Sensor Selection for Relative Acoustic Transfer Function Steered Linearly-Constrained Beamformers

Jie Zhang, *Member, IEEE*, Jun Du, *Senior Member, IEEE*, and Li-Rong Dai

***Abstract*—For multi-microphone speech enhancement, different microphones might have different contributions, assome are even marginal. This is more likely to happen in wireless acoustic sensor networks (WASNs), where somesensors might be distant. In this work, we therefore consider sensor selection for linearly-constrained beamformers. Theproposed sensor selection approach is formulated by minimizing the total output noise power and constraining thenumber of selected sensors. As the considered sensor selection problem requires the relative acoustic transfer function(RTF), the covariance whitening based RTF estimation or a direct-path RTF approximation is exploited. For a singletarget source, we can thus substitute the estimated RTF or the assumed RTF to the original problem formulation in orderto design a minimum variance distortionless response (MVDR) beamformer. Alternatively, we can integrate the two RTFsto design a linearly constrained minimum variance (LCMV) beamformer in order to alleviate the effects of RTFestimation/approximation errors. By leveraging the superiority of LCMV beamformers, the proposed approach can beapplied to the multi-source case. An evaluation using a simulated large-scale WASN demonstrates that the integration ofRTFs for the sensor selection based LCMV beamformer can be beneficial as opposed to relying on either of theindividual RTF steered sensor selection based MVDR beamformers. We conclude that the sensors that are close to thetarget source(s) and also some around the coherent interferers are more informative.**

***Index Terms*—Beamformers, convex optimization, covariance whitening, relative acoustic transfer function, sensor selection, speech enhancement, wireless acoustic sensor networks.**

I. Introduction

MICROPHONE arrays are frequently deployed in various audio applications, e.g., hearing aids (HAs) [1], teleconferencing systems [2], hands-free telephony [3], speech recognition [4], human-robot interaction [5], etc. Although the traditional array system has been widely studied over the past few decades, its configuration brings several limitations, leading to a bottleneck with respect to the speech processing performance. Usually, conventional array systems are equipped with multiple microphones, which are physically linked to a central computing unit. Rearranging such a wired and centralized array system (e.g., including a new microphone or removing a useless microphone) seems impractical. The spatial sampling capability is limited, as the location of the microphone arrays cannot be changed easily. In case the microphones are distant from the target speaker, low-quality audio recordings are obtained and the system performance degrades. Moreover, the size of the arrays should be determined by the application scenarios, for example, only a small array consisting of 2-4 microphones can be equipped by each HA.

Nowadays, with the increased popularity of using wireless devices, e.g., laptops, smartphones, we are surrounded by wireless acoustic sensor networks (WASNs). In WASNs, each node can be mounted with a single microphone or a small microphone array. Due to the capability of wireless communication, the sensor network can be organized more flexibly, either in a centralized fashion or in a distributed way [6]–[9]. The utilization of WASNs for speech processing can potentially resolve the limitations within the conventional microphone array systems. For instance, as the wireless devices can be distributed anywhere, they might be very close to the target source, resulting in high-quality recordings which are beneficial for speech enhancement. Even though the HAs can only host a rather limited number of microphones, if the external wireless devices share their measurements with the HAs, they are able to make use of more data, leading to a performance improvement [10]–[12]. However, incorporating more external sensors as a WASN in return requires a higher power consumption and computational complexity. In this context, the challenges need to be addressed are 1) *how to optimally select the most informative subset of sensors from a large-scale WASN?* and 2) *how to reconstruct the target speech signal from the incomplete observations over the WASN?*

The concept of sensor selection originates from wireless sensor networks (WSNs) [13]. Mathematically, it can be formulated by optimizing a certain performance measure subject to a constraint on the cardinality of the selected subset, or in the other way around. In principle, sensor selection is a combinatorial optimization problem. In order to perform sensor selection efficiently, some convex relaxation techniques [13]–[15] or greedy heuristics (e.g., submodularity) [16] should be leveraged. Using the selected subset of sensors can still perform source localization [14], field estimation [17], target tracking [15], etc, yet the resource consumption is saved, as much less sensors are involved. In WASNs, there are also some sensor selection algorithms that have been proposed recently for e.g., speech

Manuscript received October 13, 2020; revised January 26, 2021; accepted March 3, 2021. Date of publication March 8, 2021; date of current version March 26, 2021. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant WK2100000016, in part by the National Key R&D Program of China under Grant 2017YFB1002202, and in part by the Leading Plan of CAS under Grant XDC08010200. The associate editor coordinating the review of this manuscript and approving it for publication was M. M. Doss. *(Corresponding author: Jie Zhang.)*

The authors are with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China (USTC), Hefei 230026, China (e-mail: jzhang6@ustc.edu.cn; jundu@ustc.edu.cn; lrdai@ustc.edu.cn).

Digital Object Identifier 10.1109/TASLP.2021.3064399

enhancement [18]–[20], speech recognition [21], and speaker tracking [22]. It was shown in [20] that sensor selection is effective in reducing the power consumption and computational complexity for WASNs with a tolerable scarification of noise reduction performance.

A. Contributions

In this work, we consider to select the most informative subset of sensors from a large-scale WASN for linearly constrained beamformers based speech enhancement, such that the target speech signal can be estimated using a subset of microphone observations. The considered sensor selection problem is formulated by minimizing the total output noise power and constraining the number of selected sensors. Since the audio recordings across different microphones are highly correlated, using all measurements from the complete network is unnecessary and a large amount of resource consumption is required. By using the proposed sensor selection algorithm, the data redundancy can thus be removed to some extent and the resource consumption can be saved.

We begin with the sensor selection in the context of minimum variance distortionless response (MVDR) beamforming (SS-MVDR) for a single source estimation. As the original SS-MVDR optimization problem requires the relative acoustic transfer function (RTF), we consider two variants: 1) using an estimated RTF and 2) exploiting an assumed RTF, as the RTF can be approximated using *a priori* information. As there exist estimation errors in the estimated RTF and approximation errors in the assumed RTF, using the respective RTF would affect the optimality of sensor selection and further degrade the speech enhancement performance. Then, we integrate the two constraints associated with the estimated and assumed RTFs to design a linearly constrained minimum variance (LCMV) beamformer and consider the sensor selection criterion for such an LCMV beamformer (SS-LCMV). Both SS-MVDR and SS-LCMV can be derived as semi-definite programming problems using convex optimization techniques.

Further, we analyze that for the single target source case the obtained SS-LCMV can be approximated as an integration of two SS-MVDRs, but SS-LCMV is more robust against the RTF estimation/approximation errors. The selected subset obtained by SS-LCMV can be viewed as the intersection set between the selected subsets obtained by the two SS-MVDR methods. In case the estimated (or assumed) RTF is more reliable, the selected subset of SS-LCMV will be more similar to that by the estimated (or assumed) RTF steered SS-MVDR. Therefore, compared to SS-MVDR, the proposed SS-LCMV method is able to automatically check the reliability of involved RTFs. Due to the fact that the LCMV beamformer can handle multiple sources, we further apply the proposed SS-LCMV algorithm to a multi-source case, where the RTFs are estimated sequentially. Experimental results using a simulated WASN validate the proposed approaches. We find that the sensors close to the existing sources are more likely to be selected, as the sensors around the target source(s) are helpful for enhancing the target signal(s) and those close to the interfering source(s) (even having a low signal-to-noise ratio) are beneficial for suppressing the noise signal(s). Using the integrated RTFs can refine the selected subset of sensors and improve the noise reduction performance.

In [20], we proposed a microphone subset selection method for MVDR (MSS-MVDR) beamformer based noise reduction, which is formulated by minimizing the total power consumption over the WASN and constraining the desired noise reduction performance. In principle, MSS-MVDR is a special case of the proposed SS-MVDR problem, as the total power consumption is directly linked to the cardinality of the selected subset. On the other hand, MSS-MVDR was solved based on the assumption that the acoustic transfer function (ATF) of a single source is given. The ATF estimation error would affect the selection of MSS-MVDR significantly. Therefore, the proposed SS-MVDR can be seen as a generalization of MSS-MVDR, and the proposed SS-LCMV is an extension of MSS-MVDR, which is more practical and applicable to a more dynamic scenario. Compared to the utility-based sensor selection method that was proposed in [18], [23], the proposed method can achieve a better noise reduction performance in case the number of the selected sensors is fixed.

B. Outline and Notations

The rest of this paper is structured as follows. Section II introduces the required preliminary knowledge, including signal model, the covariance whitening based RTF estimation method and linearly-constrained beamformers (e.g., MVDR, LCMV). Section III presents the proposed sensor selection method in the context of MVDR beamforming. In Section IV, we extend the proposed method to a more general LCMV framework with a single source and multiple sources being taken into account. Section V presents the experimental results using a simulated WASN. Finally, Section VI concludes this work.

The notation used in this paper is as follows: Upper (lower) bold face letters are used for matrices (column vectors). $(\cdot)^T$ or $(\cdot)^H$ denotes (vector/matrix) transposition or conjugate transposition. $\mathbb{E}(\cdot)$ denotes the mathematical expectation operation. $\mathrm{diag}(\cdot)$ refers to a block diagonal matrix with the elements in its argument on the main diagonal. $\mathbf{1}_N$ and $\mathbf{O}_N$ denote the $N \times 1$ vector of ones and the $N \times N$ matrix with all its elements equal to zero, respectively. $\mathbf{I}_N$ is an identity matrix of size N. $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. $|\mathcal{U}|$ denotes the cardinality of the set $\mathcal{U}$.

II. Fundamentals

A. Signal Model

In this work, we consider a WASN consisting of M spatially distributed acoustic sensor nodes, which are exploited for sampling and monitoring the acoustic scene of interest. Without loss of generality, we assume that each node is composed of a single microphone. Letting l and ω, respectively, denote the frame index and the frequency index, in the short-time Fourier transform (STFT) domain, the recorded signal at the kth microphone, say $Y_k(\omega, l)$, can be written as

$$Y_k(\omega, l) = X_k(\omega, l) + N_k(\omega, l),\ k = 1, \ldots, M, \qquad (1)$$

where $X_k(\omega, l)$ denotes the target signal component at microphone k, and $N_k(\omega, l)$ the noise component recorded by the kth microphone, which can incorporate coherent noise sources (e.g., competing speakers) and incoherent noises (e.g., late reverberation of the target source, sensor thermal noise).

