Adaptive quantization for multi-channel Wiener filter-based speech enhancement in wireless acoustic sensor networks

Fernando de la Hucha Arce[1], Marc Moonen[1], Marian Verhelst[2] and Alexander Bertrand[1]

[1+2] KU Leuven, Dept. of Electrical Engineering (ESAT)
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[1] Stadius Center for Dynamical Systems, Signal Processing and Data Analytics
[2] MICAS Research Group
{fernando.delahuchaarce, marc.moonen, marian.verhelst, alexander.bertrand}@esat.kuleuven.be

***Abstract*—Speech enhancement in wireless acoustic sensor networks requires the exchange of audio signals. Since the wireless communication often dominates the nodes' energy budget, techniques for data exchange reduction are crucial. Adaptive quantization aims to optimize the bit depth of each exchanged signal according to its contribution to the speech enhancement performance. This enables the network to scale its energy and communication bandwidth requirements according to the current operating environment. The impact metric was previously proposed to predict the effect of quantization in linear minimum mean squared error (MMSE) estimation. We provide new insights on greedy adaptive quantization based on this impact metric. We achieve this by expanding the mathematical framework to include a new metric based on the gradient of the MMSE as a function of the quantization noise power. Using these tools we show how the MMSE gradient naturally leads to a greedy algorithm, and how the impact metric is a generalization of the gradient metric and a previously proposed metric. Besides, we validate the impact metric for adaptive quantization both in a simulated and in a real wireless acoustic sensor network deployed in a home environment, showing the energy savings achievable through greedy adaptive quantization.**

***Index Terms*—Wireless acoustic sensor networks, speech enhancement, adaptive quantization, energy efficiency, system reconfiguration, microphone arrays.**

I. Introduction

A wireless acoustic sensor network (WASN) is a collection of battery-powered sensor nodes where each node is equipped with a microphone or microphone array, a processing unit and a wireless communication module [1]. The nodes are distributed over an area of interest with the goal of performing a signal processing task such as noise reduction or acoustic localization. The main advantage of a WASN over a single stand-alone microphone array is its extended coverage, which is made possible by placing many microphones over the area of interest. This typically translates into a better performance, as microphone array algorithms benefit from enhanced spatial diversity. Furthermore, the deployment of a WASN often yields a higher probability to have microphones close to a sound source, which is advantageous since these microphones will record signals with high signal-to-noise ratio (SNR).

Nevertheless, WASNs pose several technical challenges that are not present in stand-alone microphone arrays, such as inter-node synchronization, delay management, communication bandwidth usage and energy efficiency. The latter, energy efficiency, is crucial to allow the network to perform its task for a reasonable period of time, since nodes are mostly powered by batteries and hence have a tight energy budget. A significant effort has been made to classify the different approaches to improve energy efficiency in wireless sensor networks (WSNs), as the optimal techniques depend on the intended WSN application. A comprehensive taxonomy of these approaches can be found in [2], and a more recent survey in [3] also considers the importance of the different techniques for specific classes of applications of WSNs.

In this paper we focus on a speech enhancement application for a WASN, where the goal is to estimate a desired speech signal while suppressing interfering sound sources and noise. In particular we focus on the multi-channel Wiener filter (MWF) [4]–[6], which is a multi-microphone noise reduction algorithm that produces a linear minimum mean squared error (MMSE) estimate of the desired speech component in the signal captured by one of the microphones. The algorithm does not rely on *a priori* knowledge of the microphone or sound source locations, which makes it suitable for a WASN since nodes are usually randomly deployed and may even be mobile (e.g. if a node is carried by a person, such as a mobile phone or a hearing aid).

A. Sensor subset selection

A substantial part of previous research on energy efficiency in WSNs has been focused on the sensor subset selection problem, which aims at using only the signals from those sensors (microphones, in the case of WASNs) that provide a significant contribution to the signal processing task at hand, while putting other sensors to sleep. This saves energy by avoiding the transmission of signals from sensors with low relevance, and allows the communication bandwidth resources to be allocated to the transmission of the signals from the most useful sensors. The sensor subset selection problem is combinatorial and thus difficult to solve in general. Due to its importance, it has been the focus of extensive research, and several techniques have been proposed to tackle it. For an overview of these techniques, the reader is directed to [7]. Recent work on sensor selection can be found in [8], [9] and references therein. In [8] the authors investigate the sensor selection problem for parameter estimation in a WSN where the sensor measurements follow a non-linear model, assuming that the measurements are independent random variables. The problem is formulated as a non-convex optimization problem and solved through convex relaxation. In [9] the authors develop a more general framework where they consider correlated measurement noise, and propose a greedy algorithm to solve the sensor selection problem based on the Fisher information matrix.

A different approach has been proposed to solve the sensor selection problem for signal estimation based on a greedy algorithm using the *utility metric* [10], [11]. The *utility* of a sensor signal is defined as the change in estimation performance when the sensor is removed from the estimation process and the corresponding estimator is subsequently reoptimized. The motivation is that the utility can be computed and tracked at a very low computational cost, which combined with the greedy approach allows to perform sensor subset selection swiftly and at low complexity, even though the solution will generally be suboptimal. Besides, the algorithm is fully data-driven and does not require any prior knowledge of the underlying measurement model, such as the microphone and source positions or the acoustic transfer functions, which indeed is generally not available in WASN applications. This priority on speed and low complexity is crucial for *adaptive* signal estimation, since the network needs to rapidly react to the changing signal conditions (e.g. sound sources moving in the case of a WASN) and has to avoid investing too much energy from the already limited budget of the nodes. This approach has been specifically applied to WASNs [12], and it has been extended to a distributed implementation of the MWF [13].

B. Adaptive quantization

While sensor subset selection does indeed help to save energy and communication bandwidth, it forces the nodes into a binary behaviour i.e., they either transmit their signals at full resolution or they are put to sleep. One technique to provide a more flexible scaling of the estimation performance and the energy consumption of the network is *adaptive quantization*, where each sensor signal is assigned a variable bit depth to encode its signal samples according to its contribution to the estimation performance. By using this technique, nodes are able to spend more or less energy on data transmission according to the estimation performance required. From the point of view of information theory, this problem can be tackled using source coding techniques. A comprehensive overview of source coding for WASNs can be found in [14], [15], where the focus is directed towards theoretical results based on rate-distortion theory.

In [16], a pragmatic approach is taken, in which a generalized version of the utility metric referred to as the *impact metric* is introduced to predict the MMSE increase in the estimation due to the quantization noise. This allows to model the effect of the quantization noise resulting from changing the bit depth of each sensor signal's samples on the estimation performance. The impact metric can be used by a heuristic algorithm to gradually decrease the bit depth in each sensor signal until a target MMSE (or corresponding SNR) is met.

C. Contributions and outline of the paper

The goal of this paper is twofold. Our first goal is to provide some new insights on greedy adaptive quantization based on the impact metric from [16]. To this end, we expand the mathematical framework for adaptive quantization in linear MMSE estimation and we apply it in a WASN with a centralized processing architecture. We consider the MMSE as a function of the quantization noise power in each sensor signal, and based on this we define a new metric for adaptive quantization based on the gradient of the MMSE. We demonstrate how this MMSE gradient naturally gives rise to a greedy algorithm. We then show how the impact metric is in fact a generalization of this gradient metric, which then also motivates the use of a greedy algorithm using the impact metric. Besides, we explain how the utility metric for sensor subset selection [10], [11] can be viewed as another limit case of the impact metric. Finally, we discuss the theoretical advantages and disadvantages of each metric, and propose a correction to improve the gradient metric.

The second goal of the paper is to validate the impact metric for adaptive quantization in a speech enhancement task in a simulated as well as in a real life WASN in a home environment. We compare the behaviour of the four metrics and show the superiority of the impact and the corrected gradient metrics over the gradient and utility metrics due to their inherent adaptation to the significance of each quantization bit. To conclude, we provide an estimation of the savings in transmission energy achievable through the use of the greedy adaptive quantization algorithm based on the aforementioned metrics.

The paper is structured as follows. In Section II, we formulate the problem statement and signal model, we briefly review the multi-channel Wiener filter for speech enhancement, and we introduce the quantization error model that is used throughout the paper. In Section III we model the effect of quantization noise in linear MMSE estimation, and show how adaptive quantization can be performed based on four metrics derived from this model (utility, impact, gradient and corrected gradient). In Section IV we show experimental results of adaptive quantization for speech enhancement performed on real recordings from a WASN. Finally, we present the conclusions in Section V.

II. Problem Statement

We consider a WASN composed of several nodes, each having one or more microphones, with K microphones in total. The signal samples of the k-th microphone signal are encoded, upon acquisition by the analog-to-digital converter, with a certain bit depth dictated by the hardware in use. We

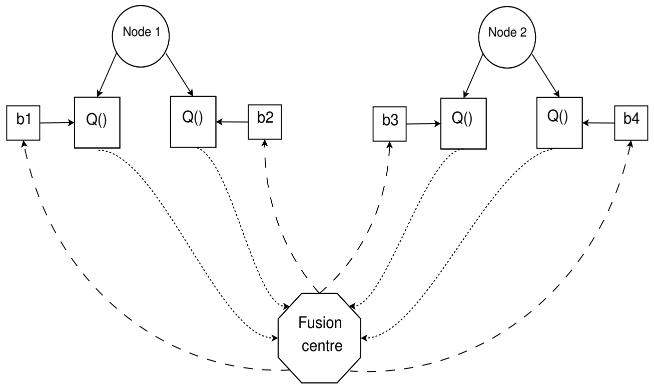

Figure 1. Example of a small WASN with adaptive quantization.

consider a centralized scheme for the network, where each node transmits its microphone signals to a fusion centre, which could be one of the nodes in the WASN or an external node with access to more computational power or energy resources. The fusion centre's task is to obtain an estimate of the desired speech component present in one of the microphone signals, which will be referred to as the reference microphone signal[1]. This speech enhancement task is solved in the fusion centre through the use of a multi-channel Wiener filter [4]–[6], which produces a linear MMSE estimate of the desired speech signal component in the reference microphone signal. We will give a brief review of the MWF in Section II-B.

