A Study on Reference Microphone Selection for Multi-Microphone Speech Enhancement

Jie Zhang, Huawei Chen, Li-Rong Dai, and Richard C. Hendriks

***Abstract*—Multi-microphone speech enhancement methods typically require a reference position with respect to which the target signal is estimated. Often, this reference position is arbitrarily chosen as one of the reference microphones. However, it has been shown that the choice of the reference microphone can have a significant impact on the final noise reduction performance. In this paper, we therefore theoretically analyze the impact of selecting a reference on the noise reduction performance with near-end noise being taken into account. Following the generalized eigenvalue decomposition (GEVD) based optimal variable span filtering framework, we find that for any linear beamformer, the output signal-to-noise ratio (SNR) taking both the near-end and far-end noise into account is reference dependent. Only when the near-end noise is neglected, the output SNR of rank-1 beamformers does not depend on the reference position. However, in general for rank-r beamformers with $r > 1$ (e.g., the multichannel Wiener filter) the performance does depend on the reference position. Based on these, we propose an optimal algorithm for microphone reference selection that maximizes the output SNR. In addition, we propose a lower-complexity algorithm that is still optimal for rank-1 beamformers, but suboptimal for the general $r > 1$ rank beamformers. Experiments using a simulated microphone array validate the effectiveness of both proposed methods and show that in terms of quality, several dB can be gained by selecting the proper reference microphone.**

***Index Terms*—Speech enhancement, multi-channel beamforming, reference microphone, relative acoustic transfer function, variable span linear filters, low-rank approximation.**

I. Introduction

DURING the last few decades, speech enhancement and noise reduction have become widely used in numerous applications. Usually, it is employed as a front-end step to improve the speech quality and speech intelligibility in audio processing scenarios, like speech recognition [1], binaural hearing aids (HAs) [2], teleconferencing systems [3], source localization [4] and mobile robot systems [5]. These applications use both single-microphone algorithms [6]–[8] and multi-microphone algorithms [9]–[12]. Compared to single-microphone noise reduction algorithms, in which only temporal (spectral) information is exploited, the multi-microphone counterpart (e.g., beamforming) generally leads to a better noise reduction performance, as both temporal and spatial information can be used.

The multi-microphone noise reduction methods can be classified into 1) linearly constrained beamforming [9], [10], [13] and 2) unconstrained beamforming [14]–[16]. Two well-known linearly constrained approaches are the linearly constrained minimum variance (LCMV) beamformer and the minimum variance distortionless response (MVDR) beamformer [10], [13]. Both are designed to minimize the output signal variance. The LCMV beamformer can take a set of linear constraints into account, while the MVDR beamformer only includes a single linear constraint to guarantee an undistorted target signal. Therefore, the MVDR beamformer can be viewed as a special case of the LCMV beamformer. Unconstrained beamforming, e.g., the multi-microphone Wiener filter (MWF) based algorithms, aim at minimizing the mean square-error (MSE) between the target signal at a reference position (typically at one of the reference microphones) and the estimated target signal at the same reference position. The MWF distorts the target signal inevitably, since no distortionless constraints are taken into account. In order to alleviate this drawback, one can add a constraint to the MWF to control the signal distortion level, leading to the speech distortion weighted MWF (SDW-MWF) [16], which can then trade-off the noise reduction capability and the signal distortion level.

Both the linearly-constrained and unconstrained beamformers require a reference position with respect to which the target signal is estimated. This could be the original source location, in which case, the beamformers become dependent on the acoustic transfer function (ATF) of the desired source from the original location to the microphones. However, often the reference position is chosen as one of the microphones, which turns the ATF into a relative acoustic transfer function (RTF). It is known that under specific conditions, the beamforming performance is not influenced by the chosen reference microphone [17]–[19]. It is known that this holds when the target source correlation matrix has rank one and the performance is measured using the output signal-to-noise ratio (SNR) defined as the ratio between the variance of the estimated target at the output of the beamformer and the variance of the processed far-end noise, i.e., the noise in the beamformer output [19].

Manuscript received xxxxx xx, 2020; revised xxxxx xx, 2020; accepted xxxxx xx, 2020. Date of publication xxxxx xx, 2020; date of current version xxxxx xx, 2020. This work is supported by National Key R&D Program of China (2017YFB1002202), the Leading Plan of CAS (XDC08010200), and the Key Science&Technology Project of Anhui Province (18030901016). The work of H. Chen is supported by National Natural Science Foundation of China under Grant No. 61971219. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. ********. (*Corresponding author: Jie Zhang*)

J. Zhang and L.-R. Dai are with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China (USTC), 230026, Hefei, China. (e-mail: j.zhang6@ustc.edu.cn; lrdai@ustc.edu.cn)

H. Chen is with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, 210016 Nanjing, China. (e-mail: hwchen@nuaa.edu.cn)

R. C. Hendriks is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands. (e-mail: r.c.hendriks@tudelft.nl)

Digital Object Identifier: **************

However, in practice, it turns out that the chosen reference microphone does influence the final performance of certain beamformers [20]. This depends on the type of beamformers, the rank of the estimated target correlation matrix and also on the performance metric that is used.

In order to increase the beamformer performance in practice, it is thus of relevance to understand the exact relation between the chosen reference microphone and the final performance. As a performance metric to optimize, we will constrain ourselves in this work to the output SNR. However, we will extend this by including also near-end noise to demonstrate the impact of reference microphone selection for more general performance metrics than the conventional output SNR. In the case of more conventional microphone arrays, the impact of choosing a reference microphone might be small [20], due to the fact that the microphones are usually spatially close. In the case of distributed microphone arrays (i.e., a wireless acoustic sensor network (WASN)), reference microphone selection can have a more severe influence on the performance [21], due to the larger spatial diversity. For instance, it was experimentally shown in [22] that choosing different reference microphones heavily affects the speech recognition accuracy (e.g., word error rates) in meeting recognition scenarios using a distributed microphone array. In [20], an approach was proposed to select the optimal reference microphone for the MWF. This method was refered to as $\mathrm{maxoSNR}$. However, this method requires to evaluate the performance of all M (the number of microphones) filters. To overcome this drawback, several sub-optimal but more practical methods were suggested in [20], including choosing the microphone that has the highest input SNR ($\mathrm{maxiSNR}$), selecting the one that is closest to the target source ($\mathrm{minDist}$) and using the microphone that has the largest input power ($\mathrm{maxEnergy}$).

Prior to presenting an improved method for reference microphone selection, we study in this work first more systematically the dependence of the output SNR on the microphone reference. To do so, we consider an extended version of the output SNR by also including the near-end noise. We will show that in general, the beamformer performance in terms of the output SNR always depends on the selected microphone reference. In addition, we show that even when the near-end noise can be neglected, the performance of general rank-r ($r > 1$) beamformers in terms of the output SNR still depends on the chosen reference microphone. Only when we consider rank-1 beamformers (e.g., the MVDR beamformer) without near-end noise, it indeed follows as already known from [19] that the output SNR is microphone reference independent. As the more general case of rank-r ($r > 1$) beamformers (e.g., the MWF) with near-end noise resembles the practical situation, it is of relevance to understand how to choose a proper reference microphone. We will show that dependent on the exact setup, the loss in performance by not selecting the optimal reference microphone can be in the order of several dB. Based on the theoretical foundings, we propose an optimal reference microphone selection approach by maximizing the output SNR of general rank-r beamformers, which is in line with the selection criterion proposed in [20] and is referred to as $\mathrm{maxoSNR}$. Instead of verifying the beamformers for all possible reference microphones, we demonstrate that the optimal reference microphone can be determined by checking the diagonal elements of two matrices, which are constructed by the generalized eigenvalues and eigenvectors of the noise and noisy correlation matrices. In addition, we present an alternative selection criterion by considering a semi-definite programming problem. Furthermore, we show that given the principal eigenvector, which is basically equivalent to the RTF in the case of a single target source, searching for its maximum absolute value gives a sub-optimal solution for the reference microphone selection. We refer this method to as maxRTF. Compared to the initial $\mathrm{maxoSNR}$ method in [20], both the proposed $\mathrm{maxoSNR}$ and maxRTF methods do not require to evaluate all possible M filters. As the proposed $\mathrm{maxoSNR}$ and [20] use the same problem formulation and achieve the same solution, but differ in solvers, we will stick to the same name in this work. In order to validate the proposed approach, we conduct experiments using a simulated microphone array. It is shown that the proposed $\mathrm{maxoSNR}$ method improves the output SNR against other (sub-optimal) strategies or naive (random) selection, without the need to evaluate the performance of all M possible filters.

The rest of this paper is structured as follows. Section II presents the required fundamental knowledge. In Section III, we summarize the MMSE-based optimal variable span filters. In Section IV, we theoretically analyze the impact of the signal rank and the reference microphone on the performance of MMSE beamformers in terms of output SNR. In Section V, we propose two reference microphone selection approaches. The proposed algorithms are validated in Section VI via numerical simulations. Finally, Section VII concludes this work.

II. Fundamentals

A. Signal model

In this work, we consider an array of M microphones. These could be part of a conventional microphone array, or, a distributed WASN. Let i and k denote the time-frame index and the frequency-bin index, respectively, in the short-time Fourier transform (STFT) domain. Assuming an additive signal model, the acoustic signal at the mth microphone is then given by

$$\begin{aligned} Y_m(i,k) &= X_m(i,k) + N_m(i,k) \\ &= a_m(k)S(i,k) + N_m(i,k), \end{aligned} \tag{1}$$

with

- $X_m(i,k)$ the target source STFT coefficient received by microphone m;
- $N_m(i,k)$ the noise STFT coefficient at the mth microphone, which might include the coherent noise (e.g., interference, reverberation) and incoherent noise (e.g., sensor self noise);
- $a_m(k)$ the ATF from the target position to the mth microphone[1];

[1] In this work, we assume that the single target source keeps static during the observation time period of interest, as tracking or estimating dynamic source(s) is beyond the scope of this paper. Under this assumption, the ATF or RTF of the target source with respect to the microphone array is time-invariant, i.e., only frequency-dependent.

