Microphone Subset Selection for MVDR Beamformer Based Noise Reduction

Jie Zhang, Sundeep Prabhakar Chepuri, Richard C. Hendriks, and Richard Heusdens

***Abstract*—In large-scale wireless acoustic sensor networks (WASNs), many of the sensors will only have a marginal contribution to a certain estimation task. Involving all sensors increases the energy budget unnecessarily and decreases the lifetime of the WASN. Using microphone subset selection, also termed as sensor selection, the most informative sensors can be chosen from a set of candidate sensors to achieve a prescribed inference performance. In this paper, we consider microphone subset selection for minimum variance distortionless response (MVDR) beamformer based noise reduction. The best subset of sensors is determined by minimizing the transmission cost while constraining the output noise power (or signal-to-noise ratio). Assuming the statistical information on correlation matrices of the sensor measurements is available, the sensor selection problem for this model-driven scheme is first solved by utilizing convex optimization techniques. In addition, to avoid estimating the statistics related to all the candidate sensors beforehand, we also propose a data-driven approach to select the best subset using a greedy strategy. The performance of the greedy algorithm converges to that of the model-driven method, while it displays advantages in dynamic scenarios as well as on computational complexity. Compared to a sparse MVDR or radius-based beamformer, experiments show that the proposed methods can guarantee the desired performance with significantly less transmission costs.**

***Index Terms*—Sensor selection, MVDR, noise reduction, sparsity, convex optimization, transmission power, greedy algorithm.**

I. Introduction

MICROPHONE arrays have become increasingly popular in many speech processing applications, e.g., hearing aids [1], teleconferencing systems [2], hands-free telephony [3], speech recognition [4], human-robot interaction [5], etc. Compared to their single-microphone counterparts, microphone arrays typically lead to an enhanced performance when detecting, localizing, or enhancing specific sound sources. This is due to the fact that with a microphone array the sound field is not only sampled in time, but also in space.

Manuscript received March 28, 2017; revised July 25, 2017; accepted December 14, 2017. Date of publication xxxxx xx, 2017; date of current version xxxxx xx, 2018. This work is supported by the China Scholarship Council and Circuits and Systems (CAS) Group, Delft University of Technology, Delft, The Netherlands. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sven Nordholm.

The authors are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: {j.zhang-7, s.p.chepuri, r.c.hendriks, r.heusdens}@tudelft.nl). Sundeep Prabhakar Chepuri is supported in part by the ASPIRE project (project 14926 within the STW OTP programme), which is financed by the Netherlands Organization for Scientific Research (NWO) and the KAUST-MIT-TUD consortium under grant OSR-2015-Sensors-2700. (*Corresponding author: Jie Zhang*)

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

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

Although traditional microphone arrays have been widely investigated, see [6] and reference therein, they do have some important limitations. Typically, conventional microphone arrays have one central processing unit, that is, a fusion center (FC), which physically connects to the microphones. Rearranging the microphones in such a conventional wired and centralized array is impractical. Moreover, usually the target source is located far away from the array, resulting in a low signal-to-noise ratio (SNR). In addition, typically, the size of conventional arrays is limited as the maximum array size is determined by the application device [7].

Recently, wireless acoustic sensor networks (WASNs) have attracted an increased amount of interest [7]–[10]. In a WASN, each sensor node is equipped with a single microphone or a small microphone array, and the nodes are spatially distributed across a specific environment. The microphone nodes communicate with their neighboring nodes or the FC using wireless links. The use of WASNs can potentially resolve the limitations encountered with the conventional arrays that were mentioned before. At first, the WASN is not constrained to any specific (fixed) array configuration. Secondly, with a WASN, the position and number of microphones is not anymore determined by the application device. Instead, microphones can be placed at positions that are difficult to reach with conventional microphones. With a WASN, the array-size limitations disappear and the network becomes scalable (i.e., larger array apertures can be achieved) [11]. The fact that microphones in the WASN sample the sound field in a much larger area can yield higher quality recordings as it is likely that some of the sensors are close to the target source and have a higher SNR. One of the bottlenecks in a WASN is the resource usage in terms of power. Transmission of data between nodes or from the nodes to the FC will influence the battery lifetime of the sensor. Although all microphones in the WASN will positively contribute to the estimation task, only a few will have a significant contribution. It is questionable whether using all microphones in the network is beneficial taking the energy usage and lifetime of the sensors into account. Instead of blindly using all sensors, selecting a subset of microphones that is most informative for an estimation task at hand can reduce the data to be processed as well as transmission costs.

In this work, we investigate spatial filtering based noise reduction using only the most informative data via *microphone subset selection*, or so-called *sensor selection*, to reach a prescribed performance with low power consumption. Sensor selection is important for data dimensionality reduction. Mathematically, sensor selection is often expressed in terms of the

following optimization problem:

$$\arg \min_{\mathbf{p}\in\{0,1\}^M} \quad f(\mathbf{p}) \quad \text{s.t.} \quad \mathbf{1}_M^T\mathbf{p} = K, \tag{1}$$

where $\mathbf{p}$ indicates whether a sensor is selected or not, and the cost function $f(\mathbf{p})$ is optimized to select the best subset of K sensors out of M available sensors. Basically, the problem in (1) is a non-convex Boolean optimization problem, which incurs a combinatorial search over all the $\binom{M}{K}$ possible combinations. Usually, it can be simplified via convex relaxation techniques [12]–[14] or using greedy heuristics, e.g., leveraging submodularity [15], [16]. When the cardinality of $\mathbf{p}$ is of more concern, the cost function and constraint in (1) can also be interchanged by minimizing the cardinality of $\mathbf{p}$, i.e., $||\mathbf{p}||_0$, while constraining the performance meansure $f(\mathbf{p})$.

In general, sensor selection can be categorized into two classes: model-driven schemes and data-driven schemes. For the model-driven schemes, sensor selection is an offline design, where the sensing operation is designed based only on the data model (even before gathering data) such that a desired ensemble inference performance is achieved. In other words, the model-driven schemes provide the selected sensors *a priori* for the inference tasks [14]. There are many applications of the model-driven schemes for sensor placement in source localization [13], power grid monitoring [17], field estimation [18], target tracking [14], to list a few. In contrast to the offline design schemes, dimensionality reduction can also be done on already acquired data by discarding, i.e., censoring, less informative samples; this is referred as data-driven schemes. Data-driven sensor selection has been applied within the context of speech processing, e.g., speech enhancement [19], [20], speech recognition [21], and target tracking by sensor scheduling [22]. In the WASNs context, due to time-varying topologies, we have typically no information about the data model (e.g., probability density function), but the online measured data (e.g., microphone recordings) are available instead. In this work, we start with the model-driven sensor selection for the spatial filtering based noise reduction problem, which is then extended to a data-driven scheme.

A. Contributions

In this paper, we consider the problem of selecting the most informative sensors for noise reduction based on the minimum variance distortionless response (MVDR) beamformer. We formulate this problem to minimize the total transmission power subject to a constraint on the performance. While the classical sensor selection problem formulation as also given in (1) puts a constraint on the number of selected sensors, in the speech enhancement context the desired number of sensors is typically unknown. Hence, the desired number of sensors heavily depends on the scenario, e.g., the number of sound sources. Within the speech enhancement context it would be more useful to relate the constraint to a certain performance in terms of the expected quality or intelligibility of the final estimated signal. We therefore reformulate the sensor selection problem to be constrained to a certain expected output performance. In such a way, the selected sensors are always the ones having the (near-)minimum transmission power.

The minimization problem is first solved by convex optimization techniques exploiting the available complete joint statistics (i.e., correlation matrices) of the microphone measurements of the complete network, such that the selected subset of microphones is optimal. This is referred as the proposed model-driven approach.

In a more practical scenario, usually it is impossible to estimate the joint statistics of the complete network beforehand due to the dynamics of the scenario. Instead, the real-time measured data is only what can be accessed. Therefore, we extend the proposed model-driven algorithm to a data-driven scheme using a greedy sensor selection strategy. The performance of the greedy approach is proven to converge to that of the model-based method from an experimental perspective. There are a few existing contributions considering microphone subset selection in the area of audio signal processing. For example, Szurley et al. greedily selected an informative subset according to the SNR gain at each individual microphone for speech enhancement [20]. Bertrand and Moonen [19] conducted greedy sensor selection based on the contribution of each sensor signal to mean squared error (MSE) cost for signal estimation. Kumatani et al. proposed a channel selection for distant speech recognition by considering the contribution of each channel to multichannel cross-correlation coefficients (MCCCs) [21]. The proposed greedy algorithm shows an advantage in computational complexity and optimality as compared to existing greedy approaches [19], [20].

B. Outline and notation

The rest of this paper is organized as follows. Sec. II introduces the signal model, the classical MVDR beamforming, and sensor selection model. Sec. III presents the problem formulation. Sec. IV presents two solvers based on convex optimization to solve the model-driven sensor selection problem. Sec. V proposes a greedy algorithm. Sec. VI illustrates the simulation results. Finally, Sec. VII 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. $\mathrm{diag}(\cdot)$ refers to a block diagonal matrix with the elements in its argument on the main diagonal. $\mathbf{1}_N$ and $\mathbf{0}_N$ 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. Preliminaries

A. Signal model

We assume a spatially distributed candidate set of M microphone sensors that collect and transmit their observations to an FC. The multi-microphone noise reduction methods considered in this paper operate in the frequency domain on a frame-by-frame basis. Let l denote the frame index and ω the frequency bin index, respectively. We assume that the user (i.e., FC) has one source of interest, while multiple interfering sources are present in the environment. Using a discrete Fourier transform (DFT) domain description, the noisy