Our main focus will be the problem of reducing the bit depth of each individual microphone signal in the WASN according to its contribution to the speech enhancement performance. The bit depth reduction leads to a reduction in the required communication bandwidth and in the node's required energy budget for wireless transmission, but it will also have an impact on the speech enhancement performance. Besides, the contribution of each node to the enhancement performance is subject to changes in the acoustic scenario, so we will focus on strategies with low computational complexity that allow the fusion centre to perform a quick decision on the desired bit depth assignment for each individual microphone. This enables each node at run-time to scale down the energy spent in wireless transmission according to the current operating environment.

An illustration of the problem is given in Figure 1, where a small network with two nodes and a fusion centre is depicted. The nodes quantize the signals of each individual microphone k with the corresponding bit depth b_k before transmission. The fusion centre performs the speech enhancement task using the transmitted quantized microphone signals (dotted lines) and takes a decision on the optimal bit depth for each communicated microphone signal (dashed lines).

In the remaining part of this section we introduce formally the signal model for the WASN, we briefly review the multi-channel Wiener filter for speech enhancement and we explain the quantization error model we will use throughout the rest of the paper.

A. Signal model

We denote the set of microphones by $\mathcal{K} = \{1, \dots, K\}$. The signal y_k captured by the k-th microphone can be described in the short-time Fourier transform domain (STFT) as

$$y_k(t,\omega) = x_k(t,\omega) + v_k(t,\omega), \quad k \in \mathcal{K}, \tag{1}$$

where t is the frame index, ω represents frequency, $x_k(t,\omega)$ is the desired speech signal component and $v_k(t,\omega)$ is the undesired noise signal component. We assume that $x_k(t,\omega)$ and $v_k(t,\omega)$ are uncorrelated. We note here that $v_k(t,\omega)$ contains all undesired sound signals, which may include speech from undesired speakers besides acoustic noise. For the sake of simplicity, we will omit the indices t and ω throughout the rest of the paper, keeping in mind that all operations take place in the STFT domain unless explicitly stated otherwise.

The fusion centre stacks all signals in the $K \times 1$ vector

$$\mathbf{y} = [y_1, y_2, \dots, y_K]^T . \tag{2}$$

The vectors $\mathbf{x}$ and $\mathbf{v}$ are defined in a similar manner, so the relationship $\mathbf{y} = \mathbf{x} + \mathbf{v}$ is satisfied.

B. Multi-channel Wiener filter

In speech enhancement, the goal is to obtain an estimate of the speech component x_{ref} present in the microphone signal y_{ref} selected as the reference. We will focus on the multi-channel Wiener filter to perform the speech enhancement task, and we will provide a brief summary in this section. For more information the reader is directed to [4]–[6].

The multi-channel Wiener filter is the linear estimator $\hat{\mathbf{w}}$ that minimizes the mean squared error (MSE)

$$J(\mathbf{w}) = E\left\{|x_{\text{ref}} - \mathbf{w}^H \mathbf{y}|^2\right\}, \tag{3}$$

where $E\{\cdot\}$ is the expectation operator and the superscript H denotes conjugate transpose. When the microphone signal correlation matrix $\mathbf{R}_{yy} = E\left\{\mathbf{y}\mathbf{y}^H\right\}$ is full rank[2], the solution to the minimization problem is given by

$$\hat{\mathbf{w}} = \mathbf{R}_{yy}^{-1}\, \mathbf{r}_{yx_{\text{ref}}}, \tag{4}$$

where $\mathbf{r}_{yx_{\text{ref}}} = E\{\mathbf{y}x^*_{\text{ref}}\}$ and the superscript $*$ denotes complex conjugation. Since we assume that $\mathbf{x}$ and $\mathbf{v}$ are uncorrelated, $\mathbf{r}_{yx_{\text{ref}}}$ is given by $\mathbf{r}_{yx_{\text{ref}}} = \mathbf{R}_{xx}\, \mathbf{c}_{\text{ref}}$, where $\mathbf{R}_{xx} = E\{\mathbf{x}\mathbf{x}^H\}$ is the desired speech correlation matrix and $\mathbf{c}_{\text{ref}}$ is the $K \times 1$ vector $\mathbf{c}_1 = [0, 0, \dots, 1, \dots, 0]^T$ where the entry corresponding to the reference microphone signal is equal to one.

The matrix $\mathbf{R}_{yy}$ can be estimated by temporal averaging, for instance using a forgetting factor or a sliding window. Temporal averaging is not possible for $\mathbf{R}_{xx}$ since the desired speech signal components $\mathbf{x}$ are not observable. In practice, the noise correlation matrix $\mathbf{R}_{vv} = E\{\mathbf{v}\mathbf{v}^H\}$ can be estimated during periods when the desired speech source is not active, as indicated by a voice activity detection (VAD) module. Since we assume that $\mathbf{x}$ and $\mathbf{v}$ are uncorrelated, it is then possible to

[1] The reference microphone does not necessarily belong to the fusion centre, the microphone of any node can be selected to be the reference.

[2] In practice, this assumption is usually satisfied because of the presence of a noise signal component in each microphone signal that is independent of other microphone signals, such as thermal noise. If this is not the case, matrix pseudoinverses have to be used instead of matrix inverses.

use the relationship $\mathbf{R}_{xx} = \mathbf{R}_{yy} - \mathbf{R}_{vv}$ to obtain an estimate of $\mathbf{R}_{xx}$. However, this is prone to robustness issues, created by oversubstraction, leading to the estimated desired speech correlation matrix not being positive semi-definite. These issues arise often in high frequencies, where the desired speech component may have very low power. To improve robustness in low SNR and non-stationary conditions, an implementation based on the generalized eigenvalue decomposition (GEVD) can be employed [17], [18].

The minimum mean squared error (MMSE) can be obtained by plugging (4) into (3) to obtain

$$J(\hat{\mathbf{w}}) = P_{\text{ref}} - \mathbf{r}_{yx_{\text{ref}}}^H \mathbf{R}_{yy}^{-1} \mathbf{r}_{yx_{\text{ref}}} = P_{\text{ref}} - \mathbf{r}_{yx_{\text{ref}}}^H \hat{\mathbf{w}} \,, \tag{5}$$

where $P_{\text{ref}} = E\left\{|x_{\text{ref}}|^2\right\}$ is the power of the desired speech signal.

C. Quantization error model

We will consider uniform quantization of the time domain samples of each microphone signal $y_k(t)$, prior to the transformation to the STFT domain. In practice, this means that the nodes transmit their time domain samples and the STFT is performed in the fusion centre. We discuss the possibility of quantizing the STFT coefficients directly prior to transmission in Section III-D. This configuration would require each node to perform the STFT over its own microphone signals and transmit the frequency domain coefficients to the fusion centre.

The quantization of a real number $d \in [-D/2, D/2]$ with b bits can be expressed as

$$Q(d) = \Delta_b \left(\left\lfloor \frac{d}{\Delta_b} \right\rfloor + \frac{1}{2} \right) , \tag{6}$$

where

$$\Delta_b = \frac{D}{2^b} \,. \tag{7}$$

In practice, the parameter D is given by the dynamic range of the analog-to-digital converter of the corresponding microphone. The quantization error, or noise, is then defined as

$$e_b = Q(d) - d \,. \tag{8}$$

The mathematical properties of the quantization noise e_b have been the subject of extensive study [19]–[21], where it has been shown that the input signal and the quantization noise are uncorrelated under certain technical conditions on the characteristic function of the input signal. Under the same conditions, the mean squared error due to quantization is given by

$$E\{|e_b|^2\} = \frac{\Delta_b^2}{12} \,. \tag{9}$$

We consider that, for the k-th microphone signal, the time domain samples of y_k are quantized with b_k bits according to (6) before being transmitted to the fusion centre. The quantization error can be expressed as

$$e_k(n) = Q(y_k(n)) - y_k(n) \,, \tag{10}$$

where n indexes the samples of frame t. The fusion centre performs the STFT and collects the results for each frequency ω and frame t in the $K \times 1$ vector $\mathbf{y}_e$ given by

$$\mathbf{y}_e = \mathbf{y} + \mathbf{e} \,, \tag{11}$$

where $\mathbf{e} = [e_1, \ldots, e_K]^T$ is the $K \times 1$ vector whose k-th element is the quantization error corresponding to the k-th microphone signal at frequency ω. Note that all K signals have been included in the quantization process. However, if the fusion centre is also equipped with microphones (e.g., it is a node of the WASN), these signals do not need to be transmitted and hence have a fixed quantization. In this case, the microphone signals from the fusion centre are removed from the adaptive quantization process, but they are still included in the estimation process.

Using the statistical properties of the quantization error [19]–[21], we will assume that every element of $\mathbf{e}$ is uncorrelated with every element of $\mathbf{y}$. Again, under certain technical conditions, the power spectrum of the quantization noise is white, i.e. its power is evenly distributed across all frequencies [19]. Although these conditions are not always satisfied in practice, particularly for quantization with only a few bits, we will combine this property with (9) to approximate the quantization noise power at each frequency as

$$p_{e_k} = L \frac{\Delta_{b_k}^2}{12} \,, \tag{12}$$

where L is the length of the discrete Fourier transform (DFT) used to implement the STFT in practice. The factor L in (12) appears as a consequence of the application to $e_k(n)$ of Parseval's theorem for the non-unitary DFT, given by

$$\sum_{n=0}^{L-1} |e_k(n)|^2 = \frac{1}{L} \sum_{m=0}^{L-1} |e_k(\omega_m)|^2 \,, \tag{13}$$

where $e_k(\omega_m)$ is the L-point DFT corresponding to $e_k(n)$. The non-unitary definition of the DFT is given by

$$Z_m = \sum_{l=0}^{L-1} z_l \, e^{-im\frac{2\pi}{L} l} \,, \tag{14}$$

where z_l is the input sequence, i is the imaginary unit and Z_m is the resulting transformed sequence. If a factor of $\frac{1}{\sqrt{L}}$ is applied to the right-hand side of (14) the DFT becomes a unitary transformation and the factor L is no longer needed in (12). In the rest of the paper we assume that the non-unitary DFT is used to implement the STFT, keeping in mind that the unitary DFT can be employed simply by re-scaling (12).