- $S(i,k)$ the target source STFT coefficient at the source position.

Often, instead of the ATF, the RTF is used. This is due to the fact that the ATF is a scaled version of the RTF and the scaling factor is hard to determine, while the RTF can be estimated using e.g., [23]–[27]. The RTF is defined as the normalized ATF with respect to an arbitrarily chosen reference microphone n, given by

$$h_{n,m}(k) = a_m(k)/a_n(k), \tag{2}$$

which can be estimated using the covariance subtraction or covariance whitening method [23]–[25]. Clearly, when $n = m$, $h_{n,m}(k) = 1$. With the RTF, the signal model in (1) can be written as

$$Y_m(i,k) = h_{n,m}(k)X_n(i,k) + N_m(i,k). \tag{3}$$

For notational brevity, we will omit the time-frequency indices (i,k) in the sequel bearing in mind that all the operations take place in the STFT domain. Using vector notation, the signal model can be written as

$$\begin{aligned}\mathbf{y} &= \mathbf{x} + \mathbf{n}\\ &= \mathbf{a}S + \mathbf{n}\\ &= \mathbf{h}_n X_n + \mathbf{n},\end{aligned} \tag{4}$$

where

$$\begin{aligned}\mathbf{y} &= [Y_1(i,k), Y_2(i,k), \ldots, Y_M(i,k)]^T,\\ \mathbf{x} &= [X_1(i,k), X_2(i,k), \ldots, X_M(i,k)]^T,\\ \mathbf{n} &= [N_1(i,k), N_2(i,k), \ldots, N_M(i,k)]^T,\\ \mathbf{a} &= [a_1(k), a_2(k), \ldots, a_M(k)]^T,\\ \mathbf{h}_n &= [h_{n,1}(f), h_{n,2}(f), \ldots, h_{n,M}(f)]^T,\end{aligned}$$

where $(\cdot)^T$ denotes the matrix/vector transpose.

B. Second-order statistics

Assuming that the target source and the noise components are mutually uncorrelated, we can formulate the correlation matrix of the microphone measurements as

$$\begin{aligned}\mathbf{\Phi_{yy}} &= \mathbb{E}\left\{\mathbf{y}\mathbf{y}^H\right\}\\ &= \mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\} + \mathbb{E}\left\{\mathbf{n}\mathbf{n}^H\right\}\\ &= \mathbf{\Phi_{xx}} + \mathbf{\Phi_{nn}},\end{aligned} \tag{5}$$

where $\mathbf{\Phi_{xx}}$ and $\mathbf{\Phi_{nn}}$ denote the correlation matrix of the signal component and the correlation matrix of the noise components, respectively, and $\mathbb{E}\{\cdot\}$ denotes mathematical expectation, and $(\cdot)^H$ the matrix/vector complex conjugate transpose. For the single target source case, $\mathbf{\Phi_{xx}}$ is a rank-1 matrix in theory, since by definition we have

$$\begin{aligned}\mathbf{\Phi_{xx}} &= \mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\}\\ &\triangleq \sigma_S^2\mathbf{a}\mathbf{a}^H \triangleq \sigma_{X_n}^2\mathbf{h}_n\mathbf{h}_n^H,\end{aligned} \tag{6}$$

where $\sigma_S^2 = \mathbb{E}\left\{|S|^2\right\}$ and $\sigma_{X_n}^2 = \mathbb{E}\left\{|X_n|^2\right\}$ denote the power spectral density (PSD) of the target source and the PSD of the signal component at the reference microphone n, respectively. However, in practice the correlation matrices $\mathbf{\Phi_{yy}}$, $\mathbf{\Phi_{nn}}$ and $\mathbf{\Phi_{xx}}$ are unknown and have to be estimated. For example, $\mathbf{\Phi_{yy}}$ can be estimated from the noisy data, $\mathbf{\Phi_{nn}}$ from the noise-only data using a voice activity detector (VAD), and $\mathbf{\Phi_{xx}}$ by subtracting the estimated $\mathbf{\Phi_{nn}}$ from $\mathbf{\Phi_{yy}}$, i.e.,

$$\hat{\mathbf{\Phi}}_{\mathbf{xx}} = \hat{\mathbf{\Phi}}_{\mathbf{yy}} - \hat{\mathbf{\Phi}}_{\mathbf{nn}}. \tag{7}$$

Due to inevitable estimation errors, the estimated correlation matrix $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$ will hardly ever be rank one, even when $\mathbf{\Phi_{xx}}$ is rank one. For that reason, we consider in the theoretical analysis of optimal reference microphone selection the case where $\mathbf{\Phi_{xx}}$ has in general rank $r \geq 1$ for rank-r approximating beamformers.

C. Problem formulation and existing approaches

For the multi-microphone noise reduction problem, the key step is designing a frequency-dependent spatial filter $\mathbf{w} = [w_1, w_2, \ldots, w_M]^T$. With such a spatial filter, the estimated speech signal can be obtained as

$$\hat{S} = \mathbf{w}^H\mathbf{y}. \tag{8}$$

The SNR after beamforming, i.e., the output SNR, is given by

$$\mathrm{oSNR}(k) = \frac{\mathbf{w}^H\mathbf{\Phi_{xx}}\mathbf{w}}{\mathbf{w}^H\mathbf{\Phi_{nn}}\mathbf{w}}, \tag{9}$$

where the denominator only contains the output noise of the beamformer, i.e., the far-end noise. In our analysis in Section IV, we will extend this definition with near-end noise, as this resembles the realistic practical setup and will be shown to significantly influence the reference microphone selection. In case $\mathbf{\Phi_{xx}}$ truly has rank $r = 1$, it is known that the output SNR as defined in (9) is microphone reference independent. However, in practice when the estimate of $\mathbf{\Phi_{xx}}$ has rank $r > 1$, the output SNR turns out to be reference dependent for general rank-r beamformers like the MWF. The most intuitive criterion of reference microphone selection is by maximizing the (measured) output SNR [20]. Suppose that the mth microphone is selected as the reference microphone. Let the corresponding spatial filter be denoted by $\mathbf{w}_m$ and the resulting output SNR by oSNR_m. The optimal reference microphone selection in the sense of maximizing the output SNR can be formulated as the following $\mathrm{maxoSNR}$ optimization problem:

$$n_k = \arg\max_m \mathrm{oSNR}_m(k). \tag{10}$$

In [20], this optimization problem (10) is solved via an exhaustive search, i.e., designing M filters and evaluating the output SNR of each filter. The exhaustive search might be problematic due to the time complexity in designing all M filters, particularly when M is large, e.g., in WASNs.

As the original $\mathrm{maxoSNR}$ requires to examine the performance of all filters, several sub-optimal low-complexity approaches were also introduced in [20].

1) $\mathrm{maxiSNR}$*:* Instead of selecting the reference based on the output SNR, it was proposed in [20] to perform the selection based on the input SNR. In this case, the reference is selected as

$$n_k = \arg\max_m \ \mathrm{iSNR}_m(k), \tag{11}$$

with the frequency-dependent input SNR defined as

$$\mathrm{iSNR}_m(k) = \frac{\sum_i |X_m(i,k)|^2}{\sum_i |N_m(i,k)|^2}. \tag{12}$$

Notice that this selection mechanism does not include the filter $\mathbf{w}_k$ and leads thus to a sub-optimal solution.

2) $\mathrm{minDist}$*:* The input SNR and the signal PSD $\sigma^2_{X_m}$ are directly related to the distance between the target source position and the microphone. The closer a microphone to the target source, the larger input SNR it obtains. An alternative sub-optimal reference selection method was therefore presented where the microphone that is closest to the target source is chosen as the reference microphone. Clearly, $\mathrm{minDist}$ depends on the source localization and microphone calibration results.

3) $\mathrm{maxEnergy}$*:* Another sub-optimal selection procedure introduced in [20] is based on choosing the microphone that has the maximum input power, i.e.,

$$n_k = \arg\max_m \sum_i |Y_m(i,k)|^2, \tag{13}$$

since in case the noise sources are far away from the microphones or the input SNRs are high, the input power is dominated by the speech component. Note that $\mathrm{maxEnergy}$ might lose validity if the noise source is close to the microphones.

Notice that the $\mathrm{maxoSNR}$, $\mathrm{maxiSNR}$ and $\mathrm{maxEnergy}$ are frequency-dependent, and thus might select a different reference at different frequency bins (i.e., *soft selection*), while $\mathrm{minDist}$ employs a *hard selection*. In this work, in order to avoid an exhaustive search, we will theoretically analyze the impact of reference microphone selection on the performance and then propose a low-complexity approach.

III. Optimal beamformer design

To guide the reader, we summarize in this section the work on optimal variable span linear filters presented in [19] and also based on the work in [28], [29]. We will use these variable span linear filters in Section IV to get more understanding on the relation between the optimal reference, the output SNR and the rank of the (estimated) correlation matrix $\mathbf{\Phi}_{\mathbf{xx}}$.

