Relative Acoustic Transfer Function Estimation in Wireless Acoustic Sensor Networks

Jie Zhang, Richard Heusdens, and Richard Christian Hendriks

***Abstract*—In this paper, we present an algorithm to estimate the relative acoustic transfer function (RTF) of a target source in wireless acoustic sensor networks (WASNs). Two well-known methods to estimate the RTF are the covariance subtraction (CS) method and the covariance whitening (CW) approach, the latter based on the generalized eigenvalue decomposition. Both methods depend on the use of the noisy correlation matrix, which, in practice, has to be estimated using limited and (in WASNs) quantized data. The bit rate and the fact that we use limited data records therefore directly affect the accuracy of the estimated RTFs. Therefore, we first theoretically analyze the estimation performance of the two approaches in terms of bit rate. Second, we propose a rate-distribution method by minimizing the power usage and constraining the expected estimation error for both RTF estimators. The optimal rate distributions are found by using convex optimization techniques. The model-based methods, however, are impractical due to the dependence on the true RTFs. We therefore further develop two greedy rate-distribution methods for both approaches. Finally, numerical simulations on synthetic data and real audio recordings show the superiority of the proposed approaches in power usage compared to uniform rate allocation. We find that in order to satisfy the same RTF estimation accuracy, the rate-distributed CW methods consume much less transmission energy than the CS-based methods.**

***Index Terms*—Sensor networks, relative transfer function, covariance subtraction, covariance whitening, model/data-driven rate distribution, quantization, convex optimization.**

I. Introduction

ACOUSTIC transfer function (ATF) identification is required by many algorithms in wireless acoustic sensor networks (WASNs), e.g., Wiener filtering [1]–[3] or beamforming [4]–[7] based noise reduction, or, sound source localization [8]. Often, instead of the ATF, algorithms use the relative acoustic transfer function (RTF) [5], which is obtained by normalizing the ATF with its value at the reference microphone. The RTF of a single desired source spans the signal subspace of interest and directly determines the formation of the target spatial autocorrelation matrix.

Manuscript received October 15, 2018; revised March 19, 2019 and May 20, 2019; accepted June 14, 2019. Date of publication June 18, 2019; date of current version June 28, 2019. This work was supported by the China Scholarship Council under Grant 201506010331. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maria de Diego. *(Corresponding author: Jie Zhang.)*

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@tudelft.nl; r.heusdens@tudelft.nl; r.c.hendriks@tudelft.nl).

Digital Object Identifier 10.1109/TASLP.2019.2923542

Assuming a perfect voice activity detector (VAD) is available, the microphone recordings can be classified into noise-only segments and speech+noise segments. During each of these periods, we can estimate the noise and noisy correlation matrices, respectively, using sample correlation matrices. Given the estimated noise and noisy correlation matrices and assuming that the target speech and noise signals are mutually uncorrelated, the low-rank target spatial correlation matrix (more strictly, with a rank equal to the number of target point sources of interest) can be obtained by subtracting the noise correlation matrix from the noisy correlation matrix. Most existing RTF estimation algorithms are based on the use of sample correlation matrices. Due to the estimation errors in the sample correlation matrices, particularly in noisy and reverberant environments, the autocorrelation matrix of the target sources will be full-rank in practice [1]. The estimation errors on the correlation matrices will directly affect the accuracy of the estimated RTFs.

In centralized WASNs, where all the network nodes are wirelessly connected to a fusion center (FC), the nodes need to quantize and transmit their microphone recordings to the FC. The quantization of the data is thus another source for inaccuracies when estimating the RTFs. Moreover, the number of quantization levels (i.e., the bit-rate) used to transmit data to the FC is one-to-one related to the required transmission power. The power usage is another point of concern in WASNs as typically the wireless sensors are battery-driven with limited power budget. The transmission power can be assumed to be exponentially related with the communication rate (e.g., in bits per sample) [9], [10]. Intuitively, the lower the rate, the less power is required, but the worse the RTF estimation, leading to a trade-off between RTF estimation accuracy and power consumption. In this paper, we investigate the relation between power usage required for data transmission in WASNs and the estimation accuracy of the RTFs (due to quantization errors, limited data when calculating samples covariance matrices and limited signal-to-noise ratio). As a result, we obtain an algorithm to estimate the RTF at prescribed accuracy, at low rate and low power usage.

Given the target speech correlation matrix, the RTF can be estimated by simply extracting its normalized first column vector, i.e., covariance subtraction (CS) [1], [11]–[14], or by calculating the normalized principal eigenvector [1], [8]. The idea behind the CS method is that the true speech correlation matrix is rank-1 under the assumption that only a single target speech point source is present. Alternatively, given the noise and noisy correlation matrices, we can first whiten the noisy correlation matrix using the noise correlation matrix, then the RTF can be

estimated by taking the normalized first column of the whitened noisy correlation matrix, or by computing the normalized principal eigenvector of the whitened noisy correlation matrix, i.e., covariance whitening (CW) [15]–[18]. Using the technique of generalized eigenvalue decomposition (GEVD) for a matrix pencil (i.e., noise and noisy correlation matrices), the CW method is then equivalent to extracting the normalized principal generalized eigenvector. In this work, we will only discuss the two extreme cases, i.e., 1) the CS method where the RTF is obtained by extracting the normalized first column vector and 2) the CW method where the RTF is obtained by calculating the normalized principle eigenvector of the whitened noisy correlation matrix, as the presented results can easily be extended to the other two cases. In the remainder of this work, we refer to these two cases as the CS and CW method, respectively. In general, the CW method can achieve better performance than the CS method, especially in severe noisy scenarios [13], [18]. However, the CS method is more appealing from an implementation point of view, since it only requires to extract the first column vector of a matrix, while the other one requires computationally more demanding matrix eigenvalue decompositions and/or matrix inversion. In [13] and [18], Markovich-Golan and Gannot analyzed the performance of the CS and CW methods using synthetic non-stationary Gaussian signals, respectively. We will take the performance analysis of both methods as the basis of the energy-aware RTF estimation procedures that are presented in this work.

A. Contributions

The contributions of this paper can be summarized as follows. Firstly, we briefly analyze the performance of the CS method and the CW method in a theoretical fashion, with quantization noise being taken into account. This is based on the work presented in [13], [18]. It is shown that the estimation errors of both methods are related to the signal-to-noise ratio (SNR), the communication rate and the number of available segments which are used to estimate the second-order statistics (SOS). We show that the CW always performs better than the CS method. This is because the performance of the CW method depends on the output SNR of a minimum variance distortionless response (MVDR) beamformer, while the CS method depends in a similar way on the input SNR, which is always lower than the MVDR output SNR.

Secondly, based on the framework presented in [19], we develop for both the CS and CW approach a model-driven rate-distribution algorithm for RTF estimation in WASNs, referred to as MDRD-CS and MDRD-CW. The model-driven problems are formulated by minimizing the total transmission costs between all microphone nodes and the FC and constraining the expected RTF estimation performance. Using convex optimization techniques, the MDRD-CS/CW problems are derived as semi-definite programs. Through distributing bit rates optimally, the transmission cost in WASNs can be saved significantly compared to a blind full-rate transmission strategy, meanwhile satisfying the prescribed desired estimation performance on the RTF. Note that the MDRD-CS/CW methods depend on the true RTF and noise SOS, which are unknown in practice. The proposed model-driven methods are thus not practical from the perspective of implementation.

To make the model-based methods practical, we further propose two corresponding data-driven methods (i.e., DDRD-CS and DDRD-CW), which are (performance-wise) near-optimal and use a greedy rate distribution strategy, but only rely on realizations. Since the microphone nodes send the quantized data to the FC frame-by-frame, we can estimate the RTF and noise SOS using the previously received segments, and then solve the model-driven problems based on the estimated RTF and noise SOS. Then, each node quantizes the new segment at the rate that is obtained by the model-driven method. As such, the data-driven methods can avoid the dependence on the true RTF and noise SOS.

Finally, the proposed approaches are validated via numerical simulations in a simulated WASN. We find that both the MDRD-CS and the MDRD-CW satisfy the performance requirement, and the DDRD-CS (or DDRD-CW) method converges to the MDRD-CS (or MDRD-CW) method when increasing the number of available segments. We conclude that the sensors that are closer to the FC are more likely to be allocated with a higher rate, since they are cheaper in transmission. Besides, we show that at higher bit-rates, redundant information is transmitted, as the performance of CS/CW-based methods does not gain much with increasing bit rate. Hence, the proposed methods can reduce the redundant bits and save energy usage compared to the unnecessary full-rate quantization. Furthermore, it is shown that given the same performance requirement, the MDRD-CW (or DDRD-CW) method consumes much less transmission energy compared to the MDRD-CS (or DDRD-CS) method.

B. Outline and Notation

The paper is structured as follows. Section II presents preliminaries on the signal model and the estimation of sample correlation matrices. In Section III, we theoretically analyze the performance of the CS/CW-based RTF estimators. Section IV formulates the rate-distributed RTF estimation problem and solves it in the context of the CS and CW methods, respectively. In Section V, we show the proposed greedy methods. The proposed methods are validated in Section VI via numerical simulations. Finally, Section 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. $(\cdot)^*$ denotes the conjugate of a complex number. $\mathrm{diag}(\cdot)$ refers to a block diagonal matrix with the elements in its argument on the main diagonal. $\mathbf{I}_N$ and $\mathbf{O}_N$ denote the identity matrix and the $N \times N$ matrix with all its elements equal to zero, respectively. $\mathbf{e}_1$ is a column vector with 1 at the first entry and zeros elsewhere. $\mathbf{0}_N$ is an $N \times 1$ all-zeros column vector. $\mathbb{E}\{\cdot\}$ denotes the statistical expectation operation. $\mathrm{Tr}(\cdot)$ and $\mathrm{rank}(\cdot)$ denote the trace and rank of a matrix, respectively. $||\cdot||_2$ denotes the ℓ_2 norm. $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. Furthermore, $\odot$ denotes the Hadamard (elementwise) product. $\hat{X}$ and $\tilde{X}$ denote the estimate of a random variable X and the corresponding estimation error, respectively.