III. Adaptive quantization for the multi-channel Wiener filter in a WASN

We now consider the effect of quantization noise on the estimation process described in the previous section. Our interest here is to study how changing the bit depth for the transmission of the microphone signal samples affects the operation of the MWF, in particular, how it affects the MMSE. The analysis of this effect will lead to a metric based on the gradient of the MMSE which, as we will show, naturally leads to a greedy adaptive quantization algorithm. We will then demonstrate how this gradient metric is a limit case of a recently proposed impact metric [16], which was already known to also generalize the utility metric proposed in [10], [11]. Besides, based on this reasoning, we propose a correction to improve the gradient metric for adaptive quantization. This analysis provides a motivation for applying a greedy algorithm based on any of these metrics, which allows to dynamically change, at any moment in time, the bit depth assigned to each microphone signal. In Section IV, we will demonstrate experimentally that the impact and the corrected gradient metrics outperform the gradient and utility metrics, due to their inherent adaptation to the difference in quantization levels corresponding to different bit depths.

A. *Effect of quantization on the minimum mean squared error*

The MWF $\hat{\mathbf{w}}_e$ based on the quantized microphone signal samples is obtained following (4) as

$$\hat{\mathbf{w}}_e = \mathbf{R}_{y_e y_e}^{-1} \mathbf{r}_{y_e x_{\text{ref}}} , \tag{15}$$

where $\mathbf{R}_{y_e y_e} = E\left\{\mathbf{y}_e \mathbf{y}_e^H\right\}$. Using (11) and the assumptions stated in section II-C, we express $\mathbf{R}_{y_e y_e}$ as

$$\mathbf{R}_{y_e y_e} = E\left\{(\mathbf{y}+\mathbf{e})(\mathbf{y}+\mathbf{e})^H\right\} = \mathbf{R}_{yy} + \mathbf{R}_{ee} . \tag{16}$$

The quantization error correlation matrix $\mathbf{R}_{ee}$ is diagonal[3] , with the k-th element of the diagonal being $E\{|e_k|^2\} = p_{e_k}$, where p_{e_k} is defined in (12). As $\mathbf{e}$ is assumed to be uncorrelated with x_{ref}, the cross correlation remains unchanged, i.e.

$$\mathbf{r}_{y_e x_{\text{ref}}} = \mathbf{r}_{y x_{\text{ref}}} . \tag{17}$$

As explained in section II-B, $\mathbf{r}_{y_e x_{\text{ref}}}$ can be computed as

$$\left(\mathbf{R}_{y_e y_e} - \mathbf{R}_{v_e v_e}\right)\mathbf{c}_{\text{ref}} = \left(\mathbf{R}_{yy} + \mathbf{R}_{ee} - \mathbf{R}_{vv} - \mathbf{R}_{ee}\right)\mathbf{c}_{\text{ref}} \tag{18}$$

$$= \left(\mathbf{R}_{yy} - \mathbf{R}_{vv}\right)\mathbf{c}_{\text{ref}} , \tag{19}$$

where $\mathbf{R}_{v_e v_e} = E\{\mathbf{v}_e \mathbf{v}_e^H\}$, which indeed confirms (17). Similarly to (5), we can now find the MMSE corresponding to $\hat{\mathbf{w}}_e$, given by

$$J_e(\hat{\mathbf{w}}_e) = P_{\text{ref}} - \mathbf{r}_{y x_{\text{ref}}}^H \left(\mathbf{R}_{yy} + \mathbf{R}_{ee}\right)^{-1} \mathbf{r}_{y x_{\text{ref}}} . \tag{20}$$

We highlight that $J_e(\hat{\mathbf{w}}_e)$ is a function of the quantization error powers p_{e_k}, which we can make explicit by rewriting the function as

$$J_e(\mathbf{p}_e) = P_{\text{ref}} - \mathbf{r}_{y x_{\text{ref}}}^H \left(\mathbf{R}_{yy} + \mathbf{R}_{ee}\right)^{-1} \mathbf{r}_{y x_{\text{ref}}} \tag{21}$$

$$= P_{\text{ref}} - \mathbf{r}_{y x_{\text{ref}}}^H \left(\mathbf{R}_{yy} + \mathit{diag}(\mathbf{p}_e)\right)^{-1} \mathbf{r}_{y x_{\text{ref}}} , \tag{22}$$

where $\mathbf{p}_e = [p_{e_1}, \ldots, p_{e_K}]^T$ is the vector of quantization error powers, and where $\mathit{diag}(\cdot)$ is the operator that generates a diagonal matrix with diagonal elements equal to the entries of the vector in its argument. Equation (22) is important because it defines the cost function that we will use as the basis for adaptive quantization, since it is the *minimum* mean squared error that can be obtained with a linear estimator (i.e., the MWF) after adding quantization noise to each microphone signal. We emphasize that (22) gives the MMSE when the MWF is first re-optimized using the quantized microphone signals, i.e., based on (15), and not the mean squared error resulting from applying the original (optimized for the non-quantized signals) MWF $\hat{\mathbf{w}}$ to the quantized microphone signals.

B. *Gradient-based approach to adaptive quantization*

The goal of adaptive quantization is to allocate a bit depth to each sensor which is smaller than (or at most equal to) an initial maximum bit depth. Since each bit depth reduction also reduces the speech enhancement performance, the goal becomes to find the bit depth allocation which uses the minimum total number of bits $\sum_k b_k$ given a maximum tolerated MMSE. Equivalently, the problem could be stated as finding the lowest MMSE with a given total number of bits $\sum_k b_k$.

The gradient of the function $J_e(\mathbf{p}_e)$ gives the direction of maximal increase of the MMSE for a given $\mathbf{p}_e$, i.e. for a given bit depth allocation. To further reduce the total number of bits beyond the bit depth allocation corresponding to $\mathbf{p}_e$, $\mathbf{p}_e$ has to be changed to $\mathbf{p}_e + \Delta\mathbf{p}_e$, where $\Delta\mathbf{p}_e$ is constrained to have non-negative entries. The corresponding MMSE increase for an infinitesimally small $\Delta\mathbf{p}_e$ is then given by the inner product of $\Delta\mathbf{p}_e$ and the gradient of $J_e(\mathbf{p}_e)$. In order to compute this gradient, we will use the intermediate step

$$\frac{\partial J_e(\mathbf{p}_e)}{\partial \mathbf{R}_{ee}} = \left(\mathbf{R}_{yy} + \mathbf{R}_{ee}\right)^{-1} \mathbf{r}_{y x_{\text{ref}}} \mathbf{r}_{y x_{\text{ref}}}^H \left(\mathbf{R}_{yy} + \mathbf{R}_{ee}\right)^{-1} , \tag{23}$$

which follows from applying the identity [22]

$$\frac{\partial \mathbf{a}^H \mathbf{X}^{-1} \mathbf{a}}{\partial \mathbf{X}} = -\mathbf{X}^{-H} \mathbf{a}\mathbf{a}^H \mathbf{X}^{-H}$$

together with the fact that $(\mathbf{R}_{yy} + \mathbf{R}_{ee})^{-1}$ is a Hermitian matrix. Equation (23) can be simplified using (15)-(17) to obtain

$$\frac{\partial J_e(\mathbf{p}_e)}{\partial \mathbf{R}_{ee}} = \hat{\mathbf{w}}_e \hat{\mathbf{w}}_e^H . \tag{24}$$

Since the matrix $\mathbf{R}_{ee}$ is diagonal, we can now find the gradient $\mathbf{g}_e$ as the diagonal of the right-hand side term in (24), i.e.

$$\mathbf{g}_e = \nabla J_e(\mathbf{p}_e) = |\hat{\mathbf{w}}_e|^2 , \tag{25}$$

where the operator $|\cdot|$ is applied element-wise to its argument.

To minimize the MMSE increase for an infinitesimally small $\Delta\mathbf{p}_e$, the inner product $\Delta\mathbf{p}_e^T \mathbf{g}_e$ has to be minimized. However, every component of $\mathbf{g}_e$ is non-negative and the vector $\Delta\mathbf{p}_e$ is also constrained to have non-negative components. Hence the best choice for $\Delta\mathbf{p}_e$ is a vector whose components are all zero except the one corresponding to the minimum element of $\mathbf{g}_e$.

This result shows that, when adding a small amount of quantization noise, it should be added to a single microphone signal instead of dividing it over multiple microphone signals. This naturally leads to a greedy algorithm, where at each step the gradient $\mathbf{g}_e$ is computed from the MWF $\hat{\mathbf{w}}_e$ using (25), after which its minimum element is identified and the bit depth for the corresponding microphone signal is reduced by q bits. Note that the above reasoning has assumed the vector $\mathbf{p}_e$ to be a continuous variable, i.e. each element of the vector can take any real value. However, the bit depth is a discrete variable and it determines the quantization noise power added to a signal. Hence, the smallest possible quantization power that can be added to a signal corresponds to reducing its bit depth by 1 bit, which is the recommended value for q in order to avoid taking a too large step. This also avoids reducing the bit depth of one signal too quickly, which may be a poor choice compared

[3] One could intuitively expect quantization to reduce the cross-correlation between the microphone signals. In the Appendix we consider a quantization model that includes this reduction and show that its effect on the MWF is equivalent to the one presented in Section III-A.

to distributing the q bit reduction over several signals. After removing a bit from the microphone signal with the smallest entry in the gradient vector, the MWF is re-optimized to the new bit depth assignment, and the gradient is recomputed. This process is continued until the MMSE exceeds a pre-defined threshold.