For the single point source case, which is characterized by the ATF $a_k(\omega)$ (relating the target source position to microphone k), the signal component is then given by

$$X_k(\omega, l) = a_k(\omega) S(\omega, l), \tag{2}$$

where $S(\omega, l)$ denotes the target signal. Note that (2) holds under the assumption that the target source keeps static, implying that the ATF of this source is time-invariant. Without loss of generality, taking the first microphone as the reference microphone, which can be chosen using a more sophisticated method in [24], and defining the RTF as

$$h_k(\omega) = a_k(\omega)/a_1(\omega), \tag{3}$$

then the signal component equals

$$X_k(\omega, l) = h_k(\omega) X_1(\omega, l), \tag{4}$$

where $X_1(\omega, l) = a_1(\omega) S(\omega, l)$. The introduction of RTF is due to the fact that in practice RTF can be estimated using covariance subtraction or covariance whitening method [25]–[27], however directly estimating ATF is still unknown. More importantly, the utilization of RTF does not degrade the beamforming performance. For notational conciseness, we will omit the frame and frequency indexes in the sequel bearing in mind that all the following operations are realized in the STFT domain. Let the vector $\mathbf{y}$ stack the microphone measurements for each time-frequency bin, i.e., $\mathbf{y} = [Y_1, Y_2, \ldots, Y_M]^T$. Similarly, we define the vectors $\mathbf{x}$, $\mathbf{n}$, $\mathbf{a}$ and $\mathbf{h}$ for stacking the signal components, noise components, the ATF and the RTF, respectively, such that the considered signal model can also be given by

$$\mathbf{y} = \mathbf{h} X_1 + \mathbf{n}. \tag{5}$$

Furthermore, we assume that the target source and the noise components are mutually uncorrelated, such that the relationship between the second-order statistics can be formulated as

$$\begin{aligned} \mathbf{\Phi_{yy}} &= \mathbb{E}\left\{\mathbf{y}\mathbf{y}^H\right\} = \mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\} + \mathbb{E}\left\{\mathbf{n}\mathbf{n}^H\right\} \\ &= \mathbf{\Phi_{xx}} + \mathbf{\Phi_{nn}}, \end{aligned} \tag{6}$$

where $\mathbf{\Phi_{xx}}$ and $\mathbf{\Phi_{nn}}$ denote the correlation matrix of the signal components and the correlation matrix of the noise components, respectively. For a single target source case, $\mathbf{\Phi_{xx}}$ is a rank-1 matrix in theory, since by definition we have

$$\mathbf{\Phi_{xx}} \triangleq \sigma_S^2 \mathbf{a}\mathbf{a}^H \triangleq \sigma_{X_1}^2 \mathbf{h}\mathbf{h}^H, \tag{7}$$

where $\sigma_S^2 = \mathbb{E}\{|S|^2\}$ and $\sigma_{X_1}^2 = \mathbb{E}\{|X_1|^2\}$ denote the power spectral density (PSD) of the target source and the PSD of the signal component at the reference microphone, respectively. In practice, these correlation matrices can be estimated using the average smoothing technique. For instance, given a perfect voice activity detector (VAD), the microphone signal can be classified into speech-absent frames and speech-plus-noise frames. During the two periods, the noise and noisy correlation matrices can be estimated, respectively.

The key procedure of the linearly constrained beamforming technique is to design a linear filter $\mathbf{w} = [w_1, w_2, \ldots, w_M]^T$, such that the estimated target signal at the chosen reference microphone can be obtained through beamforming as

$$\hat{X}_1 = \mathbf{w}^H \mathbf{y}. \tag{8}$$

B. Covariance Whitening Based RTF Estimation (CW-RTF)

For the design of such a beamformer, the RTF estimate is required. Among various RTF estimation approaches, it was shown in [27] that the eigen-decomposition based covariance whitening method has a superiority in performance, particularly in noisy and strong reverberant environments. However, it is more complex compared to, e.g., covariance subtraction, from the perspective of implementation, since the time consuming matrix inversion or matrix decomposition is involved. In this work, in order to alleviate the impact of RTF estimation on sensor selection based noise reduction, we adopt the covariance whitening method to estimate the RTF vector.

Given the microphone measurements $\mathbf{y}$ per time-frequency bin and the estimated noise correlation matrix $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$, the covariance whitening method uses $\hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1/2}$ to whiten $\mathbf{y}$ as

$$\bar{\mathbf{y}} = \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1/2} \mathbf{y}, \tag{9}$$

where $\hat{\mathbf{\Phi}}_{\mathbf{nn}}^{1/2}$ is the square root of $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$. Then, the correlation matrix of the whitened signals can be calculated using the average smoothing technique as

$$\hat{\mathbf{\Phi}}_{\bar{\mathbf{y}}\bar{\mathbf{y}}} = \frac{1}{L_y} \sum_{l=1}^{L_y} \bar{\mathbf{y}}(l) \bar{\mathbf{y}}(l)^H \tag{10}$$

$$\triangleq \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1/2} \hat{\mathbf{\Phi}}_{\mathbf{yy}} \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-H/2}, \tag{11}$$

where L_y denotes the number of speech-plus-noise segments. Note that $\hat{\mathbf{\Phi}}_{\bar{\mathbf{y}}\bar{\mathbf{y}}}$ and $\hat{\mathbf{\Phi}}_{\mathbf{yy}}$ should be estimated using the same frame set. Let $\boldsymbol{\phi}$ denote the principal eigenvector of $\hat{\mathbf{\Phi}}_{\bar{\mathbf{y}}\bar{\mathbf{y}}}$, i.e.,

$$\hat{\mathbf{\Phi}}_{\bar{\mathbf{y}}\bar{\mathbf{y}}} \boldsymbol{\phi} = \lambda_{\max} \boldsymbol{\phi}, \tag{12}$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\hat{\mathbf{\Phi}}_{\bar{\mathbf{y}}\bar{\mathbf{y}}}$. With $\boldsymbol{\phi}$, the covariance whitening based RTF estimate is given by

$$\hat{\mathbf{h}}_{\mathrm{CW}} = \frac{\hat{\mathbf{\Phi}}_{\mathbf{nn}}^{1/2} \boldsymbol{\phi}}{\mathbf{e}_1^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{1/2} \boldsymbol{\phi}}, \tag{13}$$

where $\mathbf{e}_1$ is a column vector with the first entry equal to one and zeros elsewhere. Notably, it is easy to check that the covariance whitening based RTF estimate is equivalent to the normalized principal eigenvector of the generalized eigenvalue decomposition (GEVD) of the matrix pencil $\{\hat{\mathbf{\Phi}}_{\mathbf{yy}}, \hat{\mathbf{\Phi}}_{\mathbf{nn}}\}$.

C. MVDR Beamformer

The well-known MVDR beamformer is formulated by minimizing the output noise power under a linear constraint, which is exploited for preserving the signal power that comes from the direction of interest. Mathematically, it can be designed by considering the following constrained optimization problem:

$$\mathbf{w}_{\mathrm{MVDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{h} = 1. \tag{14}$$

With such a linear constraint and using (7), we can see that $\mathbf{w}_{\mathrm{MVDR}}^H \mathbf{\Phi_{xx}} \mathbf{w}_{\mathrm{MVDR}} = \sigma_{X_1}^2$, which implies that the power of

the desired signal component at the reference microphone is preserved. Any reduction of the objective function is caused by reducing the noise power. Given the true RTF vector $\mathbf{h}$ and using the technique of Lagrangian multipliers, the MVDR beamformer can be shown to be given by [28], [29]

$$\mathbf{w}_{\mathrm{MVDR}} = (\mathbf{h}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{h})^{-1} \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{h}. \tag{15}$$

The output noise power of MVDR beamformer is given by

$$\mathbf{w}_{\mathrm{MVDR}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{w}_{\mathrm{MVDR}} = (\mathbf{h}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{h})^{-1}, \tag{16}$$

and the output signal-to-noise ratio (SNR) is given by

$$\mathrm{oSNR}_{\mathrm{MVDR}} = \sigma_{X_1}^2 \mathbf{h}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{h}. \tag{17}$$

Obviously, the optimal design of the MVDR beamformer is dependent on the true RTF, which is unknown, the optimal MVDR beamformer is thus impractical to approach. In practice, we can substitute *a priori* information on RTF into (14) to obtain near-optimal, but more practical solutions.

1) MVDR Based on the Estimated RTF (MVDR-EST): One alternative way to design a practical MVDR beamformer is based on the use of the estimated RTF vector. Substitutting the RTF estimate from (13) into (15), the resulting MVDR beamformer is then given by

$$\mathbf{w}_{\mathrm{MVDR-EST}} = (\hat{\mathbf{h}}_{\mathrm{CW}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \hat{\mathbf{h}}_{\mathrm{CW}})^{-1} \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \hat{\mathbf{h}}_{\mathrm{CW}}. \tag{18}$$

2) MVDR Based on an Assumed RTF (MVDR-ASS): In some cases, the RTF of the target source can be approximated by using *a priori* assumptions. For example, for the hearing-aid users, usually the target source is located in the front direction. In this context, the RTF of the target source can be approaximated by gain and delay values as

$$\hat{\mathbf{h}}_{\mathrm{ASS}} = \left[1, g_{21} e^{-j2\pi f \tau_{21}}, \ldots, g_{M1} e^{-j2\pi f \tau_{M1}}\right], \tag{19}$$

where g_{k1} is the attenuation coefficient which depends on the distance between the source position and the microphone pair, and $\tau_{k1}, \forall k$ denotes the time-difference of arrival (TDOA). Therefore, one can use the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$ for a practical MVDR implementation. The resulting near-optimal MVDR beamformer is then given by

$$\mathbf{w}_{\mathrm{MVDR-ASS}} = (\hat{\mathbf{h}}_{\mathrm{ASS}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \hat{\mathbf{h}}_{\mathrm{ASS}})^{-1} \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \hat{\mathbf{h}}_{\mathrm{ASS}}. \tag{20}$$

Note that under the utilization of an estimated RTF or an assumed counterpart, the signal power cannot be exactly preserved any more as using the true RTF, because there exist estimation/approximation errors.

