FUNCTION DESIGNABLE BEAMFORMER BASED ON PROBABILISTIC ASSUMPTIONS ON FILTER AND ITS AUXILIARY VARIABLES

Ryotaro Sato, Kenta Niwa, and Noboru Harada

NTT Media Intelligence Laboratories, NTT Corporation, Japan

ABSTRACT

We propose a novel beamformer design method that exploits probabilistic assumptions on auxiliary variables derived from filters and observed signals. Many conventional beamformer design methods can be understood in the context of optimization problems for some probabilistic cost functions. However, the class of cost functions used with these methods is quite limited to reflect multiple pieces of information and our demands on the filter, such as the sparsity assumption of the source signals and the low-latency constraint of the filter. We propose a method to design cost functions that incorporate multiple probabilistic assumptions. The assumptions are expressed as the sum of many convex terms, and every term has different auxiliary variables that are linearly constrained. Such cost functions can be optimized by iteratively optimizing with regard to each term alternately. This method enables us to more arbitrarily tune the beamformer. We conducted numerical simulations showing that our method effectively improves the performance from multiple perspectives.

***Index Terms*—** beamforming, array signal processing, probabilistic modeling, convex optimization, alternating direction method of multipliers (ADMM)

1. INTRODUCTION

Microphone array beamforming [1, 2] is a common technique that enhances a target signal arriving from a specific direction while suppressing surrounding noise. It has been widely applied, such as in teleconference systems, car communication systems, and smart speakers [3, 4]. Many beamformer design methods in the literature estimate the optimal filters that minimize a cost function for output noise power under certain constraints [5]. For example, the problem with the well-known minimum variance distortionless response (MVDR) beamformer [6] is the minimization of the power of the output signal subject to the constraints for the response to the target direction. Alternatively, the maximum likelihood (ML) beamformer [7] minimizes the noise power in the output signal. Many studies to improve the performance have been conducted by imposing additive constraints or penalty terms on the cost function [8–12].

In many practical situations, we expect generated beamformers to have multiple desired functions such as low latency and less distortion, while retaining as high of a noise reduction ratio as possible. These demands are modeled by imposing probability assumptions on variables derived from the filter (auxiliary variables of the filter). For example, if we know that the target source is speech, the estimated signal can be well approximated by the Laplace distribution due to its sparseness in the time-frequency domain [13–15]. Furthermore, we empirically know that the filter coefficients will change smoothly with respect to frequency bins, although they are sometimes unstable at the bins where the spatial correlation matrix is rank-deficient. Moreover, if the filter coefficients are cooperatively designed such that the frequency elements change smoothly, they will result in a low-latency filter. If we can incorporate these probabilistic assumptions into the filter estimation, it will be possible to have multiple desired functions in the generated beamformer without specializing it for only noise reduction. We will then be able to generate multifunctional beamformers. However, since the cost function forms have been limited to relatively simple classes due to the limit of application of simple optimization techniques, it has been difficult to design beamformers to have multiple desired functions by just applying conventional methods.

In this paper, we propose a cost design method that imposes probabilistic assumptions on the filter and its auxiliary variables. We note that auxiliary variables (e.g. residual noises, estimated target signals, and the discrete differences between neighbor filter elements) are mere affine transformations of the filter. If the probabilistic assumptions for variables are given by log-concave distributions [16], the product of these distributions is also a log-concave form. Its negative logarithm is then a sum of convex functions, with which the filter and its auxiliary variables are linearly associated. For such a constrained convex minimization problem, the alternating direction method of multipliers (ADMM) [17, 18] is an applicable solver to obtain optimal filters. To show that a wide class of probabilistic assumptions can be used for the auxiliary variables in the proposed method, we will generate and evaluate a speech enhancement beamformer that has low latency while retaining the high noise reduction ratio.

This paper is organized as follows. We formulate the problem and overview the conventional methods in the context of probabilistic optimization problems in Section 2. In Section 3, we propose a cost function design for a function designable beamformer and its optimization procedure. After evaluating the generated beamformers in Section 4, we conclude this paper in Section 5.

2. FORMULATION OF BEAMFORMING METHODS

In this section, we first introduce symbol notations in 2.1. We then briefly explain the conventional beamformer design methods in association with probabilistic models in 2.2.

