm 2, the spatial filter update in (43) depends on the spatial filter in the previous iteration similar to online algorithms for the generalized side-lobe canceler [42]. As a result, the spatial filter varies smoothly along time frames even when SCM changes rapidly. We expect that this property of Algorithm 2 stabilizes the online causal MPRDR beamformer.

In the online setting, we should compute the inverse of the KKT matrix for each time frame, which costs  $\mathcal{O}(FM^3)$ . To the best of the authors' knowledge, techniques for the efficient update of the inverse matrix, including the Sherman–Morrison formula [43], are not applicable to Algorithm 2. The reduction of computational cost for the matrix inverse with or without approximation is a direction of future works.

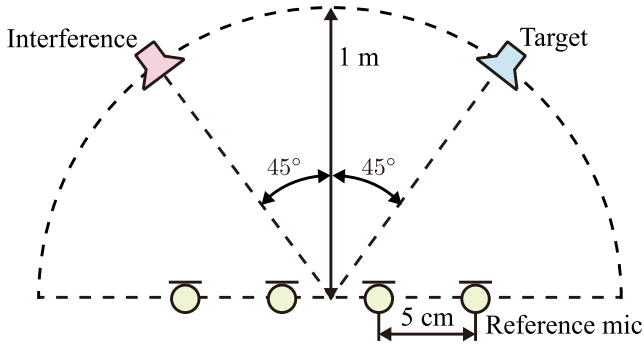


Fig. 3. Spatial arrangements of sound sources and a linear microphone array.

## V. EXPERIMENTAL EVALUATION OF CAUSAL MPDR AND MPRDR BEAMFORMER

To confirm the effectiveness of the proposed methods, we conducted several experiments in the offline setting. In Section V-A, we investigate the relation between the extraction performance and the number of iterations with various tap lengths. Section V-B demonstrates the advantage of the causality constraint. The proposed methods were compared with various low-latency beamforming methods in Section V-C.

### A. Investigation of Causal MPDR Beamformer

1) *Experimental Conditions:* In this experiment, we investigated the performance of the proposed methods with synthesized four-channel audio mixtures. Source signals of 10 s were generated by concatenating utterances of the same speakers in the Voice Conversion Challenge (VCC) 2018 dataset [44]. They were downsampled at 16 kHz. A four-channel linear microphone array was placed at the center of the room of 6.0 m × 5.0 m × 4.0 m. The target and interference speakers were around the microphone array as illustrated Fig. 3. The reverberation time was randomly sampled from [0.16, 0.32] s. Room impulse responses were synthesized by using `pyroomacoustics` toolbox [45] and convolved with the source signals.

The RTF of the target source was computed by applying the eigenvalue decomposition to the SCM of another clean source image uttered by the same speaker from the same position. STFT was implemented with the half-overlapping Hann window. The number of DFT points was in {2048, 4096, 8192, 16384}, and the window size was half of it. The causal MPDR beamformer in (14) and the causal MPrDR beamformer in (25) are abbreviated as `Prop-exact` and `Prop-relax`, respectively. The number of non-causal components  $\delta$  was set to 0, i.e., the proposed methods were causal. Algorithm 1 for `Prop-exact` and `Prop-relax` was iterated  $10^5$  times, where  $\lambda$  was set to 1.8 for faster convergence. We also computed the causal MPDR beamformer in the closed-form<sup>4</sup> given in (24). In this experiment, the SCM was computed from the entire signal.

<sup>4</sup>The closed-form method was evaluated only when the number of DFT points was 2048 and 4096 due to the limitation of computational complexity.

The extraction performance was evaluated by the scale-invariant source-to-distortion ratio (SI-SDR) [46], the wide-band extension of the perceptual evaluation of subjective quality (PESQ) [47], and the extended short-time objective intelligibility (ESTOI) [48]. To investigate the amount of distortion, we define a distortion ratio (DR) as follows:

$$\text{DR}(w_0, \dots, w_{F-1}) = 10 \log_{10} \left( \frac{F}{\sum_{f=0}^{F-1} |1 - \mathbf{a}_f^H \mathbf{w}_f|^2} \right). \quad (51)$$

This measure becomes infinity when the distortionless constraint is completely satisfied.

2) *Relation Between Extraction Performances and Number of Iterations:* SI-SDR, PESQ, ESTOI, and DR averaged over 10 audio mixtures are illustrated in Fig. 4. The performance of the proposed methods depended on the hyperparameter of the DRS method  $\rho$ . SI-SDR and PESQ for `Prop-exact` with Algorithm 1 converged faster to those for the closed-form method when  $\rho = 0.005$ . In the following experiments, we thus set  $\rho$  to 0.005 for both causal MPDR and MPrDR beamformers. Although both `Prop-exact` and `Prop-relax` require many iterations for the convergence, their SI-SDR quickly improved as illustrated in the top row of Fig. 4. Especially, `Prop-relax` achieved SI-SDR of 10 dB within 10 iterations and with the DFT points of 2048. This setting is reasonable for real-time applications with limited computational resources.

SI-SDR, PESQ, and ESTOI for `Prop-exact` decreased after a large number of iterations. This is because, in the DRS method, the causal spatial filters first intensively minimized the cost function while not satisfying the distortionless constraint. As the number of iterations increased, these filters satisfied the distortionless constraint more accurately, which resulted in lower extraction performance. On the other hand, `Prop-relax` kept high extraction performance even after a large number of iterations thanks to the relaxed distortionless constraint. The effect of the relaxation can also be confirmed in DR depicted in the bottom row of Fig. 4. `Prop-exact` improved DR along with the increase of the number of iterations and achieved DR of higher than 60 dB. That is, the distortionless constraint was substantially satisfied after a sufficient number of iterations. On the other hand, `Prop-relax` did not improve DR because we relax the distortionless constraint as in (25b)–(25c). We emphasize that `Prop-relax` does not cause a phase delay on the target signal because  $\mathbf{w}_f^H \mathbf{a}_f$  in (49b) takes a non-negative real value. To demonstrate this point, we show the normalized cross-correlation between the target signal and the extracted signal in Fig. 5. Since the peak appears at a lag of zero, no delay occurs by `Prop-relax`. High SI-SDR of `Prop-relax` also indicates no delay on the extracted signal. This is because SI-SDR only compensates for the scale mismatch between the target and extracted signals.