D. LCMV Beamformer

An important limitation within the classic MVDR beamformers is that the distortionless response corresponding to one direction (which is characterized by the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ or the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$) can be preserved. Clearly, in case the mismatch between the involved RTF and the true RTF is large, the performance of the MVDR beamformer will degrade significantly. For this, one can add more linear constraints to the MVDR optimization problems. These linear constraints are associated with multiple directions, such that the distortionless response from more RTFs can be preserved, resulting in an LCMV beamformer as

$$\mathbf{w}_{\mathrm{LCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{w} \quad \text{s.t.} \quad \mathbf{C}^H \mathbf{w} = \mathbf{b}, \tag{21}$$

where the matrix $\mathbf{C} \in \mathbb{C}^{M \times N}$ is constructed from multiple RTFs, and the vector $\mathbf{b} \in \mathbb{C}^N$ is dedicated to enforcing the distortion level for each RTF. Similarly, the LCMV beamformer can be resolved as a close-form solution:

$$\mathbf{w}_{\mathrm{LCMV}} = \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{C} (\mathbf{C}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{C})^{-1} \mathbf{b}. \tag{22}$$

The output noise power of LCMV filter is then given by

$$\mathbf{w}_{\mathrm{LCMV}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{w}_{\mathrm{LCMV}} = \mathbf{b}^H (\mathbf{C}^H \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{C})^{-1} \mathbf{b}. \tag{23}$$

Remark 1: In order to further visualize the LCMV beamformer, one can consider a special case. Given

$$\mathbf{C} = [\hat{\mathbf{h}}_{\mathrm{CW}}, \hat{\mathbf{h}}_{\mathrm{ASS}}], \quad \mathbf{b} = [1, 1]^T, \tag{24}$$

the resulting LCMV beamformer is a linear combination of the two MVDR beamformers based on the use of the estimated and assumed RTFs [30], i.e.,

$$\hat{\mathbf{w}}_{\mathrm{LCMV}} = \alpha_1 \mathbf{w}_{\mathrm{MVDR-EST}} + \alpha_2 \mathbf{w}_{\mathrm{MVDR-ASS}}, \tag{25}$$

where α_1 and α_2 depends on the estimated and assumed RTFs and the noise correlation matrix, and are detailed in [30, Sec. III-C]. Clearly, the LCMV beamformer is a generalization of the MVDR filter. Since the LCMV beamformer takes more constraints into account, less degrees of freedom are left for adjusting the filter coefficients to minimize the noise power.

III. Sensor Selection for MVDR Beamforming

In this section, we will present the proposed SS-MVDR method for speech enhancement using a subset of microphone measurements over a large-scale WASN.

A. Sensor Selection Model

The sensor selection problem is formulated by choosing a best subset of sensors in order to optimize an objective function subject to certain constraints. For this, we first introduce a Boolean selection vector

$$\mathbf{p} = [p_1, p_2, \ldots, p_M]^T \in \{0, 1\}^M,$$

where $p_k = 1, \forall k$ indicates that the kth microphone is selected, and otherwise unselected. Further, we use $K = ||\mathbf{p}||_0$ to represent the number of selected sensors with ℓ_0-norm denoting the number of non-zero elements of a vector. Letting $\mathrm{diag}(\mathbf{p})$ denote a diagonal matrix whose diagonal entries are given by $\mathbf{p}$, we can define a selection matrix $\mathbf{\Sigma}_{\mathbf{p}} \in \{0, 1\}^{K \times M}$ which is obtained by removing the all-zero rows of $\mathrm{diag}(\mathbf{p})$. Clearly, we can obtain the following properties:

$$\mathbf{\Sigma}_{\mathbf{p}} \mathbf{\Sigma}_{\mathbf{p}}^T = \mathbf{I}_K, \quad \mathbf{\Sigma}_{\mathbf{p}}^T \mathbf{\Sigma}_{\mathbf{p}} = \mathrm{diag}(\mathbf{p}). \tag{26}$$

With the selection matrix at hand, we can construct the incomplete audio measurements as

$$\mathbf{y}_{\mathbf{p}} = \mathbf{\Sigma}_{\mathbf{p}} \mathbf{y} = \mathbf{\Sigma}_{\mathbf{p}} \mathbf{x} + \mathbf{\Sigma}_{\mathbf{p}} \mathbf{n} \in \mathbb{C}^K. \tag{27}$$

Similarly, the RTF and noise correlation matrix associated with the selection sensors are given by

$$\mathbf{h}_{\mathbf{p}} = \mathbf{\Sigma}_{\mathbf{p}} \mathbf{h} \in \mathbb{C}^K, \ \hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}} = \mathbf{\Sigma}_{\mathbf{p}} \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{\Sigma}_{\mathbf{p}}^T \in \mathbb{C}^{K \times K}. \tag{28}$$

The MVDR beamformer depending on the selected sensors is then given by

$$\mathbf{w_p} = \left(\mathbf{h_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{h_p}\right)^{-1} \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{h_p}, \tag{29}$$

where we note that in general $\hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \neq \mathbf{\Sigma_p} \hat{\mathbf{\Phi}}_{\mathbf{nn}}^{-1} \mathbf{\Sigma_p}^T$, unless $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$ is diagonal (i.e., the uncorrelated noise case). As a result, the estimated target signal can be obtained as $\hat{X}_1 = \mathbf{w_p}^H \mathbf{y_p}$.

B. Problem Formulation (SS-MVDR)

Given the RTF (e.g., $\hat{\mathbf{h}}_{\mathrm{CW}}$, $\hat{\mathbf{h}}_{\mathrm{ASS}}$), the proposed sensor selection for MVDR beamforming can be formulated by minimizing the total output noise power under a constraint that the signal associated with the considered RTF is undistorted, as the following constrained optimization problem shows

$$\begin{aligned} \min_{\mathbf{p},\mathbf{w_p}} \quad & \mathbf{w_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}} \mathbf{w_p} \\ \text{s.t.} \quad & \mathbf{w_p}^H \mathbf{h_p} = 1 \\ & ||\mathbf{p}||_0 \leq K, \quad \mathbf{p} \in \{0,1\}^M, \end{aligned} \tag{30}$$

where K denotes the maximum number of sensors that can be selected, which can be assigned by users. Obviously, this is a non-convex optimization problem, because of the non-linear selection operation by selection matrix $\mathbf{\Sigma_p}$ and the Boolean variables $\mathbf{p}$. However, we can simplify it for analysis. Considering the Lagrangian function of (30) and calculating the partial derivative with respect to $\mathbf{w_p}$, we find that the MVDR beamformer is the solution. Hence, plugging the MVDR beamformer from (29) into (30), we obtain a simplified sensor selection problem:

$$\begin{aligned} \max_{\mathbf{p}} \quad & \mathbf{h_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{h_p} \\ \text{s.t.} \quad & ||\mathbf{p}||_0 \leq K, \quad \mathbf{p} \in \{0,1\}^M, \end{aligned} \tag{31}$$

By doing this, we can get rid of jointly optimizing the original problem over two variables. The simplified version only needs to consider the selection variable, which indeed is still a non-convex (combinatorial) optimization problem.

C. Convex Solver

In this section, we will resolve the proposed sensor selection problem following convex optimization techniques and using the true RTF $\mathbf{h}$. Note that the proposed solver also applies to the case of using the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ or the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$. First of all, in order to avoid the non-linearity within $\mathbf{h_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{h_p}$, we consider to decompose the matrix $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$ as

$$\hat{\mathbf{\Phi}}_{\mathbf{nn}} = \lambda \mathbf{I} + \mathbf{G}, \tag{32}$$

where the constant λ is positive and $\mathbf{G}$ is a positive semi-definite matrix. Since $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$ is always positive definite in the presence of correlated and uncorrelated noises, we can find such a decomposition via eigenvalue decomposition (EVD) of $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$. For example, λ can be chosen to be the minimum eigenvalue of $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$. Even though λ might be close to zero, as long as $\lambda > 0$, (38) will be always feasible. With λ and $\mathbf{G}$ at hand, it can be seen that

$$\hat{\mathbf{\Phi}}_{\mathbf{nn,p}} = \mathbf{\Sigma_p} \hat{\mathbf{\Phi}}_{\mathbf{nn}} \mathbf{\Sigma_p}^T = \lambda \mathbf{I}_K + \mathbf{\Sigma_p} \mathbf{G} \mathbf{\Sigma_p}^T. \tag{33}$$

Further, the objective function of (31) can be derived as

$$\mathbf{h_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{h_p} = \mathbf{h}^H \underbrace{\mathbf{\Sigma_p}^T \left(\lambda \mathbf{I}_K + \mathbf{\Sigma_p} \mathbf{G} \mathbf{\Sigma_p}^T\right)^{-1} \mathbf{\Sigma_p}}_{\mathbf{Q}} \mathbf{h}. \tag{34}$$

Using the matrix inversion lemma [31]

$$\mathbf{C}(\mathbf{B}^{-1} + \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C})^{-1} \mathbf{C}^T = \mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^T)^{-1} \mathbf{A},$$

the matrix $\mathbf{Q}$ can be represented as

$$\mathbf{Q} = \mathbf{G}^{-1} - \mathbf{G}^{-1} \left(\mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p})\right)^{-1} \mathbf{G}^{-1}. \tag{35}$$

With $\mathbf{Q}$, we can equivalently re-write (31) in an epigraph form as [32]

$$\begin{aligned} \max_{\mathbf{p},\eta} \quad & \eta \\ \text{s.t.} \quad & \eta \leq \mathbf{h}^H \mathbf{Q} \mathbf{h} \\ & ||\mathbf{p}||_0 \leq K, \quad \mathbf{p} \in \{0,1\}^M, \end{aligned} \tag{36}$$

Substituting $\mathbf{Q}$ from (35) into the constraint $\eta \leq \mathbf{h}^H \mathbf{Q} \mathbf{h}$, we obtain

$$\mathbf{h}^H \mathbf{G}^{-1} \mathbf{h} - \eta \geq \mathbf{h}^H \mathbf{G}^{-1} \left(\mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p})\right)^{-1} \mathbf{G}^{-1} \mathbf{h},$$

which can be reformulated as a symmetric linear matrix inequality (LMI) [32] using the Schur complement

$$\begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1} \mathbf{h} \\ \mathbf{h}^H \mathbf{G}^{-1} & \mathbf{h}^H \mathbf{G}^{-1} \mathbf{h} - \eta \end{bmatrix} \succeq \mathbf{O}_{M+1}, \tag{37}$$

due to the fact that the matrix $\mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p})$ is always positive definite with a positive λ and a positive semi-definite matrix $\mathbf{G}^{-1}$.