C. Alternative metrics for adaptive quantization

In this section, we will show how the gradient metric used in the previous section is a limit case of the impact metric, which has been used in [16] for adaptive quantization. This provides an intuitive explanation of why the greedy approach, which follows naturally from the gradient metric, also works well when using this impact metric, as will be demonstrated in section IV.

The impact metric from [16] was initially proposed as a generalization of the utility metric defined in [10], [11]. The *utility* of the k-th microphone signal y_k is defined as the increase in MMSE when y_k is removed from the estimation [10]. The mathematical expression of this definition is given by

$$u_k = J_{-k}(\hat{\mathbf{w}}_{-k}) - J(\hat{\mathbf{w}}) \,, \tag{26}$$

where $\hat{\mathbf{w}}_{-k}$ is the re-optimized MWF obtained with all signals except y_k. Assuming the MWF $\hat{\mathbf{w}}$ is known, then the utility of y_k is shown [10] to be equal to

$$u_k = \frac{1}{\alpha_k} |w_k|^2 \,, \tag{27}$$

where α_k is the k-th element in the diagonal of $\mathbf{R}_{yy}^{-1}$, and w_k is the k-th element of $\hat{\mathbf{w}}$.

The *impact* of the noise e_k is defined as the increase in MMSE when the uncorrelated noise signal e_k is added to y_k, while other microphone signals remain unchanged [16]. In mathematical terms the definition can be expressed as

$$I_{e_k} = J_e(\hat{\mathbf{w}}_e) - J(\hat{\mathbf{w}}) \,, \tag{28}$$

where $\hat{\mathbf{w}}_e$ is the re-optimized MWF for $\mathbf{y}_e$, as in (15), with $\mathbf{e} = [0, \ldots, e_k, \ldots, 0]^T$. In [16] the impact is shown to be equal to

$$I_{e_k} = \frac{p_{e_k}}{1 + \alpha_k p_{e_k}} |w_k|^2 \tag{29}$$

where α_k is again the k-th element in the diagonal of $\mathbf{R}_{yy}^{-1}$, w_k is the k-th element of $\hat{\mathbf{w}}$, and where p_{e_k} represents the power of the noise added to y_k, given by (12) for the case of quantization noise.

To simplify further notation and the comparison between different metrics, we consider the gradient for the case $\mathbf{p}_e = \mathbf{0}$, where $\mathbf{0}$ is the zero vector, such that (25) is rephrased as $\mathbf{g} = |\hat{\mathbf{w}}|^2$, where[4] each element is given by

$$g_k = |w_k|^2 \,. \tag{30}$$

Despite the fact that the impact (29), utility (27) and gradient (30) metrics predict a change in the *minimum* mean squared error, which implicitly requires to re-optimize the MWF, all three metrics can be calculated from the *current* MWF coefficients at almost no additional computational cost compared to the computation of $\hat{\mathbf{w}}$ itself.

Algorithm 1 Greedy adaptive quantization for MWF in WASN

```
1:  Choose a metric $m_k$ from $I_k$, $g_k$, $g_{\text{warped},k}$ or $u_k$.
2:  Initialize $D_k \, \forall k \in \mathcal{K}$ to the dynamic range of each sensor.
3:  Initialize the bit depth assignment $b_k \, \forall k \in \mathcal{K}$ to the maximum bit depth allowed by the hardware.
4:  Initialize $p_{e_k} \forall k \in \mathcal{K}$ using equation (12).
5:  while MMSE$_{\text{current}}$ < MMSE$_{\text{threshold}}$ do
6:      Each signal $y_k$ is quantized in time domain with $b_k$ bits using (6).
7:      Receive $N_{\text{fr}}$ signal frames from $y_k \, \forall k \in \mathcal{K}$.
8:      Apply STFT to the received frames.
9:      Compute $\hat{\mathbf{w}}(\omega_m) \, \forall \omega_m$ based on the quantized microphone signals using equation (15).
10:     Update[5] $p_{e_k}$ using $b_k - 1$ and equation (12) $\forall k \in \mathcal{K}$.
11:     Compute the selected metric $m_k(\omega_m) \, \forall \omega_m$ according to equation (29), (30), (31) or (27) respectively.
12:     Combine $m_k(\omega_m)$ using equation (32).
13:     Find the index $k_{\text{min}}$ of the signal with minimal $m_k$.
14:     Reduce $b_{k_{\text{min}}}$ by 1 bit.
15:     If $b_{k_{\text{min}}}$ equals 0 after the reduction, remove the $k_{\text{min}}$-th signal for subsequent iterations.
16: end while
```

By comparing (29) with (27) and (30), we see that both the gradient g_k and the utility u_k are limit cases of the impact I_{e_k} when $p_{e_k} \to 0$ and $p_{e_k} \to \infty$ respectively. Although $p_{e_k} \to 0$ would obviously give an impact equal to zero, the relative differences between the impact metric for different k become equal to those of the gradient metric.

These two limit cases can be interpreted as follows. For the utility, the interpretation is that removing the microphone signal y_k from the estimation process is similar to adding an infinite amount of noise on y_k ($p_{e_k} \to \infty$), making it completely useless, which corresponds to a removal of that channel. For the gradient, the distinction between the gradient and the impact is that the gradient characterizes the best linear approximation of the function $J_e(\mathbf{p}_e)$, while the impact computes the actual MMSE increase produced by adding the error e_k with power p_{e_k}. Since the gradient approximation is only valid in an infinitesimally small neighbourhood, it is only able to accurately capture the influence of e_k on the MMSE for small values of p_{e_k}. Besides, note that the quantization noise power p_{e_k} increases exponentially with each bit reduced, so the gradient becomes less accurate as the microphone signals are quantized with lower resolution. On the other hand, the impact metric accounts directly for p_{e_k}, which makes it inherently adaptive to the significance of each bit considered for removal. For low significance bits, the impact is close to the gradient. However, as the significance of a bit increases, the impact behaves more like the utility. By contrast, the gradient assumes that the p_{e_k} corresponding to a bit removal is the same for

[4] The comparison is valid for any $\mathbf{p}_e$, we choose this case purely to simplify the notation.

[5] The update is done with $b_k - 1$ in order for the metric to predict what would happen if the bit depth of the k-th signal is reduced by 1 bit. However only one signal gets its b_k actually reduced in step 14.

all k, or in other words it assumes that the search space is isotropic, which only holds true when all microphone signals have the same bit depth. This can be adjusted by making p_{e_k} in (22) a linear function of the resolution corresponding to the least significant bit, e.g., $\beta_k \Delta_{b_k}$, and taking the derivative with respect to β_k. This would then provide a warped gradient vector

$$\mathbf{g}_{warped} = \mathbf{D} \cdot |\hat{\mathbf{w}}|^2 \,. \tag{31}$$

where $\mathbf{D} = \operatorname{diag}(\Delta_{b_1}, \ldots, \Delta_{b_K})$. Note that this warped gradient is again an asymptotic case of the impact measure, if p_{e_k} is substituted with $\beta_k \Delta_{b_k}$ in (29), and letting $\beta_k \to 0$.

D. Frequency domain considerations

To conclude, we must turn our attention to the fact that all of the above is valid at each frequency ω. This opens the possibility to assign a different bit depth to each frequency component of each microphone signal y_k.

In Section II-C we took the approach of performing quantization in the time domain. In order to select the signal from which a bit is to be removed, we need to choose a rule to combine each metric across all frequencies. We propose to perform a sum of the metrics across all frequencies. For instance, for the impact the combined metric would be given by

$$I_k = \sum_{m=0}^{L-1} I_{e_k}(\omega_m) \,. \tag{32}$$

For the utility, gradient and warped gradient the combined metric is defined in a similar way. It is noted that one could as well use a weighted sum in (32), e.g., based on speech intelligibility weights. We provide a summary of the greedy quantization algorithm based on any of the four metrics described so far in Algorithm 1.

However, strategies to allow the assignment of a different bit depth to each frequency component can be considered, as is commonly done in audio coding, to represent the most relevant frequency components with higher accuracy. Instead of assigning a different bit depth to every single frequency bin, frequency bins can also be grouped in a set of R frequency bands $\mathbf{\Omega} = \Omega_1 \cup \cdots \cup \Omega_R$, where $\mathbf{\Omega}$ comprises all frequency bins such that $|\mathbf{\Omega}| = L$. This means that every STFT coefficient of each microphone signal y_k at the frequency band Ω_r is quantized following (6) with $b_{k,r}$ bits. The real and imaginary parts of each STFT coefficient are quantized independently. The corresponding metric can be computed in a similar way to (32) as

$$I_{k,r} = \sum_{\omega_m \in \Omega_r} I_{e_k}(\omega_m) \,, \tag{33}$$

where $I_{k,r}$ is the impact corresponding to the k-th microphone signal in the r-th frequency band. For the utility, gradient and warped gradient the combined metric is again defined in a similar way.

This configuration opens up several strategies to decide which frequency band and microphone signal will have its bit depth reduced in each iteration of the algorithm. For our discussion we consider the strategy of removing, in each iteration, one bit in each frequency bin assigned to the frequency band $\Omega_{r_{\min}}$ of the microphone signal $y_{k_{\min}}$ with minimum $I_{k,r}$. This is the most conservative greedy strategy, which can be viewed as a limit case that will generally provide a better performance compared to greedier strategies where the bit depth is reduced in multiple channels and frequency bands simultaneously. It is noted that a more conservative greedy strategy comes with the cost of a larger number of required iterations to reach a pre-defined total number of bits. In Section IV-A and IV-B we show the performance of this particular strategy applied to a speech enhancement scenario.