A. Joint diagonalization

Given two correlation matrices $\mathbf{\Phi}_{\mathbf{xx}} \in \mathbb{C}^{M\times M}$ and $\mathbf{\Phi}_{\mathbf{nn}} \in \mathbb{C}^{M\times M}$, the joint diagonalization of such a matrix pencil is equivalent to solving the generalized eigenvalue decomposition (GEVD) problem as [30]

$$\mathbf{\Phi}_{\mathbf{xx}}\mathbf{U} = \mathbf{\Phi}_{\mathbf{nn}}\mathbf{U}\mathbf{\Lambda}, \tag{14}$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M] \in \mathbb{C}^{M\times M}$ contains the generalized eigenvectors and the diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$ contains the corresponding eigenvalues. Given matrices $\mathbf{U}$ and $\mathbf{\Lambda}$, $\mathbf{\Phi}_{\mathbf{xx}}$ and $\mathbf{\Phi}_{\mathbf{nn}}$ can be jointly diagonalized as

$$\mathbf{U}^H\mathbf{\Phi}_{\mathbf{xx}}\mathbf{U} = \mathbf{\Lambda}, \tag{15}$$

$$\mathbf{U}^H\mathbf{\Phi}_{\mathbf{nn}}\mathbf{U} = \mathbf{I}_M, \tag{16}$$

where $\mathbf{I}_M$ denotes an M-dimensional identity matrix. Based on the GEVD of $\{\mathbf{\Phi}_{\mathbf{xx}}, \mathbf{\Phi}_{\mathbf{nn}}\}$ and due to the fact that $\mathbf{\Phi}_{\mathbf{nn}}$ is always positive definite, we can see that

$$\mathbf{\Phi}_{\mathbf{nn}}^{-1}\mathbf{\Phi}_{\mathbf{xx}}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \tag{17}$$

implying that $(\lambda_j, \mathbf{u}_j), \forall j$ are the right eigenpairs of $\mathbf{\Phi}_{\mathbf{nn}}^{-1}\mathbf{\Phi}_{\mathbf{xx}}$. Further, the noisy correlation matrix $\mathbf{\Phi}_{\mathbf{yy}}$ can be diagonalized as

$$\mathbf{U}^H\mathbf{\Phi}_{\mathbf{yy}}\mathbf{U} = \mathbf{\Lambda} + \mathbf{I}_M. \tag{18}$$

Therefore, $\mathbf{\Phi}_{\mathbf{xx}}$ can be diagonalized by calculating the eigenpairs based on the use of noise and noisy correlation matrices.

B. Optimal MMSE beamformer

Given a reference microphone m, the optimal minimum mean square-error (MMSE) beamformer is formulated as the following constrained optimization problem [16], [19], [31]

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbb{E}\left[|\mathbf{w}^H\mathbf{x} - X_m|^2\right] \\ \text{s.t.} \quad & \mathbb{E}\left[|\mathbf{w}^H\mathbf{n}|^2\right] \le c, \end{aligned} \tag{19}$$

where $0 \le c \le \sigma^2_{N_m}$ with $\sigma^2_{N_m}$ denoting the noise PSD at the reference microphone. Applying this MMSE beamformer to the input noisy microphone signals, the signal component at the reference microphone is estimated.

In order to formulate different types of linear beamformers as a function of the generalized eigenvectors, the solution of (19) is defined in the form

$$\mathbf{w} = \mathbf{U}\boldsymbol{\nu}, \tag{20}$$

where $\boldsymbol{\nu} \in \mathbb{C}^M$. Substituting $\mathbf{w} = \mathbf{U}\boldsymbol{\nu}$ into (19), we obtain

$$\boldsymbol{\nu} = (\mathbf{\Lambda} + \mu\mathbf{I}_M)^{-1}\mathbf{U}^H\mathbf{\Phi}_{\mathbf{xx}}\mathbf{e}_m, \tag{21}$$

where $\mathbf{e}_m$ is a column vector with the mth element equal to one and zeros elsewhere. Notice that $\mathbf{e}_m$ functions as a selection vector selecting microphone m as the reference microphone. Consequently, the optimal beamformer is thus given by

$$\mathbf{w} = \mathbf{U}(\mathbf{\Lambda} + \mu\mathbf{I}_M)^{-1}\mathbf{U}^H\mathbf{\Phi}_{\mathbf{xx}}\mathbf{e}_m, \tag{22}$$

where the Lagrange multiplier $\mu \ge 0$ is chosen such that $\boldsymbol{\nu}^H\boldsymbol{\nu} = c$. Different choices of μ can trade off the signal distortion level and noise reduction performance. The resulting beamformer is referred to as the speech distortion weighted multichannel Wiener filter (SDW-MWF) [32], [33].

C. Low-rank approximation for beamformer design

Let $P \le M$ be the rank of $\mathbf{\Phi}_{\mathbf{xx}}$. In theory, $\mathrm{Rank}(\mathbf{\Phi}_{\mathbf{xx}})$ is equal to the number of the sources of interest. However, due to the estimation errors in the noise and noisy correlation matrices, P can be greater. In many applications, one makes a rank-r approximation of $\mathbf{\Phi}_{\mathbf{xx}}$, where $r \le P \le M$ [11], [19], [28], [30]. Letting $\mathbf{U}^{-H} = \mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_M]$, we can decompose $\mathbf{\Phi}_{\mathbf{xx}}$ as

$$\mathbf{\Phi}_{\mathbf{xx}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H = \sum_{j=1}^{M} \lambda_i \mathbf{q}_j \mathbf{q}_j^H. \tag{23}$$

Further, it is easy to verify that

$$\mathbf{\Phi_{nn}} = \mathbf{Q}\mathbf{Q}^H, \quad \mathbf{Q}^H\mathbf{\Phi}_{\mathbf{nn}}^{-1}\mathbf{\Phi_{xx}} = \mathbf{\Lambda}\mathbf{Q}^H, \tag{24}$$

which means that $\mathbf{q}_i, \forall i$ are the left eigenvectors of $\mathbf{\Phi}_{\mathbf{nn}}^{-1}\mathbf{\Phi_{xx}}$. For the single speech source scenario with $\mathbf{\Phi_{xx}}$ rank-1, the normalized principal eigenvector $\mathbf{q}_1$ gives the RTF [23]–[25].

Consequently, with $\mathbf{Q}$, a rank-r approximation of $\mathbf{\Phi_{xx}}$ can be constructed by exploiting the r-maximum eigenvalues and the corresponding eigenvectors as

$$\hat{\mathbf{\Phi}}_{\mathbf{xx}} = \mathbf{Q}_r\mathbf{\Lambda}_r\mathbf{Q}_r^H = \sum_{j=1}^{r} \lambda_j \mathbf{q}_j\mathbf{q}_j^H. \tag{25}$$

Substituting the low-rank approximation of $\mathbf{\Phi_{xx}}$ into (22), the rank-r optimal MMSE beamformer is given by

$$\mathbf{w}_r = \mathbf{U}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{\Lambda}_r\mathbf{Q}_r^H\mathbf{e}_m. \tag{26}$$

Choosing particular values for r and/or μ, well-known special cases of $\mathbf{w}_r$ are obtained.

1) Classic MWF: In case $\mu = 1$ and $r = P = M$, we can see that

$$\mathbf{w}_{\mathrm{MWF}} = \underbrace{\mathbf{U}\left(\mathbf{\Lambda} + \mathbf{I}\right)^{-1}\mathbf{U}^H}_{\mathbf{\Phi}_{\mathbf{yy}}^{-1}}\underbrace{\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H}_{\mathbf{\Phi_{xx}}}\mathbf{e}_k, \tag{27}$$

since $\mathbf{U}^H\mathbf{Q} = \mathbf{I}_M$ (i.e., the left and right eigenvectors are bi-orthogonal). This filter is known as the classic MWF.

2) Rank-1 beamformer: In case $r = 1$, we obtain the rank-1 beamformer as

$$\mathbf{w}_1 = \frac{\lambda_1 q_{1m}^*}{\lambda_1 + \mu}\mathbf{u}_1, \tag{28}$$

where $q_{1m}^* = \mathbf{q}_1^H\mathbf{e}_m$ denotes the complex conjugate of the mth element of $\mathbf{q}_1$.

3) MVDR beamformer: In case $r = 1$ and $\mu = 0$, we obtain the classic MVDR beamformer as

$$\mathbf{w}_{\mathrm{MVDR}} = q_{1m}^*\mathbf{u}_1, \tag{29}$$

which is a special case of the rank-1 beamformer. By setting proper required parameters, one can obtain different variants of the optimal MMSE beamformer, e.g., see [19] for an overview.

IV. Performance analysis

In this section, we will analyze the dependence of the output SNR of the MMSE beamformers on the reference microphone m. In realistic speech communication systems, as Fig. 1 shows, it is required not only to enhance the target signal, but also to play out the enhanced speech signal for the listener. The speech quality and speech intelligibility of the beamformer output signal then also depend on the acoustic noise in the listening environment, as the enhanced signal would be acoustically mixed with the near-end noise in a noisy environment [34]. For this, we first extend the definition of the frequency-dependent output SNR to also include the near-end noise. That is,

$$\mathrm{oSNR}_m^{\mathrm{near}} = \frac{\mathbf{w}^H\mathbf{\Phi_{xx}}\mathbf{w}}{\mathbf{w}^H\mathbf{\Phi_{nn}}\mathbf{w} + \sigma_U^2}, \tag{30}$$

where σ_U^2 denotes the near-end noise variance of the noise in the environment of the listener that gets acoustically mixed with the beamformer output. In Fig. 4 and Fig. 5, we visualize the combination of the far-end and near-end scenarios.

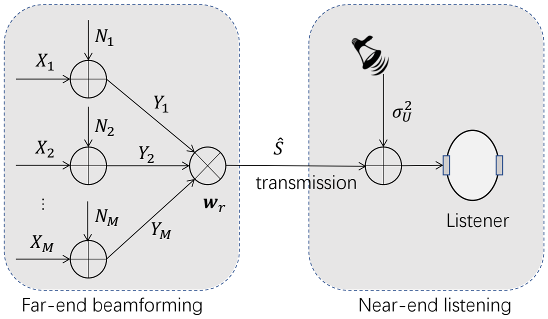


Figure 1. An illustrative example of realistic speech communication systems consisting of the far-end beamforming and near-end listening modules.

