Multi-Microphone Noise Reduction for Hearing Assistive Devices

Koutrouvelis, Andreas

DOI
10.4233/uuid:cdb32aa2-9ca4-448c-a8a0-63f458c375ff

Publication date
2018

Document Version
Final published version

Citation (APA)
Koutrouvelis, A. (2018). *Multi-Microphone Noise Reduction for Hearing Assistive Devices*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:cdb32aa2-9ca4-448c-a8a0-63f458c375ff

Multi-Microphone Noise Reduction for Hearing Assistive Devices

Multi-Microphone Noise Reduction for Hearing Assistive Devices

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
vrijdag 21 december 2018 om 12:30 uur

door

Andreas KOUTROUVELIS

Elektrotechnisch ingenieur,
Technische Universiteit Delft, Delft, Nederland
geboren te Patras, Griekenland.

Dit proefschrift is goedgekeurd door de

promotor: Dr. ir. R. Heusdens
copromotor: Dr. ir. R.C. Hendriks

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Dr. ir. R. Heusdens,	Technische Universiteit Delft, promotor
Dr. ir. R.C. Hendriks,	Technische Universiteit Delft, copromotor
Onafhankelijke leden:	
Prof. dr. ir. A. de Vries	Technische Universiteit Eindhoven
Prof. dr. ir. S. Doclo	Technische Universiteit Oldenburg, Duitsland
Prof. dr. S. Gannot	Bar-Ilan Universiteit, Israël
Prof. dr. ir. G.J.T. Leus	Technische Universiteit Delft
Prof. dr. ir. R.L. Lagendijk	Technische Universiteit Delft

This work is part of the research programme "Spatially Correct Multi-Microphone Noise Reduction Strategies suitable for Hearing Aids" with project number 13262, which is partly financed by the Netherlands Organisation for Scientific Research (NWO). In addition, this work was partly financed by the Oticon Foundation.

ISBN 978-94-6186-999-9

To my parents Ioannis and Eleni,
my sister Fani,
my love Theodora.

Contents

Summary

The paramount importance of good hearing in everyday life has driven an exploration into the improvement of hearing capabilities of (hearing impaired) people in acoustic challenging situations using hearing assistive devices (HADs). HADs are small portable devices, which primarily aim at improving the intelligibility of an acoustic source that has drawn the attention of the HAD user. One of the most important steps to achieve this is via filtering the sound recorded using the HAD microphones, such that ideally all unwanted acoustic sources in the acoustic scene are suppressed, while the target source is maintained undistorted. Modern HAD systems often consist of two collaborative (typically wirelessly connected) HADs, each placed on a different ear. These HAD systems are commonly referred to as binaural HAD systems. In a binaural HAD system, each HAD has typically more than one microphone forming a small local microphone array. The two HADs merge their microphone arrays forming a single larger microphone array. This provides more degrees of freedom for noise reduction. The multi-microphone noise reduction filters are commonly referred to as beamformers, and the beamformers designed for binaural HAD systems are commonly referred to as binaural beamformers.

Binaural beamformers typically change the magnitude and phase relations of the microphone signals by forming a beam towards the target's direction while ideally suppressing all other directions. This may alter the spatial impression of the acoustic scene, as the filtered sources now reach both ears with possibly different relative phase and magnitude differences compared to before processing. This will appear unnatural to the HAD user. Therefore, there is an increasing interest in the preservation of the spatial information (also referred to as binaural cues) of the acoustic scene after processing. Apart from the fact that binaural-cue preservation leads to a more natural impression to the user, experimental studies have shown that speech degraded by spatially separated sources has a higher intelligibility than when sources are co-located. Last but not least, incorrectly perceived spatial information can even lead to dangerous situations when, e.g., in traffic, sources are not localized correctly. Hence, it has become evident that HADs should achieve both noise reduction and binaural-cue preservation due to the aforementioned reasons. The present dissertation is mainly concerned with this particular problem and proposes several alternative binaural beamformers.

One of the biggest challenges in binaural beamforming is to exploit the available degrees of freedom to achieve optimal performance in both noise reduction and binaural-cue preservation. Typically, there is a trade-off between the two goals. Increasing noise reduction leads to worse binaural-cue preservation, while a better binaural-cue preservation implies worse noise reduction performance. The trade-off between the two goals can be based on preference or objective psychoacoustic criteria. In the current thesis we propose methods within both frameworks, where

the user can manually or semi-automatically selects the trade-off.

Noise reduction using microphones from different devices (as in binaural HAD systems) poses a difficult task on how to share the computations among the devices. The simplest way of achieving such a task is to select one of the devices as the fusion center and perform all computations centrally. Finally, in some applications the fusion center should broadcast the result to the other devices. The main limitation of such a centralized system is the lack of robustness, since the whole system depends on a single device. Moreover, the fusion center needs to store all microphone recordings from all devices which sometimes becomes impractical when there is limited storage capacity. Finally, performing all computations in a single device typically leads to a larger battery consumption of this device and larger overall delays of the system due to the limited computational power of the fusion center. Processing in such sensor networks is also very relevant to HAD systems. It becomes even more relevant nowadays where the trend is to use additional microphones from other portable devices such as mobile phones in order to increase the available degrees of freedom and therefore achieve a better trade-off between noise reduction and binaural-cue preservation.

To tackle the problems of centralized implementations, distributed (iterative) implementations are preferable which distribute the calculations over all devices. The challenge in distributed implementations is how often and how much the devices need to communicate to each other in order to converge to the same result as the centralized implementations. Preferably, they should not have larger communications costs than the centralized implementations. This is due to the fact that communication is one of the most important factors for battery consumption. In this thesis, we propose several effective distributed noise reduction methods which can tackle all the aforementioned problems of the centralized implementations and have minimal communication costs compared to other existing methods.

1

Introduction

HEARING assistive devices (HADs) [1, 2] have become increasingly important in society. Being able to hear and understand spoken messages and conversations is important when taking an active role in society. Hearing aids and cochlear implants are the most well-known examples of HADs, although also other devices equipped with a set of microphones, a processing unit, and a loudspeaker could be used as a HAD. Using the microphones, HADs acquire the acoustic signals in the environment, and, after processing, play them back using the loudspeaker. The processing aims at improving the hearing capabilities of the user in complex acoustic scenarios and optionally compensate the hearing loss in case of a hearing-impaired user. For instance, a HAD can improve the speech clarity/intelligibility of an attended talker thereby reducing the listener fatigue.

Figure 1.1, depicts a high-level overview of a HAD, which consists of four main blocks: a microphone array, a processing unit, a loudspeaker, and an optional transmitter/receiver. The microphone array is a transducer which captures the acoustic mechanical waves and converts them to analog electric signals. The processing unit first converts the analog signals to digital signals via an analog-to-digital converter (ADC), and then processes the digital signals such that they become useful to the HAD user. Finally, it converts the processed digital signals back to analog signals through a digital-to-analog converter (DAC) and then send them to the loudspeaker. The loudspeaker is a transducer which converts the analog electric signals to acoustic mechanical waves traveling in the ear canal of the HAD user. The transmitter/receiver unit is sending and receiving signals from other devices that may collaborate with the HAD.

