

Is the Ideal Ratio Mask Really the Best? — Exploring the Best Extraction Performance and Optimal Mask of Mask-based Beamformers

Atsuo Hiroe*, Katsutoshi Itoyama*[†], and Kazuhiro Nakadai*

* Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan

[†] Honda Research Institute Japan Co., Ltd., Saitama, Japan

E-mail: {hiroe,itoyama,nakadai}@ra.sc.e.titech.ac.jp

Abstract—This study investigates mask-based beamformers (BFs), which estimate filters to extract target speech using time-frequency masks. Although several BF methods have been proposed, the following aspects are yet to be comprehensively investigated. 1) Which BF can provide the best extraction performance in terms of the closeness of the BF output to the target speech? 2) Is the optimal mask for the best performance common for all BFs? 3) Is the ideal ratio mask (IRM) identical to the optimal mask? Accordingly, we investigate these issues considering four mask-based BFs: the maximum signal-to-noise ratio BF, two variants of this, and the multichannel Wiener filter (MWF) BF. To obtain the optimal mask corresponding to the peak performance for each BF, we employ an approach that minimizes the mean square error between the BF output and target speech for each utterance. Via the experiments with the CHiME-3 dataset, we verify that the four BFs have the same peak performance as the upper bound provided by the ideal MWF BF, whereas the optimal mask depends on the adopted BF and differs from the IRM. These observations differ from the conventional idea that the optimal mask is common for all BFs and that peak performance differs for each BF. Hence, this study contributes to the design of mask-based BFs.

I. INTRODUCTION

Target speech extraction is effective in improving speech intelligibility in telecommunication systems and the performance of automatic speech recognition systems [1]. Therefore, beamformers (BFs) are utilized to avoid nonlinear distortions such as musical noises and spectral distortions [2], [3]. In the last decade, combined frameworks comprising BFs and deep neural networks (DNNs), referred to as mask-based BFs, have been proposed [4]–[6]. In these frameworks, DNNs generate one or two time-frequency (TF) masks corresponding to the target, interferences, or both to inform the BF of the sound to be enhanced or suppressed. Afterward, the BF estimates a filter for extracting the target using masks. For filter estimation, the following BF methods are adopted: 1) maximum signal-to-noise ratio (max-SNR) or generalized eigenvalue (GEV) BF [4], [5], [7], 2) minimum variance distortionless response (MVDR) BF [5], [6], [8], and 3) multichannel Wiener filter (MWF) BF [9]–[11].

Several mask types have been examined to achieve improved extraction performance. Initially, ideal binary masks (IBMs) were employed to train DNNs for mask generation [4], [5].

Subsequently, ideal ratio masks (IRMs) have been utilized [6], [11], [12].

Overviewing these studies, we concluded that the following aspects are yet to be comprehensively investigated:

- 1) Which BF can achieve the best performance if an optimal mask is provided for each BF method?
- 2) Is an optimal mask common for all BF methods?
- 3) Are conventional ideal masks such as IRMs identical to an optimal mask?

Regarding these aspects, several studies such as [5], [13], and [14] considered that the mask optimal for the single-channel TF masking [12], [15] should commonly be optimal for all BF methods. However, this assumption was not verified in these studies. Moreover, they compared multiple BF methods that employed the same mask in terms of extraction performance. However, these results do not answer the first aspect unless the optimal mask is common.

The remainder of this paper is organized as follows. Sections II and III explain the related work and BF methods presented in this study, respectively. Section IV examines the method for obtaining the optimal mask for each BF method. Section V verifies the aforementioned points experimentally while Section VI discusses the experimental results. Finally, Section VII concludes the study.

II. RELATED WORK

First, we overview the history of mask-based BFs. The max-SNR, MVDR, and MWF BFs employ the statistics of the target, interferences, or both. The statistics are called target (or speech) and interference (or noise) covariance matrices. Given that the accuracy of both statistics affects the extraction performance, estimating them accurately is a fundamental issue [16], [17]. In [4], [5], both statistics were computed using two binary masks, each of which represents periods when only the target or interferences are present, and the masks were generated using a properly trained DNN. This idea was first adopted for the max-SNR and MVDR BFs [4]–[6], then applied to the MWF BF [18].