A. Rank-r beamformer with near-end noise

Using the rank-r optimal filter given in (26) with $1 \leq r \leq M$, the near-end output SNR in (30) can be calculated as

$$\mathrm{oSNR}_m^{\mathrm{near}} = \frac{\mathbf{w}_r^H\mathbf{\Phi_{xx}}\mathbf{w}_r}{\mathbf{w}_r^H\mathbf{\Phi_{nn}}\mathbf{w}_r + \sigma_U^2} = \frac{\mathbf{e}_m^H\mathbf{A}\mathbf{e}_m}{\mathbf{e}_m^H\mathbf{B}\mathbf{e}_m + \sigma_U^2}, \tag{31}$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} = \mathbf{Q}_r\mathbf{\Lambda}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{U}_r^H\mathbf{\Phi_{xx}}\mathbf{U}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{\Lambda}_r\mathbf{Q}_r^H,$$
$$\mathbf{B} = \mathbf{Q}_r\mathbf{\Lambda}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{U}_r^H\mathbf{\Phi_{nn}}\mathbf{U}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{\Lambda}_r\mathbf{Q}_r^H.$$

By inspection, we have

$$\begin{aligned}\mathbf{U}_r^H\mathbf{\Phi_{xx}}\mathbf{U}_r &= \mathbf{U}_r^H\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H\mathbf{U}_r \\ &= \left[\mathbf{I}_r \ \mathbf{0}_{r\times(M-r)}\right]\mathbf{\Lambda}\left[\mathbf{I}_r \ \mathbf{0}_{(M-r)\times r}\right]^T = \mathbf{\Lambda}_r, \\ \mathbf{U}_r^H\mathbf{\Phi_{nn}}\mathbf{U}_r &= \mathbf{U}_r^H\mathbf{Q}\mathbf{Q}^H\mathbf{U}_r = \mathbf{I}_r.\end{aligned}$$

As a consequence, we obtain

$$\begin{aligned}\mathbf{A} &= \mathbf{Q}_r\mathbf{\Lambda}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{\Lambda}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{\Lambda}_r\mathbf{Q}_r^H \\ &= \sum_{j=1}^{r}\frac{\lambda_j^3}{(\lambda_j + \mu)^2}\mathbf{q}_j\mathbf{q}_j^H,\end{aligned} \tag{32}$$

$$\begin{aligned}\mathbf{B} &= \mathbf{Q}_r\mathbf{\Lambda}_r\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\left(\mathbf{\Lambda}_r + \mu\mathbf{I}_r\right)^{-1}\mathbf{\Lambda}_r\mathbf{Q}_r^H \\ &= \sum_{j=1}^{r}\frac{\lambda_j^2}{(\lambda_j + \mu)^2}\mathbf{q}_j\mathbf{q}_j^H.\end{aligned} \tag{33}$$

The output SNR of rank-r beamformers is thus given by

$$\mathrm{oSNR}_m^{\mathrm{near}} = \frac{\sum_{j=1}^{r}\frac{\lambda_j^3}{(\lambda_j+\mu)^2}|q_{mj}|^2}{\sum_{j=1}^{r}\frac{\lambda_j^2}{(\lambda_j+\mu)^2}|q_{mj}|^2 + \sigma_U^2}, \tag{34}$$

which is clearly reference microphone dependent via the factor q_{mj} included in the summation over j.

B. Rank-r beamformer without near-end noise

If $\sigma_U^2 = 0$, i.e., the near-end noise is neglected, the far-end output SNR of the rank-r beamformer is then given by

$$\mathrm{oSNR}_m^{\mathrm{far}} = \frac{\mathbf{e}_m^H\mathbf{A}\mathbf{e}_m}{\mathbf{e}_m^H\mathbf{B}\mathbf{e}_m} = \frac{\sum_{j=1}^{r}\frac{\lambda_j^3}{(\lambda_j+\mu)^2}|q_{mj}|^2}{\sum_{j=1}^{r}\frac{\lambda_j^2}{(\lambda_j+\mu)^2}|q_{mj}|^2}, \tag{35}$$

which is still reference microphone dependent via the factor q_{mj}. This dependence implies that the selection of a reference

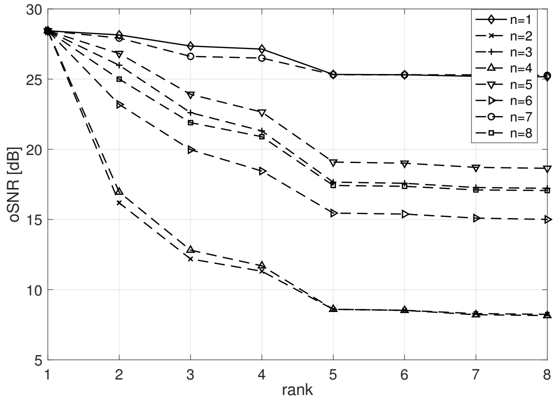


Figure 2. The narrowband far-end output SNR in terms of rank r using the nth microphone as the reference without near-end noise.

microphone also affects the noise reduction performance of general rank-r beamformers (e.g., the conventional MWF, SDW-MWF) without near-end noise, certainly when implemented with the estimated (higher rank) correlation matrices.

C. Rank-1 beamformer without near-end noise

As a special case, applying the rank-1 beamformers, the far-end output SNR is given by

$$\mathrm{oSNR}_m^{\text{far}} = \frac{\mathbf{u}_1^H \mathbf{\Phi}_{\mathbf{xx}} \mathbf{u}_1}{\mathbf{u}_1^H \mathbf{\Phi}_{\mathbf{nn}} \mathbf{u}_1} = \lambda_1. \tag{36}$$

Obviously, the rank-1 beamformer is capable of maximizing the output SNR, which equals the maximum generalized eigenvalue. Therefore, for all rank-1 beamformers (e.g., MVDR beamformer, maximum SNR beamformer), the far-end output SNR (i.e., when neglecting the near-end noise) is then reference microphone independent.

Furthermore, the output SNR of any rank-r beamformer cannot exceed the maximum eigenvalue. An illustrative example of the output SNR in terms of the rank r is shown in Fig. 2. We use a uniform linear array (ULA) consisting of $M = 8$ microphones and design the rank-r optimal beamformer given in (26) for noise reduction. In this case, one can choose any microphone as the reference. It is clear that for $r = 1$, the maximum output SNR is obtained independent of the reference microphone. For any rank-r beamformer with $2 \leq r \leq M$, the output SNR depends on the reference. With an increase in the rank, the performance decreases. This also follows from Theorem 1.

Theorem 1. *Given the same reference microphone, the far-end output SNR of rank-r MMSE beamformers satisfies [19]*

$$\lambda_1 = \mathrm{oSNR}_{r=1} \geq \mathrm{oSNR}_{r=2} \geq \cdots \geq \mathrm{oSNR}_{r=M}. \tag{37}$$

Proof. Letting $x_j = \frac{\lambda_j^2}{(\lambda_j+\mu)^2}|q_{mj}|^2 > 0, \forall j = 1, \ldots, r$, then it can be shown that

$$\begin{aligned}\mathrm{oSNR}_{r=1} - \mathrm{oSNR}_{r=2} &= \lambda_1 - \frac{\lambda_1 x_1 + \lambda_2 x_2}{x_1 + x_2} \\ &= \frac{(\lambda_1 - \lambda_2) x_2}{x_1 + x_2} \geq 0,\end{aligned}$$

since $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_M$. This can be easily generalized to show $\mathrm{oSNR}_{r=j} \geq \mathrm{oSNR}_{r=j+1}$ for $j \geq 2$. This completes the proof. □

Altogether, we can conclude that in general the output SNR of any rank-r beamformer is affected by the reference, and only when $\sigma_U^2 = 0$, the rank-1 beamformers are not affected by the reference microphone. Next, we will optimize the output SNR given in (30) for the general case via reference microphone selection.

V. Proposed reference selection approach

In this section, we will propose two reference microphone selection approaches.

A. maxoSNR

Typically, the estimated correlation matrix $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$ has a rank $\mathrm{Rank}(\hat{\mathbf{\Phi}}_{\mathbf{xx}}) > 1$ because of inaccuracies in the estimated correlation matrices (which are estimated using a limited amount of data). Based on the previous analysis, we know from (36) that for rank-1 beamformers, with the absence of near-end noise the output SNR does not depend on the reference regardless of the actual rank $\mathrm{Rank}(\hat{\mathbf{\Phi}}_{\mathbf{xx}})$. However, generally, when the near-end noise is also present, for any rank-r beamformer, it holds that they do depend on the reference microphone. Their performance is thus affected by the chosen reference for any r. The MMSE beamformers are calculated per frequency bin and for each frequency bin the narrowband SNR can be quite different depending on the reference microphone. Therefore, we first propose to optimize the frequency-dependent output SNR by selecting a reference microphone for each frequency bin individually. At the end of this subsection, this will be extended to broadband reference selection where one microphone is selected for the complete frequency range. In line with (30), the frequency-dependent optimal reference microphone can be determined by solving the following problem formulation:

$$\begin{aligned} n_k &= \arg\max_m \mathrm{oSNR}_m^{\text{near}} = \arg\max_m \frac{\mathbf{e}_m^H \mathbf{A} \mathbf{e}_m}{\mathbf{e}_m^H \mathbf{C} \mathbf{e}_m}, \\ &\text{s.t.} \quad \mathbf{1}_M^T \mathbf{e}_m = 1, \ \mathbf{e}_m \in \{0,1\}^M, \end{aligned} \tag{38}$$

where $\mathbf{1}_M$ denotes an M-dimensional all-ones column vector, and the constraints are to force that only one element in $\mathbf{e}_k$ equals one, and $\mathbf{C} = \mathbf{B} + \sigma_U^2 \mathbf{I}$. Clearly, this is a Boolean optimization problem, which can be maximized by taking the maximum of the element-wise division between the diagonal elements of $\mathbf{A}$ and $\mathbf{C}$. As an alternative, we can also solve this as a semi-definite programming (SDP) problem. To do so, we first relax (38) as