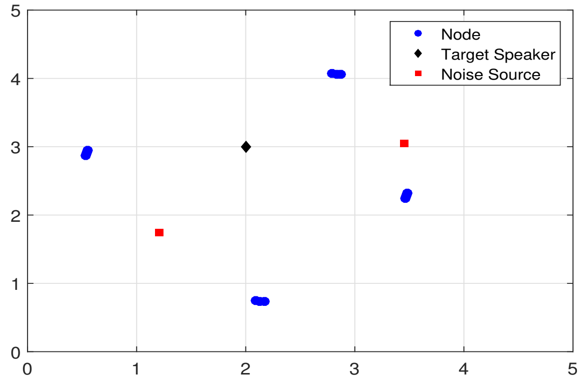

Figure 2. Acoustic scenario for the simulated room acoustic experiment.

Note that, in every iteration, the bit depth in $|\Omega_r|$ (out of L) frequency bins is reduced, which corresponds to a reduction of $|\Omega_r|/L$ bits per time domain sample. This is less than the full bit per sample reduction achieved through time domain quantization, which shows that the proposed strategy for frequency domain quantization is more conservative than the strategy for time domain quantization.

Besides, it is important to mention that frequency bands do not influence each other in the sense that the bit depth reduction in one band will not affect the decision in the rest of the bands. In the case of non-uniform bands, where each frequency band spans a different number of frequency bins, a trade-off with the transmission energy has to be considered, i.e. removing a bit from a wider frequency band will introduce more quantization noise but will result in less energy spent in transmission since the total number of bits will be lower.

IV. Experimental results

In this section we discuss the results obtained from several experiments to observe and characterize the performance of the greedy adaptive quantization algorithm based on the four metrics described in Section III. We will discuss experiments on two different audio datasets. In the first one the audio signals captured by the microphones are obtained by simulating the acoustics of a room with the image method [23]. In the second one, the audio signals were recorded using a wireless acoustic sensor network set up in a real home environment in a house in Mol, Belgium using nodes designed by researchers from the MICAS group of the Dept. of Electrical Engineering (ESAT) in KU Leuven. The details of each experiment will be discussed in Sections IV-A and IV-B. In all experiments the desired speaker audio consists of three sentences, spoken by a female speaker, from the TIMIT database [24]. The noise characteristics will be described in the section corresponding to each experiment. The sampling frequency is $f_s = 16$ kHz. The audio processing is implemented in batch mode, where the correlation matrices $\mathbf{R}_{yy}(\omega_m)$ and $\mathbf{R}_{vv}(\omega_m)$ are estimated using samples over the entire length of the microphone signals. An ideal VAD is used to exclude the influence of speech detection errors. The audio signals are divided in frames using a Hann window with 50% overlap, and the STFT is implemented using a discrete Fourier transform (DFT) of length $L = 512$. The multi-channel Wiener filter is computed based on a GEVD of $\mathbf{R}_{yy}(\omega_m)$ and $\mathbf{R}_{vv}(\omega_m)$ as in [17] since, as we mentioned in Section II-B, this method is superior to the subtraction-based implementation.

In order to assess the changes in noise reduction and speech distortion due to the bit depth reduction we will use two figures of merit, the speech intelligibility weighted signal-to-noise ratio (SI-SNR) [25] and the speech intelligibility weighted spectral distortion (SI-SD) [6]. They are based on the band importance function B_i, which expresses the importance for intelligibility of the i-th one-third octave band with center frequency $f_{c,i}$. The values for $f_{c,i}$ and B_i are defined in [26]. The definitions of the two figures of merit are given by

$$\mathrm{SNR_{SI}} = \sum_i B_i \, \mathrm{SNR}_i \tag{34}$$

$$\mathrm{SD_{SI}} = \sum_i B_i \, \mathrm{SD}_i \,. \tag{35}$$

The quantity SNR_i is the SNR (in dB) in the one-third octave band with centre frequency $f_{c,i}$. In order to account for quantization, the quantization noise in the input signals can be obtained by subtracting the clean input signal and its corresponding quantized version. The quantization error obtained is added to the noise component of each microphone, and they are filtered to obtain the noise component in the output signal, which is then used to compute the noise power at each one-third octave frequency band.

For the SI-SD, SD_i is the average spectral distortion in the one-third octave band with centre frequency $f_{c,i}$, given by

$$\mathrm{SD}_i = \int_{2^{-1/6} f_{c,i}}^{2^{1/6} f_{c,i}} \frac{|10\log_{10} G^s(f)|}{(2^{1/6} - 2^{-1/6}) f_{c,i}} df \,. \tag{36}$$

The function $G^s(f)$ is given by

$$G^s(f) = \frac{E\{X_{\text{out}}(f) X^*_{\text{out}}(f)\}}{E\{X_{\text{in}}(f) X^*_{\text{in}}(f)\}} \,, \tag{37}$$

where $X_{\text{out}}(f)$ is the speech component at the output of the MWF, and $X_{\text{in}}(f)$ is the frequency domain speech component at the reference microphone signal. A distortion value of 0 indicates undistorted speech, while larger values correspond to increased speech distortion. To account for quantization, $X_{\text{out}}(f)$ is computed by first quantizing the speech component at each microphone with the corresponding bit depth, and then applying the filter to the quantized speech components.

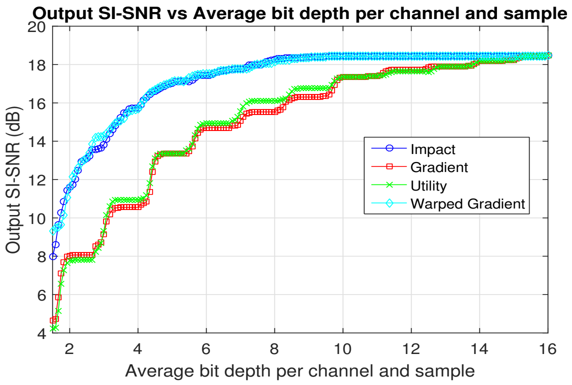

Figure 3. SI-SNR at each step of the greedy quantization algorithm using time domain quantization for the simulated room acoustic experiment.

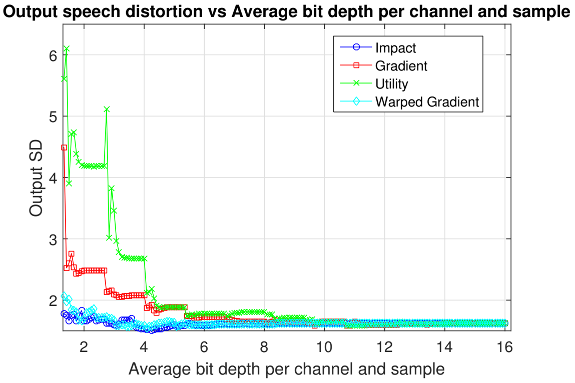

Figure 4. SI-SD at each step of the greedy quantization algorithm using time domain quantization for the simulated room acoustic experiment.

A. Simulated Room Acoustics

Our first experiment is a study of the behaviour of the greedy algorithm for adaptive quantization using simulated room acoustics. The scenario consists of a room of dimensions $5 \times 5 \times 3$ m, with a reverberation time of 0.2 s. In the room there are two babble noise sources [27] and one desired speech source. The WASN consists of four nodes, where each node is equipped with three omnidirectional microphones, such that the total number of microphone signals is $K = 12$. Independent white Gaussian noise was added to each microphone signal with a power of $2.5 \cdot 10^{-5}$, about 1% of the power of the babble noise impinging on the microphones. A 2D diagram of the acoustic scenario is depicted in Figure 2. All sources are located at a height of 1.8 m, while the nodes are placed 2 m high. The inter-microphone distance at each node is 4 cm and the sampling rate is 16 kHz. The maximum bit depth was set to 16 bits. The broadband input SNR for every microphone lies between 0 dB and 5 dB. The acoustics of the room are modeled using a room impulse response generator, which allows to simulate the impulse response between each source and each microphone using the image method [23]. The code is available online[6]. The total duration of the signals is 20 seconds.

[6] https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

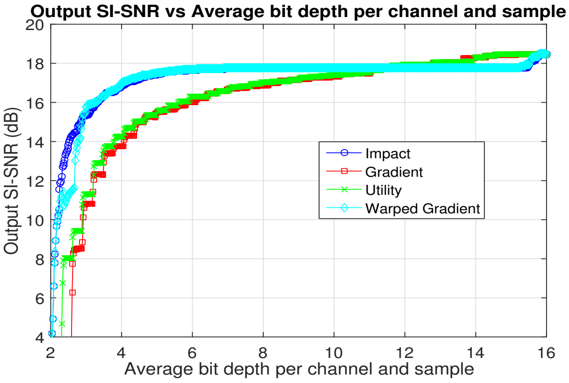

Figure 5. SI-SNR at each step of the greedy quantization algorithm with frequency domain quantization for the simulated room acoustic experiment.

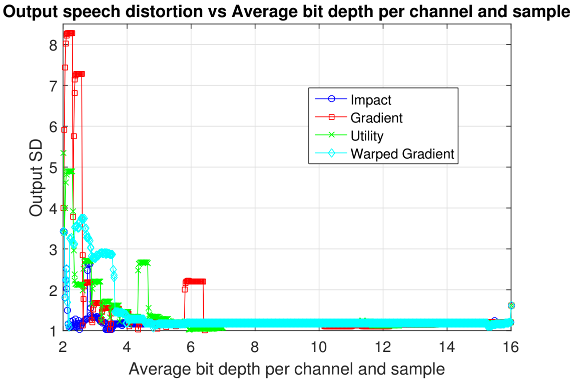

Figure 6. SI-SD at each step of the greedy quantization algorithm with frequency domain quantization for the simulated room acoustic experiment.