DFT coefficient at the k-th microphone, say $y_k(\omega, l)$, for $k = 1, 2, \cdots, M$, is given by

$$y_k(\omega, l) = x_k(\omega, l) + n_k(\omega, l), \tag{2}$$

where $x_k(\omega, l) = a_k(\omega)s(\omega, l)$ with $a_k(\omega)$ denoting the acoustic transfer function (ATF) of the target signal with respect to the k-th microphone and $s(\omega, l)$ the target source signal at the source location of interest. In (2), the component $n_k(\omega, l)$ represents the total received noise at the k-th microphone (including interfering sources and internal thermal additive noise). For notational convenience, the frequency variable ω and the frame index l will be omitted now onwards bearing in mind that the processing takes place in the frequency domain. Using vector notation, signals from M microphones are stacked in a vector $\mathbf{y} = [y_1, ..., y_M]^T \in \mathbb{C}^M$. Similarly, we define an M dimensional speech vector $\mathbf{x}$ for the speech component contained in $\mathbf{y}$ as $\mathbf{x} = \mathbf{a}s \in \mathbb{C}^M$ with $\mathbf{a} = [a_1, \cdots, a_M]^T \in \mathbb{C}^M$ denoting the steering vector which is constructed from the ATFs, and a length-M noise vector $\mathbf{n}$. As a consequence, the signal model in (2) can be compactly written as

$$\mathbf{y} = \mathbf{x} + \mathbf{n}. \tag{3}$$

Assuming that the speech and noise components are mutually uncorrelated, the correlation matrix of the received signals is given by

$$\mathbf{R_{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R_{xx}} + \mathbf{R_{nn}} \in \mathbb{C}^{M\times M}, \tag{4}$$

where $\mathbb{E}\{\cdot\}$ denotes the mathematical expectation, and $\mathbf{R_{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} = P_s\mathbf{a}\mathbf{a}^H$ with $P_s = \mathbb{E}\{|s|^2\}$ representing the power spectral density (PSD) of the target source. Notice that due to the assumption that $\mathbf{x}$ and $\mathbf{n}$ are uncorrelated, $\mathbf{R_{xx}}$ can be estimated by subtracting the noise correlation matrix $\mathbf{R_{nn}}$, which is estimated during the absence of speech from the speech-plus-noise correlation matrix $\mathbf{R_{yy}}$ [23]. In this work, we assume that a perfect voice activity detector (VAD) is available, such that the noise-only segments and the speech-plus-noise segments are classified accurately.

B. MVDR beamformer

The well-known MVDR beamformer minimizes the total output power after beamforming while simultaneously keeping the gain of the array towards the desired signal fixed. Therefore, any reduction in the output energy is obtained by suppressing interference or noise. Mathematically, this can be written as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ \mathbf{w}^H\mathbf{R_{nn}}\mathbf{w}, \quad \text{s.t.} \quad \mathbf{w}^H\mathbf{a} = 1. \tag{5}$$

The optimal solution, in a best linear unbiased estimator sense, can be obtained using the method of Lagrange multipliers, and is given by [8], [24], [25]

$$\hat{\mathbf{w}} = \frac{\mathbf{R_{nn}^{-1}}\mathbf{a}}{\mathbf{a}^H\mathbf{R_{nn}^{-1}}\mathbf{a}}. \tag{6}$$

After processing by the MVDR beamformer, the output SNR evaluated at a given time-frequency bin is given by the ratio of the variance of the filtered signal to the variance of the filtered noise

$$\begin{aligned}\text{SNR}_{\text{out}} =& \frac{\mathbb{E}\left\{|\hat{\mathbf{w}}^H\mathbf{x}|^2\right\}}{\mathbb{E}\left\{|\hat{\mathbf{w}}^H\mathbf{n}|^2\right\}} = \frac{\hat{\mathbf{w}}^H\mathbf{R_{xx}}\hat{\mathbf{w}}}{\hat{\mathbf{w}}^H\mathbf{R_{nn}}\hat{\mathbf{w}}} \\ =& P_s\mathbf{a}^H\mathbf{R_{nn}^{-1}}\mathbf{a}.\end{aligned} \tag{7}$$

C. Sensor selection model

The task of sensor selection is to determine the best subset of sensors to activate in order to minimize an objective function, subject to some constraints, e.g., the number of activated sensors or output noise power. We introduce a selection vector

$$\mathbf{p} = [p_1, p_2, ..., p_M]^T, \tag{8}$$

where $p_i \in \{0, 1\}$ with $p_i = 1$ indicating that the i-th sensor is selected. Let $K = \|\mathbf{p}\|_0$ represent the number of selected sensors with the ℓ_0-(quasi) norm referring to the number of non-zero entries in $\mathbf{p}$. Using a sensor selection matrix $\mathbf{\Phi_p}$, the selected microphone measurements can be compactly expressed as

$$\mathbf{y_p} = \mathbf{\Phi_p}\mathbf{y} = \mathbf{\Phi_p}\mathbf{x} + \mathbf{\Phi_p}\mathbf{n}, \tag{9}$$

where $\mathbf{y_p} \in \mathbb{C}^K$ is the vector containing the measurements from the selected sensors. Let $\text{diag}(\mathbf{p})$ be a diagonal matrix whose diagonal entries are given by $\mathbf{p}$, such that $\mathbf{\Phi_p} \in \{0, 1\}^{K\times M}$ is a submatrix of $\text{diag}(\mathbf{p})$ after all-zero rows (corresponding to the unselected sensors) have been removed. As a result, we can easily get the following relationships

$$\mathbf{\Phi_p}\mathbf{\Phi_p}^T = \mathbf{I}_K, \quad \mathbf{\Phi_p}^T\mathbf{\Phi_p} = \text{diag}(\mathbf{p}). \tag{10}$$

Therefore, applying the selection model to the classical MVDR beamformer in Sec. II-B, the best linear unbiased estimator for a subset of K microphones determined by $\mathbf{p}$ will be

$$\hat{\mathbf{w}}_\mathbf{p} = \frac{\mathbf{R_{nn,p}^{-1}}\mathbf{a_p}}{\mathbf{a_p}^H\mathbf{R_{nn,p}^{-1}}\mathbf{a_p}}, \tag{11}$$

where $\mathbf{a_p} = \mathbf{\Phi_p}\mathbf{a}$ is the steering vector corresponding to the selected microphones, and $\mathbf{R_{nn,p}} = \mathbf{\Phi_p}\mathbf{R_{nn}}\mathbf{\Phi_p}^T$ represents the noise correlation matrix of the selected sensors after the rows and columns of $\mathbf{R_{nn}}$ corresponding to the unselected sensors have been removed, i.e., $\mathbf{R_{nn,p}}$ is a submatrix of $\mathbf{R_{nn}}$.

III. Problem formulation

This work focuses on selecting the most informative subset of microphones for spatial filtering based noise reduction. The problem is formulated from the viewpoint of minimizing transmission cost subject to a constraint on the output performance. In particular, we express the filtering performance in terms of the output noise power, which is under the MVDR beamformer equivalent to the output SNR. However, notice that this can easily be replaced by other performance measures expressing the desired quality or intelligibility.

Let $\mathbf{c} = [c_1, c_2, ..., c_M]^T \in \mathbb{R}^M$ denote the pairwise transmission cost between each microphone and the FC. In general, the power consumption for wireless transmission can be modeled as [26]

$$c_i = c(d_i) + c_i^{(0)}, \ \forall i, \tag{12}$$

where $c(d_i)$ represents the power consumption depending on the distance d_i from the node with the i-th microphone to the FC, and $c_i^{(0)}$ is a constant depending on the power consumption of the i-th microphone itself. Based on the energy model in (12), our initial problem can be formulated as

$$\begin{aligned} \min_{\mathbf{w_p},\mathbf{p}\in\{0,1\}^M} \quad & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \quad & \mathbf{w_p}^H\mathbf{R_{nn,p}}\mathbf{w_p} \le \frac{\beta}{\alpha}, \\ & \mathbf{w_p}^H\mathbf{a_p} = 1, \end{aligned} \tag{P1}$$

where $||\cdot||_1$ denotes the ℓ_1-norm, β denotes the minimum output noise power after beamforming, and $\alpha \in (0,1]$ is an adaptive factor to control the output noise power compared to β. Note that β does not depend on the measurements of the whole network, because β/α is just a number that can be assigned by users, e.g., 40 dB, to indicate a desired performance. In (P1), the ℓ_1-norm is used to represent the total transmission costs of the network, i.e., between all the selected sensors and the FC, and it equals the inner-product $\mathbf{c}^T\mathbf{p}$ since both $\mathbf{p}$ and $\mathbf{c}$ are non-negative. Also, notice that (P1) is a general case for spatial filtering based noise reduction problems, e.g., using MVDR beamformers or linear constrained minimum variance (LCMV) beamformers [27]. In the next section, we will show how the optimization problem in (P1) can be solved using some of the properties of the MVDR beamformer.

IV. Model-driven sensor selection

In this section, we propose two slightly different ways to solve the optimization problem in (P1), firstly based on the correlation matrix $\mathbf{R_{xx}}$ and secondly based on knowledge of the steering vector $\mathbf{a}$, respectively. Both these solvers rely on the knowledge of the correlation matrices of the complete network, so that they belong to the model-driven schemes.

Considering the MVDR beamformer in (11), the output noise power using the selected sensors is given by

$$\hat{\mathbf{w}}_\mathbf{p}^H\mathbf{R_{nn,p}}\hat{\mathbf{w}}_\mathbf{p} = \left(\mathbf{a}_\mathbf{p}^H\mathbf{R}_\mathbf{nn,p}^{-1}\mathbf{a_p}\right)^{-1}, \tag{13}$$

where the constraint $\mathbf{w}_\mathbf{p}^H\mathbf{a_p} = 1$ in (P1) is implicit. Based on the fact that the MVDR beamformer keeps the speech components undistorted and suppresses the noise components, the variance of the filtered speech components can be shown to equal

$$\hat{\mathbf{w}}_\mathbf{p}^H\mathbf{R_{xx,p}}\hat{\mathbf{w}}_\mathbf{p} = P_s, \tag{14}$$

where $\mathbf{R_{xx,p}}$ denotes the submatrix of $\mathbf{R_{xx}}$ corresponding to the selected sensors. Hence, following (7) the output SNR using the selected sensors is given by

$$\begin{aligned} \mathrm{SNR}_{\mathrm{out},\mathbf{p}} &= \frac{\hat{\mathbf{w}}_\mathbf{p}^H\mathbf{R_{xx,p}}\hat{\mathbf{w}}_\mathbf{p}}{\hat{\mathbf{w}}_\mathbf{p}^H\mathbf{R_{nn,p}}\hat{\mathbf{w}}_\mathbf{p}} \\ &= P_s\mathbf{a}_\mathbf{p}^H\mathbf{R}_\mathbf{nn,p}^{-1}\mathbf{a_p} \\ &= P_s\mathbf{a}^H\mathbf{\Phi}_\mathbf{p}^T\mathbf{R}_\mathbf{nn,p}^{-1}\mathbf{\Phi_p a}. \end{aligned} \tag{15}$$

As a result, the original optimization problem in (P1) can equivalently be rewritten as

$$\begin{aligned} \min_{\mathbf{p}\in\{0,1\}^M} \quad & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \quad & P_s\mathbf{a}_\mathbf{p}^H\mathbf{R}_\mathbf{nn,p}^{-1}\mathbf{a_p} \ge \alpha\cdot\mathrm{SNR}, \end{aligned} \tag{P2}$$

where $\mathrm{SNR} = \frac{P_s}{\beta}$ represents the maximum output SNR. Both (P1) and (P2) are non-convex because of the Boolean variable $\mathbf{p}$, but also due to the non-linearity of the constraint in $\mathbf{p}$. In what follows, we will present solvers by linearizing (P2) and reformulating it using convex relaxation. Note that (P1) and (P2) are built from different perspectives (i.e., constraining the output noise power and SNR, respectively), but in the context of the MVDR beamforming, they are equivalent.