$$\begin{aligned} \max_{\mathbf{e}_m, \eta} \quad & \mathbf{e}_m^H \mathbf{A} \mathbf{e}_m / \eta \\ \text{s.t.} \quad & \mathbf{e}_m^H \mathbf{C} \mathbf{e}_m \leq \eta \\ & \mathbf{1}_M^T \mathbf{e}_m = 1, \ \mathbf{e}_m \in \{0,1\}^M, \end{aligned} \tag{39}$$

by introducing a new variable $\eta > 0$. Note that the first constraint can be re-written as a linear inequality constraint using the Schur complement [35]

$$\begin{bmatrix} \mathbf{C}^{-1} & \mathbf{e}_m \\ \mathbf{e}_m^T & \eta \end{bmatrix} \succeq \mathbf{O}_{M+1}, \tag{40}$$

due to the fact that $\mathbf{C}$ is positive definite. Furthermore, if we relax $\mathbf{e}_k$ using the continuous surrogates as $0 \leq e_m[i] \leq 1, \forall i$, we can reformulate (38) as the following SDP problem [35]:

$$\begin{aligned} \max_{\mathbf{e}_m, \eta} \quad & \mathbf{e}_m^H \mathbf{A} \mathbf{e}_m / \eta \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{e}_m \\ \mathbf{e}_m^T & \eta \end{bmatrix} \succeq \mathbf{O}_{M+1} \\ & \mathbf{1}_M^T \mathbf{e}_m = 1, \ \mathbf{e}_m \in [0,1]^M, \end{aligned} \tag{41}$$

which can be solved using a toolbox like CVX [36]. In principal, (39) can be seen as a special case of the general microphone subset selection problem proposed in [37], as only one microphone needs to be selected. The final reference microphone is given by the index of the maximum value of $\mathbf{e}_m$. The proposed selection method is performed per frequency bin, that is, the reference microphone might be changing across frequencies, thus referred to as narrowband $\mathrm{maxoSNR}$. Note that different from [20], the proposed $\mathrm{maxoSNR}$ method (either by simply checking the diagonal elements of the matrices $\mathbf{A}$ and $\mathbf{C}$ or by considering the SDP problem) does not need to design M filters and includes the effect of the near-end noise. Also, checking the diagonal elements and considering the SDP problem lead to the same reference selection.

In order to use the same microphone as the reference for all frequency bins (i.e., broadband selection), one can consider to maximize the broadband output SNR instead of the narrowband SNR as in (38). That is,

$$\begin{aligned} n &= \arg\max_m \frac{\sum_k \mathbf{e}_m^H \mathbf{A}(k) \mathbf{e}_m}{\sum_k \left(\mathbf{e}_m^H \mathbf{B}(k) \mathbf{e}_m + \sigma_U^2 \right)} \\ &= \arg\max_m \frac{\sum_k \mathbf{e}_m^H \mathbf{A}(k) \mathbf{e}_m}{\sum_k \mathbf{e}_m^H \left(\mathbf{B}(k) + \sigma_U^2 \mathbf{I}_M \right) \mathbf{e}_m} \\ &= \arg\max_m \frac{\mathbf{e}_m^H \sum_k \mathbf{A}(k) \mathbf{e}_m}{\mathbf{e}_m^H \sum_k \mathbf{C}(k) \mathbf{e}_m}, \end{aligned} \tag{42}$$

where $\mathbf{C}(k) = \mathbf{B}(k) + \sigma_U^2 \mathbf{I}_M$, subject to the constraints given in (38). Taking the summations $\sum_k \mathbf{A}(k)$ and $\sum_k \mathbf{C}(k)$ as two individual matrices, (42) can then be solved using exactly the same two techniques presented earlier in this section, which gives the optimal single reference microphone across all frequencies. We will refer to this method as the broadband $\mathrm{maxoSNR}$ reference selection method.

B. maxRTF

By ignoring the Boolean constraints in (38), (38) can be relaxed as

$$\max_{\boldsymbol{\psi}} \quad \frac{\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi}}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi} + \sigma_U^2}, \tag{43}$$

where $\boldsymbol{\psi} \in \mathbb{C}^M$. Due to the fact that the matrices $\mathbf{A}$ and $\mathbf{B}$ are positive definite, for any $\boldsymbol{\psi}$ we know that $\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi} > 0$ and $\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi} > 0$. Further, since $\mathbf{B} = \mathbf{Q}_r \boldsymbol{\Lambda}_2 \mathbf{Q}_r^H$, where

$$\boldsymbol{\Lambda}_2 = \boldsymbol{\Lambda}_r \left(\boldsymbol{\Lambda}_r + \mu \mathbf{I}_r \right)^{-1} \left(\boldsymbol{\Lambda}_r + \mu \mathbf{I}_r \right)^{-1} \boldsymbol{\Lambda}_r,$$

$\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}$ is thus bounded by

$$\frac{\lambda_r^2}{\left(\lambda_r + \mu \right)^2} \leq \boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi} \leq \frac{\lambda_1^2}{\left(\lambda_1 + \mu \right)^2}. \tag{44}$$

Therefore, we obtain

$$\frac{\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi}}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi} + \sigma_U^2} = \frac{\frac{\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi}}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}}}{1 + \frac{\sigma_U^2}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}}} \geq \frac{\frac{\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi}}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}}}{1 + \frac{\sigma_U^2 (\lambda_1 + \mu)^2}{\lambda_1^2}}, \tag{45}$$

since $\frac{\sigma_U^2}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}} \geq 0$ with equality obtained when $\boldsymbol{\psi} = \mathbf{q}_1$. As a consequence, (43) can be optimized by maximizing the scaled lower bound, i.e., solving the generalized Rayleigh quotient problem:

$$\max_{\boldsymbol{\psi}} \quad g(\boldsymbol{\psi}) = \frac{\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi}}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}}. \tag{46}$$

For this, we need the following theorem.

Theorem 2. *For $\boldsymbol{\psi} \in \mathbb{C}^M$, $g(\boldsymbol{\psi})$ is bounded by*

$$\lambda_M \leq g(\boldsymbol{\psi}) \leq \lambda_1,$$

the minimum is obtained if and only if $\boldsymbol{\psi} = \mathbf{q}_M$, and the maximum is obtained if and only if $\boldsymbol{\psi} = \mathbf{q}_1$.

Proof. From the analysis in Section IV, we know that $\mathbf{A} = \mathbf{Q}_r \boldsymbol{\Lambda}_1 \mathbf{Q}_r^H$ and $\mathbf{B} = \mathbf{Q}_r \boldsymbol{\Lambda}_2 \mathbf{Q}_r^H$, where $\boldsymbol{\Lambda}_1$ is given by

$$\begin{aligned} \boldsymbol{\Lambda}_1 &= \boldsymbol{\Lambda}_r \left(\boldsymbol{\Lambda}_r + \mu \mathbf{I}_r \right)^{-1} \boldsymbol{\Lambda}_r \left(\boldsymbol{\Lambda}_r + \mu \mathbf{I}_r \right)^{-1} \boldsymbol{\Lambda}_r \\ &= \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_r. \end{aligned}$$

Therefore, we have $\mathbf{A} = \mathbf{Q}_r \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_r \mathbf{Q}_r^H = \mathbf{B} \boldsymbol{\Lambda}_r$, since $\boldsymbol{\Lambda}_r$ is a diagonal matrix. The GEVD of the matrix pencil $\{\mathbf{A}, \mathbf{B}\}$ is then given by

$$\mathbf{A} \mathbf{Q}_r = \mathbf{B} \boldsymbol{\Lambda}_r \mathbf{Q}_r, \tag{47}$$

or equivalently by $\mathbf{B}^{-1} \mathbf{A} \mathbf{Q}_r = \boldsymbol{\Lambda}_r \mathbf{Q}_r$. Maximizing or minimizing the generalized Rayleigh quotient $\frac{\boldsymbol{\psi}^H \mathbf{A} \boldsymbol{\psi}}{\boldsymbol{\psi}^H \mathbf{B} \boldsymbol{\psi}}$ turns out to be solving the GEVD problem. Therefore, the maximum can be obtained when $\boldsymbol{\psi} = \mathbf{q}_1$ (e.g., the principal eigenvector) and the minimum is obtained when $\boldsymbol{\psi} = \mathbf{q}_M$ (i.e., the eigenvector corresponding to the minimum eigenvalue). This completes the proof. □

In this case, the optimal unknown is given by the principal eigenvector $\mathbf{q}_1$. Motivated by this, selecting the reference microphone by searching for the maximum absolute value of $\mathbf{q}_1$ gives a sub-optimal solution as

$$n_k = \arg\max_m |q_{m1}|^2. \tag{48}$$

Remark 1. *For any rank-1 MMSE beamformer, the reference dependent near-end output SNR is given by*

$$\mathrm{oSNR}_m^{near} = \frac{\mathbf{w}_1^H \boldsymbol{\Phi}_{\mathbf{xx}} \mathbf{w}_1}{\mathbf{w}_1^H \boldsymbol{\Phi}_{\mathbf{nn}} \mathbf{w}_1 + \sigma_U^2} = \frac{\alpha_m \lambda_1}{\alpha_m + \sigma_U^2}, \tag{49}$$

where

$$\alpha_m = \left(\frac{\lambda_1}{\lambda_1 + \mu} \right)^2 |q_{1m}|^2, \tag{50}$$

implying that optimizing (48) enables an optimal reference in the sense of SNR. This is due to the fact that the reference-dependent SNR monotonically increases with $|q_{1m}|^2$, i.e., maximizing oSNR_m^{near} is equivalent to optimizing $|q_{1m}|^2$ in the rank-1 case. For a higher-rank case, (48) is then sub-optimal.