Second, we mention the studies that compare mask-based BFs. At least three methods have been adopted as aforementioned, and several studies have compared two or three of them,

as presented in Table I. However, the best-performing method depends on experimental setups. Moreover, it is reported in [18] that the difference between the max-SNR and MWF BFs is marginal and the performance depends on the number of microphones used.

Third, we mention the mask types considered to be optimal for the mask-based BFs. Initially, the IBMs were considered to be optimal [4], [5]; then, the IRMs were considered optimal [6], [12], [20]–[22]. These ideas were based on the findings in the single-channel TF masking [12], [15]. However, these studies did not investigate whether these findings really apply to the mask-based BFs or the optimal mask is common for any BF method.

III. FRAMEWORK OF MASK-BASED BFs

In this study, we define the best extraction performance as the BF output closest to the target in the TF domain, considering that a significant goal of BFs is to extract (or estimate) the target. Moreover, similar to conventional TF masks, we have the constraint that mask values are nonnegative and real-valued.

Fig. 1 illustrates our idea. The vertical and horizontal axes indicate the closeness of the BF output to the target and the variation of the mask values, respectively. Although the mask values vary multidimensionally, this figure conceptually represents the variation as a single axis. In the mask-based BFs, the extraction performance should depend on the variation and exhibit the peak on a particular mask. We refer to this mask as the optimal one. Considering that multiple BF methods and mask types are employed, we can rephrase the questions mentioned in Section I as follows:

Issue 1: Which BF method has the highest peak, or are all peaks of the same height?

Issue 2: Is the optimal mask common for all BF methods, or dependent on each one?

Issue 3: Can the mask considered to be ideal achieve peak performance? More particularly, is the IRM optimal?

We assumed that the studies presented in Table I considered these issues as in Fig. 2. This indicates that the mask mentioned in each study is optimal for multiple BF methods such as BFs 1 and 2, whereas each BF method demonstrates a different height of the peak. However, this idea is yet to be verified.

Considering that this study is the first to examine these issues, we focus on the following methods for simplicity:

TABLE I

STUDIES COMPARING MULTIPLE BF METHODS. METHODS IN BOLD TYPE REPRESENT THOSE THAT PERFORMED THE BEST IN EACH STUDY. IN [18], THE PERFORMANCE DEPENDS ON THE NUMBER OF MICROPHONES. (SNR: SIGNAL-TO-NOISE RATIO, PESQ: PERCEPTUAL EVALUATION OF SPEECH QUALITY, WER: WORD ERROR RATE)

	Metric	Methods compared
Heymann+16 [5]	SNR	Max-SNR , MVDR
Boeddeker+18[13]	SNR,PESQ	Max-SNR , MVDR
Wang+18 [14]	WER	Max-SNR, MVDR, MWF
Heymann+18 [18]	WER	Max-SNR, MWF
Shimada+18 [19]	WER	MVDR, MWF

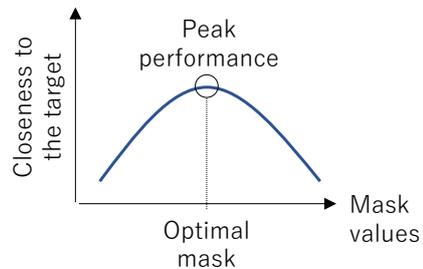


Fig. 1. Conceptual plot of the relationship between the closeness of the BF output to the target and mask values

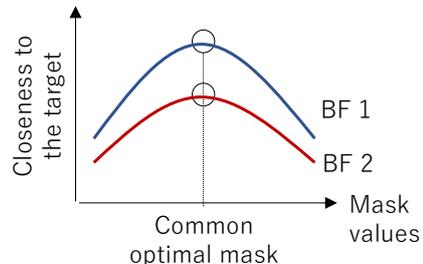


Fig. 2. Conventional ideas on the peak performance and optimal mask for multiple BF methods

- Mask-based MWF BF, which employs a single mask
- Mask-based max-SNR BF, which utilizes two masks, and its variants that employ a single mask

We do not include the MVDR BF in this study, given that the method has another issue regarding the accuracy of the estimation of the steering vector [23].