In terms of SI-SDR, PESQ, and ESTOI, `Prop-exact` achieved its best performance when the number of DFT points  $F$  was 16384. Meanwhile, the performance of `Original` was saturated when  $F = 8192$ . This is because `Prop-exact` with  $\delta = 0$  halve the actual filter length and has less DoF than `Original`. `Prop-relax` outperformed `Original` in



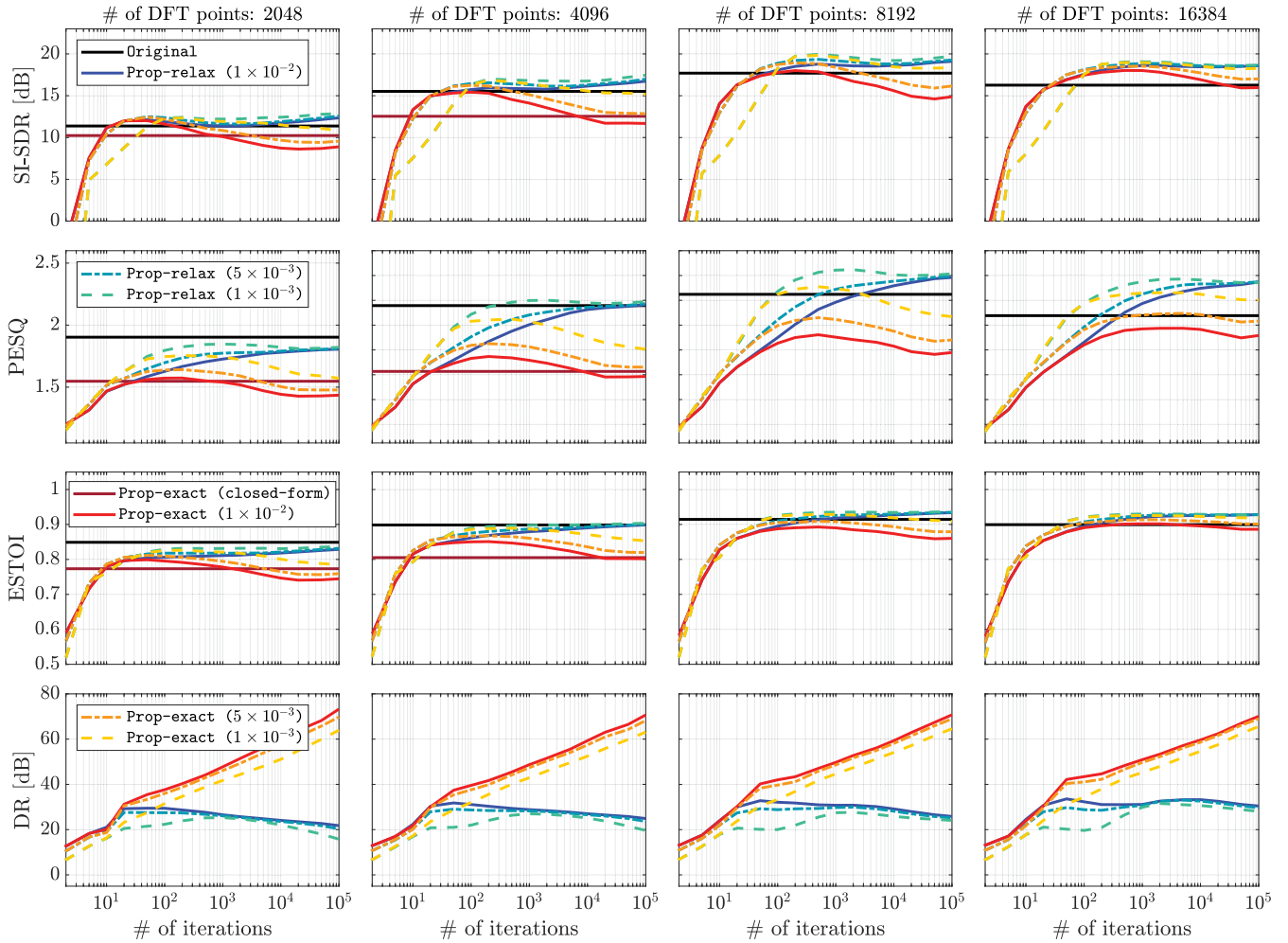


Fig. 4. Evolution of the average SI-SDR, PESQ, ESTOI, and DR with respect to the number of iterations of Algorithm 1. The black line (Original) is for the MPDR beamformer in (9). The dark red line [Prop-exact (closed-form)] corresponds to the causal MPDR beamformer given by (24). Their DRs become infinity theoretically and are omitted. In the other methods, the hyperparameter in Algorithm 1,  $\rho$ , is shown in parentheses.

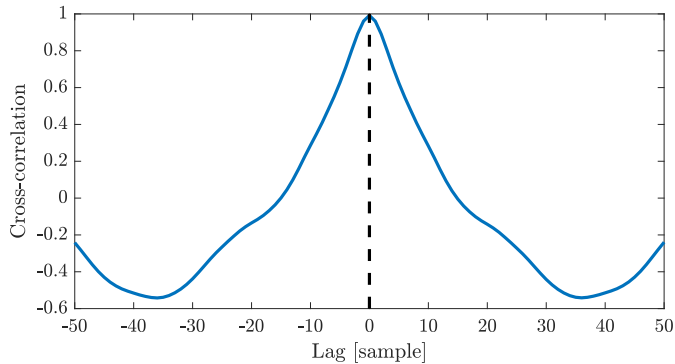


Fig. 5. Cross-correlation between the target source image at the reference microphone and the extracted signal. Black dotted line shows a lag of zero.