2.1. Problem Setting

Let us suppose that a microphone array with M omnidirectional sensors is placed in the field. By applying the beamforming filters to the recorded signals, we will emphasize the target signal arriving from an identified direction. To model the observed signals, we introduce several variables. Note that these variables are in the short-time Fourier transform (STFT) domain unless stated otherwise. Let $\boldsymbol{a}_f \in \mathbb{C}^M$ $(f = 1, \ldots, F)$ be the transfer functions from the source to microphones at every frequency bin f. We use $s_{f,t} \in \mathbb{C}$ to denote the target signal for each frame t $(t = 1, \ldots, T)$. We also denote the noise from the k-th interfering source as $n_{\mathrm{i}k,f,t} \in \mathbb{C}$. By using these notations, the observed signal $\boldsymbol{z}_{f,t}$ is modeled as instantaneous

Table 1. Comparison of representative beamformer design methods whose cost form is given by product of probabilistic distributions.

	Filter form	$-\log P_y(\boldsymbol{y}_f)$	$-\log P_e(\boldsymbol{e}_f)$	Constraint on $\tilde{\boldsymbol{w}}^*$	Given parameters
Delay Sum	$\frac{\boldsymbol{h}_s}{\boldsymbol{h}_s^{\mathsf{H}}\boldsymbol{h}_s}$	-	$\sigma_e^{-2}\|\boldsymbol{e}\|^2$	$\boldsymbol{w}^{\mathsf{H}}\boldsymbol{h}_s = 1$	$\boldsymbol{h}_s, \sigma_e^2$
Maximum Likelihood (ML)	$\frac{\mathbf{R}_e^{-1}\boldsymbol{h}_s}{\boldsymbol{h}_s^{\mathsf{H}}\mathbf{R}_e^{-1}\boldsymbol{h}_s}$	-	$\boldsymbol{e}^{\mathsf{H}}\mathbf{R}_e^{-1}\boldsymbol{e}$	$\boldsymbol{w}^{\mathsf{H}}\boldsymbol{h}_s = 1$	$\boldsymbol{h}_s, \mathbf{R}_e$
MVDR	$\frac{\mathbf{R}_z^{-1}\boldsymbol{h}_s}{\boldsymbol{h}_s^{\mathsf{H}}\mathbf{R}_z^{-1}\boldsymbol{h}_s}$	-	$\boldsymbol{e}^{\mathsf{H}}\mathbf{R}_z^{-1}\boldsymbol{e}$	$\boldsymbol{w}^{\mathsf{H}}\boldsymbol{h}_s = 1$	$\boldsymbol{h}_s, \mathbf{R}_z$
Robust Constraint on $\boldsymbol{h}_s$ [10]				$\mathrm{Re}(\boldsymbol{w}^{\mathsf{H}}\boldsymbol{h}_s) \geq 1,\ ^{\forall}\boldsymbol{h}_s \in \mathcal{E}$	$\mathcal{E}$ (an ellipsoid)

mixtures

$$\boldsymbol{z}_{f,t} = s_{f,t}\boldsymbol{a}_f + \sum_k n_{\mathrm{i}kf,t}\boldsymbol{a}_{\mathrm{i}kf} + \boldsymbol{n}_{\mathrm{b}f,t}, \tag{1}$$

where $\boldsymbol{n}_{\mathrm{b}f,t}$ represents background noise. Our task is to estimate the beamforming filters $\boldsymbol{w}_f \in \mathbb{C}^M$ such that the output $y_{f,t} \in \mathbb{C}$ will be close to $s_{f,t}$. The $y_{f,t}$ is given by

$$y_{f,t} = \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{z}_{f,t} \tag{2}$$

where the superscript $^{\mathsf{H}}$ denotes the complex conjugate transpose.

For the later part of this paper, several variables that are derived from filters and observed signals are defined. We can estimate the residual noise components included in the observed signals $\boldsymbol{e}_{f,t} \in \mathbb{C}^M$, which are given by

$$\boldsymbol{e}_{f,t} = \boldsymbol{z}_{f,t} - y_{f,t}\boldsymbol{h}_f = \boldsymbol{z}_{f,t} - (\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{z}_{f,t}) \cdot \boldsymbol{h}_f. \tag{3}$$

Here $\boldsymbol{h}_f \in \mathbb{C}^M$ denotes the array manifold vector towards the target direction. The $\boldsymbol{e}_{f,t}$ will contain interference and background noises if the beamformer output is mainly composed of the target signal.