The processing unit consists, in addition to the ADC and DAC, of three main blocks: *feedback cancellation*, *noise reduction*, and optionally *hearing-loss compensation*. In some HADs the loudspeaker is very close to the microphones so that the microphones acquire a portion of the sound produced by the loudspeakers. This effect is called (acoustical) *feedback* and causes annoying artifacts, like howling, which

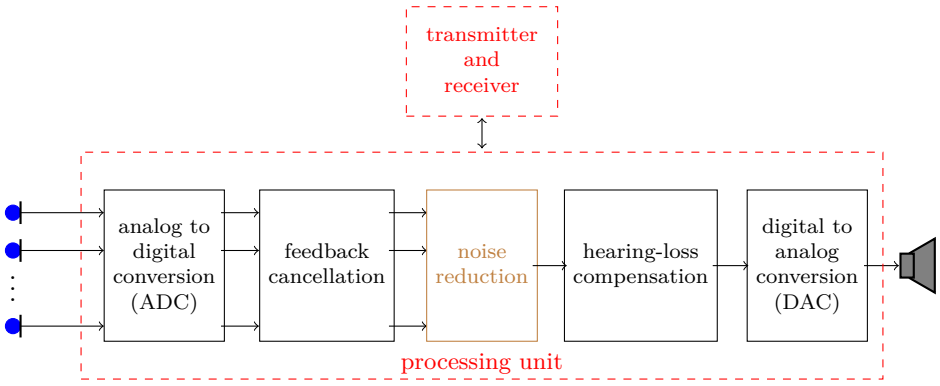

Figure 1.1: A high-level overview of a hearing assistive device (HAD). On the left-hand side, the blue circles are the microphones and on the right-hand side is the loudspeaker. The transmitter/receiver block is included when the HAD is collaborating with other external devices. The hearing-loss compensation block is included mainly in hearing-aid devices meant for hearing-impaired people.

need to be reduced through a feedback cancellation method [3–6]. After feedback cancellation, noise reduction [7–10] takes place which tries to reduce all unwanted acoustic sources while keeping the target source unaltered. This helps the HAD user to concentrate and understand the content of conversations without large listening effort. After noise reduction, for hearing impaired users, hearing-loss compensation takes place. The most common problem for hearing-impaired people is the fact that they cannot hear certain frequencies at similar low-intensity levels as normal-hearing people. Therefore, these frequencies are amplified in the hearing-loss compensation module using a frequency-dependent gain function which is based on the individual's measured hearing loss. However, since the maximum allowable loudness is more or less fixed, the dynamic range between the minimum audible loudness and the maximum allowable loudness reduces. In noisy acoustic environments, the hearing-loss compensation may be insufficient to obtain well intelligible speech, which is due to several reasons. At first, the hearing-loss compensation also amplifies acoustic noise. Even though inaudible before amplification, after amplification it can mask the target signal. Typically, problems of low-intelligible speech are further increased due to the reduced dynamic range in combination with a reduced time and frequency resolution of the impaired hearing system. As a consequence, noise reduction is needed to suppress the acoustic noise as much as possible.

In this dissertation, we mainly focus on the noise reduction block in Figure 1.1. The noise reduction performance can be increased significantly if multiple microphones are used compared to a single microphone [10]. Each sound source in the acoustic scene has a unique spatial signature which is the location of this source with respect to the locations of the microphones. The locations of the sources can be estimated from the microphone signals and exploited by the noise reduction algorithm to maintain or suppress sound sources coming from specific locations. More specifically, the sound sources coming from different locations reach the mi-

crophones at slightly different time instances and with different intensities. Thus, the multi-microphone noise reduction algorithm can properly delay and attenuate the microphone signals such that sound sources from certain locations are maintained, while others are suppressed. Although the noise reduction improves with the number of microphones, due to space and hardware limitations, usually only a few microphones (2 or 3) are used in a typical HAD.

The time and intensity differences are not only exploited by the noise reduction algorithm, but also by the human brain in order to localize sound sources. More specifically, the auditory system of the human brain exploits time and intensity differences between the two ears (see Figure 1.2), which are referred to as *binaural cues* [11], in order to localize sound. The time difference arises due to the difference in distance between the left ear and the sound source and the distance between the right ear and the sound source. This also introduces intensity differences. However, intensity differences are even more emphasized due to the presence of the head, which attenuates the sound reaching the ear which is on the opposite side of the sound source with respect to the head.

Usually, a HAD user wears two HADs. If the original time and intensity differences remain unaltered after processing, the HAD user will be still able to correctly localize the sound sources. However, as explained before, a multi-microphone noise reduction algorithm modifies the time and intensity differences of the microphone signals. As a result, without taking special measures against binaural-cue distortions, the spatial impression of the HAD user after processing will be distorted [2]. To maintain the location of the sound source unaltered after processing, the time and intensity differences should be preserved after noise reduction.

Although sometimes people use a single-device HAD system [1], also called a monaural system, most commonly HADs come in pairs with multiple microphones per device. The two HADs can work independently, or collaborate through a communication link established between them. The first system is referred to as a *bilateral* HAD system [1], while the second system is referred to as a *binaural* HAD system [1]. The binaural HAD system can provide improved noise reduction performance compared to the bilateral system because the two microphone arrays of the two HADs are merged into a single larger microphone array which can be used by the noise reduction algorithm.

An additional benefit to the improved noise reduction performance in binaural HAD systems, is the fact that by using microphone signals from both HADs, spatial information from both sides of the head is captured and can be used to provide the correct spatial impression of the acoustic scene to the HAD user. The multi-microphone noise reduction performed in binaural HAD systems is typically referred to as *binaural multi-microphone noise reduction*. Binaural multi-microphone noise reduction methods typically aim at simultaneous noise reduction and preservation of binaural cues [2], by trading noise reduction against binaural-cue preservation [2]. Therefore, the challenge is to optimally design the trade-off such that the intelligibility will be maximized without perceived binaural-cue distortions, resulting in processed signals that sound as natural as possible.

Besides naturalness and safety aspects (e.g., think of the application in a traffic

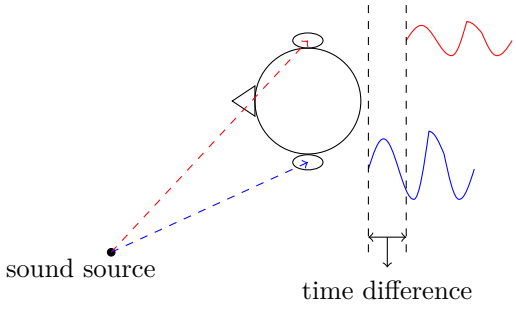

Figure 1.2: A sound source arriving at different times at both ears. It will arrive first to the left and then to the right ear. The head attenuates more the captured sound from the right ear compared to the left.

scenario) binaural cues are also important for intelligibility [2, 12]. This is due to the *binaural release from masking* effect of the auditory system [2, 12, 13]. In particular, it has been experimentally shown that if a speech signal and an interferer are co-located, it is easier for the interferer to mask the speech signal, compared to the situation where the sources are coming from different directions [12]. This observation motivated researchers even further to search for designs of binaural multi-microphone noise reduction methods that do not harm the spatial impression of the HAD user after noise reduction.

Apart from noise reduction and binaural-cue preservation, another important aspect in binaural HAD systems is power consumption. This is mainly due to the computational complexity of the noise reduction algorithms and the communication costs between different HADs. Specifically, since the two HADs in the binaural system communicate via the transmitter/receiver block (see Figure 1.1), they have to carefully select what information should be exchanged. The simplest strategy is to send all microphone signals from the left device to the right and vice versa. This strategy is very inefficient because of the large bit-rate used by the HAD system to send partly redundant information. In order to avoid large communication costs, a data-compression algorithm can be used in the transmitter/receiver block to reduce the bit-rate [14–17]. One of the main steps of the compression algorithm is the quantization of the signals. The quantization noise added to the signals needs to be controlled such that its impact is minimized, while satisfying constraints on the bit-rate and energy consumption needed for transmission [15–17]. In this dissertation we assume that the microphone signals from both hearing aids are available without quantization noise. Moreover, the computations should be shared among the devices such that the battery power consumption is reduced per device. This requires efficient distributed algorithms that solve the binaural multi-microphone noise reduction problem by first splitting this into sub-problems which are solved independently at each device, followed by combining the solutions of the sub-problems

to form the global solution [18].

There are two main categories of multi-microphone noise reduction methods [2, 8]. The first consists of *spatial filtering* methods (also known as *beamforming* methods), which minimize the output noise power such that the target signal is undistorted. The second category consists of *spatio-temporal filtering* methods, which exploit both the spatial and temporal structure of the acoustic scene. Unlike spatial filtering, spatio-temporal filtering allows distortions to the target signal, but typically achieves increased noise reduction improvement. The vast majority of multi-microphone noise reduction methods are based on linear filtering due to its low-complexity and simplicity. Typically, spatio-temporal filters can be expressed as the concatenation of a spatial filter and a single-channel temporal filter applied to the output of the spatial filter [8, 19].

The remaining part of this chapter is organized as follows. To introduce the problem of multi-microphone noise reduction, we start in Section 1.1 with a short overview on single-device (monaural) spatial filtering and its challenges. In Section 1.2, we give a short overview of single-device spatio-temporal filtering and its challenges. In Section 1.3, we review the most important binaural cues. In Section 1.4, we briefly review the most important binaural multi-microphone methods existing within the literature. In Section 1.5, we review some well-known distributed multi-microphone noise reduction approaches. In Section 1.6, we list the research questions addressed in this dissertation. In Section 1.7, we summarize the contributions of the current dissertation organized per-chapter and, finally, in Section 1.8, we list all papers that comprise the contributions of this dissertation.

1.1. Spatial Filtering

The minimum variance distortionless response beamformer (MVDR) [20, 21] is one of the simplest existing beamformers which minimizes the output noise power such that the target signal is undistorted after filtering. It is very popular because i) it is the best-performing method in terms of noise reduction among all linear spatial filters, and ii) has a closed-form expression resulting in very fast implementations. A widely-used extension of the MVDR beamfomer is the linearly-constrained minimum variance (LCMV) beamformer [22, 23] which has additional linear equality constraints compared to the MVDR beamfomer. These additional constraints can be used to have a more user-defined control on the spatial behavior of the beamformer. For instance, nulling constraints can be used in order to cancel interferers that are crucial to be eliminated.

The aforementioned two spatial filters depend on i) estimates of the multi-microphone second-order statistics of the acoustic scene, also known as the noisy cross-power spectral density matrix (CPSDM), and ii) estimates of the acoustic transfer functions (ATFs) of the target source and possibly of the interferers. The ATFs contain the spatial relationship between the sources and the microphones. Theoretically it has been shown that if the MVDR and LCMV beamformers use a perfect estimate of the ATFs of the target source, they will not distort the target signal. However, in practice this is not true due to the inaccurately estimated ATFs, and the fact that the estimated noisy CPSDM contains information about the target

source as well. This is a fundamental problem in spatial filtering which has been investigated for many years (see e.g., [24, 25] for an overview).

Although several ATF estimators [26–31] and CPSDM estimators [30, 32, 33] have been proposed, both the ATF and the CPSDM are prone to estimation errors. Many alternative robust spatial filtering methods have been proposed in order to reduce susceptibility on ATF estimation errors (see e.g., [21, 34–40]). The robust spatial filtering methods can roughly be divided into four main categories. The first category of methods introduce extra inequality/uncertainty constraints to the basic spatial filtering problem to prevent target distortions (see e.g., [39]). The second category adds a diagonal loading to the CPSDM [35, 38]. The third category uses the noise-only CPSDM (see e.g., [21, 34]) and, therefore, the target signal is not suppressed from the objective function of the filter. The fourth category uses CPSDMs which are data-independent and fixed (see e.g., [36, 37, 40]) and, similar to the noise-only CPSDM, the objective function does not suppress the target signal.

1.2. Spatio-temporal Filtering

Several spatio-temporal filters can be split into a spatial filter and a single-channel post-filter [8, 37, 41]. A commonly used post-filter is the single-channel Wiener filter [7, 8], although depending on the statistical assumptions, other post-filters are optimal [42]. The most well-known spatio-temporal filtering method is the multi-channel Wiener filter (MWF) [8], which minimizes the mean square-error between the true target signal and the estimated target signal. The MWF can be decomposed as the concatenation of an MVDR beamformer and a single-channel Wiener.

The performance of the MWF method depends on the accuracy of the estimated target and noise CPSDMs. When there are large estimation errors, there is performance degradation. One of the most unwanted aspects of performance degradation are target distortions. Apart from being able to use robust spatial filters in the decomposition of the spatio-temporal filter, one can also reduce target distortions using the speech distortion weighted MWF (SD-MWF) method [43] which has a trade-off parameter which controls the speech distortion and noise reduction performance. The MVDR and MWF methods are special cases of the SD-MWF method.

1.3. Binaural Cues

The auditory system of the brain exploits the binaural cues (e.g., intensity and time differences between the two ears) in order to localize a sound source. Binaural cues become more dominant in certain frequency bands. As such, binaural cues are typically described and analyzed in the frequency domain in which time and intensity differences are translated into phase and magnitude differences, respectively. Specifically, binaural cues can be divided into three main categories: the *interaural level difference* [11, 44], the *interaural phase difference* [11, 44], and the *interaural coherence* [45]. The first two are directional binaural cues which are responsible for the localization of the acoustic sources, while the third one is responsible for understanding the perceived width of diffuse sound fields and the perceived distance of the sound sources in the acoustic scene [45].

The difference in level (intensity) between the two ears is due to two main reasons: i) the difference in distance between the two ears and the acoustic source, and ii) the head shadow effect which becomes more important for high frequencies [11]. The interaural level difference is mainly exploited by the auditory system for frequencies above 3 kHz [11]. The phase difference mainly occurs due to the time difference of arrival of the source signal between the two ears and is mainly exploited by the auditory system for frequencies below 1.5 kHz [11]. It becomes apparent that neither the interaural level or phase differences are well exploited by the auditory system in the frequency range 1.5 to 3 kHz and, thus, the localization ability is very limited in this frequency range [11].

From Sections 1.1 and 1.2 it is clear that multi-microphone noise reduction algorithms employ phase and magnitude differences between microphones, and as such, change the original phase and magnitude of all sources. Commonly, this leads to the situation where the binaural cues of the sound sources after processing are different compared to before processing. As a result, binaural cues are distorted or lost. To overcome this, binaural multi-microphone noise reduction algorithms can be used.

1.4. Binaural multi-microphone noise reduction

Binaural multi-microphone noise reduction aims at both noise reduction and binaural-cue preservation. This means that after processing, the binaural cues introduced in Section 1.3, should be identical to those before processing. Several methods have been proposed within the literature that aim at achieving both goals, but there seems to be an inevitable trade-off between them.