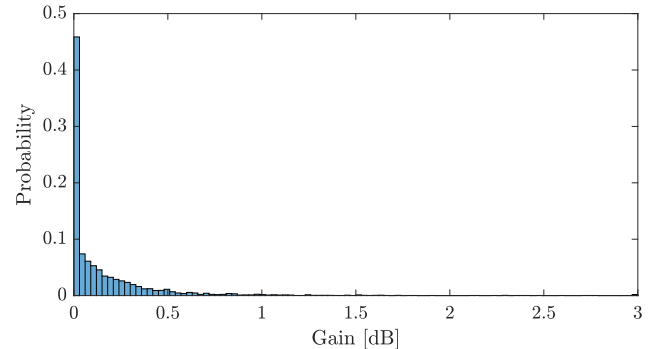


Fig. 6. Histogram of the filter gains with respect to the RTF of the target signal, i.e.,  $20 \log_{10}(1 + z_f)$ . For visibility, we limit the range up to 3 dB, and all gains above 3 dB are accumulated in the right-most bin. The maximum of  $20 \log_{10}(1 + z_f)$  was 13 dB as an outlier.

terms of SI-SDR while reducing the algorithmic delay. We will investigate this phenomenon more in Section V-B.

3) *Gain Parameter and Directivity Patterns*: To further investigate the causal MPDR beamformer, we show a histogram

of the filter gain  $20 \log_{10}(1 + z_f)$  for Prop-relax in Fig. 6. The gain was less than 2 dB in most of the frequencies. Examples of the directivities of the spatial filters are shown in Fig. 7.

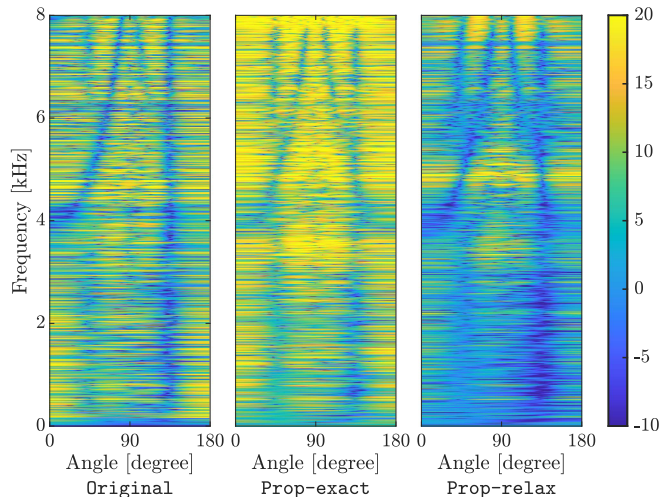


Fig. 7. Directivity patterns of the spatial filters. The target and interference sources are in  $45^\circ$  and  $135^\circ$ , respectively. The gain for the target signal is theoretically 0 dB in *Original*. It is close to 0 dB in *Prop-exact* because DR becomes high after a sufficient number of iterations, while it is equal to or greater than 0 dB in *Prop-relax*.

Compared with *Prop-exact*, *Prop-relax* had a deeper null for the direction of the interference source ( $135^\circ$ ). That is, the relaxation of the distortionless constraint is effective to suppress the interference signal. These results supported the high objective measures of *Prop-relax*.

### B. Effect of Causality Constraint

To demonstrate the advantage of the causality constraint, we evaluated the performance of the causal MPDR and MPrDR beamformers with different numbers of non-causal components. Audio mixtures were simulated as in the previous experiment. The half-overlapping Hann window was used in STFT where the number of DFT points and window length were 8192 and 4096, respectively. For both *Prop-exact* and *Prop-relax*, Algorithm 1 was iterated  $10^4$  times.

The cost function in (7a) was shown on the left side of Fig. 8. The cost functions for the proposed methods monotonically decreased as  $\delta$  increased. *Original*, which corresponds to the causal MPDR beamformer with  $\delta = 4096$ , achieved the minimum cost function. This is because larger  $\delta$  increases the DoF of the spatial filter. Meanwhile, as illustrated on the right side of Fig. 8, *Prop-exact* and *Prop-relax* achieved their best SI-SDR when the algorithmic delay was 32 ms and 6 ms, respectively. As discussed in Section III-F, beamforming in the T-F domain coincides with filter-and-sum in the time domain when  $\delta = 0$ . As  $\delta$  increases, the discrepancy arises between the T-F domain beamforming used in the cost function and the filter-and-sum for the low-latency extraction. That is, there is a trade-off between the DoF of the spatial filter and the mismatch between the two beamforming implementations. Hence, the proposed methods with a limited tap length can outperform the original MPDR beamformer.

Comparing the proposed methods, *Prop-exact* preferred larger  $\delta$  than *Prop-relax* according to the right side of Fig. 8.

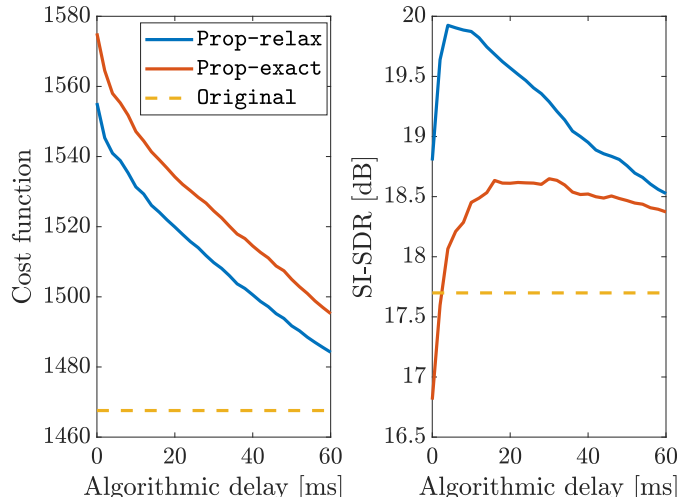


Fig. 8. Relation between algorithmic delay, cost function, and SI-SDR.