Now the non-convexity of (31) lies in the ℓ_0-norm and the Boolean constraint. For the ℓ_0-norm, one alternative is by using the ℓ_1-norm which is convex to relax it. For the Boolean constraint, it can be relaxed using continuous surrogates or semi-definite relaxation [33]. In this work, we will relax $p_k \in \{0,1\}$ to $0 \leq p_k \leq 1, \forall k$. To this end, we can represent the original SS-MVDR problem as

$$\begin{aligned} \max_{\mathbf{p},\eta} \quad & \eta \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1} \mathbf{h} \\ \mathbf{h}^H \mathbf{G}^{-1} & \mathbf{h}^H \mathbf{G}^{-1} \mathbf{h} - \eta \end{bmatrix} \succeq \mathbf{O}_{M+1} \\ & \mathbf{1}_M^T \mathbf{p} \leq K, \quad 0 \leq p_k \leq 1, \forall k, \end{aligned} \tag{38}$$

which is a semi-definite programming problem and can be efficiently solved in polynomial time using interior-point methods or some off-the-shelf solvers, like CVX [34] or SeDuMi [35]. The computational complexity of (38) is cubic in terms of M. The final Boolean solution can be obtained by randomized rounding or deterministic rounding techniques. Note that (38) is a general sensor selection problem for the RTF-steered MVDR beamformer. In practice, given the noise statistics, we can substitute the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ or the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$ into (38) to solve a specific, but more practical problem.

IV. Sensor Selection for LCMV Beamforming

Since using the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ or the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$ might distort the target signal, similar to SS-MVDR we thus propose a sensor selection based LCMV (SS-LCMV) beamformer in this section.

A. General Sensor Selection for LCMV (SS-LCMV)

The considered sensor selection based LCMV beamformer is formulated by minimizing the total output noise power under a set of linear constraints together with the cardinality constraint, as the following optimization problem shows

$$\begin{aligned} \min_{\mathbf{p},\mathbf{w_p}} \quad & \mathbf{w_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}} \mathbf{w_p} \\ \text{s.t.} \quad & \mathbf{C_p}^H \mathbf{w_p} = \mathbf{b} \\ & ||\mathbf{p}||_0 \leq K, \quad \mathbf{p} \in \{0,1\}^M, \end{aligned} \tag{39}$$

where $\mathbf{C_p} = \mathbf{\Sigma_p C} \in \mathbb{C}^{K \times N}$. Again, (39) is a non-convex combinatorial optimization problem. In order to find an efficient solver for (39), similarly to Section III, we consider its Lagrange function and derive the partial derivative with respect to $\mathbf{w_p}$. The resulting beamformer $\mathbf{w_p}$ is then given by

$$\mathbf{w_p} = \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} \left(\mathbf{C_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} \right)^{-1} \mathbf{b}, \tag{40}$$

which is the classic LCMV beamformer given in (22), but now depends on the selected subset of sensors. Under the utilization of such an LCMV beamformer, the output noise power in the objective function of (39) is given by

$$\mathbf{w_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}} \mathbf{w_p} = \mathbf{b}^H \left(\mathbf{C_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} \right)^{-1} \mathbf{b}. \tag{41}$$

Substituting (41) into (39), we can therefore get rid of optimizing the filter coefficients, and only the selection variable is unknown. The original problem can then be simplified as

$$\begin{aligned} \min_{\mathbf{p}} \quad & \mathbf{b}^H \left(\mathbf{C_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} \right)^{-1} \mathbf{b} \\ \text{s.t.} \quad & ||\mathbf{p}||_0 \leq K, \quad \mathbf{p} \in \{0,1\}^M. \end{aligned} \tag{42}$$

By introducing a new variable η, (42) can be equivalently reformulated in the following epigraph form:

$$\begin{aligned} \min_{\mathbf{p},\eta} \quad & \eta \\ \text{s.t.} \quad & \mathbf{b}^H \left(\mathbf{C_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} \right)^{-1} \mathbf{b} \leq \eta \\ & ||\mathbf{p}||_0 \leq K, \quad \mathbf{p} \in \{0,1\}^M. \end{aligned} \tag{43}$$

In order to linearize the first constraint in (43), we introduce a symmetric positive semi-definite matrix $\mathbf{T} \in \mathbb{S}_+^N$, such that it can be relaxed as two new constraints:

$$\mathbf{b}^H \mathbf{T}^{-1} \mathbf{b} \leq \eta, \tag{44}$$

$$\mathbf{C_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} \succeq \mathbf{T}. \tag{45}$$

Clearly, (44) and (45) are sufficient to obtain the first constraint in (43). Furthermore, using the Schur complement (44) can be reformulated as an LMI

$$\begin{bmatrix} \mathbf{T} & \mathbf{b} \\ \mathbf{b}^H & \eta \end{bmatrix} \succeq \mathbf{O}_{N+1}. \tag{46}$$

The left-hand side of (45) can be shown to be given by

$$\mathbf{C_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \mathbf{C_p} = \mathbf{C}^H \mathbf{Q} \mathbf{C}, \tag{47}$$

where $\mathbf{Q}$ is given by (see Section III-C)

$$\mathbf{Q} = \mathbf{G}^{-1} - \mathbf{G}^{-1} \left(\mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) \right)^{-1} \mathbf{G}^{-1}.$$

Hence, (45) can be re-written as

$$\mathbf{C}^H \mathbf{G}^{-1} \mathbf{C} - \mathbf{T} \succeq \mathbf{C}^H \mathbf{G}^{-1} \left(\mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) \right)^{-1} \mathbf{G}^{-1} \mathbf{C}, \tag{48}$$

which can further be reformulated as an LMI:

$$\begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1} \mathbf{C} \\ \mathbf{C}^H \mathbf{G}^{-1} & \mathbf{C}^H \mathbf{G}^{-1} \mathbf{C} - \mathbf{T} \end{bmatrix} \succeq \mathbf{O}_{M+N}. \tag{49}$$

In addition, we relax the cardinality constraint ℓ_0-norm in (43) using the corresponding ℓ_1-norm, and relax the Boolean constraint using the box counterpart, such that (43) can be reformulated in a semi-definite programming form as

$$\begin{aligned} \min_{\mathbf{p},\eta,\mathbf{T}} \quad & \eta \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{T} & \mathbf{b} \\ \mathbf{b}^H & \eta \end{bmatrix} \succeq \mathbf{O}_{N+1} \\ & \begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1} \mathbf{C} \\ \mathbf{C}^H \mathbf{G}^{-1} & \mathbf{C}^H \mathbf{G}^{-1} \mathbf{C} - \mathbf{T} \end{bmatrix} \succeq \mathbf{O}_{M+N} \\ & \mathbf{1}_M^T \mathbf{p} \leq K, \quad 0 \leq p_k \leq 1, \forall k, \end{aligned} \tag{50}$$

which can be efficiently solved by exploiting convex optimization techniques as before. Note that the computational complexity of solving (50) is of the order of $\mathcal{O}((M+N)^3)$. The final Boolean selection variables should be recovered by using rounding techniques.

B. Relation to SS-MVDR

For the single target source case, either applying the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ or using the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$ to the SS-MVDR optimization problem in Section III might not achieve the best subset of sensors, leading to a decrease in the noise reduction performance, due to the errors between the involved RTFs and the true one. In this case, using both RTFs and considering the sensor selection for LCMV beamforming, we can obtain an instantiation of (50). Let

$$\mathbf{C_p} = [\hat{\mathbf{h}}_{\mathrm{CW,p}}, \hat{\mathbf{h}}_{\mathrm{ASS,p}}] = [\mathbf{\Sigma_p} \hat{\mathbf{h}}_{\mathrm{CW}}, \mathbf{\Sigma_p} \hat{\mathbf{h}}_{\mathrm{ASS}}], \quad \mathbf{b} = [1,1]^T,$$

The corresponding LCMV beamformer can be computed as

$$\mathbf{w_p} = \alpha_1 \mathbf{w}_{1,\mathbf{p}} + \alpha_2 \mathbf{w}_{2,\mathbf{p}}, \tag{51}$$

where the weights α_1 and α_2 can be calculated [30], and the respective MVDR beamformers are given by

$$\mathbf{w}_{1,\mathbf{p}} = \frac{\hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \hat{\mathbf{h}}_{\mathrm{CW,p}}}{\hat{\mathbf{h}}_{\mathrm{CW,p}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \hat{\mathbf{h}}_{\mathrm{CW,p}}}, \quad \mathbf{w}_{2,\mathbf{p}} = \frac{\hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \hat{\mathbf{h}}_{\mathrm{ASS,p}}}{\hat{\mathbf{h}}_{\mathrm{ASS,p}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \hat{\mathbf{h}}_{\mathrm{ASS,p}}}.$$