Note that the relationships between the estimated variables ($y_{f,t}$, $\boldsymbol{e}_{f,t}$) and $\tilde{\boldsymbol{w}}$ are described by affine transformations. To make this apparent, a special notation is introduced. For arbitrary vector or scalar-valued variables $\boldsymbol{x}_f$ that have an index running over frequency bins, we denote the vector containing all frequency bins' information as $\tilde{\boldsymbol{x}} = (\boldsymbol{x}_1^{\mathsf{T}}, \ldots, \boldsymbol{x}_F^{\mathsf{T}})^{\mathsf{T}}$. In addition, matrices $\mathbf{F}_t$ and $\mathbf{G}_t$ for each t are introduced as

$$\mathbf{F}_t = \mathrm{diag}[\boldsymbol{z}_{1,t}, \boldsymbol{z}_{2,t}, \ldots, \boldsymbol{z}_{F,t}]^{\mathsf{T}} \in \mathbb{C}^{F\times MF}, \tag{4}$$

$$\mathbf{G}_t = \mathrm{diag}[\boldsymbol{h}_1\boldsymbol{z}_{1,t}^{\mathsf{T}}, \boldsymbol{h}_2\boldsymbol{z}_{2,t}^{\mathsf{T}}, \ldots, \boldsymbol{h}_F\boldsymbol{z}_{F,t}^{\mathsf{T}}] \in \mathbb{C}^{MF\times MF}. \tag{5}$$

With these notations, the beamformer output and residual components in the STFT domain are expressed as the following forms:

$$\tilde{\boldsymbol{y}}_t = \mathbf{F}_t\tilde{\boldsymbol{w}}^*, \tag{6}$$

$$\tilde{\boldsymbol{e}}_t = -\mathbf{G}_t\tilde{\boldsymbol{w}}^* + \tilde{\boldsymbol{z}}_t \tag{7}$$

where * denotes the complex conjugate of the original vector.

2.2. Conventional Beamformer Design Methods

The conventional beamformer design methods are organized as a probabilistic cost optimization problem. Assume that $\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{e}}$, and $\tilde{\boldsymbol{w}}^*$ are random variables and their distributions $P_y(\tilde{\boldsymbol{y}}), P_e(\tilde{\boldsymbol{e}})$, and $P_w(\tilde{\boldsymbol{w}}^*)$ are given, respectively. $P_y(\tilde{\boldsymbol{y}})$ and $P_e(\tilde{\boldsymbol{e}})$ are expected to reflect the statistical properties of the variables. On the other hand, $P_w(\tilde{\boldsymbol{w}}^*)$ is often used to directly represent the desired frequency response to the target direction. Under these assumptions, the likelihood function of $\tilde{\boldsymbol{w}}^*$ for any given observed signals $\{\tilde{\boldsymbol{z}}_t\}_{t=1}^T$ is expressed as

$$L(\tilde{\boldsymbol{w}}^* \mid \{\tilde{\boldsymbol{z}}_t\}) \propto \prod_{t=1}^{T} P_y(\tilde{\boldsymbol{y}}_t \mid \tilde{\boldsymbol{w}}^*)P_e(\tilde{\boldsymbol{e}}_t \mid \tilde{\boldsymbol{w}}^*) \cdot P_w(\tilde{\boldsymbol{w}}^*), \tag{8}$$

where the relationships between $\tilde{\boldsymbol{y}}_t$ and $\tilde{\boldsymbol{e}}_t$ and $\tilde{\boldsymbol{w}}^*$ are given by (6) and (7), respectively. Maximizing this likelihood w.r.t. $\tilde{\boldsymbol{w}}^*$ will lead to the optimal filter estimation:

$$\underset{\tilde{\boldsymbol{w}}^*}{\mathrm{minimize}}\, L(\tilde{\boldsymbol{w}}^* \mid \{\tilde{\boldsymbol{z}}_t\}). \tag{9}$$

The formulation (9) is applicable to various conventional beamformer design methods. We briefly describe the derivation of the MVDR beamformer as an example. Let us assume that the spatial correlation matrices of $\boldsymbol{z}_f$'s are given by $\mathbf{R}_f = \mathrm{E}_{\boldsymbol{z}_f}[\boldsymbol{z}_f\boldsymbol{z}_f^{\mathsf{H}}]$ ($f = 1, \ldots, F$). We also assume that $\tilde{\boldsymbol{e}}_t$ is normally distributed: $\boldsymbol{e}_{f,t} \sim \mathcal{N}(0, \mathbf{R}_f)$. In addition, the distortionless constraints on $\boldsymbol{w}_f^*$'s are imposed: $\boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{a}_f = 1$. Under these assumptions, the problem form (9) is reformulated by