Since the RTF is equivalent to the principal left eigenvector [23]–[25], we thus refer to (48) as the proposed narrowband

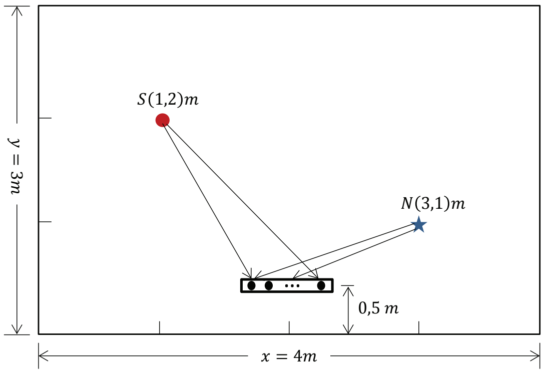


Figure 3. A microphone array based speech enhancement system.

maxRTF method. Similarly to the broadband $\mathrm{maxoSNR}$ method, we can also design a broadband maxRTF procedure by choosing the microphone whose RTF has the maximum average power over all frequencies, i.e.,

$$n = \arg\max_{m} \sum_{k=1}^{F} \left|q_{1m}(k)\right|^2 / F, \tag{51}$$

where F is the total number of frequency bins.

VI. Experimental results

In this section, the proposed reference selection algorithms for the MMSE beamformers are evaluated using a simulated microphone array. Section VI-A shows the experimental setup. In Section VI-B and Section VI-C, the instrumental speech quality and speech intelligibility are evaluated, respectively. In Section VI-D, we evaluate the performance of two often-used filters, i.e., the MWF and MVDR beamformer. The proposed $\mathrm{maxoSNR}$ method and the proposed maxRTF will be compared to the reference methods $\mathrm{maxiSNR}$ [20], $\mathrm{minDist}$ [20], $\mathrm{maxEnergy}$ [20] and a random reference selection procedure. For the $\mathrm{maxiSNR}$ and $\mathrm{maxEnergy}$ methods, which are introduced as a narrowband selection procedure in [20], the corresponding broadband versions will also be compared[2]. For the $\mathrm{minDist}$ method, we assume that the source-microphone distances are known, which in practice need to be estimated. The performance of the average random selection method is a broadband selection that is evaluated by averaging the performance that is obtained by all possible M filters. In order to clearly observe the superiority of the proposed methods over the baselines, we use “prop.” to indicate the proposed methods in the legends of graphs.

A. Experimental setup

We use a conventional ULA consisting of $M = 8$ omnidirectional microphones with a spacing of 2 cm. The microphones are indexed as $m \in \{1, 2, \cdots, 8\}$ from left to right. We consider a simulated 2D room with dimensions (4×3) m as Fig. 3 depicts, where a single target source and a coherent interfering point source are located at $(1, 2)$ m and $(3, 1)$ m, respectively. The target speech source is a 5 minute audio stream that is obtained by concatenating several speech signals originating from the TIMIT database [38]. The interfering source is a stationary Gaussian speech shaped noise signal. The sampling frequency is 16 kHz. The ATFs are generated using the toolbox in [39]. All the filtering processes take place in the STFT domain, where a square-root Hann window of 50 ms for segmentation with 50% overlap, and the estimated speech signal is recovered via inverse STFT. Due to the thermal noise of electronic devices, we model the microphone self noise using a zero-mean uncorrelated Gaussian noise at an SNR of 40 dB. The reverberation time is set to be $T_{60} = 200$ ms. The trade-off parameter is set to be $\mu = 1$. Further, the source-to-interference ratio (SIR) is set to be 0 dB. In order to focus on the influence of reference microphone selection, we assume that an ideal VAD is available, such that the microphone recordings can be classified into noise-only segments and speech-plus-noise segments, and during these two periods the correlation matrices $\mathbf{\Phi_{nn}}$ and $\mathbf{\Phi_{yy}}$ are estimated using the average smoothing method, respectively. Throughout the numerical simulations, the actual rank of the estimated autocorrelation matrix $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$ is $P = M$ due to the limited amount of measurements.

As evaluation metrics, we use the SNR gain to measure the speech quality, and the gain in short-time objective intelligibility (STOI) [40] and the gain in speech intelligibility in bits (SIIB) [41], [42] to measure the instrumental speech intelligibility. The SNR gain (denoted by $\Delta\mathrm{SNR}$) is obtained by subtracting the input SNR from the output SNR, similarly for $\Delta\mathrm{STOI}$ and $\Delta\mathrm{SIIB}$. The STOI score is to measure the instrumental intelligibility of a speech signal, which represents the correlation between the short-time temporal envelopes of the clean and enhanced (or noisy) signals, and has been shown to be highly correlated to human speech intelligibility score. The STOI score ranges from 0 to 1, and the higher it is, the more intelligible the speech is. The SIIB score measures the amount of information shared between the clean speech (i.e., the talker at the start point of a realistic speech communication system) and the degraded speech (i.e., the listener at the end point) in bits per second (bps), and has been shown to be more reliable than STOI for a larger diversity of processing conditions [42]. Similarly to STOI, the higher the SIIB score is, the more intelligible the obtained speech is. In simulations, we set the total number of frequency bins to be $F = 1024$.

B. Instrumental speech quality

In order to study the impact of the rank-r approximation of $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$ on the noise reduction performance, we first show the SNR gain of the rank-r MMSE beamformer in terms of the rank-r approximation of the beamformers summarized in Section III using different reference microphone selection methods in Fig. 4. To model the near-end noise, we add zero-mean Gaussian noise at a variance of $\sigma_U^2 = 10^{-4}$ to the beamformer output. The target to near-end noise ratio (TNNR)

[2] Each narrowband method considers reference selection for each frequency bin individually, that is, different frequency bins might use a difference microphone as the reference. The broadband $\mathrm{maxiSNR}$ method can be designed by choosing the microphone having the maximum average input SNR over all frequencies, and the broadband $\mathrm{maxEnergy}$ method by choosing the microphone having the maximum average energy, such that their reference for all frequency bins keeps the same. Note that $\mathrm{minDist}$ and the random selection method are already broadband.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.3039930, IEEE/ACM Transactions on Audio, Speech, and Language Processing

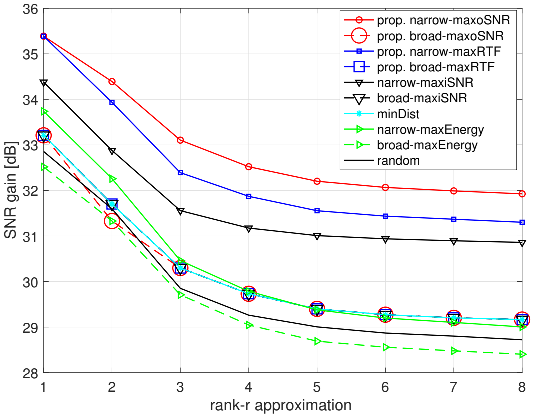


Figure 4. The SNR gain in terms of the rank r with TNNR = 40 dB.

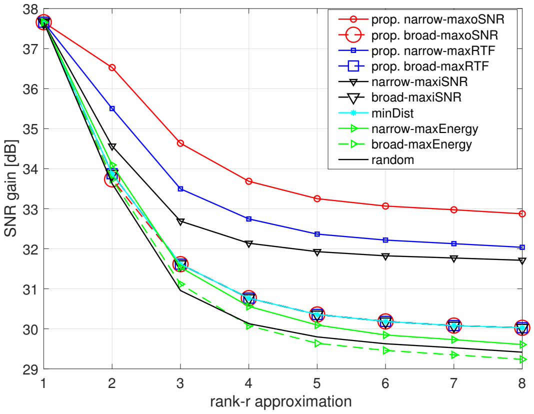


Figure 5. The SNR gain of the MMSE beamformers in terms of the rank r that is used for approximating $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$ without near-end noise, i.e., $\sigma_U^2 = 0$.

is around 40 dB. For implementation, after the noise and noisy correlation matrices $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$ and $\hat{\mathbf{\Phi}}_{\mathbf{yy}}$ are estimated, we use the r eigenvectors corresponding to the r-maximum eigenvalues and the corresponding eigenvalues (minus one) of $\{\hat{\mathbf{\Phi}}_{\mathbf{yy}}, \hat{\mathbf{\Phi}}_{\mathbf{nn}}\}$ to perform the rank-r approximation of $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$. As expected from the theoretical analysis, with an increase in the rank, the SNR gain of all comparison methods decreases. From Fig. 4, we can observe that for any rank-r case, the SNR gain depends on the reference microphone in case the near-end noise is taken into account. As expected from Section V-B, for the rank-1 case the narrowband maxoSNR and maxRTF obtain the same SNR gain, as they are equivalent in this case. The proposed narrowband maxoSNR method achieves the best performance in SNR gain, and the proposed narrowband maxRTF approach is near-optimal. In general, the narrowband selection procedure outperforms the corresponding broadband counterpart with respect to the SNR. Notably, with an increase in the rank, the selection of a reference microphone has a more severe impact on the performance of MMSE beamformers, e.g., the SNR gap between the proposed method and maxEnergy becomes larger. Interestingly, comparing the broadband methods, the performance of the proposed broadband maxoSNR, maxRTF, the broadband maxiSNR and minDist approaches overlaps, which is better than the random selection method. That is, in the broadband sense the microphone that is located closest to the target source is optimal for maximizing the SNR gain. For the full-rank case, the proposed narrowband and broadband maxoSNR methods can improve the SNR gain by 3 dB and 0.5 dB compared to the random selection, respectively.