Applying such an LCMV beamformer, the resulting total output noise power can be calculated as

$$\mathbf{w_p}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}} \mathbf{w_p} = \frac{|\alpha_1|^2}{\hat{\mathbf{h}}_{\mathrm{CW,p}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \hat{\mathbf{h}}_{\mathrm{CW,p}}} + \frac{|\alpha_2|^2}{\hat{\mathbf{h}}_{\mathrm{ASS,p}}^H \hat{\mathbf{\Phi}}_{\mathbf{nn,p}}^{-1} \hat{\mathbf{h}}_{\mathrm{ASS,p}}}$$

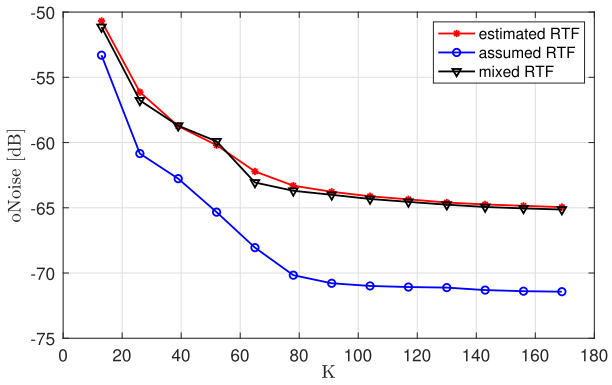

Fig. 1. The residual noise power (in dB) in terms of the number of the selected sensor using different RTFs.

$$+2\mathcal{R}\left(\frac{\alpha_1^*\alpha_2\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}}{\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}}\right), \tag{52}$$

where the operation $\mathcal{R}(\cdot)$ extracts the real part of a complex number, and the first (or second) term represents the residual noise power using the estimated (or assumed) RTF steered MVDR beamformer. The third term denotes the residual noise power using the mixed RTF. In order to analyze the function of each term, we consider the experimental setup as shown in Fig. 2(a) and randomly select K sensors to perform beamforming. The estimated RTF is obtained using the covariance whitening method, and the assumed RTF is calculated using (19), so the assumed RTF is much more accurate than the estimated one. In Fig. 1, we show the residual noise power in terms of the number of selected sensors using different RTFs. It is clear that in case the assumed RTF is more accurate, the noise power obtained by using the mixed RTF approaches that obtained using the estimated RTF. In case the estimated RTF is more accurate, the noise power can be compared similarly. Therefore, we can approximate $\mathbf{w}_\mathbf{p}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}\mathbf{w}_\mathbf{p}$ using

$$\mathbf{w}_\mathbf{p}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}\mathbf{w}_\mathbf{p}\approx\frac{(1+\mu_1)|\alpha_1|^2}{\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}}+\frac{(1+\mu_2)|\alpha_2|^2}{\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}}, \tag{53}$$

where μ_1 and μ_2 denote the confidence level of $\hat{\mathbf{h}}_{\mathrm{CW}}$ and $\hat{\mathbf{h}}_{\mathrm{ASS}}$, respectively, and $\mu_1+\mu_2=1$. In case the assumed RTF is more accurate, $\mu_1>\mu_2$; otherwise $\mu_2>\mu_1$. Note that the introduction of this approximation is to find the link between the proposed SS-MVDR and SS-LCMV methods. A more accurate derivation for (52) is left to the reader. Substituting (51) and (53) into the general LCMV problem description in (39), we arrive at

$$\begin{aligned}\min_{\mathbf{p}}\quad&\frac{(1+\mu_1)|\alpha_1|^2}{\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}}+\frac{(1+\mu_2)|\alpha_2|^2}{\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}}\\ \text{s.t.}\quad&||\mathbf{p}||_0\le K,\quad\mathbf{p}\in\{0,1\}^M,\end{aligned} \tag{54}$$

which can further be re-written in an epigraph form as

$$\begin{aligned}\min_{\mathbf{p},\eta_1,\eta_2}\quad&\frac{(1+\mu_1)|\alpha_1|^2}{\eta_1}+\frac{(1+\mu_2)|\alpha_2|^2}{\eta_2}\\ \text{s.t.}\quad&\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{CW},\mathbf{p}}\ge\eta_1\\ &\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}^H\hat{\mathbf{\Phi}}_{\mathbf{nn},\mathbf{p}}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS},\mathbf{p}}\ge\eta_2\\ &||\mathbf{p}||_0\le K,\quad\mathbf{p}\in\{0,1\}^M.\end{aligned} \tag{55}$$

Based on the decomposition of the matrix $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$ and the introduction of the matrix $\mathbf{Q}$, the two inequality constraints in (55) can be reformulated as two LMIs (similar to (37)). Following the convex relaxation strategies in Section III-C, (55) can be relaxed as a semi-definite programming problem:

$$\begin{aligned}\min_{\mathbf{p},\eta_1,\eta_2}\quad&\frac{(1+\mu_1)|\alpha_1|^2}{\eta_1}+\frac{(1+\mu_2)|\alpha_2|^2}{\eta_2}\\ \text{s.t.}\quad&\begin{bmatrix}\mathbf{G}^{-1}+\lambda^{-1}\mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1}\hat{\mathbf{h}}_{\mathrm{CW}}\\ \hat{\mathbf{h}}_{\mathrm{CW}}^H\mathbf{G}^{-1} & \hat{\mathbf{h}}_{\mathrm{CW}}^H\mathbf{G}^{-1}\hat{\mathbf{h}}_{\mathrm{CW}}-\eta_1\end{bmatrix}\succeq\mathbf{O}\\ &\begin{bmatrix}\mathbf{G}^{-1}+\lambda^{-1}\mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS}}\\ \hat{\mathbf{h}}_{\mathrm{ASS}}^H\mathbf{G}^{-1} & \hat{\mathbf{h}}_{\mathrm{ASS}}^H\mathbf{G}^{-1}\hat{\mathbf{h}}_{\mathrm{ASS}}-\eta_2\end{bmatrix}\succeq\mathbf{O}\\ &\mathbf{1}_M^T\mathbf{p}\le K,\quad 0\le p_k\le 1,\forall k.\end{aligned} \tag{56}$$

Remark 2: By inspection, (56) can be regarded as an integration of two sensor selection problems, which are designed using the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ and the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$ based MVDR beamformers, respectively. Let the estimated RTF $\hat{\mathbf{h}}_{\mathrm{CW}}$ based MVDR sensor selection problem refer to as SS-MVDR-EST, and the selected subset of sensors from (38) be denoted by $\mathcal{S}_{\mathrm{EST}}$. Let the assumed RTF $\hat{\mathbf{h}}_{\mathrm{ASS}}$ based MVDR sensor selection problem refer to as SS-MVDR-ASS, and the corresponding selected subset be denoted by $\mathcal{S}_{\mathrm{ASS}}$. Further, we refer to (56) which is based on the integration of two RTFs as SS-LCMV-INT, and the selected subset of sensors as $\mathcal{S}_{\mathrm{INT}}$. From the perspective of sensor selection, $\mathcal{S}_{\mathrm{INT}}$ should be the intersection between $\mathcal{S}_{\mathrm{EST}}$ and $\mathcal{S}_{\mathrm{ASS}}$. In order to more clearly see the link between (56) and (38), we can consider two extreme cases. In case $|\alpha_1|^2$ is too small, the second term in the objective function of (56) dominates, then SS-LCMV-INT reduces to SS-MVDR-ASS. This means that the assumed RTF approximates the true RTF well and the estimated one is not trustable. In case $|\alpha_2|^2\to 0$, the first term dominates, then SS-LCMV-INT reduces to SS-MVDR-EST. This means that the covariance whitening method provides a good RTF estimate.

C. Application to the Multi-Source Case

In the presence of multiple target speech sources, which are required to be preserved at the output of a linearly constrained beamformer, we need to design an LCMV beamformer, as the MVDR filter can only handle a single source. We assume that there are $N\ge 2$ sources, and let $\hat{\mathbf{h}}_i,\forall i$ denote the estimated RTF of the ith source with respect to the sensor nodes. Then, we can substitute

$$\mathbf{C}=[\hat{\mathbf{h}}_1,\hat{\mathbf{h}}_2,\ldots,\hat{\mathbf{h}}_N],\quad\mathbf{b}=\mathbf{1}_N, \tag{57}$$

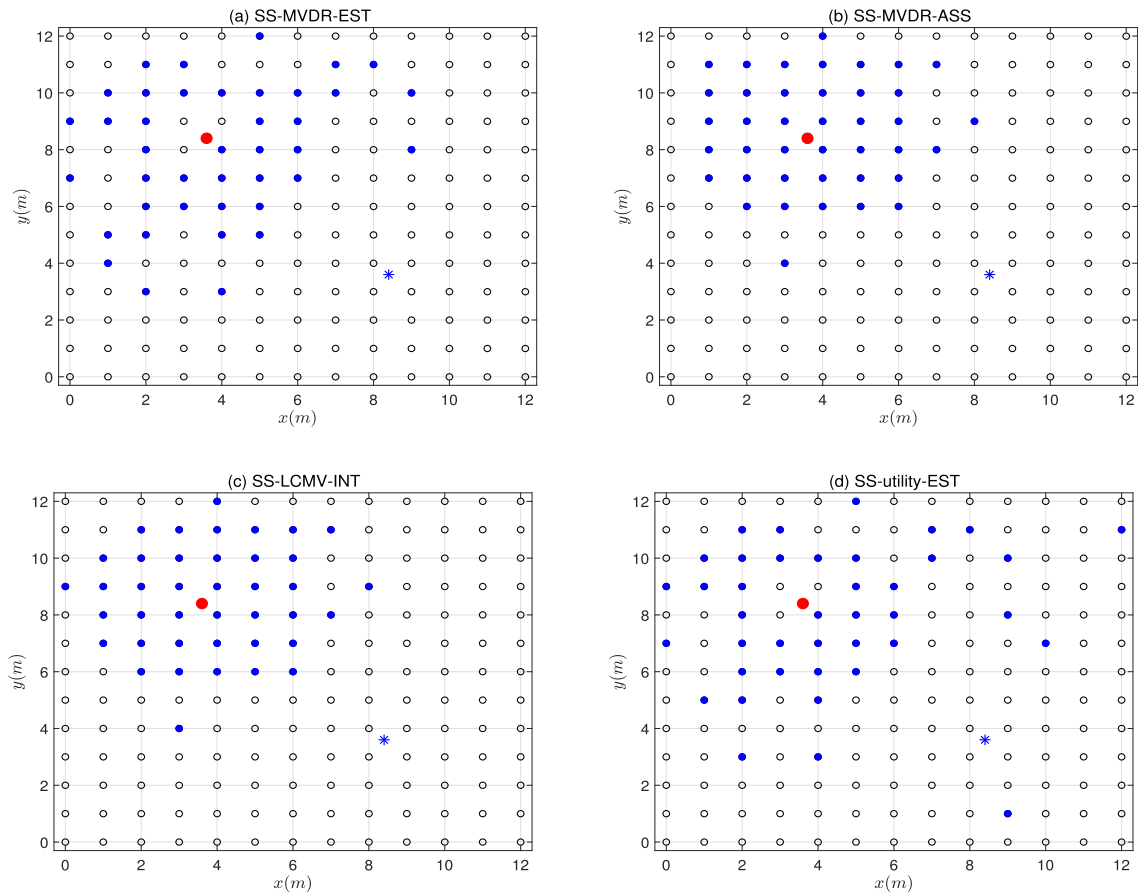

Fig. 2. Sensor selection examples where the blue sensors are activated by different approaches for $K = 40$: (a) SS-MVDR-EST, (b) SS-MVDR-ASS, (c) SS-LCMV-INT, (d) SS-utility-EST, which uses the estimated RTF for sensor selection (the selection result of SS-utility-ASS is similar).