A. Convex relaxation using $\mathbf{R_{xx}}$

From the output SNR in (15), the selection variable $\mathbf{p}$ appears at three places, that are: $\mathbf{\Phi}_\mathbf{p}^T$, $\mathbf{R}_\mathbf{nn,p}^{-1}$ and $\mathbf{\Phi_p}$. We combine these together as one new matrix $\mathbf{Q} = \mathbf{\Phi}_\mathbf{p}^T\mathbf{R}_\mathbf{nn,p}^{-1}\mathbf{\Phi_p}$. To simplify calculations, in what follows, we will rearrange $\mathbf{Q}$ such that $\mathbf{p}$ occurs only at one place. Let us first consider a decomposition of the noise covariance matrix [14], [28]

$$\mathbf{R_{nn}} = \lambda\mathbf{I}_M + \mathbf{G}, \tag{16}$$

where λ is a positive scalar and $\mathbf{G}$ is a positive definite matrix (if λ is smaller than the smallest eigenvalue of $\mathbf{R_{nn}}$, this decomposition can be easily found). The reason for choosing such a λ is to make $\mathbf{G}^{-1}+\lambda^{-1}\mathrm{diag}(\mathbf{p})$ positive definite, which will be seen after (24). Using (16), we have

$$\mathbf{R_{nn,p}} = \mathbf{\Phi_p}\left(\lambda\mathbf{I}_M + \mathbf{G}\right)\mathbf{\Phi}_\mathbf{p}^T = \lambda\mathbf{I}_K + \mathbf{\Phi_p G\Phi}_\mathbf{p}^T, \tag{17}$$

and $\mathbf{Q}$ can be reformulated as

$$\mathbf{Q} = \mathbf{\Phi}_\mathbf{p}^T\left(\lambda\mathbf{I}_K + \mathbf{\Phi_p G\Phi}_\mathbf{p}^T\right)^{-1}\mathbf{\Phi_p}. \tag{18}$$

Using the matrix inversion lemma [29, p.18]

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

we can simplify $\mathbf{Q}$ in (18) as

$$\mathbf{Q} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\left(\mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p})\right)^{-1}\mathbf{G}^{-1}. \tag{19}$$

Note that (19) is still non-linear in $\mathbf{p}$ due to the inversion operation, but $\mathbf{p}$ appears now only at one place. Based on $\mathbf{Q}$, the output SNR with sensor selection as in (15) can be calculated as [29, p.6]

$$\begin{aligned} \mathrm{SNR}_{\mathrm{out},\mathbf{p}} &\overset{(1)}{=} \mathrm{trace}\left(P_s\mathbf{a}^H\mathbf{\Phi}_\mathbf{p}^T\mathbf{R}_\mathbf{nn,p}^{-1}\mathbf{\Phi_p a}\right) \\ &\overset{(2)}{=} \mathrm{trace}\left(\mathbf{QR_{xx}}\right) \\ &\overset{(3)}{=} \mathrm{trace}\left(\mathbf{R}_\mathbf{xx}^{\frac{H}{2}}\mathbf{Q}\mathbf{R}_\mathbf{xx}^{\frac{1}{2}}\right), \end{aligned} \tag{20}$$

where the trace$(\cdot)$ operator computes the trace of a matrix, and $\mathbf{R}_\mathbf{xx}^{\frac{1}{2}}$ represents the principal square root of $\mathbf{R_{xx}}$. The second and third equality in (20) is based on trace property, which is employed to make the linear matrix inequality (LMI) in (25) symmetric. Here, we utilize the trace operation to express the output SNR as a function of $\mathbf{R_{xx}}$. The latter can be estimated using the recorded audio in practice, e.g., during the training phase, or using the correlation matrices $\mathbf{R_{yy}}$ and $\mathbf{R_{nn}}$ without the need to explicitly know the steering vector $\mathbf{a}$ or $\mathbf{a_p}$.

Secondly, in what follows we will linearize the SNR constraint in (P2). To do this, we introduce a new matrix $\mathbf{Z}$ to equivalently rewrite the constraint in (P2) as

$$\mathrm{trace}\left(\mathbf{Z} - \frac{\alpha P_s}{M\beta}\mathbf{I}_M\right) \ge 0, \tag{21}$$

$$\mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{Q}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} = \mathbf{Z}, \tag{22}$$

where the equality constraint in (22) is non-linear in $\mathbf{p}$. For linearization, we relax it to an inequality constraint

$$\mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{Q}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} \succeq \mathbf{Z}. \tag{23}$$

Note that (21) and (23) are sufficient conditions for obtaining the original constraint in (P2), this is why we utilize $\succeq$ for convex relaxation. Substituting (19) in (23), we get

$$\mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{G}^{-1}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{Z} \succeq \mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{G}^{-1}\left[\mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}\left(\mathbf{p}\right)\right]^{-1}\mathbf{G}^{-1}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}}. \tag{24}$$

Due to the positivity of λ, the positive definiteness of $\mathbf{G}$ and the Boolean vector $\mathbf{p}$, the matrix $\mathbf{G}^{-1}+\lambda^{-1}\mathrm{diag}(\mathbf{p})$ is positive definite, and this is why we chose in (16) a positive scalar λ and a positive definite matrix $\mathbf{G}$ to decompose the matrix $\mathbf{R}_{\mathbf{nn}}$. Using the Schur complement [30, p.650], we obtain a symmetric LMI of size $2M$ from (24) as

$$\begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} \\ \mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{G}^{-1} & \mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{G}^{-1}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{0}_{2M}, \tag{25}$$

which is linear in $\mathbf{p}$. Furthermore, the Boolean variable $\mathbf{p}$ can be relaxed using continuous variables $\mathbf{p} \in [0,1]^M$ or semidefinite relaxation [31]. In this work, we utilize the former way. Accordingly, (P2) can be expressed in the following form:

$$\begin{aligned} \min_{\mathbf{p},\mathbf{Z}} \quad & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \quad & \mathrm{trace}\left(\mathbf{Z} - \frac{\alpha P_s}{M\beta}\mathbf{I}_M\right) \geq 0, \\ & \begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} \\ \mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{G}^{-1} & \mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}}\mathbf{G}^{-1}\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{0}_{2M}, \\ & 0 \leq p_i \leq 1, \quad i = 1, 2, \cdots, M. \end{aligned} \tag{26}$$

The relaxed optimization problem in (26) is a semidefinite programming problem [30, p.128] and can be solved efficiently in polynomial time using interior-point methods or solvers, like CVX [32] or SeDuMi [33]. The computational complexity for solving (26) is of the order of $\mathcal{O}(M^3)$. The approximate Boolean selection variables p_i can be obtained by randomized rounding using the solution of (26) [13]. Notice that the solver in (26) depends on $\mathbf{R}_{\mathbf{xx}}$. In a practical scenario, this is unknown, but can be estimated based on estimates of the correlation matrices $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{nn}}$ as shown in (4). $\mathbf{R}_{\mathbf{yy}}$ can be estimated from the data itself, and $\mathbf{R}_{\mathbf{nn}}$ can be estimated using a VAD or noise correlation matrix estimator for the noise-only frames, see e.g., [34].

B. Solver based on the steering vector $\mathbf{a}$

Suppose the ATFs from the source to the microphones are known, the steering vectors $\mathbf{a}$ (in free field) can be constructed. With $\mathbf{a}$, the output SNR in (20) can be expressed as

$$\mathrm{SNR}_{\mathrm{out},\mathbf{p}} = P_s\mathbf{a}^H\mathbf{Q}\mathbf{a}. \tag{27}$$

Therefore, using the expression for $\mathbf{Q}$ in (19), the original constraint in (P2) can be rewritten as

$$\mathbf{a}^H\mathbf{G}^{-1}\mathbf{a} - \mathbf{a}^H\mathbf{G}^{-1}\left(\mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p})\right)^{-1}\mathbf{G}^{-1}\mathbf{a} \geq \frac{\alpha}{\beta},$$

or, reorganized as

$$\mathbf{a}^H\mathbf{G}^{-1}\mathbf{a} - \frac{\alpha}{\beta} \geq \mathbf{a}^H\mathbf{G}^{-1}\left(\mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p})\right)^{-1}\mathbf{G}^{-1}\mathbf{a}. \tag{28}$$

Using the Schur complement, (28) can be reformulated as a symmetric LMI of size $M+1$

$$\begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1}\mathbf{a} \\ \mathbf{a}^H\mathbf{G}^{-1} & \mathbf{a}^H\mathbf{G}^{-1}\mathbf{a} - \frac{\alpha}{\beta} \end{bmatrix} \succeq \mathbf{0}_{M+1}. \tag{29}$$

Accordingly, the optimization problem in (P2) is expressed as

$$\begin{aligned} \min_{\mathbf{p}} \quad & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1}\mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1}\mathbf{a} \\ \mathbf{a}^H\mathbf{G}^{-1} & \mathbf{a}^H\mathbf{G}^{-1}\mathbf{a} - \frac{\alpha}{\beta} \end{bmatrix} \succeq \mathbf{0}_{M+1} \\ & 0 \leq p_i \leq 1, \quad i = 1, 2, \cdots, M, \end{aligned} \tag{30}$$

where the Boolean variables $\mathbf{p}$ have already been relaxed using the continuous surrogates $\mathbf{p} \in [0,1]^M$, and (30) has a standard semidefinite programming form, which can also be solved by the aforementioned tools. Notice that this solver depends on knowledge on $\mathbf{a}$. To estimate (the direct path of) $\mathbf{a}$ one can use a source localization algorithm, e.g., [35]–[37], in combination with the sensor locations, or use the generalized eigenvalue decomposition of the matrices $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$ [38], [39].