A. Signal models

In this study, all signals are in the TF domain. For simplicity, the frequency index is omitted, whereas the frame index t is adopted. Let $\mathbf{x}(t) = [x_1(t), \dots, x_N(t)]^T$ be an observation vector obtained with N microphones. The observation $\mathbf{x}(t)$ can be expressed as the following summation:

$$\mathbf{x}(t) = \mathbf{s}(t) + \mathbf{n}(t), \quad (1)$$

where $\mathbf{s}(t) = [s_1(1), \dots, s_N(t)]^T$ denotes the components arriving from the target and $\mathbf{n}(t) = [n_1(1), \dots, n_N(t)]^T$ represents the residuals called interferences. Using the observation $\mathbf{x}(t)$ and BF filter \mathbf{w} , the BF output $y(t)$ is expressed as

$$y(t) = \mathbf{w}^H \mathbf{x}(t). \quad (2)$$

We use the following three covariance matrices:

$$\Phi_x = \left\langle \mathbf{x}(t)\mathbf{x}(t)^H \right\rangle_t, \quad (3)$$

$$\hat{\Phi}_s = \left\langle m_s(t)\mathbf{x}(t)\mathbf{x}(t)^H \right\rangle_t, \quad (4)$$

$$\hat{\Phi}_n = \left\langle m_n(t)\mathbf{x}(t)\mathbf{x}(t)^H \right\rangle_t, \quad (5)$$

where $m_s(t)$ and $m_n(t)$ denote TF masks for the target and interferences, respectively, and $\langle \cdot \rangle_t$ computes the average over t . Each mask comprises nonnegative real values. We

refer to these matrices as observation, target, and interference covariance matrices, respectively. Unlike Φ_x , both $\hat{\Phi}_s$ and $\hat{\Phi}_n$ are estimated matrices computed from the masks and observations without using $s(t)$ and $n(t)$.

Moreover, consider $\text{GEV}_{\max}(\mathbf{A}, \mathbf{B})$ and $\text{GEV}_{\min}(\mathbf{A}, \mathbf{B})$ to be the eigenvectors corresponding to the maximum and minimum eigenvalues in the following GEV problem, respectively:

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}. \quad (6)$$

B. Mask-based and ideal MWF BFs

The MWF BF is formulated as a problem of minimizing the mean square error (MSE) between the BF output and the corresponding reference $p(t)$ [14]:

$$\mathbf{w} = \arg \min_{\mathbf{w}} \left\langle |p(t) - \mathbf{w}^H \mathbf{x}(t)|^2 \right\rangle_t \quad (7)$$

$$= \Phi_x^{-1} \left\langle \mathbf{x}(t) \overline{p(t)} \right\rangle_t, \quad (8)$$

where $\overline{p(t)}$ denotes the conjugate of $p(t)$. The mask-based MWF BF employs the masked observation as the reference, and thus, corresponds to the case $p(t) = m_s(t)x_k(t)$ as

$$\mathbf{w}_{\text{mwf}} = \arg \min_{\mathbf{w}} \left\langle |m_s(t)x_k(t) - \mathbf{w}^H \mathbf{x}(t)|^2 \right\rangle_t \quad (9)$$

$$= \Phi_x^{-1} \left\langle m_s(t) \mathbf{x}(t) \overline{x_k(t)} \right\rangle_t, \quad (10)$$

where $x_k(t)$ denotes the observation obtained with the k th microphone.

Another significant variant is the ideal MWF BF, which provides the upper-bound extraction performance for all BFs [24]. When $s(t)$ in (1) is known, the upper-bound extraction performance can be achieved using the MWF with $p(t) = s_k(t)$. The corresponding filter is obtained as

$$\mathbf{w}_{\text{ideal}} = \arg \min_{\mathbf{w}} \left\langle |s_k(t) - \mathbf{w}^H \mathbf{x}(t)|^2 \right\rangle_t \quad (11)$$

$$= \Phi_x^{-1} \left\langle \mathbf{x}(t) \overline{s_k(t)} \right\rangle_t. \quad (12)$$

Note that the ideal MWF BF is not a particular case of the mask-based one. This is because a mask value $m_s(t)$ constrained to be real-valued and nonnegative cannot render (9) equivalent to (11), whereas $m_s(t)$ that can take any complex values can. Therefore, it is not evident whether the mask-based MWF BF can achieve the same extraction performance as the ideal alternative.