into (50) to obtain the most informative subset of sensors. We refer to this multi-source case as SS-LCMV-N.

From the implementation view of point, the RTF estimation of multiple sources can be estimated using, e.g., [36]. To improve the accuracy, we will estimate the RTF of each source successively in this work. Specifically, given the noise correlation matrix $\hat{\mathbf{\Phi}}_{\mathbf{nn}}$ which is estimated during the training phase and a perfect VAD, we can detect the speech-plus-noise segments, in which only one speech source of interest is active and the other target sources are inactive. During this period, the noisy correlation matrix can be estimated via average smoothing (e.g., Eq. (10)). Using the covariance whitening method, the RTF of this active source can thus be estimated. The RTF of other sources is estimated similarly. Note that for the multiple source case, the number of candidate sensors should be larger than the number of sources, i.e., $M > N$, such that there are $M - N$ degrees of freedom left for adjusting the beamformer coefficients to perform noise reduction.

To this end, we have shown the sensor selection for RTF-steered MVDR and LCMV beamformers. It was shown that using the estimated RTF and the assumed RTF for LCMV is an integration of two individual SS-MVDR solutions. Also, the proposed general LCMV sensor selection algorithm can be extended to the multiple source case. We summarize the proposed algorithms in Table I.

TABLE I
A SUMMARY OF SENSOR SELECTION FOR RTF-STEERED LINEARLY CONSTRAINED BEAMFORMERS

Method	Beamformer	Solver	Selected subset
SS-MVDR-EST	Eq. (18)	Eq. (38)	$\mathcal{S}_{\mathrm{EST}}$
SS-MVDR-ASS	Eq. (19)	Eq. (38)	$\mathcal{S}_{\mathrm{ASS}}$
SS-LCMV-INT	Eq. (51)	Eq. (50, 56)	$\mathcal{S}_{\mathrm{INT}}$
SS-LCMV-N	Eq. (22)	Eq. (50)	$\mathcal{S}_N$

V. EXPERIMENTS

In this section, we will validate the proposed approaches via numerical simulations. At first, we will present the experimental setup and the comparison approaches. Then the proposed methods will be applied to the single target source case. Finally, we will consider the application of the proposed SS-LCMV-N method to the multi-source scenario.

Experimental setup: Fig. 2 shows the typical experimental setting that we use in the simulations. We consider a 2D room with dimensions (12×12) m, where 169 candidate microphones are uniformly distributed. All the speech sources are originated from the TIMIT database [37], and all the noise signals from the NoiseX-92 database [38]. The room impulse responses (RIRs) of directional sources are generated using the toolbox [39]. The measurements of each microphone are synthesized by summing:

1) the source component (convolving the source signal and its RIR), 2) interference component (convolving the interferer (i.e., a competing speaker) and the corresponding RIR) and 3) the uncorrelated sensor noise (i.e., microphone self noise). The uncorrelated noise is modeled as a white Gaussian noise. The signal to interferer ratio (SIR) and the signal to uncorrelated noise ratio (SNR) are set to be 0 dB and 50 dB, respectively. The final signal-to-interferer-noise ratio (SINR) is around -2 dB. All the signals are sampled at 16 kHz. The signals are segmented using a square-root-Hann window with a length of 32 ms and 50% overlap. The reverberation time is set to be $T_{60} = 200$ ms. In order to focus on the sensor selection problem, the microphone signals are synchronized already in this work, and the noise correlation matrix is estimated during 15 seconds speech-absent period before performing the online sensor selection algorithms (e.g., using the average smoothing technique or [40]). The noise source positions are assumed to be static.

Comparison methods: In [18], [23], a utility-based sensor selection approach was proposed. Since in principle sensor selection is a combinatorial optimization problem, the utility-based method greedily removes the sensor that has the least contribution to the noise reduction task from the total sensor set (i.e., backward selection), or adds the sensor that has the largest contribution to the selected subset (i.e., forward selection). Obviously, the utility-based method can only determine the status of one sensor at each iteration. The procedure works like a sub-modularity based optimization problem [41]. Since usually the cardinality of the selected subset is rather smaller than the total number of sensors, in order to save the number of iterations, we will adopt the forward selection strategy for comparison. As the utility-based method also requires the RTFs, we will design two variants. In case the covariance whitening based RTF estimate is used, the utility-based method will be referred to as SS-utility-EST; in case the assumed RTF is exploited, it is then referred to as SS-utility-ASS.

Further, a random selection method will be compared to the proposed method, which randomly selects K sensors to perform beamforming. Using the estimated or assumed RTF, we thus obtain two variants of the random procedure, which are referred to as random-EST and random-ASS, respectively. Note that the performance of the random selection methods is averaged over 100 trails. In addition, as usually a microphone signal is dominated by the target/interfering source in case the source is close to the microphone, it is somehow reasonable that the microphones close to the source(s) are more informative for noise reduction. Therefore, we will also compare a maxEnergy method, which selects K sensors that have the largest input power. Similarly, using the estimated or assumed RTF, we can obtain maxEnergy-EST or maxEnergy-ASS.

A. Simulations for a Single Target Source

In this part, we consider the proposed sensor selection for the single target source case. The target point speech source (red dot) is located at (3.6, 8.4) m. We also place a coherent interfering source (blue star) at (8.4, 3.6) m. The variance of the signals is controlled by the SIR parameter. We use the RTF model in (19) for free fields as the assumed RTF, which is obtained by exploiting the source position in combination with the microphone positions. In reverberant environments, this modeling indeed reveals the direct-path component of the source signal. Further, we set $\mu_1 = 0.95$ and $\mu_2 = 0.05$.

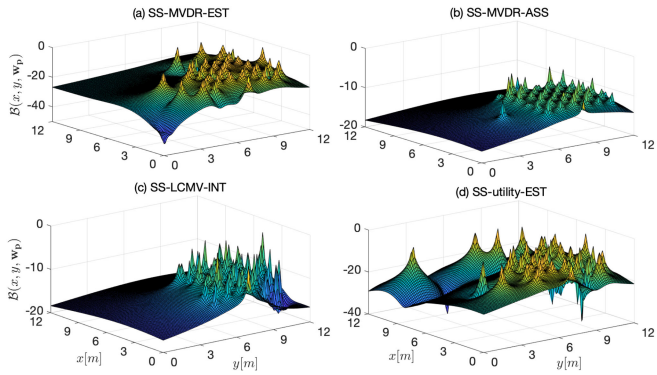

Fig. 3. Beam patterns (in dB) in terms of 2D spatial positions with $K = 40$.

Fig. 2 illustrates some typical sensor selection examples obtained by different approaches for $K = 40$. Clearly, for all comparison methods, most of the selected sensors are close to the target source, as these microphones can record high-quality signals, leading to a great contribution to speech enhancement. Comparing SS-MVDR-EST or SS-utility-EST to SS-MVDR-ASS, due to the RTF estimation errors, the obtained selected subset of sensors might not be optimal, as some sensors that are far away from the sources are also selected. Since the assumed RTF that is exploited by SS-MVDR-ASS is based on the exact use of source position and sensor positions, it can improve the selection performance, as the selected sensors are more assembled around the source position. Comparing SS-LCMV-INT to SS-MVDR-EST, it is clear that the selected subset is refined by integrating the superiority of SS-MVDR-ASS. From the perspective of set theory, the selected subset of SS-LCMV-INT can thus be viewed as the intersection set between the two subsets obtained by SS-MVDR-EST and SS-MVDR-ASS. The corresponding beam patterns for the angular frequency $\omega = 0.02\pi$ rad are shown in Fig. 3, from which we can observe that the mainlobes of SS-LCMV-INT are more concentrated to the true source position.

Fig. 4 shows the output noise power (in dB) of the comparison methods in terms of the cardinality of the selected subset for $T_{60} = 200$ ms. Comparing SS-MVDR-ASS (or SS-utility-ASS) to SS-MVDR-EST (or SS-utility-EST), it is clear that the use of the assumed RTF can improve the noise reduction performance in moderately reverberant situations, since the assumed RTF represents the accurate direct-path propagation of the target source. More importantly, the proposed SS-LCMV-INT method achieves the minimum output noise power, i.e., the best speech estimation quality. This reveals that integrating the estimated RTF and the assumed RTF is beneficial for sensor selection based linearly-constrained beamformers. Due to the fact that the proposed methods are based on a global optimization strategy, the proposed SS-MVDR-EST (or SS-MVDR-ASS) method performs better than SS-utility-EST (or SS-utility-ASS) which

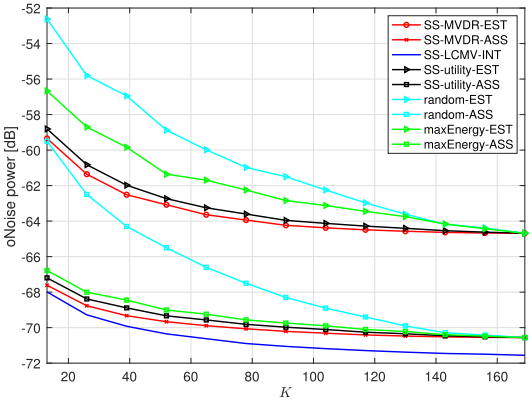

Fig. 4. The output noise power (in dB) in terms of the number of the selected sensor for $T_{60} = 200$ ms.