In Fig. 5, we show the SNR gain of the rank-1 beamformer in terms of the rank r for $\sigma_U^2 = 0$. It is clear that for $r = 1$, the SNR gain of rank-1 MMSE beamformers is reference independent without near-end noise being taken into account, and the maximum SNR gain is achieved. Furthermore, the SNR gain of the rank-1 MMSE beamformers in terms of the TNNR is shown in Fig. 6. It is clear that for the rank-1 case, the proposed narrowband maxoSNR and maxRTF methods are equivalent. The SNR gain of all reference selection methods based rank-1 MMSE beamformers increases with an increase in the TNNR (i.e., a decrease in the near-end noise variance σ_U^2). When the near-end noise is negligible, all the considered reference selection methods obtain a similar performance, that is, the reference does not affect the near-end output SNR. In case the variance of the near-end noise increases, selecting a proper reference becomes more important for rank-1 beamformers, as the performance gap between the proposed narrowband methods and other approaches becomes larger.

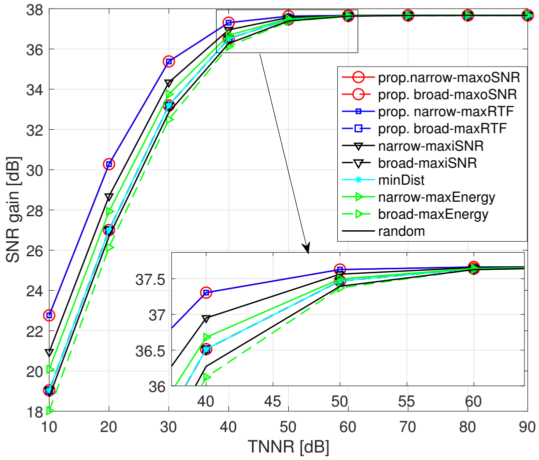


Figure 6. The SNR gain of the rank-1 MMSE beamformer in terms of TNNR.

C. Instrumental speech intelligibility

In this section, we evaluate the reference selection algorithms in terms of the predicted instrumental intelligibility with the TNNR fixed to be 40 dB. Fig. 7(a) shows the STOI gain in terms of the rank r that is used for approximating $\hat{\mathbf{\Phi}}_{\mathbf{xx}}$. The proposed broadband maxoSNR, broadband maxRTF, broadband maxiSNR and minDist methods all select the first microphone as the reference, resulting in the maximum improvement in STOI. In Fig. 7(a), it is clear that the broadband methods (except for the broadband maxEnergy and the random method) can achieve a better speech intelligibility compared to the narrowband approaches, while in Fig. 4 the narrowband method achieves a better SNR gain. All the narrowband methods (except for the narrowband maxEnergy) have a similar performance in terms of intelligibility. We can conclude that, in general, the narrowband procedure is

Table I
NEAR-END NOISE REDUCTION PERFORMANCE USING THE CLASSIC MWF ($\mu = 1, r = M$) AND THE MVDR BEAMFORMER ($\mu = 0, r = 1$).

Method	MWF ($\mu = 1, r = M$)					MVDR ($\mu = 0, r = 1$)				
	ΔSNR	ΔSTOI	ΔSIIB	ΔPESQ	RefMic (#)	ΔSNR	ΔSTOI	ΔSIIB	ΔPESQ	RefMic (#)
prop. narrow maxoSNR	**32.738**	0.350	220.61	1.423	1 (213)	**37.301**	0.352	183.43	1.384	1 (201)
prop. narrow maxRTF	31.894	0.352	195.37	1.423	1 (211)	**37.301**	0.352	206.89	1.384	1 (201)
narrow maxiSNR	31.613	0.354	199.98	1.437	1 (241)	36.986	0.355	217.09	1.402	1 (243)
narrow maxEnergy	29.517	0.334	167.21	1.345	8 (239)	36.690	0.345	218.52	1.312	8 (237)
broad maxoSNR, maxRTF broad maxiSNR, minDist	**29.905**	**0.392**	**320.76**	**1.508**	**1 (1024)**	**36.518**	**0.383**	**325.73**	**1.474**	**1 (1024)**
broad maxEnergy	29.130	0.313	158.28	1.395	8 (1024)	36.192	0.335	197.57	1.356	8 (1024)
broad random	29.319	0.354	212.77	1.471	4 (1024)	36.283	0.354	228.38	1.436	4 (1024)

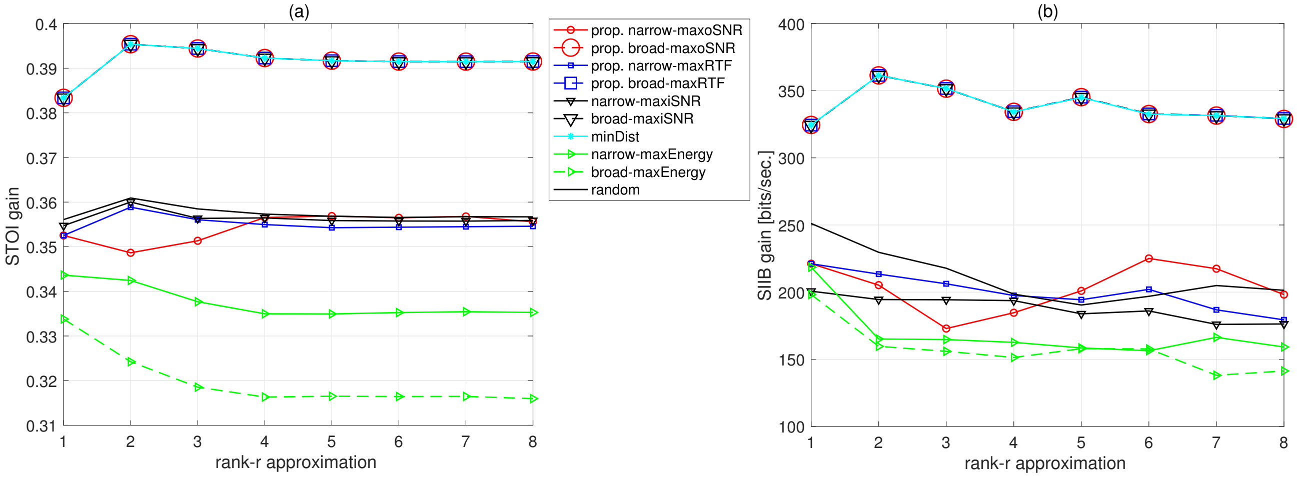


Figure 7. The STOI gain and the SIIB gain (in bits per second) in terms of the rank r with TNNR = 40 dB.

better in SNR gain, while the broadband version is better in terms of speech intelligibility. This is due to the fact that the narrowband methods change the reference microphone across frequencies, that is, the phase and magnitude of the target signal might change per frequency, which will influence the speech intelligibility. Interestingly, the narrowband $\mathrm{maxEnergy}$ (which might use different microphones as the reference across frequencies) outperforms the broadband $\mathrm{maxEnergy}$ (which uses microphone 8 as the reference for all frequencies) in both SNR and STOI, as the signal recorded by microphone 8 is dominated by the noise source. Fig. 7(b) shows the speech intelligibility in terms of the SIIB gain (in bits per second). These results are similar to Fig. 7(a). Comparing the proposed broadband $\mathrm{maxoSNR}$ to the random selection method, it is clear that apart from the signal quality in terms of SNR, the speech intelligibility can also be improved by choosing a proper reference microphone. That is, in practice the reference microphone should not be arbitrarily chosen, as this will harm the performance.

D. Evaluation of MWF and MVDR

Finally, we consider two often-used spatial filters, i.e., the classic MWF (i.e., $\mu = 1$, $r = M$) and the rank-1 MVDR beamformer (i.e., $\mu = 0$, $r = 1$) with TNNR = 40 dB. The speech enhancement performance is shown in Table I, where we also show the gain in the perceptual evaluation of speech quality (PESQ) [43], denoted by $\Delta\mathrm{PESQ}$. We also indicate the microphone index that is most frequently chosen as the reference microphone and the corresponding times it is chosen by different approaches. The performance of the broadband $\mathrm{maxoSNR}$, maxRTF, $\mathrm{maxiSNR}$ and $\mathrm{minDist}$ approaches is identical, as they all select the first microphone as the reference for all frequencies. Therefore, we show these methods in one row together in Table I. The proposed narrowband $\mathrm{maxoSNR}$ obtains the best output SNR. Given the source-microphone distance, the broadband $\mathrm{minDist}$ method obtains the best predicted speech intelligibility improvement. This is due to the fact that the closest microphone has the maximum input SNR and its recording is dominated by the clean signal component. However, the $\mathrm{minDist}$ is an impractical method, due to the unavailability of the source-microphone distance. In this case, the proposed broadband methods can be applied to obtain an informative reference. Randomly choosing a reference microphone can do better than the $\mathrm{maxEnergy}$ method, but it is still worse than using more elaborate strategies (e.g., the proposed methods, $\mathrm{maxiSNR}$ and $\mathrm{minDist}$). The broadband $\mathrm{maxEnergy}$ method uses microphone 8 as the reference, but achieves the worst performance, as this microphone is closest to the coherent interfering source and its measurement is dominated by the noise source. The conclusions in terms of PESQ gain are similar to the conclusions related to speech intelligibility gain. Altogether, we see that we gain about 3 dB in terms of SNR by selecting the right reference microphone and increase the predicted instrumental intelligibility as measured by SIIB with around 100 bps. Finally, we show the