C. Max-SNR BF and its variants

The mask-based max-SNR BF is formulated as the following maximization problem [4], [5]:

$$\mathbf{w}_{\text{snr}} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^H \hat{\Phi}_s \mathbf{w}}{\mathbf{w}^H \hat{\Phi}_n \mathbf{w}} \quad (13)$$

$$= \text{GEV}_{\max}(\hat{\Phi}_s, \hat{\Phi}_n). \quad (14)$$

Although this method originally utilizes two masks, we can derive two different variants that adopt a single mask by assuming the following relationship:

$$\hat{\Phi}_s + \hat{\Phi}_n = \Phi_x. \quad (15)$$

One variant utilizes a mask for the interferences [17], [25]:

$$\mathbf{w}_{\text{nor}} = \arg \min_{\mathbf{w}} \frac{\mathbf{w}^H \hat{\Phi}_n \mathbf{w}}{\mathbf{w}^H \Phi_x \mathbf{w}} \quad (16)$$

$$= \text{GEV}_{\min}(\hat{\Phi}_n, \Phi_x). \quad (17)$$

We refer to this as the minimum noise-to-observation ratio (min-NOR) BF. Similarly, we can derive the other, which employs a mask for the target and is expressed as

$$\mathbf{w}_{\text{sor}} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^H \hat{\Phi}_s \mathbf{w}}{\mathbf{w}^H \Phi_x \mathbf{w}} \quad (18)$$

$$= \text{GEV}_{\max}(\hat{\Phi}_s, \Phi_x). \quad (19)$$

We refer to this as the maximum signal-to-observation ratio (max-SOR) BF.

Furthermore, even without assuming (15), the min-NOR and max-SOR BFs can be equivalent to each other in terms of the filter estimation if one mask is computed from the other using (20).

$$m_s(t) + m_n(t) = \alpha, \quad (20)$$

where α is a nonnegative value, such that all the mask values are nonnegative. For example, assigning $m_s(t) = \alpha - m_n(t)$ to (18) and (4) results in the problem represented by (16) if $m_n(t)$ is nonnegative for all t .

For the max-SNR BF and its variants, the scales of the filter and BF outputs are uncertain, unlike the MWF BF. Hence, a post-process for determining the proper scales is required [16], [17]. Considering that the formulation of these BFs differs from that of the MWF BF, it is not evident whether these BFs can achieve the same extraction performance as the ideal MWF BF.

IV. OBTAINING THE OPTIMAL MASK

To explore the peak performance for each BF method, this study employs a bottom-up approach that obtains the optimal mask for each mixture of the target and interferences, instead of a priori deciding whether a particular mask type is optimal.

Let \mathcal{M} be a set of mask values adopted in a BF. This set comprises $m_s(t)$, $m_n(t)$, or both for all t , depending on the BF method employed. We can formulate \mathcal{M} as the solution to the problem of minimizing the following MSE:

$$\mathcal{M} = \arg \min_{\mathcal{M}} \left\langle |s_k(t) - y(t)|^2 \right\rangle_t, \quad (21)$$

where $s_k(t)$ and $y(t)$ denote the target included in the observation of the k th microphone and BF output, respectively. To render the mask values nonnegative and to avoid both diverging the mask values and converging them to zero, we impose the following constraints on (21):

$$m(t) \geq 0 \quad \text{for all } t, \quad (22)$$

$$\left\langle m(t)^2 \right\rangle_t = 1, \quad (23)$$

where $m(t)$ denotes $m_s(t)$ or $m_n(t)$. In (21), $y(t)$ is computed as follows. First, the BF filter $\mathbf{w}(t)$ is computed depending on the adopted BF method, then the $y(t)$ is computed using (2).

To determine the ideal scale of $y(t)$ independent of the BF method, we apply the following post-filtering process, referred to as the *ideal scaling*:

$$\gamma = \frac{\langle s_k(t) \overline{y(t)} \rangle_t}{\langle |y(t)|^2 \rangle_t}, \quad (24)$$

$$y(t) \leftarrow \gamma y(t). \quad (25)$$

Owing to the ideal scaling, the constraint represented as (23) does not affect the scale of $y(t)$.

Because \mathcal{M} , $w(t)$, and $y(t)$ depend on each other, \mathcal{M} cannot be obtained explicitly. In contrast, we employ the iterative algorithm based on gradient descent. Given that the mask estimation process adopts no explicit association between the masks and sources, the obtained masks such as $m_s(t)$ and $m_n(t)$ do not necessarily correspond to the sources such as the target and interferences.