TABLE I  
STFT CONDITION FOR EACH METHOD

	Shift size	Window size	DFT points	Low-latency
Long	2048	4096	8192	No
Short	48	96	96	Yes
WPD	48	96	96	Yes
Asym	48	4096 / 96	8192	Yes
Trunc	2048	4096	8192	Yes
<i>Prop-exact</i>	2048	4096	8192	Yes
<i>Prop-relax</i>	2048	4096	8192	Yes

In *Asym*, the lengths of the windows in STFT and iSTFT are different.

This should be because *Prop-exact* has less DoF of the spatial filter than *Prop-relax* due to the exact distortionless constraint.

### C. Comparison With Existing Low-Latency Methods

We compared the proposed methods with two T-F-domain and one time-domain methods for low-latency beamforming. One method is the MPDR beamforming with asymmetric windows (*Asym*), which reduces the algorithmic delay by using a short window in iSTFT. The asymmetric Hann window formulated in [20], [21] was used. Another T-F domain method is the weighted power minimization distortionless response (WPD) beamformer [40], [49], which is abbreviated as *WPD*. This method assumes the convolutive mixing process in the T-F domain, which allows us to use a short window in both STFT and iSTFT. The prediction delay and the tap size were tuned to 5 and 5, respectively. The time-domain method truncates the spatial filter as in (13) and performs filter-and-sum as in (12), which is abbreviated as *Trunc* [14]. We also evaluated the original MPDR beamforming with a long window (*Long*) and a short window (*Short*).

To keep the algorithmic delay 6 ms, the STFT conditions were different in each method as summarized in Table I. Although the STFT conditions of the time-domain methods were the same to *Long*, the algorithmic delay was reduced by truncating non-causal components where  $\delta$  was set to 96.

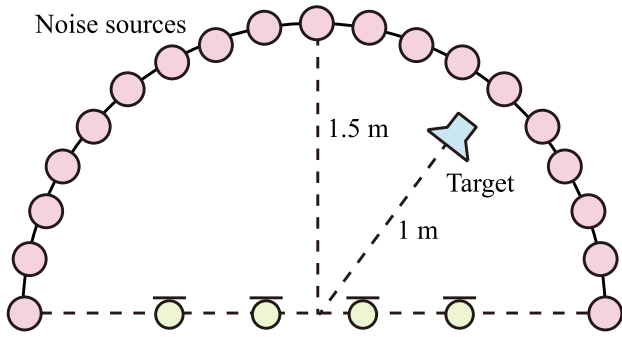


Fig. 9. Spatial arrangements for diffuse noise suppression.

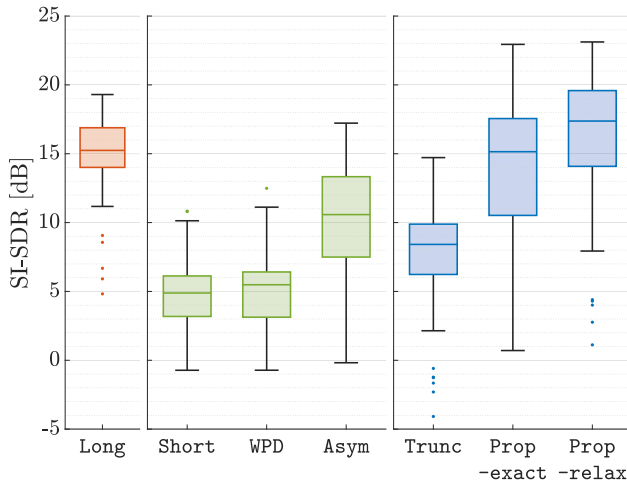


Fig. 10. Comparison of SI-SDR of various low-latency beamforming methods in target source extraction. The algorithmic delay was reduced to 6 ms except for the MPDR beamforming with a long window (Long) depicted by the red box. The T-F-domain and time-domain methods are indicated by green and blue boxes, respectively.

In contrast to the experiments in Sections V-A and V-B, we simulated 50 audio mixtures under two conditions. Under the first condition, the directions of the target and interference sources were randomly sampled, where the minimum difference between the directions was set to 5 degrees. Other settings were the same as in previous experiments. The second condition is for evaluating the performance of diffuse noise suppression. We simulated the diffuse noise by using the CHiME3 noise [50], where noise sources were equally spaced on 19 positions [51] as illustrated in Fig. 9. The target source was located at a random direction. Audio samples for both conditions are available online.<sup>5</sup>

The boxplot of SI-SDR under the first condition is illustrated in Fig. 10. By Comparing Long and Short, the performance of the MPDR beamforming significantly decreased with the short window. The existing methods improved the median SI-SDR from Short. In the T-F-domain methods, Asym performed best by using a long window in STFT. Prop-relax achieved better SI-SDR than Prop-exact thanks to relaxing the distortionless

<sup>5</sup>[Online]. Available: <https://sites.google.com/view/yoshiki-masuyama/causalmprdr>

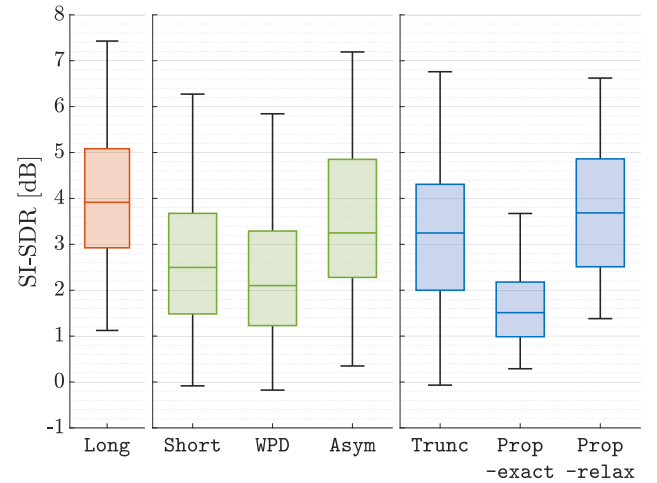


Fig. 11. Comparison of SI-SDR of various low-latency beamforming methods in diffuse noise suppression.