$$\underset{\{\boldsymbol{w}_f\}}{\mathrm{minimize}} \sum_{f=1}^{F}(\boldsymbol{w}_f - \gamma_f\mathbf{R}_f^{-1}\boldsymbol{a}_f)^{\mathsf{H}}\mathbf{R}_f(\boldsymbol{w}_f - \gamma_f\mathbf{R}_f^{-1}\boldsymbol{a}_f) \tag{10}$$

$$\text{s.t. } \boldsymbol{w}_f^{\mathsf{H}}\boldsymbol{a}_f = 1,$$

where $\gamma_f = (\boldsymbol{a}_f^{\mathsf{H}}\mathbf{R}_f^{-1}\boldsymbol{a}_f)^{-1}$. Now it is clear that the solution is the well-known MVDR beamformer $\boldsymbol{w}_{\mathrm{opt}f} = \gamma_f\mathbf{R}_f^{-1}\boldsymbol{a}_f$. Table 1 explains probabilistic assumption differences among representative conventional methods.

However, since the formulation (9) is given by the optimization w.r.t. $\tilde{\boldsymbol{w}}^*$, two problems remain. One is that the probabilistic assumption on sound signals is usually limited to simple distribution forms, e.g., normal distributions. However, normal distributions are suggested to not always be appropriate to express sound signals [13, 14, 19]. The other problem is that the class of cost functions or constraints of $\tilde{\boldsymbol{w}}^*$ is too limited to consider various probabilistic assumptions in parallel. Some studies have investigated imposing additive assumptions for $\tilde{\boldsymbol{w}}^*$ [10–12]. However, optimizing cost functions with a variety of complicated priors is generally quite difficult since such cost functions are composed of diverse $\tilde{\boldsymbol{w}}^*$-dependent terms that are complexly correlated with each other. These issues make it quite difficult to generate a beamformer that has multiple desired functions such that it simultaneously achieves low latency, stability, and high noise reduction performance, for example.

To overcome these problems, it might be useful to impose the probabilistic assumptions on the auxiliary variables such as $\tilde{\boldsymbol{y}}$ (6) and $\tilde{\boldsymbol{e}}$ (7) and express the cost function as the sum of apparently independent terms. Our considerations in this paper are based on this idea.

3. PROPOSED METHOD

A new method for a function designable beamformer is proposed in 3.1. We then discuss applying this method in a realistic situation and show its flexibility.

3.1. Problem Formulation

We propose a new method of beamformer cost design that depends on not only $\tilde{\boldsymbol{w}}^*$ but also the newly introduced auxiliary variables (e.g. $\tilde{\boldsymbol{e}}$ and $\tilde{\boldsymbol{y}}$). Here the auxiliary variables are supposed to be expressed as the affine transformations of $\tilde{\boldsymbol{w}}^*$.

We first introduce J auxiliary variables $\boldsymbol{v}_j$ ($j = 1, \ldots, J$), which are related to $\tilde{\boldsymbol{w}}^*$ by $\boldsymbol{v}_j = \mathbf{D}_j\tilde{\boldsymbol{w}}^* + \boldsymbol{b}_j$ for every j. Note that these are generalizations of $\tilde{\boldsymbol{y}}_t$ (6) and $\tilde{\boldsymbol{e}}_t$ (7); therefore, our method will include the methods argued in the previous section. To simplify

the notations, we write $\hat{\boldsymbol{v}} = (\boldsymbol{v}_1^\mathsf{T}, \ldots, \boldsymbol{v}_J^\mathsf{T})^\mathsf{T}, \hat{\mathbf{D}} = (\mathbf{D}_1^\mathsf{T}, \ldots, \mathbf{D}_J^\mathsf{T})^\mathsf{T}$, and $\hat{\boldsymbol{b}} = (\boldsymbol{b}_1^\mathsf{T}, \ldots, \boldsymbol{b}_J^\mathsf{T})^\mathsf{T}$, accordingly.

Using $\boldsymbol{v}_j$'s, suppose that the whole cost function can be expressed as

$$L(\tilde{\boldsymbol{w}}^*, \hat{\boldsymbol{v}}) = L_0(\tilde{\boldsymbol{w}}^*) + \sum_{j=1}^{J} L_j(\boldsymbol{v}_j), \tag{11}$$

where each L_j $(j = 0, \ldots, J)$ is assumed to be convex. To see the validity of this assumption, suppose that we use *log-concave* distributions for all the auxiliary variables such as $\tilde{\boldsymbol{y}}_t$'s or $\tilde{\boldsymbol{e}}_t$'s. Here, a probability distribution is called log-concave if the negative logarithm of its density function is convex. Many probabilistic distributions that are used to describe signals in audio signal processing literature belong to this class, such as the normal and Laplace distributions. Each L_j in (11) can be interpreted as the negative logarithm of the likelihood for $\boldsymbol{v}_j$; therefore, the convexity assumption is automatically satisfied for such distributions.