Equations (21), (22), and (23) might seem to solve the same problem as the ideal MWF BF, which is represented as (11), based on the iterative algorithm and consequently achieve the same extraction performance as the ideal MWF BF. However, such a perspective is incorrect, as mentioned in Sections III-B and III-C. Rather, the objective of this study is to examine to what extent the output of the mask-based BF can approach that of the ideal MWF BF if the mask best fits (21).

V. EXPERIMENTS

To clarify Issues 1–3 mentioned in Section III, we conducted the following experiments:

- 1) Exploring the peak performance for each BF method
- 2) Verifying whether the optimal mask is common for all the BF methods
- 3) Examining whether the IRM can achieve the peak performance

The following subsections first mention the dataset and system for these experiments, and then demonstrate the experimental results in order.

A. Dataset and experimental system

We employed the CHiME-3 simulated test set [26], which comprises both 330 utterances from four speakers and four background (BG) noises. In this dataset, sound data were recorded at 16 kHz with six microphones attached to a tablet device. We generated the TF domain signals using the short-time Fourier transform with window and shift lengths of 1024 and 256, respectively.

Fig. 3 illustrates the experimental system. The modules labeled *Absolute value*, *Normalization*, and *Ideal scaling* correspond to (22), (23), and (25), respectively. The observation data were generated by mixing clean speech and BG noise. To represent multiple scenarios in different SNRs, three multipliers such as 1, 2, and 4 were applied to the BG noise. We refer to these values as *BG multipliers*. The BF output $y(t)$ was generated as explained in Section IV; one or two masks were employed and the BF filter was estimated depending on the BF method, such as (10), (14), (17), and (19). Given that

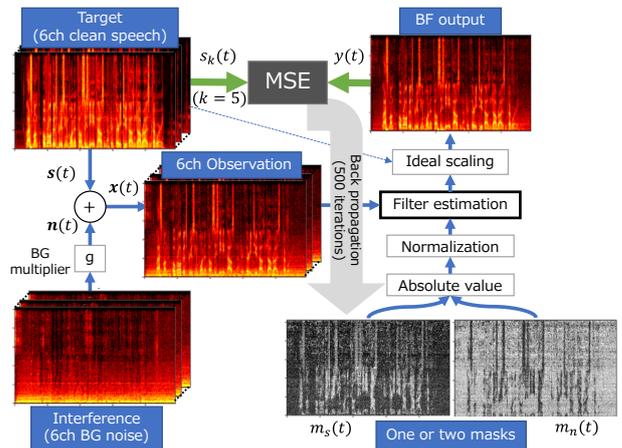


Fig. 3. Configuration of the system obtaining or employing the optimal mask for each utterance

microphone #5 is the closest to the speaker, we set k to 5 in (21) and (24). The backpropagation based on the gradient descent algorithm was utilized only in the first experiment to obtain the optimal mask for each utterance. The estimation of the BF filter and output was implemented in PyTorch [27], which supports the backpropagation of matrix operations in the complex number domain.

Table II presents average SNR scores of the observation of microphone #5 for each BG multiplier.

For the evaluation score, we adopted the source-to-distortion ratio (SDR) [28], which is provided as a performance score in this dataset, considering that this score basically represents the closeness of the BF output to the target.

B. Experiment 1: Exploring the peak performance for each BF method

First, to explore the peak performance of each BF method, we obtained the optimal mask for each utterance, which is the solution to (21) under the constraints represented as (22) and (23), using the backpropagation of 500 iterations. Table III presents SDR scores of the mask-based BFs, including those of the ideal MWF BF using $k = 5$ in (12). All the mask-based BFs practically demonstrated identical scores comparable to the ideal MWF BF. These are remarkable results, given that they were not theoretically evident, as mentioned in Sections III-B and III-C.

C. Experiment 2: Verifying whether the optimal mask is common for all the BF methods

Furthermore, to verify whether the optimal mask is common for all the BF methods, we applied the optimal mask obtained

TABLE II
SNR SCORES [dB] OF THE OBSERVATION OF MICROPHONE #5

BG multiplier		
$\times 1$	$\times 2$	$\times 4$
7.540	1.536	-4.421

in a BF method to another one. If the same score is achieved after applying the mask, we can claim that the optimal mask is the same between these two BF methods. Although 12 permutations are possible from four methods, we examined six considered to be nontrivial. In the cases that $m_s(t)$ was applied to a BF method using $m_n(t)$, and vice versa, we converted the mask values using the following rules to satisfy (20):

$$m_s(t) = \max_t \{m_n(t)\} - m_n(t), \quad (26)$$

$$m_n(t) = \max_t \{m_s(t)\} - m_s(t), \quad (27)$$

where $\max_t \{\cdot\}$ denotes the maximum value over t .