II. FUNDAMENTALS

A. Signal Model

We consider K microphones that sample the sound field consisting of one target point source, degraded by acoustic background noise. In the short-time Fourier transform (STFT) domain, letting l and ω denote the index of time frame and angular frequency, respectively, the noisy DFT coefficient at the kth microphone, say $Y_k(\omega,l), k=1,\ldots,K$, is given by

$$Y_k(\omega,l) = X_k(\omega,l) + U_k(\omega,l), \tag{1}$$

where $X_k(\omega,l) = a_k(\omega)S(\omega,l)$ with $a_k(\omega)$ the ATF of the target signal with respect to the kth microphone and $S(\omega,l)$ the DFT coefficient of the target source signal at the source location. In this work we assume that the ATF is time-invariant, i.e., the target source is assumed static, during the time period of interest. Therefore, $a_k(\omega)$ is not a function of l. In (1), the term $U_k(\omega,l)$ represents the total received noise at the kth microphone (including interfering sources and sensor noise). In this work, the noise signals contained in $U_k(\omega,l)$ are assumed stationary during the time period of interest. This assumption is not strictly necessary for the theory that we will derive. However, the expressions that we present depend on the SOS that can only be estimated if the sources are stationary for a fixed period of, say L time-frames. In a centralized WASN, we assume that a FC is employed to collect data and process the tasks at hand. In this case, the microphone nodes need to transmit their recordings to the FC, and the recordings should be quantized at specified communication rates. Taking the utilization of quantizers into account and letting $Q_k(\omega,l)$ denote the quantization noise[1] contained in the transmitted data from the kth microphone node, the quantized version of the kth microphone measurements that is received by the FC is given by

$$\hat{Y}_k(\omega,l) = X_k(\omega,l) + U_k(\omega,l) + Q_k(\omega,l). \tag{2}$$

Note that the quantization takes place in the STFT domain directly. Given a bit-rate, the real and imaginary parts of $Y_k(\omega,l)$ are quantized separately, as the bit-rate is equally distributed to the real and imaginary parts [20]. A more optimal but complicated rate distribution for quantizing complex Gaussian random variables can be found in [21]. 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, the quantized signals from the K microphones are stacked in a vector $\hat{\mathbf{y}} = [\hat{Y}_1,\ldots,\hat{Y}_K]^T \in \mathbb{C}^K$. Similarly, we define K dimensional vectors $\mathbf{y}$, $\mathbf{x}$, $\mathbf{u}$, $\mathbf{q}$ and $\mathbf{a}$ for the microphone recordings, the target speech component, the received noises by the microphones, the quantization noise and the ATFs, respectively, such that (2) can be rewritten as

$$\hat{\mathbf{y}} = \mathbf{a}S + \mathbf{u} + \mathbf{q}, \tag{3}$$

with the clean speech component given by $\mathbf{x} = \mathbf{a}S$. Furthermore, we define $\mathbf{n} = \mathbf{u} + \mathbf{q}$ as the total noise at the FC including quantization noise. Without loss of generality, we assume that the first microphone is taken as the reference microphone. The RTF can then be defined as

$$\mathbf{d} = \mathbf{a}/a_1, \tag{4}$$

where a_1 refers to the first entry of vector $\mathbf{a}$.

B. Estimating Sample Covariance Matrices

We assume that the quantization noise is uncorrelated with the microphone recording,[2] and that the noise components and the target signal are mutually uncorrelated, such that from the signal model (2), the SOS of the noisy microphone signals during speech+noise segments are given by

$$\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbb{E}\{\hat{\mathbf{y}}\hat{\mathbf{y}}^H\} = \mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{qq}}. \tag{5}$$

Further, the SOS of the noise are given by

$$\mathbf{R}_{\mathbf{nn}} = \mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{qq}}. \tag{6}$$

Assuming that the speech and noise signals are mutually uncorrelated, $\mathbf{R}_{\mathbf{xx}}$ can be calculated as

$$\begin{aligned}\mathbf{R}_{\mathbf{xx}} &\triangleq \sigma_S^2\mathbf{a}\mathbf{a}^H = \sigma_{X_1}^2\mathbf{d}\mathbf{d}^H \\ &= \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\mathbf{nn}},\end{aligned} \tag{7}$$

with $\sigma_S^2 = \mathbb{E}\{|S|^2\}$ and $\sigma_{X_1}^2 = \mathbb{E}\{|X_1|^2\}$, respectively, representing the power spectral density (PSD) of the target source and the PSD of the speech component at the reference microphone. Obviously, we have the relation $\sigma_{X_1}^2 = |a_1|^2\sigma_S^2$. Note that $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{nn}}$ are full-rank (positive definite) matrices, and $\mathrm{rank}\,(\mathbf{R}_{\mathbf{xx}}) = 1$ in a single speech point source scenario. More importantly, both $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{nn}}$ depend on $\mathbf{R}_{\mathbf{qq}}$, while $\mathbf{R}_{\mathbf{xx}}$ does not. From (5) and (6), we know that the communication rate affects $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{nn}}$ by the addition of the matrix $\mathbf{R}_{\mathbf{qq}}$. Hence, in case $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ are perfectly estimated (e.g., given sufficiently long data measurements), $\mathbf{R}_{\mathbf{qq}}$ can be eliminated by calculating $\mathbf{R}_{\mathbf{xx}}$ with the subtractive operation in (7).

In practice, given L speech+noise segments, the SOS $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ can be estimated by average smoothing, that is

$$\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{L}\sum_{l=1}^{L}\hat{\mathbf{y}}_l\hat{\mathbf{y}}_l^H. \tag{8}$$

The SOS estimator in (8) is unbiased and the corresponding estimation error is denoted by

$$\tilde{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}. \tag{9}$$

Similarly, we can estimate $\mathbf{R}_{\mathbf{nn}}$ by

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \frac{1}{|\mathcal{T}|}\sum_{l\in\mathcal{T}}\mathbf{n}_l\mathbf{n}_l^H, \tag{10}$$

where $\mathcal{T}$ indicates a set of noise-only time segments. However, to make the analysis on the CS and CW method consistent, we will assume that $\mathbf{R}_{\mathbf{nn}}$ is known and can be used to estimate the RTF vector. This could be argued for under conditions of relatively stationary noise sources. In that case, $\mathbf{R}_{\mathbf{nn}}$ can be estimated with relatively small error as sufficiently long time segments can be used. The assumption that $\mathbf{R}_{\mathbf{nn}}$ is known is required in the derivation of the CW-based RTF estimation accuracy. However, in the derivation of the CS-based RTF estimation accuracy

[1] In real-life applications, $Y_k(\omega,l)$ is already quantized, since it is acquired by the analog-to-digital converter of the kth sensor. In this case, $Q_k(\omega,l)$ would represent the error from changing the bit resolution of $Y_k(\omega,l)$.

[2] This assumption holds under high rate communication. At low rates, this can be achieved by applying subtractive dither [22], [23].

this assumption is strictly speaking not necessary and expressions can also be derived taking estimation errors on $\mathbf{R_{nn}}$ into account. In the derivation of the estimation accuracy under the CW approach it is not trivial to take both estimation errors on $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R_{nn}}$ into account. As such this is a disadvantage of the CW approach. However in order to make comparison of both methods possible, we make the same assumption in both methods. From now on we therefore assume $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is estimated and $\mathbf{R_{nn}}$ is known. However, in Section III-A, for completeness, we will give the expressions for the CS estimation accuracy when also $\mathbf{R_{nn}}$ is estimated. With $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R_{nn}}$ at hand, using (7) we can obtain the estimate of $\hat{\mathbf{R}}_{\mathbf{xx}}$ by

$$\hat{\mathbf{R}}_{\mathbf{xx}} \triangleq \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R_{nn}}, \tag{11}$$

which can be reformulated as

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \mathbf{R_{xx}} + \tilde{\mathbf{R}}_{\mathbf{xx}}, \tag{12}$$

with $\tilde{\mathbf{R}}_{\mathbf{xx}} = \tilde{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$. Although $\mathrm{rank}(\mathbf{R_{xx}}) = 1$, in practice we have $\mathrm{rank}(\hat{\mathbf{R}}_{\mathbf{xx}}) > 1$ due to the estimation error in $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$. The RTF estimators presented in the sequel are based on the SOS $\mathbf{R_{xx}}$, $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R_{nn}}$, whereas in practice these matrices are replaced by the sample correlation matrices $\hat{\mathbf{R}}_{\mathbf{xx}}$, $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$.