The MVDR and LCMV beamformers (described in Section 1.1) can be easily extended to the binaural context. The binaural MVDR (BMVDR) beamformer [2, 46] is the binaural version of the MVDR beamformer and is the best-performing binaural beamformer in terms of noise reduction among all linear binaural spatial filters. Several perceptual evaluations have shown that the BMVDR also achieves the best intelligibility among many other proposed binaural spatial filters [47, 48]. The large noise reduction improvement of the BMVDR comes with a price on the binaural-cue distortions [2, 46]. In particular, it has been shown that after processing the noisy acoustic scene with the BMVDR, the unwanted sources (interferers and background noise) are perceived as coming from the same location as the target source [2, 46]. Thus, there is no spatial separation of the target and residual noise sources after processing. The lack of spatial separation of the sources constrains the intelligibility improvement due to the vanished binaural realise of masking. More specifically, in [47, 48], it was experimentally shown that there was a significant intelligibility improvement of a non-practical oracle-based method, which has the same noise reduction performance as the BMVDR, but, unlike the BMVDR, preserves the correct spatial information. Therefore, a natural question that arises is whether there are other linear binaural spatial filters that can preserve the binaural cues of the unwanted sources with larger intelligibility improvement compared to the BMVDR.

The SD-MWF spatio-temporal method can also be easily applied in the binaural context. Specifically, the binaural SD-MWF (BSD-MWF) has been proposed in [49, 50], which can be decomposed into the BMVDR filter and a single-channel Wiener

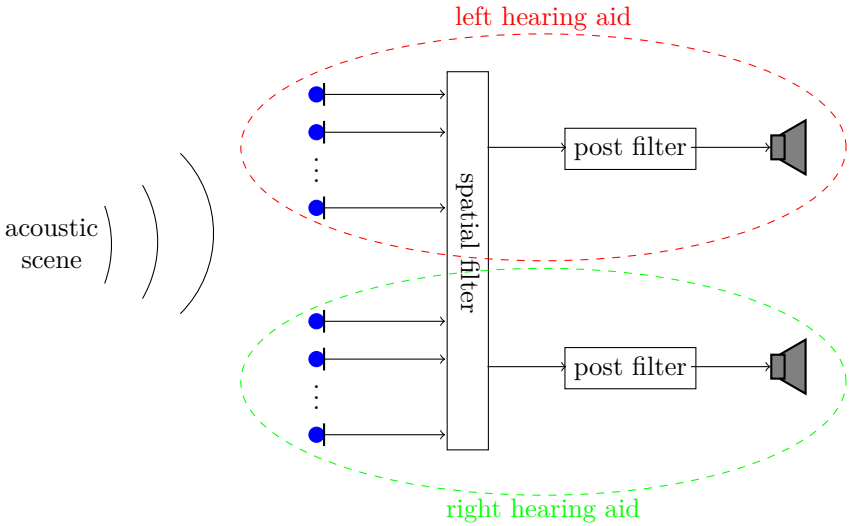

Figure 1.3: Binaural spatio-temporal filter as a concatenation of a spatial and two temporal filters. On the left-hand side the blue circles are the microphones, and on the right-hand side the gray figures are the two loudspeakers.

post-filter with speech-distortion control applied to both outputs of the BMVDR filter [51, 52]. The general structure of a binaural spatio-temporal filter is depicted in Figure 1.3. Similarly to the BMVDR beamformer, the BSD-MWF filter preserves the binaural cues of the target, but causes severe distortions to the binaural cues of the noise components.

Although, the BMVDR and BSD-MWF have good noise reduction capabilities, they both fail in preserving the binaural cues of the noise sources. As such, several other binaural spatial or spatio-temporal filters have been proposed which provide different trade-offs between noise reduction and binaural-cue preservation of the noise and can be classified mainly into three categories as depicted in Figure 1.4.

The first category consists of methods (see e.g., [46, 52–54]) that try to preserve the correct locations of the point sources, but ignore the binaural cues of the diffuse noise field. The first category can be further split into two sub-categories. In the first sub-category, there is control on the power of the interferers at the output of the filter [53], while in the second sub-category there is not [46, 52, 54]. Specifically, in the second sub-category, the interferers are suppressed as much as possible. The methods belonging to the first category use linear equality constraints to preserve the binaural cues of the interferers. This type of constraints may exhaust very quickly the degrees of freedom for noise reduction when the number of interferers needed to be preserved increases.

The second category of binaural multi-microphone noise reduction methods consists of methods that aim at preserving the binaural cues of diffuse noise fields [19, 47, 55], i.e., acoustic fields where the sound is coming from all directions. This is accomplished by using non-linear inequality constraints on the interaural coherence

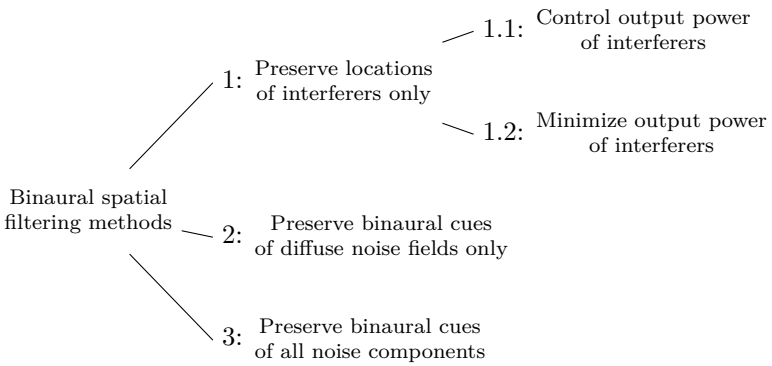

Figure 1.4: Classification of spatial and spatio-temporal filtering methods according to the binaural-cue preservation goals.

of the sound field. These methods however do not preserve the directional binaural cues of the point interfering sources. In addition, these methods do not have closed-form solutions due to the included inequality constraints and can be solved iteratively.

The third category of binaural multi-microphone noise reduction methods consists of methods that aim at both preserving the binaural cues of the interfering point sources and the diffuse noise field [19, 47, 50, 56]. In reality, these sound fields are the most common ones and, thus, these methods provide the most natural impression to the user. However, their task to preserve the spatial information from so many noise components reduces the degrees of freedom for noise reduction and makes it hard to obtain a good noise reduction performance.

For most binaural spatial filters, one of the biggest challenges is how to design the trade-off between binaural-cue preservation and noise reduction. Existing trade-offs are based on two main concepts. The first concept is flexibility in which the user can easily tune a trade-off parameter and put more emphasis on one of the two goals. The other concept is simplicity which is to have (semi) automatic methods that select the trade-off. This latter concept may be based on psychoacoustic criteria.

1.5. Distributed Multi-Microphone Noise Reduction

So far we have discussed multi-microphone methods exploiting the microphone signals from either one or two devices. In the latter case, the two devices form a (small) wireless network exchanging information. Of course this can be extended to multiple (more than two) devices, (see Figure 1.5). Such a network can be viewed as a graph where its nodes are the devices and its edges the links between the devices. If there is no edge between two devices, there is no direct communication between these devices. This does not mean necessarily that the information cannot reach both devices since other devices of the network may be connected with these devices.

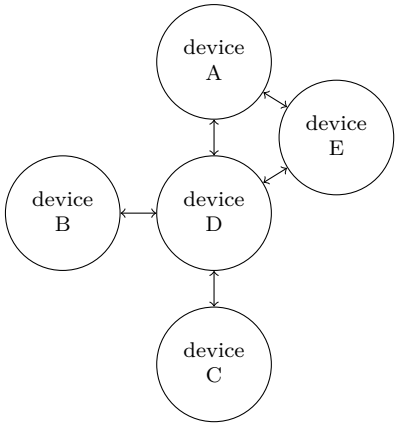

Figure 1.5: This is an example of a general cyclic network. In this example, only the nearby devices communicate with each other.

Spatial filtering requires two main steps in the computations. The first is to compute the spatial filter and the second is to apply the spatial filter to the microphone signals of all devices. In order to compute the spatial filter, the joint estimated noise or noisy CPSDM from all devices is needed. Obviously, both steps are impossible without direct or indirect communication between the devices.

Two questions naturally arise here: i) Which information should be exchanged between the devices and ii) how should the calculations be distributed over the devices. There are two main methodologies of computing a spatial filter over a network of devices. The first methodology (*centralized implementation*) is to perform all computations in a central fusion center which can be one of the devices in the network. The final outcome of the computation can then be broadcast to all other devices. The second methodology (*distributed implementation*) is to distribute the computations over all devices in the network. In the following, we will highlight the pros and cons of both methodologies and explain why in certain cases a distributed implementation is preferable over a centralized one.

There are five main aspects in which both methodologies have to be compared. The first aspect is the performance of the multi-microphone noise reduction method in terms of e.g., noise reduction. The second aspect are the communication costs which mainly depend on the bit-rate and the distance between the wirelessly connected devices. The third aspect is the computational complexity of an implementation. The fourth aspect is the limited data storage capabilities of the devices. The fifth aspect is the robustness of the entire system due to malfunction or disconnection of some nodes.

Spatial filtering requires an estimate of the joint CPSDM, which is computed using all microphone signals over a long-enough time interval and is typically adapted continuously. In a centralized implementation all nodes need to send their local microphone signals to the fusion center. This results in a huge data flooding in the

network. Therefore, a low bit-rate should be used to avoid problems with respect to the channel capacity. A low bit-rate results in large quantization noise and, consequently, in low performance of the system. After the fusion center obtains all microphone signals it needs to save them. Therefore, in case there are many devices in the network, the fusion center needs a large memory which is less practical when the fusion center is a small portable device. After storage, the fusion center needs to compute the joint CPSDM and compute the spatial filter which typically requires the inversion of a matrix with dimensions depending on the size of the network. All these heavy computations are performed in a single device. This will most likely result in fast battery consumption of the device selected as the fusion center. Moreover, such a network is very sensitive to a sudden malfunction or disconnection of the fusion center from the network. Finally, the fusion center is often required to broadcast the outcome of the filtering operation to all the other nodes which adds extra communication costs to all devices.

There are several distributed multi-microphone noise reduction methods which try to avoid some or all of the aforementioned problems of the corresponding centralized approaches. The general idea is that all multi-microphone noise reduction methods can be written in the form of an optimization problem. If the optimization problem has a seperable objective function and constraints, it is possible to have a distributed implementation [57]. Typically, distributed multi-microphone noise reduction methods are classified into two main categories.

The first category consists of sub-optimal methods, which either partially estimate the CPSDM or not estimate the CPSDM at all, but rather use a fixed spatial coherence matrix. The method proposed in [58] is the distributed implementation of the delay and sum beamformer and is based on the randomized gossip algorithm [59]. In particular, it replaces the CPSDM with the identity matrix as in the case of the centralized delay and sum beamformer. This fixed CPSDM choice makes the objective function separable and allows this method to run in general *cyclic networks* (i.e., networks that may contain loops). An example of a cyclic network is demonstrated in Figure 1.5. The method proposed in [60] is a distributed implementation of the MVDR spatial filter and is based on message passing [61, 62]. Unlike the method proposed in [58], which does not exploit the second-order statistics of the noise, the method proposed in [60] estimates the full CPSDM matrix but also adds a diagonal loading parameter to it in order to control the convergence rate. When increasing this diagonal loading parameter, the convergence becomes faster. At the extreme case, where the diagonal loading becomes large, the methods in [58] and [60] have equivalent performance. Although both methods are very simple, they are sub-optimal since they do not compute spatial filters based on the best possible estimate of the CPSDM.

The second category consists of optimal methods which typically solve the optimization problem iteratively and reach optimality (i.e., equivalent performance with the corresponding centralized optimal method) after some iterations. The method proposed in [63] is the distributed implementation of the LCMV spatial filter. Although it reaches an optimal performance after some iterations, its implementation is based only on acyclic networks (i.e., networks that are tree-structured). This con-

strain its applicability in general cyclic networks. An alternative distributed implementation of the LCMV spatial filter was proposed in [64]. This method overcomes this problem and can work in arbitrary cyclic networks, but with slower convergence rate compared to [63]. If the convergence rate is very slow, the number of iterations and, thus, the number of exchanged messages between the devices, increases drastically. To the best of our knowledgem there is no investigation available on the real difference in communication-costs between the distributed implementations and the corresponding centralized implementations as a function of the convergence rate.

It becomes clear that a distributed beamformer should be designed not only based on optimal performance but also on being applicable in general cyclic networks. Moreover, the communication and computation costs should remain as low as possible.

1.6. Research Questions

The majority of the work presented in this dissertation was done within the context of the project entitled "Spatially Correct Multi-Microphone Noise Reduction Strategies Suitable for Hearing Aids" a project funded by the Dutch technology foundation STW and the hearing-aid company Oticon A/S. This project consists of two work packages, WP1 and WP2. Most content of the current dissertation is the outcome of WP1 entitled "Spatially Optimal Multi-Channel Noise Reduction Techniques". This work package is about developing binaural noise reduction methods which provide efficient trade-offs between noise reduction and binaural-cue preservation. WP2 is entitled "Resource-Constrained Multi-Channel Noise Reduction Techniques for Hearing Aids" and focuses on resource allocation (e.g., power usage, latency etc.) when using multiple hearing aids.

This dissertation also addresses general problems of multi-microphone noise reduction which can be applied to binaural hearing-aids as well. The research questions that will be addressed in this dissertation have partially been derived from WP1 and can be formulated as follows:

Q1: Binaural multi-microphone noise reduction aims at both intelligibility improvement and binaural-cue preservation. Always, there is a trade-off between these two goals. *Can we find binaural multi-microphone noise reduction methods that can (approximately) preserve the binaural cues of all sources in the acoustic scene while at the same time improve intelligibility?*

Q2: The performance of multi-microphone noise reduction methods depends on several parameters such as the ATFs of the sources, the CPSDMs of the sources, etc. Estimation of these parameters is required. *Can we accurately estimate these parameters?*

Q3: Following research question **Q2**, another natural research question that arises is the following. *Can we develop multi-microphone noise reduction methods that are robust to estimation errors in the signal model parameters?*

1.7. Dissertation Contributions and Outline

In this section, the contributions of the dissertation are summarized per chapter.

1.7.1. Chapter 2

This chapter introduces the signal model and assumptions used in the current dissertation. We review the mathematical description, advantages and disadvantages of the most important existing (binaural) spatial and spatio-temporal filters in the literature. In addition, we review existing spatial filters that are robust against ATF estimation errors. Finally, we review some state-of-the-art distributed optimization methods which can be used to obtain spatial or spatio-temporal filters in a distributed fashion.

1.7.2. Chapter 3

In this chapter, we review in more detail the equality-constrained binaural spatial filtering methods in [46, 53, 54], which preserve (in theory) accurately the directional binaural cues of the point sources. These methods exhaust very quickly the degrees of freedom for noise reduction when the number of point sources that are to be preserved increase. Therefore, in this chapter, we propose an alternative method which replaces the equality constraints with inequality constraints [65]. While equality constraints preserve the binaural cues exactly, inequality constraints approximately preserve the binaural cues without reducing significantly the noise reduction performance.

The proposed method and the equality constrained methods in [46, 53, 54] belong to the first category of binaural multi-microphone noise reduction methods (see Figure 1.4) and aim at noise reduction and preservation of the locations of the point sources. Unlike the equality-constrained method in [46, 54], which uses one linear equality constraint per interferer to preserve its location, the proposed inequality-constrained method uses one inequality constraint per interferer. The inequality constraint depends on a parameter which controls how accurate the preservation of the location of the interferer will be. Therefore, this parameter provides a flexible trade-off between binaural-cue preservation and noise reduction. The method proposed in [65] has as a special case the BMVDR and the equality-constrained method [46, 54]. In fact, the trade-off parameter has been designed in such a way that if it is selected to provide the worst binaural-cue preservation accuracy, the BMVDR beamformer is obtained. On the other hand, if this trade-off parameter is selected to provide the best possible binaural-cue preservation accuracy, then the proposed equality-constrained method in [46, 54] is obtained as a special case.

The inequality constraints increase the feasibility set of the optimization problem compared to the equality constraints and, therefore, extra noise reduction can be achieved by sacrificing some of the binaural-cue preservation accuracy. Moreover, the proposed method can use many more constraints than the equality-constrained approaches and, thus, is more appropriate to more complicated scenarios with many more sources.

Unlike the equality-constrained method in [46, 54] which has a closed-form solution, the main drawback of the proposed inequality-constrained approach is its

non-convex problem formulation, which needs to be solved iteratively. We propose a sub-optimal successive convex optimization method to approximately solve this problem.

1.7.3. Chapter 4

The proposed method of Chapter 3 and the methods proposed in [46, 53, 54] require estimates of the ATFs of the target and the interfering sources. Several approaches have been proposed to estimate the ATFs of the sources but, unfortunately, they are based on many assumptions which are not always valid in practical acoustical scenarios. In this chapter we propose a new methodology to preserve the binaural cues of the interfering sources using pre-determined anechoic ATFs [66]. These pre-determined ATFs are related to fixed azimuths and/or elevations around the head, which cover a grid of the entire space around the head. The higher the resolution of this grid, the better the binaural-cue preservation. On the other hand, the higher the resolution, the worse the noise reduction performance will be. This methodology can be used in all methods discussed in Chapter 3. However, the most appropriate method is the inequality-constrained method [65], that we propose in Chapter 3, since it allows to use many more constraints than the equality-constrained methods and, therefore, provides a much better resolution on the grid.

Using pre-determined ATFs, we avoid to estimate the actual ATFs of the interferers, simplifying binaural spatial filtering significantly. This is very convenient, especially in acoustic environments where the sources and or the head of the hearing-aid user are moving continuously. A situation in which it is very difficult to track multiple ATFs continuously.

1.7.4. Chapter 5

Although the inequality-constrained method of Chapter 3 provides a very flexible trade-off between noise reduction and binaural-cue preservation, the complexity remains prohibitive mainly due to the multiple convex optimization problems that need to be solved per time-frequency bin.

In this chapter, we propose a less complex sub-optimal method [67] to solve the non-convex optimization problem proposed in Chapter 3. The sub-optimal method is based on the semidefinite relaxation principle [68] and requires to solve a single convex optimization problem per time-frequency bin. This method reduces significantly the computations, while at the same time achieves, in some cases, a slightly better trade-off between noise reduction and binaural-cue preservation than the sub-optimal method proposed in Chapter 3.

1.7.5. Chapter 6

Unlike Chapters 3, 4 and 5, which mainly focus on preserving the correct locations of the sources, in this chapter we propose a new binaural spatio-temporal filtering method [69], This method belongs to the third category mentioned in Section 1.4 (see Figure 1.4) and is able to preserve both the locations of the sources as well as the binaural cues of the diffuse noise field. With this method we also propose a more perceptually oriented trade-off between noise reduction and binaural-cue

preservation.

The method consists of two main phases. In the first phase the BMVDR filter is used to obtain an estimate of the target signal. If the residual noise included in this estimate dominates the target signal, then the BMVDR filter is not only needless but also harmful for the binaural cues of the residual noise. In this case, a scaled version of the noisy acoustic scene is provided at the output instead of the BMVDR output. On the other hand, if the target signal dominates the residual noise after processing, then there is no reason to preserve the binaural cues of the residual noise and, thus, the BMVDR filter is used to provide the best noise reduction performance.

In other words, this method applies as much noise reduction as possible, if the residual noise obtained after processing is not audible to the HAD user. If there is a large amount of residual noise after filtering that is audible, we just preserve its spatial cues by maintaining a scaled version of the noisy acoustic scene.

1.7.6. Chapter 7

In Chapters 3, 4 and 6, the proposed methods are evaluated using objective measures. Sometimes these measures fail to reveal all possible perceptual differences between the proposed methods. In this chapter, a subjective evaluation is provided [48], which gives more insights in the true capabilities and limitations of the proposed methods. Specifically, we evaluate the true intelligibility improvement and the true localization accuracy of the proposed methods. Note that this chapter does not subjectively compare the method proposed in Chapter 5. However, in Chapter 5, we provide separate subjective evaluations with respect to the method proposed in Chapter 5.

1.7.7. Chapter 8

In all previous chapters, we mainly discussed binaural multi-microphone noise reduction methods. In this chapter, we focus on general spatial filtering methods which can be easily applied in the context of binaural spatial filtering. This chapter is separated into two main parts. In the first part, we propose new low-complexity linearly-constrained beamformers which are robust to ATF estimation errors [67]. The optimization problems associated to the proposed beamformers have also a naturally separable objective function and constraints which makes them ideal candidates for distributed implementations. The second part of this chapter proposes two alternative distributed implementations of these beamformers [67] based on the message passing algorithm [60–62] and the primal dual method of multipliers algorithm [70].

1.7.8. Chapter 9

We can conclude from all previous chapters that there is a large interest on finding methods that can give accurate estimates of signal model parameters in practical acoustic scenes with moving and highly non-stationary sources. Among these are the ATFs and the power spectral densities of the sources and the late reverberation. The estimated PSDs combined with the estimated ATFs can provide parametric estimates of the noise and target CPSDMs. These parameters are useful in

many multi-microphone-based applications such as source separation, dereverberation, binaural multi-microphone noise reduction, source tracking and localization, room geometry estimation, etc.

In this chapter, we propose several optimization problems which can jointly estimate the aforementioned parameters using the combination of two theories: confirmatory factor analysis [71–73] and non-orthogonal joint diagonalization [26]. The combination and the careful adjustment of these two theories in the context of microphone arrays gives us a powerful tool to estimate these parameters accurately.

Specifically, the non-orthogonal joint diagonalization method proposed in [26] jointly estimates the power spectral densities and ATFs of the sources and the PSDs of the microphones self noises. Unfortunately, this method does not guarantee positive estimated PSDs leading to improper solutions. The confirmatory factor analysis method proposed in [73] can be easily adjusted to perform non-orthogonal joint diagonalization and at the same time guarantee positive estimated PSDs. We introduce additional linear constraints and simple box constraints to the parameters to be estimated to increase robustness. Finally, we bring into play the late reverberation component in our methods which has been omitted from the method proposed in [26]. The contribution of the late reverberation in the acoustic scene is typically large and not taking this into account in the optimization problem will lead to performance degradation.

1.7.9. Chapter 10

In this chapter, we draw some final conclusions of this dissertation and discuss possible directions for future investigation. We discuss several theoretical and practical open questions and give suggestions on how future research could address these questions.

1.8. List of Papers

In this section, we list all papers submitted and published during the whole period of the PhD study.

Journals

1. **A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch and R. Heusdens**, *A Fast Method for High-Resolution Voiced/Unvoiced Detection and Glottal Closure/Opening Instant Estimation of Speech*, IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**, 2 (2016).

2. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens and J. Jensen**, *Relaxed Binaural LCMV Beamforming*, IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**, 1 (2017).

3. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens and J. Jensen**, *A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks with Arbitrary Topology*, IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**, 8 (2018).

4. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens and J. Jensen**, *A Convex Approximation of the Relaxed Binaural Beamforming Optimization Problem*, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2019).

5. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens and J. Jensen**, *Joint Estimation of the Multi-Microphone Signal Model Parameters*, submitted IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Conferences

1. **A. I. Koutrouvelis, R. C. Hendriks, J. Jensen and R. Heusdens**, *Improved Multi-Microphone Noise Reduction Preserving Binaural Cues*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2016).

2. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, J. Jensen and M. Guo**, *Binaural Beamforming Using Pre-Determined Relative Acoustic Transfer Functions*, 25th European Signal Processing Conference (EUSIPCO), (2017).

3. **A. I. Koutrouvelis, J. Jensen, M. Guo, R. C. Hendriks and R. Heusdens**, *Binaural Speech Enhancement with Spatial Cue Preservation Utilising Simultaneous Masking*, 25th European Signal Processing Conference (EUSIPCO), (2017).

4. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, S. van de Paar, J. Jensen and M. Guo**, *Evaluation of Binaural Noise Reduction Methods in Terms of Intelligibility and Perceived Localization*, 26th European Signal Processing Conference (EUSIPCO), (2018).

Symposiums (Posters)

1. **A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, J. Jensen and M. Guo**, *Binaural beamforming without estimating relative acoustic transfer functions*, WIC/IEEE SP Symposium on Information Theory and Signal Processing in the Benelux (2017).

2. **A. I. Koutrouvelis, T. W. Sherson, R. Heusdens and R. C. Hendriks**, *A Novel Low-Complexity Robust Distributed Beamformer*, WIC/IEEE SP Symposium on Information Theory and Signal Processing in the Benelux (2018).

References

[1] J. M. Kates, *Digital hearing aids* (Plural publishing, 2008).

[2] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process. Mag. **32**, 18 (2015).

[3] J. M. Kates, *Feedback cancellation in hearing aids: Results from a computer simulation,* IEEE Trans. Signal Process. **39**, 553 (1991).

[4] A. Spriet, I. Proudler, M. Moonen, and J. Wouters, *Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal,* IEEE Trans. Signal Process. **53**, 3749 (2005).

[5] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, *Feedback control in hearing aids,* in *Springer handbook of speech processing* (Springer, 2008) pp. 979–1000.

[6] M. Guo, S. H. Jensen, and J. Jensen, *Novel acoustic feedback cancellation approaches in hearing aid applications using probe noise and probe noise enhancement,* IEEE Trans. Audio, Speech, Language Process. **20**, 2549 (2012).

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, 2013).

[8] M. Brandstein and D. Ward (Eds.), *Microphone arrays: signal processing techniques and applications* (Springer, 2001).

[9] 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* (Morgan & Claypool, 2013).

[10] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment* (John Wiley & Sons, 2006).

[11] W. M. Hartmann, *How we localize sound,* Physics Today **52**, 24 (1999).

[12] A. W. Bronkhorst, *The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,* Acta Acoustica **86**, 117 (2000).

[13] H. Levitt and L. R. Rabiner, *Binaural release from masking for speech and gain in intelligibility,* J. Acoust. Soc. Amer. **42**, 601 (1967).

[14] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).

[15] O. Roy and M. Vetterli, *Rate-constrained collaborative noise reduction for wireless hearing aids,* IEEE Trans. Signal Process. **57**, 645 (2009).

[16] J. Amini, R. C. Hendriks, R. Heusdens, M. Guo, and J. Jensen, *On the impact of quantization on binaural mvdr beamforming,* in *Speech Communication; 12. ITG Symposium; Proceedings of* (2016) pp. 1–5.

[17] 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. **27**, 154 (2019).

[18] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, *Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids,* IEEE Trans. Audio, Speech, Language Process. **17**, 38 (2009).

[19] 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. **26**, 1261 (2018).

[20] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. IEEE **57**, 1408 (1969).

[21] H. Cox, *Resolving power and sensitivity to mismatch of optimum array processors,* J. Acoust. Soc. Amer. **54**, 771 (1973).

[22] O. L. Frost III, *An algorithm for linearly constrained adaptive array processing,* Proceedings of the IEEE **60**, 926 (1972).

[23] B. D. Van Veen and K. M. Buckley, *Beamforming: A versatile approach to spatial filtering,* IEEE ASSP Mag. **5**, 4 (1988).

[24] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing* (John Wiley & Sons, 2004).

[25] S. A. Vorobyov, *Principles of minimum variance robust adaptive beamforming design,* ELSEVIER Signal Process. **93**, 3264 (2013).

[26] L. Parra and C. Spence, *Convolutive blind separation of non-stationary sources,* IEEE Trans. Audio, Speech, Language Process. **8**, 320 (2000).

[27] S. Gannot, D. Burshtein, and E. Weinstein, *Signal enhancement using beamforming and nonstationarity with applications to speech,* IEEE Trans. Signal Process. , 1614 (2001).

[28] S. Gannot and I. Cohen, *Speech enhancement based on the general transfer function GSC and postfiltering,* IEEE Trans. Speech Audio Process. , 561 (2004).

[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. , 1071 (2009).

[30] S. Gannot, E. Vincet, S. Markovich-Golan, and A. Ozerov, *A consolidated perspective on multi-microphone speech enhancement and source separation,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 692 (2017).

[31] B. Schwartz, S. Gannot, and E. A. P. Habets, *Two model-based em algorithms for blind source separation in noisy environments,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 2209 (2017).

[32] R. C. Hendriks and T. Gerkmann, *Noise correlation matrix estimation for multi-microphone speech enhancement,* IEEE Trans. Audio, Speech, Language Process. **20**, 223 (2012).

[33] J. Jensen and M. S. Pedersen, *Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2015) pp. 5728 – 5732.

[34] H. Cox, *Robust adaptive beamforming,* IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**, 1365 (1987).

[35] B. D. Carlson, *Covariance matrix estimation errors and diagonal loading in adaptive arrays,* **24**, 397 (1988).

[36] J. L. Flanagan, A. C. Surendran, and E. E. Jan, *Spatially selective sound capture for speech and audio processing,* ELSEVIER Speech Commun. **13**, 207 (1993).

[37] I. A. McCowan and H. Bourlard, *Microphone array post-filter based on noise field coherence,* IEEE Trans. Audio, Speech, Language Process. **11**, 709 (2003).

[38] J. Li, P. Stoica, and Z. Wang, *On robust Capon beamforming and diagonal loading,* IEEE Trans. Signal Process. **51**, 1702 (2003).

[39] R. G. Lorenz and S. P. Boyd, *Robust minimum variance beamforming,* IEEE Trans. Signal Process. **53**, 1984 (2005).

[40] S. Doclo and M. Moonen, *Superdirective beamforming robust against microphone mismatch,* IEEE Trans. Audio, Speech, Language Process. **15**, 617 (2007).

[41] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, *An optimum microphone array post-filter for speech applications,* in *ISCA Interspeech* (2006).

[42] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, *On optimal multichannel mean-squared error estimators for speech enhancement,* IEEE Signal Process. Lett. **16**, 885 (2009).

[43] S. Doclo and M. Moonen, *GSVD-based optimal filtering for single and multimicrophone speech enhancement,* IEEE Trans. Signal Process. **50**, 2230 (2002).