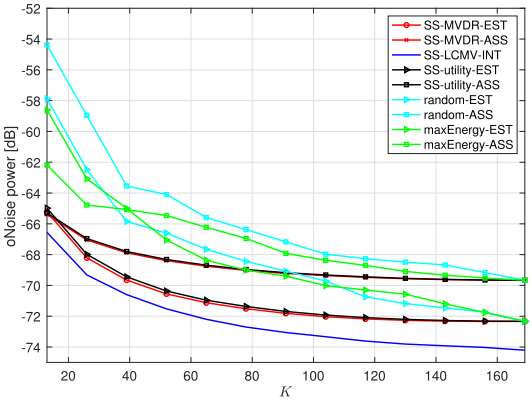

Fig. 5. The output noise power (in dB) in terms of K for $T_{60} = 800$ ms.

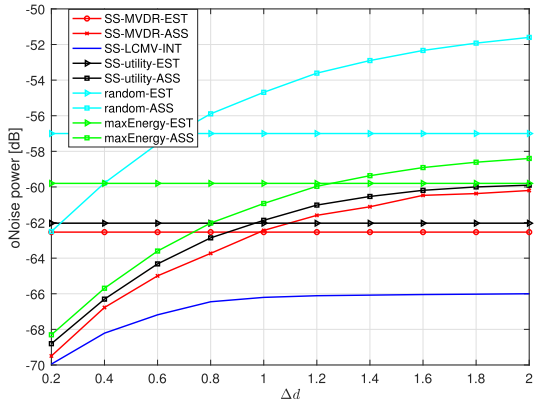

Fig. 6. The output noise power (in dB) in terms of the average distance error between the true source position and the assumed position.

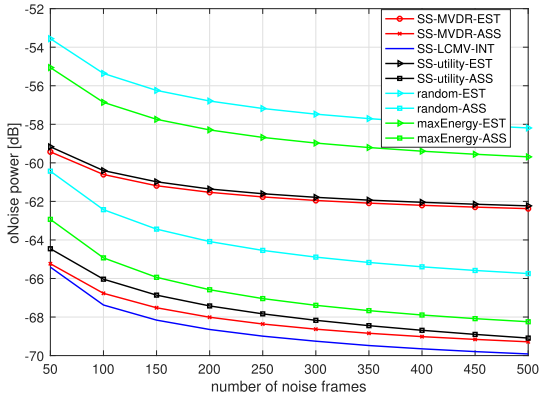

Fig. 7. The output noise power in terms of the number of noise frames.

is on a basis of greedy local selection strategy. Besides, the random selection method performs much worse than the proposed methods or the utility-based method. Also, the maxEnergy-based methods fail to improve the noise reduction performance, particularly when K is small, as they perform sensor selection only depending on the input microphone signal power, leading to many sensors that are close to the interfering source to be selected, especially in case SIR is low (e.g., SIR $= 0$ dB). Moreover, we show the output noise power in terms of K for $T_{60} = 800$ ms in Fig. 5. Due to the strong reverberation, the SS-MVDR-EST (or SS-utility-EST) method works better than SS-MVDR-ASS (or SS-utility-ASS), as in this case the estimated RTF is more accurate than the assumed one. Again, the proposed SS-LCMV-INT method achieves the best performance.

Due to the fact that in practice it is difficult to obtain the accurate RTF model in (19) depending on the source position together with the sensor locations, resulting in an inevitable error between the assumed RTF and the true RTF, we further investigate the output noise power in terms of the RTF calibration error Δd (in m) in Fig. 6. We assume that the target source is randomly located within the circle centered by the true source location with a radius of Δd. Since SS-MVDR-EST, SS-utility-EST, random-EST and maxEnergy-EST do not rely on the assumed RTF, their performance keeps constant in terms of Δd in Fig. 6. It is obvious that with an increase in the calibration error Δd, the performance of all assumed RTF based methods degrades. However, the proposed SS-LCMV-INT still achieves the best performance owing to the estimated RTF, because for a large Δd the estimated RTF is more accurate compared to the assumed one. Hence, in case the estimated RTF is more reliable, the performance of SS-LCMV-INT converges to that of SS-MVDR-EST; in case the assumed RTF is more accurate, it converges to that of SS-MVDR-ASS, which is consistent to the conclusion from (56). The proposed SS-LCMV-INT method can automatically check the reliability of the estimated and assumed RTFs and perform sensor selection based on the more reliable one. This validates that integrating the two RTFs for designing an LCMV beamformer is superior over using the respective RTF for an MVDR beamformer in the context of noise reduction.

Further, as the proposed methods together with the RTF depend on the noise correlation matrix, which has to be estimated using a limited amount of noise frames. The number of frames for estimating the noise correlation matrix will affect the RTF estimation accuracy and thus the performance of the proposed methods. We therefore show the output noise power of the comparison methods in terms of the number of noise frames in Fig. 7 with $K = 40$ and $\Delta d = 0$. Increasing the number of noise frames leads to an improvement in the noise correlation matrix estimation accuracy. It is clear that with an increase in the number of noise frames, the noise reduction performance of all

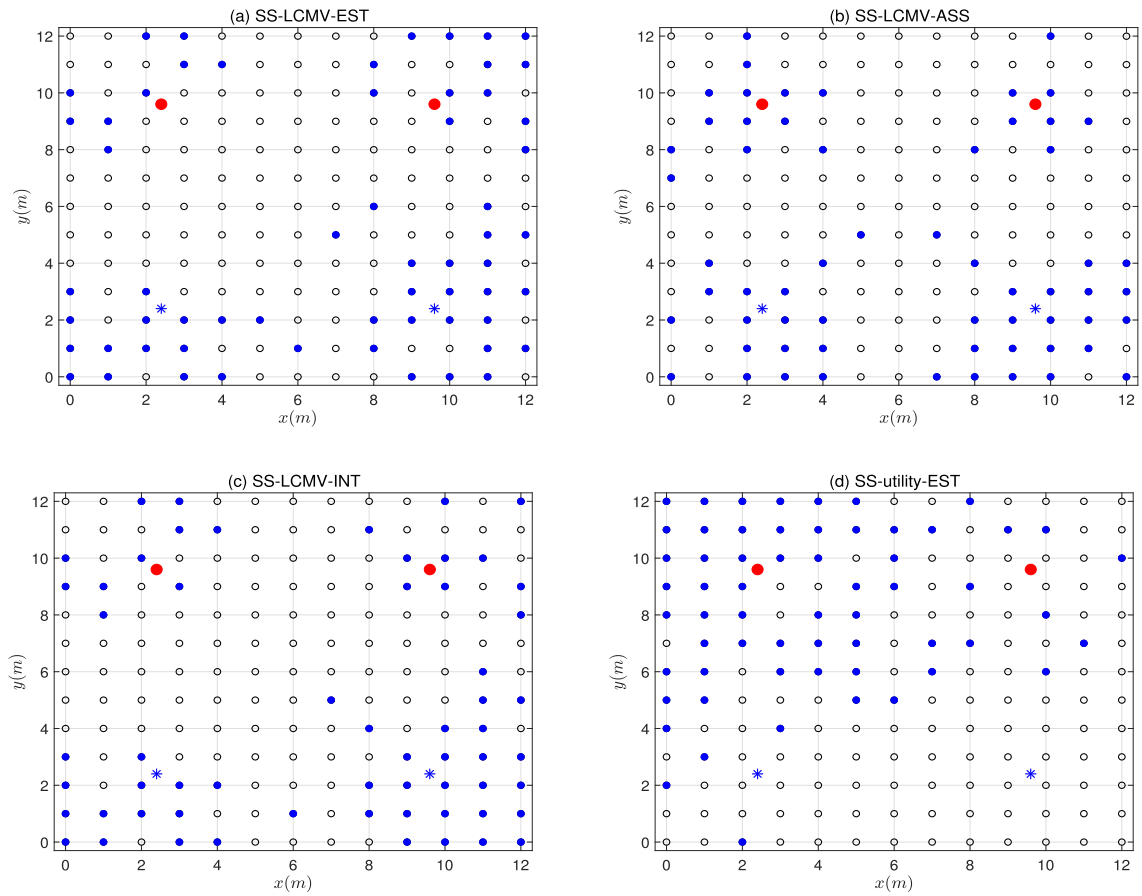

Fig. 8. Sensor selection examples where the blue sensors are activated by different approaches for $K = 60$: (a) SS-MVDR-EST, (b) SS-MVDR-ASS, (c) SS-LCMV-INT, (d) SS-utility-EST, which uses the estimated RTF for sensor selection. The considered scenario includes two target sources (red solid dots) and two coherent interfering sources (blue stars).

approaches improves, and the proposed SS-LCMV-INT method achieves the best performance.