For the SOS of the quantization noise, we assume that each microphone node employs a uniform quantizer for quantization, such that given b_k bits per sample, the PSD of the quantization noise is given by [24], [25]

$$\sigma_{q_k}^2 = \Delta_k^2/12, \forall k, \tag{13}$$

where the uniform intervals have width $\Delta_k = \mathcal{A}_k/2^{b_k}$ with $\mathcal{A}/2$ denoting the maximum absolute value of the kth microphone measurement. Assuming that the quantization noise across microphones is mutually uncorrelated, the correlation matrix of the quantization noise across microphones reads

$$\mathbf{R_{qq}} = \frac{1}{12} \times \mathrm{diag}\left(\left[\frac{\mathcal{A}_1^2}{4^{b_1}}, \frac{\mathcal{A}_2^2}{4^{b_2}}, \ldots, \frac{\mathcal{A}_K^2}{4^{b_K}}\right]\right). \tag{14}$$

III. Performance Analysis for RTF Estimators

In this section, we will theoretically analyze the RTF estimation performances of the CS method and the CW method, which is based on the work presented in [13] and [18], respectively, which we extend by taking quantization noise into account. The estimation accuracy is defined as the ratio between the expected squared norms of the error vector $\tilde{\mathbf{d}}$ and the true RTF vector as [13]

$$\epsilon \triangleq \mathbb{E}[||\tilde{\mathbf{d}}||_2^2]/||\mathbf{d}||_2^2. \tag{15}$$

A. Performance Analysis for CS Method

The CS method takes the normalized first column of the matrix $\hat{\mathbf{R}}_{\mathbf{xx}}$ as the RTF estimate [1], [11], i.e.,

$$\hat{\mathbf{d}}_{\mathrm{CS}} \triangleq \frac{\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}, \tag{16}$$

which is based on the rank-1 model for the clean-speech correlation matrix $\mathbf{R_{xx}}$. The denominator of (16) represents the signal power at the reference microphone, i.e.,

$$\hat{\sigma}_{X_1}^2 \triangleq \mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1. \tag{17}$$

In order to analyze the CS-based RTF estimator, we write the RTF estimate from (16) as

$$\hat{\mathbf{d}}_{\mathrm{CS}} = \mathbf{d} + \tilde{\mathbf{d}}_{\mathrm{CS}}. \tag{18}$$

In [18], it was shown that the estimation error term $\tilde{\mathbf{d}}_{\mathrm{CS}}$ is given by

$$\tilde{\mathbf{d}}_{\mathrm{CS}} = \frac{1}{|a_1|^2\hat{\sigma}_S^2}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\tilde{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1. \tag{19}$$

Assuming the estimation error $\tilde{\mathbf{R}}$ of the covariance matrix $\mathbf{R}$ of a Gaussian random variable when estimated as in (8) obeys a complex Wishart distribution [26], it can be shown (see [18]) that given the noise SOS $\mathbf{R_{nn}}$, the RTF estimation error ϵ_{CS} of the CS-based method from (15) is given by [13], [18]

$$\epsilon_{\mathrm{CS}} = \frac{1+\frac{1}{\eta}}{L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2} \cdot \mathrm{Tr}\left(\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R_{nn}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H\right), \tag{20}$$

where η is referred to as the signal-to-(total)noise ratio at the reference microphone, i.e.,

$$\eta \triangleq \frac{\hat{\sigma}_{X_1}^2}{\mathbf{e}_1^T\mathbf{R_{nn}}\mathbf{e}_1} = \frac{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}{\mathbf{e}_1^T\mathbf{R_{nn}}\mathbf{e}_1}. \tag{21}$$

Finally, taking the quantization noise into account as $\mathbf{R_{nn}} = \mathbf{R_{uu}} + \mathbf{R_{qq}}$, and for readability, defining

$$\mathbf{G} = \left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\left(\mathbf{R_{uu}} + \mathbf{R_{qq}}\right)\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H,$$

such that the final CS error model can be formulated as

$$\epsilon_{\mathrm{CS}} = \frac{1+\frac{1}{\eta}}{L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2} \cdot \mathrm{Tr}\left(\mathbf{G}\right). \tag{22}$$

Note that (22) differs from the one in [13] by the facts that 1) quantization noise is taken into account 2) similar as in [18] we assume $\mathbf{R_{nn}}$ to be known (estimated based on larger data records), resulting in the term $\frac{1}{\eta}$ in (22).

Further, in case $\mathbf{R_{nn}}$ is estimated based on a different number of frames, say $T = |\mathcal{T}|$ frames, that are different (independent) from the L frames used to estimate $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$, we obtain

$$\epsilon_{\mathrm{CS}} = \frac{\frac{1}{L}+\frac{1}{\eta}\left(\frac{1}{L}+\frac{1}{T}\right)}{||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2} \cdot \mathrm{Tr}\left(\mathbf{G}\right). \tag{23}$$

If $L = T$, (23) will be identical to the error model derived in [13].

B. Performance Analysis for CW Method

The CW method takes the normalized principal eigenvector of the whitened noisy covariance matrix as the estimated RTF, which is given by

$$\hat{\mathbf{d}}_{\mathrm{CW}} = \frac{\mathbf{R}_{\mathbf{nn}}^{H/2}\hat{\boldsymbol{\psi}}}{\mathbf{e}_1^T\mathbf{R}_{\mathbf{nn}}^{H/2}\hat{\boldsymbol{\psi}}}, \tag{24}$$

where $\hat{\boldsymbol{\psi}}$ is the principal eigenvector of the matrix $\hat{\mathbf{R}}_{\mathbf{zz}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{z}\mathbf{z}^H$ with $\mathbf{z} = \mathbf{R}_{\mathbf{nn}}^{-H/2}\hat{\mathbf{y}}$. In [18], it was shown that the error vector of the CW method can be approximated by

$$\tilde{\mathbf{d}}_{\mathrm{CW}} = \frac{\theta}{a_1}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}^{H/2}\tilde{\boldsymbol{\psi}}, \tag{25}$$

where $\theta = \sqrt{\mathbf{a}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{a}}$, and $\tilde{\boldsymbol{\psi}}$ denotes the estimation error vector of the principal eigenvector, and its covariance matrix is given

by [27]

$$\mathbf{\Theta}_{\boldsymbol{\psi}} = \frac{\lambda_1}{L(\lambda_1 - 1)^2}\left(\mathbf{I} - \boldsymbol{\psi}\boldsymbol{\psi}^H\right), \tag{26}$$

where $\lambda_1 = \mathbf{a}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{a}\hat{\sigma}_S^2 + 1$ denotes the principal eigenvalue, and the true principal eigenvector is given by $\boldsymbol{\psi} = \mathbf{R}_{\mathbf{nn}}^{-H/2}\mathbf{a}/\theta$. Hence, the covariance matrix of $\tilde{\mathbf{d}}_{\mathrm{CW}}$ can be formulated as

$$\begin{aligned}\mathbf{\Theta} &\overset{(a)}{=} \frac{|\theta|^2}{|a_1|^2}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}^{\frac{H}{2}}\mathbf{\Theta}_{\boldsymbol{\psi}}\mathbf{R}_{\mathbf{nn}}^{\frac{1}{2}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H \\ &\overset{(b)}{=} \frac{1 + \frac{1}{\hat{\sigma}_{X_1}^2\mathbf{d}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}}}{L\hat{\sigma}_{X_1}^2}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H, \end{aligned} \tag{27}$$

where (a) is obtained by substitution of (25) and (b) is due to the fact that $(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T)\mathbf{d} = \mathbf{0}_K$. Finally, taking the quantization noise into account, we can formulate the CW-based RTF estimation error as

$$\epsilon_{\mathrm{CW}} = \frac{\mathrm{Tr}(\mathbf{\Theta})}{||\mathbf{d}||_2^2} = \frac{1 + \frac{1}{\hat{\sigma}_{X_1}^2\mathbf{d}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}}}{L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2} \cdot \mathrm{Tr}\left(\mathbf{G}\right). \tag{28}$$

Note that in fact the term $\hat{\sigma}_{X_1}^2\mathbf{d}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}$ is the output SNR of an MVDR beamformer [4], [28]–[30].

Remark 1: By inspection, the estimation errors of both the CS method and the CW method are influenced by the SNR, frame length and communication rate, the signal power and the location of source, i.e., $||\mathbf{d}||_2^2$. The final expression in (22) or (28) differs from the one derived in [13], [18] by the fact that the quantization noise is now also taken into account. Comparing (28) to (22), the only difference lies in the SNR term. Since after the use of an MVDR beamformer, the SNR can be improved, i.e., $\eta \leq \hat{\sigma}_{X_1}^2\mathbf{d}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}$, we can conclude that the CW-based RTF estimator always achieves a higher accuracy than the CS method.

IV. Model-Driven Rate-Distributed Methods

In this section, we first present the transmission energy model, and then formulate the general rate-distributed RTF estimation problem. Finally, we propose convex optimization approaches for the resulting rate distribution problems for the CS-based and CW-based methods.

A. Transmission Energy Model

In WASNs, the sensors transmit data to the FC via wireless links, and the communication channels are inevitably corrupted by additive noise. Let us assume that the transmission channel noise is white Gaussian with PSD $V_k, \forall k$. Given a transmitted power E_k from the kth microphone node in the WASN, the received energy by the FC will be $D_k^{-r}E_k$ with D_k and r denoting the transmission distance from the kth microphone to the FC and the path loss exponent, respectively. Typically, $2 \leq r \leq 6$ [9], [31]. We assume $r = 2$ throughout this work without loss of generality. The loss in the received energy is caused by the channel power attenuation. With these, the SNR of the kth channel can be formulated as

$$\mathrm{SNR}_k = D_k^{-2}E_k/V_k, \forall k, \tag{29}$$

which is different from the acoustic noise or acoustic SNR that is mentioned before. Assuming that the transmitted speech signals are Gaussian distributed in the STFT domain, the capacity based on the Shannon theory [32] for Gaussian channels is then given by

$$b_k = \frac{1}{2}\log_2\left(1 + \mathrm{SNR}_k\right), \forall k, \tag{30}$$

which is valid for one frequency bin. To achieve reliable transmissions, b_k bits per sample at most can be transmitted from microphone k to the FC at each frequency bin. Based on the channel SNR (29) and the capacity (30), we can formulate the transmitted energy as [9], [10], [19], [20], [33]

$$E_k = D_k^2V_k(4^{b_k} - 1), \forall k. \tag{31}$$

Notice that the above energy model holds under two conditions [9], [10]: 1) band-limited input signals, and 2) the microphone recordings are quantized at the channel capacity.