Now our task is reduced to the typical constrained convex minimization problem below:

$$\underset{\hat{\boldsymbol{v}}, \tilde{\boldsymbol{w}}}{\text{minimize}}\, L_0(\tilde{\boldsymbol{w}}^*) + \sum_{j=1}^{J} L_j(\boldsymbol{v}_j) \;\text{ s.t. }\; \hat{\boldsymbol{v}} = \hat{\mathbf{D}}\tilde{\boldsymbol{w}}^* + \hat{\boldsymbol{b}}. \tag{12}$$

This problem can be solved by dividing the whole problem into a couple of subproblems that are linearly constrained, and then optimizing each term repeatedly. Although there are various algorithms that can be applied to the constrained convex optimization problems of the form (12), the ADMM-based algorithm [18] is used as an instance, shown in 3.3. In the next subsection, we specifically design the proposed cost form assuming a practical situation.

3.2. Example of Cost Functions for Multifunctional Beamformer

We now discuss the application of the proposed cost design method (12) assuming a practical situation. In the following, we consider a situation in which we want to enhance the speech signal from a speaker for live streaming. Several interfering sources also exist, which emit signals following the complex normal distributions at each frequency bin. In such a situation, we prefer a beamformer that is specialized for speech enhancement and achieves extremely low latency in the time domain.

To incorporate the known property of signal distributions into cost functions, we impose probabilistic assumptions on the auxiliary variables $\tilde{\boldsymbol{y}}_t$ and $\tilde{\boldsymbol{e}}_t$, which are defined as (6) and (7). As the distribution of $\tilde{\boldsymbol{e}}_t$, we use the normal distribution

$$P(\boldsymbol{e}_{f,t}) \propto \exp(-\boldsymbol{e}_{f,t}^H \mathbf{R}_f^{-1} \boldsymbol{e}_{f,t}). \tag{13}$$

Meanwhile, the literature shows that a speech signal is often assumed to be sparse in the STFT domain [13, 14]. Motivated by this consideration, we adopt the Laplace distribution as distributions of $\tilde{\boldsymbol{y}}_t$

$$P(y_{f,t}) \propto \exp(-\beta|y_{f,t}|) \tag{14}$$

where $\beta(>0)$ is a constant related to the variance.

We also introduce new auxiliary variables and impose on them the probabilistic distributions that induce the resultant beamformer to perform with relatively low latency. To design these terms, we first reconsider the meaning of $\tilde{\boldsymbol{w}}^*$. Most conventional wide-band beamformers estimate the optimal filter for each frequency bin individually without taking into account the relevance between different bins. However, the discontinuous or non-smooth $\tilde{\boldsymbol{w}}^*$ with respect to f may cause long impulse responses in the time domain. Furthermore, we do not prefer the filter that contains the all-pass factor, which causes unnecessary group delay. These unwanted characteristics of filters increase the values of the discrete second derivatives of the phase response with respect to f. Motivated by these considerations, we introduce new $F-2$ auxiliary variables

$$\boldsymbol{\eta}_f = \boldsymbol{w}_f^* - 2\boldsymbol{w}_{f+1}^* + \boldsymbol{w}_{f+2}^* \;\; (f = 1, \ldots, F-2),$$

and define $L_{\eta f}(\boldsymbol{\eta}_f)$ as

$$L_{\eta f}(\boldsymbol{\eta}_f) = \lambda \left\|\boldsymbol{\eta}_f\right\|_2, \tag{15}$$

which means that $\|\boldsymbol{\eta}_f\|$'s should be small enough. To the best of our knowledge, this is the first study that imposed probabilistic assumptions on the discrete second derivatives of the filter.