The obtained results are presented in Table IV. The first and second rows indicate that the optimal masks obtained in the max-SNR BF were applied to the max-SOR and min-NOR BFs, respectively. These two rows present lower scores than Experiment 1. This suggests that the optimal mask obtained in the max-SNR BF is not necessarily optimal for the max-SOR or min-NOR BFs.

The third row indicates that the optimal mask $m_s(t)$ obtained in the max-SOR BF was applied to the min-NOR BF using (27), and the fourth row indicates that the mask $m_n(t)$ obtained in the min-NOR BF was applied to the max-SOR BF using (26). The two rows practically demonstrated the same scores as the max-SOR and min-NOR BFs in Table III. This suggests that the optimal mask can be converted between the max-SOR and min-NOR BFs, despite the fact that the two masks are not the same.

The fifth and sixth rows indicate that the optimal mask $m_s(t)$ obtained in a method was applied to the other for the max-SOR and MWF BFs. These two rows demonstrated considerably lower scores than the results of the max-SOR and MWF BFs in Table III. These results suggest that the optimal mask is not common between these two BF methods.

TABLE III
PEAK PERFORMANCE FOR EACH METHOD AND SCENARIO IN SDR [dB]
COMPARED WITH THAT OF THE IDEAL MWF.

BF Method	Mask(s) used	BG multiplier		
		×1	×2	×4
Max-SNR	m_s, m_n	19.430	14.276	9.451
Max-SOR	m_s	19.426	14.276	9.451
Min-NOR	m_n	19.434	14.276	9.451
MWF	m_s	19.438	14.268	9.430
Ideal MWF		19.441	14.276	9.451

TABLE IV
EXTRACTION PERFORMANCE IN SDR [dB] AFTER APPLYING THE
OPTIMAL MASK TO ANOTHER BF METHOD

BF for mask estimation	Applied to	Mask used	BG multiplier		
			×1	×2	×4
Max-SNR	Max-SOR	m_s	17.084	13.493	9.165
	Min-NOR	m_n	18.399	14.032	9.381
Max-SOR	Min-NOR	m_s & (27)	19.426	14.276	9.451
Min-NOR	Max-SOR	m_n & (26)	19.434	14.274	9.451
Max-SOR	MWF	m_s	15.465	10.259	5.337
MWF	Max-SOR	m_s	14.230	12.092	8.458

D. Experiment 3: Examining whether the IRM can achieve the peak performance

The third experiment is for examining if the IRM can achieve the peak performance for each BF method. Considering that the term IRM was ambiguously employed in the related studies [6], [12], [20]–[22], we define several masks that utilize the ratios of the target and interferences based on [12].

The IRMs for the target and interferences are defined as

$$m_s(t) = \left(\frac{|s_k(t)|^2}{|s_k(t)|^2 + |n_k(t)|^2} \right)^\beta, \quad (28)$$

$$m_n(t) = \left(\frac{|n_k(t)|^2}{|s_k(t)|^2 + |n_k(t)|^2} \right)^\beta, \quad (29)$$

where $k = 5$ and $\beta = 1$ or 0.5. As another type of ratio mask, we employed the spectral magnitude masks (SMMs) defined as

$$m_s(t) = \frac{|s_k(t)|}{|s_k(t) + n_k(t)|}, \quad (30)$$

$$m_n(t) = \frac{|n_k(t)|}{|s_k(t) + n_k(t)|}. \quad (31)$$

In the single-channel TF masking, these masks are ideal in terms of the magnitudes of the target and interferences [12].

The obtained results are presented in Table V. In this table, the combination of the max-SOR and IRM with $\beta = 0.5$ demonstrated the best score for all the scenarios. However, these scores were lower than that in Experiment 1. The results suggest that the IRM does not achieve the peak performance for any BF method examined in this study; hence, it differs from the optimal mask for each method.

In addition, Table V indicates that the max-SNR, max-SOR, and min-NOR BFs present the same scores for the IRM with $\beta = 1$. This is because this mask type evidently satisfies both (15) and (20). Hence, these three BFs are equivalent in this case, as mentioned in Section III-C.