B. General Problem Formulation

The proposed model-driven rate-distributed RTF estimation method is formulated by minimizing the total transmission costs while constraining the RTF estimation error, which can be expressed as the following optimization problem:

$$\begin{aligned}\min_{\mathbf{b}} \quad & \sum_{k=1}^{K} D_k^2V_k(4^{b_k} - 1) \\ \text{s.t.} \quad & \epsilon_{\mathrm{CS/CW}} \leq \frac{\beta}{\alpha}, \\ & b_k \in \mathbb{Z}_+, b_k \leq b_{\max}, \forall k, \end{aligned} \tag{P1}$$

where $\epsilon_{\mathrm{CS/CW}}$ indicates the use of either ϵ_{CS} or ϵ_{CW} from (22) and (28), respectively, $\mathbb{Z}_+$ denotes a non-negative integer set, $b_{\max}$ the maximum rate, and β the optimal performance, which can be the RTF estimation error of the CS or CW-based method when all the sensor measurements are quantized at the maximum bit rate, and $\alpha \in (0, 1]$ is the parameter to control the desired performance. In practice, β/α is just a number, which can be assigned by users, not necessarily dependent on the optimal performance. By solving (P1), we can determine the optimal rate distribution that the microphone nodes can utilize to quantize their recordings, such that a desired RTF estimation accuracy is achieved with minimum energy usage. One way to solve (P1) is exhaustive search, i.e., evaluating the performance for all $(b_{\max} + 1)^K$ possible candidate rate distributions, but evidently this is intractable unless $b_{\max}$ or/and K are very small. Note that (P1) is formulated per frequency bin. Also, (P1) is non-convex due to the facts that:

- the constraint $\epsilon_{\mathrm{CS/CW}} \leq \frac{\beta}{\alpha}$ is non-linear in $\mathbf{b}$;
- the bit-rate $\mathbf{b}$ is constrained to be integer valued.

Next, we will solve (P1) using convex optimization techniques in the context of the CS and CW methods, respectively.

C. Model-Driven Rate-Distributed CS (MDRD-CS)

For the first constraint $\epsilon_{\mathrm{CS}} \leq \frac{\beta}{\alpha}$ in (P1), using the expression ϵ_{CS} from (22), we can rewrite it as

$$c_1 \cdot \left[c_2 + \mathrm{Tr}\left(\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{qq}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H\right)\right] \leq \frac{\beta}{\alpha},$$

or rearranged as

$$\mathrm{Tr}\left(\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)\mathbf{R_{qq}}\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H\right)\leq\frac{\beta}{\alpha c_1}-c_2, \tag{32}$$

where the constants c_1 and c_2 are given by

$$c_1=\frac{1+\frac{1}{\eta}}{L||\mathbf{a}||_2^2\hat{\sigma}_S^2}=\frac{1+\frac{1}{\eta}}{L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2}, \tag{33}$$

$$c_2=\mathrm{Tr}\left(\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)\mathbf{R_{uu}}\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H\right). \tag{34}$$

Clearly, (32) is non-convex and non-linear in terms of the bit rates $b_k,\forall k$. For linearization, we equivalently rewrite (32) into two new constraints by introducing a new Hermitian positive semi-definite matrix $\mathbf{Z}\in\mathbb{S}_+^K$ with $\mathbb{S}_+$ denoting the set of Hermitian positive semi-definite matrices, i.e.,

$$\mathrm{Tr}\left(\mathbf{Z}\right)\leq\frac{\beta}{\alpha c_1}-c_2, \tag{35}$$

$$\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)\mathbf{R_{qq}}\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H=\mathbf{Z}. \tag{36}$$

Now, (35) is linear in the new variable $\mathbf{Z}$, however, (36) is still non-convex in b_k. To convexify (36), we can relax it to

$$\mathbf{Z}\succeq\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)\mathbf{R_{qq}}\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H, \tag{37}$$

since (37) and (35) are sufficient to obtain the original constraint in (32). By inspection, (37) can be written as a linear matrix inequality (LMI) using the Schur complement [34, p.650], i.e.,

$$\begin{bmatrix}\mathbf{R_{qq}^{-1}} & \mathbf{I}-\mathbf{d}\mathbf{e}_1^T\\ \left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H & \mathbf{Z}\end{bmatrix}\succeq\mathbf{O}_{2K}, \tag{38}$$

where $\mathbf{R_{qq}^{-1}}$ can be computed from (14) as

$$\mathbf{R_{qq}^{-1}}=12\times\mathrm{diag}\left(\left[\frac{4^{b_1}}{\mathcal{A}_1^2},\frac{4^{b_2}}{\mathcal{A}_2^2},\ldots,\frac{4^{b_K}}{\mathcal{A}_K^2}\right]\right). \tag{39}$$

Note that (38) is not an LMI in the unknown parameters $\mathbf{b}$, but in $4^{b_k},\forall k$. Finally, we define a constant vector $\mathbf{f}=[\frac{12}{\mathcal{A}_1^2},\cdots,\frac{12}{\mathcal{A}_K^2}]^T$ and introduce a variable change $t_k=4^{b_k}\in\mathbb{Z}_+,\forall k$, such that $\mathbf{R_{qq}^{-1}}=\mathrm{diag}\left(\mathbf{f}\odot\mathbf{t}\right)$ and (38) are both linear in $\mathbf{t}$. For the integer constraint $b_k\in\mathbb{Z}_+,\forall k$, we relax it to $b_k\in\mathbb{R}_+$, i.e., $t_k\in\mathbb{R}_+,\forall k$. Altogether, we obtain a standard semi-definite programming (SDP) problem [34, p.128] as

$$\begin{aligned}\min_{\mathbf{t},\mathbf{Z}}\quad&\sum_{k=1}^K D_k^2V_k(t_k-1)\\ \text{s.t.}\quad&\mathrm{Tr}\left(\mathbf{Z}\right)\leq\frac{\beta}{\alpha c_1}-c_2,\\ &\begin{bmatrix}\mathrm{diag}\left(\mathbf{f}\odot\mathbf{t}\right) & \mathbf{I}-\mathbf{d}\mathbf{e}_1^T\\ \left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H & \mathbf{Z}\end{bmatrix}\succeq\mathbf{O}_{2K},\\ &1\leq t_k\leq 4^{b_{\max}},\ \forall k.\end{aligned} \tag{P2}$$

D. Model-Driven Rate-Distributed CW (MDRD-CW)

Applying the expression from (28) to (P1), one can consider the MDRD-CW problem. Then, the first constraint $\epsilon_{\mathrm{CW}}\leq\frac{\beta}{\alpha}$ in (P1) can be rewritten as

$$\mathrm{Tr}\left(\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)\mathbf{R_{qq}}\left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H\right)\leq\frac{\beta}{\alpha c_1'}-c_2, \tag{40}$$

where c_1' is defined by

$$c_1'=\frac{1+\frac{1}{\hat{\sigma}_{X_1}^2\mathbf{d}^H\mathbf{R_{nn}^{-1}}\mathbf{d}}}{L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2}, \tag{41}$$

and $\mathbf{R_{nn}^{-1}}$ can be calculated as

$$\begin{aligned}\mathbf{R_{nn}^{-1}}&\overset{(a)}{=}\left(\mathbf{R_{uu}}+\mathbf{R_{qq}}\right)^{-1}\\ &\overset{(b)}{=}\mathbf{R_{uu}^{-1}}-\mathbf{R_{uu}^{-1}}\left(\mathbf{R_{uu}^{-1}}+\mathbf{R_{qq}^{-1}}\right)^{-1}\mathbf{R_{uu}^{-1}},\end{aligned} \tag{42}$$

where (b) is derived from the matrix inversion lemma [35, p.18].[3] Similar to Section IV-C, by introducing a matrix $\mathbf{Z}\in\mathbb{S}_{++}^K$, (40) can equivalently be rewritten into two new constraints, e.g., (35) and (36), and the latter one can be relaxed to the LMI in (38).

Further, due to the fact that the unknown rates also sit in c_1' and c_1' is non-convex in terms of the bit rate $\mathbf{b}$, we relax (41) as

$$c_1'\geq\frac{1+\frac{1}{\hat{\sigma}_{X_1}^2\mathbf{d}^H\mathbf{R_{nn}^{-1}}\mathbf{d}}}{L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2}. \tag{43}$$

With the substitution of the expression for $\mathbf{R_{nn}^{-1}}$ from (42) into (43), we obtain

$$\delta\geq\mathbf{d}^H\mathbf{R_{uu}^{-1}}\left(\mathbf{R_{uu}^{-1}}+\mathbf{R_{qq}^{-1}}\right)^{-1}\mathbf{R_{uu}^{-1}}\mathbf{d}, \tag{44}$$

where δ is given by

$$\delta=\mathbf{d}^H\mathbf{R_{uu}^{-1}}\mathbf{d}-\frac{1/\hat{\sigma}_{X_1}^2}{c_1'L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2-1}. \tag{45}$$

Using the Schur complement, (44) can be reformulated as the following LMI:

$$\begin{bmatrix}\mathbf{R_{uu}^{-1}}+\mathbf{R_{qq}^{-1}} & \mathbf{R_{uu}^{-1}}\mathbf{d}\\ \mathbf{d}^H\mathbf{R_{uu}^{-1}} & \delta\end{bmatrix}\succeq\mathbf{O}_{K+1}. \tag{46}$$

Note that (45) is non-convex in c_1', which can be relaxed to

$$\delta\leq\mathbf{d}^H\mathbf{R_{uu}^{-1}}\mathbf{d}-\frac{1/\hat{\sigma}_{X_1}^2}{c_1'L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2-1}, \tag{47}$$

since (47) and (44) are sufficient conditions for obtaining (40). As a consequence, the MDRD-CW problem can also be formulated as an SDP problem:

$$\begin{aligned}\min_{\mathbf{t},\mathbf{Z},c_1',\delta}\quad&\sum_{k=1}^K D_k^2V_k(t_k-1)\\ \text{s.t.}\quad&\mathrm{Tr}\left(\mathbf{Z}\right)\leq\frac{\beta}{\alpha c_1'}-c_2,\\ &\begin{bmatrix}\mathrm{diag}\left(\mathbf{f}\odot\mathbf{t}\right) & \mathbf{I}-\mathbf{d}\mathbf{e}_1^T\\ \left(\mathbf{I}-\mathbf{d}\mathbf{e}_1^T\right)^H & \mathbf{Z}\end{bmatrix}\succeq\mathbf{O}_{2K},\\ &\begin{bmatrix}\mathbf{R_{uu}^{-1}}+\mathrm{diag}\left(\mathbf{f}\odot\mathbf{t}\right) & \mathbf{R_{uu}^{-1}}\mathbf{d}\\ \mathbf{d}^H\mathbf{R_{uu}^{-1}} & \delta\end{bmatrix}\succeq\mathbf{O}_{K+1},\\ &\frac{1/\hat{\sigma}_{X_1}^2}{c_1'L||\mathbf{d}||_2^2\hat{\sigma}_{X_1}^2-1}-\mathbf{d}^H\mathbf{R_{uu}^{-1}}\mathbf{d}+\delta\leq 0,\\ &1\leq t_k\leq 4^{b_{\max}},\ \forall k.\end{aligned} \tag{P3}$$

[3] $\left(\mathbf{A}+\mathbf{CBC}^T\right)^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{A}^{-1}$.