By introducing the probabilistic assumptions on variables $\tilde{\boldsymbol{e}}$ (13), $\tilde{\boldsymbol{y}}$ (14), and $\tilde{\boldsymbol{w}}^*$ (15) as shown above, the negative log likelihood of the joint probability is written as

$$L(\tilde{\boldsymbol{w}}, \hat{\boldsymbol{v}}) = \sum_{t=1}^{T}\sum_{f=1}^{F}(\boldsymbol{e}_{f,t}^\mathsf{H}\mathbf{R}_f^{-1}\boldsymbol{e}_{f,t} + \beta|y_{f,t}|) + \sum_{f=1}^{F-2}\lambda\left\|\boldsymbol{\eta}_f\right\|_2, \tag{16}$$

$$\hat{\boldsymbol{v}} = \left[\boldsymbol{e}_{1,1}, \ldots, \boldsymbol{e}_{F,T}, y_{1,1}, \ldots, y_{F,T}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{F-2}\right]. \tag{17}$$

We emphasize that all these $2FT + F - 2$ auxiliary variables are expressed as the affine transformations of $\tilde{\boldsymbol{w}}^*$; therefore, the optimization problem of (16) can be interpreted as an instance of (12).

Algorithm 1 Beamformer optimization based on ADMM

```
Initialize v̂, û, γ.
for n = 1, ..., N_iteration do
    w̃* ← arg min_{w̃*} L_0(w̃*) + (γ/2)‖D̂w̃* − v̂ + û + b̂‖_2^2
    for j = 1, ..., J do
        v_j ← prox_{L_j(·)/γ}(D_j w̃* + u_j + b_j)
        u_j ← u_j + D_j w̃* − v_j + b_j
    end for
end for
return w̃*
```

3.3. ADMM-based Optimization Algorithm

To solve the constrained minimization problem (12) and derive a fixed beamformer, we write out an iterative algorithm in this subsection. It is well known that ADMM is an effective solver for the problem form (12). This algorithm solves the original problem by solving its *dual* problem [20]. The ADMM-based method for solving (12) is written as Algorithm 1. Here, $\boldsymbol{u}_j$ is a dual variable that has the same dimension as its corresponding variable $\boldsymbol{v}_j$. In the following, we derive the concrete update rules for each variable of (17) on the basis of this algorithm.

The update rule for $\tilde{\boldsymbol{w}}^*$ is derived straightforwardly. Since $L_0(\cdot)$ vanishes in (16), it is simplified to

$$\tilde{\boldsymbol{w}}^* \leftarrow (\hat{\mathbf{D}}^\mathsf{H}\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}^\mathsf{H}(\hat{\boldsymbol{v}} - \hat{\boldsymbol{u}} - \hat{\boldsymbol{b}}). \tag{18}$$

Here $\hat{\mathbf{D}}^\mathsf{H}\hat{\mathbf{D}} = \sum_j \mathbf{D}_j^\mathsf{H}\mathbf{D}_j$ is actually a block band matrix:

$$\hat{\mathbf{D}}^\mathsf{H}\hat{\mathbf{D}} = \begin{pmatrix} \mathbf{A}_1 + \mathbf{I}_M & -2\mathbf{I}_M & \mathbf{I}_M & \cdots & 0 \\ -2\mathbf{I}_M & \mathbf{A}_2 + 5\mathbf{I}_M & -4\mathbf{I}_M & \cdots & 0 \\ \mathbf{I}_M & -4\mathbf{I}_M & \mathbf{A}_3 + 6\mathbf{I}_M & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{A}_F + \mathbf{I}_M \end{pmatrix}, \tag{19}$$

where each $\mathbf{A}_f = (1 + \|\boldsymbol{h}_f\|_2^2)\sum_t \boldsymbol{z}_{f,t}^*\boldsymbol{z}_{f,t}^\mathsf{T}$ is an $M \times M$ matrix. We can alleviate the computation cost of multiplying $(\hat{\mathbf{D}}^\mathsf{H}\hat{\mathbf{D}})^{-1}$ by precomputing the Cholesky decomposition of $\hat{\mathbf{D}}^\mathsf{H}\hat{\mathbf{D}}$ [21].

On the other hand, the update rules for the auxiliary variables are the proximal operators of the corresponding cost terms. Here, for a given convex function f, the function *proximal operator* $\mathrm{prox}_f(\cdot)$ is defined by $\mathrm{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} f(\boldsymbol{y}) + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2/2$. The update

Table 2. Simulation conditions.