TABLE V
EXTRACTION PERFORMANCE IN SDR [dB] WHEN THE IRM ($\beta = 1, 0.5$)
AND SMM WERE EMPLOYED FOR EACH BF METHOD.

BF method	Type of ideal mask	Mask(s) used	BG multiplier		
			×1	×2	×4
Max-SNR	IRM ($\beta = 1$)	m_s, m_n	18.642	13.889	9.226
	IRM ($\beta = 0.5$)		18.313	13.790	9.203
	SMM		17.973	13.529	9.060
Max-SOR	IRM ($\beta = 1$)	m_s	18.642	13.889	9.226
	IRM ($\beta = 0.5$)		18.725	13.948	9.275
	SMM		13.640	11.447	8.147
Min-NOR	IRM ($\beta = 1$)	m_n	18.642	13.889	9.226
	IRM ($\beta = 0.5$)		18.267	13.747	9.160
	SMM		17.372	12.735	8.103
MWF	IRM ($\beta = 1$)	m_s	17.316	12.788	8.375
	IRM ($\beta = 0.5$)		16.228	11.799	7.477
	SMM		17.109	12.316	7.734

VI. DISCUSSION

In this section, we discuss the experimental results for each issue mentioned in Section III.

A. Discussion on Issue 1

The results of Experiment 1 provide the answer to Issue 1. The peak performance of each BF method is practically identical and comparable to the upper bound given by the ideal MWF BF. These findings differ from the conventional idea presented in Fig. 2 and may deprive the meaning of the discussion on the BF that can achieve the best extraction performance. However, it is an open question whether these findings are applicable to any BF method and dataset. Therefore, further investigation is required.

B. Discussion on Issue 2

The results of Experiment 2 provide the answer to Issue 2. Although the optimal mask is not common, it depends on the BF method adopted. This finding differs from the conventional concept presented in Fig. 2, similar to the discussion on Issue 1. However, we consider this to be natural because the optimal mask is formulated as the solution to a different problem for each BF method, as explained in Section IV.

In addition, Table IV indicates several remarkable points. Although both the max-SOR and min-NOR BFs are derived from the max-SNR BF, as mentioned in Section III-C, the top two rows suggest that $m_s(t)$ and $m_n(t)$ are optimal for the max-SNR BF, while not for the max-SOR or min-NOR BFs. We consider that the reason for this is that the optimal masks obtained in the max-SNR BF do not satisfy (15) or (20). Hence, the max-SNR BF is not equivalent to the other two BFs in this case. In contrast, the third and fourth rows in this table suggest that the max-SOR and min-NOR BFs are equivalent if both $m_s(t)$ and $m_n(t)$ satisfy (20). Moreover, the bottom two rows exhibited lower scores than the other rows. Although $m_s(t)$ has commonly been interpreted as the mask for the target, the facts suggest that the optimal mask $m_s(t)$ for the max-SOR BF differs significantly from that of the MWF BF and vice versa.

From the discussion on Issues 1 and 2, we have obtained a novel idea for the peak performance and optimal mask among multiple BF methods, as illustrated in Fig. 4. This indicates that the optimal mask depends on the BF method adopted, while the peak performance is the same among multiple BF methods and comparable to the upper bound achieved by the ideal MWF.

C. Discussion on Issue 3

The results of Experiment 3 provide the answer to Issue 3. Although these masks have been considered to be optimal for any BF method, Table V suggests that the IRM and SMM do not achieve the upper-bound performance and differ from the optimal mask for any BF method. These findings can explain why the studies presented in Table I mentioned a different BF method as the best. Table V indicates that the BF method that appears the best depends on the mask type employed. For

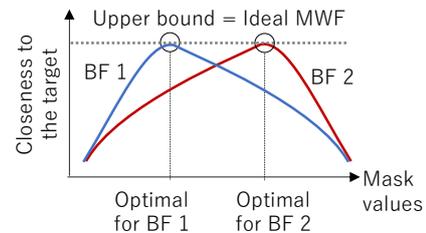


Fig. 4. Obtained concept of the peak performance and optimal mask among multiple BF methods

example, focusing on the IRM with $\beta = 0.5$ in this table, we can determine that the max-SOR BF performs the best. However, this result only suggests that this mask type is the closest to the optimal mask for the max-SOR in this dataset.