Remark 1. *The differences between (26) and (30) are threefold: 1) (30) preserves the constraint on the output SNR (or noise power), yet (26) relaxes it in a convex way by introducing an auxiliary variable* $\mathbf{Z}$*; 2) Observing the LMIs in (26) and (30), they differ in dimensions (i.e.,* $2M$ *and* $M+1$*, respectively), so (30) is computationally much more efficient; 3) The solver in (26) requires to estimate the speech correlation matrix* $\mathbf{R}_{\mathbf{xx}}$ *and the PSD* P_s *of the target source, while (30) requires the steering vector* $\mathbf{a}$*.*

Remark 2. *For a special case, when the noise is spatially uncorrelated with covariance matrix*

$$\mathbf{R}_{\mathbf{nn}} = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_M^2\right),$$

the optimization problem (P2) can be simplified to the following Boolean linear programming problem

$$\begin{aligned} \min_{\mathbf{p}} \quad & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \quad & \mathbf{a}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathrm{diag}(\mathbf{p})\mathbf{a} \geq \frac{\alpha}{\beta}. \end{aligned} \tag{31}$$

Although the above optimization problem is nonconvex in $\mathbf{p} \in \{0,1\}^M$*, it admits a simple non-iterative solution based on rank ordering. More specifically, the optimal solution to* (31) *is given by setting the entries of* $\mathbf{p}$ *corresponding to the indices*

$$\min\left\{i \in \{1, 2, \cdots, M\} | \frac{c_{[1]}}{v_{[1]}} + \cdots + \frac{c_{[i]}}{v_{[i]}} \geq \frac{\alpha}{\beta}\right\}$$

to 1, and the remaining entries of $\mathbf{p}$ *to 0, where* $v_{[1]}, \cdots, v_{[M]}$ *and* $c_{[1]}, \cdots, c_{[M]}$ *are numbers of* $v_1, v_2, \cdots, v_M$ *and* $c_1, c_2, \cdots, c_M$*, respectively, sorted in ascending order with* $v_i = c_i\sigma_i^2/|a_i|^2$ *and* a_i *being the* i*-th entry of* $\mathbf{a}$*.*

V. GREEDY SENSOR SELECTION

In Sec. IV, the sensor selection problem was solved using statistical information from the complete network, i.e., $\mathbf{R_{xx}}$ and $\mathbf{R_{nn}}$. In practice, this information is unknown and needs to be estimated from all the sensors' measurements. Hence, we call this a model-driven approach as the complete $\mathbf{R_{xx}}$ and $\mathbf{R_{nn}}$ are required as well as the transmission power from the microphones to the FC. In a practical scenario, it is undesired to estimate the statistics of the complete network up front, as this would imply a lot of data transmission for sensor nodes that might never be selected in the end as most sensors are non-informative. Moreover, in practice, the position of the FC or microphones might be changing as well. For this reason we need a selection mechanism that does not rely on knowledge of the statistics and microphone-FC distances of the complete network. Instead, we could access the measurements of neighboring sensors (close to the FC or already selected sensors). In this section, we present a greedy approach for the sensor selection based noise reduction problem, which does not require to estimate the global statistics. Therefore, the greedy algorithm can be performed online, and it belongs to the data-driven category. In Sec. VI, we will experimentally show that the data-driven and model-driven approach will converge to a similar performance.

Let r_i denote the spatial position of the i-th microphone, $\mathcal{S}_1$ a candidate set of microphones and $\mathcal{S}_2$ the selected set, respectively. The proposed greedy algorithm is summarized in Algorithm 1. Given an arbitrary initial spatial point z_0 and a transmission range R_0[1], we can initialize the candidate set $\mathcal{S}_1$ of sensors, i.e., the R_0-closest sensors to z_0. For the candidate set $\mathcal{S}_1$, we estimate the noise correlation matrix $\mathbf{R}_{\mathbf{nn},\mathcal{S}_1}$ and decompose it following (16), and then solve the optimization problem in (26) or (30). For instance, for $\mathcal{S}_1$ the optimization problem in (30) can be reformulated as

$$\begin{aligned} \min_{\mathbf{p}\in[0,1]^{K_1}} \quad & \|\mathrm{diag}(\mathbf{p}_{\mathcal{S}_1})\mathbf{c}_{\mathcal{S}_1}\|_1 \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{G}_{\mathcal{S}_1}^{-1}+\lambda_{\mathcal{S}_1}^{-1}\mathrm{diag}(\mathbf{p}_{\mathcal{S}_1}) & \mathbf{G}_{\mathcal{S}_1}^{-1}\mathbf{a}_{\mathcal{S}_1} \\ \mathbf{a}_{\mathcal{S}_1}^H\mathbf{G}_{\mathcal{S}_1}^{-1} & \mathbf{a}_{\mathcal{S}_1}^H\mathbf{G}_{\mathcal{S}_1}^{-1}\mathbf{a}_{\mathcal{S}_1}-\frac{\alpha}{\beta_{\mathcal{S}_1}} \end{bmatrix} \succeq \mathbf{0}_{K_1+1} \\ & 0\le p_i\le 1,\ \forall i\in\mathcal{S}_1, \end{aligned} \tag{32}$$

where $\beta_{\mathcal{S}_1}$ represents the output noise power of the classical MVDR beamformer using the microphones in the candidate set $\mathcal{S}_1$, which is termed as the local constraint. Notice that the adaptive factor α is the same as that in the model-driven scheme. If $\alpha \le 1$, (32) will always have a feasible solution within $\mathcal{S}_1$, the feasible set will be taken and used to define a new set $\mathcal{S}_2$ with $|\mathcal{S}_2| \le |\mathcal{S}_1|$. Then, based on the set $\mathcal{S}_2$, a new set $\mathcal{S}_1$ is formed based on the R_0-closest sensors with respect to the sensors included in the set $\mathcal{S}_2$[2]. These operations are continued until $\mathcal{S}_1$ or $\mathcal{S}_2$ does not change (i.e., until convergence has been achieved). The finally selected set $\mathcal{S}_2$ will always be smaller than the selected set for the model-driven approach from Sec. IV when using the same α. This is due to the fact that the output noise power $\beta_{\mathcal{S}_1}$ in the constraint of the greedy approach is based on the set $\mathcal{S}_1$ that is always smaller or equal to the initial set as used by the model-driven approach in (30) (where β is obtained by involving all sensors). As a result, β/α will always be smaller than $\beta_{\mathcal{S}_1}/\alpha$. In summary, $\beta/\alpha < \beta_{\mathcal{S}_1}/\alpha$. The performance of the greedy approach (after convergence) will therefore always be somewhat worse than the model-based approach, as the constraint is less tight. This can either be solved by choosing a different (larger) α for the greedy approach, or, by switching from the constraint $\beta_{\mathcal{S}_1}/\alpha$ to the constraint β/α after convergence. As an alternative, we could have used the constraint β/α within the greedy approach of (32) right from the beginning. However, in that case, in the first few iterations (32) would have no feasible solution as an insufficient amount of measurements are available to satisfy the constraint on the output noise power. As a consequence of an infeasible solution, the selected set $\mathcal{S}_2$ will keep all sensors from $\mathcal{S}_1$, of which many are actually uninformative.

In order to make the performance of the proposed greedy algorithm converge to that of the model-driven approach, we switch from $\beta_{\mathcal{S}_1}$ (local constraint) to β (global constraint) after the above iterative procedure converges (i.e., the constraint $\beta_{\mathcal{S}_1}/\alpha$ for solving (32) has been satisfied). Finally, the proposed greedy algorithm will converge to the model-driven method based on the global constraint. To conclude, the greedy algorithm includes two steps: using a locally defined constraint ($\beta_{\mathcal{S}_1}/\alpha$) and using a globally defined constraint (β/α), as summarized in Algorithm 1. Recall that the globally defined constraint, which involves β/α with β denoting the minimum output noise power after beamforming, does not need to be dependent on the measurements of the whole network. Hence, the greedy algorithm does not need to know the exact optimal performance, i.e., β. For the implementation in practice, we only need to set a number for β/α depending on the expected performance. Note that the computational complexity of each iteration is of the order of $\mathcal{O}(|\mathcal{S}_1|^3)$, and the number of iterations depends on z_0 and R_0. From the description of the algorithm, we know that both the greedy algorithm and the model-driven method have, in the end, the same constraint that must be satisfied, leading to very similar performance, which can also be found in simulations.

VI. SIMULATIONS

In this section, the proposed algorithms are experimentally evaluated. Sec. VI-A introduces three reference methods that we will use for comparison. In Sec. VI-B, the experimental setup is explained. In Sec. VI-C, the proposed model-driven sensor selection based MVDR beamformer (referred to as MD-MVDR in short) is compared with the reference methods introduced in Sec. VI-A. In Sec. VI-D, we will analyze the performance of the proposed greedy approach as a data-driven sensor selection, including the convergence behaviour, initialization and the adaptivity of a moving FC. Sec. VI-E compares the computational complexity between the model-driven method and the greedy approaches.

[1] R_0 can be defined as the wireless transmission range $\sqrt{\log(2M)/M}$ in a random geometric graph to guarantee that the network is connected with high probability [40].

[2] R_0-closest sensors with respect to the set $\mathcal{S}_2$ include all the sensors that are R_0-closest to any individual sensor in $\mathcal{S}_2$.

Algorithm 1: Greedy Sensor Selection

Step 1: initialization
  Initial point: z_0
  Transmission range: R_0
  Selected set: $\mathcal{S}_2 = \varnothing$
  Candidate set: $\mathcal{S}_1 = \{i | \|r_i - z_0\|_2 \leq R_0,\ \forall i\ \}$;
Step 2: considering local constraint
  Cardinality of the active set: $K_1 = |\mathcal{S}_1|$;
  Decomposing: $\mathbf{R}_{\mathbf{nn},\mathcal{S}_1} = \lambda_{\mathcal{S}_1}\mathbf{I}_{K_1} + \mathbf{G}_{\mathcal{S}_1}$;
  Solving (32) using the local constraint $\beta_{\mathcal{S}_1}$;
  Update:
    $\mathcal{S}_2 = \{i | p_i = 1,\ \forall i \in \mathcal{S}_1\}$;
    $\mathcal{S}_1 = \mathcal{S}_2 \cup \{i | \|r_i - r_{\mathcal{S}_2}\|_2 \leq R_0,\ \forall i\}$;
    Go to line 6 until converge;
Step 3: solving (32) using global constraint β;
  If infeasible, update
    $\mathcal{S}_2 = \mathcal{S}_1$;
    $\mathcal{S}_1 = \mathcal{S}_2 \cup \{i | \|r_i - r_{\mathcal{S}_2}\|_2 \leq R_0,\ \forall i\}$;
    Go to line 14;
  If feasible, update
    $\mathcal{S}_2 = \{i | p_i = 1,\ \forall i \in \mathcal{S}_1\}$;
    $\mathcal{S}_1 = \mathcal{S}_2 \cup \{i | \|r_i - \mathcal{S}_2\|_2 \leq R_0,\ \forall i\}$;
    Go to line 14 until converge;
Return $\mathcal{S}_2$.

A. Reference methods

Apart from the classical MVDR beamforming without sensor selection as introduced in Sec. II-B, the proposed approaches will also be compared with a weighted sparse MVDR beamformer [41]–[43], a radius-based MVDR beamformer and a utility-based greedy method [19], [20].

1) Weighted sparse MVDR beamformer: A naive alternative to sensor selection for spatial filtering is to enforce sparsity in the filter coefficients while designing the beamformer. Due to the physical nature of sound, this approach trades a small loss in SNR for a large reduction in communication power required to produce a beamformer output by reducing the active nodes. Some existing works on sparse MVDR beamformers are presented in [41]–[43]. One of our reference methods is therefore a sparse MVDR beamformer. However in order to make the comparison with the sparse MVDR beamformer fair, we use a weighting by the transmission power. Using the model of transmission costs from (12), the weighted sparse MVDR beamformer can be formulated as

$$\begin{aligned} \hat{\mathbf{w}} = &\arg\min_{\mathbf{w}}\ \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w} + \mu\|\mathbf{w}^H\mathrm{diag}(\mathbf{c})\|_0 \\ &\text{s.t.}\ \ \mathbf{w}^H\mathbf{a} = 1, \end{aligned} \tag{33}$$

where μ denotes the regularization parameter to control sparsity, and the ℓ_0-norm can be relaxed by the ℓ_1-norm or the concave surrogate based on sum-of-logarithms [13], [44]. When $\mu = 0$, it is identical to the classical MVDR beamformer in Sec. II-B. Note that a larger μ leads to a sparser $\mathbf{w}$. The product $\mathbf{w}^H\mathrm{diag}(\mathbf{c})$ indicates the pairwise transmission costs. Weighting the beamforming filter $\mathbf{w}$, the sensors with smaller transmission costs have a dominant contribution to $\mathbf{w}$ compared to sensors with larger transmission costs. From the standpoint of implementation, for each frequency bin, if $|w_i| \geq \varepsilon, \forall i$, the i-th sensor will be selected, otherwise not. Due to this "inevitable" thresholding, the resulting beamformer is not necessarily MVDR anymore. The threshold ε is chosen empirically.

Algorithm 2: Utility based greedy sensor addition

Initialization: same to **Algorithm 1**;
for $k = 1, 2, ..., M$
  Compute the gain of output noise power $\mathbf{\Delta}$ by adding each sensor in $\mathcal{S}_1 \backslash \mathcal{S}_2$ to $\mathcal{S}_2$;
  Compute utility vector: $\mathbf{g} = [\frac{\Delta_1}{c_1}, \frac{\Delta_2}{c_2}, ..., \frac{\Delta_{|\mathcal{S}_1 \backslash \mathcal{S}_2|}}{c_{|\mathcal{S}_1 \backslash \mathcal{S}_2|}}]^T$;
  $i = \arg\max_i \mathbf{g}$;
  Add sensor: $\mathcal{S}_2 = \mathcal{S}_2 \cup i$;
  Update: $\mathcal{S}_1 = \mathcal{S}_2 \cup \{i | \|r_i - r_{\mathcal{S}_2}\|_2 \leq R_0,\ \forall i\}$;
end for until $c_{\mathcal{S}_2} \geq c_T$
Return $\mathcal{S}_2$.

2) Radius-based MVDR beamformer: The goal of this article is to minimize the transmission costs while constraining the performance. A straightforward way to reduce transmission costs is by selecting the sensors close to the FC. The closer a sensor to the FC, the less transmission power is required. Hence, given a radius γ, we can involve the sensors within the circle centered by the FC for the MVDR beamfomer, which we call radius-based MVDR beamformer. An example is given in Fig. 2(a), where the blue sensors are chosen with $\gamma = 6$ m. Obviously, this approach does not take the source or interference information into account, and its performance suffers from γ and the network topology.

3) Utility based greedy sensor addition: In [20], the most informative subset of microphones is obtained by greedily removing the sensor that has the least contribution to a utility measurement (e.g., SNR gain, output noise power, MSE cost), also called backward selection. This method requires to know the statistics offline and can be considered a model-driven approach. While in [19], apart from sensor selection based on backward selection, an alternative was proposed by greedily adding the sensor that has the largest contribution to the utility (forward selection). This can be considered as an online data-driven procedure. In order to compare the proposed greedy algorithm with the state-of-the-art greedy methods, we summarize [19], [20] as the utility based greedy sensor addition shown in Algorithm 2. In this work, our focus is on the transmission costs. To measure the utility, we therefore take the ratio of the gain of the output noise power $\mathbf{\Delta}$ that is obtained by adding each sensor from $\mathcal{S}_1 \backslash \mathcal{S}_2$ to $\mathcal{S}_2$, to the transmission cost. Here, the sets $\mathcal{S}_1$ and $\mathcal{S}_2$, are respectively, defined the same as for Algorithm 1. The sensor which has the larger ratio between noise reduction and transmission cost would have the larger utility. When the transmission costs for the selected set $\mathcal{S}_2$ exceeds the maximum cost budget c_T, the algorithm is terminated. Note that this approach only adds one sensor to the selected set $\mathcal{S}_2$ per iteration, thus it may require many iterations to get an acceptable solution.

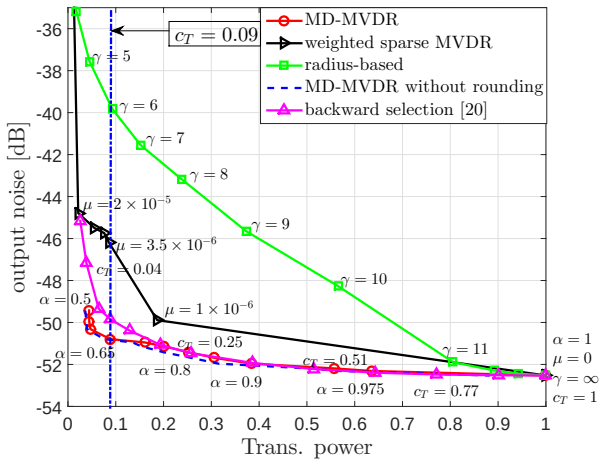


Figure 1. Output noise power in terms of transmission cost for different choices of α, μ, γ, c_T.

B. Experiment Setup

Fig. 2(a) shows the experimental setup employed in the simulations, where 169 candidate microphones are placed uniformly in a 2D room with dimensions (12×12) m. The desired speech source (red solid circle) is located at $(2.4, 9.6)$ m. The FC (black solid square) is placed at $(9, 3)$ m. Two interfering sources (blue stars) are positioned at $(2.4, 2.4)$ m and $(9.6, 9.6)$ m, respectively. The target source signal is a 10 minute long concatenation of speech signals originating from the TIMIT database [45]. The interferences are stationary Gaussian speech shaped noise sources. All signals are sampled at 16 kHz. We use a square-root Hann window of 20 ms for framing with 50% overlap. The ATFs are generated using [46] with reverberation time $T_{60} = 200$ ms. The threshold ε for the sparse MVDR beamformer is set to be 10^{-5} empirically, since the coefficients smaller than this threshold are negligible. We also model microphone self noise using zero-mean uncorrelated Gaussian noise with an SNR of 50 dB.

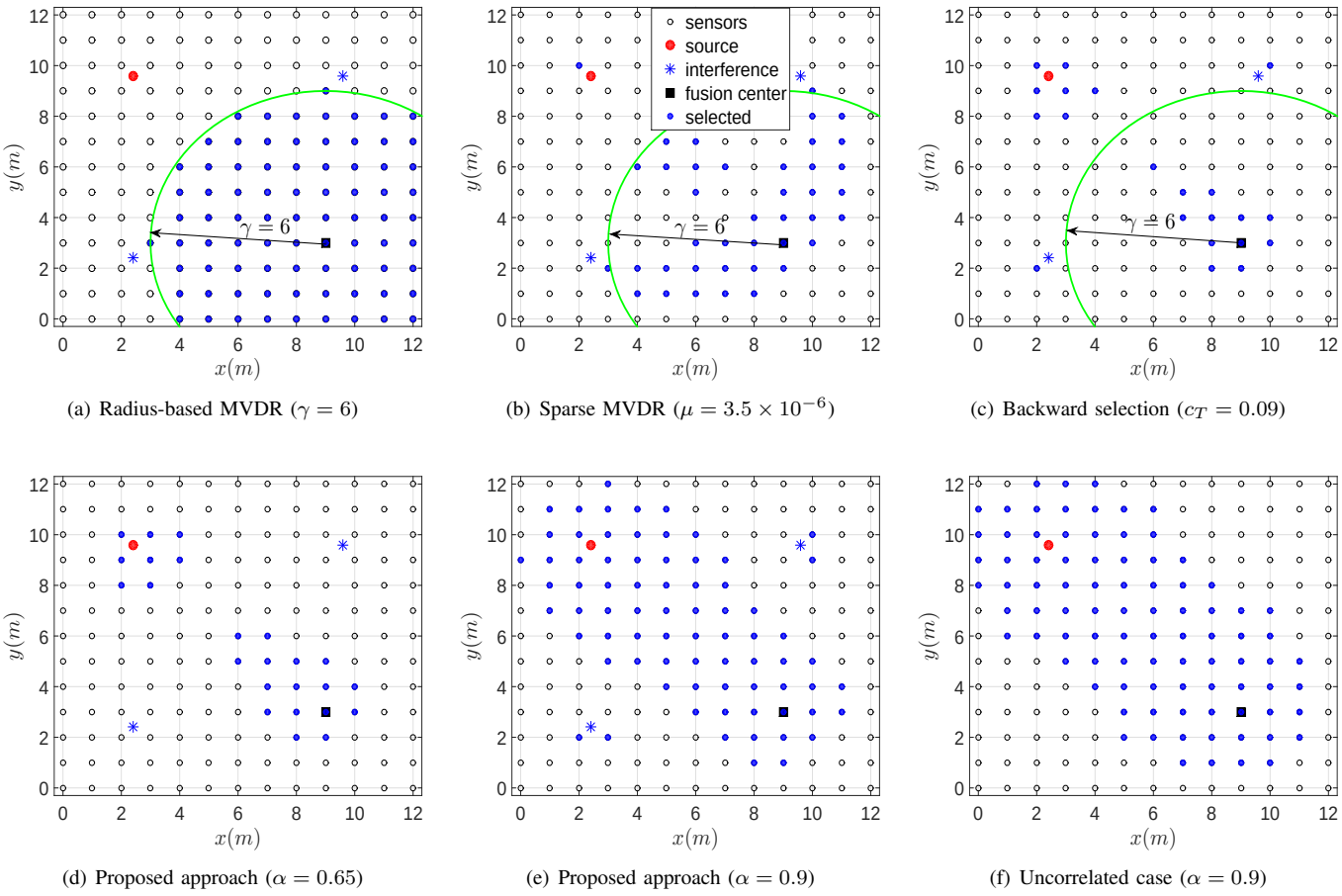
To focus on the concept of sensor selection, we assume that the ATFs (i.e., steering vector $\mathbf{a}$) are perfectly known. In practice, this can be estimated using source localization algorithms, e.g., [35], [36], in combination with the sensor locations, or, by calculating the generalized eigenvalue decomposition of the matrices $\mathbf{R_{nn}}$ and $\mathbf{R_{yy}}$ [38], [39]. For the correlation matrices, we use noise-only segments which are long enough to estimate $\mathbf{R_{nn}}$; during the speech-plus-noise segments $\mathbf{R_{yy}}$ is tracked and $\mathbf{R_{xx}}$ can be obtained by subtracting the estimate of $\mathbf{R_{nn}}$ from $\mathbf{R_{yy}}$ simultaneously. For the wireless transmission model in (12), we consider the simplest wireless transmission case, where the transmission cost between each sensor and the FC is proportional to the square of their Euclidean distance [47], and we assume that the device dependent cost $c_i^{(0)} = 0, \forall i$. In the following simulations, the transmission costs are normalized between 0 and 1 based on the total transmission costs between all the microphones and the FC.