Signal	$f_{\mathrm{s}} = 16$ kHz, 80000 samples ($T = 5$ s)
STFT	Hann window, 1024 samples (64 ms)
Frame shift	512 samples (32 ms)
Input signal	Target / (Interference + Background) : 3.5 dB, Target / Background : 10.5 dB
Room	2 m * 1.5 m, absorption rate $= 0.5$
$N_{\mathrm{Iteration}}$	100
γ	(maximum eigenvalue of $\mathbf{R}_f$'s)$^{-1}$
(λ, β)	$(0.3 f_{\mathrm{s}} T, 0.01\gamma)$, $(0, 0.01\gamma)$, $(0.3 f_{\mathrm{s}} T, 0)$

rules for $y_{f,t}$ and $\boldsymbol{\xi}_f$ can be derived using the proximal operator of the ℓ^2 norm:

$$\mathrm{prox}_{\lambda\|\cdot\|_2}(\boldsymbol{z}) = (\|\boldsymbol{z}\|_2 - \lambda)_+ \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2} \tag{20}$$

where $(\cdot)_+$ denotes $\max(\cdot, 0)$. The $\boldsymbol{e}_{f,t}$-dependent terms are just the quadratic forms. The proximal operator of such a function is easily derived from the definition. Consequently, the update rules for the auxiliary variables are written as

$$\begin{aligned}
\boldsymbol{e}_{f,t} &\leftarrow \frac{\gamma}{2}\mathbf{R}_f \left(I + \frac{\gamma}{2}\mathbf{R}_f\right)^{-1} \left(\boldsymbol{z}_{f,t} - (\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{z}_{f,t})\boldsymbol{h}_f + \boldsymbol{u}_{e,f,t}\right), \\
y_{f,t} &\leftarrow \left(1 - \frac{\beta}{\gamma|\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{z}_{f,t} + u_{y,f,t}|}\right)_+ (\boldsymbol{w}_f^{\mathsf{H}} \boldsymbol{z}_{f,t} + u_{y,f,t}), \\
\boldsymbol{\eta}_f &\leftarrow \left(1 - \frac{\lambda}{\gamma\|\boldsymbol{w}_f^* - 2\boldsymbol{w}_{f+1}^* + \boldsymbol{w}_{f+2}^* + \boldsymbol{u}_{\eta,f}\|}\right)_+ \\
&\quad \cdot (\boldsymbol{w}_f^* - 2\boldsymbol{w}_{f+1}^* + \boldsymbol{w}_{f+2}^* + \boldsymbol{u}_{\eta,f}).
\end{aligned} \tag{21}$$

As a result, the update rules if we adapt the ADMM to our constructed model are summarized as (18) and (21).

4. EXPERIMENT

4.1. Conditions

We conducted numerical simulations to evaluate the efficacy of the proposed method. Six microphones were equally spaced on a circumference of radius 2 cm. We used 100 utterances recorded from six male and female speakers as the target signals. The number of interfering sources was chosen randomly from the range $2, \ldots, 6$ for each trial, and their arrival directions were also randomly decided within the range where the relative angle to the target source is more than 60°. Every interfering source independently emits the normally distributed signal that is filtered by a low-pass filter with cutoff frequency of 1600 Hz. The recorded signal is modeled as the sum of the ones propagated from the sources and the weak white background noise generated by $\mathcal{N}(0, \sigma_{\mathrm{n}}^2 \mathbf{I})$, where $\sigma_n^2 = 10^{-6}$. All the target and interfering sources were located on the horizontal plane on which the microphones lie, at a distance of 0.5 m. All of these elements were in a rectangular room. The impulse responses were obtained by numerical simulations in advance [22]. We compared the performance of the proposed method with those of the conventional methods: the MVDR and ML beamformers. In the proposed method, the initial value of $\tilde{\boldsymbol{w}}$ is taken to be the delay sum beamformer. Two variants of the proposed method were also evaluated: one ignores the distribution for $\tilde{\boldsymbol{w}}^*$ by setting $\lambda = 0$, and other ignores the distribution for $\tilde{\boldsymbol{e}}^*$ by setting $\gamma = 0$. We evaluated the methods by using the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR). Other conditions are summarized in Table 2.

4.2. Results

Figure 1 shows the SDR and SIR improvements for each method, which represent how much the target signals were enhanced. The results indicate that the proposed method improved the mean SDR and SIR by 2.0 and 1.3 dB, respectively. The SDR and SIR can be further improved by ignoring the cost terms for $\tilde{\boldsymbol{w}}^*$ ($\lambda = 0$).

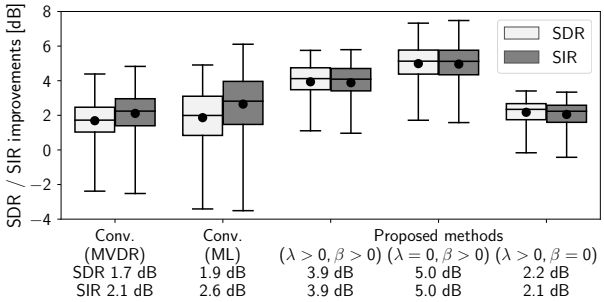

Fig. 1. Comparison of signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) improvements between conventional and proposed methods. Boxplots represent quantiles, minimum/maximum values, and median for each condition. Dots also indicate mean values. Both $\beta = 0$ and $\lambda = 0$ mean ignoring cost terms for $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{w}}^*$, respectively.