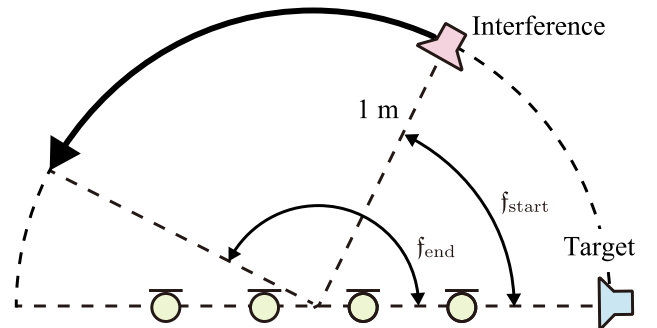


Fig. 12. Arrangement of the fixed target and moving interference sources.

constraint, and it outperformed the existing low-latency methods by a large margin.

The boxplot of SI-SDR under the second condition is illustrated in Fig. 11. The overall performance decreased under diffuse noise because the MPDR beamforming is dedicated to suppressing directional sources. Remarkably, Prop-relax was still comparable to Long while its algorithmic delay is only 6 ms. Meanwhile, Prop-exact resulted in relatively worse performance, which could be because the spatial filter that satisfies the two constraints lacks DoF for suppression.

## VI. EXPERIMENTAL EVALUATION OF ONLINE CAUSAL MPDR AND MPRDR BEAMFORMERS

In this section, we investigate the performance of the proposed online methods under a non-stationary condition.

### A. Experimental Conditions

We synthesized four-channel audio mixtures of 30 s with a moving interference source by using `gpurir` toolbox [52]. Fig. 12 shows the simulated situation where the interference source moved away from the target source. The distances from the microphone array to the target and interference sources were

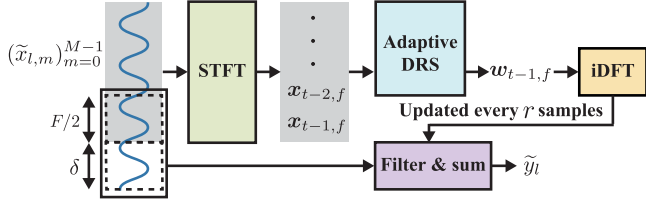


Fig. 13. Block diagram of the online causal MPDR and MPrDR beamformers. The spatial filters  $w_{t-1,f}$  are converted to the time domain and used to extract the target source signal per sample by filter-and-sum. It requires the  $F/2$  causal and  $\delta$  non-causal samples. The time-domain filters are updated in every  $r$  samples where  $r$  is the window shift in STFT.

1.0 m. The room has dimensions 6.0 m  $\times$  5.0 m  $\times$  4.0 m, and the reverberation time was randomly sampled from  $[0.16, 0.32]$ . An array is located at the center of the room, and the spacing between microphones was 5 cm.

We compared the proposed online methods (Prop-exact and Prop-relax) with the online version of the original MPDR beamformer (Original) and its truncated version (Trunc). The online MPDR beamformer was computed by replacing the time-invariant SCM in (9) with the time-varying one in (48). Hence, the spatial filter at the  $t$ th time frame  $w_{t,f}$  depends on  $x_{t,f}$ . In contrast, Trunc used the truncated version of  $w_{t-1,f}$  to extract the target signal at the subsequent  $r$  samples by the filter-and-sum in (12). The proposed methods also used the spatial filter computed from the previous time frame to extract the following target source signal as depicted in Fig. 13. Thanks to this, the total algorithmic delay of Trunc, Prop-exact, and Prop-relax is only 96 samples, while that of Original is 4096 samples. We changed the forgetting factor  $\beta$  in (48) to show the relationship between  $\beta$  and the extraction performance. A larger  $\beta$  maintains the SCM at the previous time frame more, and thus the SCM varies more smoothly. We set the number of non-causal components  $\delta$  to 96 and the number of iterations per time frame  $K$  in Algorithm 2 to 5 and 10.

### B. Experimental Results Under Non-Stationary Condition

Table II shows the SI-SDR of the extracted signals with various forgetting factors. We investigated the performance under two conditions where the end esdirection of the interference source  $f_{\text{end}}$  was different. On the contrary to the stationary condition, Original outperformed Prop-relax. This should be because it requires more DoF of the spatial filter to suppress the moving interference source. We stress that Prop-relax performed best in the low-latency methods. This result confirms the effectiveness of the online causal MPrDR beamformer for real-time applications.

Regardless of the conditions, i.e., the speed of the interference source, the performance of Original has a sharp peak at  $\beta = 0.97$ . On the other hand, Prop-exact and Prop-relax worked well even with smaller  $\beta$ . As discussed in Section IV-B, the proximal update in Algorithm 2 minimizes the time-varying cost function around the spatial filter at the previous iteration. This procedure smooths the time-varying spatial filter and results