C. Evaluation of the model-driven approach

In order to compare the state-of-the-art approaches mentioned in Sec. VI-A, we first investigate the influence of the required parameters α, μ, γ, c_T on the performance, for the proposed and the different reference methods. Fig. 1 shows the relationship between the output noise power (in dB) and the transmission power for SIR $= 0$ dB with SIR representing signal-to-interference ratio. Fig. 1 also shows the results without randomized rounding (blue dashed curve) regarded as the lower bound of the proposed method, i.e., involving the selection variable $\mathbf{p}$ (thus, no selection) for computations. As we can see that the performance of MD-MVDR is smaller than that of the MD-MVDR without rounding, the binary solution of the proposed method using randomized rounding is still satisfactory in terms of expected output noise power. We can conclude that in order to reach the same noise reduction performance, the proposed approach always requires significantly less transmission costs compared to the weighted sparse beamformer or radius-based beamformer. If the transmission power budget c_T (defined in Algorithm 2) is small, the proposed method performs better than the backward selection [20], and if c_T is large, they are comparable. Furthermore, when $\alpha = 0.65, \gamma = 6, \mu = 3.5 \times 10^{-6}$, the four approaches approximately have the same transmission power as $c_T = 0.09$. Hence, in the simulations that will follow we will compare the cases for $\alpha = 0.65, \gamma = 6, \mu = 3.5 \times 10^{-6}, c_T = 0.09$. Note that in Fig. 1, all the microphones are involved for the MVDR beamforming when $\alpha = 1, \gamma = \infty, \mu = 0, c_T = 1$. This is the optimal MVDR beamformer.

Fig. 2(a)-(d) illustrate typical sensor selection examples for one angular frequency ($\omega = \pi/256$ rad/s) of the radius-based MVDR beamformer ($\gamma = 6$), sparse MVDR beamformer ($\mu = 3.5 \times 10^{-6}$), backward selection ($c_T = 0.09$) and the proposed method ($\alpha = 0.65$), respectively. In addition, we show the radius for the radius-based MVDR, where all the sensors within this radius are selected, and thus not depicted explicitly in Fig. 2(b)-(d). For fixed sensor and source locations, it is observed that the selected sensors are the same for most frequency bins. The sensors within the green circles ($\gamma = 6$) are selected by the radius-based method, which chooses the γ-closest sensors relative to the FC for the MVDR beamformer. It can be seen that in order to save transmission power as well as to reduce noise, the proposed approach selects some microphones close to the source and some close to the FC for computation, while the sparse MVDR beamformer or radius-based method do not have this property. Although the backward selection has this property, it performs somewhat worse in noise reduction, which can be seen in Fig. 1. On one hand, the signals recorded by the microphones close to the source position are degraded less by the interfering source, and they preserve the target source better. Those microphones are helpful for enhancing the target source. On the other hand, the microphones close to the FC require less transmission power to transmit data to the FC. They are selected as they hardly add to the total transmission costs. When we increase the adaptive factor α, more sensors that are close to the interference positions are selected as well, because they carry information on the interfering sources as shown in Fig. 2(e).

Fig. 2(f) illustrates the case where interfering sources are absent, and the microphone recordings are degraded by the microphone self noise, taking the noise level SNR=50 dB. Compared to Fig. 2(e), most selected microphones are the



(a) Radius-based MVDR ($\gamma = 6$) (b) Sparse MVDR ($\mu = 3.5 \times 10^{-6}$) (c) Backward selection ($c_T = 0.09$)

(d) Proposed approach ($\alpha = 0.65$) (e) Proposed approach ($\alpha = 0.9$) (f) Uncorrelated case ($\alpha = 0.9$)

Figure 2. Microphone subset selection examples (The blue sensors are activated for the MVDR beamformer): (a) radius-based MVDR beamforming, (b) sparse MVDR beamforming, (c) backward selection [20] and (d) proposed method ($\alpha = 0.65$) for spatially correlated noises, respectively, (e) proposed method for correlated case with $\alpha = 0.9$, and (f) proposed method ($\alpha = 0.9$) for spatially uncorrelated noises only.

same, and they are more aggregate to the source position as well as to the FC. The difference is whether to select sensors that are close to the interferences. From this comparison, we can also conclude that the sensors that are close to the interference are useful for cancelling the correlated noise.

D. Evaluation of the data-driven approach

In this subsection, we will evaluate the proposed greedy approach compared to the model-driven algorithm and the utility-based method. The experimental setup is kept the same as that used for the evaluation of the model-driven approach. The advantages of the greedy algorithm will be demonstrated from three perspectives, i.e., convergence behaviour, initialization, and for a scenario with a moving FC. Note that for the greedy approach, its convergence behaviour depends on the initial point z_0 and the transmission range.

1) Convergence behaviour: In order to analyze the convergence behaviour of the proposed greedy approach, the sensor network topology in this work is viewed as a grid topology, such that its transmission range R_0 is fixed to the distance between two neighboring microphone nodes. In this part, we take the initial point z_0 at the position (9, 3) m as an example to show the convergence behaviour of the greedy algorithm. The effect of the choice of z_0 will be looked into later in this section.

Fig. 3 illustrates the proposed greedy algorithm (i.e., Algorithm 1) for $\alpha = 0.9$ using the same experimental setup of Fig. 2(e). In detail, at the 1st iteration (e.g., $k = 1$) the R_0-closest candidate set $\mathcal{S}_1$ has five sensors. Based on the local constraint three sensors (in blue) are selected to form the set $\mathcal{S}_2$. The candidate set $\mathcal{S}_1$ is then increased by adding the R_0-closest sensors with respect to $\mathcal{S}_2$. This procedure continues for the first 21 iterations. When $k = 21$, we can see that $\mathcal{S}_2$ is completely surrounded by $\mathcal{S}_1$, such that if we still use the local constraint, there would be no new sensors that can be added to $\mathcal{S}_1$, from which we conclude that the local constraint, i.e., $\beta_{\mathcal{S}_1}/\alpha$, has been satisfied. In order to satisfy the global constraint on the output noise power, the algorithm is then switched to the global constraint after the 21st iteration, i.e., β/α. Finally, three more iterations are further required the reach the expected performance.

We can see from Fig. 3, that the proposed greedy method does not blindly increase the candidate set $\mathcal{S}_1$ towards all possible directions. Instead, $\mathcal{S}_1$ is increased only in the informative direction to the source location, such that the less informative microphones are not included. Furthermore, notice that the final selected set $\mathcal{S}_2$ differs slightly from the model-

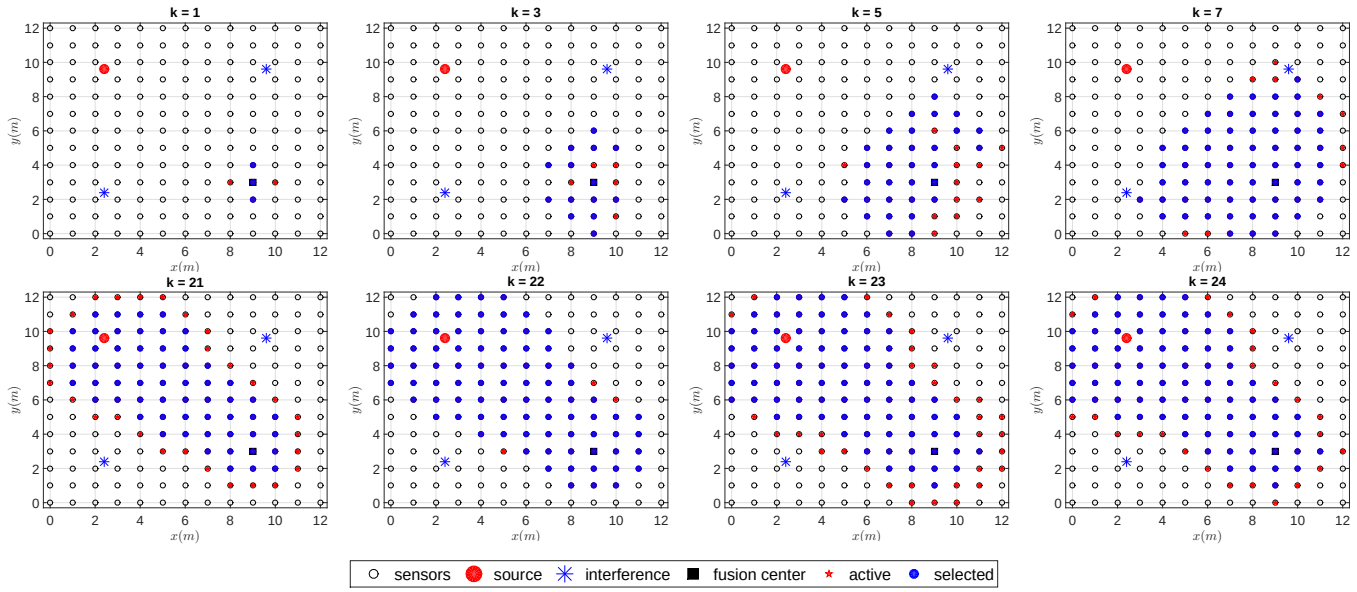


Figure 3. An illustration of the convergence behaviour for the proposed greedy algorithm (i.e., Algorithm 1). The initial point is located at (9, 3) m.