B. Application to the Multiple Source Case

In this section, we apply the proposed SS-LCMV-N method to the scenario with multiple target sources. The experimental setup is shown in Fig. 8. Two target point sources are placed at (2.4, 9.6) m and (9.6, 9.6) m, respectively. Two coherent interfering sources are located at (2.4, 2.4) m and (9.6, 2.4) m, respectively. The other required parameters are set similarly as before. Note that the RTFs of the two target sources are estimated using the covariance whitening method one by one. Given a perfect VAD and the noise statistics, we detect the segments where the first target source is active and the other target is inactive, and during this period the corresponding RTF is estimated. This procedure applies to the RTF estimation of the other target source. Again, the assumed RTFs of the target sources are approximated by the direct-path RTF model in (19) based on the source positions in combination with the microphone locations. For the comparison approaches, in case the estimated RTFs are applied to (56), the proposed SS-LCMV-N is referred to as SS-LCMV-EST. If the assumed RTF is used in (56), it is then called SS-LCMV-ASS. Similarly to the integration of the estimated RTF and the assumed RTF in the single source case, we can also integrate the estimated RTFs and assumed RTFs of the two target sources as a spanned LCMV beamformer, referred to as SS-LCMV-INT (which includes four linear equality constraints in this case).

Fig. 8 illustrates typical sensor selection examples obtained by SS-LCMV-EST, SS-LCMV-ASS, SS-LCMV-INT and SS-utility-EST, respectively. Obviously, the proposed methods achieve a superiority over the utility-based method, as the regions of the sources are detected. The proposed methods can effectively select some sensors that are close to the target sources (having a high SNR) and some that are also close to the interfering sources (having a low SNR), which are beneficial for enhancing the targets and suppressing the noise sources, respectively. The selection results of SS-LCMV-EST and SS-LCMV-ASS only differ in a very limited number of sensors, and the intersection of them results in the selection of the proposed SS-LCMV-INT approach.

VI. Conclusion

In this paper, we investigated the selection of a subset of sensors from a large amount of candidate sensors for the linearly constrained beamformers based speech enhancement issue. The proposed sensor selection problem was formulated by minimizing the total output noise power and constraining the cardinality of the selected subset, as the number of selected sensors directly affects the system complexity. In the context of both MVDR and LCMV, the considered sensor selection problem can be solved

using convex optimization techniques. The proposed method is applicable to both the single target source case and the multiple source case. It was shown that the sensors that close to the target source(s) and some close to the interfering source(s) are more likely to be selected.

Given the estimated/assumed RTF of a single source, we can use the proposed SS-MVDR-EST/SS-MVDR-ASS method to find the subset of informative sensors. By integrating the estimated RTF and the assumed RTF from the MVDR beamformers to design an LCMV beamformer, the proposed SS-LCMV-INT method is obtained. It was shown that the integration of RTFs can improves the noise reduction performance, as SS-LCMV-INT can perform sensor selection based on the reliability of the respective RTFs. As the LCMV beamformer based on the two constraints associated with the estimated and assumed RTFs can be regarded as a linear combination of the two MVDR beamformers, the selected subset by SS-LCMV-INT indeed is the intersection between the two subsets obtained by SS-MVDR-EST and SS-MVDR-ASS. In case the estimated RTFs is more reliable, the selected subset of sensors by SS-LCMV-INT is more dominated by the sensors selected by SS-MVDR-EST; otherwise the sensors selected by SS-MVDR-ASS will dominate. Therefore, the proposed SS-LCMV-INT method is more robust against the RTF estimation/approximation errors. Since the proposed method performs sensor selection per frequency bin, in order to further reduce the time complexity, we will consider sensor selection across frequencies in the future.

As the proposed method depends on the noise correlation matrix and the estimated RTF and in practice the estimation of these parameters also consumes a certain amount of transmission energy, the proposed method is thus model-based. Given a WASN and required parameters, the proposed method can thus be applied to include a subset of microphones for speech enhancement. In case the parameters are unknown *a priori*, we can add an initialization step for parameter estimation and then use the proposed method for speech enhancement. Compared to the case of full WASN inclusion for both parameter estimation and noise reduction, the proposed method can still save power consumption. It would be more practical to design a data-driven method, which can adaptively increase the selected sensor subset from an initial point in the WASN. This can be implemented by combining the proposed method with the greedy sensor selection method in [20] and data-based parameter estimation approaches in [27]. As the focus of this work is mainly on the impact of RTF on the sensor selection based beamforming, we will leave this combination as a part of future research. In dynamic scenarios, online parameter estimation and sensor scheduling should be taken into account.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the associate editor for their helpful remarks and constructive suggestions, as they are very insightful for improving the quality of this paper. Some Matlab examples related to this paper.[1]

[1][Online]. Available: https://cas.tudelft.nl/Repository/

REFERENCES

[1] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output. i. fixed-processing systems," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 529–542, Nov. 1997.

[2] Q. Zou, X. Zou, M. Zhang, and Z. Lin, "A robust speech detection algorithm in a microphone array teleconferencing system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 3025–3028.

[3] G. Huang, J. Benesty, I. Cohen, and J. Chen, "A simple theory and new method of differential beamforming with uniform linear microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1079–1093, Mar. 2020.

[4] D. C. Moore and I. A. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, pp. V–497.

[5] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4771–4783, Sep. 2015.

[6] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Proc. IEEE Symp. Commun. Veh. Technol.*, 2011, pp. 1–6.

[7] Y. Zeng and R. C. Hendriks, "Distributed delay and sum beamformer for speech enhancement via randomized gossip," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 260–273, Jan. 2014.

[8] J. Zhang, A. I. Koutrouvelis, R. Heusdens, and R. C. Hendriks, "Distributed rate-constrained LCMV beamforming," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 675–679, May 2019.

[9] G. Huang, J. Benesty, I. Cohen, and J. Chen, "Differential beamforming on graphs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 901–913, Feb. 2020.

[10] J. Zhang, R. Heusdens, and R. C. Hendriks, "Rate-distributed binaural LCMV beamforming for assistive hearing in wireless acoustic sensor networks," in *Proc. IEEE 10th Sensor Array Multichannel Signal Process. Workshop*, 2018, pp. 460–464.

[11] J. Amini, R. C. Hendriks, R. Heusdens, M. Guo, and J. Jensen, "Asymmetric coding for rate-constrained noise reduction in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 154–167, Jan. 2019.

[12] J. Amini, R. C. Hendriks, R. Heusdens, M. Guo, and J. Jensen, "Spatially correct rate-constrained noise reduction for binaural hearing aids in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2731–2742, Oct. 2020.

[13] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[14] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for nonlinear measurement models," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.

[15] S. Liu, S. P. Chepuri, M. Fardad, E. Masazade, G. Leus, and P. K. Varshney, "Sensor selection for estimation with correlated measurement noise," *IEEE Trans. Signal Process*., vol. 64, no. 13, pp. 3509–3522, Jul. 2016.

[16] D. Golovin, M. Faulkner, and A. Krause, "Online distributed sensor selection," in *Proc. ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, 2010, pp. 220–231.

[17] H. Zhang, J. Moura, and B. Krogh, "Dynamic field estimation using wireless sensor networks: Tradeoffs between estimation error and communication cost," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2383–2395, Jun. 2009.

[18] A. Bertrand and M. Moonen, "Efficient sensor subset selection and link failure response for linear MMSE signal estimation in wireless sensor networks," in *Proc. EURASIP Eur. Signal Process. Conf.*, 2010, pp. 1092–1096.

[19] J. Szurley, A. Bertrand, M. Moonen, P. Ruckebusch, and I. Moerman, "Energy aware greedy subset selection for speech enhancement in wireless acoustic sensor networks," in *Proc. EURASIP Eur. Signal Process. Conf.*, 2012, pp. 789–793.

[20] 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, Lang. Process.*, vol. 26, no. 3, pp. 550–563, Mar. 2018.

[21] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition, " in *Proc. Int. Workshop Hands-Free Speech Commun*., 2011, pp. 1–6.

[22] Y. He and K. P. Chong, "Sensor scheduling for target tracking in sensor networks," in *Proc. IEEE Conf. Decis. Control*, 2004, pp. 743–748.

[23] A. Bertrand, J. Szurley, P. Ruckebusch, I. Moerman, and M. Moonen, "Efficient calculation of sensor utility and sensor removal in wireless sensor networks for adaptive signal estimation and beamforming," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5857–5869, Nov. 2012.
[24] J. Zhang, H. Chen, L. R. Dai, and R. C. Hendriks, "A study on reference microphone selection for multi-microphone speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 671–683, Nov. 2020.
[25] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 544–548.
[26] 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 *Proc. EURASIP Eur. Signal Process. Conf.*, 2018, pp. 2513–2517.
[27] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1507–1519, Oct. 2019.
[28] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE Proc. IRE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[29] 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, Apr. 1988.
[30] R. Ali, T. V. Waterschoot, and M. Moonen, "Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2288–2300, Dec. 2019.
[31] K. B. Petersen *et al.*, "The matrix cookbook," *Tech. Univ. Denmark*, vol. 7, 2008.
[32] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, U.K.: Cambridge Univ. Press, 2004.
[33] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
[34] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008.
[35] J. F. Sturm, "Using SeDuMi: A Matlab toolbox for optimization over symmetric cones," *Optim. Methods Softw.*, vol. 11, no. 1–4, pp. 625–653, 1999.
[36] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multi-microphone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, Jul. 2019.
[37] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *Nat. Inst. Standards Technol.*, vol. 15, pp. 29–50, 1988.
[38] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.
[39] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., vol. 2, no. 2.4, pp. 1–21, 2006.
[40] R. C. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.
[41] A. Krause and D. Golovin, "Submodular function maximization," *Tractability: Practical Approaches Hard Problems*, vol. 3, no. 19, pp. 71–104, 2014.

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

Jun Du (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2009 to 2010, he was with iFlytek Research as a Team Leader, working on speech recognition. From 2010 to 2013, he joined Microsoft Research Asia as an Associate Researcher, working on handwriting recognition, OCR. Since 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has authored or coauthored more than 150 papers. His main research interests include speech signal processing and pattern recognition applications. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and a Member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of the 2018 IEEE Signal Processing Society Best Paper Award. His team won several champions of CHiME-4/CHiME-5/CHiME-6 Challenge, SELD Task of 2020 DCASE Challenge, and DIHARD-III Challenge.

Li-Rong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 1983, the M.S. degree from the Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, China, in 1997. In 1993, he joined USTC. He is currently a Professor with the School of Information Science and Technology, USTC. He has authored or coauthored more than 50 papers in the areas of his research interests, which include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition.