Energy-aware noise reduction for wireless acoustic sensor networks

Zhang, Jie

DOI
10.4233/uuid:7461ee1c-1f76-43aa-b8bb-8da6f57c3528

Publication date
2020

Document Version
Final published version

Citation (APA)
Zhang, J. (2020). *Energy-aware noise reduction for wireless acoustic sensor networks*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:7461ee1c-1f76-43aa-b8bb-8da6f57c3528

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Energy-Aware Noise Reduction for Wireless Acoustic Sensor Networks

ENERGY-AWARE NOISE REDUCTION FOR WIRELESS ACOUSTIC SENSOR NETWORKS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
woensdag 15 januari 2020 om 10:00 uur

door

Jie ZHANG

Master of Science in Computer Applied Technology,
Peking University, Beijing, China.
geboren te Anhui, China.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. ir. R. Heusdens
promotor: Dr. ir. R. C. Hendriks

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. R. Heusdens,	Technische Universiteit Delft, Netherlands
Dr. ir. R. C. Hendriks,	Technische Universiteit Delft, Netherlands

Onafhankelijke leden:

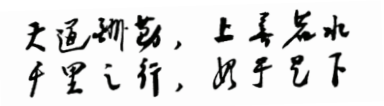
Prof. dr. J. Jensen	Aalborg Universitet, Denmark
Prof. dr. -Ing. T. Gerkmann	Universität Hamburg, Germany
Prof. dr. ir. A. Bertrand	Katholieke Universiteit Leuven, Belgium
Prof. dr. ir. G.J.T. Leus	Technische Universiteit Delft, Netherlands
Prof. dr. A. Hanjalic	Technische Universiteit Delft, Netherlands

This work described in this thesis was financially supported by China Scholarship Council (CSC) under Grant 201506010331 and in part by the Circuits and Systems (CAS) group, Delft University of Technology, Delft, The Netherlands.

Keywords: Microphone subset selection, rate distribution, noise reduction, binaural cue preservation, distributed algorithms, relative acoustic transfer function, quantization, bit-rate, power consumption, energy efficiency, wireless acoustic sensor networks.

ISBN 978-94-6366-239-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.



To Zhenzhen, who gives me ∞ love and support.
To Chi, who gives me ∞ hope for life.
To the people, who are working hard for their dreams.

CONTENTS

SUMMARY

In speech processing applications, e.g., speech recognition, hearing aids (HAs), video conferencing, and human-computer interaction, speech enhancement or noise reduction is an essential front-end task, as the recorded speech signals are inevitably corrupted by interference, including coherent/incoherent noise and reverberation. Traditional noise reduction algorithms are mostly based on spatial filtering techniques using a microphone array. The performance of the noise reduction algorithms scales with the number of microphones that are involved in filtering, but a large-sized microphone array cannot be mounted in many realistic systems, e.g., HAs. In the last few decades, with a great development in micro-electro-mechanical systems, wireless devices are more and more commonly-used in our daily life, like the smartphone, laptop, wireless HA, and ipad. These devices have acoustic sensors equipped and a capability of wireless communication, leading to a wireless acoustic sensor network (WASN). The WASN can be organized in a centralized fashion where all the devices are only allowed to connect with a fusion center (FC), or in a decentralized way where the devices are connected with the close-by counterparts via wireless links. This WASN can resolve the disadvantages of the traditional microphone array systems, since the wireless devices can be placed anywhere in the vicinity and one device is able to make use of measurements from other external devices. More importantly, the acoustic scene can be sampled more comprehensively, resulting in a potential improvement in noise reduction performance.

Due to the fact that these wireless devices are usually battery powered, it is desirable that the noise reduction task is accomplished before each device uses up its power budget, such that the life-time of the network can be improved. It is therefore important to make use of the total power budget as efficiently as possible. The power usage in terms of data transmission is related to the number of sensors, the distance and the transmission rate between two communicating nodes. In this thesis, we will mainly focus on saving the total power usage over the WASN while maintaining an expected signal/parameter estimation performance.

First, we consider a strategy of sensor selection for improving the WASN energy efficiency, since the total power usage is directly affected by the number of sensors, as the more sensors that are involved in spatial filtering, the higher power usage is required for data aggregation in a WASN. The sensor selection problem is formulated as minimizing the total power usage given a constraint on the output noise variance. Under the utilization of a minimum variance distortionless response (MVDR) beamformer, the optimal subset of sensors can be found by using convex optimization techniques. Then, the selected sensors will use full-rate quantization to send their measurements to the FC for the subsequent beamforming. Experimental results show that the sensors close to the target source(s), those around the FC and some next to the coherent noise sources are more likely to be chosen.

Second, we consider a strategy of rate distribution for improving the WASN energy

efficiency, since the power usage is also related to the communication rate. The aforementioned sensor selection is actually a *hard decision* on the status of the sensors, while rate distribution allows for a *soft decision* on the sensors. In other words, we now allow the sensors to communicate with the FC at any possible rate between zero and a pre-defined maximum value. Such a rate distribution problem is formulated similarly, i.e., minimizing the total power usage subject to a constraint on the desired noise reduction performance, but the optimization unknowns are the integer rates rather than the Boolean selection variables. Both sensor selection and rate distribution can save the power usage and guarantee the expected noise reduction performance. By leveraging the multiple decision strategy, rate distribution can further reduce the power usage compared to sensor selection. Further, we consider a more complicated but practical scenario of a large-scale WASN consisting of HAs. For the HA user, it is necessary not only to suppress the interfering sources, but also to preserve the binaural cues of all existing directional sources. The binaural linearly constrained minimum variance (LCMV) beamformer is capable of performing joint noise reduction and binaural cue preservation. The proposed rate-distribution algorithm can thus easily be applied in this scenario by substituting the binaural LCMV beamformer into the original problem formulation. In addition, since the centralized implementation is not robust against changes in the network topology, particularly if the FC drops out from the network, we extend the considered rate distribution approach to a fully decentralized fashion.

Finally, we consider the rate distribution problem in the context of estimating relative acoustic transfer function (RTF), since the beamformers rely on the RTF information. More importantly the sensor selection or rate distribution method that was proposed before is based on the RTF. Estimating the RTF can be achieved by exploiting the noise and noisy correlation matrices, while estimating these correlation matrices requires a large amount of data transfer. Hence, rate distribution is an option for saving the power usage in RTF estimation. For this, we consider two well-known RTF estimation approaches, i.e., covariance subtraction (CS) and covariance whitening (CW), and analyze their performance in terms of bit rate. Following the rate-distribution formulation in the context of noise reduction, we also propose to minimize the total power usage under a constraint on an expected RTF estimation accuracy. We find that the resulting rate distribution is mainly affected by the distance between the sensors and the FC and the signal-to-noise ratio. It is shown that many bits in microphone recordings are redundant and the full-rate transmission is certainly unnecessary.

SAMENVATTING

In spraakverwerkingstoepassingen, zoals spraakherkenning, hoorapparaten (HAs), videoconferenties, en de interactie tussen mens en computer, zijn spraakversterking of ruisreductie een belangrijke front-end taak. Het is namelijk onontkoombaar is dat de opgenomen spraaksignalen interferentie bevatten, waaronder coherente/incoherente ruis en reverberatie. Traditionele ruisreductie algoritmes zijn vooral gebaseerd op spatiele filter technieken met een microfoon array. De prestaties van ruisreductie algoritmes schalen met het aantal microfoons die worden gebruikt voor het filteren. Echter, grote microfoon-arrays kunnen vanwege de afmetingen vaak niet worden gecombineerd met toepassingen zoals HAs. In de laatste decennia, door een sterke ontwikkeling in micro-elektro-mechanische system, zijn draadloze apparaten steeds normaler geworden in ons dagelijks leven. Denk hierbij aan bijvoorbeeld de smartphone, laptop, draadloze HAs, en de iPad. Dergelijke apparaten zijn uitgerust met akoestische sensoren en kunnen draadloos communiceren, en vormen zo een draadloos akoestisch sensor netwerk (WASN). Een WASN kan worden geordend op een gecentraliseerde manier waar alle apparaten alleen mogen verbinden met een fusiecentrum (FC), of op een gedecentraliseerde manier waar apparaten zijn verbonden met hun nabijliggende tegenhangers via draadloze verbindingen. Een dergelijk WASN kan de nadelen van traditionele microfoon-array systemen opheffen, aangezien de draadloze apparaten overal in de nabijheid kunnen worden geplaatst, en elk apparaat gebruik kan maken van de metingen van andere externe apparaten. Belangrijker is dat de akoestische omgeving beter kan worden bemonsterd, wat resulteert in een potentiele verbetering van de ruisreductie prestaties.

Doordat deze draadloze apparaten meestal in energie worden voorzien door een batterij, is het wenselijk dat ruisreductie wordt bewerkstelligd voordat elk apparaat zijn vermogensbudget verbruikt, zodat de levensduur van het netwerk kan worden verbeterd. Het vermogensverbruik in termen van datatransmissie is gerelateerd aan het aantal sensoren, de afstand, en transmissiesnelheid tussen twee communicerende nodes. In deze scriptie zullen we ons vooral concentreren op het besparen van het totale energieverbruik in de WASN, terwijl de verwachte signaal/parameter-schatting in stand wordt gehouden.

Als eerste beschouwen we een sensorselectie-strategie om de energie-efficiëntie van het WASN te verbeteren, aangezien het totale energieverbruik direct gerelateerd is aan het aantal sensoren, alsmede doordat een hoger energieverbruik nodig is voor data-aggregatie wanneer er meer sensoren zijn betrokken bij het spatiele filteren. Dit sensorselectie probleem is geformuleerd als het minimaliseren van het totale energieverbruik gegeven een beperking op de output ruisvariantie. Door gebruik te maken van een minimum variance distortionless response (MVDR) beamformer kan de optimale subset van sensoren worden gevonden door convexe optimalisatietechnieken. De geselecteerde sensoren kunnen dan hun metingen kwantiseren en naar de FC sturen om te beamformen. Experimentele resultaten tonen aan dat de sensoren dichtbij de akoesti-

sche bron, de sensoren dichtbij de FC, en de sensoren nabij coherente ruisbronnen met meer waarschijnlijkheid worden gekozen.

Als tweede overwegen we een strategie voor de herverderling van de bit-rate om de energie-efficiëntie van de WASN te verbeteren. Het energieverbruik wordt namelijk ook bepaald door de communicatie bit-rate. De eerder beschreven sensorselectie is in feite een *harde beslissing* over de status van de sensoren, terwijl bit-rate verdeling het mogelijk maakt om een *zachte beslissing* te nemen. Met andere woorden, we staan nu toe dat de sensoren kunnen communiceren met de FC met elke transmissiesnelheid tussen nul bits en een gegeven maximale waarde. Een dergelijk bit-rate herverdelingsprobleem wordt op een gelijkaardige manier geformuleerd als het sensor selectie probleem. Namelijk als het minimaliseren van het totale energieverbruik onderworpen aan een beperking op de gewenste ruisreductie prestaties, waarbij de optimalisatievariabelen nu gehele getallen zijn in plaats van Booleaanse selectievariabelen. Zowel sensorselectie als bit-rate herverdeling kunnen energie besparen, en waarborgen de verwachte ruisreductie prestaties. Door gebruik te maken van een meerkeuzige besluitstrategie, kan snelheidsverdeling het energieverbruik verder verlagen vergeleken met sensorselectie. We overwegen ook een gecompliceerder maar praktisch scenario van een grootschalig WASN die bestaat uit HAs. Voor een HA gebruiker is het niet alleen nodig om interfererende bronnen te onderdrukken, maar ook om binaurale signalen van alle bestaande directionele bronnen te behouden. De binaural linearly constrained minimum variance (LCMV) beamformer is in staat om gelijktijdig ruis te onderdrukken en binaurale signalen te behouden. Het voorgestelde bit-rate herverdeling algoritme kan dus makkelijk worden toegepast op dit scenario door de binaurale LCMV beamformer in de originele probleemstelling te vervangen. De gecentraliseerde implementatie is echter niet robuust tegen veranderingen in de netwerktopologie, vooral als de FC uitvalt. Daarom breiden we deze bit-rate herverderling uit naar een volledig gedecentraliseerde vorm.

Tenslotte beschouwen we het bit-rate verdelingsprobleem in de context van het schatten van de relatieve akoestische overdrachtsfunctie (RTF), waar de beamformers gebruik van maken. Sterker nog, de eerder voorgestelde sensor selectie en snelheidsverdeling methode zijn gebaseerd op de RTF. Het schatten van de RTF van worden behaald door de ruis en ruizige correlatiematrices te exploiteren, alhoewel het schatten van deze correlatiematrices een grote dataoverdracht vereist. Daarom is bit-rate verdeling een manier om energie te besparen bij het schatten van de RTF. Om dit te doen overwegen we twee bekende benaderingen voor het bepalen van de RTF, namelijk covariance subtraction (CS) en covariance whitening (CW), en analyseren hun prestaties aangaande de bit-rate. In aansluiting op de bit-rate herverdelingsformulering in de context van ruisreductie, stellen wij ook voor om het totale energieverbruik te minimaliseren gegeven een beperking op de verwachte nauwkeurigheid van de schatting van de RTF. We bemerken dat de resulterende bit-rate herverdeling vooral wordt bepaald door de afstand tussen de sensoren en de FC, alsmede de signaal-ruisverhouding. We tonen aan dat veel bits in microfoonopnames overbodig zijn en dat een maximale bit-rate zeker onnodig is.

1

INTRODUCTION

DURING the last few decades, noise reduction, often-time called speech enhancement, has been widely investigated. In many audio processing applications, e.g., speech recognition [1, 2], teleconferencing systems [3], sound source localization [4, 5, 6], mobile robot systems [7, 8], to list a few, it can be exploited as a front-end process to improve the signal-to-noise ratio (SNR) for subsequent tasks. Other important applications of noise reduction are the improvement of speech intelligibility for hearing-impaired listeners [9] and to increase the recognition rate of speech recognition systems [1, 2]. With regard to the noise reduction problem, both single-microphone algorithms [10, 11, 12, 13] and multi-microphone algorithms [14, 15, 16, 17, 18, 19] can be exploited. For the single-microphone noise reduction algorithms, only temporal (spectral) information contained in the input signal is exploited. For the multi-microphone algorithms, also called beamforming, the sound field is sampled both in time and in space, so that both temporal and spatial information can be used. The multi-microphone techniques can thus achieve a great improvement in the noise reduction performance compared to the single-microphone counterpart.

1.1. CONVENTIONAL MULTI-MICROPHONE NOISE REDUCTION

Conventional multi-microphone noise reduction systems are mostly based on the utilization of a microphone array, as Fig. 1.1 depicts. The microphone array provides multi-microphone audio measurements, from which both temporal and spatial information can be employed. In general, the multi-microphone noise reduction methods can be categorized into two classes: 1) linearly constrained beamforming [14, 15, 20] and 2) unconstrained beamforming [21, 22, 23]. The most well-known linearly constrained approach is the linearly constrained minimum variance (LCMV) beamformer [15, 20], which minimizes the output signal variance subject to a set of linear constraints. For example, these linear constraints can be used to steer a beam having a response of one into the directions of the sources of interest and steer a beam having a response of zero into the directions of the interferers, such that the power of the target sources can exactly be preserved and the noise signals can be entirely suppressed. Due to the explicit

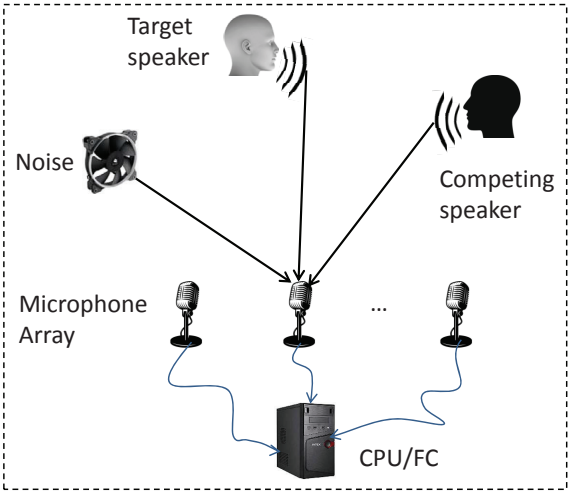


Figure 1.1: A typical example of the traditional microphone array based speech processing systems.

matrix inversion involved in calculating the LCMV beamformer coefficients, the generalized sidelobe canceler (GSC) [15, 24] is often used as an alternative formulation of the LCMV beamformer, which can be implemented more efficiently. As a special case of the LCMV beamformer, the minimum variance distortionless response (MVDR) beamformer minimizes the output signal variance such that the target signal is undistorted after spatial filtering, i.e., the MVDR beamformer only takes into account the linear constraint associated with a single target source. Hence, in the multi-microphone case the MVDR beamformer has degrees-of-freedom left to adjust the filter coefficients, leading to a better noise reduction performance. The LCMV beamformer can cope with multiple sources at the cost of sacrificing the noise reduction capability. Unconstrained beamforming, e.g., multi-microphone Wiener filter (MWF) based algorithms, is based on the use of a minimum mean square-error (MMSE) estimator, which minimizes the expected mean square-error (MSE) between the ground truth of the target signal (or the target signal at a reference microphone) and the estimated target signal (or the estimated target at the same reference microphone). The MWF can achieve a better noise reduction performance than the linearly constrained beamformers, yet it would also distort the target signal inevitably, since no constraints related to the target sources/interferers are taken into account. In order to alleviate this drawback, one can add a constraint for the MMSE estimator to control the signal distortion level, leading to the speech distortion weighted MWF (SD-MWF) [23], which can then trade-off the noise reduction capability and the signal distortion level.

In order to implement the aforementioned multi-microphone noise reduction algorithms, usually the second order statistics (SOS), e.g., noise correlation matrix and noisy correlation matrix, and the acoustic transfer functions (ATFs) are required. For estimating these parameters, data transmission and data processing are necessary. Given a perfect voice activity detector (VAD), the microphone measurements can be classified

into noise-only segments and speech-plus-noise segments. The noise and noisy correlation matrices can be estimated during these two periods using sample covariance matrices [25, 26, 27]. The ATFs characterize the channel responses from the sources to the receivers, which might include a direct-path component and a series of reflections in a reverberant environment. Instead of using the ATF for beamforming directly, the relative acoustic transfer function (RTF) can also be used [28, 29, 30]. The RTF is defined as the normalized ATF with respect to an arbitrarily chosen reference microphone. In practice, the errors in estimating these involved parameters would significantly affect the performance of the aforementioned multi-microphone noise reduction algorithms [31].

There are several limitations of conventional microphone array based noise reduction systems. From the perspective of system design, an obvious drawback of such traditional microphone arrays is the fact that it is impractical to rearrange the microphones in such a wired array, since all the microphones are physically linked. For instance, it is not convenient to add a new microphone to the array system. Due to the fixed array layout and the fact that the array cannot be placed anywhere, the awareness of sensing the acoustical scene is limited, in particular when the speech sources of interest are far away from the microphone array. Moreover, the size of the conventional arrays is another limitation to their practical usage, as typically the maximum array size is determined by the application at hand. For instance, binaural hearing aids (HAs) can only host a small number of microphones (usually 2-4 microphones per HA) [32].

1.2. WIRELESS ACOUSTIC SENSOR NETWORK

Nowadays, we are surrounded by portable devices, e.g., smartphones, laptops, hands-free telephony kits, binaural HAs, each equipped with one or several microphones. These devices can be positioned anywhere in the vicinity of interest. With the help of wireless communication capabilities, the devices can communicate (or can be connected) with other devices or a (remote) fusion center (FC), resulting in a wireless microphone network or so-called wireless acoustic sensor network (WASN). Fig. 1.2 illustrates a typical example of WASNs, which includes several smartphones, laptops, an HA and a microphone array. Note that each wireless device uses an analog-to-digital converter (ADC) to convert the analog acoustic signals to the digital versions that can be processed subsequently. Also, the radio frequency (RF) module which is usually a small electronic device is utilized to transmit and/or receive radio signals between two devices.

The utilization of WASNs can potentially overcome the limitations in the context of traditional microphone array systems and bring several benefits for audio processing applications. Firstly, the wireless devices can be placed at locations difficult to reach with conventional wired microphone arrays. The WASNs can thus sample and monitor a much larger acoustical scene. With such sensor placement, some sensor nodes might be close to the target speaker location and have a higher SNR. As a result, these sensors can record high-quality audio measurements that could be very beneficial. Secondly, the WASNs do not have the array-size limitation. For example, even though the HA applications require small-sized microphone arrays, the hearing assistive devices can still make use of the data measurements from other external devices, if these devices can transmit their recordings to the HAs via wireless links. With these advantages, it is expected that WASNs might be the next generation for audio acquisition systems [33]. Further, the

1

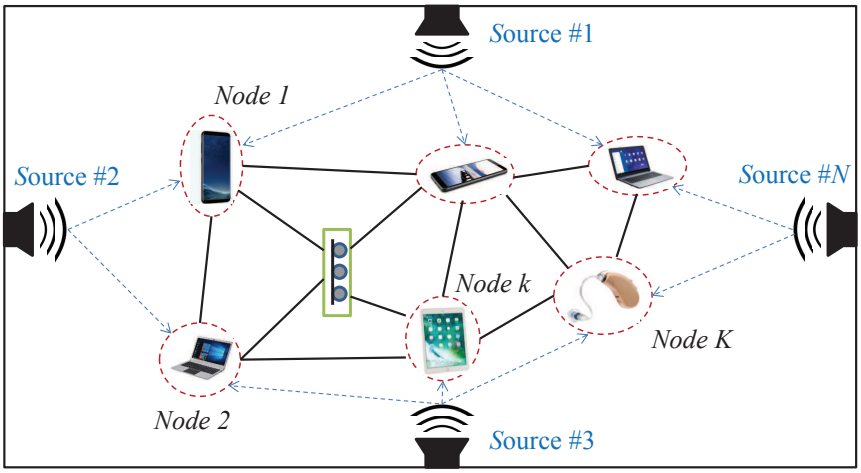


Figure 1.2: A typical example of WASNs, which consists of a couple of wireless devices, e.g., smartphones, laptops, an HA, a microphone array. The nodes can communicate with the close-by neighbors.

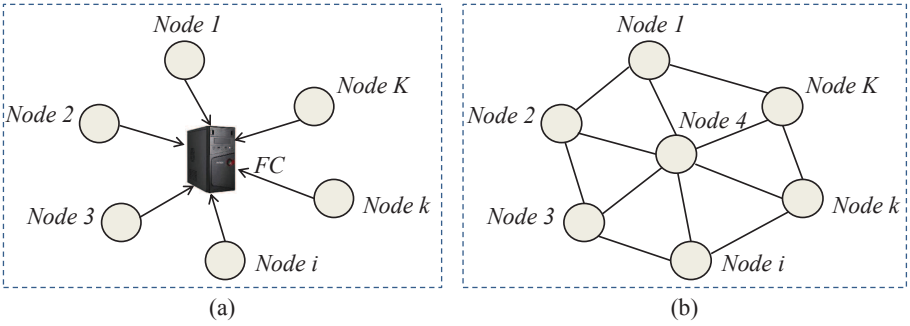


Figure 1.3: (a) The topology of a centralized microphone network (or centralized WASN), where all the microphone nodes are physically (or wirelessly) connected to the FC. (b) The topology of a distributed WASN, where each sensor node is allowed to communicate with the neighboring nodes.

WASNs can be constructed more flexibly, e.g., either in a centralized scheme or in a decentralized fashion. For the centralized WASNs as shown in Fig. 1.3(a), all the devices are connected to an FC (similar to the conventional microphone array case, but via wireless links). In this case, the FC collects the data measurements from all other sensors and conducts all computations. For the decentralized WASNs as shown in Fig. 1.3(b), there is no FC and the sensor nodes are connected to their neighboring nodes only. In this case, all the sensors have to collaborate together to complete the tasks of interest, resulting in a frequent information exchange over the WASN.

1.3. ENERGY-AWARE NOISE REDUCTION IN WASNS

An important challenge for signal processing in wireless sensor networks (WSNs) in general and WASNs in particular is the energy consumption, as in practice the sensors are usually battery driven with a limited energy budget. When a sensor node is depleted of energy, it will die and will be disconnected from the network. This will degrade the performance of the application significantly if such a sensor is very informative, e.g., having a high SNR. Hence, the life expectation of the WASNs is directly affected by the power consumption. It is important to make use of the energy budget as efficiently as possible, such that the network lifetime can be maximized. Generally speaking, the power within the context of WASNs is consumed by two processes: data transmission and data processing [34, 35]. The data transmission occurs between all the nodes and the FC in the centralized setup, or between neighboring node pairs in the decentralized setup. The data processing is conducted at the FC end in the centralized case, or distributed over the nodes in the decentralized case.

The power consumption of each individual device depends on the transmission energy and the power for having it activated. If a sensor is turned off, it has no power consumption. If a sensor is turned on, its power consumption will then be the summation of the power for having the sensor activated, the transmission energy and the processing power. The transmission energy of the activated sensors depends on the transmission distance, transmission rate (in bits per sample) and the noise power spectral density (PSD) of the communication channel [36, 37, 38]. The larger the transmission distance (the transmission rate or channel noise PSD), the higher the transmission energy. In addition, the total power consumption over the WASNs is the summation of the power consumption of all devices, and the number of the activated sensors will thus affect the total power consumption as well. Therefore, in order to improve the energy efficiency of noise reduction techniques or signal parameter estimation algorithms in WASNs, different strategies can be designed from different perspectives.

1.4. RESEARCH QUESTIONS

In this section, we will propose several research questions that will be discussed in this dissertation, together with the motivations behind them.

As the devices in the WASN are equipped with a limited battery resource, they should use the power resource as efficiently as possible in order to prolong the lifetime of the network. Extracting the clean target signal(s) from the mixed noisy sensor measurements in a WASN is required by many applications, and can be achieved using multi-microphone spatial filtering techniques, e.g., MVDR, LCMV, MWF as mentioned in Sec. 1.1. Let $f(\mathbf{w},\mathbf{x})$ denote a cost function representing the total power consumption over the WASNs, i.e., the total transmission costs between all the sensor nodes or power needed to keep sensors turned on. Obviously, the total power consumption will depend on the applied beamformer weights $\mathbf{w}$ (e.g., having a weight of zero for a particular sensor implies no transmission of data is necessary). Further, it is expected that the total power consumption depends on some additional parameters $\mathbf{x}$, which can represent transmission bit-rate or selection variables. In addition, let $g(\mathbf{w},\mathbf{x})$ denote the performance or distortion metric, e.g., output noise power, output SNR or output intelligibility of the

multi-microphone spatial filter when applying filter $\mathbf{w}$. We can then formulate the following two related constrained optimization problems, that are,

$$\underset{\mathbf{w},\mathbf{x}}{\text{minimize}}\ f(\mathbf{w},\mathbf{x}) \quad \text{subject to } g(\mathbf{w},\mathbf{x}) \leq \beta, \tag{1.1}$$

$$\underset{\mathbf{w},\mathbf{x}}{\text{minimize}}\ g(\mathbf{w},\mathbf{x}) \quad \text{subject to } f(\mathbf{w},\mathbf{x}) \leq C, \tag{1.2}$$

where C denotes the total power budget. Notice that in (1.1) we considered $g(\mathbf{w},\mathbf{x})$ to be a distortion with β the maximum allowable distortion, but in the case where $g(\mathbf{w},\mathbf{x})$ represents a performance metric, the inequality sign should be replaced by a larger or equal sign. As a result, the optimization problem (1.1) can be interpreted as the following research question:

Q1: Given a prescribed performance, can we design an effective strategy for saving the power consumption over WASNs?

Depending on the exact physical meaning of the vector variable $\mathbf{x}$ in (1.1), we can investigate different optimization strategies, leading to several varieties of research question **Q1**. Firstly, it is possible that some nodes are closer to the target sources, having a higher SNR and some nodes are closer to the interferers having a lower SNR. Although including more sensors in the beamformer will generally increase the noise reduction performance, it will also consume more transmission power, because all the sensor nodes have to transmit their data to the FC in a centralized WASN. Moreover, clearly, not all sensors are as informative. To achieve a certain expected performance, it could be that we do not need to use all the measurements from all the sensors, i.e., a subset of the sensors might be sufficient. Instead of blindly using all the sensors, selecting the most informative subset of sensors for noise reduction algorithms would significantly decrease the amount of the transmitted data, leading to a saving of transmission cost and communication bandwidth. Therefore, from the perspective of sensor selection, the research question **Q1** can be made more specific as

Q1.1: Given a certain expected performance, can we choose a subset of microphone nodes that minimizes the power consumption for beamforming?

From the perspective of signal acquisition, the sensor measurements are already quantized via ADCs. In case we use full-rate transmission for the raw data, as is typically done, a larger amount of energy usage will be required compared to the situation where signals are quantized at lower rates, obviously, at the cost of introducing more quantization noise. The wireless transmission power is directly related to the bit rate (e.g., an exponential relationship). This makes it worth to take into account the bit-rate allocation among the different sensor nodes before transmission. Given the desired performance, it is possible that certain information is redundant and lower rates are sufficient and more energy efficient. Making use of the bit-rate budget as efficiently as possible would be an effective way to save the energy consumption. Therefore, from the perspective of rate distribution, the research question **Q1** can also be further specified as

Q1.2: Given a certain expected performance, how to efficiently distribute the bits for signal quantization in order to reduce power consumption?

Since the topology of the considered WASN could be time-varying, it is more preferable to organize the network in a decentralized way, resulting in the requirement of distributed beamforming based noise reduction algorithms. Given the research question **Q1**, it is then natural to ask whether

Q2: Given a prescribed noise reduction performance, how to design an efficient data transmission strategy between nodes to reduce the power consumption for distributed beamforming?

One of the potential applications of WASNs are hearing aids (HAs). In addition to performing noise reduction, HAs typically have to satisfy certain constraints on the preservation of the spatial sound information. These are often referred to as spatial cues. In such WASNs for HAs, it is thus required to jointly perform noise reduction and spatial cue preservation for the HA users. The additional microphones in the WASN offer additional advantages over the use of a conventional pair of HAs. Among these advantages is the improved ability of noise reduction (or improved speech intelligibility), and, the improved ability to preserve binaural spatial cues of interfering sources. Roughly speaking, it holds that the more sensors are involved, the higher the degrees of freedom to perform jointly noise reduction and spatial cue preservation. However, incorporating all the existing devices in the WASN at full quanization rate might consume a larger than necessary amount of transmission power. This leads to the following hearing-aid related research question:

Q3: For the hearing-aid devices, how to efficiently make use of the measurements from external devices to jointly achieve noise reduction and binaural cue preservation?

Typically, multi-microphone noise reduction algorithms require knowledge on the ATFs or RTFs of the target sources with respect to the devices. Depending on the exact formulation, this can be implicit (via a dependency on the target correlation matrix), or explicit. In practice, the ATF or RTF is unknown and needs to be estimated. Within the WASN context, this comes with transmission and quantization of data and raises the question what the optimal rate distribution in terms of energy consumption is in order to obtain a prescribed performance. For the RTF estimation problem, two well-known methods are available. These are the covariance subtraction (CS) method [39, 40, 41, 42, 43] and covariance whitening (CW) method [18, 29, 44, 45]. Both approaches require estimates of correlation matrices. In a centralized setup, estimating these two matrices is performed via average smoothing over a sufficiently long period of sensor measurements after all the measurements are quantized and transmitted to the FC. Hence, similar to the noise reduction problem, there is a trade-off between the RTF estimation accuracy and the total energy consumption, leading to the following research question

1

Q4: Given a prescribed RTF estimation accuracy, can we design an effective data transmission strategy for saving the power consumption over WASNs?

1.5. STRUCTURE OF THE DISSERTATION

In this section, we will present the structure of this dissertation by summarizing the contribution of each included chapter.

1.5.1. CHAPTER 2: BACKGROUND

This chapter will give a more mathematical description of the fundamental knowledge and the research questions that are discussed in this dissertation. First, we present the general signal model, sensor selection model, rate distribution problem and assumptions that are used throughout the dissertation. Furthermore, we review the conventional multi-microphone spatial filtering based noise reduction algorithms (e.g., MVDR, LCMV) and a distributed implementation of the linearly-constrained beamformers. In addition, the CS and CW methods for RTF estimation are presented.

1.5.2. CHAPTER 3: MICROPHONE SUBSET SELECTION

This chapter answers research question **Q1** from the perspective of sensor selection, i.e., corresponding to **Q1.1**. In this chapter, we consider microphone subset selection for MVDR beamforming based multi-microphone noise reduction in WASNs. The traditional sensor selection problem is usually formulated by optimizing the performance measure subject to a constraint on the cardinality of the selected sensors, or the other way around. However, in the context of WASNs, we might not know how many sensors need to be included. Further, the energy usage is a vital concern within the context of WASNs. Therefore, we reformulate the sensor selection problem by minimizing the total transmission cost between all the sensor nodes and the FC and constraining the output noise power. Optimizing this sensor selection problem results in the best subset of sensors that satisfies the noise reduction performance and has the minimum transmission power.

For the proposed sensor selection problem, we present two methods for solving it. First, following convex optimization techniques, we derive the initial problem as a semi-definite optimization problem, which is based on the correlation matrices of the microphone measurements of the complete network or the ATFs. Given the correlation matrices or the ATFs, the sensor selection problem can be solved, which is called *model-driven sensor selection*. However, this model-driven method is impractical, since it depends on the statistical information of the complete network which is usually unavailable. In practice, we even do not know how many sensors are present in the WASNs, due to the fact that the wireless devices are free to join or leave the network. In order to make the proposed model-based method practical and avoid estimating the statistics beforehand, we further propose a greedy sensor selection approach, which is called *data-driven sensor selection*. It is shown that the performance of the greedy approach converges to that of the model-driven method, while it displays advantages in dynamic scenarios (e.g., with a moving FC). The sensors close to the target source(s), those close to the FC and some close to the interferers are more likely to be selected, since they have a higher SNR for

signal enhancement, a shorter distance for reducing transmission cost, and more information on noise sources for noise suppression, respectively.

1.5.3. CHAPTER 4: CENTRALIZED RATE DISTRIBUTION

This chapter answers research question **Q1** from the perspective of rate distribution, i.e., corresponding to **Q1.2**. In Chapter 3, we consider the use of sensor selection strategy to reduce the total transmission cost over the WASNs, which means that the decision on a sensor's status is *binary*, i.e., selected or not selected. If a sensor is selected, it will use full-rate quantization to communicate with the FC; if not, it will be turned off (or zero rate is allocated). In this chapter, we consider a more general selection strategy, which is called *rate distribution*. Differing from the sensor selection, rate distribution allows for a *soft decision* on the sensors, i.e., the sensor measurements can be quantized at any bit rate from zero to the maximum bit rate. Only if a sensor is allocated with zero bits, it is not selected from the perspective of sensor selection; otherwise it is selected. As the transmission power between the sensors and the FC is related to the bit rate, we can also reduce the energy consumption by optimizing the rate distribution.

Similar to the problem formulation in Chapter 3, in this chapter we minimize the total transmission power between all the sensors and the FC subject to a constraint on the output noise power, which is an integer optimization problem. Now, the optimization variable is not the binary selection variable anymore, but an integer valued bit-rate vector. Using convex optimization techniques and under the utilization of an LCMV beamformer, the rate distribution problem can also be derived as a semi-definite program. Additionally, in this chapter we investigate the relationship between sensor selection and rate allocation in a theoretical fashion. It can be shown that rate allocation is a generalization of sensor selection. More specifically, the sensor selection problem can be solved by considering the rate allocation problem. The best microphone subset can be determined by thresholding the bit rates, e.g., the sensors whose rates are larger than a certain threshold should be chosen for the sensor selection method. We also propose a bisection method for determining this threshold. Experimental results in simulated WASNs show that the sensors that are closer to the sources and the FC will be allocated with higher rates. Given the same constraint on noise reduction performance, if we neglect the power for having a sensor activated, the rate allocation method can always save more transmission power than the sensor selection method. However, if we take the power for having a sensor activated into account, this will not be always the case. More specifically, if this power is small, rate distribution is more cheaper in energy usage; otherwise sensor selection is more economical in transmission.

1.5.4. CHAPTER 5: DECENTRALIZED RATE DISTRIBUTION

This chapter answers research question **Q2**, i.e., rate distribution in the context of distributed beamforming. The centralized organization of WASNs has several limitations. Firstly, the amount of data that needs to be transmitted and saved at the FC scales up with the network size, which is a heavy load to the FC. Secondly, all the computations are performed at a single node and a disconnection of the FC will cause full collapse of the network. Thirdly, it will be very power demanding if the FC is far away from the sensors. In order to avoid these limitations, decentralized algorithms are preferred, since in

the decentralized setting, the beamformer calculation is distributed over all the nodes and the information exchange takes place between two neighboring nodes.

In this chapter, we present for the rate-distributed LCMV beamforming that was proposed in Chapter 4 a corresponding decentralized solution. We decentralize the obtained LCMV filter structure by exploiting an imposed block diagonal form of the noise correlation matrix. To calculate the beamformer weights in a decentralized fashion, the transmission rate between two neighboring nodes needs to be determined. For this, we reformulate the centralized rate distribution problem in a node-separable form, then we conclude that each node can determine its quantization rate locally without any information exchange. In a simulated WASN, we show that the proposed decentralized algorithm can achieve the same noise reduction performance as the centralized method, but consumes less power. In the decentralized setting, the sensors having a higher SNR will be allocated with a higher rate compared to the sensors having a lower SNR.

1.5.5. CHAPTER 6: RATE-DISTRIBUTED BINAURAL LCMV BEAMFORMING

This chapter investigates the situation where an HA is part of a bigger WASN and simultaneous noise reduction and preservation of spatial information is desired. With this application, we demonstrate a possible application of the rate-distribution LCMV beamforming method that was proposed in Chapter 4. More specifically, we study research question **Q3** and provide a strategy to trade-off the noise reduction performance versus spatial cue preservation capability via optimizing the quantization rate distribution.

In detail, the problem formulation remains the same as what we considered in Chapter 4, while now the FC is assumed to be one of the HAs, i.e., all the other devices should transmit their measurements to this HA at a certain rate. As the BLCMV beamformer can jointly perform noise reduction and spatial cue preservation, we substitute the BLCMV beamformer to the general rate-distribution problem, leading to the proposed rate distributed BLCMV beamforming problem in the binaural context. For comparison, we also apply the sensor selection method that was proposed in Chapter 3 to this binaural context. It is shown that in order to achieve the same noise reduction performance, the rate-distribution method has to activate more sensors, each at a much lower rate than the maximum rate, resulting in a saving of power consumption and a better spatial cue preservation compared to the sensor selection method.

1.5.6. CHAPTER 7: RELATIVE TRANSFER FUNCTION ESTIMATION

This chapter answers research question **Q4**. From the previous chapters, we can conclude that rate distribution is an effective way for saving the power consumption over WASNs. RTFs are required for practically any beamforming algorithm and can be calculated from the correlation matrices. However, in practice, correlation matrices are unknown as well and need to be estimated. Estimating the correlation matrices requires a large amount of data aggregation. As a result, the transmission rate will also affect the RTF estimation accuracy directly. Following the idea of optimizing the rate distribution that was used in the previous chapters, we propose rate-distributed RTF estimation methods in this chapter.

As the CS and CW methods are the most often-used methods for estimating RTFs, we first analyze the estimation accuracy of these two methods in terms of the quantization

rate. Then, we propose to minimize the total transmission power between all the sensor nodes and the FC, subject to a constraint on the RTF estimation accuracy. Substituting the error models of the two methods to the general problem formulation, we obtain two corresponding semi-definite programs for rate distribution, which are *model-driven approaches*. From the derivations, we find that the model-driven methods are based on the true RTF vector, which limits their practical usage. To alleviate this drawback, we further propose two corresponding *data-driven approaches*. Due to the fact that in practice the sensors send quantized data to the FC on a segment-by-segment basis, the FC can estimate the parameters that are required by the model-driven methods using the previously received segments and calculate the rate distribution by solving the model-driven optimization problems, then the sensors can use the obtained rate to transmit the new segment. In a simulated WASN, it is shown that to satisfy the same RTF estimation performance, the rate-distributed CW methods need less rate budgets, i.e., less transmission power, than the CS-related methods. With increasing the number of segments, the performance of the data-driven methods converges to that of the corresponding model-based approaches.

1.5.7. CHAPTER 8: CONCLUSIONS

In this chapter, we draw some final conclusions of this dissertation. In addition, we describe some open challenges and interesting questions. Also, we give some suggestions towards these open topics for future research.

1.6. LIST OF PAPERS

In this section, all the papers published during the PhD study are summarized.

JOURNALS

1. **J. Zhang**, R. Heusdens and R. C. Hendriks, Relative acoustic transfer function estimation in wireless acoustic sensor networks, *IEEE/ACM Trans. Audio, Speech, Language Process.*, 27(10): 1507–1519, 2019.

2. **J. Zhang**, A. I. Koutrouvelis, R. Heusdens and R. C. Hendriks, Distributed rate-constrained LCMV beamforming, *IEEE Signal Processing Letters*, 26(5): 675–679, 2019.

3. **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, Language Process.*, 26(11): 2015–2026, 2018.

4. **J. Zhang**, S. P. Chepuri, R. C. Hendriks and R. Heusdens, Microphone subset selection for MVDR beamformer based noise reduction, *IEEE/ACM Trans. Audio, Speech, Language Process.*, 26(3): 550–563, 2018.

5. C. Pang, H. Liu, **J. Zhang** and X. Li, Binaural sound localization based on reverberation weighting and generalized parametric mapping, *IEEE/ACM Trans. Audio, Speech, Language Process.*, 25(8): 1618–1632, 2017.

1

CONFERENCES

1. **J. Zhang**, R. Heusdens and R. C. Hendriks, Sensor selection and rate distribution based beamforming for wireless acoustic sensor networks, *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, Sept. 2019.

2. **J. Zhang**, R. Heusdens and R. C. Hendriks, Rate-distributed binaural LCMV beamforming for assistive hearing in wireless acoustic sensor networks, *IEEE 10th Sensor Array and multi-microphone Signal Processing Workshop (SAM)*, pp. 460–464, Sheffield, UK, July, 2018. (**Best student paper award**)

3. **J. Zhang**, R. C. Hendriks and R. Heusdens, Structured total least squares based internal delay estimation for distributed microphone auto-localization, *IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, Xi'an, China, Sept. 2016. (**Finalist best student paper contest**)

4. **J. Zhang**, R. C. Hendriks and R. Heusdens, Greedy gossip algorithm with synchronous communication for wireless sensor networks, *The 6th Joint WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux*, pp. 228–235, Louvain-la-Neuve, Belgium, May, 2016.

5. **J. Zhang**, R. Heusdens and R. C. Hendriks, Low-rate relative transfer function estimation in energy-aware wireless acoustic sensor networks, *Audio Analysis Workshop*, Aalborg University, Denmark, Aug. 2018.

6. **J. Zhang**, R. Heusdens and R. C. Hendriks, Rate-distributed spatial filtering based noise reduction in wireless acoustic sensor networks, *The 8th WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux*, University of Twente, Enschede, the Netherlands, May, 2018.

7. **J. Zhang**, S. P. Chepuri, R. C. Hendriks and R. Heusdens, Microphone subset selection for spatial filtering based noise reduction with multiple target sources, *The 7th WIC/IEEE Symposium on Information Theory and Signal Processing in the Benelux*, Delft University of Technology, Delft, the Netherlands, May, 2017.

2

BACKGROUND

CHAPTER 1 presented a high-level problem description for multi-microphone noise reduction in WASNs and the motivations behind that. In order to concisely look into the different research questions, some preliminaries are required. In this chapter, we therefore give an overview of the background knowledge required to read this thesis, including the signal model, classic spatial filtering techniques, sensor selection, quantization, binaural LCMV beamforming with spatial cue preservation, distributed LCMV beamforming and classic RTF estimation methods. This background knowledge is required for reading the remaining chapters of this dissertation.

2.1. SIGNAL MODEL

We consider a WASN consisting of M microphone nodes that are involved to monitor and sample the sound field of interest. Note that in practice each node can be equipped with a single microphone or a small microphone array. Assume that I target sources and J interfering sources are present in the environment. Let $s_i(t), i = 1, \cdots, I$ and $u_j(t), j = 1, \cdots, J$, respectively, denote the ith target source signal and the jth interfering source signal in the time domain. Due to the presence of reverberation, the source signals propagate to the microphone nodes through a direct path and a series of reflection paths as illustrated in Fig. 2.1. In the time domain the microphone recording $y_k(t)$ can be given by

$$y_k(t) = \sum_{i=1}^{I} \left(s_i * \ddot{a}_{ik}\right)(t) + \sum_{j=1}^{J} \left(u_j * \ddot{h}_{jk}\right)(t) + v_k(t), k = 1, \cdots, M, \tag{2.1}$$

where $*$ denotes convolution, $\ddot{a}_{ik}(t)$ denotes the room impulse response (RIR) from the ith target source location to the kth microphone node, $\ddot{h}_{jk}(t)$ the RIR of the jth interfering source with respect to the kth microphone node, and $v_k(t)$ the spatially uncorrelated noise at the kth microphone node, e.g., sensor-self noise.

In the short-time Fourier transform (STFT) domain, let l denote the frame index and ω the angular frequency index, respectively. Let $S_i(\omega, l)$, $U_j(\omega, l)$ and $V_k(\omega, l)$ denote the

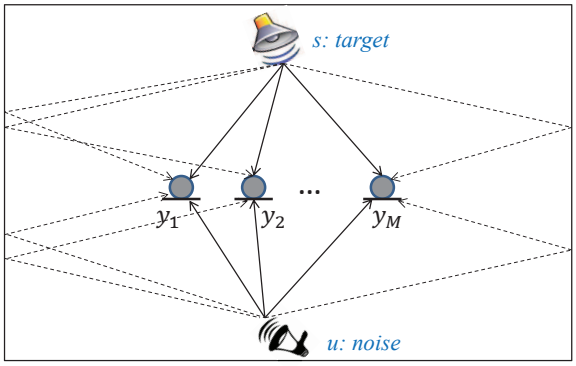


Figure 2.1: An illustrative model for signal acquisition in WASNs: the target and interfering sources propagate to the microphone nodes through a direct path and a couple of reflection paths.

STFT coefficients of $s_i(t)$, $u_j(t)$ and $v_k(t)$, respectively. The corresponding STFT-domain description of the time-domain signal $y_k(t)$ is then given by

$$Y_k(\omega,l)=\sum_{i=1}^{I}S_i(\omega,l)a_{ik}(\omega,l)+\sum_{j=1}^{J}U_j(\omega,l)h_{jk}(\omega,l)+V_k(\omega,l),k=1,\cdots,M, \tag{2.2}$$

where $a_{ik}(\omega,l)$ (or $h_{jk}(\omega,l)$) is the discrete Fourier transform (DFT) of $\ddot{a}_{ik}(t)$ (or $\ddot{h}_{jk}(t)$), which is then called the acoustic transfer function (ATF). Throughout this dissertation, we assume that the ATFs of all existing sources are time-invariant, that is, the ATFs are only frequency dependent, such that the index l can be neglected for $a_{ik}(\omega,l)$ and $h_{jk}(\omega,l)$. This assumption is approximately true in case the sources keep static and the RIRs are shorter than the length of the STFT analysis window. For longer RIRs, e.g., in strong reverberant environments, a more accurate signal model is required. For the sake of notational brevity, we will neglect the frequency index ω and the frame index l in the sequel as all operations are performed per frequency band and per time frame independently.

Using vector notation, we stack for each frequency bin the microphone recordings in an M-dimensional vector $\mathbf{y}=[Y_1,Y_2,\cdots,Y_M]^T\in\mathbb{C}^M$ where $(\cdot)^T$ denotes matrix/vector transposition. Similarly, we define M-dimensional vectors:

$$\mathbf{x}_i=\begin{bmatrix}S_ia_{i1}\\S_ia_{i2}\\\vdots\\S_ia_{iM}\end{bmatrix},\quad \mathbf{a}_i=\begin{bmatrix}a_{i1}\\a_{i2}\\\vdots\\a_{iM}\end{bmatrix},\quad \mathbf{n}_j=\begin{bmatrix}U_jh_{j1}\\U_jh_{j2}\\\vdots\\U_jh_{jM}\end{bmatrix},\quad \mathbf{h}_j=\begin{bmatrix}h_{j1}\\h_{j2}\\\vdots\\h_{jM}\end{bmatrix},\quad \mathbf{v}=\begin{bmatrix}V_1\\V_2\\\vdots\\V_M\end{bmatrix},$$

for the ith target source received by the WASN, the ATFs of the ith target source with respect to the WASN, the jth interfering source received by the WASN, the ATFs of the jth interfering source with respect to the WASN, and the uncorrelated noise components, respectively, such that we can compactly rewrite the signal model in (2.2) as

$$\mathbf{y}=\sum_{i=1}^{I}\mathbf{x}_i+\sum_{j=1}^{J}\mathbf{n}_j+\mathbf{v}, \tag{2.3}$$

where $\mathbf{x}_i = \mathbf{a}_i S_i$ and $\mathbf{n}_j = \mathbf{h}_j U_j$. Further, we can collect the ATFs of the target sources in a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_I] \in \mathbb{C}^{M \times I}$. Similarly, the ATFs of the interfering sources can be collected as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_J] \in \mathbb{C}^{M \times J}$. Then, (2.3) can also be expressed as

$$\mathbf{y} = \underbrace{\mathbf{A}\mathbf{s}}_{\mathbf{x}} + \underbrace{\mathbf{H}\mathbf{u} + \mathbf{v}}_{\mathbf{n}}, \tag{2.4}$$

where $\mathbf{s} = [S_1, S_2, \cdots, S_I]^T \in \mathbb{C}^I$ and $\mathbf{u} = [U_1, U_2, \cdots, U_J]^T \in \mathbb{C}^J$.

We assume that the target sources and the interfering sources are mutually uncorrelated, and the sources are zero-mean, such that the relationship between the correlation matrices can be given by

$$\mathbf{R}_{\mathbf{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R}_{\mathbf{xx}} + \underbrace{\mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{vv}}}_{\mathbf{R}_{\mathbf{nn}}}, \tag{2.5}$$

where

$$\mathbf{R}_{\mathbf{xx}} = \sum_{i=1}^{I} \mathbb{E}\{\mathbf{x}_i \mathbf{x}_i^H\} = \sum_{i=1}^{I} \sigma_{S_i}^2 \mathbf{a}_i \mathbf{a}_i^H = \mathbf{A}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{A}^H, \tag{2.6}$$

$$\mathbf{R}_{\mathbf{uu}} = \sum_{j=1}^{J} \mathbb{E}\{\mathbf{n}_i \mathbf{n}_i^H\} = \sum_{j=1}^{J} \sigma_{U_i}^2 \mathbf{h}_j \mathbf{h}_j^H = \mathbf{H}\mathbf{\Sigma}_{\mathbf{u}}\mathbf{H}^H, \tag{2.7}$$

where $\mathbf{\Sigma}_{\mathbf{x}} = \text{diag}([\sigma_{S_1}^2, \cdots, \sigma_{S_I}^2])$ with $\sigma_{S_i}^2 = \mathbb{E}\{|S_i|^2\}$ denoting the variance of the ith target source at a particular frequency bin, and $\mathbf{\Sigma}_{\mathbf{u}} = \text{diag}([\sigma_{U_1}^2, \cdots, \sigma_{U_J}^2])$ with $\sigma_{U_j}^2 = \mathbb{E}\{|U_j|^2\}$ the variance of the jth interfering source. As the sources are assumed to be zero-mean, $\sigma_{S_i}^2$ (or $\sigma_{U_j}^2$) also represent the power spectral density (PSD) of S_i (or U_j). In (2.5), the second-order statistics (SOS) of all disturbances are included in $\mathbf{R}_{\mathbf{nn}}$. In theory, $\mathbf{R}_{\mathbf{xx}}$ is a rank-I matrix, $\mathbf{R}_{\mathbf{uu}}$ is a rank-J matrix, and $\mathbf{R}_{\mathbf{yy}}$ is full-rank due to the presence of the uncorrelated noise components.

In practice, these correlation matrices can be estimated using average smoothing. Given a voice activity detector (VAD), the microphone recordings can be classified into noise-only segments and speech-plus-noise segments. During the noise-only period, the noise correlation matrix can be estimated, like

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \frac{1}{L_n} \sum_{l=1}^{L_n} \mathbf{n}(l)\mathbf{n}(l)^H. \tag{2.8}$$

Similarly, during the speech-plus-noise period, the noisy correlation matrix can be estimated, like

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \frac{1}{L_y} \sum_{l=1}^{L_y} \mathbf{y}(l)\mathbf{y}(l)^H. \tag{2.9}$$

Note that the L_n segments for estimating $\mathbf{R}_{\mathbf{nn}}$ and the L_y segments for estimating $\mathbf{R}_{\mathbf{yy}}$ are different. After $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$ are estimated, the correlation matrix of the clean signal components can be obtained by subtracting $\hat{\mathbf{R}}_{\mathbf{nn}}$ from $\hat{\mathbf{R}}_{\mathbf{yy}}$, i.e.,

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}}, \tag{2.10}$$

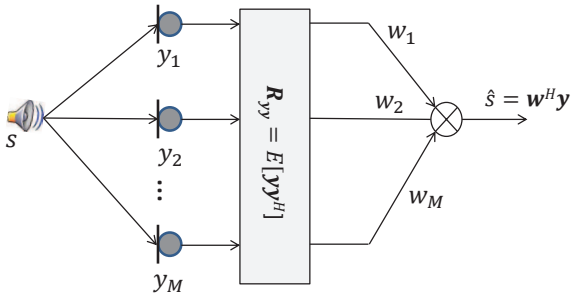


Figure 2.2: The spatial beamforming structure for multichannel noise reduction algorithms.

since $\mathbf{R}_{\mathbf{xx}} \triangleq \mathbf{R}_{\mathbf{yy}} - \mathbf{R}_{\mathbf{nn}}$ by definition. In practice, there are errors in estimating the matrices $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$, leading to a full-rank matrix $\hat{\mathbf{R}}_{\mathbf{xx}}$. Note that a more accurate estimate of $\hat{\mathbf{R}}_{\mathbf{xx}}$ can be obtained using the generalized eigenvalue decomposition (GEVD) of the matrix pencil $(\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}})$ [41].

2.2. SPATIAL FILTERING

Linearly constrained minimum variance (LCMV) beamforming is a well-known and widely-used multichannel spatial filtering technique. The LCMV beamformer can be illustrated by a multiple input single output system as Fig. 2.2 depicts with filter coefficients $\mathbf{w} = [w_1, w_2, \cdots, w_M]^T$. The filter coefficients are adjusted based on the statistics of the noise signal and can be determined by solving the following constrained optimization problem [14, 15, 20]:

$$\mathbf{w}_{\mathrm{LCMV}} = \arg\min_{\mathbf{w}} J(\mathbf{w}), \quad \text{subject to } \mathbf{\Lambda}^H\mathbf{w} = \mathbf{f}, \tag{2.11}$$

where the cost function is given by

$$J(\mathbf{w}) = \mathbb{E}\{|\mathbf{w}^H\mathbf{n}|^2\} = \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w}, \tag{2.12}$$

and $\mathscr{U}$ equality constraints with $\mathbf{f} = [f_1, f_2, \cdots, f_{\mathscr{U}}]^T \in \mathbb{C}^{\mathscr{U}}$ and $\mathbf{\Lambda} \in \mathbb{C}^{M\times\mathscr{U}}$ are taken into account. Applying the technique of Lagrange multipliers, a closed-form solution to (2.11) can be found as

$$\mathbf{w}_{\mathrm{LCMV}} = \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f}. \tag{2.13}$$

The structure of $\mathbf{\Lambda}$ and $\mathbf{f}$ should be specified according to the requirements of the application. For example, in case $\mathbf{\Lambda} = \mathbf{A}$ and $\mathbf{f} = \mathbf{1}_I$ with $\mathbf{1}_I$ denoting an I-dimensional all-ones column vector, the LCMV beamformer will be used to preserve the signals that come from the directions that are characterized by the ATFs in $\mathbf{A}$ and try to suppress the signals that come from all other directions. In a slightly alternative formulation, $\mathbf{\Lambda}$ and $\mathbf{f}$ can also be used to cancel (null) certain interferers, or, to preserve spatial cues in a binaural hearing aid setting [46, 47, 48, 49], which will be discussed in Sec. 2.5 in detail.

After being processed by an LCMV beamformer, the output signal is thus given by

$$\hat{S} = \mathbf{w}_{\mathrm{LCMV}}^H\mathbf{y}. \tag{2.14}$$

and the power (or variance) of the output noise signal can be computed as

$$J(\mathbf{w}) = \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w} = \mathbf{f}^H \left(\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{\Lambda} \right)^{-1} \mathbf{f}, \tag{2.15}$$

and the output signal-to-noise (SNR) can be calculated by

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\mathbf{w}^H \mathbf{R}_{\mathbf{xx}} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}}. \tag{2.16}$$

Remark 1. *In case* $\mathbf{\Lambda} = \mathbf{A}$ *with* $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_I] \in \mathbb{C}^{M \times I}$ *and* $\mathbf{f} = \mathbf{1}_I$ *are used in the general LCMV beamforming problem formulation, that is, the LCMV beamformer is used to exactly preserve the power of the target sources by constraining* $\boldsymbol{A}^H \mathbf{w} = \mathbf{1}_I$*, optimizing (2.11) is equivalent to*

$$\mathbf{w}_{\mathrm{LCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}, \quad \text{subject to } \boldsymbol{A}^H \mathbf{w} = \mathbf{1}_I. \tag{2.17}$$

Suppose the noise signal and the target sources are mutually uncorrelated, the LCMV beamformer is equivalent to the minimum power distortionless response (MPDR) beamformer, which is given by[1]

$$\mathbf{w}_{\mathrm{MPDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{yy}} \mathbf{w}, \quad \text{subject to } \boldsymbol{A}^H \mathbf{w} = \mathbf{1}_I, \tag{2.18}$$

since $\mathbf{w}^H \mathbf{R}_{\mathbf{yy}} \mathbf{w} = \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w} + \mathbf{w}^H \mathbf{R}_{\mathbf{xx}} \mathbf{w} = \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w} + \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{x}})$ *where* $\mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{x}})$ *is a constant that only depends on the power of the source signals with* $\mathrm{Tr}(\cdot)$ *denoting the trace operation. In this case, the LCMV beamformer is given by*

$$\mathbf{w} = \mathbf{R}_{\mathbf{yy}}^{-1} \boldsymbol{A} \left(\boldsymbol{A}^H \mathbf{R}_{\mathbf{yy}}^{-1} \boldsymbol{A} \right)^{-1} \mathbf{1}_I = \mathbf{R}_{\mathbf{nn}}^{-1} \boldsymbol{A} \left(\boldsymbol{A}^H \mathbf{R}_{\mathbf{nn}}^{-1} \boldsymbol{A} \right)^{-1} \mathbf{1}_I, \tag{2.19}$$

and the corresponding output noise power is given by

$$J(\mathbf{w}) = \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w} = \mathbf{1}_I^H \left(\boldsymbol{A}^H \mathbf{R}_{\mathbf{nn}}^{-1} \boldsymbol{A} \right)^{-1} \mathbf{1}_I. \tag{2.20}$$

Furthermore, the output SNR can be derived as

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\mathbf{w}^H \mathbf{R}_{\mathbf{xx}} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}} = \frac{\mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{x}})}{\mathbf{1}_I^H \left(\boldsymbol{A}^H \mathbf{R}_{\mathbf{nn}}^{-1} \boldsymbol{A} \right)^{-1} \mathbf{1}_I}, \tag{2.21}$$

since we have $\mathbf{w}^H \mathbf{R}_{\mathbf{xx}} \mathbf{w} = \mathbf{w}^H \mathbf{A} \mathbf{\Sigma}_{\mathbf{x}} \mathbf{A}^H \mathbf{w} = \mathbf{1}_I^H \mathbf{\Sigma}_{\mathbf{x}} \mathbf{1}_I = \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{x}})$.

Remark 2. *The well-known minimum variance distortionless response (MVDR) beamformer is a special case of the LCMV beamformer. Suppose that there is only one source of interest which is characterized by the ATF vector* $\mathbf{a}$*. As we wish to only preserve the power of*

[1] Strictly speaking, minimizing $\mathbf{w}^H \mathbf{R}_{\mathbf{yy}} \mathbf{w}$ is not equivalent to minimizing $\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}$ under the same linear constraint, due to the estimation errors in the matrices $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{nn}}$. Here, we assume that the statistics are perfectly estimated, such that they are equivalent.

this source and cancel out all the other existing sources, the LCMV beamforming problem in (2.11) can be reformulated into the following special case

$$\mathbf{w}_{\text{MVDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}, \quad \text{subject to } \mathbf{a}^H \mathbf{w} = 1. \tag{2.22}$$

Applying the method of Lagrange multipliers, the solution of the MVDR filter is given by

$$\mathbf{w}_{\text{MVDR}} = \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{a} \left(\mathbf{a}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{a} \right)^{-1}. \tag{2.23}$$

Similarly, after being filtered by an MVDR beamformer, the output SNR can be shown as

$$\text{SNR}_{\text{out}} = \frac{\mathbf{w}^H \mathbf{R}_{\mathbf{xx}} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}} = \frac{\sigma_S^2}{\left(\mathbf{a}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{a} \right)^{-1}} = \sigma_S^2 \mathbf{a}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{a}, \tag{2.24}$$

with the output noise power $\mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w} = \left(\mathbf{a}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{a} \right)^{-1}$ *and* σ_S^2 *the PSD of the single target source.*

2.3. SENSOR SELECTION MODEL

Sensor selection is a sparse sensing techniques for signal inference [50, 51], which chooses a subset of sensors from a much larger set. From the perspective of signal acquisition, it chooses a subset of measurements corresponding to the selected sensors. Given M sensors, sensor selection can be realized by designing a selection vector

$$\mathbf{p} = [p_1, p_2, \cdots, p_M]^T \in \{0,1\}^M, \tag{2.25}$$

where $p_k = 1, \forall k$ indicates that the kth sensor is selected, and $p_k = 0$ means that the kth sensor is not selected. We can use $K = ||\mathbf{p}||_0$ to represent the number of the selected sensors with the ℓ_0-(quasi) norm denoting the number of non-zero entries in vector $\mathbf{p}$.

Let diag($\mathbf{p}$) denote a diagonal matrix whose diagonal elements are given by $\mathbf{p}$. With the matrix diag($\mathbf{p}$) at hand, we can further construct a selection matrix $\mathbf{\Phi}_{\mathbf{p}} \in \{0,1\}^{K \times M}$ which is obtained by removing the all-zero rows of diag($\mathbf{p}$), i.e., the rows corresponding to the unselected sensors and the fat matrix $\mathbf{\Phi}_{\mathbf{p}}$ consisting of a subset of the rows of diag($\mathbf{p}$). Based on this construction, we can see that the following two properties hold

$$\mathbf{\Phi}_{\mathbf{p}}^T \mathbf{\Phi}_{\mathbf{p}} = \text{diag}(\mathbf{p}), \quad \mathbf{\Phi}_{\mathbf{p}} \mathbf{\Phi}_{\mathbf{p}}^T = \mathbf{I}_K, \tag{2.26}$$

where $\mathbf{I}_K$ is a K-dimensional identity matrix.

Considering the original signal model in (2.4) and using the selection matrix, the selected sensor measurements can be given by

$$\mathbf{y}_{\mathbf{p}} = \mathbf{\Phi}_{\mathbf{p}} \mathbf{y} = \mathbf{\Phi}_{\mathbf{p}} \mathbf{x} + \mathbf{\Phi}_{\mathbf{p}} \mathbf{n}. \tag{2.27}$$

Further, the correlation matrices of the selected measurements can be written as

$$\mathbf{R}_{\mathbf{yy},\mathbf{p}} = \mathbb{E}\{\mathbf{\Phi}_{\mathbf{p}} \mathbf{y} \mathbf{y}^H \mathbf{\Phi}_{\mathbf{p}}^H\} = \mathbf{\Phi}_{\mathbf{p}} \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} \mathbf{\Phi}_{\mathbf{p}}^H = \mathbf{\Phi}_{\mathbf{p}} \mathbf{R}_{\mathbf{yy}} \mathbf{\Phi}_{\mathbf{p}}^H. \tag{2.28}$$

Similarly, we can see that

$$\mathbf{R}_{\mathbf{xx},\mathbf{p}} = \mathbf{\Phi}_\mathbf{p}\mathbf{R}_{\mathbf{xx}}\mathbf{\Phi}_\mathbf{p}^H, \quad \mathbf{R}_{\mathbf{nn},\mathbf{p}} = \mathbf{\Phi}_\mathbf{p}\mathbf{R}_{\mathbf{nn}}\mathbf{\Phi}_\mathbf{p}^H. \tag{2.29}$$

This reveals that $\mathbf{R}_{\mathbf{yy},\mathbf{p}}$, $\mathbf{R}_{\mathbf{xx},\mathbf{p}}$ and $\mathbf{R}_{\mathbf{nn},\mathbf{p}}$ are the sub-matrices of $\mathbf{R}_{\mathbf{yy}}$, $\mathbf{R}_{\mathbf{xx}}$ and $\mathbf{R}_{\mathbf{nn}}$, respectively. For instance, $\mathbf{R}_{\mathbf{yy},\mathbf{p}}$ can be obtained by removing the rows and columns of $\mathbf{R}_{\mathbf{yy}}$ corresponding to the unselected sensors.

The considered sensor selection model can be taken into account for the spatial filtering techniques in Sec. 2.2. For example, the classical MVDR beamforming problem can be extended as

$$\begin{aligned}\mathbf{w}_\mathbf{p} &= \arg\min_{\mathbf{w}_\mathbf{p}} \mathbf{w}_\mathbf{p}^H\mathbf{R}_{\mathbf{nn},\mathbf{p}}\mathbf{w}_\mathbf{p} \\ &= \arg\min_{\mathbf{w}_\mathbf{p}} \mathbf{w}_\mathbf{p}^H\mathbf{\Phi}_\mathbf{p}\mathbf{R}_{\mathbf{nn}}\mathbf{\Phi}_\mathbf{p}^H\mathbf{w}_\mathbf{p}, \quad \text{subject to } \mathbf{a}_\mathbf{p}^H\mathbf{w}_\mathbf{p} = 1,\end{aligned} \tag{2.30}$$

where $\mathbf{a}_\mathbf{p} = \mathbf{\Phi}_\mathbf{p}\mathbf{a}$ denotes the ATF vector for the selected sensors. Following the derivation of the classic MVDR beamformer, the sensor selection based MVDR beamformer is given by

$$\mathbf{w}_\mathbf{p} = \frac{\mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1}\mathbf{a}_\mathbf{p}}{\mathbf{a}_\mathbf{p}^H\mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1}\mathbf{a}_\mathbf{p}}, \tag{2.31}$$

and the output noise power is given by

$$\mathbf{w}_\mathbf{p}^H\mathbf{R}_{\mathbf{nn},\mathbf{p}}\mathbf{w}_\mathbf{p} = \left(\mathbf{a}_\mathbf{p}^H\mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1}\mathbf{a}_\mathbf{p}\right)^{-1}. \tag{2.32}$$

Note that in general $\mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \neq \mathbf{\Phi}_\mathbf{p}\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Phi}_\mathbf{p}^H$, so $\mathbf{w}_\mathbf{p} \neq \mathbf{\Phi}_\mathbf{p}\mathbf{w}_{\text{MVDR}}$. Only in case the noise components across microphones are mutually uncorrelated, i.e., $\mathbf{R}_{\mathbf{nn}}$ is a diagonal matrix, we will have $\mathbf{w}_\mathbf{p} = \mathbf{\Phi}_\mathbf{p}\mathbf{w}_{\text{MVDR}}$.

In conventional sensor selection problems, the cardinality of the selected sensors is of more interest [52, 53, 51, 50, 54, 55]. Usually, the selection strategy is designed by minimizing the cardinality and constraining the inference performance, as the following optimization problem shows

$$\begin{aligned}\underset{\mathbf{w}_\mathbf{p},\mathbf{p}\in\{0,1\}^M}{\text{minimize}} \quad & ||\mathbf{p}||_0 \\ \text{subject to} \quad & g(\mathbf{w}_\mathbf{p},\mathbf{p}) \leq \beta,\end{aligned} \tag{2.33}$$

which is a special case of (1.1). Given the cardinality of $\mathbf{p}$, i.e., it is known how many sensors should be involved, and (2.33) can be reformulated equivalently by interchanging the objective and the constraint function as

$$\begin{aligned}\underset{\mathbf{w}_\mathbf{p},\mathbf{p}\in\{0,1\}^M}{\text{minimize}} \quad & g(\mathbf{w}_\mathbf{p},\mathbf{p}) \\ \text{subject to} \quad & ||\mathbf{p}||_0 = K\end{aligned} \tag{2.34}$$

Both (2.33) and (2.34) are non-convex combinatorial optimization problems, which is caused by the Boolean constraint on the selection variable $\mathbf{p}$, i.e., the cardinality function

$||\mathbf{p}||_0$ is non-convex in $\mathbf{p}$. One way to approach the optimal solution is by evaluating the performance of all the $\binom{M}{K}$ possible combinations. Obviously, this is computationally intractable unless K and M are small.

As we stated in Chapter 1, in large-scale WASNs the cardinality of the selected sensors is of less interest, since it might be the case that we even do not know how many sensors are available. Instead, saving the total power consumption is our goal. In many applications, it is more natural to provide a certain expected performance, e.g., a certain output noise power, a certain speech recognition precision, or a certain predicted speech intelligibility performance. Hence, following the general problem description in (1.1), we can reformulate the sensor selection for MVDR beamforming based noise reduction problem as

$$\begin{aligned} \underset{\mathbf{w_p},\mathbf{p}\in\{0,1\}^M}{\text{minimize}} \quad & f(\mathbf{w_p},\mathbf{p}) = \sum_{k=1}^{M} p_k c_k \\ \text{subject to} \quad & g(\mathbf{w_p},\mathbf{p}) = \left(\mathbf{a}_\mathbf{p}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{a}_\mathbf{p}\right)^{-1} \leq \beta, \end{aligned} \tag{2.35}$$

where c_k denotes the transmission power from the kth sensor to a fusion center (FC). Again, our reformulated sensor selection problem is still non-convex essentially due to the Boolean constraint.

In order to solve the aforementioned sensor selection problems, there are two approaches that can be applied: *model-driven schemes* and *data-driven schemes*. The model-driven methods, e.g., [52, 53, 51, 50, 55], are based on the use of the statistics of the complete network which needs to be estimated before online data gathering, such that the convex optimization techniques can be applied to find the optimal subset. That is, the model-driven sensor selection can be regarded as an offline design. The selected subset of sensors is thus *a priori* knowledge of locations where the sensors should be placed, i.e., *sensor placement*, such that a prescribed estimation performance is guaranteed. However, in many applications the statistics of the measurements of the complete network is not always available, so that data-driven methods, e.g., *greedy approaches* [54, 56, 57, 58, 59, 60], should be considered for searching a near-optimal solution. In case the cost function in the sensor selection problems is submodular [54, 58, 61, 62, 63], submodularity-based greedy optimization can be exploited. If the submodularity cannot be leveraged, utility-based greedy methods can be used [59, 60, 64], e.g., at each step by adding the sensor which has the largest contribution to improve the output SNR or by removing the sensor that has the least contribution.

2.4. UNIFORM QUANTIZATION

In this work we consider (noisy) microphone recordings that are sampled (i.e., discretized) and subsequently quantized to a discrete set of levels. This leads to the introduction of quantization noise. As we study in this work the trade-off between performance and energy consumption for beamforming in WASNs, we do not only consider transmission of information at the maximum bit-rate, but also use more coarse quantizers. This will introduce additional noise (quantization noise) into the beamforming problem. In this section we therefore give a brief overview on the effect of quantization on the signal

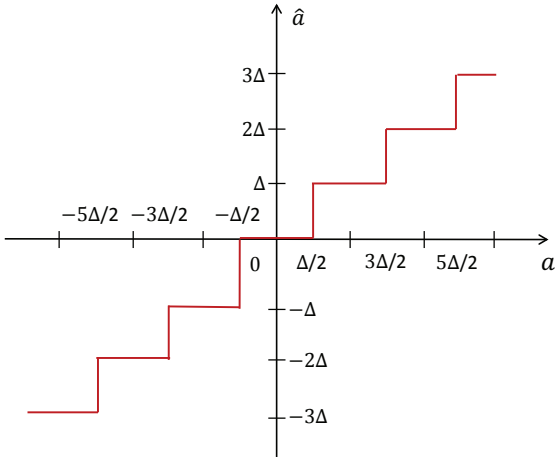


Figure 2.3: A two dimensional representation of a uniform quantizer with eight fixed quantization cells.

model introduced in Sec. 2.1.

Let a be a realization of a continuous random variable A. Any a is quantized to a value $\hat{a} = Q(a)$, where $\hat{a}$ is a realization of random variable $\hat{A}$, the reproduction of A. This leads to quantization noise $n_q = \hat{a} - a$. Given b bits per sample that can be used to quantize a and suppose the random variable A is bounded by $a \in [-\mathscr{A}/2, \mathscr{A}/2]$, i.e., $\mathscr{A}$ denotes the range of A, we can divide the signal range into 2^b quantization cells. With these cells, any value of a can be mapped. Note that the quantization cells can be divided uniformly or non-uniformly, leading to a *uniform quantizer* or *non-uniform quantizer*, respectively [65, 66]. In this dissertation, we will only consider the use of uniform quantizers for quantizing the microphone recordings. Fig. 2.3 shows a two dimensional representation of a uniform quantizer with eight fixed cells.

Suppose the kth sensor uses b_k bits per sample for quantizing its measurements and $y_k \in [-\mathscr{A}_k/2, \mathscr{A}_k/2]$ with $\mathscr{A}_k/2$ denoting the maximum absolute value of the kth microphone signal. Under the utilization of a uniform quantizer[2], we can then construct 2^{b_k} uniform intervals (or cells) which have a width

$$\Delta_k = \frac{\mathscr{A}_k}{2^{b_k}}, k = 1, 2, \ldots, M. \tag{2.36}$$

The reproduction of y_k is then given by

$$\hat{y}_k = \Delta_k \times \text{round}\left(\frac{y_k}{\Delta_k}\right), \quad k = 1, 2, \ldots, M, \tag{2.37}$$

where round$(\cdot)$ returns the nearest integer of its argument, and the quantization noise is

[2] In practice, the microphone recordings are already quantized, since they are sensed by the ADCs. Here, we introduce this secondary quantization, such that the transmission energy from sensors to the FC can be decreased compared to merely transmitting the raw full-rate data. In this case, this quantization noise represents the error from changing the bit resolution.

given by

$$q_k = \hat{y}_k - y_k. \tag{2.38}$$

As the quantization noise is taken into account, the signal models presented in Sec. 2.1 need to be modified by adding an additional variable $q_k(\omega, l)$ or a vector $\mathbf{q} = [q_1, \ldots, q_M]^T$. Note that the signal range $\mathcal{A}_k$ might be different from sensor-to-sensor which is only related to its own signal observations.

Further, the PSD or variance of the quantization noise is given by [67, 68, 69]

$$\sigma^2_{q_k} = \frac{\Delta_k^2}{12} = \frac{1}{12} \times \left(\frac{\mathcal{A}_k}{2^{b_k}}\right)^2, \quad k = 1, 2, \ldots, M, \tag{2.39}$$

and the correlation matrix of the quantization noise across microphones is given by

$$\mathbf{R_{qq}} = \text{diag}\left(\sigma^2_{q_1}, \sigma^2_{q_2}, \ldots, \sigma^2_{q_M}\right). \tag{2.40}$$

Assuming that the quantization noise and the acoustic signals are mutually uncorrelated[3], the correlation matrix of the quantized microphone signals $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M]^T$ reads

$$\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbb{E}\{\hat{\mathbf{y}}\hat{\mathbf{y}}^H\} = \mathbf{R_{xx}} + \underbrace{\mathbf{R_{nn}} + \mathbf{R_{qq}}}_{\mathbf{R_{n+q}}}. \tag{2.41}$$

Similar to the estimation of the matrices $\mathbf{R_{yy}}$ and $\mathbf{R_{nn}}$, we can estimate $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R_{n+q}}$ using the sample correlation matrices during the quantized speech-plus-noise segments and the quantized noise-only segments, respectively. Since both $\mathbf{R_{n+q}}$ and $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ include the statistics of the quantization noise, given sufficiently long noise and noisy time intervals, the quantization noise will affect $\mathbf{R_{n+q}}$ and $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ in the same fashion by adding the same matrix $\mathbf{R_{qq}}$. As a consequence, the quantization noise will not affect the estimation of $\mathbf{R_{xx}}$ due to the subtraction operation as long as the quantization rate does not change. This is the basic motivation on how to estimate the RTFs under low communication rates, which is presented in Chapter 7.

2.5. BINAURAL LCMV BEAMFORMING

In Chapter 6, we consider the application of the presented theory on rate-distributed spatial filtering in WASNs in Chapter 4 involving a pair of hearing aids (HAs). We therefore introduce in this section the binaural LCMV (BLCMV) beamformer.

For the HA users, apart from noise reduction, spatial awareness of the interfering sources is necessary. Identifying the location of the interfering sources (they could be important warning signals) for normal hearing people is very natural, while it is hard for hearing-impaired listeners. As the source location is characterized by the spatial cue information, spatial cue preservation is thus of great significance for HA users. Making use of the measurements from the external devices in WASNs can improve the performance of such joint noise reduction and spatial cue preservation for HA applications.

[3] This assumption is true for high-rate quantization. For low rate, quantization subtractive dithering can be used to meet the correlation assumptions [70, 71, 72]. The dither signal which is known at the receiver side and the quantization noise are independent and identically distributed (i.i.d.) processes.

In the frequency domain the spatial cues of a source include the interaural phase difference (IPD) [73], interaural level difference (ILD) [73] and interaural coherence (IC) [74]. The IPD is caused by the interaural time difference (ITD), which measures the time difference of arrival (TDOA) of the source between two ears. The ILD measures the magnitude difference between binaural signals because of the acoustical shadowing effect of the head. The IPD (or ITD) and ILD are directional spatial cues, which are frequently-used in binaural source localization algorithms [75, 76, 77, 6, 78]. In general, the IPD (or ITD) based localization has ambiguity for high frequency bands (e.g., above 1.5 kHz) due to phase unwrapping [76, 78], while the ILD-based localization has large estimation variance at low frequency bands, so usually they are jointly used in practice. The IC information is important for determining the width of sound fields.

In general, binaural cue preservation is achieved by sacrificing the noise reduction capability. In other words, the more spatial cues are being preserved, the less degrees-of-freedom are left for noise reduction. For directional sources, spatial cue preservation can be achieved by preserving the interaural transfer function (ITF) which is defined as the ratio of the ATFs relating the source and the two ears, since ILD, IPD and ITD are the magnitude response, phase response and group delay of the ITF, respectively. In order to jointly suppress noise sources and preserve the spatial cues, the multi-microphone spatial filtering techniques mentioned in Sec. 1.1 can still be used by adding more constraints. For a single target source, the binaural MVDR (BMVDR) beamformer is a natural extension of the traditional MVDR beamformer in the binaural setup [79, 80]. For the BMVDR beamforming, usually each HA has to transmit its measurements via a wireless link to the other HA, such that more data are available at each device for obtaining a better noise reduction performance, and two BMVDR beamformers are then computed at the two ears, respectively. Since only two linear constraints associated with the target source are taken into account, the BMVDR beamformer can only preserve the binaural cues of the target source. On the other hand, the BMVDR beamformer has more degrees-of-freedom for filter design and can obtain a better noise reduction performance than other linearly-constrained binaural filters. In order to further preserve the binaural cues of the interfering sources, one can add more linear constraints associated with the ITFs of the interfering sources to the BMVDR beamformers [80]. In case there are multiple target sources of interest, the binaural LCMV (BLCMV) beamformer (which, similarly, is an extension of the conventional LCMV beamforming method in the binaural context) is an alternative solution [46, 47]. The MWF can also be extended to the binaural setting to preserve the binaural cues, e.g., by enforcing the constraints associated with the ITFs of all the sources of interest [48]. Again, the binaural MWF (BMWF) would distort the target source(s) inevitably. Moreover, the IC can also be preserved by using binaural filtering techniques [46, 81]. For most binaural filtering algorithms, the most essential challenge is how to obtain an acceptable noise reduction performance versus spatial cue preservation trade-off.

The BLCMV beamformer is an extension of the classic LCMV beamformer in the binaural setting (e.g., for binaural hearing-aids) for jointly performing noise suppression and binaural cue preservation of all the present point sources [46, 47]. Given M sensors, the BLCMV beamformer is defined as the concatenation of two LCMV beamformers at

the two HAs, i.e.,

$$\mathbf{w}_{\text{BLCMV}} = [\mathbf{w}_L^T, \mathbf{w}_R^T]^T \in \mathbb{C}^{2M}, \tag{2.42}$$

where L and R are used to indicate the left and the right ear beamformer, respectively. Suppose one of the HAs is the FC, all the sensors will send their measurements to this HA, and the noise statistics $\mathbf{R_{nn}} \in \mathbb{C}^{M\times M}$ can be estimated at this HA. Due to the fact that there exists a wireless link between the two HAs, one HA can also send the information of the noise statistics to the other HA [70, 82, 83], such that the final noise correlation matrix can be constructed as

$$\tilde{\mathbf{R}}_{\mathbf{nn}} = \begin{bmatrix} \mathbf{R_{nn}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R_{nn}} \end{bmatrix} \in \mathbb{C}^{2M\times 2M}. \tag{2.43}$$

The BLCMV beamformer $\mathbf{w}_{\text{BLCMV}}$ is thus a $2M$-dimensional vector. Preserving the binaural cues (e.g., ILD, IPD) can be achieved by preserving the ITF, since the ILD and IPD can be derived from the ITF as

$$\text{ILD} = |\text{ITF}|^2, \quad \text{IPD} = \angle\text{ITF}, \tag{2.44}$$

where $\angle$ denotes the argument or angle. Given I target sources and J interfering sources, the ITFs of the existing sources with respect to the two reference microphones are defined by

$$\text{ITF}^{\text{in}}_{\mathbf{x}_i} = \frac{a_{iL}}{a_{iR}}, \forall i; \qquad \text{ITF}^{\text{in}}_{\mathbf{n}_j} = \frac{h_{iL}}{h_{iR}}, \forall j, \tag{2.45}$$

where L and R are used for indicating the reference microphones at the two ears.

For the I target sources, we want not only to preserve their spatial cues, but also to keep them undistorted. To do this, we can constrain the two LCMV beamformers by

$$\mathbf{w}_L^H \mathbf{a}_i = a_{iL}, \quad \mathbf{w}_R^H \mathbf{a}_i = a_{iR}. \tag{2.46}$$

Based on these two constraints, the spatial cues of the target sources are perserved, since the input and output ITFs are identical, i.e.,

$$\text{ITF}^{\text{out}}_{\mathbf{x}_i} = \frac{\mathbf{w}_L^H \mathbf{a}_i}{\mathbf{w}_R^H \mathbf{a}_i} = \frac{a_{iL}}{a_{iR}} \Longrightarrow \text{ITF}^{\text{in}}_{\mathbf{x}_i} = \text{ITF}^{\text{out}}_{\mathbf{x}_i}, i = 1, 2, \ldots, I. \tag{2.47}$$

In addition, the power of the output source signals are given by

$$\mathbf{w}_L^H \mathbf{R_{xx}} \mathbf{w}_L = \sum_{i=1}^{I} |a_{iL}|^2 \sigma_{S_i}^2, \quad \mathbf{w}_R^H \mathbf{R_{xx}} \mathbf{w}_R = \sum_{i=1}^{I} |a_{iR}|^2 \sigma_{S_i}^2, \tag{2.48}$$

which is the clean signal power at the reference microphones. This differs from the classic LCMV beamformer which preserves the power of the original source signals, as shown in Remark 1. Combining the constraints in (2.46) for all the I target sources, we can compactly express the linear constraints as

$$\mathbf{\Lambda}_1^H \mathbf{w} = \mathbf{f}_1, \tag{2.49}$$

where

$$\mathbf{\Lambda}_1 = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix} \in \mathbb{C}^{2M\times 2I}, \quad \mathbf{f}_1 = \left[a_{1L},\ldots,a_{IL},a_{1R},\ldots,a_{IR}\right]^H \in \mathbb{C}^{2I}.$$

On the other hand, for the J interfering sources, we only need to preserve their binaural cues (ideally, not audible in the output of the BLCMV beamformer), as the binaural cues are essential for localizing the sources. To obtain this, the BLCMV beamformers have to satisfy the constraint

$$\mathrm{ITF}^{\mathrm{in}}_{\mathbf{n}_j} = \mathrm{ITF}^{\mathrm{out}}_{\mathbf{n}_j} \Longrightarrow \frac{h_{jL}}{h_{jR}} = \frac{\mathbf{w}_L^H\mathbf{h}_j}{\mathbf{w}_R^H\mathbf{h}_j}, j = 1,\cdots,J. \tag{2.50}$$

Note that for each interfering source, one constraint is required, while for each target source two linear constraints are required as shown in (2.46). Further, (2.50) can be rewritten into the following linear equality form:

$$\mathbf{w}_L^H\mathbf{h}_j h_{jR} - \mathbf{w}_R^H\mathbf{h}_j h_{jL} = 0, j = 1,\cdots,J. \tag{2.51}$$

Combining (2.51) for all interfering sources, we can compactly express the linear equality constraints for preserving their binaural cues as

$$\mathbf{\Lambda}_2^H\mathbf{w} = \mathbf{f}_2, \tag{2.52}$$

where

$$\mathbf{\Lambda}_2 = \begin{bmatrix} \mathbf{h}_1 h_{1R} & \cdots & \mathbf{h}_J h_{JR} \\ -\mathbf{h}_1 h_{1L} & \cdots & -\mathbf{h}_J h_{JL} \end{bmatrix} \in \mathbb{C}^{2M\times J}, \quad \mathbf{f}_2 = [0,0,\ldots,0]^T \in \mathbb{R}^J.$$

To this end, we can see that given I target sources and J interfering sources, for the design of BLCMV beamformers, $2I+J$ linear constraints need to be satisfied, which can be written in a more compact form by combining (2.49) and (2.52) together as

$$\mathbf{\Lambda}^H\mathbf{w} = \tilde{\mathbf{f}}, \tag{2.53}$$

where

$$\mathbf{\Lambda} = \left[\begin{array}{c|c} \mathbf{\Lambda}_1 & \mathbf{\Lambda}_2 \end{array}\right] \in \mathbb{C}^{2M\times(2I+J)}, \quad \tilde{\mathbf{f}} = \left[\begin{array}{c|c} \mathbf{f}_1^T & \mathbf{f}_2^T \end{array}\right]^T \in \mathbb{C}^{2I+J}.$$

As a consequence, the general BLCMV beamforming problem for joint noise reduction and spatial cue preservation can mathematically be formulated as

$$\mathbf{w}_{\mathrm{BLCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\tilde{\mathbf{R}}_{\mathbf{nn}}\mathbf{w}, \quad \text{subject to} \quad \mathbf{\Lambda}^H\mathbf{w} = \tilde{\mathbf{f}}. \tag{2.54}$$

Assuming that the matrix $\mathbf{\Lambda}$ is full-column rank (i.e., $\mathrm{rank}(\mathbf{\Lambda}) = 2I+J$)[4], in case $2M \geq 2I+J$, the linear system $\mathbf{\Lambda}^H\mathbf{w} = \tilde{\mathbf{f}}$ is underdetermined. There are $2I+J$ degrees-of-freedom (DOF) dedicated to spatial cue preservation, and $2M-2I-J$ DOF left for reducing the objective function (i.e., performing noise reduction) by adjusting the filter

[4] This is true when all the existing sources have different ATFs (e.g., in different directions).

coefficients. Using the method of Lagrange multipliers, we can find a closed-form solution to the BLCMV problem as

$$\mathbf{w}_{\text{BLCMV}} = \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}\right)^{-1}\tilde{\mathbf{f}} \in \mathbb{C}^{2M}, \quad \text{for} \quad 2M \geq 2I+J. \tag{2.55}$$

Obviously, the more DOF focus on preserving spatial cues, the less DOF are left for noise reduction, leading to a trade-off between noise reduction and spatial cue preservation. Moreover, in case $2M = 2I+J$, i.e., $\mathbf{\Lambda}$ is a full-rank square matrix, the linear system $\mathbf{\Lambda}^H\mathbf{w} = \tilde{\mathbf{f}}$ is determined, and the unique solution is given by

$$\mathbf{w}_{\text{BLCMV}} = \mathbf{\Lambda}^{-H}\tilde{\mathbf{f}} \in \mathbb{C}^{2M}, \quad \text{for} \quad 2M = 2I+J. \tag{2.56}$$

However, in this case there will be no DOF left for controlling the output noise power. In case $2M \leq 2I+J$, the linear system $\mathbf{\Lambda}^H\mathbf{w} = \tilde{\mathbf{f}}$ is overdetermined. In this dissertation, we will stick to the case when $2M \geq 2I+J$ to reach the goal of joint noise reduction and spatial cue preservation. After being filtered by the BLCMV beamformer in (2.55), the output noise power is given by

$$\mathbf{w}^H\tilde{\mathbf{R}}_{\mathbf{nn}}\mathbf{w} = \tilde{\mathbf{f}}^H\left(\mathbf{\Lambda}^H\tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}\right)^{-1}\tilde{\mathbf{f}}. \tag{2.57}$$

2.6. DISTRIBUTED SPATIAL FILTERING

In order to improve the robustness of the spatial filtering techniques against the variation and scalability of the network topology, a distributed implementation is required. To do this, we first model the WASN as a graph $\mathcal{G} = (\mathcal{K}, \mathcal{E})$, where $\mathcal{K} = \{1, 2, \ldots, K\}$ denotes the set of nodes (devices) and $\mathcal{E}$ the set of edges (i.e., wireless links) between the nodes. If and only if $(i, j) \in \mathcal{E}$, the ith and jth nodes can directly communicate with each other. We assume that the WASN is a connected graph, that is, there always exists a route from one node to any other node in the graph. Suppose each node k consists of M_k microphones, we thus have $M = \sum_{k=1}^{K} M_k$ microphones in total. Let $\mathcal{N}_k$ denote the set which contains all the neighboring nodes of node k but does not include node k itself.

2.6.1. DISTRIBUTED LCMV BEAMFORMING

In a reverberant environment, the reverberation of a sound source consists of early reflections and the late reverberations. Only the early reverberation is useful for enhancing the speech intelligibility [84]. This means that the noise component $\mathbf{n}$ that was presented in Sec. 2.1 can be a summation of the late reverberation of the target sources, and the early and late reverberation of the interfering sources and the uncorrelated noise. Further, in [85], it was shown that the late reverberation is highly correlated across the microphones within a node, while it is much less correlated across the microphones at different nodes, since the microphones within a node are spatially close and the microphones at different nodes are more distant. Due to this, we use $\mathbf{R}_{\mathbf{zz}}$ to represent the statistics of the early reflections of the interfering sources, and $\mathbf{R}_{\mathbf{uu}}$ for the statistics of all the late reverberation of all sources and the uncorrelated sensor noise. More importantly, it was shown in [85] that $\mathbf{R}_{\mathbf{uu}}$ can be approximated by a *block-diagonal* matrix.

Assuming that the early reverberation and the late reverberation are mutually uncorrelated, the correlation matrix of all noise components can be given by

$$\mathbf{R}_{\mathbf{nn}} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\} = \underbrace{\mathbf{R}_{\mathbf{zz}}}_{\text{corr}} + \underbrace{\mathbf{R}_{\mathbf{uu}}}_{\text{uncorr}}. \tag{2.58}$$

The classic LCMV beamforming problem in Sec. 2.2 can be reformulated as

$$\begin{aligned}\mathbf{w} &= \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w}, \quad \text{subject to } \mathbf{\Lambda}^H \mathbf{w} = \mathbf{f} \\ &= \arg\min_{\mathbf{w}} \mathbf{w}^H \left(\mathbf{R}_{\mathbf{zz}} + \mathbf{R}_{\mathbf{uu}}\right) \mathbf{w}, \quad \text{subject to } \mathbf{\Lambda}^H \mathbf{w} = \mathbf{f} \\ &\approx \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{uu}} \mathbf{w}, \quad \text{subject to } \mathbf{\Lambda}^H \mathbf{w} = \mathbf{f},\end{aligned} \tag{2.59}$$

where $\approx$ is due to the fact that given enough DOFs if the linear system $\mathbf{\Lambda}^H \mathbf{w} = \mathbf{f}$ is exactly satisfied, the filter $\mathbf{w}$ is orthogonal to the ATFs of the interfering sources. With such an LCMV beamformer, the correlated noise components can entirely be suppressed, resulting in $\mathbf{w}^H \mathbf{R}_{\mathbf{zz}} \mathbf{w} = 0$. Hence, we can simply use $\mathbf{R}_{\mathbf{uu}}$ for the filter design.

In order to solve the centralized LCMV beamforming problem given by (2.59) in a distributed fashion, we first need to write it into a node-separable form. To do this, we split the filter vector $\mathbf{w}$, the matrix $\mathbf{R}_{\mathbf{uu}}$, the ATF matrix $\mathbf{\Lambda}$ over nodes as [85]

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_K \end{bmatrix}, \quad \mathbf{R}_{\mathbf{uu}} = \begin{bmatrix} \mathbf{R}_{\mathbf{u},1} & & & \\ & \mathbf{R}_{\mathbf{u},1} & & \\ & & \ddots & \\ & & & \mathbf{R}_{\mathbf{u},K} \end{bmatrix}, \quad \mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \\ \vdots \\ \mathbf{\Lambda}_K \end{bmatrix},$$

where $\mathbf{w}_k \in \mathbb{C}^{M_k}$, $\mathbf{R}_{\mathbf{u},k} \in \mathbb{C}^{M_k \times M_k}$ and $\mathbf{\Lambda}_k \in \mathbb{C}^{M_k \times \mathscr{U}}$, such that (2.59) can equivalently be rewritten as

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{k=1}^{K} \mathbf{w}_k^H \mathbf{R}_{\mathbf{u},k} \mathbf{w}_k, \quad \text{subject to } \sum_{k=1}^{K} \mathbf{\Lambda}_k^H \mathbf{w}_k = \mathbf{f}, \tag{2.60}$$

where both the objective function and the constraint are separated in nodes. In general, the optimal solution of an optimization problem, e.g., (2.60), cannot be approached by optimizing the node-specific sub-problems locally, unless the sub-problems are independent to each other. Clearly, this is not the case when considering (2.60). In order to solve (2.60) in a distributed fashion, we consider the real-valued Lagrangian function of (2.60), which is given by

$$L(\mathbf{w}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \left[\mathbf{w}_k^H \mathbf{R}_{\mathbf{u},k} \mathbf{w}_k - 2\Re\left(\boldsymbol{\mu}^H \left(\mathbf{\Lambda}_k^H \mathbf{w}_k - \frac{\mathbf{f}}{K} \right) \right) \right], \tag{2.61}$$

where $\boldsymbol{\mu} \in \mathbb{C}^{\mathscr{U}}$ is a vector with Lagrangian multipliers, $\Re(\cdot)$ returns the real part, and the vector $\mathbf{f}$ is partitioned into K equal parts. Since $L(\mathbf{w}, \boldsymbol{\mu})$ is convex in terms of $\mathbf{w}$, the optimal filter is the minimizer of $L(\mathbf{w}, \boldsymbol{\mu})$, i.e.,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\mu}). \tag{2.62}$$

Then, by setting the derivative of $L(\mathbf{w}, \boldsymbol{\mu})$ with respect to $\bar{\mathbf{w}}_k$ to zero (where $\bar{\mathbf{w}}_k$ is the conjugate of $\mathbf{w}_k$), we can resolve the filter vector as

$$\mathbf{w}_k^* = \mathbf{R}_{\mathbf{u},k}^{-1} \boldsymbol{\Lambda}_k \boldsymbol{\mu}, \tag{2.63}$$

which depends on the global dual variables. To find the optimal dual variables, we can go from the primal domain to the dual domain by substituting (2.63) to the Lagrangian function, such that the dual optimization probelm is given by

$$\begin{aligned} \boldsymbol{\mu}^* &= \arg\max_{\boldsymbol{\mu}} \; L(\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*, \boldsymbol{\mu}) \\ &= \arg\max_{\boldsymbol{\mu}} \; -\sum_{k=1}^{K} \boldsymbol{\mu}^H \boldsymbol{\Lambda}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \boldsymbol{\Lambda}_k \boldsymbol{\mu} + 2\Re\left(\boldsymbol{\mu}^H \mathbf{f}\right), \end{aligned} \tag{2.64}$$

since the dual function is concave in terms of $\boldsymbol{\mu}$. For notational brevity, we define $\mathbf{G}_k = \boldsymbol{\Lambda}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \boldsymbol{\Lambda}_k, \forall k$. Suppose that each node k has a copy of $\boldsymbol{\mu}$, say $\boldsymbol{\mu}_k$, then the dual optimization problem reduces to an average consensus problem, as

$$\boldsymbol{\mu} = \arg\min_{\boldsymbol{\mu}_k} \sum_{k=1}^{K} \left(\boldsymbol{\mu}_k^H \mathbf{G}_k \boldsymbol{\mu}_k - \frac{2}{K} \Re\left(\boldsymbol{\mu}_k^H \mathbf{f}\right) \right) \quad \text{subject to } \boldsymbol{\mu}_k = \boldsymbol{\mu}_m, \forall (k,m) \in \mathcal{E}, \tag{2.65}$$

which can be solved by using alternating direction method of multipliers (ADMM) [86], primal-dual method of multipliers (PDMM) [87] or randomized gossip algorithms [88].

DISTRIBUTED LCMV BEAMFORMING VIA PDMM

The PDMM algorithm was recently proposed for solving general distributed optimization problems, which was originally named by *Bi-ADMM* [89]. To derive the update equations of PDMM, we first rewrite (2.65) as

$$\min_{\boldsymbol{\mu}_k} \sum_{k=1}^{K} \left(\boldsymbol{\mu}_k^H \mathbf{G}_k \boldsymbol{\mu}_k - \frac{2}{K} \Re\left(\boldsymbol{\mu}_k^H \mathbf{f}\right) \right) \quad \text{s.t. } \boldsymbol{\mu}_k - \boldsymbol{\mu}_m = 0, \forall (k,m) \in \mathcal{E}. \tag{2.66}$$

The augmented Lagrangian function of (2.66) is given by

$$\begin{aligned} \Delta L(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \Bigg(& \boldsymbol{\mu}_k^H \mathbf{G}_k \boldsymbol{\mu}_k - 2\Re\left(\boldsymbol{\mu}_k^H \frac{\mathbf{f}}{K}\right) - \Re\left(\boldsymbol{\mu}_k^H \sum_{m \in \mathcal{N}_k} \frac{k-m}{|k-m|} \boldsymbol{\gamma}_{m|k}\right) \\ & + \sum_{m \in \mathcal{N}_k} \frac{\rho}{2} \left\| \boldsymbol{\mu}_k - \boldsymbol{\mu}_m \right\|_2^2 \Bigg), \end{aligned} \tag{2.67}$$

where ρ is a positive step size and $\boldsymbol{\gamma}$ contains directional edge variables. For example, both $\boldsymbol{\gamma}_{m|k}$ and $\boldsymbol{\gamma}_{k|m}$ are associated with the edge $(k,m) \in \mathcal{E}$, while the former one is computed at node m and the latter one is computed at node k. The PDMM algorithm is an iterative method, i.e., given two unknown variables at each iteration the update equation for one unknown is obtained by minimizing the augmented Lagrangian while fixing the other unknown. In our case, at the $(t+1)$th iteration, we can use the information of $\boldsymbol{\gamma}^{(t)}$

to obtain the update expression for $\boldsymbol{\mu}^{(t+1)}$, which turns out

$$\begin{aligned}\boldsymbol{\mu}_k^{(t+1)} &= \arg\min_{\boldsymbol{\mu}_k} \Delta L(\boldsymbol{\mu}, \boldsymbol{\gamma}^{(t)}) \\ &= \left(\mathbf{G}_k + \rho|\mathcal{N}_k|\mathbf{I}\right)^{-1} \left[\sum_{m\in\mathcal{N}_k}\left(\frac{k-m}{|k-m|}\boldsymbol{\gamma}_{m|k}^{(t)} + \rho\boldsymbol{\mu}_m^{(t)}\right) + \frac{\mathbf{f}}{K}\right]. \end{aligned} \tag{2.68}$$

In [87] , it was shown that the node-specific (2.68) can be updated simultaneously or in an asynchronous fashion. After $\boldsymbol{\mu}_k^{(t+1)}$ is calculated, the edge variables can be updated as

$$\boldsymbol{\gamma}_{k|m}^{(t+1)} = \boldsymbol{\gamma}_{m|k}^{(t)} - \rho\frac{k-m}{|k-m|}\left(\boldsymbol{\mu}_k^{(t+1)} - \boldsymbol{\mu}_m^{(t)}\right), \tag{2.69}$$

which is obtained by minimizing the augmented Lagrangian in terms of $\boldsymbol{\gamma}_{k|m}^{(t)}$. The PDMM update procedure can be terminated until reaching the convergence. From (2.68) and (2.69), we can see that in each iteration not only the dual variable $\boldsymbol{\mu}$ needs to be transmitted between nodes, also the edge variables $\boldsymbol{\gamma}$. In order to reduce the transmission cost, we can make use of the information from previous iterations. For instance, rethinking (2.69), we can easily see that

$$\boldsymbol{\gamma}_{m|k}^{(t)} = \boldsymbol{\gamma}_{k|m}^{(t-1)} - \rho\frac{m-k}{|m-k|}\left(\boldsymbol{\mu}_m^{(t)} - \boldsymbol{\mu}_k^{(t-1)}\right). \tag{2.70}$$

Substituting (2.70) into (2.68), we obtain

$$\boldsymbol{\mu}_k^{(t+1)} = \left(\mathbf{G}_k + \rho|\mathcal{N}_k|\mathbf{I}\right)^{-1} \left[\sum_{m\in\mathcal{N}_k}\left(\frac{k-m}{|k-m|}\boldsymbol{\gamma}_{k|m}^{(t-1)} + 2\rho\boldsymbol{\mu}_m^{(t)} - \rho\boldsymbol{\mu}_k^{t-1}\right) + \frac{\mathbf{f}}{K}\right]. \tag{2.71}$$

Similarly, substituting (2.70) into (2.69), we obtain

$$\boldsymbol{\gamma}_{k|m}^{(t+1)} = \boldsymbol{\gamma}_{k|m}^{(t-1)} + \rho\frac{k-m}{|k-m|}\left(2\boldsymbol{\mu}_m^{(t)} - \boldsymbol{\mu}_k^{(t+1)} - \boldsymbol{\mu}_k^{(t-1)}\right). \tag{2.72}$$

As a consequence, we can get rid of transmitting the edge variables. Hence, at each iteration the transmission energy is only spent for broadcasting the dual variable $\boldsymbol{\mu}$.

After the dual variable is determined, the optimal local filters can be calculated using (2.63). Then, calculating the final beamformer output also turns to an average consensus problem as

$$\min_X \sum_{k=1}^{K}\left(X_k - \mathbf{w}_k^H\mathbf{y}_k\right)^2 \text{ subject to } X_k = X_m, \forall(k,m)\in\mathcal{E}. \tag{2.73}$$

This problem can also be solved iteratively using the PDMM algorithm with the update equations given by [85]

$$X_k^{(t+1)} = \frac{\mathbf{w}_k^H\mathbf{y}_k + \sum_{m\in\mathcal{N}_k}\left(\frac{k-m}{|k-m|}\theta_{m|k}^{(t)} + \rho X_m^{(t)}\right)}{1+\rho|\mathcal{N}_k|}, \tag{2.74a}$$

$$\theta_{k|m}^{(t+1)} = \theta_{m|k}^{(t)} - \rho\frac{k-m}{|k-m|}\left(X_k^{(t+1)} - X_m^{(t)}\right), \tag{2.74b}$$

where $\theta_{k|m}$ is the edge variable defined similarly as before.

2.6.2. DISTRIBUTED MVDR BEAMFORMING

The distributed MVDR beamforming is easier to implement compared to the LCMV beamforming. Suppose that one single target source is to be enhanced which is characterized by the ATF vector $\mathbf{a}$, the MVDR beamforming problem can be formulated as

$$\begin{aligned}\mathbf{w} &= \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{uu}} \mathbf{w}, \quad \text{subject to } \mathbf{a}^H \mathbf{w} = 1 \\ &= \arg\min_{\mathbf{w}} \sum_{k=1}^{K} \mathbf{w}_k^H \mathbf{R}_{\mathbf{u},k} \mathbf{w}_k, \quad \text{subject to } \sum_{k=1}^{K} \mathbf{a}_k^H \mathbf{w}_k = 1,\end{aligned} \tag{2.75}$$

where $\mathbf{a} = [\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_K^T]^T$. The optimal local filters are given by

$$\mathbf{w}_k^* = \frac{\mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{a}_k}{\sum_{k=1}^{K} \mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{a}_k}, \tag{2.76}$$

since the matrix $\mathbf{R}_{\mathbf{uu}}$ is block-diagonal, whose inverse can be computed by inverting each block separately. With the local filters, the beamformer output can be calculated as

$$\hat{S} = \mathbf{w}^H \mathbf{y} = \frac{\sum_{k=1}^{K} \mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{y}_k}{\sum_{k=1}^{K} \mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{a}_k} = \frac{\frac{1}{K}\sum_{k=1}^{K} \mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{y}_k}{\frac{1}{K}\sum_{k=1}^{K} \mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{a}_k}. \tag{2.77}$$

Since both $\mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{y}_k$ and $\mathbf{a}_k^H \mathbf{R}_{\mathbf{u},k}^{-1} \mathbf{a}_k$ can be calculated at each node locally, the calculations of the denominator and the numerator are indeed two averaging consensus problems, which can be solved using the randomized gossip algorithms, e.g., [90]. Note that the distributed MVDR beamforming problem can also be solved by applying the aforementioned PDMM algorithm.

2.7. RTF ESTIMATION

All aforementioned multi-microphone beamforming algorithms require the ATF, or, depending on the formulation, the RTF. Identifying the RTFs is a necessary step prior to beamforming. In literature, there are two often-used RTF estimation emthods, namely covariance subtraction (CS) [39, 40, 41, 42, 43] and covariance whitening (CW) [18, 29, 44, 45]. Assuming that a single target source is of interest whose RTF needs to be identified and the target signal and all the existing interferences are mutually uncorrelated, we can estimate the noisy correlation matrix, denoted by $\hat{\mathbf{R}}_{\mathbf{yy}}$, during the speech-plus-noise segments using (2.9) and estimate the noise correlation matrix, denoted by $\hat{\mathbf{R}}_{\mathbf{nn}}$, during the noise-only segments using (2.8). Given the noise and noisy correlation matrices, we can further estimate the correlation matrix of the clean signal components, denoted by $\hat{\mathbf{R}}_{\mathbf{xx}}$, using matrix subtraction as in (2.10), which is approximately rank-1 for the single target source case.

Based on the rank-1 assumption and using the fact that $\mathbf{R}_{\mathbf{xx}} \triangleq \sigma_S^2 \mathbf{a}\mathbf{a}^H$, where σ_S^2 and $\mathbf{a}$ denote the PSD and the ATF vector of the target source, respectively, we define the RTF vector as

$$\mathbf{d} = \mathbf{a}/a_1, \tag{2.78}$$

where a_1 refers the first element of $\mathbf{a}$. That is, the RTF is defined as the normalized ATF with respect to the reference microphone (here we choose the first microphone as the

reference microphone without loss of generality). Then, we can write $\mathbf{R_{xx}}$ in terms of the RTF as

$$\mathbf{R_{xx}} \triangleq \sigma_{X_1}^2 \mathbf{d}\mathbf{d}^H, \tag{2.79}$$

where $\sigma_{X_1}^2 = \sigma_S^2 |a_1|^2$ represents the clean signal power at the reference microphone. The CS method takes the normalized first column of $\hat{\mathbf{R}}_{\mathbf{xx}}$ as the estimated RTF, i.e.,

$$\hat{\mathbf{d}}_{\mathrm{CS}} \triangleq \frac{\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}{\mathbf{e}_1^T \hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}, \tag{2.80}$$

where $\mathbf{e}_1 = [1, 0, \ldots, 0]^T$ with 1 at the first entry and zeros elsewhere. In addition, the estimate of the signal power at the reference microphone is given by

$$\hat{\sigma}_{X_1}^2 \triangleq \mathbf{e}_1^T \hat{\mathbf{R}}_{\mathbf{xx}} \mathbf{e}_1. \tag{2.81}$$

In high SNR environments, the CS method can provide a good RTF estimate, while its performance degrades significantly in severe noisy scenarios.

The CW method is realized based on the utilization of eigenvalue decomposition (EVD). Specifically, given the estimate of the noise correlation matrix $\hat{\mathbf{R}}_{\mathbf{nn}}$, the quantized microphone measurements that are received by the FC are first whitened by

$$\ddot{\mathbf{y}} = \mathbf{R}_{\mathbf{nn}}^{-H/2} \hat{\mathbf{y}}. \tag{2.82}$$

Then, the correlation matrix of the whitened microphone signals can be estimated by

$$\hat{\mathbf{R}}_{\ddot{\mathbf{y}}\ddot{\mathbf{y}}} = \frac{1}{L_y} \sum_{l=1}^{L_y} \ddot{\mathbf{y}}\ddot{\mathbf{y}}^H, \tag{2.83}$$

similar as (2.9) during the speech-plus-noise segments. Letting $\hat{\boldsymbol{\psi}}$ denote the principal eigenvector of the matrix $\hat{\mathbf{R}}_{\ddot{\mathbf{y}}\ddot{\mathbf{y}}}$, the RTF estimate of the CW method is then given by the normalized principal eigenvector, i.e.,

$$\hat{\mathbf{d}}_{\mathrm{CW}} = \frac{\hat{\mathbf{R}}_{\ddot{\mathbf{y}}\ddot{\mathbf{y}}}^{H/2} \hat{\boldsymbol{\psi}}}{\mathbf{e}_1^T \hat{\mathbf{R}}_{\ddot{\mathbf{y}}\ddot{\mathbf{y}}}^{H/2} \hat{\boldsymbol{\psi}}}. \tag{2.84}$$

It can be shown that the CW method is equivalent to the generalized eigenvalue decomposition (GEVD) of the matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$, i.e., $\hat{\mathbf{d}}_{\mathrm{CW}}$ can also be given by the normalized generalized principal eigenvector of $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$. In general, the CW method can achieve a better RTF estimation accuracy compared to the CS method, especially in noisy environments [42, 45]. However, the CS method is more appealing from the implementation point of view, because it has a much lower computational complexity, as it only needs to take the first column of a matrix, while more computationally demanding matrix EVD and/or matrix inversion is required by the CW method. The performance analysis of both methods can be found in [42, 45, 91]. Note that the CS method can also choose any other column of the speech correlation matrix, so its performance depends on the selection of the reference microphone.

1st&2nd phase	subtraction	whitening
extracting 1st column	CS	CW-1
EVD	EVD-CS	CW

Table 2.1: The classification of all the existing RTF estimation methods for a single target source case.

In fact, the CS and CW methods are two extreme cases from the perspective of implementation. Specifically, the CS method has a low complexity and a low estimation accuracy; the CW method has a high complexity and a high estimation accuracy. Based on this, we can also define two alternative RTF estimation approaches. Using the EVD operation, the first alternative is obtained by taking the principal normalized eigenvector of the matrix $\hat{\mathbf{R}}_{\mathbf{xx}}$ as the RTF estimate [43, 92], referred to as EVD-CS. The second alternative is given by taking the normalized first column vector of the correlation matrix $\hat{\mathbf{R}}_{\ddot{\mathbf{y}}\ddot{\mathbf{y}}}$ of the whitened microphone signals and multiplying with $\hat{\mathbf{R}}_{\mathbf{nn}}^{H/2}$, referred to as CW-1. These four RTF estimation approaches are summarized in Table 2.1. In general, the eigen decomposition based methods perform better than the methods that simply extract the first column vector, as it was shown in [41] that approximating the matrix estimate $\hat{\mathbf{R}}_{\mathbf{xx}}$ by a rank-1 matrix via its principal eigenvector can improve the performance of the Wiener filter, compared to using its first column for the rank-1 approximation. To conclude, the CS method method achieves the worst performance, and the CW method the best performance, with the CW-1 and EVD-CS having intermediate performance. From the implementation efficiency point of view, the ordering is reversed.

3

MICROPHONE SUBSET SELECTION FOR MVDR BEAMFORMER BASED NOISE REDUCTION

This chapter is based on the article published as "Microphone subset selection for MVDR beamformer based noise reduction" by J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens in *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 550–563, 2018.

3.1. INTRODUCTION

MICROPHONE arrays have become increasingly popular in many speech processing applications, e.g., hearing aids [32], teleconferencing systems [93], hands-free telephony [94], speech recognition [1], human-robot interaction [6], etc. Compared to their single-microphone counterparts, microphone arrays typically lead to an enhanced performance when detecting, localizing, or enhancing specific sound sources. This is due to the fact that with a microphone array the sound field is not only sampled in time, but also in space.

Although traditional microphone arrays have been widely investigated, see [95] and reference therein, they do have some important limitations. Typically, conventional microphone arrays have one central processing unit, that is, a fusion center (FC), which physically connects to the microphones. Rearranging the microphones in such a conventional wired and centralized array is impractical. Moreover, usually the target source is located far away from the array, resulting in a low signal-to-noise ratio (SNR). In addition, typically, the size of conventional arrays is limited as the maximum array size is determined by the application device [33].

Recently, wireless acoustic sensor networks (WASNs) have attracted an increased amount of interest [33, 96, 90, 97]. In a WASN, each sensor node is equipped with a single microphone or a small microphone array, and the nodes are spatially distributed across a specific environment. The microphone nodes communicate with their neighboring nodes or the FC using wireless links. The use of WASNs can potentially resolve the limitations encountered with the conventional arrays that were mentioned before. At first, the WASN is not constrained to any specific (fixed) array configuration. Secondly, with a WASN, the position and number of microphones is not anymore determined by the application device. Instead, microphones can be placed at positions that are difficult to reach with conventional microphones. With a WASN, the array-size limitations disappear and the network becomes scalable (i.e., larger array apertures can be achieved) [35]. The fact that microphones in the WASN sample the sound field in a much larger area can yield higher quality recordings as it is likely that some of the sensors are close to the target source and have a higher SNR. One of the bottlenecks in a WASN is the resource usage in terms of power. Transmission of data between nodes or from the nodes to the FC will influence the battery lifetime of the sensor. Although all microphones in the WASN will positively contribute to the estimation task, only a few will have a significant contribution. It is questionable whether using all microphones in the network is beneficial taking the energy usage and lifetime of the sensors into account. Instead of blindly using all sensors, selecting a subset of microphones that is most informative for an estimation task at hand can reduce the data to be processed as well as transmission costs.

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

$$\arg \min_{\mathbf{p} \in \{0,1\}^M} \; f(\mathbf{p}) \quad \text{s.t.} \quad \mathbf{1}_M^T \mathbf{p} = K, \tag{3.1}$$

where $\mathbf{p}$ indicates whether a sensor is selected or not, and the cost function $f(\mathbf{p})$ is op-

timized to select the best subset of K sensors out of M available sensors. Basically, the problem in (1) is a non-convex Boolean optimization problem, which incurs a combinatorial search over all the $\binom{M}{K}$ possible combinations. Usually, it can be simplified via convex relaxation techniques [52, 53, 55] or using greedy heuristics, e.g., leveraging submodularity [98, 54]. When the cardinality of $\mathbf{p}$ is of more concern, the cost function and constraint in (1) can also be interchanged by minimizing the cardinality of $\mathbf{p}$, i.e., $||\mathbf{p}||_0$, while constraining the performance meansure $f(\mathbf{p})$.

In general, sensor selection can be categorized into two classes: model-driven schemes and data-driven schemes. For the model-driven schemes, sensor selection is an offline design, where the sensing operation is designed based only on the data model (even before gathering data) such that a desired ensemble inference performance is achieved. In other words, the model-driven schemes provide the selected sensors *a priori* for the inference tasks [55]. There are many applications of the model-driven schemes for sensor placement in source localization [53], power grid monitoring [99], field estimation [100], target tracking [55], to list a few. In contrast to the offline design schemes, dimensionality reduction can also be done on already acquired data by discarding, i.e., censoring, less informative samples; this is referred as data-driven schemes. Data-driven sensor selection has been applied within the context of speech processing, e.g., speech enhancement [101, 60], speech recognition [102], and target tracking by sensor scheduling [103]. In the WASNs context, due to time-varying topologies, we have typically no information about the data model (e.g., probability density function), but the online measured data (e.g., microphone recordings) are available instead. In this work, we start with the model-driven sensor selection for the spatial filtering based noise reduction problem, which is then extended to a data-driven scheme.

3.1.1. CONTRIBUTIONS

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

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

In a more practical scenario, usually it is impossible to estimate the joint statistics of the complete network beforehand due to the dynamics of the scenario. Instead, the real-time measured data is only what can be accessed. Therefore, we extend the pro-

posed model-driven algorithm to a data-driven scheme using a greedy sensor selection strategy. The performance of the greedy approach is proven to converge to that of the model-based method from an experimental perspective. There are a few existing contributions considering microphone subset selection in the area of audio signal processing. For example, Szurley et al. greedily selected an informative subset according to the SNR gain at each individual microphone for speech enhancement [60]. Bertrand and Moonen [101] conducted greedy sensor selection based on the contribution of each sensor signal to mean squared error (MSE) cost for signal estimation. Kumatani et al. proposed a channel selection for distant speech recognition by considering the contribution of each channel to multichannel cross-correlation coefficients (MCCCs) [102]. The proposed greedy algorithm shows an advantage in computational complexity and optimality as compared to existing greedy approaches [60, 101].

3.1.2. OUTLINE AND NOTATION

The rest of this paper is organized as follows. Sec. 3.2 introduces the signal model, the classical MVDR beamforming, and sensor selection model. Sec. 3.3 presents the problem formulation. Sec. 3.4 presents two solvers based on convex optimization to solve the model-driven sensor selection problem. Sec. 3.5 proposes a greedy algorithm. Sec. 3.6 illustrates the simulation results. Finally, Sec. 3.7 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. diag$(\cdot)$ refers to a block diagonal matrix with the elements in its argument on the main diagonal. $\mathbf{1}_N$ and $\mathbf{0}_N$ denote the $N \times 1$ vector of ones and the $N \times N$ matrix with all its elements equal to zero, respectively. $\mathbf{I}_N$ is an identity matrix of size N. $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. $|\mathscr{U}|$ denotes the cardinality of the set $\mathscr{U}$.

3.2. PRELIMINARIES

3.2.1. SIGNAL MODEL

We assume a spatially distributed candidate set of M microphone sensors that collect and transmit their observations to an FC. The multi-microphone noise reduction methods considered in this paper operate in the frequency domain on a frame-by-frame basis. Let l denote the frame index and ω the frequency bin index, respectively. We assume that the user (i.e., FC) has one source of interest, while multiple interfering sources are present in the environment. Using a discrete Fourier transform (DFT) domain description, the noisy DFT coefficient at the k-th microphone, say $Y_k(\omega, l)$, for $k = 1, 2, \cdots, M$, is given by

$$Y_k(\omega, l) = X_k(\omega, l) + N_k(\omega, l), \tag{3.2}$$

where $X_k(\omega, l) = a_k(\omega)S(\omega, l)$ with $a_k(\omega)$ denoting the acoustic transfer function (ATF) of the target signal with respect to the k-th microphone and $S(\omega, l)$ the target source signal at the source location of interest. In (3.2), the component $N_k(\omega, l)$ represents the total received noise at the k-th microphone (including interfering sources and internal thermal additive noise). For notational convenience, the frequency variable ω and the frame index l will be omitted now onwards bearing in mind that the processing takes

place in the frequency domain. Using vector notation, signals from M microphones are stacked in a vector $\mathbf{y} = [Y_1, ..., Y_M]^T \in \mathbb{C}^M$. Similarly, we define an M dimensional speech vector $\mathbf{x}$ for the speech component contained in $\mathbf{y}$ as $\mathbf{x} = \mathbf{a}S \in \mathbb{C}^M$ with $\mathbf{a} = \left[a_1, \cdots, a_M\right]^T \in \mathbb{C}^M$ denoting the steering vector which is constructed from the ATFs, and a length-M noise vector $\mathbf{n}$. As a consequence, the signal model in (3.2) can be compactly written as

$$\mathbf{y} = \mathbf{x} + \mathbf{n}. \tag{3.3}$$

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

$$\mathbf{R}_{\mathbf{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{nn}} \in \mathbb{C}^{M\times M}, \tag{3.4}$$

where $\mathbb{E}\{\cdot\}$ denotes the mathematical expectation, and $\mathbf{R}_{\mathbf{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} = \sigma_S^2\mathbf{a}\mathbf{a}^H$ with $\sigma_S^2 = \mathbb{E}\{|S|^2\}$ representing the power spectral density (PSD) of the target source. Notice that due to the assumption that $\mathbf{x}$ and $\mathbf{n}$ are uncorrelated, $\mathbf{R}_{\mathbf{xx}}$ can be estimated by subtracting the noise correlation matrix $\mathbf{R}_{\mathbf{nn}}$, which is estimated during the absence of speech from the speech-plus-noise correlation matrix $\mathbf{R}_{\mathbf{yy}}$ [104]. In this work, we assume that a perfect voice activity detector (VAD) is available, such that the noise-only segments and the speech-plus-noise segments are classified accurately.

3.2.2. MVDR BEAMFORMER

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

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w}, \ \text{s.t.} \ \mathbf{w}^H\mathbf{a} = 1. \tag{3.5}$$

The optimal solution, in a best linear unbiased estimator sense, can be obtained using the method of Lagrange multipliers, and is given by [14, 20, 96]

$$\hat{\mathbf{w}} = \frac{\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{a}}. \tag{3.6}$$

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

$$\begin{aligned}\text{SNR}_{\text{out}} &= \frac{\mathbb{E}\left\{|\hat{\mathbf{w}}^H\mathbf{x}|^2\right\}}{\mathbb{E}\left\{|\hat{\mathbf{w}}^H\mathbf{n}|^2\right\}} = \frac{\hat{\mathbf{w}}^H\mathbf{R}_{\mathbf{xx}}\hat{\mathbf{w}}}{\hat{\mathbf{w}}^H\mathbf{R}_{\mathbf{nn}}\hat{\mathbf{w}}} \\ &= \sigma_S^2\mathbf{a}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{a}.\end{aligned} \tag{3.7}$$

3.2.3. SENSOR SELECTION MODEL

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

$$\mathbf{p} = [p_1, p_2, ..., p_M]^T, \tag{3.8}$$

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

$$\mathbf{y_p} = \mathbf{\Phi_p y} = \mathbf{\Phi_p x} + \mathbf{\Phi_p n}, \tag{3.9}$$

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

$$\mathbf{\Phi_p \Phi_p}^T = \mathbf{I}_K, \quad \mathbf{\Phi_p}^T \mathbf{\Phi_p} = \text{diag}(\mathbf{p}). \tag{3.10}$$

Therefore, applying the selection model to the classical MVDR beamformer in Sec. 3.2.2, the best linear unbiased estimator for a subset of K microphones determined by $\mathbf{p}$ will be

$$\hat{\mathbf{w}}_\mathbf{p} = \frac{\mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1}\mathbf{a_p}}{\mathbf{a}_\mathbf{p}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1}\mathbf{a_p}}, \tag{3.11}$$

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

3.3. PROBLEM FORMULATION

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

Let $\mathbf{c} = [c_1, c_2, ..., c_M]^T \in \mathbb{R}^M$ denote the pairwise transmission cost between each microphone and the FC. In general, the power consumption for wireless transmission can be modeled as [105]

$$c_i = c(d_i) + c_i^{(0)}, \ \forall i, \tag{3.12}$$

where $c(d_i)$ represents the power consumption depending on the distance d_i from the node with the i-th microphone to the FC, and $c_i^{(0)}$ is a constant depending on the power consumption of the i-th microphone itself. Based on the energy model in (3.12), our initial problem can be formulated as

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

where $||\cdot||_1$ denotes the ℓ_1-norm, β denotes the minimum output noise power after beamforming, and $\alpha \in (0,1]$ is an adaptive factor to control the output noise power compared to β. Note that β does not depend on the measurements of the whole network, because β/α is just a number that can be assigned by users, e.g., 40 dB, to indicate a desired performance. In (P1), the ℓ_1-norm is used to represent the total transmission costs of the network, i.e., between all the selected sensors and the FC, and it equals the inner-product $\mathbf{c}^T\mathbf{p}$ since both $\mathbf{p}$ and $\mathbf{c}$ are non-negative. Also, notice that (P1) is a general case for spatial filtering based noise reduction problems, e.g., using MVDR beamformers or linear constrained minimum variance (LCMV) beamformers [106]. In the next section, we will show how the optimization problem in (P1) can be solved using some of the properties of the MVDR beamformer.

3.4. MODEL-DRIVEN SENSOR SELECTION

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

Considering the MVDR beamformer in (3.11), the output noise power using the selected sensors is given by

$$\hat{\mathbf{w}}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}} \hat{\mathbf{w}}_{\mathbf{p}} = \left(\mathbf{a}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{a}_{\mathbf{p}}\right)^{-1}, \tag{3.13}$$

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

$$\hat{\mathbf{w}}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{xx},\mathbf{p}} \hat{\mathbf{w}}_{\mathbf{p}} = \sigma_S^2, \tag{3.14}$$

where $\mathbf{R}_{\mathbf{xx},\mathbf{p}}$ denotes the submatrix of $\mathbf{R}_{\mathbf{xx}}$ corresponding to the selected sensors. Hence, following (3.7) the output SNR using the selected sensors is given by

$$\begin{aligned} \mathrm{SNR}_{\mathrm{out},\mathbf{p}} &= \frac{\hat{\mathbf{w}}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{xx},\mathbf{p}} \hat{\mathbf{w}}_{\mathbf{p}}}{\hat{\mathbf{w}}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}} \hat{\mathbf{w}}_{\mathbf{p}}} \\ &= \sigma_S^2 \mathbf{a}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{a}_{\mathbf{p}} \\ &= \sigma_S^2 \mathbf{a}^H \mathbf{\Phi}_{\mathbf{p}}^T \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{\Phi}_{\mathbf{p}} \mathbf{a}. \end{aligned} \tag{3.15}$$

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

$$\begin{aligned} &\min_{\mathbf{p}\in\{0,1\}^M} \quad \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ &\text{s.t.} \quad \sigma_S^2 \mathbf{a}_{\mathbf{p}}^H \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{a}_{\mathbf{p}} \geq \alpha \cdot \mathrm{SNR}, \end{aligned} \tag{P2}$$

where $\mathrm{SNR} = \frac{\sigma_S^2}{\beta}$ represents the maximum output SNR. Both (P1) and (P2) are non-convex because of the Boolean variable $\mathbf{p}$, but also due to the non-linearity of the constraint in $\mathbf{p}$. In what follows, we will present solvers by linearizing (P2) and reformulating

it using convex relaxation. Note that (P1) and (P2) are built from different perspectives (i.e., constraining the output noise power and SNR, respectively), but in the context of the MVDR beamforming, they are equivalent.

3.4.1. CONVEX RELAXATION USING $\mathbf{R_{xx}}$

From the output SNR in (3.15), the selection variable $\mathbf{p}$ appears at three places, that are: $\mathbf{\Phi}_\mathbf{p}^T$, $\mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1}$ and $\mathbf{\Phi}_\mathbf{p}$. We combine these together as one new matrix $\mathbf{Q} = \mathbf{\Phi}_\mathbf{p}^T \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{\Phi}_\mathbf{p}$. To simplify calculations, in what follows, we will rearrange $\mathbf{Q}$ such that $\mathbf{p}$ occurs only at one place. Let us first consider a decomposition of the noise covariance matrix [55, 51]

$$\mathbf{R_{nn}} = \lambda \mathbf{I}_M + \mathbf{G}, \tag{3.16}$$

where λ is a positive scalar and $\mathbf{G}$ is a positive definite matrix (if λ is smaller than the smallest eigenvalue of $\mathbf{R_{nn}}$, this decomposition can be easily found). The reason for choosing such a λ is to make $\mathbf{G}^{-1} + \lambda^{-1}\text{diag}(\mathbf{p})$ positive definite, which will be seen after (3.24). Using (3.16), we have

$$\mathbf{R}_{\mathbf{nn},\mathbf{p}} = \mathbf{\Phi}_\mathbf{p} \left(\lambda \mathbf{I}_M + \mathbf{G}\right) \mathbf{\Phi}_\mathbf{p}^T = \lambda \mathbf{I}_K + \mathbf{\Phi}_\mathbf{p} \mathbf{G} \mathbf{\Phi}_\mathbf{p}^T, \tag{3.17}$$

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

$$\mathbf{Q} = \mathbf{\Phi}_\mathbf{p}^T \left(\lambda \mathbf{I}_K + \mathbf{\Phi}_\mathbf{p} \mathbf{G} \mathbf{\Phi}_\mathbf{p}^T\right)^{-1} \mathbf{\Phi}_\mathbf{p}. \tag{3.18}$$

Using the matrix inversion lemma [107, p.18]

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

we can simplify $\mathbf{Q}$ in (3.18) as

$$\mathbf{Q} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\left(\mathbf{G}^{-1} + \lambda^{-1}\text{diag}(\mathbf{p})\right)^{-1} \mathbf{G}^{-1}. \tag{3.19}$$

Note that (3.19) is still non-linear in $\mathbf{p}$ due to the inversion operation, but $\mathbf{p}$ appears now only at one place. Based on $\mathbf{Q}$, the output SNR with sensor selection as in (3.15) can be calculated as [107, p.6]

$$\begin{aligned} \text{SNR}_{\text{out},\mathbf{p}} &\overset{(1)}{=} \text{trace}\left(\sigma_S^2 \mathbf{a}^H \mathbf{\Phi}_\mathbf{p}^T \mathbf{R}_{\mathbf{nn},\mathbf{p}}^{-1} \mathbf{\Phi}_\mathbf{p} \mathbf{a}\right) \\ &\overset{(2)}{=} \text{trace}\left(\mathbf{Q}\mathbf{R}_{\mathbf{xx}}\right) \\ &\overset{(3)}{=} \text{trace}\left(\mathbf{R}_{\mathbf{xx}}^{\frac{H}{2}} \mathbf{Q} \mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}}\right), \end{aligned} \tag{3.20}$$

where the trace$(\cdot)$ operator computes the trace of a matrix, and $\mathbf{R}_{\mathbf{xx}}^{\frac{1}{2}}$ represents the principal square root of $\mathbf{R_{xx}}$. The second and third equality in (3.20) is based on trace property, which is employed to make the linear matrix inequality (LMI) in (3.25) symmetric. Here, we utilize the trace operation to express the output SNR as a function of $\mathbf{R_{xx}}$. The latter can be estimated using the recorded audio in practice, e.g., during the training phase,

or using the correlation matrices $\mathbf{R_{yy}}$ and $\mathbf{R_{nn}}$ without the need to explicitly know the steering vector $\mathbf{a}$ or $\mathbf{a_p}$.

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

$$\text{trace}\left(\mathbf{Z}-\frac{\alpha\sigma_S^2}{M\beta}\mathbf{I}_M\right)\geq 0, \tag{3.21}$$

$$\mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{Q}\mathbf{R_{xx}^{\frac{1}{2}}}=\mathbf{Z}, \tag{3.22}$$

where the equality constraint in (3.22) is non-linear in $\mathbf{p}$. For linearization, we relax it to an inequality constraint

$$\mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{Q}\mathbf{R_{xx}^{\frac{1}{2}}}\succeq\mathbf{Z}. \tag{3.23}$$

Note that (3.21) and (3.23) are sufficient conditions for obtaining the original constraint in (P2), this is why we utilize $\succeq$ for convex relaxation. Substituting (3.19) in (3.23), we get

$$\mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{G}^{-1}\mathbf{R_{xx}^{\frac{1}{2}}}-\mathbf{Z}\succeq\mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{G}^{-1}\left[\mathbf{G}^{-1}+\lambda^{-1}\text{diag}\left(\mathbf{p}\right)\right]^{-1}\mathbf{G}^{-1}\mathbf{R_{xx}^{\frac{1}{2}}}. \tag{3.24}$$

Due to the positivity of λ, the positive definiteness of $\mathbf{G}$ and the Boolean vector $\mathbf{p}$, the matrix $\mathbf{G}^{-1}+\lambda^{-1}\text{diag}(\mathbf{p})$ is positive definite, and this is why we chose in (3.16) a positive scalar λ and a positive definite matrix $\mathbf{G}$ to decompose the matrix $\mathbf{R_{nn}}$. Using the Schur complement [108, p.650], we obtain a symmetric LMI of size $2M$ from (3.24) as

$$\begin{bmatrix}\mathbf{G}^{-1}+\lambda^{-1}\text{diag}(\mathbf{p}) & \mathbf{G}^{-1}\mathbf{R_{xx}^{\frac{1}{2}}}\\ \mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{G}^{-1} & \mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{G}^{-1}\mathbf{R_{xx}^{\frac{1}{2}}}-\mathbf{Z}\end{bmatrix}\succeq\mathbf{0}_{2M}, \tag{3.25}$$

which is linear in $\mathbf{p}$. Furthermore, the Boolean variable $\mathbf{p}$ can be relaxed using continuous variables $\mathbf{p}\in[0,1]^M$ or semidefinite relaxation [109]. In this work, we utilize the former way. Accordingly, (P2) can be expressed in the following form:

$$\begin{aligned}&\min_{\mathbf{p},\mathbf{Z}}\ \|\text{diag}(\mathbf{p})\mathbf{c}\|_1\\ &\text{s.t. trace}\left(\mathbf{Z}-\frac{\alpha\sigma_S^2}{M\beta}\mathbf{I}_M\right)\geq 0,\\ &\quad\begin{bmatrix}\mathbf{G}^{-1}+\lambda^{-1}\text{diag}(\mathbf{p}) & \mathbf{G}^{-1}\mathbf{R_{xx}^{\frac{1}{2}}}\\ \mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{G}^{-1} & \mathbf{R_{xx}^{\frac{H}{2}}}\mathbf{G}^{-1}\mathbf{R_{xx}^{\frac{1}{2}}}-\mathbf{Z}\end{bmatrix}\succeq\mathbf{0}_{2M},\\ &\quad 0\leq p_i\leq 1,\ \ i=1,2,\cdots,M.\end{aligned} \tag{3.26}$$

The relaxed optimization problem in (3.26) is a semidefinite programming problem [108, p.128] and can be solved efficiently in polynomial time using interior-point methods or solvers, like CVX [110] or SeDuMi [111]. The computational complexity for solving (3.26) is of the order of $\mathcal{O}(M^3)$. The approximate Boolean selection variables p_i can be obtained by randomized rounding using the solution of (3.26) [53]. Notice that the solver in (3.26) depends on $\mathbf{R_{xx}}$. In a practical scenario, this is unknown, but can be estimated based on estimates of the correlation matrices $\mathbf{R_{yy}}$ and $\mathbf{R_{nn}}$ as shown in (3.4). $\mathbf{R_{yy}}$ can be estimated from the data itself, and $\mathbf{R_{nn}}$ can be estimated using a VAD or noise correlation matrix estimator for the noise-only frames, see e.g., [25].

3.4.2. SOLVER BASED ON THE STEERING VECTOR **a**

Suppose the ATFs from the source to the microphones are known, the steering vectors **a** (in free field) can be constructed. With **a**, the output SNR in (3.20) can be expressed as

$$\mathrm{SNR}_{\mathrm{out},\mathbf{p}} = \sigma_S^2 \mathbf{a}^H \mathbf{Q} \mathbf{a}. \tag{3.27}$$

Therefore, using the expression for **Q** in (3.19), the original constraint in (P2) can be rewritten as

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

or, reorganized as

$$\mathbf{a}^H \mathbf{G}^{-1} \mathbf{a} - \frac{\alpha}{\beta} \geq \mathbf{a}^H \mathbf{G}^{-1} \left(\mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) \right)^{-1} \mathbf{G}^{-1} \mathbf{a}. \tag{3.28}$$

Using the Schur complement, (3.28) can be reformulated as a symmetric LMI of size $M+1$

$$\begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1} \mathbf{a} \\ \mathbf{a}^H \mathbf{G}^{-1} & \mathbf{a}^H \mathbf{G}^{-1} \mathbf{a} - \frac{\alpha}{\beta} \end{bmatrix} \succeq \mathbf{0}_{M+1}. \tag{3.29}$$

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

$$\begin{aligned} \min_{\mathbf{p}} \ & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \ & \begin{bmatrix} \mathbf{G}^{-1} + \lambda^{-1} \mathrm{diag}(\mathbf{p}) & \mathbf{G}^{-1} \mathbf{a} \\ \mathbf{a}^H \mathbf{G}^{-1} & \mathbf{a}^H \mathbf{G}^{-1} \mathbf{a} - \frac{\alpha}{\beta} \end{bmatrix} \succeq \mathbf{0}_{M+1} \\ & 0 \leq p_i \leq 1, \ \ i = 1, 2, \cdots, M, \end{aligned} \tag{3.30}$$

where the Boolean variables **p** have already been relaxed using the continuous surrogates $\mathbf{p} \in [0,1]^M$, and (3.30) has a standard semidefinite programming form, which can also be solved by the aforementioned tools. Notice that this solver depends on knowledge on **a**. To estimate (the direct path of) **a** one can use a source localization algorithm, e.g., [112, 113, 114], in combination with the sensor locations, or use the generalized eigenvalue decomposition of the matrices $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$ [44, 29].

Remark 3. *The differences between (3.26) and (3.30) are threefold: 1) (3.30) preserves the constraint on the output SNR (or noise power), yet (3.26) relaxes it in a convex way by introducing an auxiliary variable* **Z**; *2) Observing the LMIs in (3.26) and (3.30), they differ in dimensions (i.e.,* $2M$ *and* $M+1$*, respectively), so (3.30) is computationally much more efficient; 3) The solver in (3.26) requires to estimate the speech correlation matrix* $\mathbf{R}_{\mathbf{xx}}$ *and the PSD* σ_S^2 *of the target source, while (3.30) requires the steering vector* **a**.

Remark 4. *For a special case, when the noise is spatially uncorrelated with covariance matrix*

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

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

$$\begin{aligned} \min_{\mathbf{p}} \ & \|\mathrm{diag}(\mathbf{p})\mathbf{c}\|_1 \\ \text{s.t.} \ & \mathbf{a}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathrm{diag}(\mathbf{p}) \mathbf{a} \geq \frac{\alpha}{\beta}. \end{aligned} \tag{3.31}$$

Although the above optimization problem is nonconvex in $\mathbf{p} \in \{0,1\}^M$*, it admits a simple non-iterative solution based on rank ordering. More specifically, the optimal solution to (3.31) is given by setting the entries of* $\mathbf{p}$ *corresponding to the indices*

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

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

3.5. GREEDY SENSOR SELECTION

In Sec. 3.4, the sensor selection problem was solved using statistical information from the complete network, i.e., $\mathbf{R_{xx}}$ and $\mathbf{R_{nn}}$. In practice, this information is unknown and needs to be estimated from all the sensors' measurements. Hence, we call this a model-driven approach as the complete $\mathbf{R_{xx}}$ and $\mathbf{R_{nn}}$ are required as well as the transmission power from the microphones to the FC. In a practical scenario, it is undesired to estimate the statistics of the complete network up front, as this would imply a lot of data transmission for sensor nodes that might never be selected in the end as most sensors are non-informative. Moreover, in practice, the position of the FC or microphones might be changing as well. For this reason we need a selection mechanism that does not rely on knowledge of the statistics and microphone-FC distances of the complete network. Instead, we could access the measurements of neighboring sensors (close to the FC or already selected sensors). In this section, we present a greedy approach for the sensor selection based noise reduction problem, which does not require to estimate the global statistics. Therefore, the greedy algorithm can be performed online, and it belongs to the data-driven category. In Sec. 3.6, we will experimentally show that the data-driven and model-driven approach will converge to a similar performance.

Let r_i denote the spatial position of the i-th microphone, $\mathscr{S}_1$ a candidate set of microphones and $\mathscr{S}_2$ the selected set, respectively. The proposed greedy algorithm is summarized in Algorithm 1. Given an arbitrary initial spatial point z_0 and a transmission range R_0[1], we can initialize the candidate set $\mathscr{S}_1$ of sensors, i.e., the R_0-closest sensors to z_0. For the candidate set $\mathscr{S}_1$, we estimate the noise correlation matrix $\mathbf{R_{nn},\mathscr{S}_1}$ and decompose it following (3.16), and then solve the optimization problem in (3.26) or (3.30). For instance, for $\mathscr{S}_1$ the optimization problem in (3.30) can be reformulated as

$$\begin{aligned} &\min_{\mathbf{p}\in[0,1]^{K_1}} \quad \|\mathrm{diag}(\mathbf{p}_{\mathscr{S}_1})\mathbf{c}_{\mathscr{S}_1}\|_1 \\ &\text{s.t.} \quad \begin{bmatrix} \mathbf{G}_{\mathscr{S}_1}^{-1} + \lambda_{\mathscr{S}_1}^{-1}\mathrm{diag}(\mathbf{p}_{\mathscr{S}_1}) & \mathbf{G}_{\mathscr{S}_1}^{-1}\mathbf{a}_{\mathscr{S}_1} \\ \mathbf{a}_{\mathscr{S}_1}^H\mathbf{G}_{\mathscr{S}_1}^{-1} & \mathbf{a}_{\mathscr{S}_1}^H\mathbf{G}_{\mathscr{S}_1}^{-1}\mathbf{a}_{\mathscr{S}_1} - \frac{\alpha}{\beta_{\mathscr{S}_1}} \end{bmatrix} \succeq \mathbf{0}_{K_1+1} \\ &\qquad 0 \leq p_i \leq 1, \ \forall i \in \mathscr{S}_1, \end{aligned} \tag{3.32}$$

[1] R_0 can be defined as the wireless transmission range $\sqrt{\log(2M)/M}$ in a random geometric graph to guarantee that the network is connected with high probability [115].

where $\beta_{\mathcal{S}_1}$ represents the output noise power of the classical MVDR beamformer using the microphones in the candidate set $\mathcal{S}_1$, which is termed as the local constraint. Notice that the adaptive factor α is the same as that in the model-driven scheme. If $\alpha \leq 1$, (3.32) will always have a feasible solution within $\mathcal{S}_1$, the feasible set will be taken and used to define a new set $\mathcal{S}_2$ with $|\mathcal{S}_2| \leq |\mathcal{S}_1|$. Then, based on the set $\mathcal{S}_2$, a new set $\mathcal{S}_1$ is formed based on the R_0-closest sensors with respect to the sensors included in the set $\mathcal{S}_2$[2]. These operations are continued until $\mathcal{S}_1$ or $\mathcal{S}_2$ does not change (i.e., until convergence has been achieved). The finally selected set $\mathcal{S}_2$ will always be smaller than the selected set for the model-driven approach from Sec. 3.4 when using the same α. This is due to the fact that the output noise power $\beta_{\mathcal{S}_1}$ in the constraint of the greedy approach is based on the set $\mathcal{S}_1$ that is always smaller or equal to the initial set as used by the model-driven approach in (3.30) (where β is obtained by involving all sensors). As a result, β/α will always be smaller than $\beta_{\mathcal{S}_1}/\alpha$. In summary, $\beta/\alpha < \beta_{\mathcal{S}_1}/\alpha$. The performance of the greedy approach (after convergence) will therefore always be somewhat worse than the model-based approach, as the constraint is less tight. This can either be solved by choosing a different (larger) α for the greedy approach, or, by switching from the constraint $\beta_{\mathcal{S}_1}/\alpha$ to the constraint β/α after convergence. As an alternative, we could have used the constraint β/α within the greedy approach of (3.32) right from the beginning. However, in that case, in the first few iterations (3.32) would have no feasible solution as an insufficient amount of measurements are available to satisfy the constraint on the output noise power. As a consequence of an infeasible solution, the selected set $\mathcal{S}_2$ will keep all sensors from $\mathcal{S}_1$, of which many are actually uninformative.

In order to make the performance of the proposed greedy algorithm converge to that of the model-driven approach, we switch from $\beta_{\mathcal{S}_1}$ (local constraint) to β (global constraint) after the above iterative procedure converges (i.e., the constraint $\beta_{\mathcal{S}_1}/\alpha$ for solving (3.32) has been satisfied). Finally, the proposed greedy algorithm will converge to the model-driven method based on the global constraint. To conclude, the greedy algorithm includes two steps: using a locally defined constraint ($\beta_{\mathcal{S}_1}/\alpha$) and using a globally defined constraint (β/α), as summarized in Algorithm 1. Recall that the globally defined constraint, which involves β/α with β denoting the minimum output noise power after beamforming, does not need to be dependent on the measurements of the whole network. Hence, the greedy algorithm does not need to know the exact optimal performance, i.e., β. For the implementation in practice, we only need to set a number for β/α depending on the expected performance. Note that the computational complexity of each iteration is of the order of $\mathcal{O}(|\mathcal{S}_1|^3)$, and the number of iterations depends on z_0 and R_0. From the description of the algorithm, we know that both the greedy algorithm and the model-driven method have, in the end, the same constraint that must be satisfied, leading to very similar performance, which can also be found in simulations.

3.6. SIMULATIONS

In this section, the proposed algorithms are experimentally evaluated. Sec. 3.6.1 introduces three reference methods that we will use for comparison. In Sec. 3.6.2, the exper-

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

Algorithm 1: Greedy Sensor Selection

Step 1: initialization
Initial point: z_0
Transmission range: R_0
Selected set: $\mathscr{S}_2 = \varnothing$
Candidate set: $\mathscr{S}_1 = \{i | \|r_i - z_0\|_2 \leq R_0,\ \forall i\ \}$;
Step 2: considering local constraint
Cardinality of the active set: $K_1 = |\mathscr{S}_1|$;
Decomposing: $\mathbf{R}_{\mathbf{nn},\mathscr{S}_1} = \lambda_{\mathscr{S}_1}\mathbf{I}_{K_1} + \mathbf{G}_{\mathscr{S}_1}$;
Solving (3.32) using the local constraint $\beta_{\mathscr{S}_1}$;
Update:
$\mathscr{S}_2 = \{i | p_i = 1,\ \forall i \in \mathscr{S}_1\}$;
$\mathscr{S}_1 = \mathscr{S}_2 \cup \{i | \|r_i - r_{\mathscr{S}_2}\|_2 \leq R_0,\ \forall i\}$;
Go to Step 2 until converge;
Step 3: solving (3.32) using global constraint β;
If infeasible, update
$\mathscr{S}_2 = \mathscr{S}_1$;
$\mathscr{S}_1 = \mathscr{S}_2 \cup \{i | \|r_i - r_{\mathscr{S}_2}\|_2 \leq R_0,\ \forall i\}$;
Go to line 14;
If feasible, update
$\mathscr{S}_2 = \{i | p_i = 1,\ \forall i \in \mathscr{S}_1\}$;
$\mathscr{S}_1 = \mathscr{S}_2 \cup \{i | \|r_i - \mathscr{S}_2\|_2 \leq R_0,\ \forall i\}$;
Go to line 14 until converge;
Return $\mathscr{S}_2$.

imental setup is explained. In Sec. 3.6.3, the proposed model-driven sensor selection based MVDR beamformer (referred to as MD-MVDR in short) is compared with the reference methods introduced in Sec. 3.6.1. In Sec. 3.6.4, we will analyze the performance of the proposed greedy approach as a data-driven sensor selection, including the convergence behaviour, initialization and the adaptivity of a moving FC. Sec. 3.6.5 compares the computational complexity between the model-driven method and the greedy approaches.

3.6.1. REFERENCE METHODS

Apart from the classical MVDR beamforming without sensor selection as introduced in Sec. 3.2.2, the proposed approaches will also be compared with a weighted sparse MVDR beamformer [116, 117, 118], a radius-based MVDR beamformer and a utility-based greedy method [101, 60].

WEIGHTED SPARSE MVDR BEAMFORMER

A naive alternative to sensor selection for spatial filtering is to enforce sparsity in the filter coefficients while designing the beamformer. Due to the physical nature of sound, this approach trades a small loss in SNR for a large reduction in communication power required to produce a beamformer output by reducing the active nodes. Some existing

works on sparse MVDR beamformers are presented in [116, 117, 118]. One of our reference methods is therefore a sparse MVDR beamformer. However in order to make the comparison with the sparse MVDR beamformer fair, we use a weighting by the transmission power. Using the model of transmission costs from (3.12), the weighted sparse MVDR beamformer can be formulated as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ \mathbf{w}^H \mathbf{R}_{\mathbf{nn}} \mathbf{w} + \mu \|\mathbf{w}^H \mathrm{diag}(\mathbf{c})\|_0 \quad \text{s.t. } \mathbf{w}^H \mathbf{a} = 1, \tag{3.33}$$

where μ denotes the regularization parameter to control sparsity, and the ℓ_0-norm can be relaxed by the ℓ_1-norm or the concave surrogate based on sum-of-logarithms [53, 119]. When $\mu = 0$, it is identical to the classical MVDR beamformer in Sec. 3.2.2. Note that a larger μ leads to a sparser $\mathbf{w}$. The product $\mathbf{w}^H \mathrm{diag}(\mathbf{c})$ indicates the pairwise transmission costs. Weighting the beamforming filter $\mathbf{w}$, the sensors with smaller transmission costs have a dominant contribution to $\mathbf{w}$ compared to sensors with larger transmission costs. From the standpoint of implementation, for each frequency bin, if $|w_i| \geq \varepsilon, \forall i$, the i-th sensor will be selected, otherwise not. Due to this "inevitable" thresholding, the resulting beamformer is not necessarily MVDR anymore. The threshold ε is chosen empirically.

RADIUS-BASED MVDR BEAMFORMER

The goal of this article is to minimize the transmission costs while constraining the performance. A straightforward way to reduce transmission costs is by selecting the sensors close to the FC. The closer a sensor to the FC, the less transmission power is required. Hence, given a radius γ, we can involve the sensors within the circle centered by the FC for the MVDR beamfomer, which we call radius-based MVDR beamformer. An example is given in Fig. 3.2(a), where the blue sensors are chosen with $\gamma = 6$ m. Obviously, this approach does not take the source or interference information into account, and its performance suffers from γ and the network topology.

UTILITY BASED GREEDY SENSOR ADDITION

In [60], the most informative subset of microphones is obtained by greedily removing the sensor that has the least contribution to a utility measurement (e.g., SNR gain, output noise power, MSE cost), also called backward selection. This method requires to know the statistics offline and can be considered a model-driven approach. While in [101], apart from sensor selection based on backward selection, an alternative was proposed by greedily adding the sensor that has the largest contribution to the utility (forward selection). This can be considered as an online data-driven procedure. In order to compare the proposed greedy algorithm with the state-of-the-art greedy methods, we summarize [101, 60] as the utility based greedy sensor addition shown in Algorithm 2. The utility based greedy sensor removal can be summarized similarly. In this work, our focus is on the transmission costs. To measure the utility, we therefore take the ratio of the gain of the output noise power $\mathbf{\Delta}$ that is obtained by adding each sensor from $\mathcal{S}_1 \setminus \mathcal{S}_2$ to $\mathcal{S}_2$, to the transmission cost. Here, the sets $\mathcal{S}_1$ and $\mathcal{S}_2$, are respectively, defined the same as for Algorithm 1. The sensor which has the larger ratio between noise reduction and transmission cost would have the larger utility. When the transmission costs for the selected set $\mathcal{S}_2$ exceeds the maximum cost budget c_T, the algorithm is terminated. Note that this

Algorithm 2: Utility based greedy sensor addition

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

approach only adds one sensor to the selected set $\mathscr{S}_2$ per iteration, thus it may require many iterations to get an acceptable solution.

3.6.2. EXPERIMENT SETUP

Fig. 3.2(a) shows the experimental setup employed in the simulations, where 169 candidate microphones are placed uniformly in a 2D room with dimensions (12×12) m. The desired speech source (red solid circle) is located at (2.4, 9.6) m. The FC (black solid square) is placed at (9, 3) m. Two interfering sources (blue stars) are positioned at (2.4, 2.4) m and (9.6, 9.6) m, respectively. The target source signal is a 10 minute long concatenation of speech signals originating from the TIMIT database [120]. The interferences are stationary Gaussian speech shaped noise sources. All signals are sampled at 16 kHz. We use a square-root Hann window of 20 ms for framing with 50% overlap. The ATFs are generated using [121] with reverberation time $T_{60} = 200$ ms. The threshold ε for the sparse MVDR beamformer is set to be 10^{-5} empirically, since the coefficients smaller than this threshold are negligible. We also model microphone self noise using zero-mean uncorrelated Gaussian noise with an SNR of 50 dB.

To focus on the concept of sensor selection, we assume that the ATFs (i.e., steering vector $\mathbf{a}$) are perfectly known. In practice, this can be estimated using source localization algorithms, e.g., [112, 113], in combination with the sensor locations, or, by calculating the generalized eigenvalue decomposition of the matrices $\mathbf{R_{nn}}$ and $\mathbf{R_{yy}}$ [44, 29]. For the correlation matrices, we use noise-only segments which are long enough to estimate $\mathbf{R_{nn}}$; during the speech-plus-noise segments $\mathbf{R_{yy}}$ is tracked and $\mathbf{R_{xx}}$ can be obtained by subtracting the estimate of $\mathbf{R_{nn}}$ from $\mathbf{R_{yy}}$ simultaneously. For the wireless transmission model in (3.12), we consider the simplest wireless transmission case, where the transmission cost between each sensor and the FC is proportional to the square of their Euclidean distance [34], and we assume that the device dependent cost $c_i^{(0)} = 0, \forall i$. In the following simulations, the transmission costs are normalized between 0 and 1 based on the total transmission costs between all the microphones and the FC.

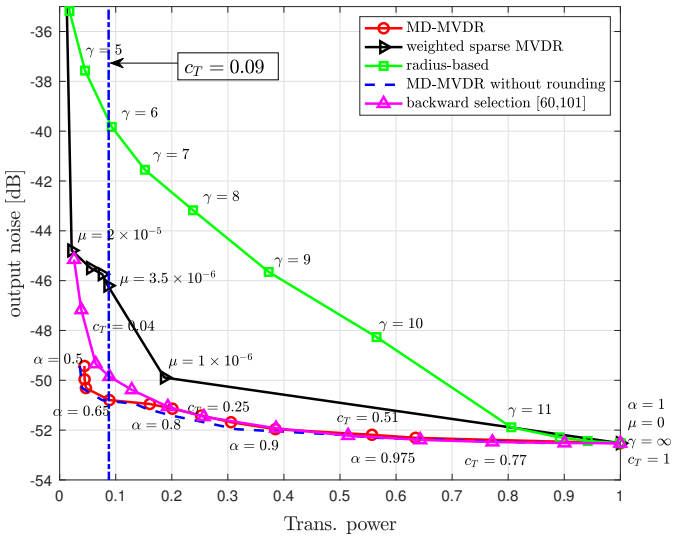


Figure 3.1: Output noise power in terms of transmission cost for different choices of α, μ, γ, c_T.

3.6.3. EVALUATION OF THE MODEL-DRIVEN APPROACH

In order to compare the state-of-the-art approaches mentioned in Sec. 3.6.1, we first investigate the influence of the required parameters α, μ, γ, c_T on the performance, for the proposed and the different reference methods. Fig. 3.1 shows the relationship between the output noise power (in dB) and the transmission power for SIR = 0 dB with SIR representing signal-to-interference ratio. Fig. 3.1 also shows the results without randomized rounding (blue dashed curve) regarded as the lower bound of the proposed method, i.e., involving the selection variable $\mathbf{p}$ (thus, no selection) for computations. As we can see that the performance of MD-MVDR is smaller than that of the MD-MVDR without rounding, the binary solution of the proposed method using randomized rounding is still satisfactory in terms of expected output noise power. We can conclude that in order to reach the same noise reduction performance, the proposed approach always requires significantly less transmission costs compared to the weighted sparse beamformer or radius-based beamformer. If the transmission power budget c_T (defined in Algorithm 2) is small, the proposed method performs better than the backward selection [60], and if c_T is large, they are comparable. Furthermore, when $\alpha = 0.65, \gamma = 6, \mu = 3.5 \times 10^{-6}$, the four approaches approximately have the same transmission power as $c_T = 0.09$. Hence, in the simulations that will follow we will compare the cases for $\alpha = 0.65, \gamma = 6, \mu = 3.5 \times 10^{-6}, c_T = 0.09$. Note that in Fig. 3.1, all the microphones are involved for the MVDR beamforming when $\alpha = 1, \gamma = \infty, \mu = 0, c_T = 1$. This is the optimal MVDR beamformer.

Fig. 3.2(a)-(d) illustrate typical sensor selection examples for one angular frequency ($\omega = \pi/256$ rad/s) of the radius-based MVDR beamformer ($\gamma = 6$), sparse MVDR beamformer ($\mu = 3.5 \times 10^{-6}$), backward selection ($c_T = 0.09$) and the proposed method ($\alpha = 0.65$), respectively. In addition, we show the radius for the radius-based MVDR, where all the sensors within this radius are selected, and thus not depicted explicitly in Fig. 3.2(b)-

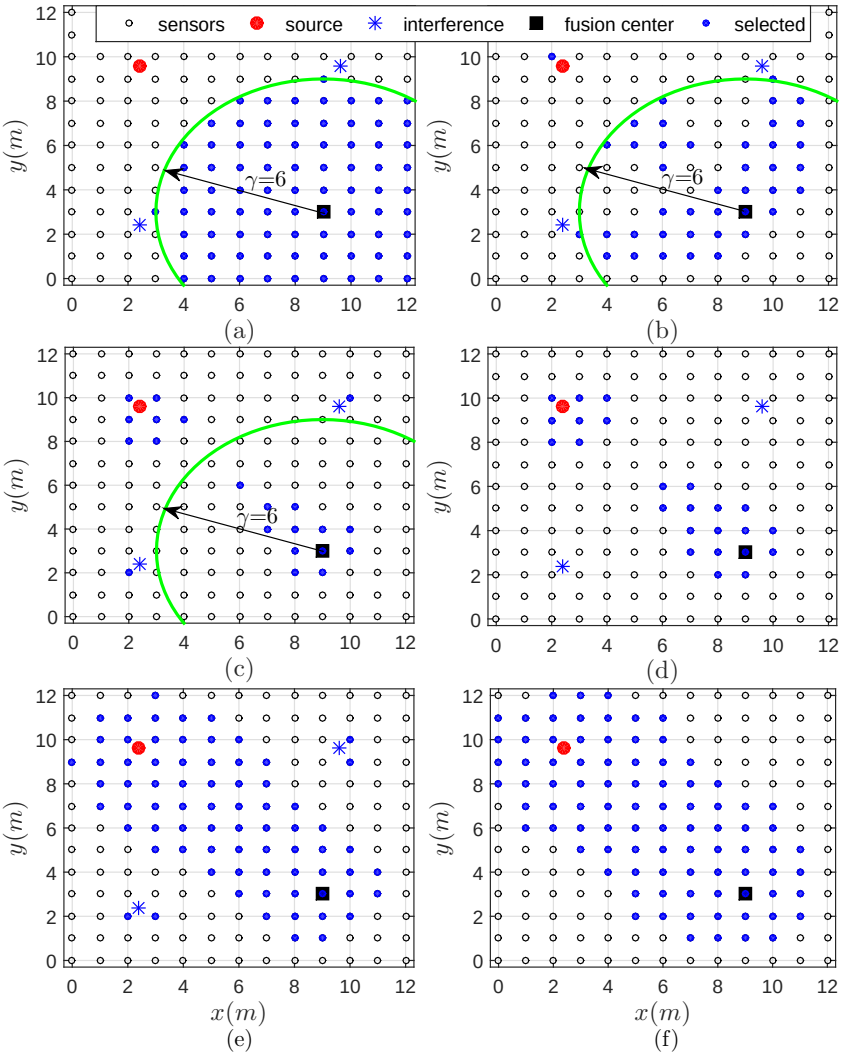


Figure 3.2: Microphone subset selection examples (The blue sensors are activated for the MVDR beamformer): (a) radius-based MVDR beamforming, (b) sparse MVDR beamforming, (c) backward selection [60] and (d) proposed method ($\alpha = 0.65$) for spatially correlated noises, respectively, (e) proposed method for correlated case with $\alpha = 0.9$, and (f) proposed method ($\alpha = 0.9$) for spatially uncorrelated noises only.

(d). For fixed sensor and source locations, it is observed that the selected sensors are the same for most frequency bins. The sensors within the green circles ($\gamma = 6$) are selected by the radius-based method, which chooses the γ-closest sensors relative to the FC for the MVDR beamformer. It can be seen that in order to save transmission power as well as to reduce noise, the proposed approach selects some microphones close to the source and some close to the FC for computation, while the sparse MVDR beamformer

or radius-based method do not have this property. Although the backward selection has this property, it performs somewhat worse in noise reduction, which can be seen in Fig. 3.1. On one hand, the signals recorded by the microphones close to the source position are degraded less by the interfering source, and they preserve the target source better. Those microphones are helpful for enhancing the target source. On the other hand, the microphones close to the FC require less transmission power to transmit data to the FC. They are selected as they hardly add to the total transmission costs. When we increase the adaptive factor α, more sensors that are close to the interference positions are selected as well, because they carry information on the interfering sources as shown in Fig. 3.2(e).

Fig. 3.2(f) illustrates the case where interfering sources are absent, and the microphone recordings are degraded by the microphone self noise, taking the noise level SNR = 50 dB. Compared to Fig. 3.2(e), most selected microphones are the same, and they are more aggregate to the source position as well as to the FC. The difference is whether to select sensors that are close to the interferences. From this comparison, we can also conclude that the sensors that are close to the interference are useful for cancelling the correlated noise.

3.6.4. EVALUATION OF THE DATA-DRIVEN APPROACH

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

CONVERGENCE BEHAVIOUR

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

Fig. 3.3 illustrates the proposed greedy algorithm (i.e., Algorithm 1) for $\alpha = 0.9$ using the same experimental setup of Fig. 3.2(e). In detail, at the 1st iteration (e.g., $k = 1$) the R_0-closest candidate set $\mathscr{S}_1$ has five sensors. Based on the local constraint three sensors (in blue) are selected to form the set $\mathscr{S}_2$. The candidate set $\mathscr{S}_1$ is then increased by adding the R_0-closest sensors with respect to $\mathscr{S}_2$. This procedure continues for the first 21 iterations. When $k = 21$, we can see that $\mathscr{S}_2$ is completely surrounded by $\mathscr{S}_1$, such that if we still use the local constraint, there would be no new sensors that can be added to $\mathscr{S}_1$, from which we conclude that the local constraint, i.e., $\beta_{\mathscr{S}_1}/\alpha$, has been satisfied. In order to satisfy the global constraint on the output noise power, the algorithm is then switched to the global constraint after the 21st iteration, i.e., β/α. Finally, three more iterations are further required the reach the expected performance.

We can see from Fig. 3.3, that the proposed greedy method does not blindly increase

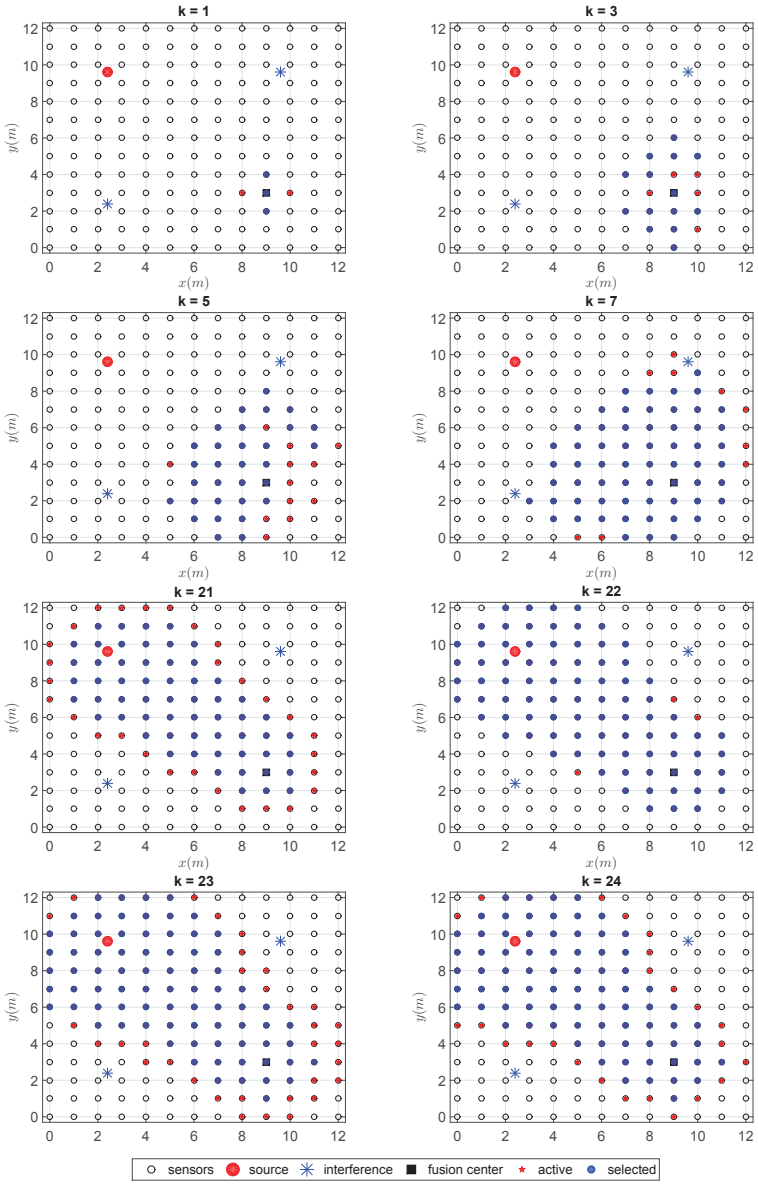


Figure 3.3: An illustration of the convergence behaviour for the proposed greedy algorithm (i.e., Algorithm 1). The initial point is located at (9, 3) m.

the candidate set $\mathscr{S}_1$ towards all possible directions. Instead, $\mathscr{S}_1$ is increased only in the informative direction to the source location, such that the less informative microphones are not included. Furthermore, notice that the final selected set $\mathscr{S}_2$ differs slightly from the model-driven approach in Fig. 3.2(e), as the greedy approach does not select the sen-

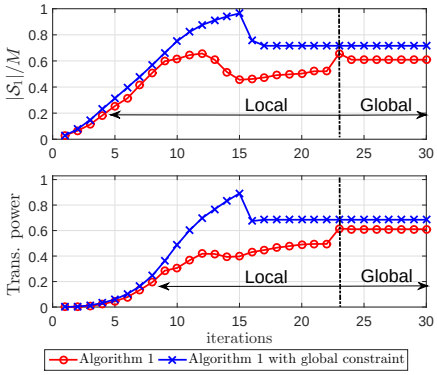


Figure 3.4: Cardinality of candidate set and transmission power vs iterations.

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

In Fig. 3.4, we show the ratio of cardinality of the candidate set $\mathcal{S}_1$ to the total number of sensors M and transmission power per iteration. The combination of the global and local constraint is compared to a greedy algorithm that uses only the global constraint for Algorithm 1. Using only the global constraint, $\mathcal{S}_1$ would blindly increase towards all directions. Clearly, we see that by using a combination between the local and the global constraint, much less sensors are included per iteration, such that the transmission power is kept low.

INITIALIZATIONS

In this part, we will show the effect of the initial point z_0 on the convergence rate. Fig. 3.5 illustrates the output noise power (in dB) in terms of iterations for four different initializations, i.e., centre (6, 6) m, source position (2.4, 9.6) m, interference position (2.4, 2.4) m and FC (9, 3) m. The red dashed line represents the performance of the model-driven algorithm proposed in Sec. 3.4, which selects the most informative sensors from all the possible candidates. The black dashed line denotes the performance of the classical MVDR beamformer using all microphones. The magenta curve shows the proposed greedy algorithm for the MVDR beamformer. The blue dashed curve denotes the performance of the utility-based algorithm [60, 101]. The output noise power of the greedy algorithm includes two steps: local constraint ($\beta_{\mathcal{S}_1}/\alpha$) and global constraint (β/α). The moment that the constraint is switched from the local to the global constraint is indicated by the red marker "×". When executing the local constraint, the output noise power decreases fastest for the initialization at the source position and slowest for the FC

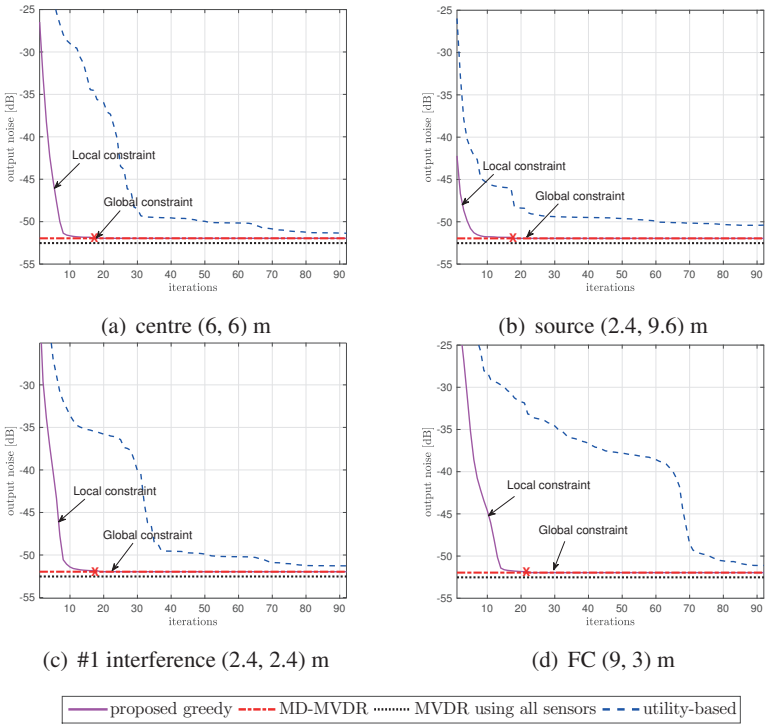


Figure 3.5: Output noise power in terms of iterations for different initial point z_0: (a) centre, (b) source, (c) interference, (d) FC.

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

Furthermore, from Fig. 3.5 we observe that the proposed greedy algorithm converges with much less iterations as compared to the utility-based method, because the latter only selects one sensor in each iteration. Note that in the comparisons the total transmission cost budgets for the two approaches are kept the same. Also, there is no guarantee for the utility-based method to fulfill the expected noise reduction performance. Given the same transmission cost budget, the proposed greedy algorithm can therefore obtain more reduction in noise power and converge much faster in terms of iterations.

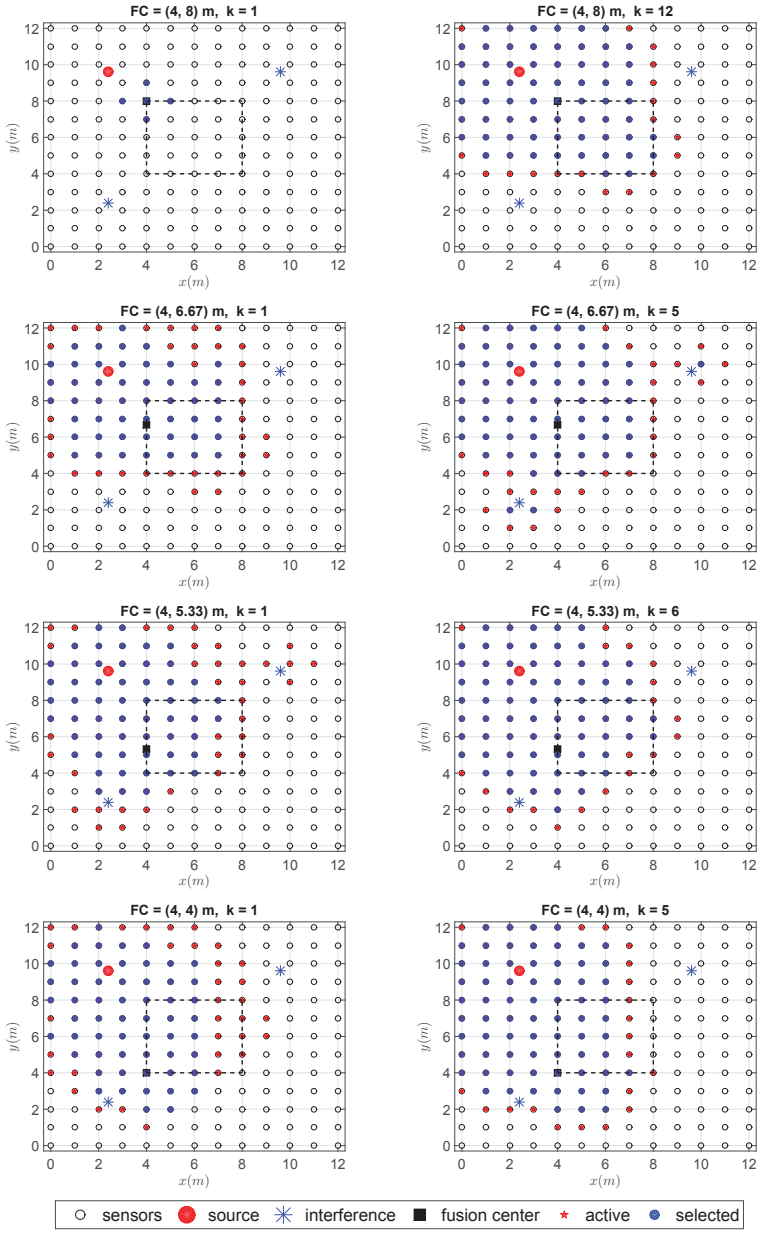


Figure 3.6: An illustration of the sensor selection based on the proposed greedy algorithm for the moving FC.

MOVING FC

In this part, we will show the advantage of the greedy algorithm in a dynamic scenario with a moving FC. In practice, the FC could be moving, because usually it is regarded

as a mobile user. Fig. 3.6 shows an example of greedy sensor selection for a moving FC, where the FC moves along the black dashed rectangle. The starting point is located at (4, 8) m, and at this position it takes 12 steps (9 steps for the local constraint and 3 steps for the global constraint) for the greedy algorithm to converge to a feasible informative set. The changing trend of the previous 11 steps is similar to Fig. 3.3, so we merely show the results of the steps 1 and 12 in the left top subplot in Fig. 3.6. The FC then slowly moves to the next position (4, 6.67) m. For the second position, we use the selected microphone set from the first position to update the candidate set, and then solve (3.32). It is found that only 5 iterations (1 for the local constraint and 4 for the global constraint) are required to obtain convergence. Subsequently, the FC continues moving. For the next positions, the greedy algorithm only requires about 6 iterations to converge. Hence, in the dynamic scenario with a moving FC, the proposed greedy approach can significantly save computational resources. Since the interferences are Gaussian shaped noise sources, once the noise correlation matrix $\mathbf{R_{nn}}$ is estimated using the noise-only segments before the FC starts to move, it can still be used for the subsequent positions of the FC. Hence, for the moving FC case, we only need to update $\mathbf{R_{yy}}$ or $\mathbf{R_{xx}}$ based on the real-time recordings. It is also noteworthy that the FC is not a microphone and the ATFs (i.e., the steering vector $\mathbf{a}$) stay the same even when the FC is moving, since the positions of microphones and the target source are fixed.

An interesting phenomenon occurs in Fig. 3.6. As the FC moves further away from the source, we can clearly see the importance of the sensors that are close to the interference. When the FC is located at (4, 6.67) m, two sensors close to the interference are also selected. This cannot be distinguished when FC = (4, 8) m, where the FC is closer to the source. Hence, we can conclude that the sensors that are close to the source, to the FC and to the interference are informative, and they are helpful to enhance the target source, to save transmission costs and to cancel the interfering sources, respectively.

3.6.5. COMPLEXITY ANALYSIS

In this subsection, we will compare the computational complexity of the greedy algorithms to that of the model-driven approach. For the model-driven approach, its complexity is of the order of $\mathcal{O}(M^3)$, so we use M^3 in the worst case for analysis without loss of generality. For the proposed greedy algorithm (i.e., Algorithm 1), suppose that J iterations are required to converge, in each iteration its complexity is of the order of $\mathcal{O}(|\mathcal{S}_1|^3)$, thus we can use $\sum_{j=1}^{J}|\mathcal{S}_1|^3$ to represent its computational complexity. For the utility-based greedy algorithm (i.e., Algorithm 2), we can find that its computational complexity is of the order of $\mathcal{O}(|\mathcal{S}_2|^2(|\mathcal{S}_1|-|\mathcal{S}_2|))$ for each iteration from [101], thus $\sum_{j=1}^{J}|\mathcal{S}_2|^2(|\mathcal{S}_1|-|\mathcal{S}_2|)$ can be exploited to represent its total complexity.

Fig. 3.7 compares the execution time of the two aforementioned greedy strategies. The execution time is normalized by the runtime of model-driven method, whose runtime is 1 as benchmark. From Fig. 3.7, we can see that the execution time of the proposed greedy algorithm depends on the initial point z_0, as it will be more expensive for the initial points that are further from the target source. Furthermore, for most initial points the proposed algorithm is computationally more efficient than the utility-based method, because we need much less iterations (20 iterations compared to 90 iterations approximately which has already been demonstrated in Fig. 3.5).

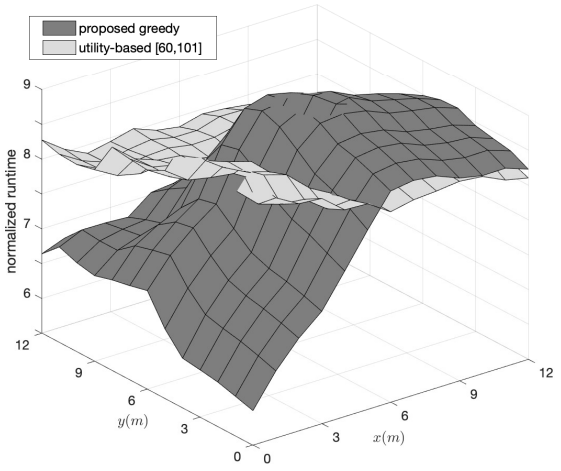


Figure 3.7: True normalized execution time in terms of different initial points.

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

3.7. CONCLUSION

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

algorithm is still effective to handle the microphone subset selection problem.

4

RATE-DISTRIBUTED SPATIAL FILTERING BASED NOISE REDUCTION IN WASNS

This chapter is based on the article published as "Rate-Distributed Spatial Filtering Based Noise Reduction in Wireless Acoustic Sensor Networks" by J. Zhang, R. Heusdens, and R. C. Hendriks in *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2015–2026, 2018.

4.1. INTRODUCTION

RECENTLY, wireless acoustic sensor networks (WASNs) have attracted an increasing amount of interest [33, 90, 97]. Compared to conventional microphone arrays with a fixed configuration, WASNs have advantages in array-size limitation and scalability of the networks. In a WASN, each sensor node is equipped with a single microphone or a small microphone array, and the nodes are spatially distributed across a specific environment. Due to the fact that the microphone nodes in a WASN can be placed anywhere, the sound field is sampled in a much larger area. It is possible that some of the nodes are close to the target source(s) and have higher signal-to-noise ratio (SNR), such that higher quality recordings can be obtained. In a WASN, the microphone nodes are connected to their neighboring nodes or a fusion center (FC) using wireless links, resulting in a distributed or centralized framework, respectively. In this work, we will mainly focus on the centralized framework, where each node samples and quantizes the microphone recordings, and transmits them to a remote FC. The tasks of interest, e.g., signal estimation or binaural cue preservation, are assumed to occur at the FC.

In WASNs, each sensor node is usually battery powered having a limited energy budget. It is therefore important to take the energy consumption into account in the design of algorithms. Generally, the energy usage within the context of WASNs can be linked to two processes: data transmission and data processing [34, 35]. The data transmission occurs between the nodes and the FC, and data processing at the FC end. Usually, data exchange is more expensive than data processing in terms of energy usage.

In order to reduce the energy usage in WASNs, there are two techniques that can be employed: sensor selection [53, 122, 52, 101, 60, 59, 58] and rate allocation [82, 70, 57]. For sensor selection, the most informative subset of sensors is chosen by maximizing a performance criterion while constraining the cardinality of the selected subset, or by minimizing the cardinality while constraining the performance. In this way, the number of sensors contained in the selected subset can be much smaller than the total set of sensors, resulting in a sparse selection. Due to the fact that only the selected sensors need to transmit their recordings to the FC, sensor selection is an effective way to save the energy usage.

Compared to sensor selection, rate allocation allows for a more smooth operating curve as sensors are not selected to only operate at full rate or zero rate (when not selected), but at any possible rate. For rate allocation, the idea is to allocate higher rates to the more informative sensors while lower or zero rates are allocated to the others. There are many studies on quantization for signal estimation in the context of wireless sensor networks, see [123, 124] and reference therein, typically under the assumption that the measurement noise across sensors is mutually uncorrelated. These models are not suitable for realistic audio applications, e.g., speech enhancement, where the noise is typically correlated across sensors because of the presence of directional interfering sources. In [125, 70], the effect of a bit-rate constraint was investigated for noise reduction in WASNs. In [82], rate-constrained collaborative noise reduction for wireless hearing aids (HAs) was studied from an information-theoretic standpoint, resulting in an information transmission strategy between two nodes. However, the approach proposed in [82] requires full binaural statistics which are difficult to estimate in a practical setting. In [57], a greedy quantization method was proposed for speech signal estima-

tion based on a so-called signal utility, which indeed represents the importance of microphone recordings. However, it only decreases/increases one bit for a node at each iteration, resulting in low convergence speed.

The difference between sensor selection and rate allocation problems lies in binary versus more smooth decisions. Given a maximum bit rate, the sensor selection approaches choose a subset of sensors first, and the selected sensors then communicate with the FC using the maximum rate. That is, each sensor only makes a binary decision on the communication rate, i.e., zero or maximum rate. In contrast to sensor selection, rate allocation approaches can execute multiple decisions on the rate, i.e., any bit rate can be fractional from zero bit rate to the maximum bit rate. If a sensor is allocated zero bits, it will not be selected. Hence, in general, rate allocation approaches do not lead to a WASN that is as sparse as the one that is obtained by the sensor selection approaches, but they can better reduce energy consumption used for transmission. On the other hand, sensor selection approaches could save more energy usage for data processing at the FC end, as typically less measurements are involved in computations.

In this work, we will only consider the energy usage for data transmission and neglect the energy usage for other processes. The wireless transmission power is regarded as a function of the distance between sensor nodes and the FC and the rate (i.e., bits per sample) which is used to quantize the signals to be transmitted. We intend to reduce energy usage from the perspective of rate allocation for spatial filtering based noise reduction in WASNs. The total wireless transmission costs are minimized by constraining the performance of the output noise power. Using a linearly constrained minimum variance (LCMV) beamformer, the problem is solved by convex optimization techniques. After the bit rates are determined, each microphone node uniformly quantizes and transmits its recordings to the FC for the signal processing tasks at hand.

4.1.1. CONTRIBUTIONS

The contributions of the paper can be summarized as follows. Firstly, we design a rate allocation strategy for rate-distributed LCMV (RD-LCMV) beamforming in WASNs by minimizing the energy usage and constraining the noise reduction performance. The original non-convex optimization problem is relaxed using convex relaxation techniques and reformulated as semi-definite programming. Based on numerical results in simulated WASNs, we find that the microphone nodes that are close to the sources (including target sources and interferers) and the FC are more likely to be allocated with more bit rates, because they have more information on SNR and cost less energy, respectively.

Secondly, we extend the model-driven microphone subset selection approach for minimum variance distortionless response (MD-MVDR) beamformer from [122] to the LCMV beamforming framework (referred as MD-LCMV). By doing so, we find the link between rate allocation and sensor selection problems, i.e., rate allocation is a generalization of sensor selection. In [122], the best microphone subset is chosen by minimizing the total transmission costs and constraining the noise reduction performance, where the transmission cost between each node and the FC is only considered as a function of distance. The selected microphone will communicate with the FC using the maximum bit rate. The energy model of the approach in the current paper is more general as compared to that in [122]. Based on the rates obtained by the proposed RD-LCMV approach,

the best microphone subset of MD-LCMV can be determined by putting a threshold on the rates, e.g., the sensors whose rates are larger than this threshold are chosen.

Finally, numerical simulations demonstrate that the selected microphone subsets resulting from thresholding the rates from the RD-LCMV method and directly applying MD-LCMV are completely the same. Both RD-LCMV and MD-LCMV can guarantee a given performance requirement, but RD-LCMV shows a superiority in energy efficiency.

4.1.2. OUTLINE AND NOTATION

The rest of this paper is organised as follows. Sec. 4.2 presents preliminary knowledge on the signal model, uniform quantization, the used energy model and LCMV beamforming. In Sec. 4.3, the problem formulation and a solver for the RD-LCMV optimization are given. Sec. 4.4 extends the sensor selection for MVDR beamforming from [122] to the LCMV beamforming framework and discusses the link between sensor selection and rate allocation problems. Sec. 4.5 shows the application of the proposed RD-LCMV method within the WASNs. Finally, Sec. 4.6 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. diag$(\cdot)$ refers to a block diagonal matrix with the elements in its argument on the main diagonal. $\mathbf{1}_N$ and $\mathbf{O}_N$ denote the $N \times 1$ vector of ones and the $N \times N$ matrix with all its elements equal to zero, respectively. $\mathbf{I}_N$ is an identity matrix of size N. $\mathbb{E}\{\cdot\}$ denotes the statistical expectation operation. $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. Finally, $\odot$ denotes the Hadamard (elementwise) product.

4.2. PRELIMINARIES

4.2.1. SIGNAL MODEL

We consider a spatially distributed candidate set of M microphone sensors that collect, quantize and transmit their observations to an FC. In the short-term Fourier transform (STFT) domain, let l denote the frame index and ω the frequency bin index, respectively. We assume that there are $\mathscr{I}$ speech sources of interest, while $\mathscr{J}$ interfering sources are potentially present in the environment. Using an STFT-domain description, the noisy DFT coefficient of the quantized signal which is to be transmitted to the FC at the kth microphone, say $\hat{Y}_k(\omega, l), k = 1, 2, \cdots, M$, is given by

$$\hat{Y}_k(\omega, l) = Y_k(\omega, l) + Q_k(\omega, l), \forall k, \tag{4.1}$$

where $Q_k(\omega, l)$ denotes the quantization noise which is assumed to be uncorrelated with the microphone recording[1] $Y_k(\omega, l)$[2], given by

$$Y_k(\omega, l) = \sum_{i=1}^{\mathscr{I}} \underbrace{a_{ik}(\omega) S_i(\omega, l)}_{X_{ik}(\omega, l)} + \sum_{j=1}^{\mathscr{J}} \underbrace{b_{jk}(\omega) U_j(\omega, l)}_{N_{jk}(\omega, l)} + V_k(\omega, l), \tag{4.2}$$

[1] This assumption holds under high rate communication. Under low rate, this can be achieved using subtractive dither [70, 71].

[2] In real-life applications, y_k is already quantized, since it is acquired by the analog-to-digital converter (ADC) of the kth microphone. In this case, Q_k would represent the error from changing the bit resolution of Y_k.

with

- $a_{ik}(\omega)$ denoting the acoustic transfer function (ATF) of the ith target signal with respect to the kth microphone;
- $S_i(\omega,l)$ and $X_{ik}(\omega,l)$, the ith target source at the source location and the ith target source at the kth microphone, respectively;
- $b_{jk}(\omega)$ the ATF of the jth interfering source with respect to the kth microphone;
- $U_j(\omega,l)$ and $N_{ik}(\omega,l)$, the jth interfering source at the source location and the jth interference source at the kth microphone, respectively;
- $V_k(\omega,l)$ uncorrelated noise at the kth microphone.

Notice that in (4.2), we assume that the ATFs are shorter than the length of the STFT window, such that the ATFs can be modelled as a multiplicative factor that varies with frequency in the STFT domain. For longer ATFs, a more accurate signal model is required for each frequency band, e.g., see [126]. For notational convenience, we will omit the frequency variable ω and the frame index l now onwards bearing in mind that the processing takes place in the STFT domain. Using vector notation, the M channel signals are stacked in a vector $\hat{\mathbf{y}} = [\hat{Y}_1, ..., \hat{Y}_M]^T \in \mathbb{C}^M$. Similarly, we define M dimensional vectors $\mathbf{y}, \mathbf{x}_i, \mathbf{n}_j, \mathbf{v}, \mathbf{q}$ for the microphone recordings, the ith target component, the jth interfering component, the additive noise and the quantization noise, respectively, such that the signal model in (4.1) can compactly be written as

$$\hat{\mathbf{y}} = \mathbf{y} + \mathbf{q} = \sum_{i=1}^{\mathscr{I}} \mathbf{x}_i + \sum_{j=1}^{\mathscr{J}} \mathbf{n}_j + \mathbf{v} + \mathbf{q}, \tag{4.3}$$

where $\mathbf{x}_i = \mathbf{a}_i s_i \in \mathbb{C}^M$ and $\mathbf{n}_j = \mathbf{b}_j u_j \in \mathbb{C}^M$ with

$$\mathbf{a}_i = [a_{i1}, a_{i2}, \cdots, a_{iM}]^T, \quad \mathbf{b}_j = [b_{j1}, b_{j2}, \cdots, b_{jM}]^T.$$

Alternatively, if we stack the ATFs for the target sources and the interfering sources, in matrices, the microphone recordings can also be written like,

$$\mathbf{y} = \mathbf{A}\mathbf{s} + \mathbf{B}\mathbf{u} + \mathbf{v}, \tag{4.4}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_{\mathscr{I}}^T \end{bmatrix}^T \in \mathbb{C}^{M\times\mathscr{I}}, \quad \mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{\mathscr{I}} \end{bmatrix} \in \mathbb{C}^{\mathscr{I}}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \vdots \\ \mathbf{b}_{\mathscr{J}}^T \end{bmatrix}^T \in \mathbb{C}^{M\times\mathscr{J}}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{\mathscr{J}} \end{bmatrix} \in \mathbb{C}^{\mathscr{J}}.$$

In order to focus on the concept of rate-distributed noise reduction, we assume in this work that the ATFs of the existing sources (i.e., $\mathbf{A}$ and $\mathbf{B}$) are known.

Assuming that the target signals and the interferers are mutually uncorrelated, the correlation matrix of the recorded signals is given by

$$\mathbf{R}_{\mathbf{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R}_{\mathbf{xx}} + \underbrace{\mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{vv}}}_{\mathbf{R}_{\mathbf{nn}}} \in \mathbb{C}^{M\times M}, \tag{4.5}$$

where $\mathbf{R}_{\mathbf{xx}} = \sum_{i=1}^{\mathscr{I}} \mathbb{E}\{\mathbf{x}_i\mathbf{x}_i^H\} = \sum_{i=1}^{\mathscr{I}} \sigma_{s_i}^2 \mathbf{a}_i\mathbf{a}_i^H = \mathbf{A}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{A}^H$ with $\sigma_{s_i}^2 = \mathbb{E}\{|S_i|^2\}$ the power spectral density (PSD) of the ith target source and $\mathbf{\Sigma}_{\mathbf{x}} = \text{diag}\left([\sigma_{s_1}^2, \cdots, \sigma_{s_{\mathscr{I}}}^2]\right)$. Similarly, $\mathbf{R}_{\mathbf{uu}} = \sum_{j=1}^{\mathscr{J}} \mathbb{E}\{\mathbf{n}_i\mathbf{n}_i^H\} = \sum_{j=1}^{\mathscr{J}} \sigma_{u_i}^2 \mathbf{b}_j\mathbf{b}_j^H = \mathbf{B}\mathbf{\Sigma}_{\mathbf{u}}\mathbf{B}^H$ with $\sigma_{u_i}^2 = \mathbb{E}\{|U_i|^2\}$ the PSD of the jth interfering source and $\mathbf{\Sigma}_{\mathbf{u}} = \text{diag}\left([\sigma_{u_1}^2, \cdots, \sigma_{u_{\mathscr{J}}}^2]\right)$. The correlation matrix of all disturbances including quantization noise in the quantized signals $\hat{\mathbf{y}}$ is given by

$$\mathbf{R}_{\mathbf{n+q}} = \mathbf{R}_{\mathbf{nn}} + \mathbf{R}_{\mathbf{qq}}, \tag{4.6}$$

under the assumption that the received noises and quantization noise are mutually uncorrelated. In practice, $\mathbf{R}_{\mathbf{n+q}}$ can be estimated using the quantized noise-only segments of sufficient duration, and $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ can be estimated using the quantized speech+noise segments[3].

4.2.2. UNIFORM QUANTIZATION

The uniform quantization of a real number $a \in [-\frac{\mathscr{A}_k}{2}, \frac{\mathscr{A}_k}{2}]$ with $\mathscr{A}_k/2$ denoting the maximum absolute value of the kth microphone signal using b_k bits can be expressed as

$$Q(a) = \Delta_k \left(\left\lfloor \frac{a}{\Delta_k} \right\rfloor + \frac{1}{2} \right), \quad k = 1, \cdots, M, \tag{4.7}$$

where the uniform intervals have width $\Delta_k = \mathscr{A}_k/2^{b_k}$. Note that $\mathscr{A}_k$ is different from sensor to sensor which is determined by its own signal observations. Each sensor should inform its $\mathscr{A}_k$ to the FC by communication. Considering the case of uniform quantization, the variance or PSD of the quantization noise is approximately given by [68, 69]

$$\sigma_{q_k}^2 = \Delta_k^2/12, \quad k = 1, \cdots, M, \tag{4.8}$$

and the correlation matrix of the quantization noise across microphones reads

$$\mathbf{R}_{\mathbf{qq}} = \frac{1}{12} \times \text{diag}\left(\left[\frac{\mathscr{A}_1^2}{4^{b_1}}, \frac{\mathscr{A}_2^2}{4^{b_2}}, ..., \frac{\mathscr{A}_M^2}{4^{b_M}} \right] \right). \tag{4.9}$$

[3] Note that both $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{n+q}}$ have quantization noise included, i.e., $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbf{R}_{\mathbf{yy}} + \mathbf{R}_{\mathbf{qq}}$ and $\mathbf{R}_{\mathbf{n+q}} = \mathbf{R}_{\mathbf{nn}} + \mathbf{R}_{\mathbf{qq}}$. Given sufficiently long noise and noisy segments, the quantization noise will influence $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{nn}}$ in the same fashion by adding a same matrix $\mathbf{R}_{\mathbf{qq}}$. Therefore, the estimation of $\mathbf{R}_{\mathbf{xx}}$ is not dependent on the communication rate, because it is obtained by subtracting $\mathbf{R}_{\mathbf{n+q}}$ from $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$.

4.2.3. TRANSMISSION ENERGY MODEL

We assume that the noise on the communication channels between the sensors and the FC is additive and white Gaussian with PSD V_k. The channel power attenuation factor is d_k^r, where d_k is the transmission distance from the kth microphone to the FC and r is the path loss exponent (typically $2 \leq r \leq 6$) [36, 127]. Without loss of generality, we assume $r = 2$ in this work. The SNR[4] of the kth channel then is

$$\mathrm{SNR}_k = d_k^{-2} E_k / V_k, \tag{4.10}$$

where E_k represents the transmitted energy of the kth microphone node per sample. Assuming Gaussian distributions for the noise and transmitted signal, the maximum capacity of such a communication channel for a specific time-frequency bin is given by the Shannon theory [128]

$$b_k = \frac{1}{2}\log_2\left(1 + \mathrm{SNR}_k\right), \tag{4.11}$$

which implies that b_k bits per sample at most can reliably be transmitted from microphone k to the FC. Based on the SNR_k and b_k, the transmission energy from microphone k to the FC for a specific time-frequency bin can be formulated as

$$E_k = d_k^2 V_k (4^{b_k} - 1), \tag{4.12}$$

which is a commonly used transmission model [36, 37, 38]. The above transmission energy model holds under two conditions [36, 38]: 1) in the context of spectrum-limited applications (e.g., audio signal processing); 2) under the assumption that we quantize the microphone recordings at the channel capacity, which is in fact an ideal source/channel coding scheme, such that the quantized signals perfectly fit in the channel capacity.

4.2.4. LCMV BEAMFORMING

The well-known LCMV beamformer is a typical spatial filtering technique where the output noise energy is minimized under a set of linear constraints. These constraints can be used to preserve target sources, or steer zeros in the direction of interferences (i.e., to suppress noise signals). In the context of binaural noise reduction [47, 46, 80], LCMV beamforming can also be used to preserve certain interaural relations in order to preserve spatial cues. Mathematically, the LCMV beamformer can be formulated as

$$\hat{\mathbf{w}}_{\mathrm{LCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}} \mathbf{w}, \quad \text{s.t.} \quad \mathbf{\Lambda}^H \mathbf{w} = \mathbf{f}, \tag{4.13}$$

which has $\mathscr{U}$ equality constraints with $\mathbf{f} = [f_1, f_2, \cdots, f_{\mathscr{U}}]^T \in \mathbb{C}^{\mathscr{U}}$ and $\mathbf{\Lambda} \in \mathbb{C}^{M \times \mathscr{U}}$. More specifically, in case the LCMV beamformer is employed to suppress noise, matrix $\mathbf{\Lambda}$ can be constructed using $\mathbf{A}$ and all the entries in $\mathbf{f}$ are non-zero values [14, 20, 129]; in case the LCMV beamformer is used for joint noise reduction and spatial cue preservation in a binaural setup, $\mathbf{\Lambda}$ is constructed using the matrices $\mathbf{A}$ and $\mathbf{B}$, and the vector $\mathbf{f}$ will have some zeros corresponding to the interfering sources [47, 46]. To make the framework proposed in this paper more general, we therefore do not specify the structure of $\mathbf{\Lambda}$ or

[4] The SNR mentioned in this section is used to measure the noise level over the communication channels, which is different from the acoustic noise or acoustic SNR that will be discussed in the experiments.

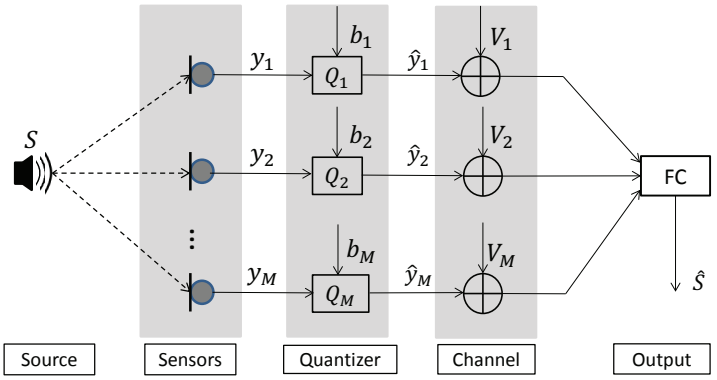


Figure 4.1: A typical communication model in WASNs.

f, which should be chosen according to the requirements in applications. The closed-form solution to (4.13), which can be found by applying Lagrange multipliers, is given by [14, 20, 129]

$$\hat{\mathbf{w}}_{\text{LCMV}} = \mathbf{R}_{\mathbf{n+q}}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\mathbf{R}_{\mathbf{n+q}}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f}. \tag{4.14}$$

The output noise power after LCMV beamforming can be shown to be given by [129]

$$\hat{\mathbf{w}}^H\mathbf{R}_{\mathbf{n+q}}\hat{\mathbf{w}} = \mathbf{f}^H\left(\mathbf{\Lambda}^H\mathbf{R}_{\mathbf{n+q}}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f}. \tag{4.15}$$

4.3. RATE-DISTRIBUTED LCMV BEAMFORMING

4.3.1. GENERAL PROBLEM FORMULATION

Fig. 4.1 shows a typical communication model in WASNs, which is considered in this work. The microphone recordings are quantized with specified bit rates and then transmitted to an FC through noisy communication channels. The FC conducts noise reduction and outputs the estimated target signal(s). In this work, we are interested in minimizing the transmission costs by allocating bit rates to microphones to achieve a prescribed noise reduction performance. Our initial goal can be formulated in terms of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b}} \quad & \sum_{k=1}^{M} d_k^2 V_k (4^{b_k} - 1) \\ \text{s.t.} \quad & \mathbf{w}^H\mathbf{R}_{\mathbf{n+q}}\mathbf{w} \le \frac{\beta}{\alpha} \\ & \mathbf{\Lambda}^H\mathbf{w} = \mathbf{f}, \\ & b_k \in \mathbb{Z}_+, \quad b_k \le b_0, \forall k, \end{aligned} \tag{P1}$$

where β denotes the minimum output noise power that can be achieved when all sensors use full-rate quantization, $\alpha \in (0,1]$ is to control a certain expected performance, $\mathbb{Z}_+$

denotes a non-negative integer set, and b_0 the maximum rate per sample of each microphone signal. The unknown variable $\mathbf{b}$ is implicit in the output noise power $\mathbf{w}^H\mathbf{R}_{\mathbf{n}+\mathbf{q}}\mathbf{w}$. Note that (P1) is a general form for the rate-distributed spatial filtering based noise reduction problem. Also, β/α does not depend on the rate allocation strategy or statistics of the whole sensor network, because β/α is just a number that can be assigned by users, e.g., 40 dB, to indicate an expected performance. By solving (P1), we can determine the optimal rate distribution that each microphone can utilize to quantize its recordings, such that the noise reduction system achieves a desired performance with minimum energy usage. One simple method to solve (P1) is exhaustive search, i.e., evaluating the performance for all $(b_0+1)^M$ choices for the rate distribution, but evidently this is intractable unless b_0 or M is very small. Next, we will find an efficient solver for (P1).

4.3.2. SOLVER FOR RATE-DISTRIBUTED LCMV BEAMFORMING

In this section, we will reformulate (P1) in the context of LCMV beamforming. Considering the utilization of an LCMV beamformer for noise reduction, the second constraint in (P1) is automatically satisfied. Substituting the solution of the LCMV beamformer from (4.14) into (P1), we get the following simplified optimization problem:

$$\begin{aligned} \min_{\mathbf{b}} \ & \sum_{k=1}^{M} d_k^2 V_k (4^{b_k}-1) \\ \text{s.t.} \ & \mathbf{f}^H \left(\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda}\right)^{-1} \mathbf{f} \leq \frac{\beta}{\alpha} \\ & b_k \in \mathbb{Z}_+, \quad b_k \leq b_0, \forall k, \end{aligned} \tag{P2}$$

where the bit rates $\mathbf{b}$ are implicit in the output noise power $\mathbf{f}^H \left(\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda}\right)^{-1} \mathbf{f}$, which is clearly non-convex and non-linear in terms of $\mathbf{b}$. In what follows, we will explicitly express $\mathbf{f}^H \left(\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda}\right)^{-1} \mathbf{f}$ in $\mathbf{b}$ and reformulate (P2) by semi-definite relaxation.

First of all, the first inequality constraint in (P2) is equivalent to the following two new constraints by introducing a new Hermitian positive definite matrix $\mathbf{Z} \in \mathbb{S}_{++}^{\mathscr{U}}$ with $\mathbb{S}_{++}^{\mathscr{U}}$ denoting a set for Hermitian positive definite matrices of dimension $\mathscr{U} \times \mathscr{U}$, i.e.,

$$\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda} = \mathbf{Z}, \tag{4.16}$$

$$\mathbf{f}^H \mathbf{Z}^{-1} \mathbf{f} \leq \frac{\beta}{\alpha}. \tag{4.17}$$

The inequality (4.17) can be rewritten as a linear matrix inequality (LMI) using the Schur complement [108, p.650], i.e.,

$$\begin{bmatrix} \mathbf{Z} & \mathbf{f} \\ \mathbf{f}^H & \frac{\beta}{\alpha} \end{bmatrix} \succeq \mathbf{O}_{\mathscr{U}+1}. \tag{4.18}$$

However, the equality constraint in (4.16) is clearly non-convex in terms of the unknowns $\mathbf{b}$. We therefore relax it to

$$\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda} \succeq \mathbf{Z}, \tag{4.19}$$

since (4.17) and (4.19) are sufficient conditions to obtain the original constraint in (P2), and we use $\succeq$ in (4.19) for convex relaxation.

Then, in order to linearize (4.19) in $\mathbf{b}$, we calculate $\mathbf{R}_{\mathbf{n+q}}^{-1}$ as

$$\mathbf{R}_{\mathbf{n+q}}^{-1} = \left(\mathbf{R}_{\mathbf{nn}} + \mathbf{R}_{\mathbf{qq}}\right)^{-1} = \mathbf{R}_{\mathbf{nn}}^{-1} - \mathbf{R}_{\mathbf{nn}}^{-1}\left(\mathbf{R}_{\mathbf{nn}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1}\right)^{-1}\mathbf{R}_{\mathbf{nn}}^{-1}, \tag{4.20}$$

where the second equality is derived from the matrix inversion lemma [107, p.18]

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

Substitution of the expression for $\mathbf{R}_{\mathbf{n+q}}^{-1}$ from (4.20) into (4.19), we obtain

$$\mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} - \mathbf{Z} \succeq \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\left(\mathbf{R}_{\mathbf{nn}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1}\right)^{-1}\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}. \tag{4.21}$$

Using the Schur complement, we obtain the following LMI[5]

$$\begin{bmatrix} \mathbf{R}_{\mathbf{nn}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1} & \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} \\ \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1} & \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{M+\mathscr{U}}, \tag{4.22}$$

where $\mathbf{R}_{\mathbf{qq}}^{-1}$ can be computed from (4.9) as

$$\mathbf{R}_{\mathbf{qq}}^{-1} = 12 \times \mathrm{diag}\left(\left[\frac{4^{b_1}}{\mathscr{A}_1^2}, \frac{4^{b_2}}{\mathscr{A}_2^2}, ..., \frac{4^{b_M}}{\mathscr{A}_M^2}\right]\right). \tag{4.23}$$

For notational convenience, we define a constant vector $\mathbf{e} = \left[\frac{12}{\mathscr{A}_1^2}, \cdots, \frac{12}{\mathscr{A}_M^2}\right]$. Further, we introduce a variable change $t_k = 4^{b_k} \in \mathbb{Z}_+, \forall k$, such that $\mathbf{R}_{\mathbf{qq}}^{-1} = \mathrm{diag}(\mathbf{e} \odot \mathbf{t})$ and (4.22) are both linear in $\mathbf{t}$. In order to convexify 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 arrive at

$$\min_{\mathbf{t},\mathbf{Z}} \sum_{k=1}^{M} d_k^2 V_k (t_k - 1) \tag{4.24}$$

$$\text{s.t.} \quad \begin{bmatrix} \mathbf{Z} & \mathbf{f} \\ \mathbf{f}^H & \frac{\beta}{\alpha} \end{bmatrix} \succeq \mathbf{O}_{\mathscr{U}+1}, \tag{4.24a}$$

$$\begin{bmatrix} \mathbf{R}_{\mathbf{nn}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1} & \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} \\ \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1} & \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{M+\mathscr{U}}, \tag{4.24b}$$

$$1 \le t_k \le 4^{b_0}, \ \forall k, \tag{4.24c}$$

which is a standard semi-definite programming problem [108, p.128] and can be solved efficiently in polynomial time using interior-point methods or solvers, like CVX [110] or SeDuMi [111]. The computational complexity for solving (4.24) is of the order of $\mathscr{O}((M+\mathscr{U})^3)$. After (4.24) is solved, the allocated bit rates can be resolved by $b_k = \log_4 t_k, \forall k$ which are continuous values.

[5] Note that (4.22) is not an LMI essentially, because it is not linear in the unknown parameters $\mathbf{b}$. Here, we call it LMI for convenience, since it looks like an LMI and is linear in $4^{b_k}, \forall k$.

4.3.3. RANDOMIZED ROUNDING

The solution provided by the semi-definite program in (4.24) consists of continuous values. A straightforward and often used technique to resolve the integer bit rates is by simply rounding, in which the integer estimates are given by $\text{round}(b_k), \forall k$ where the round$(\cdot)$ operator rounds its arguments towards the nearest integer. However, there is no guarantee that the integer solution obtained by this rounding technique always satisfies the performance constraint. Hence, we utilize a variant rounding technique, i.e., randomized rounding [53], to the estimates obtained from (4.24). Specifically, letting $\text{ceil}(b_k) - b_k$ and $1 - \text{ceil}(b_k) + b_k, \forall k$ denote the probabilities for b_k to be the nearest lower integer and the nearest upper integer, respectively, where the ceil$(\cdot)$ operator rounds its arguments towards the nearest upper integer, then we can randomly round b_k to the nearest upper/lower integer based on its probability distribution and the prescribed performance requirement. Usually, such a randomized rounding procedure needs to be performed multiple times, and the best solution is then selected. Alternatively, we can simply use $\text{ceil}(b_k), \forall k$ to resolve the integer rates. However, this is suboptimal compared to the randomized rounding technique due to more unnecessary energy usage.

4.4. RELATION TO MICROPHONE SUBSET SELECTION

In this section, we will show the relation between rate allocation and sensor selection. To do so, we first represent the rate-distributed LCMV beamforming in (4.24) as a Boolean optimization problem, and then we extend the sensor selection based MVDR beamformer from [122] to the LCMV beamforming framework. We find that sensor selection is a special case of the rate allocation problem. Finally, we propose a bisection algorithm that can be used to obtain the sensor selection results as in [122] based on the rate allocation method.

4.4.1. REPRESENTATION OF RATE-DISTRIBUTED LCMV BEAMFORMING

In this subsection, we will represent the rate-distributed LCMV beamforming in (4.24) from the perspective of Boolean optimization. This representation turns out to be very useful when comparing the rate-distributed LCMV beamforming framework to the LCMV beamforming based sensor selection framework. Setting $p_k = t_k/4^{b_0}, \forall k$ in (4.24), we obtain the following equivalent form

$$\min_{\mathbf{p},\mathbf{Z}} \; 4^{b_0} \sum_{k=1}^{M} p_k V_k d_k^2 - \varepsilon \tag{4.25}$$

$$\text{s.t.} \quad \begin{bmatrix} \mathbf{Z} & \mathbf{f} \\ \mathbf{f}^H & \frac{\beta}{\alpha} \end{bmatrix} \succeq \mathbf{O}_{\mathscr{U}+1}, \tag{4.25a}$$

$$\begin{bmatrix} \mathbf{R}_{\mathbf{nn}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1} & \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} \\ \mathbf{\Lambda}^H \mathbf{R}_{\mathbf{nn}}^{-1} & \mathbf{\Lambda}^H \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{M+\mathscr{U}}, \tag{4.25b}$$

$$0 \le p_k \le 1, \forall k, \tag{4.25c}$$

where $\mathbf{R}_{\mathbf{qq}}^{-1} = 4^{b_0}\text{diag}(\mathbf{e} \odot \mathbf{p})$ and $\varepsilon = \sum_{k=1}^{M} d_k^2 V_k$ which is an irrelevant constant that does not depend on the optimization variables. Note that for (4.25), optimizing the objective

function is equivalent to minimizing $\sum_{k=1}^{M} p_k V_k d_k^2$. Given the solution of (4.25), the rates to be allocated can be resolved by $b_k = \log_4 p_k + b_0, \forall k$ and the randomized rounding technique in Sec. 4.3.3.

Remark 5. *From the perspective of optimization, (4.24) and (4.25) are equivalent, i.e., both are semi-definite programming problems with the same computational complexity and can provide the optimal rate distribution. However, apart from the function of rate allocation, (4.25) gives an insight to sensor selection, because its unknowns* $\mathbf{p}$ *are continuous values between 0 and 1. Hence, if we apply the randomized rounding technique to the continuous* $\mathbf{p}$*, we can obtain a Boolean solution which can indicate whether a sensor is selected or not. In other words, if we are interested in sparsity-aware networks instead of energy-aware ones, (4.25) can be employed to select the best microphone subset.*

4

Based on the representation of rate-distributed LCMV beamforming in (4.25), next we will find the relation between rate allocation and sensor selection.

4.4.2. MODEL-DRIVEN LCMV BEAMFORMING

In [122], we considered the problem of microphone subset selection based noise reduction in the context of MVDR beamforming. We minimized the transmission costs by constraining to a desired noise reduction performance. The transmission cost was related to the distance between each microphone and the FC. In the case the number of constraints in (4.13) is reduced to a single constraint preserving a single target, the LCVM beamformer reduces to a special case, i.e., the MVDR beamformer. Hence, mathematically, the original sensor selection problem in [122] can be extended by adding more linear constraints to obtain the following optimization problem

$$\begin{aligned} \min_{\mathbf{w_p},\mathbf{p}} \ & \sum_{k=1}^{M} p_k d_k^2 \\ \text{s.t.} \ & \mathbf{w_p}^H \mathbf{R}_{\mathbf{n+q},\mathbf{p}} \mathbf{w_p} \leq \frac{\beta}{\alpha}, \\ & \mathbf{\Lambda}_\mathbf{p}^H \mathbf{w_p} = \mathbf{f}, \end{aligned} \tag{4.26}$$

where $\mathbf{p} = [p_1, \cdots, p_M]^T \in \{0,1\}^M$ are selection variables to indicate whether a sensor is selected or not, $\mathbf{w_p}$ denotes the coefficients of the LCMV beamformer corresponding to the selected sensors, $\mathbf{\Lambda_p}$ is a submatrix of $\mathbf{\Lambda}$ which was defined in (4.13), and other parameters are defined similarly as in (P1). Note that the transmission cost in (4.26) is only influenced by the transmission distance, since we assume that all the selected sensors use a full-rate quantization, such that we do not need the ideal source/channel coding assumption for the sensor selection problem and the channel noise $V_k, \forall k$ is neglected. Suppose that for the microphone subset selection problem, all the candidate sensors use the maximum rates, i.e., b_0 bits per sample, to communicate with the FC, such that $\mathbf{R_{n+q}} = \mathbf{R_{nn}} + \mathbf{R_{qq}}$ and $\mathbf{R_{qq}} = \frac{1}{12} \times \text{diag}\left(\left[\frac{\mathscr{A}_1^2}{4^{b_0}}, \frac{\mathscr{A}_2^2}{4^{b_0}}, ..., \frac{\mathscr{A}_M^2}{4^{b_0}}\right]\right)$. The problem (4.26) is called model-driven LCMV beamforming, because it is based on the statistical knowledge $\mathbf{R_{n+q}}$.

We will show that the optimization problem in (4.26) can be solved by considering (4.25). Let diag($\mathbf{p}$) be a diagonal matrix whose diagonal entries are given by $\mathbf{p}$, such that

$\mathbf{\Phi_p} \in \{0,1\}^{K\times M}$ is a submatrix of diag($\mathbf{p}$) after all-zero rows (corresponding to the unselected sensors) have been removed. As a result, we can easily get the following relationships

$$\mathbf{\Phi_p}\mathbf{\Phi_p}^T = \mathbf{I}_K, \;\; \mathbf{\Phi_p}^T\mathbf{\Phi_p} = \text{diag}(\mathbf{p}). \tag{4.27}$$

Therefore, applying the selection model to the classical LCMV beamformer in (4.14), the best linear unbiased estimator for a subset of K microphones determined by $\mathbf{p}$ will be

$$\hat{\mathbf{w}}_\mathbf{p} = \mathbf{R}_{\mathbf{n+q},\mathbf{p}}^{-1}\mathbf{\Lambda_p}\left(\mathbf{\Lambda_p}^H\mathbf{R}_{\mathbf{n+q},\mathbf{p}}^{-1}\mathbf{\Lambda_p}\right)^{-1}\mathbf{f}, \tag{4.28}$$

where $\mathbf{R}_{\mathbf{n+q},\mathbf{p}} = \mathbf{\Phi_p}\mathbf{R}_{\mathbf{n+q}}\mathbf{\Phi_p}^T$ represents the total noise correlation matrix of the selected sensors after the rows and columns of $\mathbf{R}_{\mathbf{n+q}}$ corresponding to the unselected sensors have been removed, i.e., $\mathbf{R}_{\mathbf{n+q},\mathbf{p}}$ is a submatrix of $\mathbf{R}_{\mathbf{n+q}}$.

Applying the result in (4.28) to (4.26) yields a simplified optimization problem based on the LCMV beamformer as

$$\begin{aligned} \min_{\mathbf{p}} \;& \sum_{k=1}^{M} p_k d_k^2 \\ \text{s.t.} \;& \mathbf{w}_\mathbf{p}^H\mathbf{R}_{\mathbf{n+q},\mathbf{p}}\mathbf{w}_\mathbf{p} \le \frac{\beta}{\alpha}, \end{aligned} \tag{4.29}$$

where similar to (4.15) the output noise power is given by

$$\mathbf{w}_\mathbf{p}^H\mathbf{R}_{\mathbf{n+q},\mathbf{p}}\mathbf{w}_\mathbf{p} = \mathbf{f}^H\left(\mathbf{\Lambda_p}^H\mathbf{R}_{\mathbf{n+q},\mathbf{p}}^{-1}\mathbf{\Lambda_p}\right)^{-1}\mathbf{f}. \tag{4.30}$$

By introducing a symmetric PSD matrix $\mathbf{Z} \in \mathbb{S}_+^{\mathscr{M}}$, we can rewrite the constraint in (4.29) into two new constraints in a similar way as in the previous section, i.e.,

$$\mathbf{\Lambda}^H\mathbf{R}_{\mathbf{n+q}}^{-1}\mathbf{\Lambda} = \mathbf{Z}, \tag{4.31}$$

$$\mathbf{f}^H\mathbf{Z}^{-1}\mathbf{f} \le \frac{\beta}{\alpha}. \tag{4.32}$$

The inequality in (4.32) can be rewritten as an LMI using the Schur complement, which is identical to (4.25a). Also, similar to Sec. 4.3, we relax the equality constraint in (4.31) to

$$\mathbf{\Lambda_p}^H\mathbf{R}_{\mathbf{n+q},\mathbf{p}}^{-1}\mathbf{\Lambda_p} \succeq \mathbf{Z}, \tag{4.33}$$

due to the non-convexity. The left-hand side of (4.33) can be calculated as

$$\begin{aligned} \mathbf{\Lambda_p}^H\mathbf{R}_{\mathbf{n+q},\mathbf{p}}^{-1}\mathbf{\Lambda_p} &\overset{(a)}{=} \mathbf{\Lambda}^H\mathbf{\Phi_p}^T\mathbf{R}_{\mathbf{n+q},\mathbf{p}}^{-1}\mathbf{\Phi_p}\mathbf{\Lambda} \\ &\overset{(b)}{=} \mathbf{\Lambda}^H\mathbf{\Phi_p}^T\left(\mathbf{\Phi_p}\mathbf{R}_{\mathbf{n+q}}\mathbf{\Phi_p}^T\right)^{-1}\mathbf{\Phi_p}\mathbf{\Lambda} \\ &\overset{(c)}{=} \mathbf{\Lambda}^H\mathbf{\Phi_p}^T\left(\mathbf{\Phi_p}\mathbf{R}_{\mathbf{nn}}\mathbf{\Phi_p}^T + \underbrace{\mathbf{\Phi_p}\mathbf{R}_{\mathbf{qq}}\mathbf{\Phi_p}^T}_{\mathbf{Q}}\right)^{-1}\mathbf{\Phi_p}\mathbf{\Lambda} \\ &\overset{(d)}{=} \mathbf{\Lambda}^H\left[\mathbf{R}_{\mathbf{nn}}^{-1} - \mathbf{R}_{\mathbf{nn}}^{-1}\left(\mathbf{R}_{\mathbf{nn}}^{-1} + \mathbf{\Phi_p}^T\mathbf{Q}^{-1}\mathbf{\Phi_p}\right)^{-1}\mathbf{R}_{\mathbf{nn}}^{-1}\right]\mathbf{\Lambda} \\ &\overset{(e)}{=} \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} - \mathbf{\Lambda}^H\mathbf{R}_{\mathbf{nn}}^{-1}\left(\mathbf{R}_{\mathbf{nn}}^{-1} + 4^{b_0}\text{diag}(\mathbf{p}\odot\mathbf{e})\right)^{-1}\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}, \end{aligned} \tag{4.34}$$

where (c) constructs $\mathbf{\Phi_p R_{qq} \Phi_p^T}$ as a new diagonal matrix $\mathbf{Q} \in \mathbb{R}^{K\times K}$ whose diagonal entries correspond to the selected sensors, (d) is derived based on the matrix inversion lemma [107, p.18][6], and (e) holds when $\mathbf{p}$ contains Boolean variables.

Substitution of (4.34) into (4.33) and using the Schur complement, we can obtain an LMI which will be identical to (4.25b). Altogether, we then reformulate the sensor selection problem for the LCMV beamforming as the following semi-definite program:

$$\min_{\mathbf{p},\mathbf{Z}} \sum_{k=1}^{M} p_k d_k^2 \tag{4.35}$$

$$\text{s.t.} \begin{bmatrix} \mathbf{Z} & \mathbf{f} \\ \mathbf{f}^H & \frac{\beta}{\alpha} \end{bmatrix} \succeq \mathbf{O}_{\mathscr{U}+1}, \tag{4.35a}$$

$$\begin{bmatrix} \mathbf{R_{nn}^{-1}} + \mathbf{R_{qq}^{-1}} & \mathbf{R_{nn}^{-1}\Lambda} \\ \mathbf{\Lambda}^H \mathbf{R_{nn}^{-1}} & \mathbf{\Lambda}^H \mathbf{R_{nn}^{-1}\Lambda} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{M+\mathscr{U}}, \tag{4.35b}$$

$$0 \le p_k \le 1, \forall k, \tag{4.35c}$$

where the Boolean variables $p_k, \forall k$ have already been relaxed by continuous surrogates. Comparing the rate allocation problem in (4.25) with the sensor selection problem in (4.35), we see that they only have difference in the cost functions. Intuitively, the sensor selection problem is equivalent to the rate allocation problem when all the communication channels have the same noise power, e.g., $V_k = 1, \forall k$. Based on this observation, it can be concluded that the sensor selection problem can be solved by the rate allocation algorithm. In other words, the proposed rate allocation approach is a generalization of the sensor selection method in [122].

4.4.3. THRESHOLD DETERMINATION BY BISECTION ALGORITHM

In Sec. 4.4.2, we have shown the relationship between the rate allocation problem and sensor selection, i.e., the former is a generalization of the latter problem, from a theoretical perspective. From this, we know that the best subset of microphones can be identified by the solution of rate distribution. Now, the essential question remaining is how to determine the selected sensors as in [122], based on the rate distribution presented in the current work. Here, we propose a bisection algorithm for threshold determination.

In detail, given the rate distribution $b_k, \forall k$ which is the solution of the problem (4.24) and the maximum rate b_0, first we set the threshold $T = \frac{b_0}{2}$, such that we choose a subset of sensors, say $\mathscr{S}$, whose rate is larger than T, that is, $\mathscr{S} = \{k | b_k \ge T\}$. If the performance using the sensors contained in the set $\mathscr{S}$, say τ, is larger than $\frac{\beta}{\alpha}$, we decrease T and update $\mathscr{S}$; if $\tau < \frac{\beta}{\alpha}$, we will increase T. This procedure continues until $\frac{\beta}{\alpha} - \tau \le \epsilon$ where ϵ is a predefined very small positive number. Furthermore, the best subset of microphones can also be found by solving the optimization problem in (4.25), while we need to apply the randomized rounding technique to resolve the Boolean variables $\mathbf{p}$.

[6] Based on the Woodbury identity $\left(\mathbf{A}+\mathbf{CBC}^T\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{A}^{-1}$, we can see that $\mathbf{C}\left(\mathbf{B}^{-1}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^T = \mathbf{A} - \mathbf{A}\left(\mathbf{A}+\mathbf{CBC}^T\right)^{-1}\mathbf{A}$. Taking $\mathbf{A} = \mathbf{R_{nn}^{-1}}, \mathbf{B} = \mathbf{Q}^{-1}$ and $\mathbf{C} = \mathbf{\Phi_p^T}$ and applying the Woodbury identity to the right side of the third equality in (4.34), we can obtain the fourth equality.

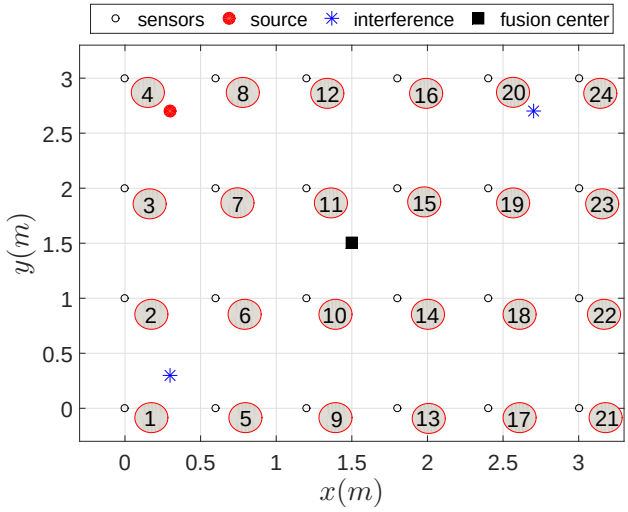


Figure 4.2: A typical WASN in a 2D scenario, where the indexes of microphones are labelled.

4.5. NUMERICAL RESULTS

In this section, we will show some numerical results for the proposed algorithm in terms of noise reduction in WASNs.

4.5.1. SINGLE TARGET SOURCE

Fig. 4.2 shows the experimental setup employed in the simulations, where 24 candidate microphones are placed uniformly in a 2D room with dimensions (3×3) m. The desired speech source (red solid circle) is located at $(0.3, 2.7)$ m. The FC (black solid square) is placed at the centre of the room. Two interfering sources (blue stars) are positioned at $(0.3, 0.3)$ m and $(2.7, 2.7)$ m, respectively. The target source signal is a 10 minute long concatenation of speech signals originating from the TIMIT database [120]. The interferences are stationary Gaussian speech shaped noise sources. The uncorrelated noise is modeled as microphone self noise at an SNR of 50 dB. All signals are sampled at 16 kHz. We use a square-root Hann window of 20 ms for framing with 50% overlap. The acoustic transfer functions are generated using [121] with reverberation time $T_{60} = 200$ ms. In order to focus on the rate-distributed spatial filtering issue, we assume that a perfect voice activity detector (VAD) is available in the sequel. Also, the microphone-to-FC distance $d_k, \forall k$ and the channel noise $V_k, \forall k$ are assumed to be known, e.g., $V_k = 1, \forall k$ without loss of generality. For the noise correlation matrix $\mathbf{R_{nn}}$, it is estimated at the FC end using sufficiently long noise-only segments when each node communicates with the FC at the maximum rate b_0 or larger.

An example of bit-rate allocation obtained by the rate-distributed LCMV beamforming and model-driven sensor selection based MVDR beamforming (referred to as MD-MVDR in short) [122] is shown in Fig. 4.3 with $\alpha = 0.8$. Since only one target source of interest exists, the optimization problem in (4.24) for the proposed method reduces to

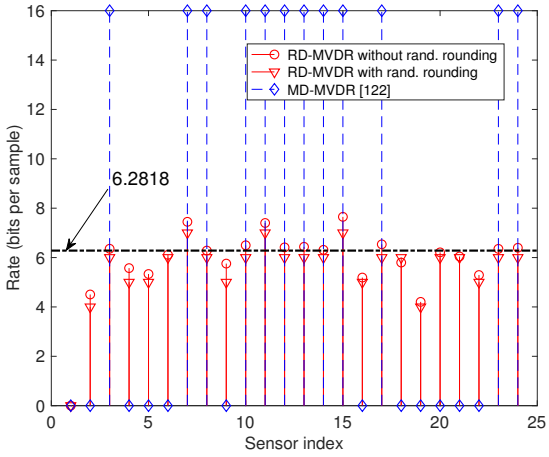


Figure 4.3: Example of rate allocation by the proposed approach (RD-MVDR) and sensor selection method (MD-MVDR). For the latter method, the selected sensors are allocated with b_0 bits, i.e., 16 bits per sample.

rate-distributed MVDR beamforming, which is referred to as RD-MVDR in short. From Fig. 4.3, it is observed that in order to fulfill the same performance, the proposed RD-MVDR method activates more sensors than the MD-MVDR. The MD-MVDR has a smaller cardinality of the selected subset. However, each active sensor obtained by RD-MVDR is allocated with a much lower bit-rate per sample compared to the maximum rates, i.e., $b_0 = 16$ bits. Also, the sensors that are close to the target source and the FC are more likely to be allocated with higher bit-rates, because they have a higher SNR and less energy costs, respectively. More importantly, we find a threshold for the rate distribution of RD-MVDR, e.g., 6.2818 bits, using the bisection algorithm from Sec. 4.4.3, and the active sensors whose rates are larger than this threshold are completely the same as the best subset obtained using the MD-MVDR algorithm. This phenomenon supports the conclusion that we have made in Sec. 4.4, i.e., the best microphone subset selection problem can be resolved by the rate allocation algorithm. Hence, given the solution of rate distribution, to find out the best microphone subset is equivalent to determining a bit-rate threshold.

In order to show the comparison of the proposed method in terms of noise reduction and energy usage, we also show the output noise power (in dB) and energy usage ratio (EUR) in terms of α in Fig. 4.4, where the indicator EUR is defined by

$$\text{EUR}_i = E_i / E_{\max}, \quad i \in \{\text{RD-MVDR, MD-MVDR}\}, \tag{4.36}$$

where E_i denotes the energy used by the RD-MVDR or MD-MVDR method, and $E_{\max}$ the maximum transmission energy when all the sensors are involved and communicate with the FC using b_0 bits. Clearly, the lower the EUR, the better the energy efficiency. In Fig. 4.4, we also compare to the desired maximum noise power, i.e., $10\log_{10}\frac{\beta}{\alpha}$. Note that β denotes the output noise power when using all sensors. Although this is hard to calculate in practice, in the simulations it can be estimated by including all sensors

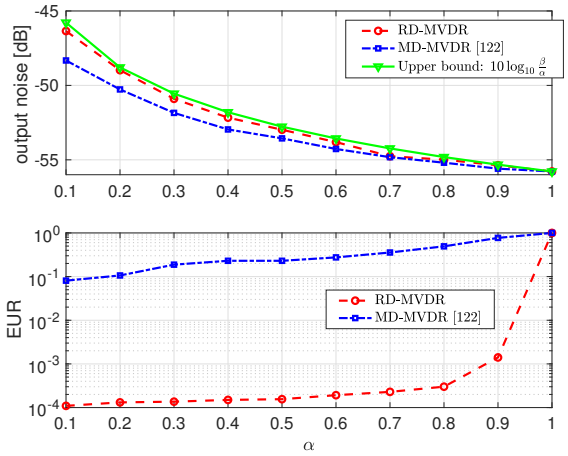


Figure 4.4: Output noise power and energy usage ratio (EUR) in terms of α. In the log-domain, the gap between the desired performance (i.e., β/α) and the maximum performance when using all sensors (i.e., β) will be $-10\log_{10}\alpha$.

and allocating each with b_0 bits. In practical applications, we just need to set a value for $10\log_{10}\frac{\beta}{\alpha}$, e.g., 40 dB, to constrain the desired performance. From Fig. 4.4, it follows that both RD-MVDR and MD-MVDR satisfy the performance requirement (i.e., below the upper bound $10\log_{10}\frac{\beta}{\alpha}$), while RD-MVDR is more efficient in the sense of energy usage, which is also explicit in the rate distribution in Fig. 4.3.

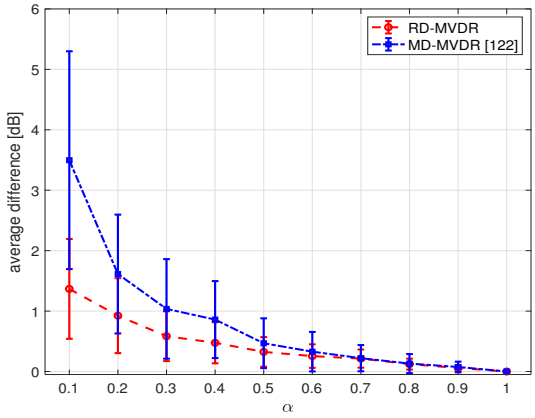


Figure 4.5: The average difference between the desired performance $10\log_{10}\frac{\beta}{\alpha}$ and output noise power of the RD-MVDR or MD-MVDR method in terms of α with random source/FC positions.

4.5.2. MONTE-CARLO SIMULATIONS

In order to give a more comprehensive comparison between rate allocation and sensor selection, we conduct Monte-Carlo simulations to show their average noise reduction performance. Considering the experimental setup in Fig. 4.2, we fix the microphone placement and the positions of the two interfering sources, but randomly choose the positions for the single target source and the FC. In Fig. 4.5, we show the average difference between the performance requirement $10\log_{10}\frac{\beta}{\alpha}$ and the output noise power of the RD-MVDR/MD-MVDR method in terms of the performance controller α, i.e., $10\log_{10}\frac{\beta}{\alpha}$ minus the output noise power of the RD-MVDR/MD-MVDR method, which is always positive. The results are averaged over 200 trails. It can be seen that with increasing α, the average difference for both RD-MVDR and MD-MVDR decreases. Compared to the MD-MVDR method, the RD-MVDR method achieves a smaller difference for all α-values, that is, the performance of the proposed rate-distributed approach is closer to the performance requirement.

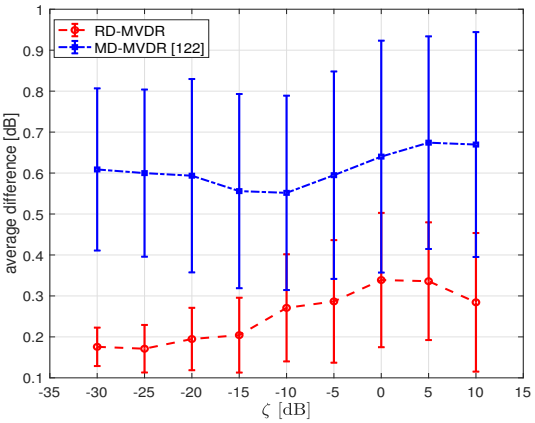


Figure 4.6: The average difference between the desired performance $10\log_{10}\frac{\beta}{\alpha}$ and output noise power of the RD-MVDR or MD-MVDR method in terms of the ATF errors with fixed source/FC positions and $\alpha = 0.6$.

In addition, in practice the ATFs are usually estimated by the generalized eigenvalue decomposition of the matrices $\mathbf{R_{nn}}$ and $\mathbf{R_{yy}}$ [29, 44]. The ATF estimation accuracy is affected by the estimation of the second-order statistics, i.e., VAD and available speech-absence/speech-presence durations. In order to analyze the robustness of the proposed approach to the ATF estimation errors in realistic scenarios, we conduct Monte-Carlo simulations. Considering that the ATF estimation of a single source (the setup is similar to Fig. 4.2) is given by $\hat{\mathbf{a}} = \mathbf{a} + \tilde{\mathbf{a}}$, where $\mathbf{a}$ and $\tilde{\mathbf{a}}$ represent the true ATF and the estimation error, respectively, we define

$$\zeta = 10\log_{10}\frac{\mathbb{E}[||\tilde{\mathbf{a}}||^2]}{||\mathbf{a}||^2}, \tag{4.37}$$

to measure the level of the estimation error. Given ζ in dB, we can generate $\tilde{\mathbf{a}}$ randomly based on zero-mean complex Gaussian distributions. Fig. 4.6 shows the average difference between the performance requirement and the aforementioned methods in terms

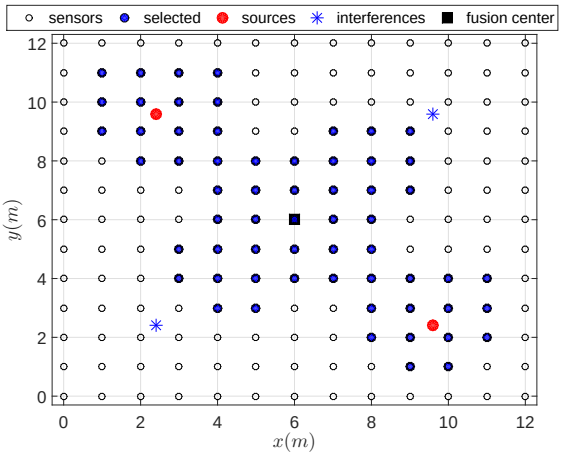


Figure 4.7: A larger-scale WASN, which consists of 169 microphone uniformly placed in a (12×12) m 2D room. The sensors are labelled from bottom to top and from left to right, which is similar to the labeling in Fig. 4.2. The selected microphones are obtained by solving (4.35) for $\alpha = 0.8$.

of the ATF estimation error ζ with fixed source/FC positions and $\alpha = 0.6$. The results are also averaged over 200 trails. Clearly, the performance of the MD-MVDR method is further away from the desired performance. With increasing ζ, the mean values of the average performance difference do not change too much, but the corresponding variances increase gradually. Hence, the proposed method is robust against the ATF estimation errors.

4.5.3. MULTIPLE TARGET SOURCES

In order to further investigate the noise reduction capability of the proposed algorithm for multiple target sources, we consider a larger-scale WASN as Fig. 4.7 shows, which consists of 169 microphones uniformly placed in a 2D room with dimensions (12×12) m. The FC is placed at the center of the room. Two target sources are located at $(2.4, 9.6)$ m and $(9.6, 2.4)$ m, respectively. Two interfering sources are located at $(2.4, 2.4)$ m and $(9.6, 9.6)$ m, respectively. Fig. 4.8 shows the rate distribution, where the proposed method (referred as RD-LCMV) is compared to the model-driven sensor selection method (referred as MD-LCMV in Sec. 4.4.2), which is solved by the bisection algorithm in Sec. 4.4.3. Similar to Fig. 4.3, the sensors that are close to the target sources and FC are allocated with higher rates. The 85th microphone node is allocated with the highest rate, e.g., 16 bits, because it is exactly located at the position of the FC. Also, it is shown that the best microphone subset by MD-LCMV can be determined by finding the optimal threshold for the solution of RD-LCMV (i.e., 3.7812 bits). Furthermore, we plot the sensor selection result that is obtained by solving (4.35) in Fig. 4.7. Comparing the sensors selected by solving (4.35) as shown in Fig. 4.7 to the sensors that are selected by applying the bisection algorithm to the solution of the RD-LCMV algorithm as shown in Fig. 4.8, we see that both sets are completely identical. This also validates the relationship between sensor selection and the rate allocation problem.

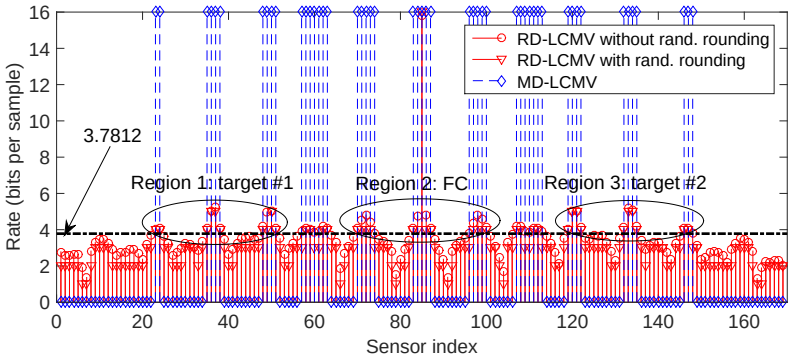


Figure 4.8: Rate distribution for the larger-scale WASN in Fig. 4.7 with $\alpha = 0.8$. The MD-LCMV problem is solved by the bisection algorithm using the results of RD-LCMV. Clearly, the sensors within three regions that are close to the targets and the FC are allocated with higher rates.

To summarize, the rate allocation algorithms (RD-LCMV or RD-MVDR) activate more sensors than the sensor selection algorithms (MD-MVDR or MD-LCMV) in general, but each activated sensor is allocated with a much lower bit-rate. Hence, from the perspective of energy usage for data transmission, the rate allocation algorithms consume less energy.

4.6. CONCLUSION

In this paper, we investigated the rate-distributed spatial filtering based noise reduction problem in energy-aware WASNs. A good strategy for bit-rate allocation can significantly save the energy costs, and meanwhile achieve a prescribed noise reduction performance as compared to a blindly uniform allocation for the best microphone subset obtained by the sensor selection approach. The problem was formulated by minimizing the total transmission costs subject to the constraint on a desired performance. In the context of LCMV beamforming, we formulated the problem as a semi-definite program (i.e., RD-LCMV). Further, we extended the model-driven sensor selection approach in [122] for the LCMV beamforming (i.e., MD-LCMV). It was shown that the rate allocation problem is a generalization of sensor selection, e.g., the best subset of microphones can be chosen by determining the optimal threshold for the rates that are obtained by the RD-LCMV or RD-MVDR algorithm. In WASNs, based on numerical validation, we found that the microphones that are close to the source(s) and the FC are allocated with higher rates, because they are helpful for signal estimation and for reducing energy usage, respectively.

5

DISTRIBUTED RATE-CONSTRAINED LCMV BEAMFORMING

This chapter is based on the article published as "Distributed Rate-Constrained LCMV Beamforming" by J. Zhang, A. I. Koutrouvelis, R. Heusdens, and R. C. Hendriks in *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 675-679, 2019.

5.1. INTRODUCTION

RECENTLY, several beamforming algorithms have been proposed for wireless acoustic sensor networks (WASNs), e.g., [130, 131, 132, 133, 134, 96, 30, 90, 135]. The calculations are done either in a centralized way [130, 131, 132, 133] or in a distributed way [134, 96, 30, 90, 135]. In the centralized case, all the sensor nodes need to transmit their measurements to a fusion center (FC), and the FC performs all computations. There are several limitations on the centralized approach. First, the amount of data that needs to be sent and saved in the FC scales up with the network size. Moreover, with an FC, all operations are performed in a single node, which, in case of disconnection from the network, will cause full collapse of the system. In contrast, the decentralized implementation distributes calculations over the nodes in the WASN, which could overcome the limitations of the centralized approaches.

In WASNs, usually the sensors are battery powered with a limited energy budget. To reduce the energy consumption of beamforming algorithms, one could apply sensor selection [53, 122, 52] or rate allocation [82, 57, 136, 49] to reduce the amount of transmitted information. Rate allocation is more general than sensor selection, as it allows for multiple decisions on the status of sensors. However, sensor selection and rate allocation methods typically work in a centralized fashion, which is, as argued above, undesirable due to scalability and instability issues. In this letter we therefore investigate a decentralized solution for rate-distributed beamforming.

In [85], a distributed linearly constrained minimum variance (LCMV) beamforming method for WASNs was proposed. This method block-diagonalizes the noise/noisy correlation matrix using linear equality constraints, leading to an efficient distributed implementation for the LCMV beamformer. However, this method does not take into account the quantization noise introduced during the communication between the devices. Nor does it take the energy usage due to transmission into account. The rate-distributed LCMV (RD-LCMV) beamformer proposed in [136] is an effective method to reduce the transmission costs over WASNs. It optimally distributes rates to the sensors by minimizing the transmission power under a constraint on the noise reduction performance. However, the RD-LCMV method was derived in a centralized way. This is less efficient with respect to transmission energy if the FC is far away from the WASN.

In this paper our contribution is twofold. First, we solve the rate-allocation problem introduced in [136] for the distributed beamformer proposed in [85]. As the beamformer output highly depends on the quantization noise, we allocate the rates between the devices such that the distributed LCMV beamformer in [85] guarantees a pre-defined performance. Secondly, we propose a distributed solution to the RD-LCMV problem introduced in [136]. Experiments in a simulated WASN validate the proposed decentralized method, i.e., the expected noise reduction performance is achieved with a saving of transmission costs compared to the centralized implementation.

5.2. FUNDAMENTALS

5.2.1. SIGNAL MODEL

We consider a connected WASN consisting of K nodes, where each node $k \in \mathcal{K} = \{1, ..., K\}$, with $\mathcal{K}$ the set of node indices, has $M_k, \forall k$ microphones. In total, we have $M = \sum_{k=1}^{K} M_k$

microphones that acquire the sound field consisting of one target source degraded by acoustic background noise. Let $\mathscr{E}$ denote the set of edges of the network and $\mathscr{N}_k$ the set of neighbouring nodes of node k. If and only if $(i,j) \in \mathscr{E}$, the ith and jth nodes can communicate with each other directly. Let l and ω denote the index of time frame and angular frequency, respectively. In the short-term Fourier transform (STFT) domain, the noisy STFT coefficient at the κth microphone, say $Y_\kappa(\omega,l), \forall \kappa$, is given by

$$Y_\kappa(\omega,l) = X_\kappa(\omega,l) + N_\kappa(\omega,l), \tag{5.1}$$

where $X_\kappa(\omega,l) = a_\kappa(\omega)S(\omega,l)$ with $a_\kappa(\omega)$ the acoustic transfer function (ATF) of the target signal with respect to the κth microphone and $S(\omega,l)$ the STFT coefficient of the target source signal at the source location. In reverberant environments, the ATF consists of early reverberation (typically the first 50 ms) and late reverberation components [137, 84]. Only the early reflections of the target source are beneficial for improving the speech intelligibility. Therefore, in (5.1), the total noise $N_\kappa(\omega,l)$ received by microphone κ is given by

$$N_\kappa(\omega,l) = Z_\kappa(\omega,l) + U_\kappa(\omega,l), \tag{5.2}$$

where $Z_\kappa(\omega,l)$ denotes the correlated noise components including the early reflections of all interfering sources, and $U_k(\omega,l)$ the remaining noise components including the late reverberation from all sources and the sensor noise. For notational brevity, the frequency variable ω and the frame index l will be omitted now onwards. Using vector notation, the M channel signals are stacked in a vector $\mathbf{y} = [Y_1,...,Y_M]^T \in \mathbb{C}^M$. Similarly, we define M-dimensional vectors $\mathbf{x},\mathbf{n},\mathbf{z},\mathbf{u},\mathbf{a}$ for the clean speech component, the total noise, the correlated noise, remaining noise and ATF, respectively, such that the signal model in (5.1) can compactly be written as

$$\mathbf{y} = \mathbf{x}+\mathbf{n} = \mathbf{x}+\mathbf{z}+\mathbf{u}, \tag{5.3}$$

where $\mathbf{x} = \mathbf{a}S$. To focus on the concept of rate-distributed noise reduction, we assume in this work that the ATFs of all sources are known. In a centralized setting, the RTF can be estimated using covariance substraction or covariance whitening method [45]. In the distributed setting this can be estimated using [138, 139, 140, 141]. Further, we assume that all sources are mutually uncorrelated, and the early reflections and late reverberation are also mutually uncorrelated (which is strictly speaking true under the assumption that the STFT coefficients S across time are uncorrelated), such that the second-order statistics (SOS) of the noise components can be written as

$$\mathbf{R}_\mathbf{n} = \mathbb{E}[|\mathbf{n}|^2] = \mathbf{R}_\mathbf{z} + \mathbf{R}_\mathbf{u}, \tag{5.4}$$

where $\mathbb{E}\{\cdot\}$ denotes the statistical expectation operation. Estimation of $\mathbf{R}_\mathbf{n}(l)$ can be done during target-free periods. This is true under the assumption that the DTF coefficients S across time are uncorrelated, because if the late and early reverberations fall in different time frames, then the late reflections in time frame l are uncorrelated with the early reflections in the same frame l.

5.2.2. CENTRALIZED LCMV BEAMFORMING

The LCMV beamformer [14, 20, 142, 129] is widely used in array processing. The filter coefficients are designed to minimize the output noise power subject to a set of linear constraints,

$$\mathbf{w}_{\text{LCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{\mathbf{n}} \mathbf{w}, \quad \text{s.t.} \quad \mathbf{\Lambda}^H \mathbf{w} = \mathbf{f}. \tag{5.5}$$

The closed-form solution to (5.5) is given by [14, 20, 142, 129]

$$\mathbf{w}_{\text{LCMV}} = \mathbf{R}_{\mathbf{n}}^{-1} \mathbf{\Lambda} \left(\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}}^{-1} \mathbf{\Lambda} \right)^{-1} \mathbf{f}. \tag{5.6}$$

Notably, the linear constraints in (5.5) can be used to preserve target sources, eliminate interfering sources [14, 20, 142, 129], or preserve the spatial cues of the sound field [47, 46, 49].

In general, the microphones within a single node are spatially close, while the microphones at different nodes in a WASN are typically more distant. In [85], it was argued that the late reverberation is highly correlated in the first case, while much less correlated in the latter case. Hence, it was suggested that the SOS $\mathbf{R}_{\mathbf{u}}$ can be approximated by a *block-diagonal* matrix where each block corresponds to the SOS of the late reverberation of one node only and the microphone self-noise. By properly using the constraints in the LCMV framework to cancel the early components contained in $\mathbf{z}$ and leveraging the block-diagonal structure of the SOS, the LCMV beamforming problem in (5.5) can be implemented in a distributed fashion. Hence, as in [85], in this work we specify $\mathbf{f} = [1, 0, \cdots, 0]^T \in \mathbb{C}^{r+1}$ (r is the number of interferers), and $\mathbf{\Lambda} = [\mathbf{a}, \mathbf{b}_1, \cdots, \mathbf{b}_r] \in \mathbb{C}^{M \times (r+1)}$ consisting of ATF vectors with $\mathbf{b}_j, \forall\, j$ the ATF of the jth interfering source. Clearly, with such a set of linear constraints $\mathbf{\Lambda}^H \mathbf{w} = \mathbf{f}$ and given enough degrees-of-freedom, the power of the target source is preserved and the power of the correlated sources can entirely be suppressed. As a result, the output noise power after LCMV beamforming can be shown to be given by [129]

$$\mathbb{E}\left[|\mathbf{w}^H \mathbf{n}|^2 \right] = \mathbb{E}\left[|\mathbf{w}^H \mathbf{u}|^2 \right] = \mathbf{w}^H \mathbf{R}_{\mathbf{u}} \mathbf{w}, \tag{5.7}$$

due to the fact that $\mathbf{b}_j^H \mathbf{w} = 0, \forall\, j$. That is, any decrease in the objective function of (5.5) is caused by reducing the uncorrelated noise components. As a result, the matrix $\mathbf{R}_{\mathbf{n}}$ can be replaced by $\mathbf{R}_{\mathbf{u}}$. In the sequel, we will use the block-diagonal approximation of $\mathbf{R}_{\mathbf{u}}$ for the design of algorithms.

5.3. DISTRIBUTED LCMV BEAMFORMING WITH QUANTIZATION NOISE

Given the block-diagonal matrix $\mathbf{R}_{\mathbf{u}}$, by using (5.7) and the constraints to null the early components contained in $\mathbf{z}$, the centralized LCMV beamforming problem in (5.5) can be written in the following node separable form:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{k=1}^{K} \mathbf{w}_k^H \mathbf{R}_{\mathbf{u},k} \mathbf{w}_k, \text{ s.t. } \sum_{k=1}^{K} \mathbf{\Lambda}_k^H \mathbf{w}_k = \mathbf{f}, \tag{5.8}$$

where $\mathbf{w}_k \in \mathbb{C}^{M_k}$, $\mathbf{\Lambda}_k \in \mathbb{C}^{M_k \times (r+1)}$ and $\mathbf{R}_{\mathbf{u},k} = \mathbb{E}[\mathbf{u}_k \mathbf{u}_k^H] \in \mathbb{C}^{M_k \times M_k}$ with $\mathbf{u}_k \in \mathbb{C}^{M_k}$ denote the elements of $\mathbf{w}$, the rows of $\mathbf{\Lambda}$ and the kth block of the matrix $\mathbf{R}_{\mathbf{u}}$, respectively. The

subscript k is used to indicate the components associated with node k. Considering the real-valued Lagrangian function of (5.8), we can obtain the optimal local LCMV filter, given by [85]

$$\mathbf{w}_k^* = \mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{\Lambda}_k\boldsymbol{\mu}^*, \tag{5.9}$$

where $\boldsymbol{\mu}^* \in \mathbb{C}^{r+1}$ is a vector with Lagrangian multipliers. Clearly, the optimal local LCMV filter $\mathbf{w}_k^*$ depends on the global optimal dual variables $\boldsymbol{\mu}^*$. To determine $\boldsymbol{\mu}^*$, one can consider the dual optimization problem of (5.8), given by

$$\boldsymbol{\mu}^* = \arg\max_{\boldsymbol{\mu}} -\sum_{k=1}^{K} \boldsymbol{\mu}^H\mathbf{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{\Lambda}_k\boldsymbol{\mu} + 2\Re\left(\boldsymbol{\mu}^H\mathbf{f}\right), \tag{5.10}$$

where $\Re(\cdot)$ returns the real part. For notational simplicity, we define $\mathbf{G}_k = \mathbf{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{\Lambda}_k, \forall k$. To optimize (5.10) in a distributed fashion, we introduce $\boldsymbol{\mu}_k, \forall k$ to denote the local version of $\boldsymbol{\mu}$ at each node. With this, (5.10) is equivalent to

$$\min_{\boldsymbol{\mu}_k} \sum_{k=1}^{K}\left(\boldsymbol{\mu}_k^H\mathbf{G}_k\boldsymbol{\mu}_k - \frac{2}{K}\Re\left(\boldsymbol{\mu}_k^H\mathbf{f}\right)\right) \text{ s.t. } \boldsymbol{\mu}_k = \boldsymbol{\mu}_m,$$

for all $(k,m) \in \mathcal{E}$. The resulting problem can be solved using randomized gossip [88], ADMM [86] or PDMM [87]. For instance, as shown in [85], the PDMM update procedure for the $(i+1)$th iteration can be summarized as

$$\begin{aligned}\boldsymbol{\mu}_k^{(i+1)} =& \left(\mathbf{G}_k + \rho|\mathcal{N}_k|\mathbf{I}\right)^{-1} \\ &\times\left[\sum_{m\in\mathcal{N}_k}\left(\frac{k-m}{|k-m|}\boldsymbol{\gamma}_{m|k}^{(i)} + \rho\boldsymbol{\mu}_m^{(i)}\right) + \frac{\mathbf{f}}{K}\right],\end{aligned} \tag{5.11a}$$

$$\boldsymbol{\gamma}_{k|m}^{(i+1)} = \boldsymbol{\gamma}_{m|k}^{(i)} - \rho\frac{k-m}{|k-m|}\left(\boldsymbol{\mu}_k^{(i+1)} - \boldsymbol{\mu}_m^{(i)}\right), \tag{5.11b}$$

where $\boldsymbol{\gamma}_{k|m}$ and $\boldsymbol{\gamma}_{m|k}$ are the direct-edge variables computed at nodes k and m, respectively, associated with the edge $(k,m) \in \mathcal{E}$, $\mathbf{I}$ denotes the identity matrix, and ρ is a positive step size. Note that in (5.11), by substituting the update equation for $\boldsymbol{\gamma}_{m|k}^{(i)}$, we can get rid of transmitting the edge variables. As such, updating the edge variables can be performed by broadingcasting $\boldsymbol{\mu}_k^{(i)}$. The iterative procedure can be terminated until $|\boldsymbol{\mu}_k^{(i)} - \boldsymbol{\mu}_m^{(i)}| < \epsilon$ where ϵ is a small positive number.

In [143, 144], the convergence of PDMM was shown in the presence of quantization noise. Due to quantization, the dual variables exchanged among nodes are noisy, i.e., $\hat{\boldsymbol{\mu}}_k^{(i)} = \boldsymbol{\mu}_k^{(i)} + \tilde{\boldsymbol{\mu}}_k^{(i)}$, where $\tilde{\boldsymbol{\mu}}_k^{(i)}$ denotes the quantization noise which is assumed to be zero-mean[1]. Using the above PDMM update equations, the LCMV filter from (5.9) in iteration i is given by

$$\hat{\mathbf{w}}_k^{(i)} = \mathbf{w}_k^{(i)} + \tilde{\mathbf{w}}_k^{(i)} = \mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{\Lambda}_k\left(\boldsymbol{\mu}_k^{(i)} + \tilde{\boldsymbol{\mu}}_k^{(i)}\right), \tag{5.12}$$

where $\tilde{\mathbf{w}}_k^{(i)} = \mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(i)}$ is the error caused by quantization. After the local filters are obtained, calculating the beamformer output reduces to an average consensus problem

[1] This assumption holds when subtractive dithering based uniform quantization is used. The dither signal, which is known at the receiver side, and the quantization noise are i.i.d. processes.

as

$$\min_{X} \sum_{k=1}^{K} \left(X_k - \hat{\mathbf{w}}_k^H \mathbf{y}_k \right)^2 \text{ s.t. } X_k = X_m, \forall (k,m) \in \mathcal{E}. \tag{5.13}$$

The PDMM update equations for (5.13) can be found in [85]. Note that for stationary signals, the update procedure in (5.11) is time-invariant, while (5.13) is always both time and frequency dependent. To reduce the communication costs, we will next derive how to find the optimal quantization rate distribution for iteratively calculating the local filters and beamforming.

5.4. PROPOSED DISTRIBUTED RATE ALLOCATION

In [136], the centralized rate-distributed LCMV (RD-LCMV) beamforming problem was formulated by minimizing the total transmission energy between all sensor nodes and the FC and constraining the resulting output noise power. Let the transmission power from node k to a neighboring node m for a single time-frequency bin be $d_k^2 V_{km}(4^{b_k}-1)$, where $0 \le b_k \le b_0, \forall k$ denotes the integer rate that is used by the node k, and d_k and V_{km} denote the transmission range and the channel noise power spectral density (PSD) between node k and node m, respectively [36, 37, 38]. Assuming that in each iteration we randomly (e.g., at a probability of $\frac{1}{K}$) pick one node of the WASN that broadcasts information to all of its neighboring nodes, such that the expected transmission power per iteration can be given by

$$g(\mathbf{b}) = \frac{1}{K} \sum_{k=1}^{K} d_k^2 V_k \left(4^{b_k} - 1 \right), \tag{5.14}$$

where V_k is the mean value of $V_{km}, m \in \mathcal{N}_k$. Assuming that I iterations are used for calculating the filters through (5.11) and J iterations for beamforming in (5.13), respectively, the original RD-LCMV problem in [136] can be reformulated as

$$\min_{\mathbf{b}} \; g(\mathbf{b}) \text{ s.t. } \sum_{k=1}^{K} \left(\mathbb{E}\left[|\hat{\mathbf{w}}_k^{(I)H} \mathbf{u}_k|^2 \right] + \mathbb{E}\left[\zeta_{X_k}^{(J)} \right] \right) \le \frac{\beta}{\alpha}, \tag{P1}$$

where $\alpha \in (0,1]$ is the parameter to control the expected performance, $\mathbb{E}[\zeta_{X_k}^{(J)}]$ denotes the primal mean-squared error (MSE) caused by quantizing X_k in calculating the beamformer output, i.e., $\zeta_{X_k}^{(J)} = |X_k - Q_{b_k}^{(J)}(X_k)|^2$ with $Q_{b_k}^{(J)}(X_k)$ denoting the quantized X_k using b_k bits. Further, the filter $\hat{\mathbf{w}}_k^{(I)}$ was given in (5.12), and $\beta = \sum_{k=1}^{K} \mathbb{E}[|\mathbf{w}_k^{(I)H} \mathbf{u}_k|^2]$ denotes the minimum output noise power (i.e., without quantization noise). In (P1), the term $\mathbb{E}[|\hat{\mathbf{w}}_k^{(I)H} \mathbf{u}_k|^2]$ denotes the residual acoustic noise and the residual noise of the beamformer due to quantizing $\boldsymbol{\mu}_k$. Note that $\zeta_{X_k}^{(J)}$ depends on the number of iterations and the topology of the network. Since the beamforming is performed iteratively with quantization, the quantization noise $\zeta_{X_k}^{(J)}$ will accumulate at each iteration. However, in [143], it was shown that in case of quantization with sufficiently small fixed cell width (e.g., uniform quantization), the error accumulates but the growth is so slow that it can be considered constant over the iteration range of interest. That is, the primal MSE $\mathbb{E}[\zeta_{X_k}^{(J)}]$ can be approximated by

$$\mathbb{E}[\zeta_{X_k}^{(J)}] \approx C\sigma_k^2, \forall k, \tag{5.15}$$

where σ_k^2 denotes the noise variance depending on the bit rate and the quantization range, and C is a constant which only depends on the topology of the network and is $\mathcal{O}(K)$.

The noise power at node k in (P1) can be calculated by

$$\mathbb{E}\left[|\hat{\mathbf{w}}_k^{(I)H}\mathbf{u}_k|^2\right] \overset{(a)}{=} \mathbb{E}\left[\left(\mathbf{w}_k^{(I)}+\tilde{\mathbf{w}}_k^{(I)}\right)^H \mathbf{u}_k\mathbf{u}_k^H\left(\mathbf{w}_k^{(I)}+\tilde{\mathbf{w}}_k^{(I)}\right)\right]$$
$$\overset{(b)}{=} \mathbb{E}\left[\mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\mathbf{w}_k^{(I)}\right] + 2\mathbb{E}\left[\Re(\mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\tilde{\mathbf{w}}_k^{(I)})\right] + \mathbb{E}\left[\tilde{\mathbf{w}}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\tilde{\mathbf{w}}_k^{(I)}\right],$$

where we note that $\sum_{k=1}^K \mathbb{E}\left[\mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\mathbf{w}_k^{(I)}\right] = \beta$.

Proposition 1. *If the quantization noise $\tilde{\boldsymbol{\mu}}_k^{(I)}$ and the acoustic noise $\mathbf{u}_k$ are independent, we have*

$$\mathbb{E}\left[\Re(\mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\tilde{\mathbf{w}}_k^{(I)})\right] = 0, \quad \mathbb{E}\left[\tilde{\mathbf{w}}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\tilde{\mathbf{w}}_k^{(I)}\right] = \mathrm{Tr}\left(\mathbf{G}_k\mathbf{R}_{\tilde{\boldsymbol{\mu}}_k}\right),$$

where $\mathbf{R}_{\tilde{\boldsymbol{\mu}}_k} = \mathbb{E}\left[\tilde{\boldsymbol{\mu}}_k^{(I)}\tilde{\boldsymbol{\mu}}_k^{(I)H}\right]$ and $\mathrm{Tr}(\cdot)$ returns the trace of a matrix.

Proof. The proof follows from the observation that $\mathbb{E}(AB) = \mathbb{E}(A)\mathbb{E}(B)$ if A and B are independent (and $\mathbb{E}(\mathbf{a}^H\mathbf{b}\mathbf{b}^H\mathbf{a}) = \mathbb{E}(\mathrm{Tr}(\mathbf{b}\mathbf{b}^H\mathbf{a}\mathbf{a}^H)) = \mathrm{Tr}(\mathbb{E}(\mathbf{a}\mathbf{a}^H)\mathbb{E}(\mathbf{b}\mathbf{b}^H))$ if $\mathbf{a}$ and $\mathbf{b}$ are independent vectors). Specifically, let $f(X)$ denote the probability density function of a random variable X. If the quantization noise $\tilde{\boldsymbol{\mu}}_k^{(I)}$ and acoustic noise $\mathbf{u}_k$ are independent, we can see that $f(\tilde{\boldsymbol{\mu}}_k^{(I)}, \mathbf{u}_k) = f(\tilde{\boldsymbol{\mu}}_k^{(I)})f(\mathbf{u}_k)$. Then, the expectations can be calculated as

$$\mathbb{E}\left[\Re\left(\mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\tilde{\mathbf{w}}_k^{(I)}\right)\right] = \Re\left(\mathbb{E}\left[\mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)}\right]\right)$$
$$= \Re\left\{\int_{\mathbf{u}_k}\int_{\tilde{\boldsymbol{\mu}}_k^{(I)}} \mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)} f(\tilde{\boldsymbol{\mu}}_k^{(I)}, \mathbf{u}_k)\, d\mathbf{u}_k\, d\tilde{\boldsymbol{\mu}}_k^{(I)}\right\}$$
$$= \Re\left\{\int_{\mathbf{u}_k} \mathbf{w}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k f(\mathbf{u}_k)\int_{\tilde{\boldsymbol{\mu}}_k^{(I)}} \tilde{\boldsymbol{\mu}}_k^{(I)} f(\tilde{\boldsymbol{\mu}}_k^{(I)})\, d\tilde{\boldsymbol{\mu}}_k^{(I)}\, d\mathbf{u}_k\right\} = 0,$$

since $\mathbb{E}\left[\tilde{\boldsymbol{\mu}}_k^{(I)}\right] = \int_{\tilde{\boldsymbol{\mu}}_k^{(I)}} \tilde{\boldsymbol{\mu}}_k^{(I)} f(\tilde{\boldsymbol{\mu}}_k^{(I)})\, d\tilde{\boldsymbol{\mu}}_k^{(I)} = 0$. In addition, we have

$$\mathbb{E}\left[\tilde{\mathbf{w}}_k^{(I)H}\mathbf{u}_k\mathbf{u}_k^H\tilde{\mathbf{w}}_k^{(I)}\right] = \mathbb{E}\left[\tilde{\boldsymbol{\mu}}_k^{(I)H}\boldsymbol{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{u}_k\mathbf{u}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)}\right]$$
$$= \int_{\mathbf{u}_k}\int_{\tilde{\boldsymbol{\mu}}_k^{(I)}} \tilde{\boldsymbol{\mu}}_k^{(I)H}\boldsymbol{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\mathbf{u}_k\mathbf{u}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)} f(\tilde{\boldsymbol{\mu}}_k^{(I)}, \mathbf{n}_k)\, d\mathbf{u}_k\, d\tilde{\boldsymbol{\mu}}_k^{(I)}$$
$$= \int_{\tilde{\boldsymbol{\mu}}_k^{(I)}} \tilde{\boldsymbol{\mu}}_k^{(I)H}\boldsymbol{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\underbrace{\int_{\mathbf{u}_k}\mathbf{u}_k\mathbf{u}_k^H f(\mathbf{u}_k)\, d\mathbf{u}_k}_{=\mathbb{E}\left[\mathbf{u}_k\mathbf{u}_k^H\right]=\mathbf{R}_{\mathbf{u},k}}\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)} f(\tilde{\boldsymbol{\mu}}_k^{(I)})\, d\tilde{\boldsymbol{\mu}}_k^{(I)}$$
$$= \int_{\tilde{\boldsymbol{\mu}}_k^{(I)}} \tilde{\boldsymbol{\mu}}_k^{(I)H}\boldsymbol{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)} f(\tilde{\boldsymbol{\mu}}_k^{(I)})\, d\tilde{\boldsymbol{\mu}}_k^{(I)}$$
$$= \mathbb{E}\left[\tilde{\boldsymbol{\mu}}_k^{(I)H}\boldsymbol{\Lambda}_k^H\mathbf{R}_{\mathbf{u},k}^{-1}\boldsymbol{\Lambda}_k\tilde{\boldsymbol{\mu}}_k^{(I)}\right] = \mathrm{Tr}\left(\mathbf{G}_k\mathbf{R}_{\tilde{\boldsymbol{\mu}}_k}\right),$$

which completes the proof. □

To this end, we can see that

$$\sum_{k=1}^{K}\mathbb{E}\left[|\hat{\mathbf{w}}_k^{(I)H}\mathbf{u}_k|^2\right]=\beta+\sum_{k=1}^{K}\mathrm{Tr}\left(\mathbf{G}_k\mathbf{R}_{\tilde{\boldsymbol{\mu}}_k}\right). \tag{5.16}$$

Further, we use fixed-rate uniform quantizers for all iterations to quantize the dual variables, such that the SOS of the quantization noise $\tilde{\boldsymbol{\mu}}_k^{(I)}$ can be given by [136, 68, 57]

$$\mathbf{R}_{\tilde{\boldsymbol{\mu}}_k}=\mathbb{E}\left[\tilde{\boldsymbol{\mu}}_k^{(i)}\tilde{\boldsymbol{\mu}}_k^{(i)H}\right]=\frac{1}{12}\times\frac{\mathscr{A}^2}{4^{b_k}}\mathbf{I}_{r+1},\forall i, \tag{5.17}$$

where $\mathscr{A}=\max|\boldsymbol{\mu}^*|$ and is pre-defined. Similarly, we have $\sigma_k^2=\frac{1}{12}\times\frac{\mathscr{B}_k^2}{4^{b_k}}$ with $\mathscr{B}_k$ the expected dynamic range of the optimal beamformer output. Using a variable change $1\leq t_k=4^{b_k}\leq 4^{b_0},\forall k$ and the property in (5.15), (P1) can be simplified as

$$\min_{\mathbf{t}}\ g(\mathbf{b})\ \text{s.t.}\ \sum_{k=1}^{K}\left[\mathrm{Tr}\left(\mathbf{G}_k\right)\mathscr{A}^2+\mathscr{B}_k^2C\right]/t_k\leq\delta, \tag{P2}$$

where $\delta=12\left(\frac{\beta}{\alpha}-\beta\right)$. By solving the KKT condition $\frac{\partial\mathscr{L}(\mathbf{t},\lambda)}{\partial t_k}=0$, the optimal solution to (P2) can be found as

$$t_k^*=\sqrt{\lambda\left(\mathscr{A}^2\mathrm{Tr}(\mathbf{G}_k)+\mathscr{B}_k^2C\right)/d_k^2V_k}, \tag{5.18}$$

which only depends on the Lagrange multiplier λ. To determine λ, one can consider the dual problem of (P2). Substituting (5.18) into (P2), we obtain the dual problem as

$$\min_{\lambda}\ \sum_{k=1}^{K}\left(\frac{\delta}{K}\lambda-2\sqrt{\lambda\left(\mathscr{A}^2\mathrm{Tr}(\mathbf{G}_k)+\mathscr{B}_k^2C\right)d_k^2V_k}+d_k^2V_k\right), \tag{5.19}$$

which is quadratic in $\sqrt{\lambda}$ and the constraint on δ is partitioned into K equal parts. As a result, we can see that the optimal global multiplier is given by

$$\lambda^*=\frac{1}{K^2}\left(\sum_{k=1}^{K}\sqrt{\lambda_k}\right)^2, \tag{5.20}$$

where the local λ_k is defined by

$$\lambda_k=K^2\left(\mathscr{A}^2\mathrm{Tr}(\mathbf{G}_k)+\mathscr{B}_k^2C\right)d_k^2V_k/\delta^2,\forall k. \tag{5.21}$$

Clearly, determining λ^* turns into an averaging problem, since λ_k can be computed separately at each node. Then, we can use PDMM to calculate the average consensus of $\sqrt{\lambda_k}$ that is required by (5.20). This requires a large amount of information exchange. To avoid this, we can consider using the locally optimal λ_k from (5.21) only, instead of the globally optimal λ^*. Substituting (5.21) into (5.18), we obtain the rate distribution as

$$t_k=K\left(\mathscr{A}^2\mathrm{Tr}(\mathbf{G}_k)+\mathscr{B}_k^2C\right)/\delta, \tag{5.22}$$

which reveals that by using local λ_k, the rate can be determined locally without any information exchange and it only depends on the noise power. However, this might affect the global optimality of the rate distribution, which will be studied experimentally. Notably, the final rates should be resolved by $b_k=\log_4 t_k,\forall k$ and randomized rounding as in [136].

5.5. NUMERICAL RESULTS

Fig. 5.1 shows a simulated WASN in a 2D room with dimensions (6×4) m. We consider $K = 21$ nodes and each node has $M_k = 3, \forall k$ microphones. We set $\rho = 0.5$ and $C = 21$. One target source is located at $(2,3)$ m. Five noise sources are randomly placed around the WASN. The duration of all sources is 10 minutes. All sources originate from the TIMIT database [120]. The sensor noise is modeled as white Gaussian noise at an SNR of 50 dB. The sampling frequency is 16 kHz. A square-root-Hann window of 50 ms for framing with 50% overlap is applied to the signals. The ATFs are generated using [121] with reverberation time $T_{60} = 200$ ms. The 21st node is assumed to be the FC for the centralized RD-LCMV method [136], i.e., all other nodes are only connected to this FC. When we cal-

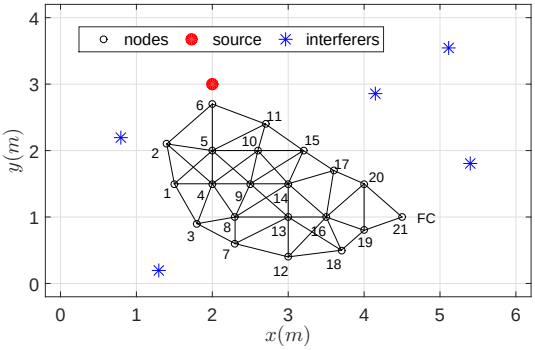


Figure 5.1: Experimental setup, where the last node is assumed to be the FC for the centralized RD-LCMV method [136].

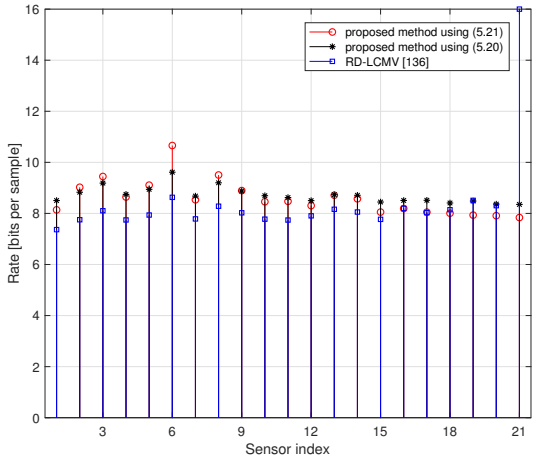


Figure 5.2: Rate distribution for one frequency bin with $\alpha = 0.8$ and $b_0 = 16$ bits per sample.

culate the dual variable $\boldsymbol{\mu}$ using PDMM from (5.11), the warm-start procedure proposed

in [85] is employed to achieve an acceptable precision of PDMM within a finite number of iterations. Fig. 5.2 shows a rate-distribution example of the proposed method and the centralized method [136] for $\alpha = 0.8$. For the proposed method, the nodes that have higher SNR are allocated with higher rate, e.g., node 6. For the centralized method [136], the nodes that are closer to the FC are allocated with higher rate. In addition, we show the output noise power and transmission cost averaged over frequencies in terms of α in Fig. 5.3. The energy of the RD-LCMV method is used for transmitting the raw audio realizations. For the proposed method, if we use the local λ_k in (5.21) to determine the rate distribution, the energy is only used for transmitting the dual variable $\boldsymbol{\mu}$ and calculating the beamformer output; if the rate distribution is computed using (5.18) with the global λ^* from (5.20), some extra energy needs to be spent for calculating λ^*. Clearly, both the centralized method and the proposed decentralized method satisfy the desired noise reduction performance, while the proposed method using (5.21)-(5.22) consumes less energy, since each sensor node only needs to communicate with the neighboring nodes, instead of with the remote FC. This reveals that using the local λ_k is effective for the energy usage versus performance trade-off in spite of scarifying rate optimality. Note that in general a global optimization problem cannot be approached by optimizing local sub-problems separately. We considered optimizing the local problems in this work, as the simulation results show that it gives a better energy usage versus performance trade-off.

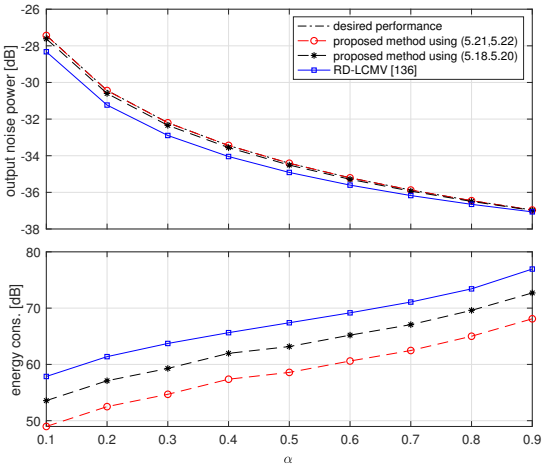


Figure 5.3: Output noise power and transmission energy in terms of α.

5.6. CONCLUSION

In this work, we solved the rate-distributed LCMV beamforming problem in [136] in a fully distributed fashion. The quantization rates were determined locally without any information exchange. Numerical results show the superiority of the proposed method in energy usage. More importantly, the decentralized implementation is more robust against the network variation compared to the centralized method.

6

RATE-DISTRIBUTED BLCMV BEAMFORMING FOR ASSISTIVE HEARING IN WASNS

This chapter is based on the article published as "Rate-Distributed Binaural LCMV Beamforming for Assistive Hearing in Wireless Acoustic Sensor Networks" by J. Zhang, R. Heusdens and R. C. Hendriks in *the 10th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 460–464, Sheffield, UK, 2018.

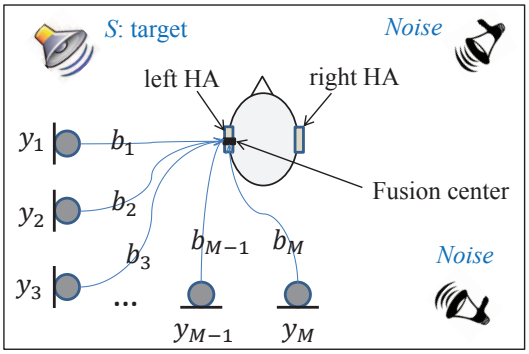


Figure 6.1: A general binaural HA configuration in WASN.

6.1. INTRODUCTION

With the introduction of wireless communication, binaural processing for hearing assistive devices has attracted an increasing interest, e.g., [145, 146, 147]. The traditional hearing-aid (HA) configuration consists of two HAs that are mounted on the two ears, but operate independently. Although this setup can very well suppress noise, it does not take interaural information between the two HAs into account. That is, traditional HAs cannot preserve the spatial cues in the sound field. However, in many scenarios the user needs to be able to identify the direction of the audible sound sources, which can be obtained from the spatial cues (e.g., interaural level/phase difference).

In order to jointly suppress noise and preserve spatial cues, several binaural HA algorithms have been proposed assuming the availability of wireless communication channels, e.g., [148, 47, 149, 141]. In this work, we consider a general framework where the HAs are part of a bigger wireless acoustic sensor network (WASN) with additional assistive wireless microphones, see Fig. 6.1. The microphones can thus be part of the HA itself, or positioned somewhere in the vicinity. The microphone recordings are transmitted via wireless links to a fusion center (FC), which we consider in this work to be one of the HAs, see Fig. 6.1. Subsequently, the FC computes the binaural outputs for both HAs and transmits the output to the contralateral HA. As such, the FC can preserve the interaural information in the binaural outputs. The larger number of microphones in such a setup can potentially lead to both better noise reduction and spatial cue preservation. These advantages of binaural HAs in a WASN setup come with higher battery costs for transmission of data, and, introduction of quantization noise. These facts are typically neglected in most contributions on binaural speech enhancement, with the exception of e.g., [82, 70, 46, 80, 150].

In practice, HAs and assistive microphones in a WASN are battery driven, so that the trade off between the increased performance and energy usage for communication over such WASNs should be taken into account. Typically, the network lifetime needs to be maximized. In order to reduce the energy usage, generally there are two techniques that can be employed: *sensor selection* [53, 122, 52] and *rate allocation* [82, 70, 57, 136]. Sensor selection approaches lead to sparse networks, as only the most informative sensors are involved such that the energy usage in terms of data processing is saved effectively.

Compared to sensor selection, rate allocation approaches can be used to distribute communication rates optimally to save the energy usage in terms of data transmission, since the transmission power between nodes and the FC is directly affected by the rate. The relationship between sensor selection and rate allocation was investigated in [136].

In this work, we apply the rate allocation approach in [136] to a binaural HA setting in a WASN. The problem is formulated by minimizing the total transmission power and constraining the noise reduction performance. The spatial cues are preserved using linear constrains within a binaural linearly constrained minimum variance (BLCMV) beamformer framework. Simulations show that although both the sensor selection and rate allocation approaches satisfy the performance requirement, the proposed rate allocation method is more efficient in energy usage and can preserve more interferers' spatial cues by including more sensors, each at a relatively low rate.

6.2. FUNDAMENTALS

6.2.1. SIGNAL MODEL

In this work, we assume that there are M microphones that are monitoring the sound field, see e.g. Fig. 6.1, where the FC allocates bit rates to each microphone node and computes the binaural output for each HA. In the short-term Fourier transform (STFT) domain, let l denote the frame index and ω the angular frequency bin. The noisy DFT coefficient of the quantized signal which is to be transmitted to the FC is given by

$$\hat{Y}_k(\omega,l) = Y_k(\omega,l) + Q_k(\omega,l), \quad k = 1,2,\cdots,M, \tag{6.1}$$

where $Q_k(\omega,l)$ denotes the quantization noise which is assumed to be uncorrelated with the microphone recording[1] $y_k(\omega,l)$ given by

$$Y_k(\omega,l) = \sum_{i=1}^{\mathscr{I}} \underbrace{a_{ik}(\omega)S_i(\omega,l)}_{X_{ik}(\omega,l)} + \sum_{j=1}^{\mathscr{J}} \underbrace{h_{jk}(\omega)U_j(\omega,l)}_{N_{jk}(\omega,l)} + V_k(\omega,l),$$

where $a_{ik}(\omega)$ denotes the acoustic transfer function (ATF) of the ith target signal with respect to the kth microphone; $S_i(\omega,l)$ and $X_{ik}(\omega,l)$, the ith target source at the source location and at the kth microphone, respectively; $h_{jk}(\omega)$ the ATF from the jth interferer to the kth microphone; $U_j(\omega,l)$ and $N_{ik}(\omega,l)$, the jth interferer at the source location and at the kth microphone, respectively; $V_k(\omega,l)$ the kth microphone self noise. For notational brevity, we will omit the frequency variable ω and the frame index l now onwards. Using vector notation, the M channel signals are stacked in a vector $\hat{\mathbf{y}} = [\hat{Y}_1,...,\hat{Y}_M]^T$. Similarly, we define the vectors $\mathbf{y}$, $\mathbf{x}_i$, $\mathbf{n}_j$, $\mathbf{v}$, $\mathbf{q}$ for the microphone recordings, the ith target component, the jth interfering component, the additive noise and the quantization noise, respectively. Using this notation, (6.1) can be written compactly as

$$\hat{\mathbf{y}} = \sum_{i=1}^{\mathscr{I}} \mathbf{x}_i + \sum_{j=1}^{\mathscr{J}} \mathbf{n}_j + \mathbf{v} + \mathbf{q} = \mathbf{As} + \mathbf{Hu} + \mathbf{v} + \mathbf{q}, \tag{6.2}$$

[1] This assumption holds under high rate communication. At low rates, this can be achieved by subtractive dither [70, 71].

where $\mathbf{x}_i = \mathbf{a}_i s_i \in \mathbb{C}^M$ and $\mathbf{n}_j = \mathbf{h}_j u_j \in \mathbb{C}^M$ with

$$\mathbf{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{iM} \end{bmatrix}, \quad \mathbf{h}_j = \begin{bmatrix} h_{j1} \\ h_{j2} \\ \vdots \\ h_{jM} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_{\mathcal{I}}^T \end{bmatrix}^T, \quad \mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_{\mathcal{I}} \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \vdots \\ \mathbf{h}_{\mathcal{J}}^T \end{bmatrix}^T, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{\mathcal{J}} \end{bmatrix}.$$

We assume in this work that the ATFs of the present sources (i.e., $\mathbf{A}$ and $\mathbf{H}$) are known. In practice, the target ATFs can be estimated using the generalized eigenvalue decomposition of the noise and noisy correlation matrices. The ATFs of the interferers can be replaced by pre-determined ATFs as in [151], at the cost of a small increase of the errors on the spatial cues. Assuming that all sources are mutually uncorrelated, the second-order statistics are then given by

$$\mathbf{R}_{\mathbf{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} = \mathbf{R}_{\mathbf{xx}} + \underbrace{\mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{vv}}}_{\mathbf{R}_{\mathbf{nn}}} \in \mathbb{C}^{M\times M}, \tag{6.3}$$

where $\mathbf{R}_{\mathbf{xx}} = \sum_{i=1}^{\mathcal{I}} \mathbb{E}\{\mathbf{x}_i \mathbf{x}_i^H\}$ and $\mathbf{R}_{\mathbf{uu}} = \sum_{j=1}^{\mathcal{J}} \mathbb{E}\{\mathbf{n}_i \mathbf{n}_i^H\}$. In practice, $\mathbf{R}_{\mathbf{nn}}$ can be estimated using noise-only frames, and $\mathbf{R}_{\mathbf{yy}}$ during the speech+noise frames. The total noise second-order statistics in $\hat{\mathbf{y}}$ is given by $\mathbf{R}_{\mathbf{n+q}} = \mathbf{R}_{\mathbf{nn}} + \mathbf{R}_{\mathbf{qq}}$, under the assumption that the received noise sources and quantization noise are mutually uncorrelated. In case sensors utilize uniform quantizers to quantize their recordings, $\mathbf{R}_{\mathbf{qq}}$ then reads [57, 136, 69]

$$\mathbf{R}_{\mathbf{qq}} = \frac{1}{12}\mathrm{diag}\left(\left[\frac{\mathscr{A}_1^2}{4^{b_1}}, \frac{\mathscr{A}_2^2}{4^{b_2}}, ..., \frac{\mathscr{A}_M^2}{4^{b_M}}\right]\right), \tag{6.4}$$

where $\mathscr{A}_k = \max\{|y_k|\}$ and $b_k, \forall k$ denotes the bit rate used by the kth microphone node. Note that the quantization in the sequel takes place in the STFT domain, e.g., the real and imaginary parts of the complex STFT coefficients are quantized separately.

6.2.2. BLCMV BEAMFORMING WITH BINAURAL CUE PRESERVATION

In [47], a general BLCMV beamforming framework was proposed for joint noise reduction and binaural cue preservation. Mathematically, this problem was formulated as

$$\hat{\mathbf{w}}_{\mathrm{BLCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \tilde{\mathbf{R}}_{\mathbf{n+q}} \mathbf{w}, \quad \text{s.t.} \quad \mathbf{\Lambda}^H \mathbf{w} = \tilde{\mathbf{f}}, \tag{6.5}$$

where

$$\tilde{\mathbf{R}}_{\mathbf{n+q}} = \begin{bmatrix} \mathbf{R}_{\mathbf{n+q}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\mathbf{n+q}} \end{bmatrix} \in \mathbb{C}^{2M\times 2M}, \tag{6.6}$$

$$\begin{aligned} \mathbf{\Lambda} &= \left[\begin{array}{c|c} \mathbf{\Lambda}_1 & \mathbf{\Lambda}_2 \end{array}\right] \in \mathbb{C}^{2M\times(2\mathcal{I}+\mathcal{J})} \\ &= \left[\begin{array}{cc|ccc} \mathbf{A} & \mathbf{0} & \mathbf{h}_1 h_{1R} & \cdots & \mathbf{h}_{\mathcal{J}} h_{\mathcal{J}R} \\ \mathbf{0} & \mathbf{A} & -\mathbf{h}_1 h_{1L} & \cdots & -\mathbf{h}_{\mathcal{J}} h_{\mathcal{J}L} \end{array}\right], \end{aligned} \tag{6.7}$$

$$\tilde{\mathbf{f}} = \left[\begin{array}{c|c} \mathbf{f}_1^H & \mathbf{f}_2^H \end{array}\right]^T \in \mathbb{C}^{2\mathscr{I}+\mathscr{J}}$$
$$= \left[\begin{array}{cccc|cccc} a_{1L}^* & \cdots & a_{\mathscr{I}L}^* & a_{1R}^* & \cdots & a_{\mathscr{I}R}^* & 0 \cdots & 0 \end{array}\right]^T,$$

and the BLCMV beamformer is the concatenation of the LCMV beamformers at the two HAs, i.e., $\mathbf{w}_{\text{BLCMV}} = [\mathbf{w}_L^T \ \mathbf{w}_R^T]^T$. In the BLCMV formulation, L and R are used to indicate the left and right beamformer or reference microphone for the two ears, respectively. Information on the spatial cues is contained in the interaural transfer function (ITF). The ITF of the ith target source with respect to the reference microphones can be defined as $\text{ITF}_{\mathbf{x}_i} = \frac{a_{iL}}{a_{iR}}, \forall i$, and the ITF of interferers can be defined similarly. Accordingly, we can see that the constraint $\mathbf{\Lambda}^H\mathbf{w} = \tilde{\mathbf{f}}$ in (6.5) consists of two components: 1) a constraint on the exact preservation of the $\mathscr{I}$ target sources, i.e., $\mathbf{\Lambda}_1^H\mathbf{w} = \mathbf{f}_1$, for which we know that full preservation requires

$$\text{ITF}_{\mathbf{x}_i}^{\text{in}} = \text{ITF}_{\mathbf{x}_i}^{\text{out}} = \frac{a_{iL}}{a_{iR}}, i = 1, \cdots, \mathscr{I}; \tag{6.8}$$

2) A constraint on the preservation of the $\mathscr{J}$ interferers, i.e., $\mathbf{\Lambda}_2^H\mathbf{w} = \mathbf{f}_2$, for which we know that preserving the spatial cues requires

$$\text{ITF}_{\mathbf{n}_j}^{\text{in}} = \text{ITF}_{\mathbf{n}_j}^{\text{out}} = \frac{h_{jL}}{h_{jR}} = \frac{\mathbf{w}_L^H\mathbf{h}_j}{\mathbf{w}_R^H\mathbf{h}_j}, j = 1, \cdots, \mathscr{J}. \tag{6.9}$$

With the preservation of ITFs in (6.8-6.9), the binaural cues, e.g., interaural level difference (ILD) and interaural phase difference (IPD) are also preserved, because ILD and IPD are derived from ITF as

$$\text{ILD} = |\text{ITF}|^2, \quad \text{IPD} = \angle\text{ITF}. \tag{6.10}$$

Using the method of Lagrange multipliers, the closed-form solution of the above BLCMV problem is given by

$$\hat{\mathbf{w}}_{\text{BLCMV}} = \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1}\mathbf{\Lambda}(\mathbf{\Lambda}^H\tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1}\mathbf{\Lambda})^{-1}\tilde{\mathbf{f}} \in \mathbb{C}^{2M}. \tag{6.11}$$

For more details on BLCMV beamforming with binaural cue preservation, we refer to [47, 46, 80, 152, 153] and references therein.

6.3. RATE-DISTRIBUTED BLCMV BEAMFORMING

6.3.1. GENERAL PROBLEM FORMULATION

Let V_k be the noise power spectral density (PSD) at the kth communication channel and d_k the distance over which transmission takes place. The transmission energy model is then given by [136]

$$g(\mathbf{b}) = \sum_{k=1}^{M} d_k^2 V_k (4^{b_k} - 1), \tag{6.12}$$

where $\mathbf{b} = [b_1, \cdots, b_M]^T$. The above energy model holds under two conditions [36, 37, 38]: 1) in the context of band-limited applications (e.g., audio processing); 2) the microphone recordings are quantized at the channel capacity for reliable transmission. In this work, we intend to minimize $g(\mathbf{b})$ by allocating bit rates to microphone nodes, such that a

prescribed noise reduction performance is obtained. With this, our initial problem can be formulated as:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b}} \ & g(\mathbf{b}) = \sum_{k=1}^{M} d_k^2 V_k (4^{b_k} - 1) \\ \text{s.t.} \ & \mathbf{w}^H \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}} \mathbf{w} \leq \frac{\beta}{\alpha} \\ & \mathbf{\Lambda}^H \mathbf{w} = \tilde{\mathbf{f}}, \quad b_k \in \mathbb{Z}_+, \quad b_k \leq b_0, \forall k, \end{aligned} \tag{P1}$$

where β denotes the minimum output noise power that can be achieved, $\alpha \in (0,1]$ is to control the expected performance, $\mathbb{Z}_+$ denotes the set of non-negative integers, and b_0 the maximum number of bits per sample of each microphone signal. Note that β/α does not depend on the rate allocation strategy or statistics of the sensor network, because β/α is just a number that can be assigned by the users, e.g., 40 dB, to indicate a certain expected performance. By solving (P1), we can determine the optimal rate distribution that each microphone can utilize to quantize its recordings, such that the noise reduction system achieves a desired performance with minimum energy usage. One simple method to solve (P1) is exhaustive search, i.e., evaluating the performance for all $(b_0+1)^M$ choices for the rate distribution, but evidently this is intractable unless b_0 or M is very small. In the next section, we will propose an efficient solver for (P1) in the context of BLCMV beamforming.

6.3.2. SOLVER FOR RATE-DISTRIBUTED BLCMV BEAMFORMING

Substituting the solution of the BLCMV beamformer from (6.11) to the general problem formulation in (P1), we can obtain a simplified optimization problem for rate-distributed BLCMV beamforming as

$$\begin{aligned} \min_{\mathbf{b}} \ & g(\mathbf{b}) = \sum_{k=1}^{M} d_k^2 V_k (4^{b_k} - 1) \\ \text{s.t.} \ & \tilde{\mathbf{f}}^H (\mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda})^{-1} \tilde{\mathbf{f}} \leq \frac{\beta}{\alpha} \\ & b_k \in \mathbb{Z}_+, \quad b_k \leq b_0, \forall k, \end{aligned} \tag{P2}$$

where $\mathbf{b}$ is implicit in the output noise power $\tilde{\mathbf{f}}^H (\mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda})^{-1} \tilde{\mathbf{f}}$, which is non-convex and non-linear in terms of $\mathbf{b}$. In what follows, we will explicitly express $\tilde{\mathbf{f}}^H (\mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda})^{-1} \tilde{\mathbf{f}}$ in terms of $\mathbf{b}$.

First of all, in order to reformulate (P2) as a convex optimization problem, we introduce a symmetric positive semi-definite matrix $\mathbf{Z} \in \mathbb{S}_+^{2\mathcal{J}+\mathcal{J}}$ with $\mathbb{S}_+$ denoting the set of symmetric positive semi-definite matrices, such that the first inequality constraint in (P2) can be recast to the following two new constraints equivalently, i.e.,

$$\mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda} = \mathbf{Z}, \tag{6.13}$$

$$\tilde{\mathbf{f}}^H \mathbf{Z}^{-1} \tilde{\mathbf{f}} \leq \frac{\beta}{\alpha}. \tag{6.14}$$

The inequality (6.14) can be rewritten as a linear matrix inequality (LMI) using the Schur

complement [108, p.650], i.e.,

$$\begin{bmatrix} \mathbf{Z} & \mathbf{f} \\ \mathbf{f}^H & \frac{\beta}{\alpha} \end{bmatrix} \succeq \mathbf{O}_{2\mathscr{I}+\mathscr{J}+1}. \tag{6.15}$$

However, the equality constraint in (6.13) is both non-linear and non-convex in the unknown $\mathbf{b}$. The non-convexity can be tackled by relaxing it to

$$\mathbf{\Lambda}^H \mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1} \mathbf{\Lambda} \succeq \mathbf{Z}, \tag{6.16}$$

since (6.14) and (6.16) are sufficient to obtain the original constraint in (P2). In order to linearize (6.16) in $\mathbf{b}$, we calculate $\tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1}$ as

$$\begin{aligned} \tilde{\mathbf{R}}_{\mathbf{n}+\mathbf{q}}^{-1} &= (\tilde{\mathbf{R}}_{\mathbf{nn}} + \tilde{\mathbf{R}}_{\mathbf{qq}})^{-1} \\ &= \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1} - \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}(\tilde{\mathbf{R}}_{\mathbf{nn}}^{-1} + \tilde{\mathbf{R}}_{\mathbf{qq}}^{-1})^{-1}\tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}, \end{aligned} \tag{6.17}$$

where the second equality is derived from the matrix inversion lemma [107, p.18][2], and

$$\tilde{\mathbf{R}}_{\mathbf{nn}} = \begin{bmatrix} \mathbf{R}_{\mathbf{nn}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\mathbf{nn}} \end{bmatrix}, \quad \tilde{\mathbf{R}}_{\mathbf{qq}} = \begin{bmatrix} \mathbf{R}_{\mathbf{qq}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\mathbf{qq}} \end{bmatrix}. \tag{6.18}$$

Substituting $\mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1}$ from (6.17) into (6.16), we obtain

$$\mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1} \mathbf{\Lambda} - \mathbf{Z} \succeq \mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}(\tilde{\mathbf{R}}_{\mathbf{nn}}^{-1} + \tilde{\mathbf{R}}_{\mathbf{qq}}^{-1})^{-1}\tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{\Lambda}. \tag{6.19}$$

Using the Schur complement, we obtain the following LMI

$$\begin{bmatrix} \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1} + \tilde{\mathbf{R}}_{\mathbf{qq}}^{-1} & \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} \\ \mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1} & \mathbf{\Lambda}^H \tilde{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{\Lambda} - \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{2M+2\mathscr{I}+\mathscr{J}}, \tag{6.20}$$

where $\tilde{\mathbf{R}}_{\mathbf{qq}}^{-1} = \begin{bmatrix} \mathbf{R}_{\mathbf{qq}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\mathbf{qq}}^{-1} \end{bmatrix}$ and $\mathbf{R}_{\mathbf{qq}}^{-1}$ can be calculated from (6.4) directly. For notational convenience, we define a constant vector $\mathbf{e} = [\frac{12}{\mathscr{A}_1^2}, \cdots, \frac{12}{\mathscr{A}_M^2}]$. Further, we introduce a variable change $t_k = 4^{b_k} \in \mathbb{Z}_+, \forall k$, such that $\mathbf{R}_{\mathbf{qq}}^{-1} = \mathrm{diag}(\mathbf{e} \odot \mathbf{t})$ and (6.20) are both linear in $\mathbf{t}$. In order to convexify 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 arrive at

$$\begin{aligned} &\min_{\mathbf{t},\mathbf{Z}} \; g(\mathbf{t}) = \sum_{k=1}^{M} d_k^2 V_k (t_k - 1) \\ &\text{s.t. } (6.15), (6.20), 1 \le t_k \le 4^{b_0}, \forall k, \end{aligned} \tag{6.21}$$

which is a standard semi-definite programming problem [108, p.128] and which can be solved efficiently in polynomial time using interior-point methods or solvers, e.g., CVX [110].

After (6.21) is solved, the allocated bit rates can be resolved by $b_k = \log_4 t_k, \forall k$ which are continuous values. In order to resolve the final integer rates, we apply the randomized rounding technique [53, 122, 136] to the solution of (6.21).

[2] $(\mathbf{A}+\mathbf{CBC}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} + \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{A}^{-1}$

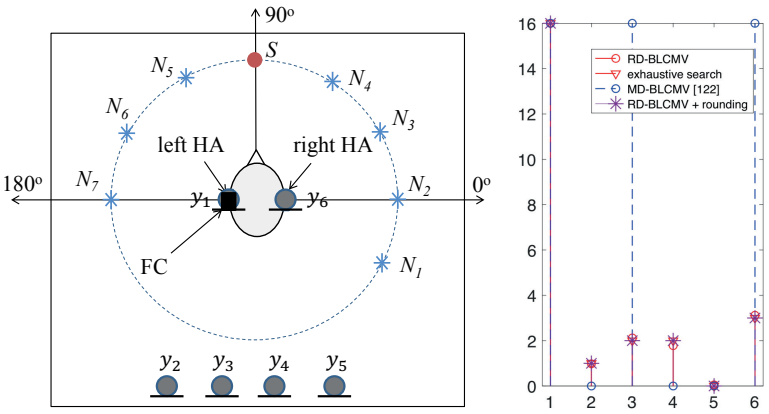


Figure 6.2: Experimental setup and rate distribution obtained by RD-BLCMV, MD-BLCMV and exhaustive search, respectively.

6.4. SIMULATION RESULTS

For the experiments, we place in total $M = 6$ microphones in a 2D room with dimensions (3×4) m, see Fig. 6.2 (left). From these M microphones, one is placed at each ear. These two microphones are taken as the reference microphone for the two HAs. The other four wireless microphones are placed as a (wireless) uniform linear array, whose x-coordinates are given by $\{0.9, 1.3, 1.7, 2.1\}$ m, and the y-coordinate is set to 1 m. The radius of the head of a user who wears the HAs is assumed to be 10 cm, and the FC is positioned at the left HA, i.e., the coordinate of the FC is $(1.4, 2.5)$ m. The ATFs are generated using [121] with reverberation time $T_{60} = 200$ ms. We consider one target source of interest which is located in front of the head. The target source signal is a 10 minute long concatenation of speech signals originating from the TIMIT database [120]. Seven speech shaped Gaussian interfering sources are present, and are placed at $-30°$, $0°$, $30°$, $60°$, $120°$, $150°$ and $180°$, respectively. All the sources are distributed on a circle (radius = 1 m) centered at the head and the elevation is set to be $0°$. All the signals are sampled at 16 kHz. We use a square-root Hann window of 20 ms for framing with 50% overlap. Microphone self noise is modeled at a signal-to-noise ratio (SNR) of 50 dB. The signal-to-(total)interference ratio (SIR) is set to be 0 dB. Furthermore, the PSD of the communication channel noise sources V_k in (6.21) are assumed to be the same for all microphones.

For comparison, we use the model-driven BLCMV (MD-BLCMV) based microphone selection [122] and a solution to (P2) based on exhaustive searching as reference methods. MD-BLCMV is an extension of [122] to our binaural setup. In order to validate the optimality of the proposed method, the exhaustive search is employed to find out the optimal rate distribution. For the maximum rate $b_0 = 16$ bits and six microphones, there are 17^6 combinations for the exhaustive search. Fig. 6.2 (right) shows the rate distribution obtained by (6.21) (i.e., the proposed method referred as RD-BLCMV), MD-BLCMV and exhaustive search, respectively. The performance control parameter α for all methods is set to be 0.8. We observe that the proposed RD-BLCMV method is very close to the optimal solution that is obtained by the exhaustive search, and if we post-process the

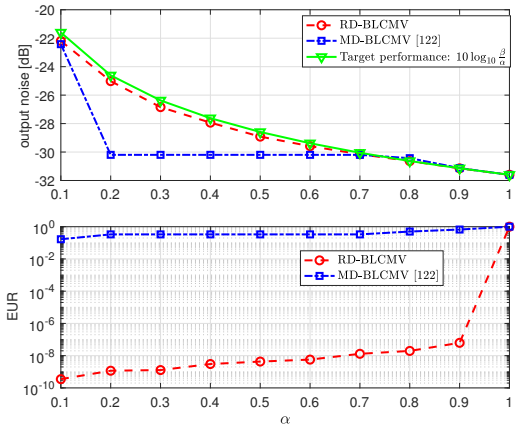


Figure 6.3: Output noise power and energy efficiency in terms of α.

results from RD-BLCMV using randomized rounding, they are completely the same. For RD-BLCMV, five sensors are activated, and the first one is allocated with highest rate (i.e., 16 bits per sample), because it has no communication cost to the FC and has a higher SNR. The rates of the other activated sensors obtained by RD-BLCMV are much smaller than 16 bits, resulting in a saving of transmission costs. The MD-BLCMV method selects only three microphones, but each operates at the maximum rate of 16 bits per sample.

Next, we compare the output noise power and energy usage ratio (EUR) in terms of the performance control parameter α. The EUR is defined as the ratio between the energy used by RD-BLCMV or MD-BLCMV and the maximum transmission energy when all sensors are involved and each quantizes at b_0 bits. By inspection of Fig. 6.3, we see that both RD-BLCMV and MD-BLCMV [122] satisfy the desired amount of noise reduction, but RD-BLCMV is much closer to the target performance $10\log_{10}\frac{\beta}{\alpha}$, particularly when $0.2 \leq \alpha \leq 0.6$. Actually, for these α-values, the two microphones at the ears and the third microphone are chosen for MD-BLCMV, so that the output noise power and energy efficiency of MD-BLCMV remains the same over this α-range. More importantly, RD-BLCMV has much better energy efficiency compared to MD-BLCMV.

Fig. 6.4 shows the total preservation errors of the binaural cues (e.g., ILD and IPD) in terms of the number of activated interferers[3]. The errors ΔILD and ΔIPD are defined as

$$\Delta\text{ILD} = \sum_{j=1}^{\mathscr{J}} \sum_{\omega} \left(\text{ILD}_j(\omega) - \widetilde{\text{ILD}}_j(\omega)\right)^2, \quad \Delta\text{IPD} = \sum_{j=1}^{\mathscr{J}} \sum_{\omega} \left(\text{IPD}_j(\omega) - \widetilde{\text{IPD}}_j(\omega)\right)^2,$$

where $\widetilde{\text{ILD}}_j(\omega)$ or $\widetilde{\text{IPD}}_j(\omega)$ denotes the ILD or IPD of the jth interfering source contained in the beamformer output. The RD-BLCMV method is compared to a BMVDR beam-

[3] In [47], it was shown that the binaural cues of at most $2M-2\mathscr{I}-1$ interferers can be preserved with M microphones using the BLCMV beamformer in (6.5). In our case with $M=6$ microphones, $\mathscr{I}=1$ target source and $\mathscr{J}=7$ interferers, the binaural cues of both the target source and all the interferers can be preserved by BLCMV or RD-BLCMV because $2M-3 > \mathscr{J}$, and the degree of freedom devoted to noise reduction is $2M-\mathscr{J}-2=3$.

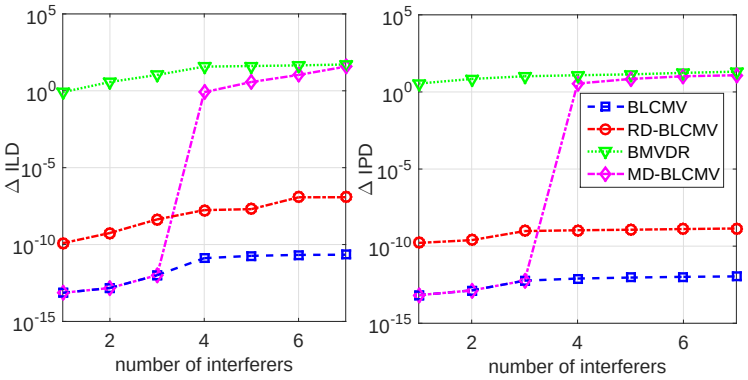


Figure 6.4: Preservation errors versus the number of activated interferers.

former [80], a BLCMV framework [47] and the MD-BLCMV beamformer. The BMVDR method is the worst preserving algorithm, as it does partially consider binaural cue preservation constraints which are associated with the target source. More specifically, for the BMVDR method, the left and right MVDR beamformers can be formulated as

$$\mathbf{w}_L = \frac{\mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1}\mathbf{a}a_L^*}{\mathbf{a}^H\mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1}\mathbf{a}}, \quad \mathbf{w}_R = \frac{\mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1}\mathbf{a}a_R^*}{\mathbf{a}^H\mathbf{R}_{\mathbf{n}+\mathbf{q}}^{-1}\mathbf{a}} \tag{6.22}$$

for one target source that is identified by the ATF $\mathbf{a}$. Clearly, we have $\mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}} = \mathrm{ITF}_{\mathbf{x}}^{\mathrm{out}} = \frac{a_L}{a_R}$ using the BMVDR beamformers. However, $\mathrm{ITF}_{\mathbf{n}_j}^{\mathrm{out}} = \frac{\mathbf{w}_L^H\mathbf{h}_j}{\mathbf{w}_R^H\mathbf{h}_j} = \frac{a_L}{a_R} = \mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}}, \forall j$, which implies that the output binaural cues of the interfering sources collapse to the binaural cues of the target source. Hence, the BMVDR beamformer cannot preserve any binaural cues of interferers. The BLCMV method shows the best performance. However, it does not take the quantization into account and includes all microphones. This means it will be able to keep the spatial cues of all present sources, however, at the high battery cost of full-rate transmission. The MD-BLCMV method uses a hard selection, e.g., if it selects a subset of microphones that is too small, it will not be able to preserve the spatial cues of all sources. The RD-BLCMV approach applies the rate distribution and thus has a soft decision of microphones. In that sense, it usually activates more microphones (at the cost of more quantization noise), but this might lead to more degrees of freedom to preserve more spatial cues, while still satisfying the target noise reduction performance. In addition, all the methods can preserve the spatial cues of the target source because of the constraint $\mathbf{\Lambda}_1^H\mathbf{w} = \mathbf{f}_1$ being taken into account. From Fig. 6.4, we see that with an increasing number of interferers, the errors of RD-BLCMV or BLCMV only slightly increase, but the errors of MD-BLCMV suddenly increase when there are more than 3 interferers. This is because the BLCMV beamformers can preserve the binaural cues of up to $2M - 2\mathscr{J} - 1$ interferers using M microphones [47]. Using hard decisions on microphone selection, the degrees of freedom are much lower than when we use the rate allocation which is a soft decision. Therefore, the RD-BLCMV beamformer allows to use more constraints to preserve interferers than the MD-BLCMV beamformer: 7 versus

3 interferers in Fig. 6.2. Furthermore, similar to the BMVDR, the output binaural cues of the {4,5,6,7}th interferer based on MD-BLCMV will also collapse to those of the target source.

6.5. CONCLUSION

In this work, we studied rate-distributed BLCMV beamforming for wireless binaural hearing aids. The proposed method was formulated by minimizing the energy usage and constraining the noise reduction performance. Under the utilization of a BLCMV beamformer, the problem was solved by semi-definite programming with the capability of joint noise reduction and binaural cue preservation. The proposed method can achieve better energy efficiency by distributing bit rates, and preserve more interferers' spatial cues by activating more sensors as compared to sensor selection approaches.

7

Relative Acoustic Transfer Function Estimation in WASNs

This chapter is based on the article published as "Relative Acoustic Transfer Function Estimation in Wireless Acoustic Sensor Networks" by J. Zhang, R. Heusdens, and R. C. Hendriks in *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1507–1519, 2019.

7.1. INTRODUCTION

ACOUSTIC transfer function (ATF) identification is required by many algorithms in wireless acoustic sensor networks (WASNs), e.g., Wiener filtering [41, 64, 154] or beamforming [129, 28, 95, 132] based noise reduction, or, sound source localization [155]. Often, instead of the ATF, algorithms use the relative acoustic transfer function (RTF) [28], 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.

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 speech+noise 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 [41]. 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) [36, 38]. 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) [39, 40, 41, 42, 43], or by calculating the normalized principal eigenvector [155, 41]. 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 whiten-

ing (CW) [18, 29, 44, 45]. 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 [42, 45]. 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 [42] and [45], 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.

7.1.1. 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 [42, 45]. 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 [136], 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.

7.1.2. OUTLINE AND NOTATION

The paper is structured as follows. Sec. 7.2 presents preliminaries on the signal model and the estimation of sample correlation matrices. In Sec. 7.3, we theoretically analyze the performance of the CS/CW-based RTF estimators. Sec. 7.4 formulates the rate-distributed RTF estimation problem and solves it in the context of the CS and CW methods, respectively. In Sec. 7.5, we show the proposed greedy methods. The proposed methods are validated in Sec. 7.6 via numerical simulations. Finally, Sec. 7.7 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. 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. Tr$(\cdot)$ and 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.

7.2. FUNDAMENTALS

7.2.1. 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 $Y_k(\omega, l), k = 1, \cdots, K$, is

given by

$$Y_k(\omega, l) = X_k(\omega, l) + U_k(\omega, l), \tag{7.1}$$

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 (7.1), 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

$$\hat{Y}_k(\omega, l) = X_k(\omega, l) + U_k(\omega, l) + Q_k(\omega, l). \tag{7.2}$$

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 [49]. A more optimal but complicated rate distribution for quantizing complex Gaussian random variables can be found in [156]. 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 $\hat{\mathbf{y}} = [\hat{Y}_1, ..., \hat{Y}_K]^T \in \mathbb{C}^K$. 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 (7.2) can be rewritten as

$$\hat{\mathbf{y}} = \mathbf{a}S + \mathbf{u} + \mathbf{q}, \tag{7.3}$$

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

$$\mathbf{d} = \mathbf{a}/a_1, \tag{7.4}$$

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

[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 $Y_k(\omega, l)$.

7.2.2. 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 (7.2), the SOS of the noisy microphone signals during speech+noise segments are given by

$$\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \mathbb{E}\{\hat{\mathbf{y}}\hat{\mathbf{y}}^H\} = \mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{qq}}. \tag{7.5}$$

Further, the SOS of the noise are given by

$$\mathbf{R}_{\mathbf{nn}} = \mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{qq}}. \tag{7.6}$$

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

$$\begin{aligned}\mathbf{R}_{\mathbf{xx}} &\triangleq \sigma_S^2 \mathbf{a}\mathbf{a}^H = \sigma_{X_1}^2 \mathbf{d}\mathbf{d}^H \\ &= \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\mathbf{nn}},\end{aligned} \tag{7.7}$$

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 $\operatorname{rank}\left(\mathbf{R}_{\mathbf{xx}}\right) = 1$ 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 (7.5) and (7.6), 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 (7.7).

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

$$\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{L}\sum_{l=1}^{L} \hat{\mathbf{y}}(l)\hat{\mathbf{y}}(l)^H. \tag{7.8}$$

The SOS estimator in (7.8) is unbiased and the corresponding estimation error is denoted by

$$\tilde{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}. \tag{7.9}$$

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

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \frac{1}{|\mathcal{T}|}\sum_{l\in\mathcal{T}} \mathbf{n}(l)\mathbf{n}(l)^H, \tag{7.10}$$

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

[2] This assumption holds under high rate communication. At low rates, this can be achieved by applying subtractive dither [70, 71].

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 this assumption is strictly speaking not necessary and expressions can also be derived taking estimation errors on $\mathbf{R}_{\mathbf{nn}}$ into account. In the derivation of the estimation accuracy under the CW approach it is not trivial to take both estimation errors on $\tilde{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{nn}}$ 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 $\tilde{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is estimated and $\mathbf{R}_{\mathbf{nn}}$ is known. However, in Sec. 7.3, for completeness, we will give the expressions for the CS estimation accuracy when also $\mathbf{R}_{\mathbf{nn}}$ is estimated. With $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{nn}}$ at hand, using (7.7) we can obtain the estimate of $\hat{\mathbf{R}}_{\mathbf{xx}}$ by

$$\hat{\mathbf{R}}_{\mathbf{xx}} \triangleq \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\mathbf{nn}}, \tag{7.11}$$

which can be reformulated as

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \mathbf{R}_{\mathbf{xx}} + \tilde{\mathbf{R}}_{\mathbf{xx}}, \tag{7.12}$$

with $\tilde{\mathbf{R}}_{\mathbf{xx}} = \tilde{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$. Although $\text{rank}(\mathbf{R}_{\mathbf{xx}}) = 1$, in practice we have $\text{rank}(\hat{\mathbf{R}}_{\mathbf{xx}}) > 1$ 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 $\mathbf{R}_{\mathbf{xx}}$, $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ and $\mathbf{R}_{\mathbf{nn}}$, 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 [68, 69]

$$\sigma_{q_k}^2 = \Delta_k^2/12, \forall k, \tag{7.13}$$

where the uniform intervals have width $\Delta_k = \mathscr{A}_k/2^{b_k}$ with $\mathscr{A}/2$ 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

$$\mathbf{R}_{\mathbf{qq}} = \frac{1}{12} \times \text{diag}\left(\left[\frac{\mathscr{A}_1^2}{4^{b_1}}, \frac{\mathscr{A}_2^2}{4^{b_2}}, ..., \frac{\mathscr{A}_K^2}{4^{b_K}} \right] \right). \tag{7.14}$$

7.3. 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 [42] and [45], 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 [42]

$$\epsilon \triangleq \mathbb{E}[||\tilde{\mathbf{d}}||_2^2]/||\mathbf{d}||_2^2. \tag{7.15}$$

7.3.1. 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 [39, 41], i.e.,

$$\hat{\mathbf{d}}_{\mathrm{CS}} \triangleq \frac{\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1}, \tag{7.16}$$

which is based on the rank-1 model for the clean-speech correlation matrix $\mathbf{R}_{\mathbf{xx}}$. The denominator of (7.16) represents the signal power at the reference microphone, i.e.,

$$\hat{\sigma}^2_{X_1} \triangleq \mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1. \tag{7.17}$$

In order to analyze the CS-based RTF estimator, we write the RTF estimate from (7.16) as

$$\hat{\mathbf{d}}_{\mathrm{CS}} = \mathbf{d} + \tilde{\mathbf{d}}_{\mathrm{CS}}. \tag{7.18}$$

In [45], it was shown that the estimation error term $\tilde{\mathbf{d}}_{\mathrm{CS}}$ is given by

$$\tilde{\mathbf{d}}_{\mathrm{CS}} = \frac{1}{|a_1|^2\sigma_S^2}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\tilde{\mathbf{R}}_{\mathbf{xx}}\mathbf{e}_1. \tag{7.19}$$

Assuming the estimation error $\tilde{\mathbf{R}}$ of the covariance matrix $\mathbf{R}$ of a Gaussian random variable when estimated as in (7.8) obeys a complex Wishart distribution [157], it can be shown (see [45]) that given the noise SOS $\mathbf{R}_{\mathbf{nn}}$, the RTF estimation error ϵ_{CS} of the CS-based method from (7.15) is given by [42, 45]

$$\epsilon_{\mathrm{CS}} = \frac{1+\frac{1}{\eta}}{L||\mathbf{d}||_2^2\sigma^2_{X_1}} \cdot \mathrm{Tr}\left(\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H\right), \tag{7.20}$$

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

$$\eta \triangleq \frac{\sigma^2_{X_1}}{\mathbf{e}_1^T\mathbf{R}_{\mathbf{nn}}\mathbf{e}_1} = \frac{\mathbf{e}_1^T\mathbf{R}_{\mathbf{xx}}\mathbf{e}_1}{\mathbf{e}_1^T\mathbf{R}_{\mathbf{nn}}\mathbf{e}_1}. \tag{7.21}$$

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

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

such that the final CS error model can be formulated as

$$\epsilon_{\mathrm{CS}} = \frac{1+\frac{1}{\eta}}{L||\mathbf{d}||_2^2\sigma^2_{X_1}} \cdot \mathrm{Tr}\left(\mathbf{G}\right). \tag{7.22}$$

Note that (7.22) differs from the one in [42] by the facts that 1) quantization noise is taken into account 2) similar as in [45] we assume $\mathbf{R}_{\mathbf{nn}}$ to be known (estimated based on larger data records), resulting in the term $\frac{1}{\eta}$ in (7.22).

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 $\mathbf{R_{\hat{y}\hat{y}}}$, we obtain

$$\epsilon_{\text{CS}} = \frac{\frac{1}{L} + \frac{1}{\eta}\left(\frac{1}{L} + \frac{1}{T}\right)}{||\mathbf{d}||_2^2 \sigma_{X_1}^2} \cdot \text{Tr}\left(\mathbf{G}\right). \tag{7.23}$$

If $L = T$, (7.23) will be identical to the error model derived in [42].

7.3.2. 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

$$\hat{\mathbf{d}}_{\text{CW}} = \frac{\mathbf{R}_{\mathbf{nn}}^{H/2}\hat{\boldsymbol{\psi}}}{\mathbf{e}_1^T \mathbf{R}_{\mathbf{nn}}^{H/2}\hat{\boldsymbol{\psi}}}, \tag{7.24}$$

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 [45], it was shown that the error vector of the CW method can be approximated by

$$\tilde{\mathbf{d}}_{\text{CW}} = \frac{\theta}{a_1}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}^{H/2}\tilde{\boldsymbol{\psi}}, \tag{7.25}$$

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 [158]

$$\Theta_{\boldsymbol{\psi}} = \frac{\lambda_1}{L(\lambda_1 - 1)^2}\left(\mathbf{I} - \boldsymbol{\psi}\boldsymbol{\psi}^H\right), \tag{7.26}$$

where $\lambda_1 = \mathbf{a}^H \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{a}\sigma_S^2 + 1$ 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 $\tilde{\mathbf{d}}_{\text{CW}}$ can be formulated as

$$\begin{aligned}\Theta &\overset{(a)}{=} \frac{|\theta|^2}{|a_1|^2}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}^{\frac{H}{2}}\Theta_{\boldsymbol{\psi}}\mathbf{R}_{\mathbf{nn}}^{\frac{1}{2}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H \\ &\overset{(b)}{=} \frac{1 + \frac{1}{\sigma_{X_1}^2 \mathbf{d}^H \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}}}{L\sigma_{X_1}^2}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{nn}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H,\end{aligned} \tag{7.27}$$

where (a) is obtained by substitution of (7.25) 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

$$\epsilon_{\text{CW}} = \frac{\text{Tr}(\Theta)}{||\mathbf{d}||_2^2} = \frac{1 + \frac{1}{\sigma_{X_1}^2 \mathbf{d}^H \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}}}{L||\mathbf{d}||_2^2 \sigma_{X_1}^2} \cdot \text{Tr}\left(\mathbf{G}\right). \tag{7.28}$$

Note that in fact the term $\sigma_{X_1}^2 \mathbf{d}^H \mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{d}$ is the output SNR of an MVDR beamformer [122, 142, 133, 129].

Remark 6. *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 (7.22) or (7.28) differs from the one derived in [42, 45] by the fact that the quantization noise is now also taken into account. Comparing (7.28) to (7.22), the only difference lies in the SNR term. Since after the use of an MVDR beamformer, the SNR can be improved, i.e.,* $\eta \leq \sigma_{X_1}^2 \mathbf{d}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{d}$*, we can conclude that the CW-based RTF estimator always achieves a higher accuracy than the CS method.*

7.4. 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.

7.4.1. 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, $2 \leq r \leq 6$ [36, 127]. 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

$$\mathrm{SNR}_k = D_k^{-2} E_k / V_k, \forall k, \tag{7.29}$$

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 [128] for Gaussian channels is then given by

$$b_k = \frac{1}{2} \log_2 \left(1 + \mathrm{SNR}_k\right), \forall k, \tag{7.30}$$

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 (7.29) and the capacity (7.30), we can formulate the transmitted energy as [36, 37, 38, 136, 49]

$$E_k = D_k^2 V_k (4^{b_k} - 1), \forall k. \tag{7.31}$$

Notice that the above energy model holds under two conditions [36, 38]: 1) band-limited input signals, and 2) the microphone recordings are quantized at the channel capacity.

7.4.2. 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:

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

where $\epsilon_{\text{CS/CW}}$ indicates the use of either ϵ_{CS} or ϵ_{CW} from (7.22) and (7.28), 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 $\epsilon_{\text{CS/CW}} \leq \frac{\beta}{\alpha}$ 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.

7.4.3. MODEL-DRIVEN RATE-DISTRIBUTED CS (MDRD-CS)

For the first constraint $\epsilon_{\text{CS}} \leq \frac{\beta}{\alpha}$ in (P1), using the expression ϵ_{CS} from (7.22), we can rewrite it as

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

or rearranged as

$$\text{Tr}\left(\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T \right) \mathbf{R}_{\mathbf{qq}} \left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T \right)^H \right) \leq \frac{\beta}{\alpha c_1} - c_2, \tag{7.32}$$

where the constants c_1 and c_2 are given by

$$c_1 = \frac{1 + \frac{1}{\eta}}{L ||\mathbf{a}||_2^2 \sigma_S^2} = \frac{1 + \frac{1}{\eta}}{L ||\mathbf{d}||_2^2 \sigma_{X_1}^2}, \tag{7.33}$$

$$c_2 = \text{Tr}\left(\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T \right) \mathbf{R}_{\mathbf{uu}} \left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T \right)^H \right). \tag{7.34}$$

Clearly, (7.32) is non-convex and non-linear in terms of the bit rates $b_k, \forall k$. For linearization, we equivalently rewrite (7.32) 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.,

$$\mathrm{Tr}(\mathbf{Z}) \leq \frac{\beta}{\alpha c_1} - c_2, \tag{7.35}$$

$$\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{qq}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H = \mathbf{Z}. \tag{7.36}$$

Now, (7.35) is linear in the new variable $\mathbf{Z}$, however, (7.36) is still non-convex in b_k. To convexify (7.36), we can relax it to

$$\mathbf{Z} \succeq \left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{qq}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H, \tag{7.37}$$

since (7.37) and (7.35) are sufficient to obtain the original constraint in (7.32). By inspection, (7.37) can be written as a linear matrix inequality (LMI) using the Schur complement [108, p.650], i.e.,

$$\begin{bmatrix} \mathbf{R}_{\mathbf{qq}}^{-1} & \mathbf{I} - \mathbf{d}\mathbf{e}_1^T \\ \left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H & \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{2K}, \tag{7.38}$$

where $\mathbf{R}_{\mathbf{qq}}^{-1}$ can be computed from (7.14) as

$$\mathbf{R}_{\mathbf{qq}}^{-1} = 12 \times \mathrm{diag}\left(\left[\frac{4^{b_1}}{\mathscr{A}_1^2}, \frac{4^{b_2}}{\mathscr{A}_2^2}, ..., \frac{4^{b_K}}{\mathscr{A}_K^2}\right]\right). \tag{7.39}$$

Note that (7.38) is not an LMI in the unknown parameters $\mathbf{b}$, but in $4^{b_k}, \forall k$. Finally, we define a constant vector $\mathbf{f} = [\frac{12}{\mathscr{A}_1^2}, \cdots, \frac{12}{\mathscr{A}_K^2}]^T$ and introduce a variable change $t_k = 4^{b_k} \in \mathbb{Z}_+, \forall k$, such that $\mathbf{R}_{\mathbf{qq}}^{-1} = \mathrm{diag}(\mathbf{f} \odot \mathbf{t})$ and (7.38) 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 [108, p.128] as

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

7.4.4. MODEL-DRIVEN RATE-DISTRIBUTED CW (MDRD-CW)

Applying the expression from (7.28) 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

$$\mathrm{Tr}\left(\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)\mathbf{R}_{\mathbf{qq}}\left(\mathbf{I} - \mathbf{d}\mathbf{e}_1^T\right)^H\right) \leq \frac{\beta}{\alpha c_1'} - c_2, \tag{7.40}$$

where c_1' is defined by

$$c_1' = \frac{1 + \frac{1}{\sigma_{X_1}^2 \mathbf{d}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{d}}}{L||\mathbf{d}||_2^2 \sigma_{X_1}^2}, \tag{7.41}$$

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

$$\begin{aligned} \mathbf{R}_{\mathbf{nn}}^{-1} &\overset{(a)}{=} \left(\mathbf{R}_{\mathbf{uu}} + \mathbf{R}_{\mathbf{qq}}\right)^{-1} \\ &\overset{(b)}{=} \mathbf{R}_{\mathbf{uu}}^{-1} - \mathbf{R}_{\mathbf{uu}}^{-1}\left(\mathbf{R}_{\mathbf{uu}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1}\right)^{-1} \mathbf{R}_{\mathbf{uu}}^{-1}, \end{aligned} \tag{7.42}$$

where (b) is derived from the matrix inversion lemma [107, p.18][3]. Similar to Sec. 7.4.3, by introducing a matrix $\mathbf{Z} \in \mathbb{S}_{++}^K$, (7.40) can equivalently be rewritten into two new constraints, e.g., (7.35) and (7.36), and the latter one can be relaxed to the LMI in (7.38).

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 (7.41) as

$$c_1' \geq \frac{1 + \frac{1}{\sigma_{X_1}^2 \mathbf{d}^H \mathbf{R}_{\mathbf{nn}}^{-1} \mathbf{d}}}{L||\mathbf{d}||_2^2 \sigma_{X_1}^2}. \tag{7.43}$$

With the substitution of the expression for $\mathbf{R}_{\mathbf{nn}}^{-1}$ from (7.42) into (7.43), we obtain

$$\delta \geq \mathbf{d}^H \mathbf{R}_{\mathbf{uu}}^{-1}\left(\mathbf{R}_{\mathbf{uu}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1}\right)^{-1} \mathbf{R}_{\mathbf{uu}}^{-1} \mathbf{d}, \tag{7.44}$$

where δ is given by

$$\delta = \mathbf{d}^H \mathbf{R}_{\mathbf{uu}}^{-1} \mathbf{d} - \frac{1/\sigma_{X_1}^2}{c_1' L||\mathbf{d}||_2^2 \sigma_{X_1}^2 - 1}. \tag{7.45}$$

Using the Schur complement, (7.44) can be reformulated as the following LMI:

$$\begin{bmatrix} \mathbf{R}_{\mathbf{uu}}^{-1} + \mathbf{R}_{\mathbf{qq}}^{-1} & \mathbf{R}_{\mathbf{uu}}^{-1} \mathbf{d} \\ \mathbf{d}^H \mathbf{R}_{\mathbf{uu}}^{-1} & \delta \end{bmatrix} \succeq \mathbf{O}_{K+1}. \tag{7.46}$$

Note that (7.45) is non-convex in c_1', which can be relaxed to

$$\delta \leq \mathbf{d}^H \mathbf{R}_{\mathbf{uu}}^{-1} \mathbf{d} - \frac{1/\sigma_{X_1}^2}{c_1' L||\mathbf{d}||_2^2 \sigma_{X_1}^2 - 1}, \tag{7.47}$$

since (7.47) and (7.44) are sufficient conditions for obtaining (7.40). As a consequence,

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

the MDRD-CW problem can also be formulated as an SDP problem:

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

Remark 7. *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 [110] or SeDuMi [111]. 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 [136, 53], the complexity of which is linear in* K*.*

7.5. 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 $\mathbf{R}_{\mathbf{uu}}$. Although we can estimate $\sigma_{X_1}^2$, SNR and $\mathbf{R}_{\mathbf{uu}}$ 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],

[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 Sec. 7.2.2 in order to make the analysis on the CS and CW methods consistent.

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 (7.14), (7.8) and (7.11), respectively. Also, we can compute the signal power $\hat{\sigma}^2_{X_1}$ and the SNR at the reference microphone $\hat{\eta}$ using (7.17) and (7.21), 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 (7.16). Using this information, we can update the constants c_1 and c_2 as

$$\hat{c}_1 = \frac{1+\frac{1}{\hat{\eta}}}{l\|\hat{\mathbf{d}}\|_2^2\hat{\sigma}^2_{X_1}}, \tag{7.48}$$

$$\hat{c}_2 = \mathrm{Tr}\Big((\mathbf{I}-\hat{\mathbf{d}}\mathbf{e}_1^T)(\mathbf{R}_{\mathbf{nn}}-\hat{\mathbf{R}}_{\mathbf{qq}})(\mathbf{I}-\hat{\mathbf{d}}\mathbf{e}_1^T)^H\Big), \tag{7.49}$$

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 $\hat{\mathbf{R}}_{\mathbf{uu}} = \mathbf{R}_{\mathbf{nn}} - \hat{\mathbf{R}}_{\mathbf{qq}}$. Based on these, we can update the rate distribution by solving (P2), i.e.,

$$\begin{aligned} \min_{\mathbf{t},\mathbf{Z}} \quad & \sum_{k=1}^{K} D_k^2 V_k(t_k-1) \\ \text{s.t.} \quad & \mathrm{Tr}(\mathbf{Z}) \le \frac{\beta}{\alpha\hat{c}_1} - \hat{c}_2, \\ & \begin{bmatrix} \mathrm{diag}(\mathbf{f}\odot\mathbf{t}) & \mathbf{I}-\hat{\mathbf{d}}\mathbf{e}_1^T \\ \left(\mathbf{I}-\hat{\mathbf{d}}\mathbf{e}_1^T\right)^H & \mathbf{Z} \end{bmatrix} \succeq \mathbf{O}_{2K}, \\ & 1 \le t_k \le 4^{b_{\max}}, \ \ \forall k. \end{aligned} \tag{7.50}$$

Note that (7.50) 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 (7.50) 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 3[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 $(\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}, \mathbf{R}_{\mathbf{nn}})$ corresponding to the maximum eigenvalue. Note that when the number of frames $l \ll L$, it is possible that (7.50) is infeasible due to insufficient segments for estimating the SOS. To circumvent the infeasibility, we can relax β in (7.50) using

$$\hat{\beta} = L\beta/l, \tag{7.51}$$

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

[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 $\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{P}\sum_{i=l-P}^{l}\hat{\mathbf{y}}_i\hat{\mathbf{y}}_i^H$, where P denotes the number of frames from the past that we want to include. If the sources are completely stationary, then $P = l-1$.

Algorithm 3: 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;
$\hat{\mathbf{R}}_{\mathbf{qq}} = \frac{1}{12} \times \text{diag}([\frac{\mathscr{A}_1^2}{4^{b_1}}, \frac{\mathscr{A}_2^2}{4^{b_2}}, ..., \frac{\mathscr{A}_M^2}{4^{b_M}}])$;
$\hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} = \frac{1}{l}\sum_{\iota=1}^{l} \hat{\mathbf{y}}_\iota \hat{\mathbf{y}}_\iota^H$;
$\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{R}}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R_{uu}} - \hat{\mathbf{R}}_{\mathbf{qq}}$;
$\hat{\sigma}^2_{X_1} = |a_1|^2 \hat{\sigma}^2_S = \mathbf{e}_1^T \hat{\mathbf{R}}_{\mathbf{xx}} \mathbf{e}_1$;
$\hat{\eta} = \frac{\hat{\sigma}^2_{X_1}}{\mathbf{e}_1^T(\hat{\mathbf{R}}_{\mathbf{qq}} + \mathbf{R_{uu}})\mathbf{e}_1}$;
Case 1: DDRD-CS
$\hat{\mathbf{d}}_{\text{CS}} = \hat{\sigma}^{-2}_{X_1} \hat{\mathbf{R}}_{\mathbf{xx}} \mathbf{e}_1$;
$\hat{c}_1 = \frac{1+\frac{1}{\hat{\eta}}}{l\|\hat{\mathbf{d}}\|_2^2 \hat{\sigma}^2_{X_1}}$;
$\hat{c}_2 = \text{Tr}((\mathbf{I} - \hat{\mathbf{d}}_{\text{CS}}\mathbf{e}_1^T)\mathbf{R_{uu}}(\mathbf{I} - \hat{\mathbf{d}}_{\text{CS}}\mathbf{e}_1^T)^H)$;
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 $\mathbf{b}_{\text{CS}}, \mathbf{b}_{\text{CW}}, \hat{\mathbf{d}}_{\text{CS}}, \hat{\mathbf{d}}_{\text{CW}}$

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., $\mathscr{O}(K^3)$, and the complexity for all the frames is of the order of $\mathscr{O}(LK^3)$.

7.6. 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.

7.6.1. SIMULATIONS ON SYNTHETIC DATA

Fig. 7.1 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 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

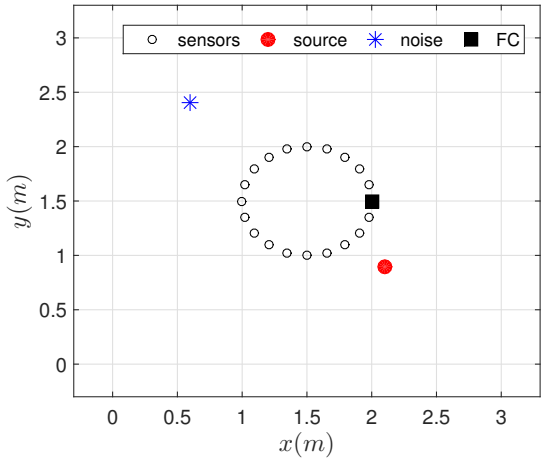


Figure 7.1: An illustration of experimental setting with 20 microphones. The FC and the first microphone are placed at the same position.

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.

[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.

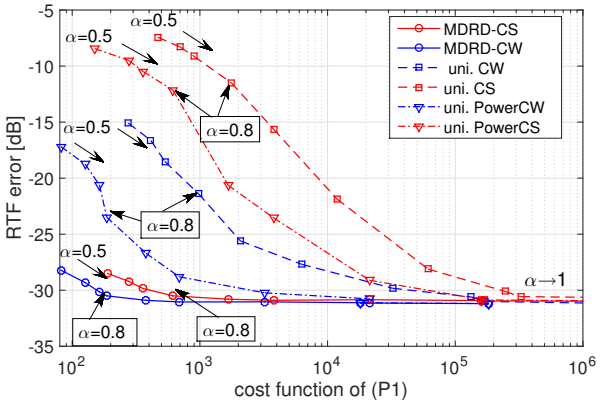


Figure 7.2: RTF error and transmission cost of the model-based methods in terms of α. The cost function in x-axis means the total transmission power per frame. The "total" refers to the summation of transmission costs over microphones and "per frame" indicates the average over L frames.

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 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. 7.2 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 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. 7.3(a) shows the rate distributions obtained by the proposed MDRD-CS/CW from Fig. 7.2 at $\alpha = 0.8$. Clearly, to fulfil a desired RTF estimation performance $\epsilon_{\text{CS/CW}} \leq \frac{\beta}{\alpha}$, 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,

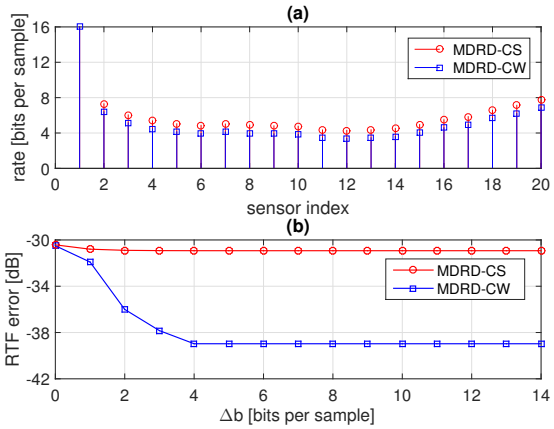


Figure 7.3: (a) An example for rate distribution when $\alpha = 0.8$ and (b) RTF accuracy in terms of rate increment.

the sensors that are closer to the FC are allocated with a higher rate. In Fig. 7.3(b), we show an example on how the RTF accuracy changes by further increasing the rate, starting from the optimal distributions given in Fig. 7.3(a). The resulting RTF accuracy is plotted as a function of the rate increment Δb. For $\Delta b = 0$, we use the optimal rate distribution given in Fig. 7.3(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.

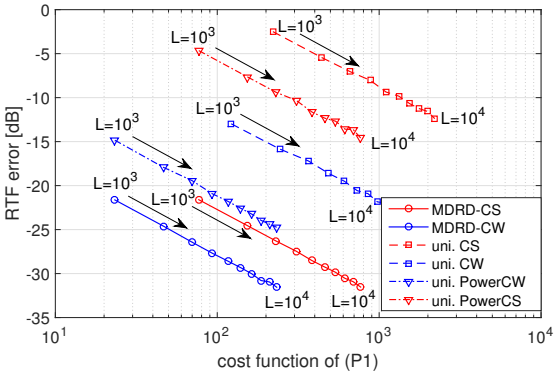


Figure 7.4: RTF error and transmission cost of model-driven methods in terms of the number of available segments for $\alpha = 0.8$. The cost function in x-axes means the total transmission power per frame.

Fig. 7.4 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.

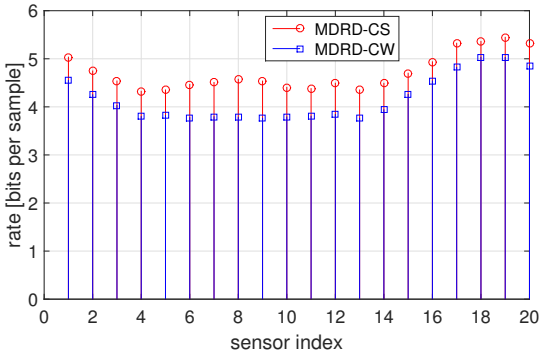


Figure 7.5: Rate distributions of the proposed model-driven methods for the scenario where the FC is located at the center of the room and $\alpha = 0.8$.

For further studying other influence factors on the proposed model-driven rate distribution approaches, we place the FC in Fig. 7.1 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. 7.5 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 higher SNR (e.g., sensor 18 which is closest to the target source) is allocated with a higher rate. This reveals that the cleaner the microphone measurements are, the more bits are required for quantization. Comparing the ranges of the distributed rates between Fig. 7.5 and Fig. 7.3(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.

7

EVALUATION OF DDRD-CS/CW METHODS

Fig. 7.6 compares the proposed DDRD-CS/CW methods to the model-driven versions, uni.CS/CW and uni.PowerCS/CW. For each segment, the uni.CS/CW methods use uniform rate allocation, and 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. 7.7 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 (7.50) 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.

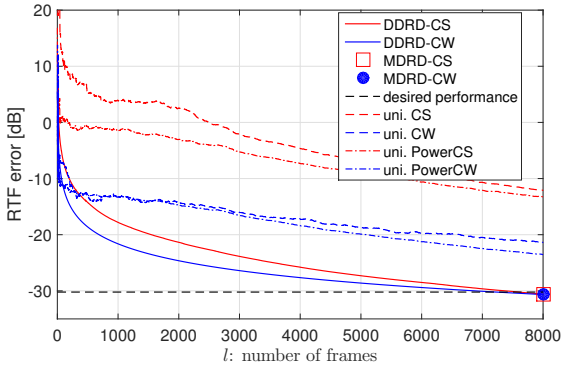


Figure 7.6: RTF accuracy of the data-driven methods for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 8000$.

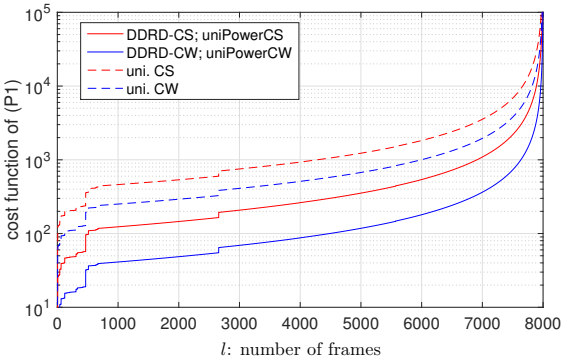


Figure 7.7: Transmission cost of the data-driven methods per frame for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 8000$. The y-axes means the total transmission power per frame.

7.6.2. 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. 7.1. The single target source is a speech signal originating from the TIMIT database [120]. 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 [121] with reverberation time $T_{60} = 200$ ms.

At first, we show the RTF estimation performance of the proposed methods in Fig. 7.8 for $\alpha = 0.8$. This is a similar comparison as in Fig. 7.6, but now using real speech signals. The total number of segments is $L = 500$. We can see that similar to the synthetic

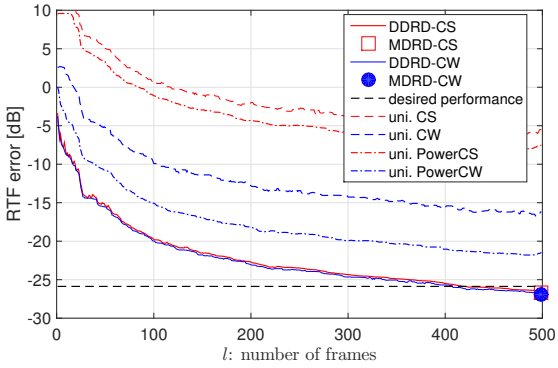


Figure 7.8: RTF estimation performance of the proposed methods using the real speech recordings for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 500$.

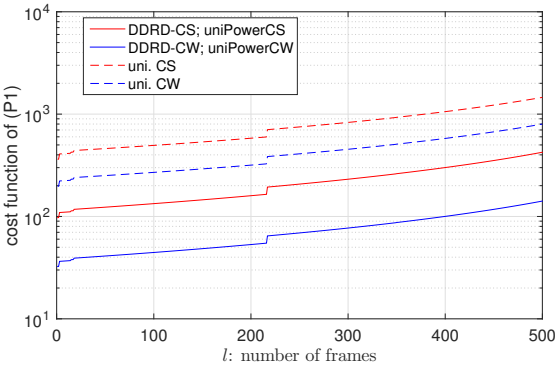


Figure 7.9: Transmission cost per frame of the proposed methods using the real speech recordings for $\alpha = 0.8$. The total number of received frames (i.e., x-axis) increases from 1 to $L = 500$. The cost function in y-axes means the total transmission power per frame.

data case in Fig. 7.6, 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. 7.9.

Secondly, we validate the application of the proposed methods in multiple reverberation conditions. The performance is 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. 7.10 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. 7.5 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 transmis-

sion 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. 7.10.

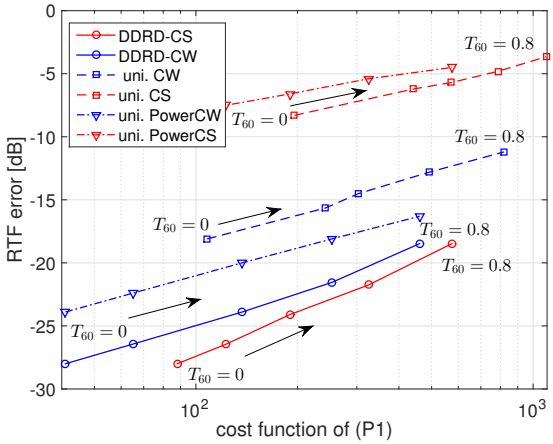


Figure 7.10: RTF estimation accuracy and transmission cost of the proposed methods for multiple reverberation conditions with $\alpha = 0.8$. The cost function in x-axes means the average transmission power per frame.

Finally, since the RTF performance is also affected by the source location (e.g., see Eqs. (7.22, 7.28)), 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. 7.11. 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.

7.7. CONCLUSION

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

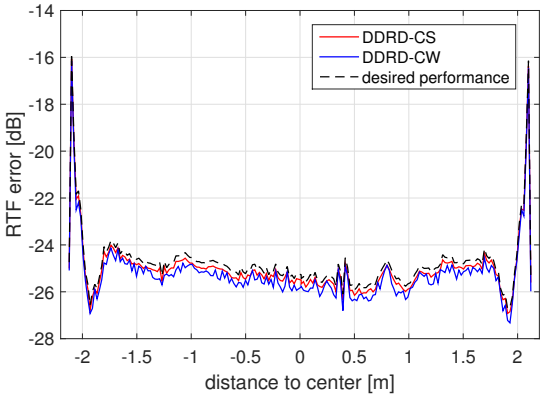


Figure 7.11: RTF error of the proposed methods in terms of the distance from the target source to the center of the room, i.e., (1.5, 1.5) m, for $\alpha = 0.8$.

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 higher 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.

8

CONCLUSION AND FUTURE RESEARCH

Throughout this thesis, we have presented several energy-efficient multi-microphone noise reduction algorithms for wireless acoustic sensor networks (WASNs) to minimize the energy consumption while achieving a prescribed noise reduction performance. In this chapter, we will conclude this dissertation by discussing the considered research questions that were posed in Chapter 1. Also, we will discuss some assumptions and restrictions that were made throughout the thesis, which need to be taken into account in future work.

8.1. CONCLUSIONS AND DISCUSSIONS

In this section, the research questions that were proposed in Sec. 1.4 will be discussed.

8.1.1. MICROPHONE SUBSET SELECTION

As discussed in Chapter 1, the devices that form the WASNs are equipped with a limited battery resource. It is therefore essential to save the power consumption in order to prolong the network lifetime. We therefore proposed the following research question.

Q1: Given a prescribed performance, can we design an effective strategy for saving the power consumption over WASNs?

In order to answer this question, we considered two strategies: sensor selection and rate distribution, since the total power consumption in terms of data transmission is directly affected by 1) the number of sensors that are involved in multi-microphone noise reduction and 2) the transmission rate that is used for quantizing the microphone measurements.

At first, we investigated in Chapter 3 microphone subset selection. Typically, some sensors are closer to the target source(s) than others. This yields differences in the local

signal-to-noise ratios (SNRs) as observed by each device. At the same time, some devices are closer to the fusion center (FC) than others, which affects the required transmission power. It was therefore proposed in Chapter 1 to investigate the following sub-research question of **Q1**:

Q1.1: Given a certain expected performance, can we choose a subset of microphone nodes that minimizes the power consumption for beamforming?

To answer this research question, we proposed in Chapter 3 a microphone subset selection strategy for minimum variance distortionless response (MVDR) beamforming based noise reduction in WASNs. The considered sensor selection problem was formulated by minimizing the total transmission cost between all the sensors and the FC under a constraint on the expected output noise power. This problem formulation differs from the classic sensor selection problem, like [52, 53], which constrains the cardinality of the subset of the selected sensors, in the sense that in large-scale WASNs the cardinality of the selected subset is not of interest, but the desired noise reduction performance. The considered sensor selection problem was first solved by using convex optimization techniques. This method was called model-driven sensor selection, since it leverages the noise second-order statistics (SOS) of the complete sensor network. However, in practice the SOS of the whole network is not given beforehand, which needs to be estimated, making the model-based method impractical. To overcome this drawback, we further proposed a greedy method, which is an online data-driven method.

Simulation results showed that the proposed methods can choose a better subset of microphones in the sense that the total transmission cost is minimized and the expected noise reduction performance is guaranteed compared to other reference approaches. For the proposed methods, the sensors that are close to the target source, those close to the FC and some close to the coherent interfering sources are more likely to be selected, as they have a higher SNR for enhancing the target signal, are cheaper in transmission and are informative for suppressing the noise sources, respectively. In addition, the practical greedy approach converges to the model-based method in terms of noise reduction performance, while the final selected sets of both methods might be different.

Although the proposed sensor selection problem was solved under the assumption that only a single target source is present, the derived methods can also be applied to the scenario with multiple target sources, e.g., see Chapter 4 and 6. Note that for the multiple source case, a linearly constrained minimum variance (LCMV) beamformer has to be used rather than an MVDR beamformer [136, 49, 159]. Also, it is worth noting that although the proposed model-based method is impractical, it can still suggest how to place the sensors optimally when the SOS of the noise field are given, which is known as *sensor placement* [160].

8.1.2. RATE DISTRIBUTION

In Chapter 3, we answered the question how to optimize the transmission cost over a WASN with a satisfactory noise reduction performance from the perspective of sensor selection. As a more practical case, in a centralized WASN the devices need to quantize their raw data measurements and then send them to the FC via a wireless link. For quantization, a certain bit-rate budget is used. The sensor selection problem proposed

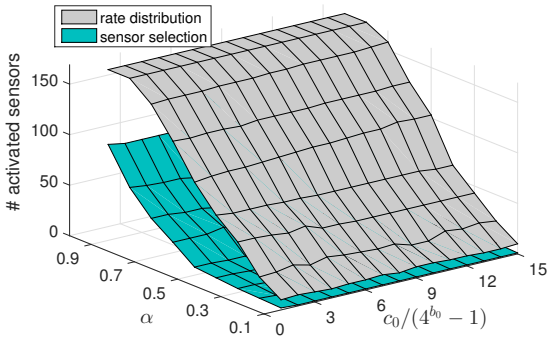


Figure 8.1: The cardinality of the activated sets of rate distribution and sensor selection in terms of the performance controller α and normalized microphone self-cost $c_0/(4^{b_0}-1)$.

in Chapter 3 did not take quantization into account, that is, each considered sensor is only allowed to operate at full rate (when selected) or zero rate (when not selected) implicitly. Since the transmission energy from a sensor to the FC is an exponential function in terms of bit-rate, it is possible that if we take quantization into account and optimize the transmission cost in terms of bit-rates, the total energy consumption could further be reduced compared to that of the sensor selection method. That led in Chapter 1 to the following sub-question of research question **Q1**, that is,

Q1.2: Given a certain expected performance, how to efficiently distribute the bits for signal quantization in order to reduce power consumption?

This question was answered in Chapter 4. To formulate this problem, we first wrote the transmission energy from a sensor to the FC as a function of the bit-rate using the channel coding theory. Similar to the sensor selection problem, we then minimized the total transmission cost under a constraint on an expected level of the output noise power. The considered rate distribution problem was solved by using convex optimization techniques, which resulted in a rate-distributed LCMV (RD-LCMV) beamformer. Furthermore, the relationship between sensor selection and rate distribution was derived in a theoretical fashion, as rate distribution is a generalization of sensor selection. Specifically, rate distribution allows sensors to have multiple (i.e., *soft*) decisions, while sensor selection makes a binary (i.e., *hard*) decision in terms of bit-rates. The optimal microphone subset of the sensor selection method can be found by thresholding the rate distribution of the proposed rate allocation method, and this threshold can be determined by using, e.g., the bisection method.

Simulation results showed that the proposed RD-LCMV method distributes higher bit-rates to the more informative sensors, e.g., those close to the target sources, those close to the FC and some close to the coherent interfering sources. Also, the relationship between these two energy-efficient approaches can clearly be observed from the rate distributions. It was shown in Chapter 4 that rate distribution is always more efficient in energy usage than the sensor selection method by ignoring the microphone self-cost, e.g., the cost for having a sensor activated. However, the comparison of energy efficiency

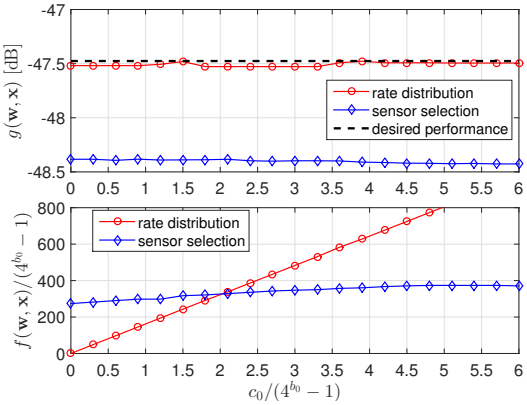


Figure 8.2: Noise reduction performance $g(\mathbf{w},\mathbf{x})$ and normalized power consumption $f(\mathbf{w},\mathbf{x})/(4^{b_0}-1)$ in terms of the normalized microphone self-cost $c_0/(4^{b_0}-1)$ with $\alpha = 0.6$ for sensor selection and rate distribution. Here, we use $\mathbf{w}$ to indicate the beamformer, and $\mathbf{x}$ the sensor selection or rate distribution variable.

will not always be the case if the microphone self-cost, say c_0, is taken into account. For instance, considering the experimental setup in Fig. 4.7, we show the cardinality of the activated sets of the two approaches in Fig. 8.1, where the microphone self-cost is normalized by $4^{b_0}-1$ with b_0 denoting the maximum bit rate. It can be seen that the rate distribution method always has more activated sensors, i.e., the sensors whose allocated rate is non-zero, then depending on the microphone self-cost, the rate distribution method might consume more transmission cost. To see this, we show a more fair comparison between the two methods in Fig. 8.2, where both the total power consumption and the microphone self-cost are normalized by $4^{b_0}-1$. It is clear that in case the cost for have a sensor activated is smaller than a certain threshold, rate distribution is more energy efficient; otherwise sensor selection consumes less power.

All the aforementioned sensor selection and rate distribution approaches are derived in a centralized fashion. That is, an FC is required to collect data measurements and conduct all the computations, and all other devices are only allowed to communicate with the FC. The disadvantage of such a centralized configuration is obvious. If the FC is disconnected from the WASN, the whole network collapses. If the FC is far away from a node in large-scale WASN, the long communication distance will lead to a large power consumption. Therefore, it is worth rethinking the proposed energy-efficient algorithms in a distributed way. This led to the next research question posed in Chapter 1, that is,

Q2: Given a prescribed noise reduction performance, how to design an efficient data transmission strategy between nodes to reduce the power consumption for distributed beamforming?

In order to answer this question, we extended the proposed rate distribution method in a decentralized WASN in Chapter 5. In the decentralized setup, the nodes are connected with other close-by nodes and the calculations as well as transmissions are distributed over nodes. For the distributed implementation, we first reformulated the orig-

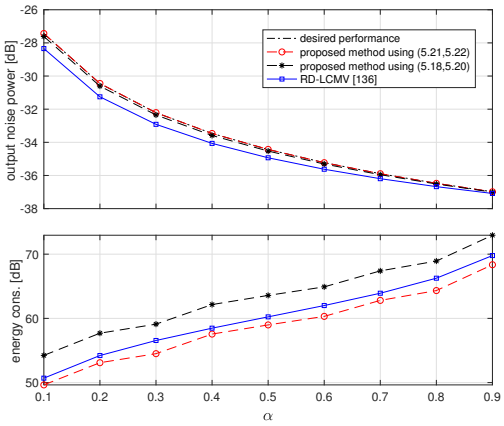


Figure 8.3: Output noise power $g(\mathbf{w},\mathbf{x})$ and power consumption $f(\mathbf{w},\mathbf{x})$ in terms of α for the centralized and decentralized rate-constrained methods.

inal rate optimization problem in a node-separable form and found a rate distribution using local calculations, which yet satisfies the performance requirement. The resulting rate distribution was then used for quantizing the relevant parameters within an optimal LCMV beamforming scheme. Simulation results showed that both the centralized RD-LCMV method and the distributed approach can satisfy the requirement on the noise reduction performance, while the distributed method consumes less transmission energy than the centralized method if the FC is further away from the WASN.

Notably, it is possible that the centralized method is more energy efficient if the FC is close to the WASN. To see this, we further consider the scenario in Fig. 5.1, where the FC is assumed to be located at the 9th node (i.e., approximately at the center of the network). The corresponding energy usage versus noise reduction performance tradeoff is shown in Fig. 8.3. In this case, it is obvious that now the centralized method consumes less energy compared to Fig. 5.3 where the FC is located at the last node, and the superiority of the distributed method in energy consumption becomes smaller. However, the decentralized implementation is more robust against the network variation compared to the centralized counterpart, since the FC is not always available.

Furthermore, we applied the proposed energy-efficient algorithms to a more general WASN, where wireless hearing-aid (HA) devices are a part of the larger WASN. For the HA users, not only noise reduction for improving the speech intelligibility is required, but also preserving the spatial cues of the interfering sources. These spatial cues, including interaural phase difference (IPD) and interaural level difference (ILD) are useful for localizing the corresponding sources. This is related to the third research question posed in Chapter 1, that is,

Q3: For the hearing-aid devices, how to efficiently make use of the measurements from external devices to jointly achieve noise reduction and binaural cue preservation?

This question was answered in Chapter 6, where a rate-distributed binaural LCMV

(RD-BLCMV) beamformer was proposed. The problem formulation is identical to the original rate distribution that was considered in Chapter 4, i.e., minimizing the total transmission cost and constraining the output noise power. The main difference lies in the fact that in this binaural context, we used a binaural LCMV (BLCMV) beamformer [47] for jointly suppressing the existing noise sources and preserving their spatial cues, while the classic LCMV beamformer was used for only enhancing the target sources in the general rate distribution problem in Chapter 4. Following the derivations of the RD-LCMV problem, the proposed RD-BLCMV beamforming problem was also solved using convex optimization techniques. Theoretically, the more sensors that are involved in the design of the BLCMV beamformer, the more degrees of freedom (DOF) that are available for preserving the spatial cues, i.e., more spatial cues can be preserved. On the other hand, if more DOF are spent on preserving the spatial cues, less DOF are left for noise reduction, leading to a trade-off between the two goals [47]. To clearly see this trade-off, we compared the proposed RD-BLCMV method in the binaural context to the sensor selection method that was proposed in Chapter 3. Simulation results showed that given the same expected noise reduction performance, the RD-BLCMV method can preserve more spatial cues of the interfering sources compared to the sensor selection method, since usually the rate distribution method activates more sensors, resulting in a higher amount of DOF.

8.1.3. LOW-RATE RELATIVE TRANSFER FUNCTION ESTIMATION

To perform the beamforming based noise reduction algorithms, e.g., MVDR and LCMV, the acoustic transfer function (ATF) or relative (acoustic) transfer function (RTF) is required for the design of beamformers. The sensor selection and rate distribution algorithms presented in Chapters 3-6 are based on the assumption that the ATF or RTF of the target source(s) with respect to the acoustic devices is known. Therefore, in order to apply the proposed sensor selection or rate distribution method, it is necessary to estimate the RTF information beforehand. Also, the RTF estimation accuracy will certainly affect the performance of other subsequent algorithms, e.g., sensor selection, rate distribution and beamforming. Noting that throughout this dissertation, our focus is on saving the total transmission cost over the WASN, it is thus interesting to investigate the following research question in the context of estimating the RTF.

Q4: Given a prescribed RTF estimation accuracy, can we design an effective data transmission strategy for saving the power consumption over WASNs?

This question was answered in Chapter 7. Assuming that a single target speech source is present, the cross power spectral density matrix of the speech component, say $\mathbf{R}_{\mathbf{xx}}$, is a rank-1 matrix, which in practice can be estimated by subtracting the noise correlation matrix, say $\mathbf{R}_{\mathbf{nn}}$ from the noisy correlation matrix, say $\mathbf{R}_{\mathbf{yy}}$. Given a perfect voice activity detector (VAD), the microphone signals can be classified into speech-plus-noise segments and noise-only segments, and during these two periods the noisy and noise correlation matrices can be estimated, respectively. Taking the quantization noise into account and assuming that the quantization noise and the microphone measurements are mutually uncorrelated, both $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$ at the FC end will include the quantization noise statistics. Due to the subtraction operation, the statistics of the quantization noise

will be eliminated in estimating $\mathbf{R}_{\mathbf{xx}}$ if the quantization noise affects $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$ in the same manner. This means that independent of the quantization rate that is used for transmission, $\mathbf{R}_{\mathbf{xx}}$ can be estimated perfectly given sufficiently long segments. However, due to the errors in estimating $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{R}_{\mathbf{yy}}$, the quantization rate will affect the estimation of $\mathbf{R}_{\mathbf{xx}}$. This is the motivation for the proposed RTF estimation approaches under a low communication rate.

In order to find the optimal rate distribution for estimating the RTF, we first analyzed the performance of two often-used RTF estimation approaches, i.e., covariance subtraction (CS) and covariance whitening (CW) [45] in terms of bit-rate. The CS method takes the normalized first column vector of the estimate of the matrix $\mathbf{R}_{\mathbf{xx}}$ as the estimated RTF. The CW method extracts the normalized principal eigenvector of the correlation matrix of the whitened microphone signals by the noise correlation matrix as the estimated RTF. Based on the performance analysis, it is shown that the performance of both methods is influenced by the SNR, as the higher the SNR is, the more accurately the RTF is estimated. Specifically, the CS method is related to the *prior* SNR of the microphone signals, while the CW method is affected by the *posterior* SNR, which is the output SNR of an MVDR beamformer and always higher than the prior SNR. This reveals that the CW method can always achieve a higher RTF estimation accuracy.

Followed by the problem formulation of the rate distribution approach, we then optimized the rate distribution by minimizing the total transmission cost and constraining the RTF estimation error. The considered rate distribution problem can be solved by applying convex optimization techniques under the utilization of either the CS method or the CW method, which we referred to as model-driven approaches. However, the model-based methods rely on the noise statistics and the true RTF, which need to be estimated in practice. This makes the model-based methods impractical. To overcome this drawback, we further proposed a greedy (data-driven) strategy, which is a practical method. Due to the fact that the sensors send data to the FC frame-by-frame, we can estimate the required parameters by the model-based methods using previously received frames and then solve the model-driven rate optimization problem to find the optimal rate distribution that will be used for quantizing the next frame.

The proposed approaches were validated using both synthetic data and natural speech signals. Simulation results showed that given the same expected RTF accuracy, the rate-distributed CW method always consumes less transmission cost compared to the rate-distributed CS method. To achieve the same RTF accuracy, the proposed methods consume less transmission cost compared to the uniform rate distribution or uniform power distribution methods. In fact, many bits in the sensor measurements are redundant, as increasing the rate distribution does not lead to a distinct decrease in the RTF error. It was also shown that the performance of the proposed greedy approach converges to that of the model-based method with increasing number of received frames. In general, the sensors that are closer to the FC and the sensors that have a higher SNR should be allocated with a higher rate, and the distance between the sensors and the FC is more relevant to the considered rate optimization problem than the SNR.

Altogether, the contributions of this thesis can be combined to form a complete energy-aware noise reduction system for WASNs. This system is depicted in Fig. 8.4, which is built in the short-time Fourier transfer (STFT) domain. Note that from the per-

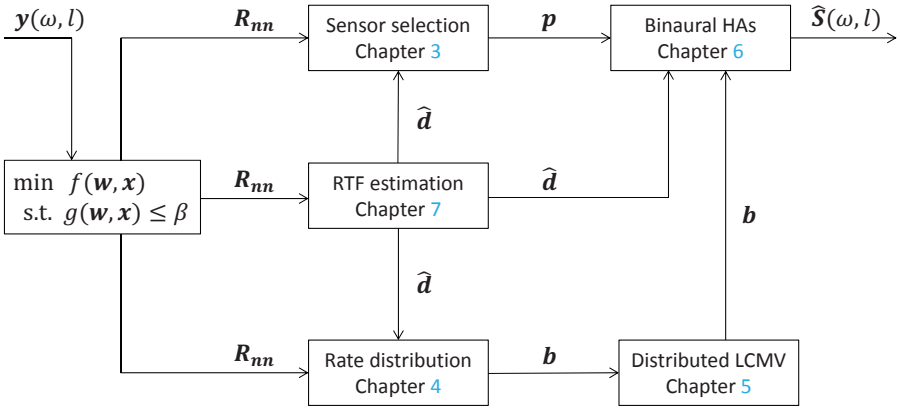


Figure 8.4: Energy-aware noise reduction scheme for WASNs based on contributions of this thesis.

spective of implementation, prior to applying the sensor selection or rate distribution based energy-aware noise reduction algorithms, the proposed energy-efficient RTF estimation approaches should be employed.

8.2. FUTURE RESEARCH

Based on the conclusions that we have drawn and the assumptions that we made in this thesis, we now give some suggestions that are worth studying for future research.

SENSOR SCHEDULING

The microphone subset selection method proposed in Chapter 3 is based on the assumption that the target source(s) is static, although the scenario with a moving FC was considered. This is not always the case in a realistic environment. It is more natural that the target, i.e., a speaker, is moving slowly in the environment, e.g., a speaker in a lecture room. In this case, the target source signal is not time invariant any more and the ATF of the source with respect to the WASN is changing over time. For one observation time slot, the best microphone subset that is obtained by the proposed sensor selection method might not be optimal for the next time slot. As a result, the microphone subset selection problem in this time-varying scenario should be treated as a continuous selection scheme or sensor scheduling problem, as Fig. 8.5 shows. For such a sensor scheduling problem, if the target is moving slowly, in each time slot we can use the classic CS or CW method to estimate the RTF information, which will then be applied to the proposed sensor selection method to find the best microphone subset for the current time slot. It means that the proposed sensor selection method will be repeated for every time slot. If the target moves faster, some lower-complexity algorithms should be considered, e.g., the sensor scheduling problem is cast as a partially observable Markov decision process as in [161]. In general, these algorithms follow the procedure of *prediction-correction*, i.e., using the previous measurements together with the previous solution to predict the current solution and then using the current measurement to correct the current solution

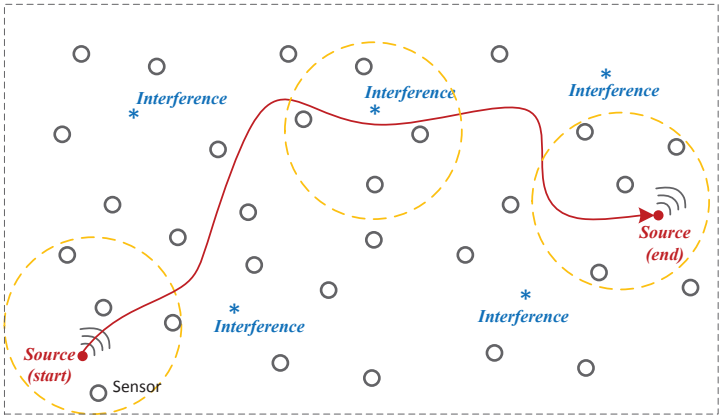


Figure 8.5: An illustration of sensor scheduling in large-scale WASNs for a moving target.

at each time step.

OPTIMAL RATE DISTRIBUTION

Strictly speaking, the rate distribution method proposed in Chapter 4 is not optimal, because it only takes the noise statistics into account. In general, the measurements across sensors are highly correlated. If we take the correlation between sensor measurements into account, the rate distribution of the proposed method can further be optimized, leading to more savings of the energy consumption for the WASNs. In [82], Roy and Vetterli proposed an optimal rate-constrained collaborative noise reduction algorithm for wireless binaural HAs. Specifically, they took the signals at one HA as the side information and encoded the signals at the other HA based on the side information, such that the minimum quantization rate can be found. This is the optimal rate allocation solution for two devices, however, it is not straightforward to extend the algorithm to a more general WASN consisting of a large number of sensors. Therefore, optimal rate distribution in WASNs is still an open research question. Nevertheless, some near-optimal solutions have been proposed recently by using sequential coding, e.g., [162], or asymmetric coding, e.g., [163].

JOINT SENSOR SELECTION AND RATE DISTRIBUTION

From a more consolidated point of view, neither the proposed microphone subset selection nor rate distribution is the optimal strategy for minimizing the power consumption over WASNs, because they are independently considered. One way to further reduce the power consumption of the proposed sensor selection or rate distribution method is by cascading the two strategies, e.g., the bit-rates of the selected sensors of the sensor selection method can be optimized using the proposed rate distribution method, or in the other way around. However, this is still sub-optimal in terms of power consumption, as the sensor selection variables and the bit-rates are not independent. Ignoring the channel noise power spectral density (PSD) in the power model that was used throughout this dissertation, the minimum power consumption should be jointly optimized over

the sensor selection variables and the bit-rates. In [164], a joint sensor selection and power allocation strategy for energy harvesting wireless sensor networks was proposed, which could be borrowed to the considered WASN scenario.

SAMPLING RATE OFFSET ESTIMATION

In WASNs the sensor nodes usually capture acoustic sources asynchronously, since each device has an independent oscillator for timing, resulting in a sampling rate offset (SRO) inevitably compared to the global sampling rate. The SRO can be viewed as a combination of clock offset and clock skew. Using the unsynchronized measurements directly will degrade the performance of audio processing algorithms [165]. It is thus required to take the estimation of the SRO (or resynchronization) into account before designing the noise reduction filters. The methods of synchronization in the context of WASNs can be classified into two categories: correlation maximization (CM) [166, 97] and least-squares coherence drift (LCD) [167]. The CM method treats the SRO as a time scaling in the (wideband) correlation function and the SRO is determined by globally searching the peak of the correlation function. The CM can achieve a good SRO estimation accuracy, while it is very time consuming. The LCD method proposed in [167] is based on the fact that the SRO leads a phase drift to the coherence between the asynchronous microphone signals and builds a linear phase model by considering two successive speech segments. Compared to CM, the complexity of the LCD method is much lower. Hence, it is worth investigating an accurate and low-complexity SRO estimation approach.

List of Abbreviations

WASN	wireless acoustic sensor network
FC	fusion center
ADC	analog-to-digital converter
CPU	central processing unit
MVDR	minimum variance distortionless response
BMVDR	binaural minimum variance distortionless response
LCMV	linear constrained minimum variance
BLCMV	binaural linear constrained minimum variance
DOF	degrees of freedom
MWF	multichannel Wiener filter
SD-MWF	speech distortion weighted multichannel Wiener filter
GSC	generalized sidelobe canceler
MSE	mean square-error
MMSE	minimum mean square-error
SNR	signal-to-noise ratio
MSE	mean squared error
MCCC	multichannel cross-correlation coefficient
DFT	discrete Fourier transform
STFT	short-time Fourier transform
ATF	acoustic transfer function
RTF	relative (acoustic) transfer function
PSD	power spectral density
VAD	voice activity detector
LMI	linear matrix inequality
SDP	semi-definite programming
MD-MVDR	model-driven minimum variance distortionless response
MD-BLCMV	model-driven binaural linear constrained minimum variance
DD-MVDR	data-driven minimum variance distortionless response
RD-MVDR	rate-distributed minimum variance distortionless response
RD-LCMV	rate-distributed linear constrained minimum variance
RD-BLCMV	rate-distributed binaural linear constrained minimum variance
EUR	energy usage ratio
HA	hearing aid
ITF	interaural transfer function
ILD	interaural level difference
ITD	interaural time difference
IPD	interaural phase difference
ADMM	alternating direction method of multipliers
PDMM	primal-dual method of multipliers
CS	covariance subtraction
CW	covariance whitening
SOS	second-order statistics
EVD	eigenvalue decomposition
GEVD	generalized eigenvalue decomposition
MDRD-CS	model-driven rate-distributed covariance subtraction
DDRD-CS	data-driven rate-distributed covariance subtraction
MDRD-CW	model-driven rate-distributed covariance whitening
DDRD-CW	data-driven rate-distributed covariance whitening

BIBLIOGRAPHY

[1] D. C. Moore and I. A. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, vol. 5, pp. V–497.

[2] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140, 2012.

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

[4] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, vol. 1, pp. 187–190.

[5] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

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

[7] J-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2003, vol. 2, pp. 1228–1233.

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

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

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

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

[12] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

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

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

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

[16] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1988, pp. 2578–2581.

[17] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1391–1400, 1986.

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

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

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

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

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

[23] 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.

[24] J. Bourgeois and W. Minker, "Time-domain beamforming and blind source separation," *Lecture Notes in Electrical Engineering. Springer-Verlag*, 2009.

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

[26] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 1072–1077, 2010.

[27] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," *EURASIP J. Advances Signal Process.*, vol. 2009, pp. 7, 2009.

[28] 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, 2001.

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

[30] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed GSC beamforming using the relative transfer function," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 1274 – 1278.

[31] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 397–401, 1988.

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

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

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

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

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

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

[38] 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, 2009.

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

[40] 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, Language Process.*, vol. 18, no. 2, pp. 342–355, 2010.

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

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

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

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

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

[46] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 543–558, 2016.

[47] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 137–152, 2017.

[48] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2384–2397, 2015.

[49] J. Zhang, R. Heusdens, and R. C. Hendriks, "Rate-distributed binaural LCMV beamforming for assistive hearing in wireless acoustic sensor networks," in *IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018.

[50] P. P. Vaidyanathan and P. Pal, "Sparse sensing with co-prime samplers and arrays," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 573–586, 2011.

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

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

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

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

[55] S. Liu, S. P. Chepuri, M. Fardad, E. Masazade, G. Leus, and P. K. Varshney, "Sensor selection for estimation with correlated measurement noise," *IEEE Trans. Signal Process.*, 2016.

[56] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *IEEE Conf. on Decision and Control*, 2010, pp. 2572–2577.

[57] 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 Process. Lett.*, vol. 23, no. 9, pp. 1202–1206, 2016.

[58] M. Contino, S. P. Chepuri, and Geert Leus, "Near-optimal greedy sensor selection for MVDR beamforming with modular budget constraint," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2017, pp. 1981–1985.

[59] 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.

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

[61] A. Krause and C. Guestrin, "Near-optimal observation selection using submodular functions," in *Proc. of Association for the Advancement of Artificial Intelligence*, 2007, vol. 7, pp. 1650–1654.

[62] A. Krause and D. Golovin, "Submodular function maximization," *Tractability: Practical Approaches to Hard Problems*, vol. 3, no. 19, pp. 8, 2012.

[63] S. Fujishige, *Submodular functions and optimization*, vol. 58, Elsevier, 2005.

[64] 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, 2012.

[65] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[66] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[67] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Information Theory*, vol. 14, no. 5, pp. 676–683, 1968.

[68] 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. 25, no. 5, pp. 442–448, 1977.

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

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

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

[72] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, 1992.

[73] L. A. Jeffress, "A place theory of sound localization.," *J. Comparative, Physiological Psychology*, vol. 41, no. 1, pp. 35, 1948.

[74] K. Kurozumi and K. Ohgushi, "The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality," *J. Acoust. Soc. Amer.*, vol. 74, no. 6, pp. 1726–1733, 1983.

[75] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[76] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[77] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[78] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1618–1632, 2017.

[79] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[80] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2449–2464, 2015.

[81] D. Marquardt and S. Doclo, "Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 7, pp. 1261–1274, 2018.

[82] O. Roy and M. Vetterli, "Rate-constrained collaborative noise reduction for wireless hearing aids," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 645–657, 2009.

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

[84] J. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[85] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1434–1448, 2018.

[86] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[87] G. Zhang and R. Heusdens, "Distributed optimization using the primal-dual method of multipliers," *IEEE Trans. Signal and Information Process. over Networks*, vol. 4, no. 1, pp. 173–187, 2018.

[88] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Information Theory*, vol. 14, no. SI, pp. 2508–2530, 2006.

[89] G. Zhang and R. Heusdens, "Bi-alternating direction method of multipliers over graphs," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 3571–3575.

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

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

[92] X.-F. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 320–324.

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

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

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

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

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

[98] D. Golovin, M. Faulkner, and A. Krause, "Online distributed sensor selection," in *ACM/IEEE Int. Conf. on Inform. Process. in Sensor Networks*, 2010, pp. 220–231.

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

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

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

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

[103] Y. He and K. P. Chong, "Sensor scheduling for target tracking in sensor networks," in *IEEE Conf. on Decision and Control*, 2004, vol. 1, pp. 743–748.

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

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

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

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

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

[109] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20, 2010.

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

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

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

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

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

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

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

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

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

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

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

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

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

[123] J.-J. Xiao, S. Cui, Z.-Q. Luo, and A. J. Goldsmith, "Power scheduling of universal decentralized estimation in sensor networks," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 413–422, 2006.

[124] S. Cui, J.-J. Xiao, A. J. Goldsmith, Z.-Q. Luo, and H. V. Poor, "Estimation diversity and energy efficiency in distributed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4683–4695, 2007.

[125] T. C. Lawin-Ore and S. Doclo, "Analysis of rate constraints for MWF-based noise reduction in acoustic sensor networks," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 269–272.

[126] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1818–1829, 2010.

[127] 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, 2002.

[128] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[129] 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, 2010.

[130] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 661–676, 2011.

[131] V. M. Tavakoli, J. R. Jensen, R. Heusdens, J. Benesty, and M. G. Christensen, "Ad hoc microphone array beamforming using the primal-dual method of multipliers," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 1088–1092.

[132] 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, Language Process.*, vol. 25, no. 4, pp. 692–730, 2017.

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

[134] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 233 –246, Jan. 2012.

[135] T. Sherson, W. B. Kleijn, and R. Heusdens, "A distributed algorithm for robust LCMV beamforming.," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 101–105.

[136] 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, Language Process.*, vol. 26, no. 11, pp. 2015–2026, 2018.

[137] J. L. Flanagan, A. C. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, no. 1-2, pp. 207–222, 1993.

[138] S Markovich Golan, Sharon Gannot, and Israel Cohen, "A reduced bandwidth binaural MVDR beamformer," in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, 2010.

[139] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 4–20, 2015.

[140] A. Bertrand and M. Moonen, "Distributed adaptive generalized eigenvector estimation of a sensor signal covariance matrix pair in a fully connected sensor network," *ELSEVIER Signal Process.*, vol. 106, pp. 209–214, 2015.

[141] A. Hassani, J. Plata-Chaves, M. H. Bahari, M. Moonen, and A. Bertrand, "Multitask wireless sensor network for joint distributed node-specific signal enhancement, LCMV beamforming and DOA estimation," *IEEE Journal of Selected Topics in Signal Process.*, vol. 11, no. 3, pp. 518–533, 2017.

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

[143] D. H. M. Schellekens, T. Sherson, and R. Heusdens, "Quantisation effects in PDMM: A first study for synchronous distributed averaging," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 4237–4241.

[144] J. A. G. Jonkman, T. Sherson, and R. Heusdens, "Quantisation effects in distributed optimisation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 3649–3653.

[145] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP J. Advances Signal Process.*, vol. 2006, pp. 175–175, 2006.

[146] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 125, no. 1, pp. 360–371, 2009.

[147] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 1075–1084, 2017.

[148] T. J. Klasen, T. Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1579–1585, 2007.

[149] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *ELSEVIER Signal Process.*, vol. 107, pp. 68–81, 2015.

[150] S. Srinivasan, A. Pandharipande, and K. Janse, "Beamforming under quantization errors in wireless binaural hearing aids," *EURASIP J. Audio, Speech, Music Process.*, vol. 2008, no. 1, pp. 824797, 2008.

[151] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, J. Jensen, and M. Guo, "Binaural beamforming using pre-determined relative acoustic transfer functions," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2017, pp. 1–5.

[152] A. Bertrand and M. Moonen, "Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3447–3459, 2013.

[153] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 233–246, 2012.

[154] 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. and Mobile Comput.*, vol. 2017, 2017.

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

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

[157] N. R. Goodman, "Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction)," *The Annals of mathematical statistics*, vol. 34, no. 1, pp. 152–177, 1963.

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

[159] J. Zhang, R. Heusdens, and R. C. Hendriks, "Sensor selection and rate distribution based beamforming for wireless acoustic sensor networks," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, 2019.

[160] S. P. Chepuri and G. Leus, "Continuous sensor placement," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 544–548, 2015.

[161] G. K. Atia, V. V. Veeravalli, and J. A. Fuemmeler, "Sensor scheduling for energy-efficient target tracking in sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4923–4937, 2011.

[162] H. Viswanathan and T. Berger, "Sequential coding of correlated sources," *IEEE Trans. Information Theory*, vol. 46, no. 1, pp. 236–246, 2000.

[163] J. Amini, R. C. Hendriks, R. Heusdens, M. Guo, and J. Jensen, "Asymmetric coding for rate-constrained noise reduction in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 154–167, 2018.

[164] M.l Calvo-Fullana, J. Matamoros, and C. Antón-Haro, "Sensor selection and power allocation strategies for energy harvesting wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3685–3695, 2016.

[165] J. Schmalenstroeer and R. Haeb-Umbach, "Insights into the interplay of sampling rate offsets and MVDR beamforming," in *13th ITG-Symposium Speech Communication*. VDE, 2018, pp. 1–5.

[166] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 571–582, 2016.

[167] M. H. Bahari, A. Bertrand, and M. Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 3, pp. 674–686, 2017.

ACKNOWLEDGEMENTS

Time is always flying. The four years PhD life is a long journey, but it is also a short period. At the first time when I stepped on this country four years ago, I could not imagine what is waiting for me in front or what to do after these four years. Life is always unpredictable and no one can find the optimal trajectory for life, but only the local optimal solutions. Fortunately, I have achieved my local optimum of these four years. All I am I owe to the help from the people around me. Without their encouragement and help, I could not finish this thesis. I would therefore like to express my gratitude to all of them.

First of all, I would like to express my sincere gratitude to my family, particularly my wife Zhenzhen. Four years ago, you quit your previous job and came to this unacquainted country to support my dream. You abandoned all you had and started from zero. This is a great scarification for me and the only reason behind this decision is love. I keep this grace in mind all the time and turn it to the motivation for achieving my doctoral degree. I believe that my success in research would be the best reward to your unconditional love. Your understanding and support throughout these four years are everywhere, especially in spirit and finance. I also would like to thank my parents, you have given me the life and built the harbor where I can recharge whenever feeling at sea. More importantly, in the first one and half years of my PhD, you helped to raise my baby without any complaints, because during that period Zhenzhen and I did not have enough budget for supporting a three-person family and everything here was still uncertain. Honestly speaking, this period is difficult for you, for us, and also for the little baby. I would like to thank my daughter Chi as well, even though you have brought me much more things to do everyday apart from research. With you, my PhD life is not only doing research, but also learning how to be a dad. It is you who lets me know the importance of balance between family and work, it is you who makes me realize the responsibility of life, and it is you who brings me infinite hopes for the future. Over these four years, I feel that I am growing up again along with your growth up.

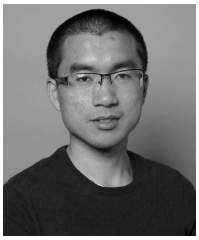
Second, I wish to express my deep thanks to my daily supervisor and promotor Richard Hendriks and my promotor Richard Heusdens. Thank you for giving me a chance to do my PhD at TU Delft. You always point out my weaknesses, clarify my confusions, and give me valuable remarks for improving my work. Without the step by step supervision from both of you, it is impossible for me to finish this thesis. It is you who make me clear about the four years research plan. It is you who let me know that the quantity of publications is not the main goal in my PhD period, but the quality of publications. It is you who have taught me how to do the right research in a right way and how to be a right researcher. Additionally, I want to mention that Richard Hendriks, a father of three kids and an efficient project manager, inspires me how to find the balance between family and work. From the view point of research, you are my supervisors; from the view point of life, you are my mentors. Once the teachers, always the teachers in life. Also, I would like to thank Prof. Alle-Jan van der Veen and Prof. Geert Leus. Although I did not discuss

much with them about my progress, they still gave me many valuable suggestions during the seminars.

Finally, I would like to thank my collaborators, e.g., Sundeep, Andreas. I feel very honored to collaborate with you and accomplish two journal papers with you. Working in the CAS group makes me very enjoyable because there are many smart and friendly colleagues around me. Many thanks to all of you. I will never forget every moment with you, especially some funny daily discussions, playing chess, the weekly sports, house warmings, the CAS outing and the Christmas dinner. With you around me, my PhD life never gets boring, but super interesting. Also, I would like to thank Antoon, Minaksie and Irma, who provide us with lots of supports outside research.

There is no never-ending feast, and we always have to say goodbye at a certain moment. Thanks to my family, thanks to my friends, thanks to time and thanks to all that have ever come to my life. I will cherish these unforgettable four years forever.

CURRICULUM VITÆ



Jie Zhang was born in Anhui Province, China, in 1990. He received the B.Sc. degree (with honors) in Electrical information Science and Technology from Yunnan University, China, in 2012 and the M.Sc. degree (with honors) from the School of Electronics and Computer Engineering, Shenzhen Graduate School, Peking University, Beijing, China, in 2015. He is currently working toward the Ph.D. degree in the Circuits and Systems (CAS) Group at the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands. His PhD research is funded by the China Scholarship Council (No. 201506010331) and in part by the CAS Group, Delft University of Technology, Delft, The Netherlands. During his PhD, he received the Best Student Paper Award from the IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM), Sheffield, UK, 2018 and a student travel grant from IEEE Signal Processing Society. His general research interests include multimicrophone speech processing for noise reduction, enhancement and sound source localization, binaural auditory, energy-aware wireless (acoustic) sensor networks.