TABLE II  
AVERAGE SI-SDR FOR ONLINE METHODS

Method	Forgetting factor $\beta$						
	0.30	0.50	0.70	0.90	0.95	0.97	0.99
Condition 1 ( $f_{\text{start}}, f_{\text{end}} = (60^\circ, 150^\circ)$ )							
Original	-1.26	-0.05	1.61	4.75	6.1	<b>6.39</b>	5.09
Trunc	1.06	2.14	3.61	5.55	<b>5.73</b>	5.38	3.98
Prop-exact (5)	2.94	2.98	3.08	<b>3.31</b>	3.29	3.12	2.42
Prop-exact (10)	2.39	2.48	2.66	3.05	<b>3.13</b>	3.00	2.25
Prop-relax (5)	5.68	5.72	<b>5.77</b>	5.62	5.23	4.82	3.72
Prop-relax (10)	5.76	5.84	<b>5.94</b>	5.82	5.42	4.99	3.80
Condition 2 ( $f_{\text{start}}, f_{\text{end}} = (60^\circ, 300^\circ)$ )							
Original	-1.28	-0.04	1.64	4.82	6.28	<b>6.79</b>	5.76
Trunc	1.18	2.34	3.82	5.66	<b>5.89</b>	5.74	4.58
Prop-exact (5)	3.30	3.33	3.43	<b>3.65</b>	3.62	3.46	2.84
Prop-exact (10)	2.61	2.72	2.92	3.38	<b>3.46</b>	3.33	2.65
Prop-relax (5)	5.73	5.80	<b>5.89</b>	5.86	5.6	5.31	4.34
Prop-relax (10)	5.71	5.84	5.99	<b>6.00</b>	5.75	5.46	4.42

The bold font indicates the best SI-SDR for each method. For the proposed methods, the number in brackets indicates the number of iterations per time frame.

in stable performances even with aggressive updates of the SCM with small  $\beta$ .

To demonstrate the feasibility of the proposed methods, we investigated the computational time of Original and Prop-relax. Our experiment was performed using Python 3.10.9 with Intel Core i9-7900X CPU. Original took 5.9 ms to calculate the spatial filter per time frame. Meanwhile, Prop-relax took 10.4 ms to compute the inverse of the KKT matrix in (44) per time frame and 2.5 ms for each iteration in Algorithm 2. As the window shift is 128 ms, the real-time factor for 10 iterations per time frame was less than 0.3. This computational time does not include the actual target source extraction by filter-and-sum, but it can be performed efficiently [53], [54]. Although the overall computational complexity is relatively large for current hearing aids, we expect the complexity to be sufficiently small for several advanced devices including augmented reality headsets [55], [56].

## VII. CONCLUSION

We proposed a low-latency beamforming method named causal MPDR beamforming. The proposed method adds the causality constraint to the optimization problem of the original MPDR beamformer and performs beamforming in the time domain. In addition, the causal MPrDR beamformer was derived by relaxing the distortionless constraint, and we applied the DRS method to its optimization problem. Through the experiments, we revealed that the causality constraint can improve extraction performance. Furthermore, we presented the online causal MPrDR beamformer and validated its effectiveness through an experiment under non-stationary conditions.

### APPENDIX A

#### CONCORDANCE BETWEEN FILTER-AND-SUM IN THE TIME DOMAIN AND ISTFT OF BEAMFORMING IN (6)

*Proposition 1:* Let us assume the window length is a multiple of the window shift:  $N/r = Q \in \mathbb{N}$ . We assume that the window in STFT satisfies the perfect reconstruction property [57]

without windowing in iSTFT as follows:

$$\sum_{q=0}^{Q-1} g_{n+rq} = 1, \quad (52)$$

for all  $n = 0, \dots, r-1$ . We further assume  $\delta = 0$  and  $F = Q'N = QQ'r$  where  $Q' \in \mathbb{N}$  is greater than one. Then, the output of the time-domain filter-and-sum coincides with the iSTFT of the result of the T-F-domain beamforming  $y_{t,f}$ :

$$\tilde{y}_l = \sum_{m=0}^{M-1} \sum_{n'=0}^{F/2-1} \tilde{w}_{n',m} \tilde{x}_{l-n',m}, \quad (53)$$

$$= \sum_{t=t_l-QQ'+1}^{t_l} \sum_{f=0}^{F-1} y_{t,f} e^{2\pi i f(l-rt)/F}, \quad (54)$$

except for both ends of  $\tilde{y}_l$ , where  $t_l = \lfloor l/r \rfloor$ , and  $\lfloor \cdot \rfloor$  is the floor function.

*Proof:* Let us consider the windowed signal at the  $t$ th time frame with zero padding:

$$\tilde{d}_{t,l,m} = \begin{cases} \tilde{x}_{l,m} g_{l-rt} & (l \in [rt, rt + N - 1]) \\ 0 & (l \notin [rt, rt + N - 1]) \end{cases}, \quad (55)$$

where  $l = 0, \dots, L-1$ . According to the perfect reconstruction property in (52), the overlap-add of the windowed signal  $\tilde{d}_{t,l,m}$  reconstructs the original signal  $\tilde{x}_{l,m}$  as follows:

$$\sum_{t=t_l-QQ'+1}^{t_l} \tilde{d}_{t,l,m} = \sum_{t=t_l-QQ'+1}^{t_l} \tilde{d}_{t,l,m}, \quad (56)$$

$$= \sum_{t=t_l-QQ'+1}^{t_l} \tilde{x}_{l,m} g_{l-rt}, \quad (57)$$

$$= \sum_{q=0}^{Q-1} \tilde{x}_{l,m} g_{l+rq-rt_l}, \quad (58)$$

$$= \tilde{x}_{l,m}, \quad (59)$$

for each  $l$  except for the edge of the  $\tilde{x}_{l,m}$ . In (56), we use that  $\tilde{d}_{t,l,m}$  is the zero-padded windowed signal. Thanks to the linearity of convolution, the channel-wise linear convolution of  $\tilde{x}_{l,m}$  and  $\tilde{w}_{n',m}$  coincides with the overlap-add of the convolution of  $\tilde{d}_{t,l,m}$  and  $\tilde{w}_{n',m}$ :

$$\sum_{n'=0}^{F/2-1} \tilde{w}_{n',m} \tilde{x}_{l-n',m} = \sum_{n=0}^{F/2-1} \tilde{w}_{n',m} \left( \sum_{t=t_l-QQ'+1}^{t_l} \tilde{d}_{t,l-n',m} \right), \quad (60)$$

$$= \sum_{t=t_l-QQ'+1}^{t_l} \left( \sum_{n'=0}^{F/2-1} \tilde{w}_{n',m} \tilde{d}_{t,l-n',m} \right). \quad (61)$$