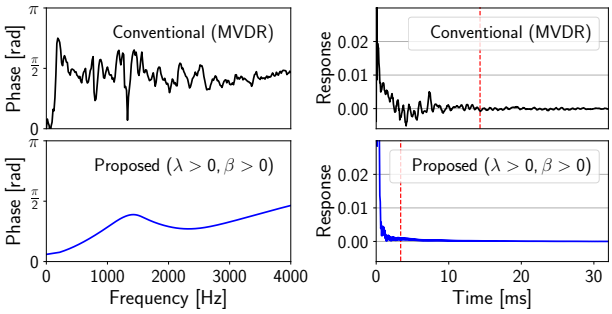

Fig. 2. (Left) Phase responses of filters estimated by both conventional (MVDR) and proposed methods. (Right) Filters' impulse responses, normalized by their ℓ^1 norms. Red dotted lines also indicate times when cumulative sum of absolute amplitudes is equal to 90% of their total sum.

To confirm the improvement of latency characteristics, we also compared the phase characteristics and waveforms of the filters estimated by both methods, as shown in Figure 2. These results suggest that the proposed cost terms for $\tilde{\boldsymbol{w}}^*$ were effective in reducing the latency, as expected. The conventional method formed a long and noisy tail, while the proposed method formed a relatively smooth filter by introducing the relationships among different frequency bins.

From these results, we can conclude that beamformers generated with the proposed method have multiple characteristics such as better SDR/SIR improvement and less latency, compared with those generated with conventional methods.

5. CONCLUSION

We proposed a probabilistic beamformer cost design method to give a beamformer multiple desired functions. Its auxiliary variables (e.g. residual noise, output signal) are affine transformations of filter coefficients. By imposing log-concave probabilistic assumptions for each variable, the cost form is consequently a linearly constrained convex minimization problem. To solve it, we applied an ADMM-based solver to obtain the optimal filter. On the basis of the proposed cost formulation, we showed a design example and confirmed that the obtained beamformer remarkably enhanced the SDR and SIR with low latency compared with conventional methods. For future work, cost terms appropriate for each situation or application of other optimization algorithms will be considered in detail.

6. REFERENCES

[1] Michael Brandstein and Darren Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, June 2001.

[2] J. C. Chen, Kung Yao, and R. E. Hudson, "Source localization and beamforming," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, March 2002.

[3] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.

[4] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 436–443.

[5] Harry L Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*, John Wiley & Sons, 2004.

[6] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug 1969.

[7] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, Sept 2004.

[8] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, October 1987.

[9] S. A. Vorobyov, A. B. Gershman, and Zhi-Quan Luo, "Robust adaptive beamforming using worst-case performance optimization: a solution to the signal mismatch problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 313–324, Feb 2003.

[10] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1684–1696, May 2005.

[11] S. A. Vorobyov, Yue Rong, and A. B. Gershman, "Robust adaptive beamforming using probability-constrained optimization," in *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*, July 2005, pp. 934–939.

[12] C. Chen and P. P. Vaidyanathan, "Quadratically constrained beamforming robust against direction-of-arrival mismatch," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 4139–4150, Aug 2007.

[13] Joon-Hyuk Chang and Nam Soo Kim, "Voice activity detection based on complex laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, April 2003.

[14] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 1, pp. I–253–I–256.

[15] Bowon Lee, Ton Kalker, and Ronald W Schafer, "Maximum-likelihood sound source localization with a multivariate complex Laplacian distribution," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), Seattle, USA*, 2008.

[16] Mark Bagnoli and Ted Bergstrom, "Log-concave probability and its applications," *Economic Theory*, vol. 26, no. 2, pp. 445–469, Aug 2005.

[17] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[18] N. Komodakis and J. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, Nov 2015.

[19] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, Nov 2012.

[20] Ernest K Ryu and Stephen Boyd, "Primer on monotone operator methods," *Appl. Comput. Math*, vol. 15, no. 1, pp. 3–43, 2016.

[21] G. Meurant, "A review on the inverse of symmetric tridiagonal and block tridiagonal matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 3, pp. 707–728, 1992.

[22] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 351–355.