Remark 2: Both the MDRD-CS problem in (P2) and the MDRD-CW problem in (P3) can be solved in polynomial time using interior-point methods or solvers, like CVX [36] or SeDuMi [37]. The computational complexity for solving both problems is of the order of $\mathcal{O}(K^3)$. After (P2) or (P3) is solved, the allocated bit rates can be resolved by $b_k = \log_4 t_k, \forall k$. Since the solution of (P2) or (P3) are continuous values, we need to further refine the rates. We recommend to utilize randomized rounding, since this technique can guarantee that the integer solution obtained in this way always satisfies the performance requirement. The randomized rounding technique is detailed in [19], [38], the complexity of which is linear in K.

V. Greedy Rate-Distributed Methods

Strictly speaking, the MDRD-CS/CW estimators proposed in the previous section are not practical, since the rate-distribution solver in (P2) or (P3) depends on the signal power $\sigma_{X_1}^2$, the true RTF $\mathbf{d}$, SNR and noise SOS $\mathbf{R_{uu}}$. Although we can estimate $\sigma_{X_1}^2$, SNR and $\mathbf{R_{uu}}$ in practice using the microphone measurements, we have no knowledge on $\mathbf{d}$. However, the model-driven methods can provide a lower bound on the optimal rate distribution that we can achieve with the constraint on the RTF estimation performance. Based on the model-driven estimators, we will propose two practical low-rate RTF estimators in this section, which are referred to as the data-driven rate-distributed CS/CW methods (i.e., DDRD-CS and DDRD-CW, respectively). In what follows, we will take the DDRD-CS algorithm as an example to clarify the proposed greedy methods, because the updating procedures for both methods are similar.

Due to the fact that the microphone nodes quantize and transmit their recordings to the FC on a frame-by-frame basis, we can update the rate distribution at the FC end using the previously received data and estimated RTF. In detail, for the first time frame,[4] we initialize the bit rates at the maximum rate, and the microphone nodes quantize data at the initial rates. At the FC end, we can estimate the initial correlation matrices $\hat{\mathbf{R}}_{\mathbf{qq}}$, $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\hat{\mathbf{R}}_{\mathbf{xx}}$ using (14), (8) and (11), respectively. Also, we can compute the signal power $\hat{\sigma}_{X_1}^2$ and the SNR at the reference microphone $\hat{\eta}$ using (17) and (21), respectively. Based on the estimate of $\hat{\mathbf{R}}_{\mathbf{xx}}$, we can extract its normalized first column as the estimated RTF, i.e., $\hat{\mathbf{d}}_{\mathrm{CS}}$, using (16). Using this information, we can update the constants c_1 and c_2 as

$$\hat{c}_1 = \frac{1 + \frac{1}{\hat{\eta}}}{l||\hat{\mathbf{d}}||_2^2 \hat{\sigma}_{X_1}^2}, \tag{48}$$

$$\hat{c}_2 = \mathrm{Tr}\left((\mathbf{I} - \hat{\mathbf{d}}\mathbf{e}_1^T)(\mathbf{R_{nn}} - \hat{\mathbf{R}}_{\mathbf{qq}})(\mathbf{I} - \hat{\mathbf{d}}\mathbf{e}_1^T)^H\right), \tag{49}$$

where l denotes the number of received segments by the FC, e.g., in the initial case $l = 1$, and the estimate of the acoustic noise statistics is given by $\hat{\mathbf{R}}_{\mathbf{uu}} = \mathbf{R_{nn}} - \hat{\mathbf{R}}_{\mathbf{qq}}$. Based on these, we can update the rate distribution by solving (P2), i.e.,

$$\begin{aligned} \min_{\mathbf{t},\mathbf{Z}} \quad & \sum_{k=1}^{K} D_k^2 V_k(t_k - 1) \\ \text{s.t.} \quad & \mathrm{Tr}\,(\mathbf{Z}) \le \frac{\beta}{\alpha \hat{c}_1} - \hat{c}_2, \\ & \begin{bmatrix} \mathrm{diag}\,(\mathbf{f} \odot \mathbf{t}) & \mathbf{I} - \hat{\mathbf{d}}\mathbf{e}_1^T \\ \left(\mathbf{I} - \hat{\mathbf{d}}\mathbf{e}_1^T\right)^H & \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{2K}, \\ & 1 \le t_k \le 4^{b_{\max}}, \ \forall k. \end{aligned} \tag{50}$$

Note that (50) is an instantaneous optimization problem of (P2) for one specific frame, as $\hat{c}_1$, $\hat{c}_2$ and $\hat{\mathbf{d}}$ need to be updated frame-by-frame and they get more accurate with more frames received by the FC.

Subsequently, the microphone nodes quantize the next frame at the recently obtained bit rates. The FC then updates the SOS and the parameters required by (50) using the past segments together with the newly received measurements in a similar way. This procedure will continue until all the frames at the microphone end have been transmitted. This data-driven approach is summarized in Algorithm 1,[5] where we also include the DDRD-CW method. The proposed DDRD-CW method is obtained by replacing the CS-steps using the CW-steps, e.g., $\hat{\mathbf{d}}$ is the normalized eigenvector of the matrix pencil $(\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \mathbf{R_{nn}})$ corresponding to the maximum eigenvalue. Note that when the number of frames $l \ll L$, it is possible that (50) is infeasible due to insufficient segments for estimating the SOS. To circumvent the infeasibility, we can relax β in (50) using

$$\hat{\beta} = L\beta/l, \tag{51}$$

such that the constraint $\mathrm{Tr}\,(\mathbf{Z}) \le \frac{\hat{\beta}}{\alpha \hat{c}_1} - \hat{c}_2$ gradually becomes tighter when increasing the number of frames, resulting in an increase in the bit-rates per frame that are required for quantization. To this end, we can conclude that the complexity of the greedy approaches for each frame is the same as the model-driven methods, i.e., $\mathcal{O}(K^3)$, and the complexity for all the frames is of the order of $\mathcal{O}(LK^3)$.

VI. Experimental Results

In this section, we evaluate the RTF estimation performance of the proposed methods using synthetic data and natural speech data. Note that in simulations, the matrix $\mathbf{R_{nn}}$ is already estimated using sufficiently long noise-only segments.

A. Simulations on Synthetic Data

Fig. 1 shows the experimental setup, where $K = 20$ candidate microphones are placed in a 2D room with dimensions (3×3) m. The microphones are distributed uniformly on a circle with the origin at (1.5, 1.5) m and a radius of 0.5 m. The FC (black

[4]Note that for the proposed rate distribution methods, we only need to transmit the speech+noise segments, since the statistics of the acoustic noise is assumed known in this work. This is the assumption that we made in Section II-B in order to make the analysis on the CS and CW methods consistent.

[5]The current setup assumes the sources to be stationary in both time and space. For non-stationary sources, e.g., moving sources, Algorithm 1 should be modified as $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{P}\sum_{\iota=l-P}^{l} \hat{\mathbf{y}}_\iota \hat{\mathbf{y}}_\iota^H$, where P denotes the number of frames from the past that we want to include. If the sources are completely stationary, then $P = l - 1$.

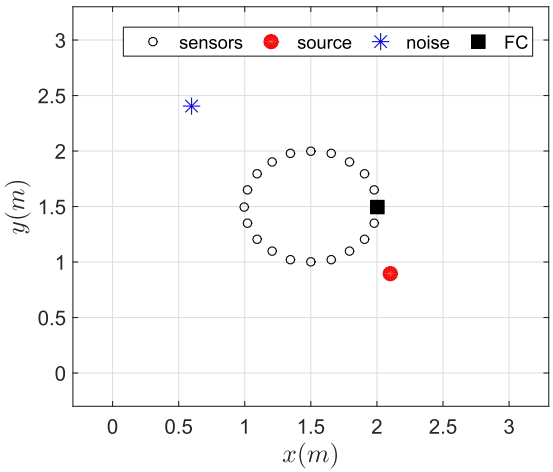


Fig. 1. An illustration of experimental setting with 20 microphones. The FC and the first microphone are placed at the same position.

Algorithm 1: DDRD-CS/CW Methods.