In Figure 3 and Figure 4 we can see the SI-SNR and SI-SD at each iteration of the greedy adaptive quantization algorithm presented in Algorithm 1 based on the four metrics discussed. In this experiment the quantization is performed in the time domain, as explained in Section II-C, such that each time domain sample of the microphone signal y_k is quantized using its allocated bit depth b_k. Note that both the SI-SNR and the SI-SD are plotted versus the average bit depth per sample and channel at each iteration, given by $\sum_k b_k/K$. In terms of SI-SNR, the impact metric performs better than both the utility and the gradient, as we expected due to its inherent adaptability to the significance of each bit for different bit depths. The same can be said about the warped gradient, which performs better than the uncorrected gradient and close to the impact due to the correction to account for the significance of each bit. In terms of distortion, there is no clear winner when the total number of bits is high. However, the impact and the warped gradient introduce the least distortion as the number of bits decreases.

We now turn our attention to quantization in the frequency domain, where each microphone signal y_k has a bit depth $b_{k,r}$ allocated to its frequency band Ω_r, as explained in Section III-D. The STFT coefficient at each frequency bin $\omega_m \in \Omega_r$ is quantized using $b_{k,r}$ bits. In each iteration, one frequency band at one microphone signal has its bit depth $b_{k_{\min},r_{\min}}$ reduced by one. The pair $(k_{\min}, r_{\min})$ is given by the channel and band with minimum impact (or corresponding metric). For this experiment we considered $R = 4$ uniform frequency bands, each spanning $\frac{L}{4}$ frequency bins. The bit allocation $b_{k,r}$ of any band can be reduced to a minimum of 2 bits. If all bands of a microphone signal y_k are assigned 2 bits, the signal is removed from the estimation process for subsequent iterations. In Figure 5 and Figure 6 we can again see the SI-SNR and SI-SD at each iteration of the greedy adaptive quantization algorithm. The two figures of merit are plotted versus the average bit depth per sample and channel $\sum_k \bar{b}_k/K$, where $\bar{b}_k = \frac{1}{R}\sum_r b_{k,r}$. We can observe again the impact and the warped gradient performing better in terms SI-SNR, which is consistent with our previous experiment. However, the decay in SI-SNR for the utility and the gradient is less pronounced, and the region where their performance is similar to the impact and the warped gradient is larger. In terms of speech distortion the results are also consistent with the previous experiment in the sense that there is no clear winner, although the impact seems to perform better as the number of bits decreases for this particular experiment.

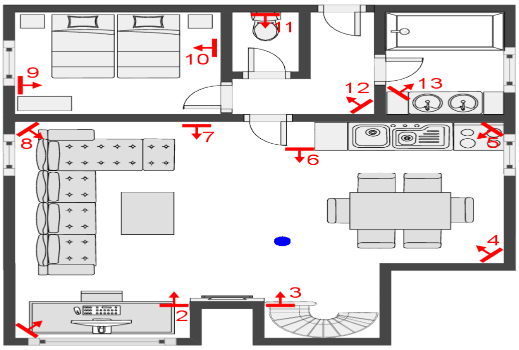

Figure 7. Schematic in 2D of the house used for the WASN recordings, with the desired speaker in blue and the WASN nodes in red.

B. Experiments on Real Recordings

In order to further compare the four metrics for greedy adaptive quantization, we turn our attention to an audio scenario where the signals are recorded using a real life wireless acoustic sensor network set up in a house in Mol, Belgium, consisting of 6 nodes with 4 microphones per node. A 2D schematic of the whole house can be seen in Figure 7, although only the living room was used for this experiment. The acoustic scenario consisted of one loudspeaker acting as the desired speaker (represented by the blue circle) and a kitchen fan (located in the top right corner of the living room in the 2D schematic) acting as the noise source. Only the nodes marked 1, 2, 3, 6, 7 and 8 were used for this experiment. The speech signal for the loudspeaker consisted of three sentences from the TIMIT [24] database, spoken by a female speaker. The total duration of the recording was 23 seconds.

The microphones employed were Sonion N8AC03 (analog), and the inter-microphone distance at each node was 5 cm. A picture of one node with the location of the microphones indicated is shown in Figure 8. The sampling frequency was

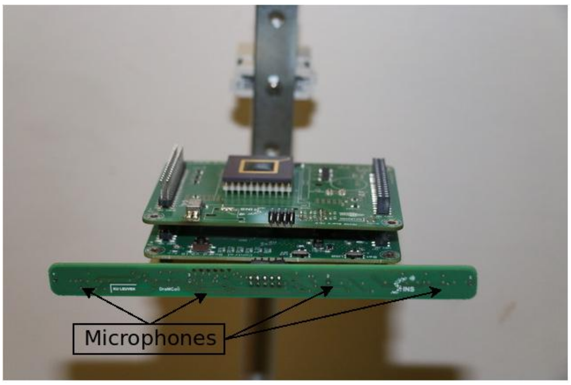

Figure 8. One node of the WASN used to make the recordings.

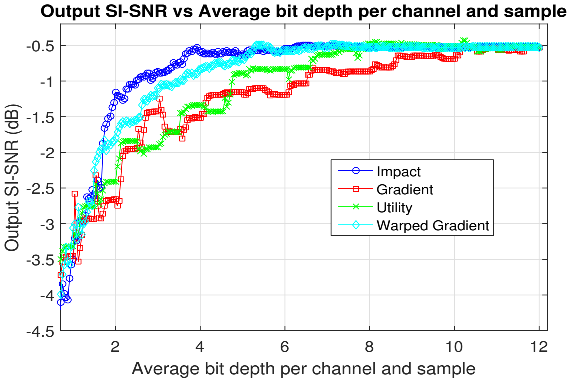

Figure 9. SI-SNR achieved at each step of the greedy quantization algorithm for the real recordings.

$f_s = 16$ kHz, and the analog-to-digital converter of every node was configured to use a bit depth of 12 bits for acquisition. The micro-controller unit in each node is the Wonder Gecko EFM32WG980 from Silicon Labs [28], which is used for sampling and sending data to a Raspberry Pi 3 [29] via USB. The Raspberry Pi at each node is used to upload the audio samples to a USB drive. A picture of one node can be seen in Figure 8. The nodes were synchronized once every second using a pulse that was sent through coaxial cable and triggered by a GPS/DCF receiver. The recorded audio signals were stored and subsequently processed using the MATLAB software as described at the beginning of Section IV. We implemented the processing off-line to focus on the characterization of the performance of the bit depth reduction algorithm and the comparison of the different metrics using real audio data.

In Figure 9 we can see the results of the SI-SNR of the output signal estimated from the MWF using the recorded audio signals. In this case, quantization was performed in the time domain. The SI-SNR of the input microphone signals lied between -16 and -7 dB. The noise power for the SI-SNR calculation was computed using the non-speech segments. The greedy adaptive quantization algorithm was stopped when the total number of bits used was 20 bits. It can be observed that the impact metric again outperforms the gradient and the utility metrics, and provides a smoother way of downscaling the WASN performance, in agreement with the results from section IV-A. Besides, the warped gradient performs very close to the impact due to the correction to account for the significance of each bit, again in agreement with the results from section IV-A. We would like to note that the impact and the warped gradient outperforming the gradient and the utility, as we can observe both in Figure 3 and in Figure 9, agrees with the theoretical discussion of Section III-C, where we describe the limitations of each metric. The four metrics achieve a similar performance only in the high resolution regime, where the samples from every signal are encoded with a high bit depth and the bits removed have low significance.

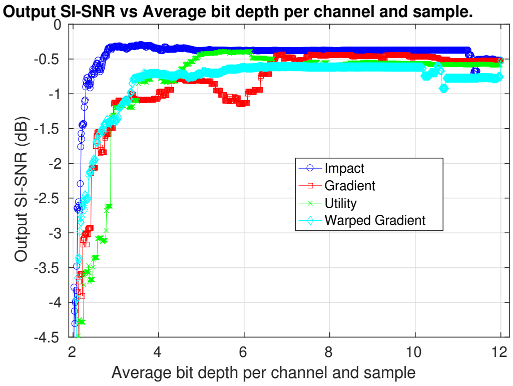

Figure 10. SI-SNR at each step of the greedy quantization algorithm using frequency domain quantization for the real recordings.

Finally, we turn again our attention to quantization in the frequency domain, as explained in Section III-D. We followed the same strategy as in the previous section, where we consider $R = 4$ uniform frequency bands, each spanning $\frac{L}{4}$ frequency bins. In Figure 10 we can see the behaviour of the SI-SNR for this experiment, where a slower decay compared to the evolution in Figure 9 is observed. Although the impact outperforms the rest of the metrics, the four metrics diverge less from each other compared to the time domain quantization as seen in Figure 9. We note that for this experiment the warped gradient performs worse than the utility and the gradient.

C. Analysis of Energy Consumption

To conclude, we focus on estimating the energy savings that can be achieved in communication by reducing the bit depth assignment of the microphone signals using the greedy adaptive quantization algorithm. This estimation is based on the power consumption of the WASN hardware nodes we used to record the audio signals. We employ a simplified model for the average energy $\mathcal{E}_{RF}$ required to transmit L_{RF} bits from one node to the fusion centre given by

$$\mathcal{E}_{RF} = \frac{P_{RF}}{d_{RF}} L_{RF} \,, \tag{38}$$

where d_{RF} is the data rate in bits per second and P_{RF} is the average power consumed by the radio module in active status. We note that (38) provides only an approximation of the required transmission energy since it ignores some factors such as the retransmission of lost packets. However, a detailed model for the transmission energy is outside the scope of this paper. The interested reader can find more advanced methods in [30].