driven approach in Fig. 2(e), as the greedy approach does not select the sensors that are close to the interfering sources, but it selects more sensors close to the target source. Hence, convergence towards the model-based approach is obtained in the sense of performance, but not in terms of selected sensors as the solution is not necessarily unique. In general, given an expected noise reduction performance and transmission power budget, it could be that more than one microphone subset are satisfactory. So for the proposed greedy approach, we cannot guarantee that the final selected subset is unique or entirely the same as the model-driven approach, but we can make sure that they have a similar performance.

In Fig. 4, we show the ratio of cardinality of the candidate set $\mathcal{S}_1$ to the total number of sensors M and transmission power per iteration. The combination of the global and local constraint is compared to a greedy algorithm that uses only the global constraint for Algorithm 1. Using only the global constraint, $\mathcal{S}_1$ would blindly increase towards all directions. Clearly, we see that by using a combination between the local and the global constraint, much less sensors are included per iteration, such that the transmission power is kept low.

2) Initializations: In this part, we will show the effect of the initial point z_0 on the convergence rate. Fig. 5 illustrates the output noise power (in dB) in terms of iterations for four different initializations, i.e., centre (6, 6) m, source position (2.4, 9.6) m, interference position (2.4, 2.4) m and FC (9, 3) m. The red dashed line represents the performance of the model-driven algorithm proposed in Sec. IV, which selects the most informative sensors from all the possible candidates. The black dashed line denotes the performance of the classical MVDR beamformer using all microphones. The magenta curve shows the proposed greedy algorithm for the MVDR beamformer. The blue dashed curve denotes the performance of the utility-based algorithm [19], [20]. The output noise power of the greedy algorithm includes two steps: local constraint ($\beta_{\mathcal{S}_1}/\alpha$) and global constraint (β/α). The moment that the constraint is switched from the local to the global constraint is indicated

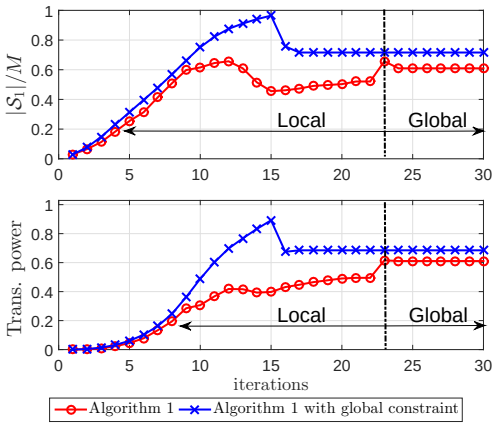
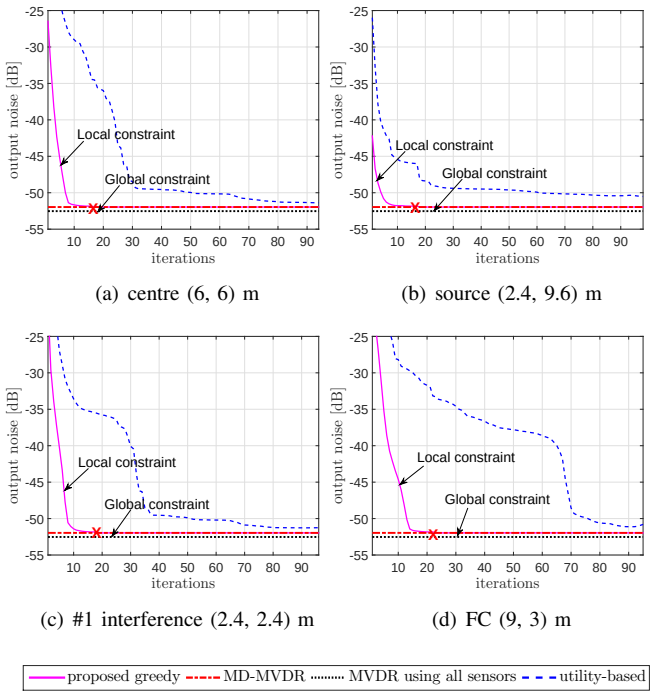


Figure 4. Cardinality of candidate set and transmission power vs iterations.

by the red marker "×". When executing the local constraint, the output noise power decreases fastest for the initialization at the source position and slowest for the FC initialization. This is due to the fact that the sensors that are close to the source are more informative for speech enhancement. After the algorithm converges based on the local constraint, by switching to the global constraint, the output noise decreases further until it reaches the performance of the model-driven approach. Hence, from a perspective of performance, the proposed greedy algorithm converges to the model-driven method. In addition, if the point of initialization is closer to the source position, the convergence is faster. To conclude, the initialization only influences the convergence rate, and it does not affect the final performance. More importantly, for all the cases of initialization, the proposed greedy approach converges to the model-driven method in the sense of performance.

Furthermore, from Fig. 5 we observe that the proposed greedy algorithm converges with much less iterations as compared to the utility-based method, because the latter only selects one sensor in each iteration. Note that in the comparisons the total transmission cost budgets for the two approaches are



(a) centre (6, 6) m (b) source (2.4, 9.6) m

(c) #1 interference (2.4, 2.4) m (d) FC (9, 3) m

Figure 5. Output noise power in terms of iterations for different initial point z_0: (a) centre, (b) source, (c) interference, (d) FC.

kept the same. Also, there is no guarantee for the utility-based method to fulfil the expected performance on the output noise power. Given the same transmission cost budget, the proposed greedy algorithm can therefore obtain more reduction in noise power and converge much faster in terms of iterations.

3) Moving FC: In this part, we will show the advantage of the greedy algorithm in a dynamic scenario with a moving FC. In practice, the FC could be moving, because usually it is regarded as a mobile user. Fig. 6 shows an example of greedy sensor selection for a moving FC, where the FC moves along the black dashed rectangle. The starting point is located at (4, 8) m, and at this position it takes 12 steps (9 steps for the local constraint and 3 steps for the global constraint) for the greedy algorithm to converge to a feasible informative set. The changing trend of the previous 11 steps is similar to Fig. 3, so we merely show the results of the steps 1 and 12 in the left top subplot in Fig. 6. The FC then slowly moves to the next position (4, 6.67) m. For the second position, we use the selected microphone set from the first position to update the candidate set, and then solve (32). It is found that only 5 iterations (1 for the local constraint and 4 for the global constraint) are required to obtain convergence. Subsequently, the FC continues moving. For the next positions, the greedy algorithm only requires about 6 iterations to converge. Hence, in the dynamic scenario with a moving FC, the proposed greedy approach can significantly save computational resources. Since the interferences are Gaussian shaped noise sources, once the noise correlation matrix $\mathbf{R_{nn}}$ is estimated using the noise-only segments before the FC starts to move, it can still be used for the subsequent positions of the FC. Hence, for the moving FC case, we only need to update $\mathbf{R_{yy}}$ or $\mathbf{R_{xx}}$ based on the real-time recordings. It is also noteworthy that the FC is not

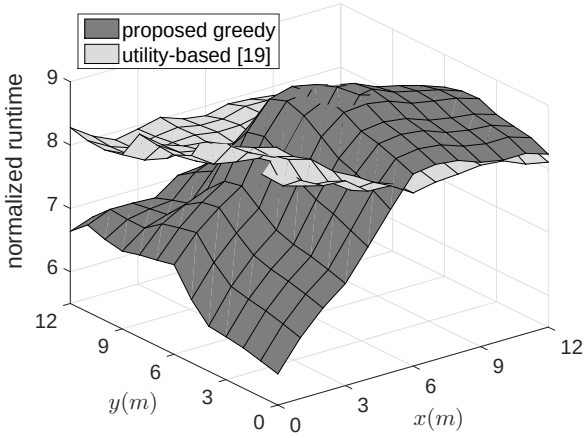


Figure 7. Execution time in terms of different initial points.

a microphone and the ATFs (i.e., the steering vector $\mathbf{a}$) stay the same even when the FC is moving, since the positions of microphones and the target source are fixed.

An interesting phenomenon occurs in Fig. 6. As the FC moves further away form the source, we can clearly see the importance of the sensors that are close to the interference. When the FC is located at (4, 6.67) m, two sensors close to the interference are also selected. This cannot be distinguished when FC = (4, 8) m, where the FC is closer to the source. Hence, we can conclude that the sensors that are close to the source, to the FC and to the interference are informative, and they are helpful to enhance the target source, to save transmission costs and to cancel the interfering sources, respectively.

E. Complexity analysis

In this subsection, we will compare the computational complexity of the greedy algorithms to that of the model-driven approach. For the model-driven approach, its complexity is of the order of $\mathcal{O}(M^3)$, so we use M^3 in the worst case for analysis without loss of generality. For the proposed greedy algorithm (i.e., Algorithm 1), suppose that J iterations are required to converge, in each iteration its complexity is of the order of $\mathcal{O}(|\mathcal{S}_1|^3)$, thus we can use $\sum_{j=1}^{J}|\mathcal{S}_1|^3$ to represent its computational complexity. For the utility-based greedy algorithm (i.e., Algorithm 2), we can find that its computational complexity is of the order of $\mathcal{O}(|\mathcal{S}_2|^2(|\mathcal{S}_1|-|\mathcal{S}_2|))$ for each iteration from [19], thus $\sum_{j=1}^{J}|\mathcal{S}_2|^2(|\mathcal{S}_1|-|\mathcal{S}_2|)$ can be exploited to represent its total complexity.