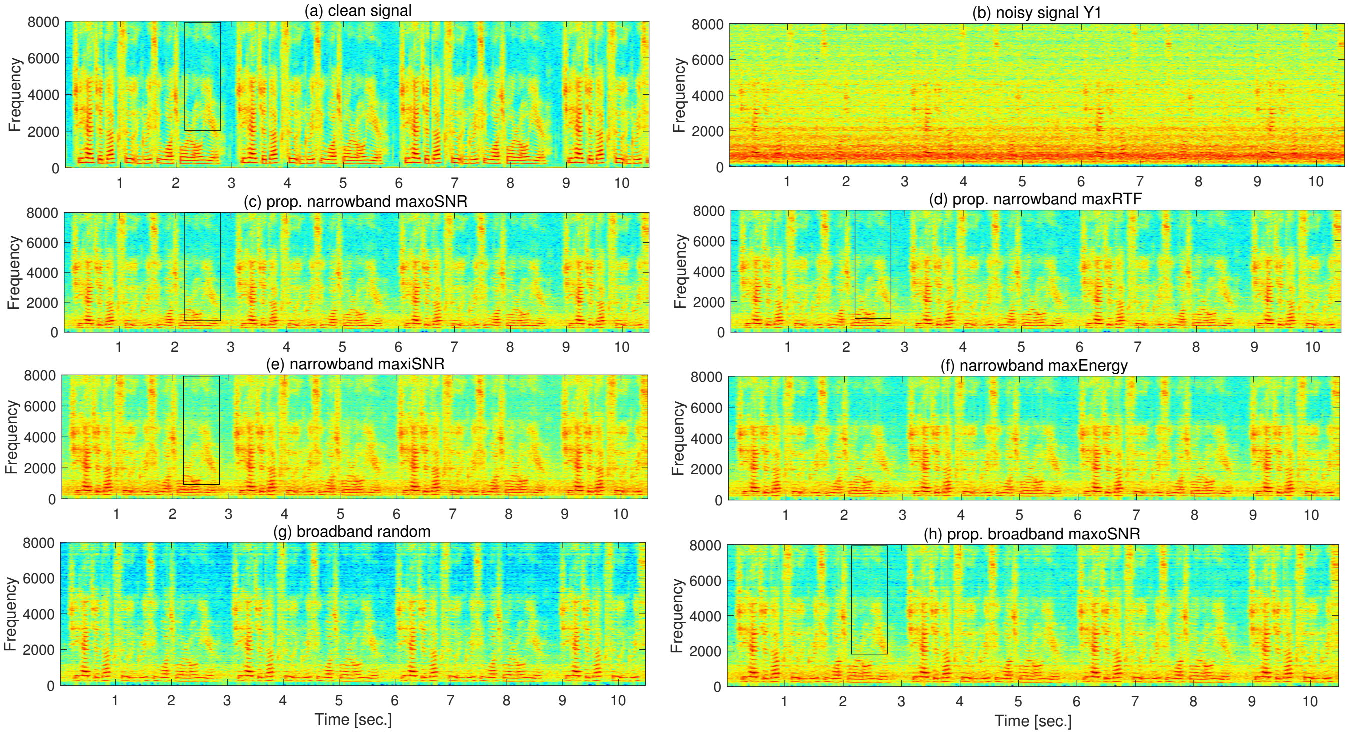


Figure 8. Spectrograms: (a) clean signal, (b) noisy signal at microphone 1, enhanced signals using (c) narrowband $\mathrm{maxoSNR}$, (d) narrowband maxRTF, (e) $\mathrm{maxiSNR}$, (f) $\mathrm{maxEnergy}$, (g) random selection, and (h) broadband $\mathrm{maxoSNR}$. Note that the enhanced signals of the broadband $\mathrm{maxoSNR}$, maxRTF, $\mathrm{maxiSNR}$ and $\mathrm{minDist}$ approaches are the same, as they use the same microphone as the reference.

spectrograms of the clean, noisy and enhanced signals using different reference selection approaches for the MWF in Fig. 8. It is obvious that the spectrograms of the proposed methods are more similar to that of the clean signal than that of comparison approaches, particularly in the square area.

VII. Conclusions

In this paper, we systematically investigated the impact of choosing a reference microphone on the spatial filtering based multi-microphone noise reduction problem. From theoretical analysis, we found that for any rank-r MMSE beamformer, the near-end output SNR including the near-end noise depends on the reference. If and only if the near-end noise is neglected, the output SNR of rank-1 beamformers (e.g., MVDR) is reference independent. The proposed narrowband $\mathrm{maxoSNR}$ method is optimal for MMSE beamformers in SNR. In addition, the proposed narrowband maxRTF approach is sub-optimal in terms of SNR. For the rank-1 beamformers, $\mathrm{maxoSNR}$ and maxRTF are equivalent. The broadband version of both methods reduces to the optimal $\mathrm{minDist}$ case, i.e., selecting the microphone closest to the target source as the reference for all frequencies. Using a simulated microphone array, it was shown that the proposed narrowband $\mathrm{maxoSNR}$ and maxRTF approaches can improve the signal SNR as compared to other practical reference microphone selection methods. In general, the narrowband selection procedure can improve the SNR, while the broadband counterpart is beneficial for improving the speech intelligibility. It is reasonable that the proposed methods are also valid in other more complex acoustic scenarios, as the proposed theory was built without strict assumptions on the number of sources, the positional relationship between the target source and the interfering source, etc.

VIII. Acknowledgements

The authors wish to thank the Associate Editor and anonymous reviewers for their constructive suggestions that helped to improve the presentation of this work.

References

[1] D. C. Moore and I. A. McCowan, “Microphone array speech recognition: Experiments on overlapping speech in meetings,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. V–497.

[2] Y. Hu and P. C. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms,” *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[3] F. Khalil, J. P. Jullien, and A. Gilloire, “Microphone array for sound pickup in teleconference systems,” *Journal of the Audio Engineering Society*, vol. 42, no. 9, pp. 691–700, 1994.

[4] H. Wang and P. Chu, “Voice source localization for automatic camera pointing system in videoconferencing,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, pp. 187–190.

[5] J-M. Valin, J. Rouat, and F. Michaud, “Enhanced robot audition based on microphone array source separation with post-filter,” in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2004, vol. 3, pp. 2123–2128.

[6] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[7] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, 1999.

[8] R. C. Hendriks, T. Gerkmann, and J. Jensen, “DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art,” *Synthesis Lectures on Speech and Audio Process.*, vol. 9, no. 1, pp. 1–80, 2013.

[9] O. L. Frost III, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[10] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas, Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[11] 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, 2002.

[12] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, 2003.

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

[14] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2013.

[15] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, pp. 9–41. Springer, 2005.

[16] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, 2007.

[17] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, Springer Science & Business Media, 2005.

[18] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "A framework for speech enhancement with ad hoc microphone arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 6, pp. 1038–1051, 2016.

[19] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 631–644, 2016.

[20] T. C. Lawin-Ore and S. Doclo, "Reference microphone selection for MWF-based noise reduction using distributed microphone arrays," in *ITG-Fachtagung Sprachkommun.* VDE, 2012, pp. 1–4.

[21] S. Stenzel, J. Freudenberger, and G. Schmidt, "A minimum variance beamformer for spatially distributed microphones using a soft reference selection," in *Int. Workshop Hands-Free Speech Commun.*, 2014, pp. 127–131.

[22] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Comparison of reference microphone selection algorithms for distributed microphone array based speech enhancement in meeting recognition scenarios," in *Int. Workshop Acoustic Sig. Enhance. (IWAENC)*, 2018, pp. 316–320.

[23] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 544–548.

[24] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2018, pp. 2513–2517.

[25] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1507–1519, 2019.

[26] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[27] X.-F. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2171–2186, 2016.

[28] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[29] S. Kotti, R. Heusdens, and R. C. Hendriks, "Clock-offset and microphone gain mismatch invariant beamforming," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, Amsterdam, the Netherlands, 2020.

[30] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, 1995.

[31] Y. Ephraim and H. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, 1995.

[32] J. W. A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multichannel wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 7, pp. 2367–2387, 2004.

[33] S. Doclo, J. W. A. Spriet, and M. Moonen, *Speech distortion weighted multichannel Wiener filtering techniques for noise reduction, Springer Series on Signals and Communication Technology*, Springer, 2014.

[34] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility enhancement based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1694–1708, 2017.

[35] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[36] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008.

[37] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone subset selection for MVDR beamformer based noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 550–563, 2018.

[38] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, vol. 15, pp. 29–50, 1988.

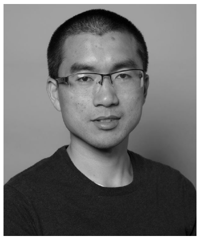
[39] E. A. P. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep.*, vol. 2, no. 2.4, pp. 1, 2006.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[41] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, 2018.

[42] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[43] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ) : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P. 862*, 2001.



Jie Zhang was born in Anhui Province, China, in 1990. He received the B.Sc. (with honors), M.Sc. (with honors), and Ph.D. degrees in electrical engineering from the Yunnan University, Yunnan, the Peking University (PKU), Beijing, P. R. China and the Delft University of Technology (TU Delft), Delft, The Netherlands, in 2012, 2015 and 2020, respectively. He is currently an Assistant Professor in the National Engineering Laboratory for Speech and Language Information Processing, Faculty of Information Science and Technology, University of Science and Technology of China (USTC), Hefei, China. He received the Best Student Paper Award for his publication at the 10th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2018) in Sheffield, UK. His current research interests include multi-microphone speech enhancement, sound source localization, binaural auditory, speech recognition, and speech processing over wireless (acoustic) sensor networks.

Richard Christian Hendriks was born in Schiedam, The Netherlands. He received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively. He is currently an Associate Professor in the Circuits and Systems (CAS) Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. His main research interest is on biomedical signal processing, and, audio and speech processing, including speech enhancement, speech intelligibility improvement and intelligibility modelling. In March 2010, he received the prestigious VENI grant for his proposal Intelligibility Enhancement for Speech Communication Systems. He obtained several best paper awards, among which the IEEE Signal Processing Society best paper award in 2016. He is an Associate Editor for the IEEE/ACM Trans. on Audio, Speech, and Language Processing and the EURASIP Journal on Advances in Signal Processing.