We will first discuss the case where quantization is performed in the time domain, that is, the bit depth assigned to the microphone signal y_k is equal for every frequency.

The number of bits L_{RF} needed for the transmission of an audio frame of length L samples from microphone signal y_k can be calculated as follows

$$L_{RF,k} = b_k L + n_{\text{pkt},k} L_{\text{overhead}} \,, \tag{39}$$

where b_k is the bit depth assigned to the microphone signal y_k, L_{overhead} is the length in bits of the headers containing protocol information and $n_{\text{pkt},k}$ is the number of packets necessary to fit L samples from y_k according to the network protocol rules.

The radio module of the nodes we used to acquire our audio recordings consists of an IEEE 802.15.4 standard compliant radio from Atmel (AT86RF233) in combination with an ARM Cortex M4 microcontroller. In active mode, the power consumption is $P_{RF} = 41.8$ mW at $d_{RF} = 1$ Mbps. The packet in the IEEE 802.15.4 standard consists of 127 payload bytes and 6 header bytes [31]. The 127 bytes include 2 CRC bytes and 125 bytes of actual data plus headers originating from higher layers (such as, e.g., IPv6 for the network layer and UDP for the transport layer). We will assume that 25 bytes correspond to headers from higher layers. This leads to each packet carrying 33 bytes of overhead and a maximum of 100 bytes of data corresponding to audio samples. The number of packets necessary to transmit L audio samples encoded with bit depth b_k is then given by

$$n_{\text{pkt},k} = \left\lceil \frac{b_k L}{8 \cdot 100} \right\rceil \,. \tag{40}$$

As we have explained in Algorithm 1, when a signal is assigned 0 bits, it gets removed from the estimation process for subsequent iterations. We are interested in calculating the total energy spent in the transmission of L samples per microphone signal included in the estimation process, which is given by

$$\mathcal{E}_{T,\text{frame}} = \sum_{k \in \mathcal{K}_a} \mathcal{E}_{RF,k} \,, \tag{41}$$

where $\mathcal{E}_{RF,k}$ is computed using (38) and (39) and $\mathcal{K}_a$ is the subset of $\mathcal{K}$ containing the indexes of the microphone signals included in the estimation process. However, we also have to consider the messages the fusion centre needs to send to the nodes every iteration to inform them of which microphone signal y_k will have its bit depth b_k reduced. These messages are limited in size since only the index of the signal whose bit depth needs to be reduced has to be communicated to the nodes. The length of one fusion centre packet in bits is given by

$$L_{\text{FC}} = L_{\text{overhead}} + 8 \,, \tag{42}$$

where we assume that the message contains one byte of payload. The energy spent in the transmission of these packages is related to the speed of refreshment of the bit depth allocation algorithm, that is, the rate at which the network performs the iterations required by the algorithm. We will denote this rate by $r_{\text{refr}} \in (0, 1]$, which is given by the inverse of the number of frames the network waits between two consecutive iterations of the bit depth allocation algorithm. A value of 1 means that we change the bit depth allocation every frame, and a value of 0.5 every two frames. Following (38) the average energy per frame required to transmit the fusion centre packet is given by

$$\mathcal{E}_{FC} = \frac{P_{RF}}{d_{RF}} L_{\text{FC}}\, r_{\text{refr}} \,. \tag{43}$$

We can then modify (41) to include $\mathcal{E}_{FC}$ so that the total energy spent by the network in the duration of one frame is

$$\mathcal{E}_T = \sum_{k \in \mathcal{K}_a} \mathcal{E}_{RF,k} + (N_{\text{nodes}} + 1)\mathcal{E}_{FC} \,, \tag{44}$$

where N_{nodes} is the number of nodes in the network, and which is included to account for the energy spent by the nodes in the reception of the packet. Note that it is implicitly assumed here that the energy spent in the reception of a packet is on the same order of magnitude of the energy spent for its transmission. This assumption is valid in short distances [32], which can be expected in the context of a WASN. A quick calculation of the ratio between $\mathcal{E}_{FC}$ and $\mathcal{E}_{RF,k}$ for $L = 512$, $b_k = 8$, $L_{\text{overhead}} = 264$ bits (corresponding to 33 bytes) and $r_{\text{refr}} = 1$ yields roughly 5%. While this is only an approximate energy model and other concerns related to communications may arise due to the speed of refreshment, such as the use of bandwidth or the need for retransmissions, from the point of view of energy we can conclude that even for fast rates, i.e. one iteration per frame, the reduction of transmission energy is not jeopardized by the refreshment rate in most situations. In practice, deciding on a value for the refreshment rate r_{refr} depends on the dynamics of the acoustic scenario, e.g. in a scenario with moving sources it may be interesting to have a high rate to be able to track the sources, while in a static scenario a lower rate can be sufficient.

We turn our attention now to quantization with a different bit depth in each of the R frequency bands. This leads to each microphone signal y_k having a bit depth $b_{k,r}$ assigned for each frequency band Ω_r. The number of bits L_{RF} needed for the transmission of $\frac{L}{2}$ complex STFT coefficients from microphone signal y_k can be calculated following (39) as

$$L_{RF,k} = \sum_{r=1}^{R} b_{k,r} L_r + n_{\text{pkt},k} L_{\text{overhead}} = \tag{45}$$

$$L_{RF,k} = \bar{b}_k L + n_{\text{pkt},k} L_{\text{overhead}} \,, \tag{46}$$

where L_r is the number of frequency bins included in band Ω_r, and $\bar{b}_k$ is the average number of bits assigned to microphone signal y_k, which is given by

$$\bar{b}_k = \frac{\sum_{r=1}^{R} b_{k,r} L_r}{L} \tag{47}$$

The number of packets necessary is now given by

$$n_{\text{pkt},k} = \left\lceil \frac{\bar{b}_k L}{8 \cdot 100} \right\rceil \,. \tag{48}$$

We note that, since each payload byte allows the fusion centre 256 combinations of channel and frequency band indexes, a packet of very similar length to the one we considered in (42) can be used in this case to let the fusion centre inform

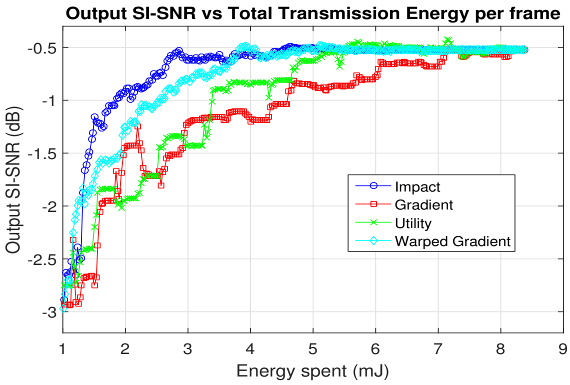

Figure 11. SI-SNR vs Total transmission energy spent in the duration of one frame in the case of time domain quantization.

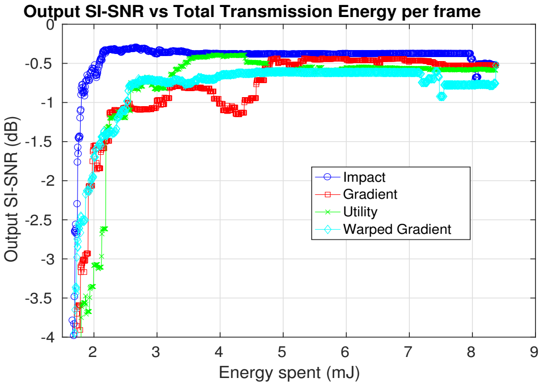

Figure 12. SI-SNR vs Total transmission energy spent in the duration of one frame in the case of frequency domain quantization.

the nodes of where to remove bits. While the quantization in several frequency bands allows for extra granularity, the energy analysis shown above applies in a straightforward manner by considering the average number of bits $\bar{b}_k$ in place of b_k.

Finally, in Figure 11 the resulting SI-SNR (the same as in Figure 9) is plotted versus the total energy spent in transmission calculated from (44). Similarly, in Figure 12 we show the resulting SI-SNR (the same as in Figure 10) plotted versus the total energy spent in transmission calculated following the energy analysis for frequency domain quantization shown above. These graphs illustrate the estimated transmission energy savings which can be achieved through the use of the greedy adaptive quantization algorithm. For time domain quantization, from Figure 11 it can be observed that the total transmission energy can be reduced roughly by half without a meaningful loss in performance, and cut by four for a small loss of 1 dB. For frequency domain quantization the savings are potentially higher since the total transmission energy can be reduced roughly to one third without meaningful loss in performance.

V. Conclusions

We have provided a better understanding of adaptive quantization for speech enhancement in wireless acoustic sensor networks based on the previously proposed impact metric. We have done so by extending the mathematical framework of adaptive quantization in linear MMSE estimation, where we have proposed a metric based on the gradient of the MMSE and demonstrated how this metric naturally leads to a greedy approach. Moreover, we have shown that the impact metric is a generalization of the gradient metric, where the gradient is a limit case of the impact. We also propose a correction to improve the gradient metric by considering the significance of each quantization bit for different bit depths. Besides, the impact also generalizes a utility metric previously proposed for sensor subset selection. Through the use of a simulated and a real life environment we have assessed the superiority of the impact and the corrected gradient metrics over the gradient and the utility metrics due to their adaptability to the significance of each quantization bit. Besides, we have provided an estimation of the possible energy savings achievable through the use of the greedy adaptive quantization algorithm based on any of the studied metrics. In future work, an extension of this approach to a distributed speech enhancement algorithm will be explored, hence going beyond the centralized setting targeted in this work. Another important research direction will be the incorporation of psychoacoustic characteristics of human hearing to the bit depth allocation algorithm in order to improve the allocation in different frequency bands.