Next, we consider  $\tilde{e}_{t,n,m} = \tilde{d}_{t,n+rt,m}$  for  $n = 0, \dots, F-1$ . We emphasize that the latter half of  $\tilde{e}_{t,n,m}$  is zero according to the definition of  $\tilde{d}_{t,l,m}$  in (55). Since actual tap length of the

spatial filter  $\tilde{w}_{n',m}$  is  $F/2$  due to  $\delta = 0$ , the linear convolution of  $\tilde{e}_{t,n,m}$  and  $\tilde{w}_{n',m}$  coincides with their circular convolution:

$$\sum_{n'=0}^{F/2-1} \tilde{w}_{n',m} \tilde{e}_{t,n-n',m} = \tilde{w}_{n',m} \circledast \tilde{e}_{t,n,m}, \quad (62)$$

for all  $n = 0, \dots, F-1$ , where  $\circledast$  denotes the circular convolution with the DFT points of  $F$ . This is because the sum of the supports of  $\tilde{e}_{t,n,m}$  and  $\tilde{w}_{n',m}$  is equal to or less than  $F$  [58].

According to (61) and (62), the channel-wise linear convolution of the audio mixture  $\tilde{x}_{l,m}$  and the filter  $\tilde{w}_{n',m}$  coincides with the overlap-add of the result of the frame-wise circular convolution, i.e., iSTFT of  $\bar{w}_{f,m} x_{t,f,m}$ . Finally, thanks to the linearity of the summation, the output of the time-domain filter-and-sum in (53) is equal to the iSTFT of the sum of the channel-wise filtered STFT coefficients  $y_{t,f} = \sum_{m=0}^{M-1} \bar{w}_{f,m} x_{t,f,m}$ . That is, the iSTFT of the result of the T-F-domain beamforming coincides with  $\tilde{y}_l$  except for both ends.

## REFERENCES

- [1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.
- [2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2008.
- [3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
- [4] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.
- [5] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.
- [6] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2004.
- [7] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.
- [8] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. IEEE Int. Workshop Acoust. Signal Enhance.*, 2012, pp. 1–4.
- [9] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.
- [10] W. C. Liao, Z. Q. Luo, I. Merks, and T. Zhang, "An effective low complexity binaural beamforming algorithm for hearing aids," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.
- [11] E. Hadad et al., "Comparison of two binaural beamforming approaches for hearing aids," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 236–240.
- [12] W. Pu, J. Xiao, T. Zhang, and Z. Luo, "A penalized inequality-constrained minimum variance beamformer with applications in hearing aids," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 175–179.
- [13] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010.
- [14] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 216–220.
- [15] R. Landgraf, J. Köhler-Kaeß, C. Lüke, O. Niebuhr, and G. Schmidt, "Can you hear me now? Reducing the Lombard effect in a driving car using an in-car communication system," in *Proc. 8th Int. Conf. Speech Prosody*, 2016, pp. 1–5.

- [16] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 506–510.
- [17] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online source separation and noise reduction based on joint optimization with dereverberation," in *Proc. IEEE Eur. Signal Process. Conf.*, 2021, pp. 1000–1004.
- [18] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *J. Amer. Acad. Audiol.*, vol. 11, no. 6, pp. 330–336, Jun. 2000.
- [19] M. A. Stone, B. C. J. Moore, K. Meisenbacher, and R. P. Derleth, "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," *Ear Hear.*, vol. 29, no. 4, pp. 601–617, Aug. 2008.
- [20] S. U. N. Wood and J. Rouat, "Unsupervised low latency speech enhancement with RT-GCC-NMF," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 332–346, May 2019.
- [21] S. Wang, G. Naithani, A. Politis, and T. Virtanen, "Deep neural network based low-latency speech separation with asymmetric analysis-synthesis window pair," in *Proc. IEEE Eur. Signal Process. Conf.*, 2021, pp. 301–305.
- [22] J. Chua and W. B. Kleijn, "A low latency approach for blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1280–1294, Aug. 2019.
- [23] P. L. Combettes and J. C. Pesquet, "A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 564–574, Dec. 2007.
- [24] Y. Masuyama, K. Yamaoka, Y. Kinoshita, and N. Ono, "Causal distortionless response beamforming by alternating direction method of multipliers," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 585–590.
- [25] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
- [26] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.
- [27] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 1981–1985.
- [28] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 5210–5214.
- [29] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.
- [30] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 5, pp. 960–971, May 2019.
- [31] Y. Masuyama, M. Togami, and T. Komatsu, "Multichannel loss function for supervised speech source separation by mask-based beamforming," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2708–2712.
- [32] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 681–685.
- [33] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 531–535.
- [34] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 6697–6701.
- [35] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proc. 2nd Berkeley Symp. Math. Statist., Prob.*, 1950, pp. 481–492.
- [36] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, U.K.: Cambridge Univ. Press., 2018.
- [37] I. Yamada, S. Gandy, and M. Yamagishi, "Sparsity-aware adaptive filtering based on a douglas-rachford splitting," in *Proc. IEEE Eur. Signal Process. Conf.*, 2011, pp. 1929–1933.
- [38] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.
- [39] M. Togami, "Simultaneous optimization of forgetting factor and time-frequency mask for block online multi-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 2702–2706.
- [40] T. Nakatani and K. Kinoshita, "Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 111–115.
- [41] M. Yamagishi and I. Yamada, "Exploiting sparsity in feed-forward active noise control with adaptive douglas-rachford splitting," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–6.
- [42] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas, Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
- [43] H. V. Henderson and S. R. Searle, "On deriving the inverse of a sum of matrices," *SIAM Rev.*, vol. 23, no. 1, pp. 53–60, Jan. 1981.
- [44] J. Lorenzo-Trueba et al., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey*, 2018, pp. 195–202.
- [45] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 351–355.
- [46] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 626–630.
- [47] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, ITU-T Standard P.862.2, International Telecommunication Union, Geneva, Switzerland, 2007.
- [48] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [49] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, Jun. 2019.
- [50] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understandings*, 2015, pp. 504–511.
- [51] Y. Kondo, Y. Kondo, N. Takamune, D. Kitamura, and H. Saruwatari, "Deficient-basis-complementary rank-constrained spatial covariance matrix estimation based on multivariate generalized gaussian distribution for blind speech extraction," *EURASIP J. Adv. Signal Process.*, vol. 2022, Sep. 2022, Art. no. 88.
- [52] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tool. Appl.*, vol. 80, pp. 5653–5671, Oct. 2020.
- [53] A. Torger and A. Farina, "Real-time partitioned convolution for ambiphonics surround sound," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 195–198.
- [54] N. Jillings, J. D. Reiss, and R. Stables, "Zero-delay large signal convolution using multiple processor architectures," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 339–343.
- [55] P. Guiraud et al., "An introduction to the speech enhancement for augmented reality (spear) challenge," in *Proc. IEEE Int. Workshop Acoust. Signal Enhance.*, 2022, pp. 1–5.
- [56] K. Sekiguchi, A. A. Nugraha, Y. Du, Y. Bando, M. Fontaine, and K. Yoshii, "Direction-aware adaptive online neural speech enhancement with an augmented reality headset in real noisy conversational environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots, Syst.*, 2022, pp. 9266–9273.
- [57] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.
- [58] A. Oppenheim, R. Schaffer, and J. Buck, *Discrete-Time Signal Processing*. London, U.K.: Pearson, 1999.