Furthermore, the fact that the IRM is not optimal imposes another issue on us. This is how the optimal mask is interpreted and represented as a formula. However, this is also an open question.

VII. CONCLUSIONS

In this study, we investigated mask-based BFs such as the max-SNR, max-SOR, min-NOR, and MWF BFs. To explore the peak performance for each BF method, we obtained the optimal mask for each utterance by minimizing the MSE between the BF output and target. We experimentally verified that these four methods have the same peak performance as the upper bound provided by the ideal MWF BF. Via additional experiments that applied the optimal mask across BF methods, we determined that the optimal mask differed for the BF method used. However, the mask values can be converted between the max-SOR and min-NOR BFs. These findings differed from the conventional idea that the optimal mask would be common and the peak performance would depend on the BF method. We verified that the IRM did not achieve the peak performance for these four BFs. Hence, this mask type was not optimal. We expect that these findings will contribute to the improvement of mask-based BFs.

Given that these findings are currently experimental, in the future, we would attempt to establish their theoretical background and investigate whether these apply to other BF methods and datasets.

The experimental system has been shared in https://github.com/hreshare/optimal_beamformers/.

REFERENCES

- [1] S. J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Sept, pp. 1571–1575, 2018.
- [2] M. Mizumachi and M. Origuchi, "Advanced delay-and-sum beamformer with deep neural network," *22nd International Congress on Acoustics (ICA)*, 2016.

- [3] M. Mizumachi, "Neural network-based broadband beamformer with less distortion," *Proceedings of International Congress on Acoustics (ICA 2019)*, p. 2760, Sep. 2019.
- [4] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3RD CHiME challenge," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*, no. June 2016, pp. 444–451, 2016.
- [5] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.
- [6] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 1981–1985, 2016.
- [7] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised training of neural mask-based beamforming," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, International Speech Communication Association, 2019, pp. 1253–1257.
- [8] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.
- [9] S. Stenzel, T. C. Lawin-Ore, J. Freudenberger, and S. Doclo, "A multichannel wiener filter with partial equalization for distributed microphones," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, unknown, Oct. 2013, pp. 1–4.
- [10] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.
- [11] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 66–70.
- [12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," en, *IEEE/ACM Trans Audio Speech Lang Process*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
- [13] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural Mask-Based beamforming for Far-Field speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6697–6701.
- [14] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments," *Comput. Speech Lang.*, vol. 49, pp. 37–51, May 2018.
- [15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," en, *IEEE/ACM Trans Audio Speech Lang Process*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
- [16] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 1, 2007, pp. I–41–I–44.
- [17] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.
- [18] J. Heymann, M. Bacchiani, and T. N. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 6722–6726.
- [19] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5734–5738.
- [20] Y.-H. Tu, J. Du, L. Sun, F. Ma, and C.-H. Lee, "On design of robust deep models for CHiME-4 multi-channel speech recognition with multiple configurations of array microphones," in *Interspeech 2017*, ISCA: ISCA, Aug. 2017.
- [21] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on Teacher–Student deep learning using improved speech presence probability for Noise-Robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, Dec. 2019.
- [22] S. Chakrabarty and E. A. P. Habets, "Time–Frequency masking based online Multi-Channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 4, pp. 787–799, Aug. 2019.
- [23] Y. H. Tu, J. Du, L. Sun, and C. H. Lee, "LSTM-based iterative mask estimation and post-processing for multichannel speech enhancement," *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, vol. 2018-Febru, no. December, pp. 488–491, 2018.
- [24] J. Malek, Z. Koldovský, and M. Bohac, "Block-online multi-channel speech enhancement using deep neural

- network-supported relative transfer function estimates,” *IET Signal Proc.*, vol. 14, no. 3, pp. 124–133, May 2020.
- [25] A. Hiroe, “Similarity-and-Independence-Aware beamformer with iterative casting and boost start for target source extraction using reference,” *IEEE Open Journal of Signal Processing*, vol. 3, pp. 1–20, 2022.
- [26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Analysis and outcomes,” *Comput. Speech Lang.*, 2017.
- [27] A. Paszke, S. Gross, F. Massa, *et al.*, “PyTorch: An imperative style, High-Performance deep learning library,” in *Advances in Neural Information Processing Systems* 32, Curran Associates, Inc., 2019, pp. 8024–8035.
- [28] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.