Fig. 7 compares the execution time of the two aforementioned greedy strategies. The execution time is normalized by the runtime of model-driven method, whose runtime is 1 as benchmark. From Fig. 7, we can see that the execution time of the proposed greedy algorithm depends on the initial point z_0, as it will be more expensive for the initial points that are further from the target source. Furthermore, for most initial points the proposed algorithm is computationally more efficient than the utility-based method, because we need much less iterations (20 iterations compared to 90 iterations approximately which has already been demonstrated in Fig. 5).

Although the computational complexity of the greedy algorithms could be larger than that of the model-driven algorithm,

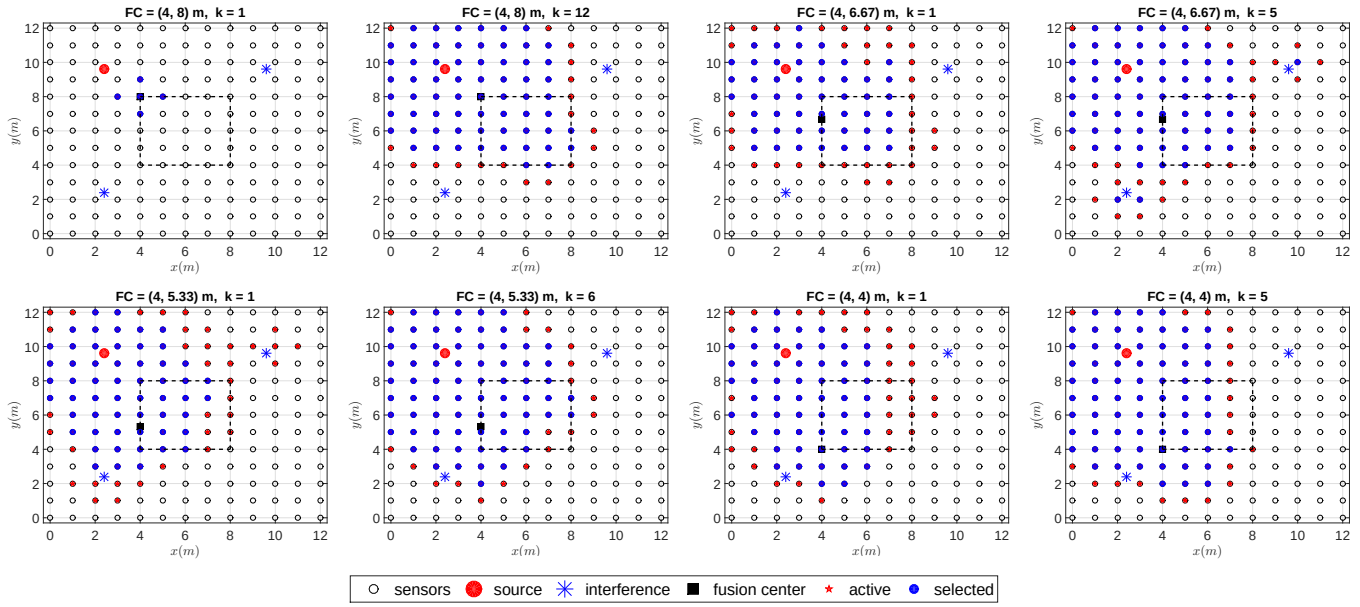


Figure 6. An illustration of the sensor selection based on the proposed greedy algorithm for the moving FC.

it belongs to the data-driven schemes. That is, we do not need to know the number of microphones in an environment, and it is unnecessary to inform all microphones to transmit their recorded data to the FC to estimate the statistics beforehand. Instead, it is only required to include the closest neighboring microphone nodes gradually, the FC then updates the statistics and decides the informative subset. Hence, compared to the model-driven method which is suitable for static environments, the greedy approach can be applied to dynamic scenarios, especially with infinite candidate microphones.

VII. Conclusions

In this work, we considered selecting the most informative microphone subset for the MVDR beamfomer based noise reduction. The proposed strategies were formulated through minimizing the transmission cost with the constraint on noise reduction performance. Firstly, if the statistics (e.g., the estimates of noise correlation matrices) are available, the microphone subset selection can be solved in a model-driven scheme by utilizing the convex optimization techniques. Additionally, in order to make the sensor selection capable of dynamic environments, a greedy approach in a data-driven scheme was proposed as an extension of the model-driven method. The performance of the proposed greedy algorithm converges to that of the model-driven approach. More importantly, it works more effectively in dynamic environments (e.g., with a moving FC). We concluded that in order to enhance the speech source as well as to save transmission costs, the sensors close to the source signal, those close to the FC and some close to the interferences are of larger probability to be selected, and they are helpful to enhance the target source, to save transmission costs and to cancel the interfering source, respectively. In a more general WASN, the network could consist of larger number of microphone nodes, which makes the model-driven approach impractical. The greedy algorithm is still a possible alternative to handle the microphone subset selection problem.

VIII. Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful remarks and constructive suggestions. The MATLAB code for this paper is available at the authors' website http://cas.tudelft.nl.

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, 1997.

[2] Q. Zou, X. Zou, M. Zhang, and Z. Lin, "A robust speech detection algorithm in a microphone array teleconferencing system," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, vol. 5, pp. 3025–3028.

[3] J. G. Ryan and R. A. Goubran, "Application of near-field optimum microphone arrays to hands-free mobile telephony," *IEEE Trans. on Vehicular Technology*, vol. 52, no. 2, pp. 390–400, 2003.

[4] 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, vol. 5, 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, 2015.

[6] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[7] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011, pp. 1–6.

[8] R. Heusdens, G. Zhang, R. C. Hendriks, Y. Zeng, and W. B. Kleijn, "Distributed MVDR beamforming for (wireless) microphone networks using message passing," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, 2012.

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

[10] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 3, pp. 651–661, 2017.

[11] Y. Jennifer, M. Biswanath, and G. Dipak, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292 – 2330, 2008.

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

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

[14] S. Liu, S. P. Chepuri, M. Fardad, E. Maşazade, 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, 2016.

[15] D. Golovin, M. Faulkner, and A. Krause, "Online distributed sensor selection," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2010, pp. 220–231.

[16] S. Rao, S. P. Chepuri, and G. Leus, "Greedy sensor selection for non-linear models," in *IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2015, pp. 241–244.

[17] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 74–85, 2004.

[18] 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, 2009.

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

[20] 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 *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2012, pp. 789–793.

[21] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *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 *IEEE Conf. on Decision and Control*, 2004, vol. 1, pp. 743–748.

[23] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 4266–4269.

[24] Otis Lamont Frost III, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[25] 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.

[26] Q. Wang, M. Hempstead, and W. Yang, "A realistic power consumption model for wireless sensor network devices," in *2006 3rd annual IEEE communications society on sensor and ad hoc communications and networks*, 2006, vol. 1, pp. 286–295.

[27] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[28] S. P. Chepuri and G. Leus, "Sparse sensing for distributed detection," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1446–1460, 2015.

[29] K. B. Petersen, M. S. Pedersen, et al., "The matrix cookbook," *Technical University of Denmark*, vol. 7, pp. 15, 2008.

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

[31] 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, 2010.

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

[33] J. F. Sturm, "Using SeDuMi 1.02: a Matlab toolbox for optimization over symmetric cones," *Optimization methods and software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[34] R. C Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 223–233, 2012.

[35] M. Pollefeys and D. Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data.," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 2445–2448.

[36] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, 2012.

[37] J. Zhang, R. C. Hendriks, and R. Heusdens, "Structured total least squares based internal delay estimation for distributed microphone auto-localization," in *Int. Workshop Acoustic Signal Enhancement (IWAENC)*, 2016.

[38] 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.

[39] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[40] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[41] M. Wax and Y. Anu, "Performance analysis of the minimum variance beamformer," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 928–937, 1996.

[42] Y. Zhang, B. P. Ng, and Q. Wan, "Sidelobe suppression for adaptive beamforming with sparse constraint on beam pattern," *Electronics Letters*, vol. 44, no. 10, pp. 615–616, 2008.

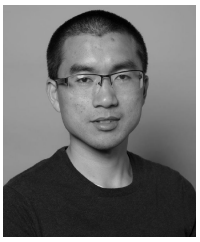
[43] M. O'Connor, W. B. Kleijn, and T. Abhayapala, "Distributed sparse MVDR beamforming using the bi-alternating direction method of multipliers," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 106–110.

[44] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

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

[46] E. A. P. Habets, "Room impulse response generator," Tech. Rep.

[47] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer networks*, vol. 38, no. 4, pp. 393–422, 2002.



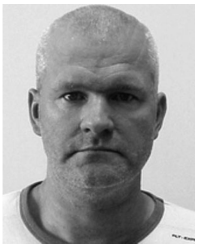
Jie Zhang was born in Anhui Province, China, in 1990. He received the M.Sc. degree from the School of Electronics and Computer Engineering, Shenzhen Graduate School, Peking University, Beijing, China. He is currently working toward the Ph.D. degree in the Circuits and Systems Group at the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands.

His current research interests include multimicrophone speech processing for noise reduction, enhancement and sound source localization, binaural auditory, energy-aware wireless (acoustic) sensor networks.

Sundeep Prabhakar Chepuri (M'16) received his M.Sc. degree (*cum laude*) in electrical engineering and Ph.D. degree (*cum laude*) from the Delft University of Technology, The Netherlands, in July 2011 and January 2016, respectively. He has held positions at Robert Bosch, India, during 2007-2009, and Holst Centre/IMEC-NL, The Netherlands, during 2010-2011. He is currently with the Circuits and Systems Group at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, The Netherlands.

Dr. Chepuri received the Best Student Paper Award for his publication at the ICASSP 2015 conference in Australia. Currently, he is an Associate Editor of the *EURASIP Journal on Advances in Signal Processing*. His general research interest lies in the field of mathematical signal processing, statistical inference, sensor networks, and wireless communications.

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 Assistant 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 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*.



Richard Heusdens received the M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively. Since 2002, he has been an Associate Professor in the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. In the spring of 1992, he joined the digital signal processing group at the Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group of Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio/speech signal processing activities within the ICT group. He held visiting positions at KTH (Royal Institute of Technology, Sweden) in 2002 and 2008 and was a guest professor at Aalborg University from 2014-2016. He is involved in research projects that cover subjects such as audio and acoustic signal processing, speech enhancement, and distributed signal processing.