**Yoshiki Masuyama** (Graduate Student Member, IEEE) received the B.E. and M.E. degrees from Waseda University, Tokyo, Japan in 2019 and 2021, respectively. He is currently working toward the Ph.D. degree with Tokyo Metropolitan University, Tokyo.



**Taishi Nakashima** (Student Member, IEEE) received the B.E. degree in engineering from Osaka University, Osaka, Japan, in 2019 and the M.S. degree in informatics from Tokyo Metropolitan University, Tokyo, Japan, in 2021. He is currently working toward the Ph.D. degree Tokyo Metropolitan University. He has received JSPS Research Fellowship (DC1) in April 2021. His research interests include blind source separation and acoustic signal processing. He is an esteemed Student Member of the Acoustical Society of Japan (ASJ) and the IEEE Signal Processing Society (SPS). He was the the 24th Best Student Presentation Award of ASJ, 16th IEEE SPS Japan Student Conference Paper Award in 2022, and Top 3% Recognition at ICASSP 2023.



**Kouei Yamaoka** (Student Member, IEEE) received the B.Sc. and M.E. degrees in information engineering and engineering from the University of Tsukuba, Tsukuba, Japan, in 2017 and 2019, respectively. He is currently working toward the Ph.D. degree with Tokyo Metropolitan University, Hino, Japan. His research interests include acoustic signal processing, signal enhancement, source localization, and asynchronous distributed microphone array. Mr. Yamaoka is a Member of the Acoustical Society of Japan.



**Nobutaka Ono** (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo, Tokyo, Japan, in 1996, 1998, and 2001, respectively. He was a Research Associate with the University of Tokyo in 2001, and became a Lecturer in 2005. He was also an Associate Professor with the National Institute of Informatics, Tokyo, in April 2011, and became a Professor in 2017. In 2017, he was with Tokyo Metropolitan University, Hino, Japan. He is the author or co-author of more than 280 articles in international journal papers and peer-reviewed conference proceedings. His research interests include acoustic signal processing, especially microphone array processing, source localization and separation, machine learning, and optimization algorithms. He was a Tutorial Speaker with ISMIR 2010 and ICASSP 2018. Dr. Ono is a Senior Member of IEEE Signal Processing Society and Member of the Acoustical Society of Japan, Institute of Electronics, Information and Communications Engineers, Information Processing Society of Japan, and Society of Instrument and Control Engineers, Tokyo. He was the Chair of Signal Separation Evaluation Campaign Evaluation Committee in 2013 and 2015, the Technical Program Chair of IWAENC 2018, General Chair of DCASE 2020 workshop, and a Member of IEEE Audio and Acoustic Signal Processing Technical Committee from 2014 to 2019. From 2012 to 2015, he was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is currently the Vice Chair of IEEE Signal Processing Society Tokyo Joint Chapter. He was the recipient of the Awaya Award from ASJ in 2007, Igarashi Award at the Sensor Symposium from IEEJ in 2004, Best Paper Award at IEEE ISIE in 2008, Measurement Division Best Paper Award from SICE in 2013, Best Paper Award in IEEE IS3C in 2014, Excellent Paper award in IIHMSP in 2014, Unsupervised Learning ICA pioneer Award from SPIE, DSS in 2015, Sato Paper Award from ASJ in 2000 and 2018, two TAF Telecom System Technology Awards in 2018, and Best Paper Award in APSIPA ASC in 2018.



**Yuma Kinoshita** (Member, IEEE) received the B.Eng., M.Eng., and the Ph.D. degrees from Tokyo Metropolitan University, Hino, Japan, in 2016, 2018, and 2020 respectively. In April 2020, he started to work with Tokyo Metropolitan University, as a project Assistant Professor. He moved to Tokai University, Japan, as an Associate Professor/Lecturer in April 2022. He became a Project Associate Professor with Tokyo Metropolitan University in April 2023. His research interests include the area of signal processing, image processing, and machine learning. He is a

Member of APSIPA, IEICE, and ASJ. He was the recipient of the IEEE ISPAACS Best Paper Award, in 2016, IEEE Signal Processing Society Japan Student Conference Paper Award, in 2018, IEEE Signal Processing Society Tokyo Joint Chapter Student Award, in 2018, IEEE GCCE Excellent Paper Award (Gold Prize), in 2019, IWAIT Best Paper Award, in 2020, and APSIPA ASC 2021 Best Paper Award. He was a Registration Chair of DCASE2020 Workshop.