Require: $\mathbf{R_{uu}}$;
Initialize: $b_k = b_{\max}, \forall k$;
for $l = 1 : L$ **do**
  Transmit the lth noisy segment using b_k bits;
  $\hat{\mathbf{R}}_{\mathbf{qq}} = \frac{1}{12} \times \text{diag}([\frac{\mathcal{A}_1^2}{4^{b_1}}, \frac{\mathcal{A}_2^2}{4^{b_2}}, \ldots, \frac{\mathcal{A}_M^2}{4^{b_M}}])$;
  $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{l}\sum_{\iota=1}^{l} \hat{\mathbf{y}}_\iota \hat{\mathbf{y}}_\iota^H$;
  $\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R_{uu}} - \hat{\mathbf{R}}_{\mathbf{qq}}$;
  $\hat{\sigma}^2_{X_1} = |a_1|^2 \hat{\sigma}^2_S = \mathbf{e}_1^T \hat{\mathbf{R}}_{\mathbf{xx}} \mathbf{e}_1$;
  $\hat{\eta} = \frac{\hat{\sigma}^2_{X_1}}{\mathbf{e}_1^T(\hat{\mathbf{R}}_{\mathbf{qq}} + \mathbf{R_{uu}})\mathbf{e}_1}$;
  Case 1: DDRD-CS
    $\hat{\mathbf{d}}_{\text{CS}} = \hat{\sigma}_{X_1}^{-2} \hat{\mathbf{R}}_{\mathbf{xx}} \mathbf{e}_1$;
    $\hat{c}_1 = \frac{1+\frac{1}{\hat{\eta}}}{l\|\hat{\mathbf{d}}\|_2^2 \hat{\sigma}^2_{X_1}}$;
    $\hat{c}_2 = \text{Tr}((\mathbf{I} - \hat{\mathbf{d}}_{\text{CS}}\mathbf{e}_1^T)\mathbf{R_{uu}}(\mathbf{I} - \hat{\mathbf{d}}_{\text{CS}}\mathbf{e}_1^T)^H)$;
    update $\mathbf{b}_{\text{CS}}$ by solving (P2);
  Case 2: DDRD-CW
    $\hat{\mathbf{d}}_{\text{CW}} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}^{H/2}\hat{\boldsymbol{\psi}}}{\mathbf{e}_1^T \hat{\mathbf{R}}_{\mathbf{nn}}^{H/2}\hat{\boldsymbol{\psi}}}$;
    $\hat{c}_2 = \text{Tr}((\mathbf{I} - \hat{\mathbf{d}}_{\text{CW}}\mathbf{e}_1^T)\mathbf{R_{uu}}(\mathbf{I} - \hat{\mathbf{d}}_{\text{CW}}\mathbf{e}_1^T)^H)$;
    update $\mathbf{b}_{\text{CW}}$ and c_1' by solving (P3);
end for
return $\mathbf{b}_{\text{CS}}, \mathbf{b}_{\text{CW}}, \hat{\mathbf{d}}_{\text{CS}}, \hat{\mathbf{d}}_{\text{CW}}$

solid square) is assumed to be at the first microphone node, i.e., (2, 1.5) m. As the first node is considered to be the FC, it can be assumed that it always quantizes at the maximum rate, since it does not cost any transmission energy. The sensors are indexed in an anti-clockwise order. One target source (red solid circle) and one interfering source (blue star) are positioned at $(2.1,\ 0.9)$ m and (0.6, 2.4) m, respectively. We assume that the positions of all sources and microphones do not change. In this section, the simulations are performed directly in the STFT domain at a single frequency bin using a synthetic non-stationary Gaussian source signal and synthetic ATFs. The target source is modelled as $S(\omega, l) \sim \mathcal{CN}(0, \sigma_S^2(l))$ (i.e., the real and imaginary parts of $S(\omega, l)$ are both zero-mean Gaussian distributed with variance $\sigma_S^2(l)$). The non-stationarity is realized by varying the variance as $\sigma_S^2(l) \sim 0.5e^{0.5}$ (which is a scaled exponential random variable with an average of one, i.e., $\sigma_S^2 = 1$), such that the resulting average variance of the target source is one. The interference consists of a stationary coherent source and spatially-white sensor noise. We employ the SNR to measure the ratio between the variances of the target source and the sensor noise. Signal-to-interferer ratio (SIR) is used to measure the ratio between the variances of the target source and the interfering sources. The ATFs of the sources are modelled as a summation of a direct-path component and reflection components modelled as a complex Gaussian random variable[6]. The ratio between the power of the direct-path component and the reflections power is denoted as direct-to-reverberation ratio (DRR). The simulation parameters are set as follows: $b_{\max} = 16$ bits per sample, SNR $= 20$ dB, SIR $= 0$ dB, DRR $= 30$ dB and the number of frames $L = 8000$. The channel noise PSD is set to be $V_k = 1, \forall k$. Note that the level of SNR or SIR is averaged over time, since the variance of the target source is time-variant. We set β in (P1) to the estimation error of the classical CS method when each sensor quantizes at the maximum bit rate. The presented results are averaged over 100 Monte-Carlo trials. In order to focus on the rate-distributed RTF estimation problem, we assume that the internal clocks of the sensors are synchronized.

1) Evaluation of MDRD-CS/CW Methods: To study the performance of the rate distribution, we compare the proposed MDRD-CS/CW methods to the CS/CW methods using a uniform rate allocation (referred to as uni.CS and uni.CW, respectively). For instance, given the rate distribution b_k obtained by the MDRD-CS method, the uni.CS method distributes $\text{round}(\sum_{k=1}^{K} b_k/K)$ bits to each sensor and estimates the RTF using the classic CS method. Similarly, the uni.CW method is based on the rate distribution that is obtained by the MDRD-CW method. In addition, we also compare uni.PowerCS/CW methods, which distribute the total transmission powers that are consumed by the MDRD-CS/CW methods uniformly to all the sensors, respectively. As such, the uni.PowerCS (or uni.PowerCW) method uses the same amount of transmission energy as the proposed MDRD-CS (or MDRD-CW) approach, but most likely with different bit-rate distributions. Fig. 2 shows the RTF estimation error and transmission cost parameterized by α. Clearly, the better the accuracy, the more transmission cost is required. Hence, the proposed methods can trade-off the performance and energy usage by controlling the parameter α. From the simulations it follows that the proposed MDRD-CS/CW methods always satisfy the performance requirement. Moreover, their transmission costs are always much lower compared to the full-rate quantization (i.e., when $\alpha = 1$) or uniform

[6]The direct path is characterized by the gain and delay values. The gain can be viewed as the reciprocal of the distance from the source to the sensors, and the delay (in number of samples) is caused by the propagation of the source. Using the power of the direct-path component and the DRR parameter, we can calculate the power (or variance) of the reflection components. Then, the reflection components can be generated as zero-mean complex Gaussian random variables.

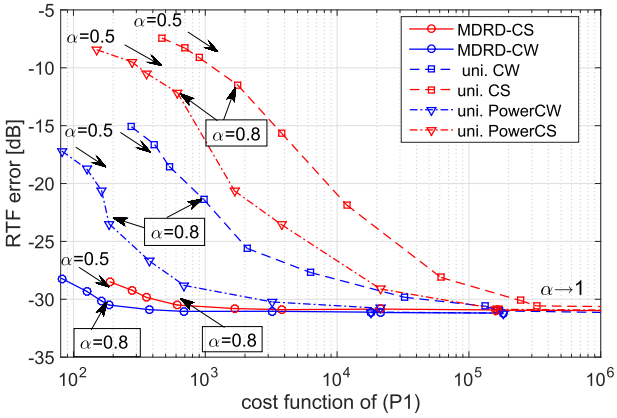


Fig. 2. RTF error and transmission cost of the model-based methods in terms of α. The cost function in x-axes means the total transmission power per frame. The "total" refers to the summation of transmission costs over microphones and "per frame" indicates the average over L frames.

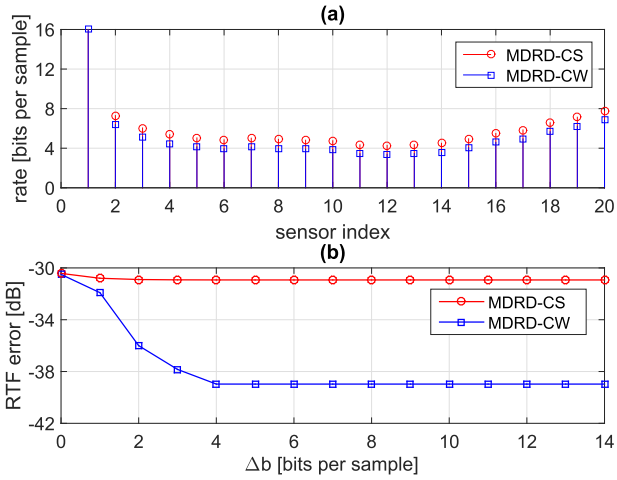


Fig. 3. (a) An example for rate distribution when $\alpha = 0.8$. (b) RTF accuracy in terms of rate increment.

rate allocation. Given the same RTF performance requirement, the MDRD-CW method consumes much less transmission energy than the MDRD-CS method. In other words, given the same power budget, the CW method always performs better than the CS method.

Fig. 3(a) shows the rate distributions obtained by the proposed MDRD-CS/CW from Fig. 2 at $\alpha = 0.8$. Clearly, to fulfil a desired RTF estimation performance $\epsilon_{\mathrm{CS/CW}} \leq \frac{\beta}{\alpha}$, we do not need full-rate quantization for all the sensors, as the optimal rate distributions are far below the maximum rate $b_{\max}$ per sensor. Given the same performance requirement, the MDRD-CW method needs less bit rates than the MDRD-CS method. Sensor one is allocated the maximum number of bits, as this is the FC and no additional transmission energy is required. Further, we see that in order to save transmission energy, the sensors that are closer to the FC are allocated with a higher rate. In Fig. 3(b), we show an example on how the RTF accuracy changes by further increasing the rate, starting from the optimal distributions given in Fig. 3(a). The resulting RTF accuracy is plotted as

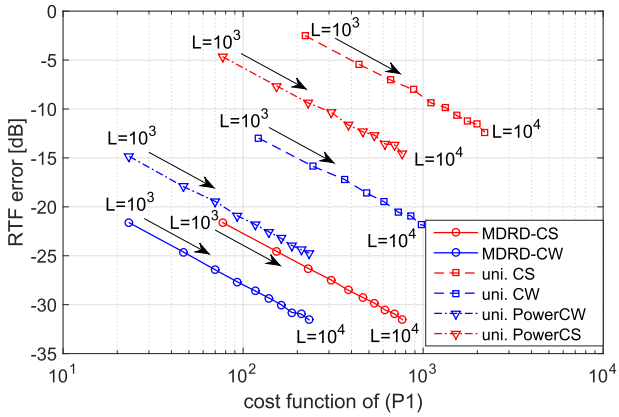


Fig. 4. RTF error and transmission cost of model-driven methods in terms of the number of available segments for $\alpha = 0.8$. The cost function in x-axes means the total transmission power per frame.

a function of the rate increment Δb. For $\Delta b = 0$, we use the optimal rate distribution given in Fig. 3(a). Then, for $\Delta b > 0$, we increase each $b_k, \forall k$ by Δb bits per sample. The resulting rate is upper-bounded by $b_{\max}$, i.e., the bit rates are increased to $b_k = \min\left(b_{\max}, b_k + \Delta b\right), \forall k$. Obviously, by increasing the bit-rate, we do not gain significantly in the RTF accuracy of the MDRD-CS method, which reveals that many bits are redundant and it is unnecessary to use full-rate quantization. Notably, the performance gain (e.g., 8 dB) in the MDRD-CW method is caused by the fact that β is set as the best performance of the classic CS method.