VI. Acknowledgments

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of Research Fund KU Leuven BOF/STG-14-005 and CoE PFV/10/002 (OPTEC), and FWO nr. G.0931.14 'Design of distributed signal processing algorithms and scalable hardware platforms for energy-vs-performance adaptive wireless acoustic sensor networks' and IWT SBO project SINS: Sound interfacing through the swarm. The scientific responsibility is assumed by the authors. The authors declare that there is no conflict of interest regarding the publication of this paper.

The authors would like to thank Steven Lauwereins for the design of the WASN nodes and Gert Dekkers for setting up the WASN and helping in the measurement process.

Appendix

The model for the effect of quantization noise on the MWF developed in Section III-A relies on the quantization noise being uncorrelated with the input microphone signals $\mathbf{y}$ and with the desired speech signal components $\mathbf{x}$ to establish equations (16) and (17). However, one might intuitively expect the quantization of microphone signal y_k to reduce the cross-correlation with the other microphone signals $y_m \in \mathcal{K} \setminus \{k\}$. This would lead to a decrease in the off-diagonal elements in $\mathbf{R}_{y_e y_e}$ compared to the off-diagonal elements in $\mathbf{R}_{yy}$.

This can be considered by using an alternative model for quantization such that (11) is substituted by

$$\mathbf{y}_q = \mathbf{A}\left(\mathbf{y} + \mathbf{e}\right), \tag{49}$$

where $\mathbf{A}$ is the $K \times K$ diagonal matrix

$$\mathbf{A} = \text{diag}\left(\sqrt{\rho_1}, \ldots, \sqrt{\rho_K}\right)$$

with elements given by

$$\rho_k = \frac{p_k}{p_k + p_{e_k}}\,, \tag{50}$$

where $p_k = E\{|y_k|^2\}$. Note that this factor re-scales each quantized microphone signal to its original power, since quantization might be expected not to increase the microphone signal power. The corresponding microphone signal correlation matrix $\mathbf{R}_{y_q y_q}$ is then given by

$$\begin{aligned}\mathbf{R}_{y_q y_q} &= E\left\{\mathbf{A}(\mathbf{y}+\mathbf{e})(\mathbf{y}+\mathbf{e})^H\mathbf{A}^H\right\} & (51)\\ &= \mathbf{A}(\mathbf{R}_{yy}+\mathbf{R}_{ee})\mathbf{A}^H & (52)\\ &= \mathbf{A}\left(\mathbf{R}_{yy}+\text{diag}(\mathbf{p}_e)\right)\mathbf{A}^H\,. & (53)\end{aligned}$$

As we can observe from (50) and (51), the off-diagonal elements of the k-th column of $\mathbf{R}_{y_q y_q}$ are the off-diagonal elements of the k-th column of $\mathbf{R}_{yy}$ multiplied by ρ_k, while the elements in the main diagonal of $\mathbf{R}_{y_q y_q}$ are equal to those of $\mathbf{R}_{yy}$. In summary, $\mathbf{R}_{y_q y_q}$ models the effect of quantization as a decrease in the cross-correlation between the microphone signals (hence the decrease in the off-diagonal elements), while their powers (given by the main diagonal elements) remain unchanged.

The cross-correlation $\mathbf{r}_{y_q x_{\text{ref}}}$ can be obtained by using (49) as

$$\begin{aligned}\mathbf{r}_{y_q x_{\text{ref}}} &= E\{\mathbf{y}_q x^*_{\text{ref}}\} = E\{\mathbf{A}(\mathbf{y}+\mathbf{e})x^*_{\text{ref}}\} & (54)\\ &= \mathbf{A}E\{\mathbf{y}\,x^*_{\text{ref}}\} + \mathbf{A}E\{\mathbf{e}\,x^*_{\text{ref}}\} = \mathbf{A}\mathbf{r}_{yx_{\text{ref}}}\,, & (55)\end{aligned}$$

where we have assumed that $\mathbf{e}$ and x_{ref} are uncorrelated. Following (5) and (20) we can express the MMSE $J_q(\hat{\mathbf{w}}_q)$ obtained from the MWF computed based on $\mathbf{y}_q$ as

$$J_q(\hat{\mathbf{w}}_q) = P_{\text{ref}} - \mathbf{r}^H_{y_q x_{\text{ref}}}\mathbf{R}^{-1}_{y_q y_q}\mathbf{r}_{y_q x_{\text{ref}}}\,. \tag{56}$$

Using (51) and (54) we find

$$\begin{aligned}J_q(\hat{\mathbf{w}}_q) &= P_{\text{ref}} - \mathbf{r}^H_{y_q x_{\text{ref}}}\mathbf{R}^{-1}_{y_q y_q}\mathbf{r}_{y_q x_{\text{ref}}} & (57)\\ &= P_{\text{ref}} - \mathbf{r}^H_{yx_{\text{ref}}}\mathbf{A}^H\mathbf{A}^{-H}\left(\mathbf{R}_{yy}+\mathbf{R}_{ee}\right)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{r}_{yx_{\text{ref}}} & (58)\\ &= P_{\text{ref}} - \mathbf{r}^H_{yx_{\text{ref}}}\left(\mathbf{R}_{yy}+\mathbf{R}_{ee}\right)^{-1}\mathbf{r}_{yx_{\text{ref}}}\,. & (59)\end{aligned}$$

which coincides with (20), proving that

$$J_q(\hat{\mathbf{w}}_q) = J_e(\hat{\mathbf{w}}_e)\,. \tag{60}$$

We can then conclude from the derivation presented above that modeling the effect of quantization noise through (11) or (49) leads to the same MMSE and thus to the same impact and gradient metric. Therefore there is no dilemma between the two models regarding the effect of the quantization of the microphone signals on the MWF.

References

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

[2] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537 – 568, 2009.

[3] T. Rault, A. Bouabdallah, and Y. Challal, "Energy efficiency in wireless sensor networks: A top-down survey," *Computer Networks*, vol. 67, pp. 104 – 122, 2014.

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

[5] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, July 2011.

[6] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 487–503, July 2005.

[7] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, Feb 2009.

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

[9] S. Liu, S. P. Chepuri, M. Fardad, E. Maşazade, G. Leus, and P. K. Varshney, "Sensor selection for estimation with correlated measurement noise," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3509–3522, July 2016.

[10] A. Bertrand and M. Moonen, "Efficient sensor subset selection and link failure response for linear MMSE signal estimation in wireless sensor networks," in *Proc. of the European signal processing conference (EUSIPCO)*, Aalborg - Denmark, August 2010, pp. 1092–1096.

[11] 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 Transactions on Signal Processing*, vol. 60, no. 11, pp. 5857–5869, Nov 2012.

[12] J. Szurley, A. Bertrand, M. Moonen, P. Ruckebusch, and I. Moerman, "Utility based cross-layer collaboration for speech enhancement in wireless acoustic sensor networks," in *2011 19th European Signal Processing Conference*, Aug 2011, pp. 235–239.

[13] J. Szurley, A. Bertrand, P. Ruckebusch, I. Moerman, and M. Moonen, "Greedy distributed node selection for node-specific signal estimation in wireless sensor networks," *Signal Processing*, vol. 94, pp. 57 – 73, 2014.

[14] A. Zahedi, J. Østergaard, S. H. Jensen, S. Bech, and P. Naylor, "Audio coding in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 141 – 152, 2015, Special Issue on ad hoc microphone arrays and wireless acoustic sensor networks, Special Issue on Fractional Signal Processing and Applications.

[15] A. Zahedi, "Source coding for wireless distributed microphones in reverberant environments," Ph.D. dissertation, Aalborg University, 2016.

[16] F. de la Hucha Arce, F. Rosas, M. Moonen, M. Verhelst, and A. Bertrand, "Generalized signal utility for LMMSE signal estimation with application to greedy quantization in wireless sensor networks," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1202–1206, Sept 2016.

[17] R. Serizel, M. Moonen, B. V. Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, April 2014.

[18] A. Hassani, A. Bertrand, and M. Moonen, "GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2557–2572, May 2016.

[19] A. Sripad and D. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 5, pp. 442–448, Oct 1977.

[20] R. M. Gray, "Quantization noise spectra," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1220–1244, Nov 1990.

[21] R. M. Gray and D. L. Neuhof, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct 1998.

[22] K. B. Petersen and M. S. Pedersen, "The Matrix Cookbook," Nov 2012, version 20121115. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “TIMIT acoustic-phonetic continuous speech corpus,” Linguistic Data Consortium, Tech. Rep., 1993.

[25] J. Greenberg, P. Peterson, and P. Zurek, “Intelligibility-weighted measures for speech-to-interference ratio and speech system performance,” *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, November 1993.

[26] A. S.3.5-1997, “American national standard methods for calculation of the speech intelligibility index,” Acoust. Soc. America, Tech. Rep., June 1997.

[27] Auditec, “Auditory tests (revised), compact disc,” 1997, Auditec St.Louis, MO.

[28] S. Labs, “EFM32 Wonder Gecko 32-bit ARM Cortex-M4 microcontroller,” 2017, http://www.silabs.com/products/mcu/32-bit/efm32-wonder-gecko.

[29] R. P. Foundation, “Raspberry Pi 3,” 2017, https://www.raspberrypi.org/products/raspberry-pi-3-model-b/.

[30] F. Rosas, R. D. Souza, M. E. Pellenz, C. Oberli, G. Brante, M. Verhelst, and S. Pollin, “Optimizing the code rate of energy-constrained wireless communications with HARQ,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 191–205, Jan 2016.

[31] “IEEE Standard for Low-Rate Wireless Networks IEEE 802.15.4,” 2015.

[32] F. Rosas and C. Oberli, “Modulation and SNR optimization for achieving energy-efficient communications over short-range fading channels,” *IEEE Trans. on Wireless Communications*, vol. 11, no. 12, pp. 4286–4295, December 2012.