Fig. 4 compares the RTF accuracy and the energy usage parameterized by the number of segments L for $\alpha = 0.8$. Clearly, the more segments for estimating the correlation matrices, the more accurately the CS/CW-based estimators perform and the more transmission costs required. To achieve the same RTF estimation performance, the proposed methods consume much less transmission cost.

For further studying other influence factors on the proposed model-driven rate distribution approaches, we place the FC in Fig. 1 at the center of the room, such that all the microphone nodes have the same distance from the FC. The locations of the target source and the noise source are fixed, that is, only the SNRs across microphones vary from each other. Fig. 5 shows an example of the resulting rate distributions for such a scenario. We can clearly see that the SNR does affect the rate distributions, as roughly the sensor having a lower SNR (e.g., sensor 18 which is closest to the interfering source) is allocated with a higher rate. This reveals that the more noisy the microphone measurements are, the more bits are required for quantization. Comparing the ranges of the distributed rates between Fig. 5 and Fig. 3(a), it can be concluded that the distance between a sensor and the FC is more relevant than the SNR for the proposed rate optimization problems.

2) Evaluation of DDRD-CS/CW Methods: Fig. 6 compares the proposed DDRD-CS/CW methods to the model-driven versions, $\mathrm{uni.CS/CW}$ and $\mathrm{uni.PowerCS/CW}$. For each segment, the $\mathrm{uni.CS/CW}$ methods use uniform rate allocation, and

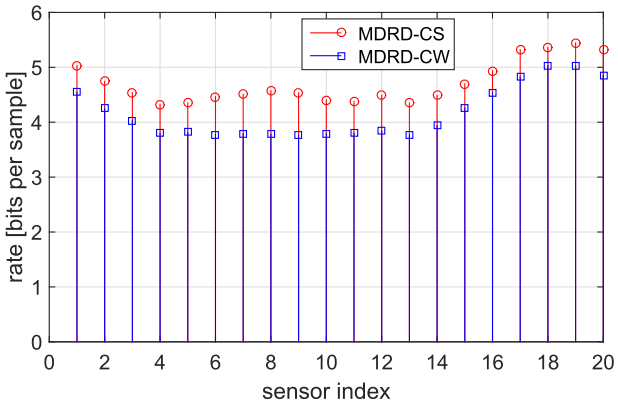


Fig. 5. Rate distributions of the proposed model-driven methods for the scenario where the FC is located at the center of the room and $\alpha = 0.8$.

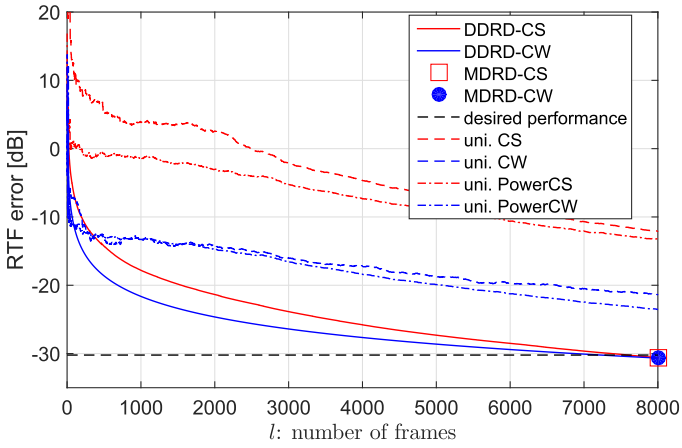


Fig. 6. RTF accuracy of the data-driven methods for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 8000$.

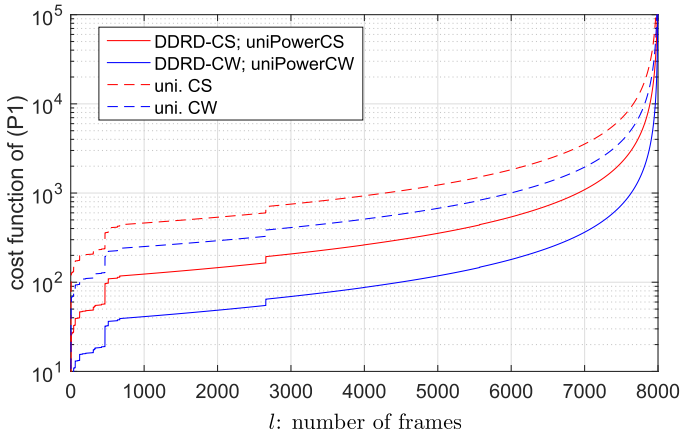


Fig. 7. Transmission cost of the data-driven methods per frame for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 8000$. The y-axes means the total transmission power per frame.

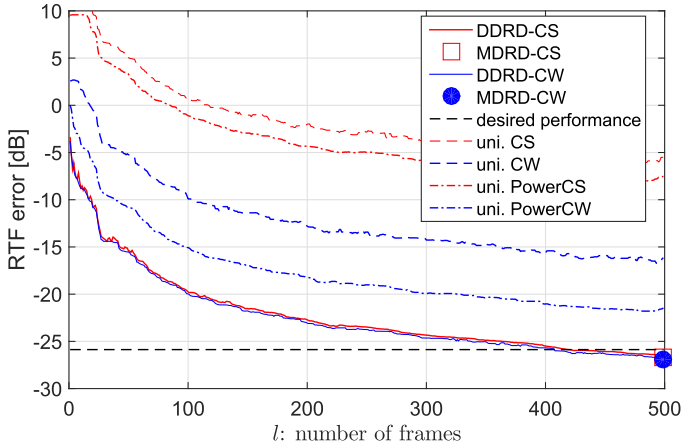


Fig. 8. RTF estimation performance of the proposed methods using the real speech recordings for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 500$.

uni.PowerCS/CW use uniform power allocation as before. Clearly, by increasing the number of available segments, the DDRD-CS method and the DDRD-CW method converge to the MDRD-CS method and the MDRD-CW method in terms of performance, respectively. The proposed DDRD-CW method converges faster. Note that the final rate distributions of the MDRD-CS (or MDRD-CW) method and the DDRD-CS (or DDRD-CW) method are not necessary to be the same. Fig. 7 shows the transmission cost per frame of the data-driven methods as a function of the number available frames. The cost of the DDRD-CS/CW methods gradually increases, which is caused by the relaxation $\hat{\beta} = L\beta/l$ for overcoming the infeasibility of (50) when $l \ll L$. Since the constraint $\mathrm{Tr}\,(\mathbf{Z}) \leq \frac{\hat{\beta}}{\alpha \hat{c}_1} - \hat{c}_2$ gradually gets tighter by increasing the number of frames, more and more bits are needed to fulfill the performance requirement. More importantly, the DDRD-CS/CW methods use much less transmission energy than the uni.CS/CW methods.

B. Simulations on Natural Speech Data

In this section, we will show the performance of the proposed methods using natural speech data in a simulated WASN. The experimental setup is same as Fig. 1. The single target source is a speech signal originating from the TIMIT database [39]. The coherent interfering source is a stationary Gaussian speech shaped noise signal. The microphone self noise is modeled as uncorrelated noise at an SNR of 50 dB. All signals are sampled at 16 kHz. We use a square-root Hann window of 100 ms for framing with 50% overlap. The real RTFs are generated using [40] with reverberation time $T_{60} = 200$ ms.

At first, we show the RTF estimation performance of the proposed methods in Fig. 8 for $\alpha = 0.8$. This is a similar comparison as in Fig. 6, but now using real speech signals. The total number of segments is $L = 500$. We can see that similar to the synthetic data case in Fig. 6, the DDRD-CS and DDRD-CW methods converge to MDRD-CS and MDRD-CW in the sense of RTF accuracy, respectively. Both methods satisfy the performance requirement. Similarly, the transmission cost per frame is shown in Fig. 9.

Secondly, we validate the application of the proposed methods in multiple reverberation conditions. The performance is

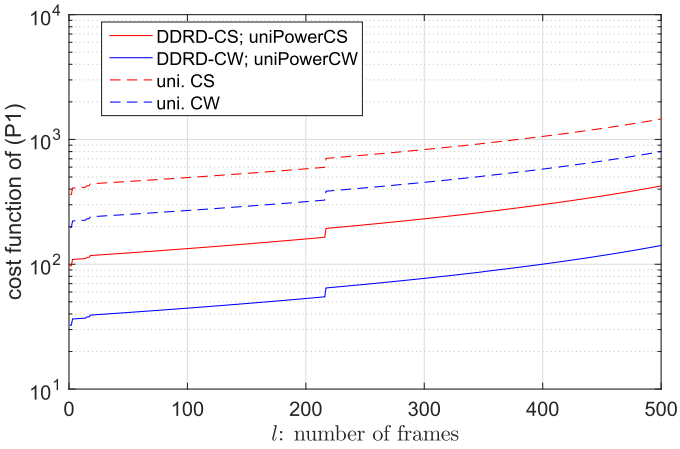


Fig. 9. Transmission cost per frame of the proposed methods using the real speech recordings for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 500$. The cost function in y-axes means the total transmission power per frame.

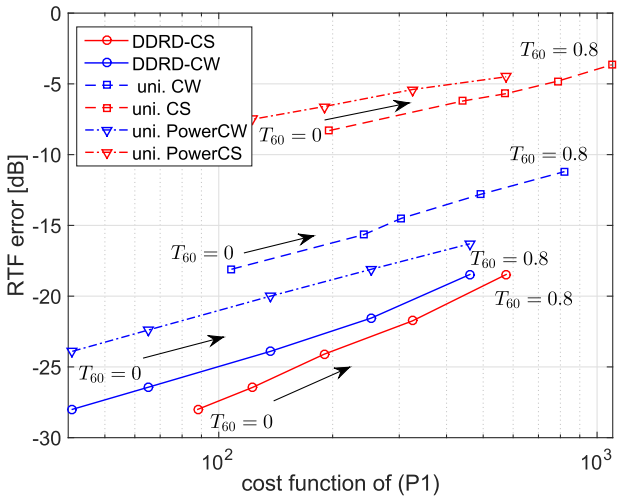


Fig. 10. RTF estimation accuracy and transmission cost of the proposed methods for multiple reverberation conditions with $\alpha = 0.8$. The cost function in x-axes means the average transmission power per frame.

examined for different values of T_{60}, selected from $\{0, 200, 400, 600, 800\}$ ms. The RTF estimation accuracy and the average transmission power per frame of the proposed methods and the reference methods are shown in Fig. 10 for $\alpha = 0.8$. Note that in reverberant environments, the early and late reverberations of the source signal might fall into different frames, since the frame length is fixed. When estimating the noisy correlation matrix and updating the RTF estimate frame-by-frame, the late reverberation of the interfering source will thus be regarded as another source of noise. Increasing the level of reverberation will lead to a lower long-term SIR. As Fig. 5 shows that the sensors with a lower SNR should be allocated with a higher rate, the proposed methods need to distribute more bits to the sensors, i.e., more transmission power, in a more reverberant environment. Also, that is why with an increase in the reverberation time, both the RTF estimation error and the transmission power increase in Fig. 10.

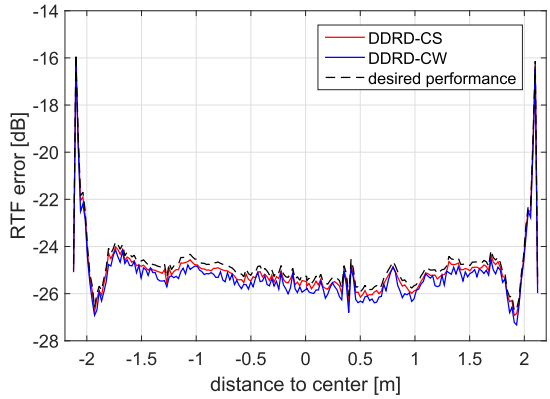


Fig. 11. RTF error of the proposed methods in terms of the distance from the target source to the center of the room, i.e., (1.5, 1.5) m, for $\alpha = 0.8$.

Finally, since the RTF performance is also affected by the source location (e.g., see Eqs. (22, 28)), we further evaluate the RTF performance for different positions of the target source. To do so, we randomly place the target source on the diagonal of the room, i.e., on the line from the bottom-left corner to the top-right corner. The RTF estimation performance in terms of the distance from the target source to the center of the sensor array is shown in Fig. 11. The proposed CS/CW-based methods obtain a similar performance variation in terms of the source location. Clearly, the proposed approaches achieve a better RTF estimation performance when the sources are located in the near-field, since the SNR is higher in this case.

VII. Concluding Remarks

In this work, we investigated the RTF estimation problem using the CS/CW methods under low bit-rate. Taking quantization noise into account, we showed that the estimation errors of both methods are influenced by the SNR, the number of available frames and the bit rate. Motivated by this, we formulated to minimize the energy usage for data transmission between sensors and the FC by constraining the RTF estimation performance, such that the optimal rate distribution can be found for the sensors to quantize their measurements. The problem was first solved by semi-definite programming, which was called MDRD-CS/CW. Since the proposed model-based methods are not practical (they depend on the true RTF), we further proposed two corresponding greedy approaches (i.e., DDRD-CS/CW). We can conclude that

- Both the model-based methods and the greedy methods satisfy the performance requirement on the RTF estimation, more importantly, with a significant saving of transmission cost compared to the full-rate quantization or uniform rate allocation;
- The performance of the greedy method converges to that of the model-based method with increasing the number of available frames;
- Given the same performance bound, the proposed CW-based methods need less bit rates, resulting in less energy consumption compared to the CS-based methods;

- The resulting rate distributions are affected by the distance, the SNR, etc. In general, the sensors that are closer to the FC are allocated with a higher rate because they are cheaper in data transmission, and the sensors that have a lower SNR should be allocated with a higher rate.

The benefits of the proposed approaches can be concluded as

- The considered methods can provide an effective strategy for saving the energy consumption over WASNs through distributing the quantization rates.
- The proposed methods can remove the redundant bits contained in the raw microphone measurements and be applied in noisy/reverberant environments.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their helpful remarks and constructive suggestions that helped to improve the presentation of this paper.

REFERENCES

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

[2] 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.

[3] F. de la Hucha Arce, M. Moonen, M. Verhelst, and A. Bertrand, "Adaptive quantization for multichannel Wiener filter-based speech enhancement in wireless acoustic sensor networks," *Wireless Commun. Mobile Comput.*, vol. 2017, 2017, Art. no. 3173196.

[4] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[6] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1, Springer Science & Business Media, 2008.

[7] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

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

[9] S. Shah and B. Beferull-Lozano, "Adaptive quantization for multihop progressive estimation in wireless sensor networks," in *Proc. EURASIP Europ. Signal Process. Conf.*, 2013, pp. 1–5.

[10] Y. Huang and Y. Hua, "Energy planning for progressive estimation in multihop sensor networks," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 4052–4065, Oct. 2009.

[11] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[12] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[13] 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.

[14] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Int. Workshop Hands-Free Speech Commun.*, 2017, pp. 11–15.

[15] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

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

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

[18] 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 Europ. Signal Process. Conf.*, 2018, pp. 2513–2517.

[19] J. Zhang, R. Heusdens, and R. C. Hendriks, "Rate-distributed spatial filtering based noise reduction in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2015–2026, Nov. 2018.

[20] 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.

[21] W. Pearlman, "Polar quantization of a complex gaussian random variable," *IEEE Trans. Commun.*, vol. 27, no. 6, pp. 892–899, Jun. 1979.

[22] J. Amini, R. C. Hendriks, R. Heusdens, M. Guo, and J. Jensen, "On the impact of quantization on binaural MVDR beamforming," in *Proc. 12th ITG Symp. Speech Commun*, 2016, pp. 1–5.

[23] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[24] A. Sripad and D. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 5, pp. 442–448, Oct. 1977.

[25] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1220–1244, Nov. 1990.

[26] N. R. Goodman, "Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction)," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 152–177, 1963.

[27] P. Stoica and T. Söderström, "Eigenelement statistics of sample covariance matrix in the correlated data case," *Digit. Signal Process.*, vol. 7, no. 2, pp. 136–143, 1997.

[28] 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.

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

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

[31] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 17–29, Mar. 2002.

[32] C. E. Shannon, "Communication in the presence of noise," *Proc. IEEE*, vol. 86, no. 2, pp. 447–457, 1998.

[33] Y. Huang and Y. Hua, "Multihop progressive decentralized estimation in wireless sensor networks," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1004–1007, Dec. 2007.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[35] K. B. Petersen *et al.*, *The Matrix Cookbook*. vol. 7. Tech. Univ. Denmark, 2008.

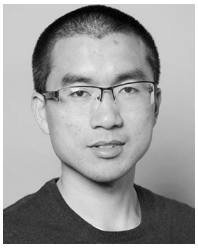
[36] M. Grant, S. Boyd, and Y. Ye, "CVX: MATLAB software for disciplined convex programming," 2008.

[37] J. F. Sturm, "Using SeDuMi 1.02: A MATLAB toolbox for optimization over symmetric cones," *Optim. Methods Softw.*, vol. 11, no. 1–4, pp. 625–653, 1999.

[38] 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.

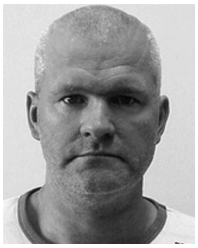
[39] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *Nat. Inst. Standards Technol.*, vol. 15, pp. 29–50, 1988.

[40] E. A. P. Habets, "Room impulse response generator," Tech. Univ. Eindhoven, Tech. Rep. 2.4, 2006, vol. 2.



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 with the Circuits and Systems Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands.

His current research interests include multimicrophone speech enhancement and sound source localization, binaural auditory, energy-aware 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), Sheffield, U.K.



Richard Heusdens received the M.Sc. and Ph.D. degrees from the Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively. Since 2002, he has been an Associate Professor with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. In Spring 1992, he joined the Digital Signal Processing Group, 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 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 with Aalborg University from 2014 to 2016. He is involved in research projects that cover subjects such as audio and acoustic signal processing, speech enhancement, and distributed signal processing.

Richard Christian Hendriks was born in Schiedam, The Netherlands. He received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively. He is currently an Associate Professor with the Circuits and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology.

His main research interest is on biomedical signal processing, and, audio and speech processing, including speech enhancement, speech intelligibility improvement and intelligibility modeling. In March 2010, he received the prestigious VENI grant for his proposal "Intelligibility Enhancement for Speech Communication Systems." He was the recipient of several best paper awards, including the IEEE Signal Processing Society Best Paper Award in 2016. He is currently an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the *EURASIP Journal on Advances in Signal Processing*.