[44] J. Blauert, *Spatial hearing: the psychophysics of human sound localization* (MIT press, 1997).

[45] 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. **74**, 1726 (1983).

[46] 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. **23**, 2449 (2015).

[47] D. Marquardt, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg (2015).

[48] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, S. van de Par, J. Jensen, and M. Guo, *Evaluation of binaural noise reduction methods in terms of intelligibility and perceived localization,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2018).

[49] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2005) pp. 29–32.

[50] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[51] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, *Theoretical analysis of binaural multimicrophone noise reduction techniques,* IEEE Trans. Audio, Speech, Language Process. **18**, 342 (2010).

[52] 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. **23** (2015).

[53] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE/ACM Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[54] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[55] D. Marquardt, V. Hohmann, , and S. Doclo, *Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids,* IEEE/ACM Trans. Audio, Speech, Language Process. **23**, 2162 (2015).

[56] J. Thiemann, M. Muller, D. Marquard, S. Doclo, and S. van der Par, *Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene,* EURASIP J. Advances Signal Process. (2016).

[57] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, *Distributed optimization and statistical learning via the alternating direction method of multipliers,* Foundations and Trends® in Machine learning **3**, 1 (2011).

[58] Y. Zeng and R. C. Hendriks, *Distributed delay and sum beamformer for speech enhancement via randomized gossip,* IEEE/ACM Trans. Audio, Speech, Language Process. **22**, 260 (2014).

[59] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms,* IEEE Trans. on Information Theory **52**, 2508 (2006).

[60] 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 Signal Enhancement (IWAENC)* (2012) pp. 1–4.

[61] G. Zhang and R. Heusdens, *Linear coordinate-descent message passing for quadratic optimization,* Neural computation **24**, 3340 (2012).

[62] G. Zhang and R. Heusdens, *Convergence of generalized linear coordinate-descent message-passing for quadratic optimization,* in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on* (2012) pp. 1997–2001.

[63] A. Bertrand and M. Moonen, *Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission,* IEEE Trans. Signal Process. **61**, 3447 (2013).

[64] J. Szurley, A. Bertrand, and M. Moonen, *Topology-independent distributed adaptive node-specific signal estimation in wireless sensor networks,* **3**, 130 (2017).

[65] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[66] 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.

[67] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *A convex approximation of the relaxed binaural beamforming optimization problem,* IEEE/ACM Trans. Audio, Speech, Language Process. (2019).

[68] L. Vandenberghe and S. Boyd, *Semidefinite programming,* SIAM review **38**, 49 (1996).

[69] A. I. Koutrouvelis, J. Jensen, M. Guo, R. C. Hendriks, and R. Heusdens, *Binaural speech enhancement with spatial cue preservation utilising simultaneous masking,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2017) pp. 598–602.

[70] G. Zhang and R. Heusdens, *Distributed optimization using the primal-dual method of multipliers,* **4**, 173 (2018).

[71] K. G. Jöreskog and D. N. Lawley, *New methods in maximum likelihood factor analysis,* British J. Math. Statist. Psycol. **21**, 85 (1968).

[72] K. G. Jöreskog, *A general approach to confirmatory maximum likelihood factor analysis,* **34**, 183 (1969).

[73] K. G. Jöreskog, *Simultaneous factor analysis in several populations,* **36**, 409 (1971).

2

Background

The goal of this chapter is to provide the reader with a sufficient background for the remaining chapters. While Chapter 1 gave a more high-level overview of the multi-microphone noise reduction problem and binaural-cue preservation without mathematical expressions, this chapter is a more formal overview including mathematical details of the signal model, assumptions, and problem formulations. It also reviews the mathematical theory of existing spatial or spatio-temporal filtering methods in both monaural and binaural contexts.

This chapter is organized as follows. In Section 2.1, we introduce the signal acquisition model in the time domain. In Section 2.2, we present the signal model in the frequency domain which is used in this dissertation. We also list a few assumptions accompanying the introduced signal model. In Section 2.3, we review the most popular monaural spatial and spatio-temporal filters and show possible connections between them. We also explain how estimation errors of (relative) acoustic transfer functions (ATFs) can effect the performance of a spatial filter including the most popular solutions to tackle this problem. In addition, we review some popular distributed signal processing techniques which can be applied to spatio-temporal filtering approaches. In Section 2.4, we explain the main concept of the binaural multi-microphone noise reduction problem and review the most well-known binaural spatial and spatio-temporal filtering methods.

2.1. Signal Acquisition

Assume that there is a single point source in the acoustic scene and a microphone array of M microphones as depicted in Figure 2.1. The point source signal is denoted by $s(t)$, and the signal acquired by the i-th microphone is denoted by $y_i(t)$. There are five reasons why the signals $s(t)$ and $y_i(t)$ differ. The first reason is that there is a time delay for the signal $s(t)$ when traveling to the i-th microphone. The second reason is that the source signal captured at the i-th microphone is attenuated when traveling from the source location to the microphone. The third reason is that when

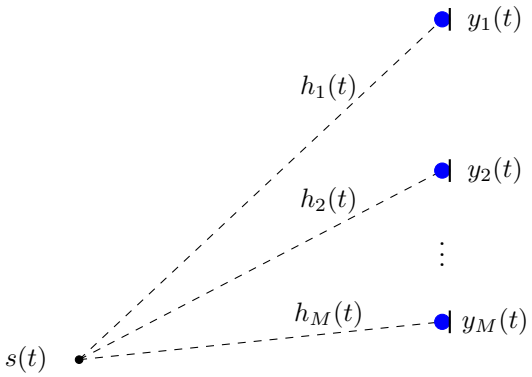

Figure 2.1: Acquisition of a single point source signal $s(t)$ from M microphones.

the point source is within a reverberant enclosure (e.g., a room), the microphone will not only acquire the delayed and attenuated point source signal, but also its reflections coming from the boundaries of the enclosure. The fourth reason is that there may exist obstacles between the source position and the microphones and, thus, the signal will be diffracted around the obstacles. For instance, in the example of HADs one of the obstacles is the head and torso of the user. The fifth and the last reason is that the microphone may not be omnidirectional, but has a specific directivity pattern which attenuates the point source signal differently from different directions.

In general, if we assume that the point source is not moving, we can model all the aforementioned channel factors as a causal linear time-invariant system with impulse response $h_i(t)$. The relationship between the point source signal and the signal acquired by the i-th microphone is given by

$$y_i(t) = h_i(t) * s(t), \quad \text{for } i = 1, \cdots, M, \tag{2.1}$$

where $*$ denotes convolution. The system between every microphone and the point source is different, i.e., $h_i(t) \neq h_j(t), \forall i \neq j$. In practice, sometimes, point sources are moving and, thus, the system is time variant. To overcome this problem, we can assume that for short-time intervals, which are called *time-frames*, the system is time invariant and, thus, we can still use (2.1) within a time-frame. In very reverberant environments $h_i(t)$ is very long (see Figure 2.2) and in most cases longer than a time-frame. The part of the impulse response that is within the time-frame is called *early impulse response* and the component which exceeds the time-frame is called *late impulse response*. In this case, within a time-frame ($t_1 \leq t \leq t_2$), the signal model in (2.1) becomes

$$y_i(t) = h_i(t) * s(t) + l(t), \quad \text{for } t_1 \leq t \leq t_2, \quad i = 1, \cdots, M, \tag{2.2}$$

where $l(t)$ is the accumulated late reverberation component of the signal from all previous time-frames. The late reverberation signal $l(t)$ results in the creation of

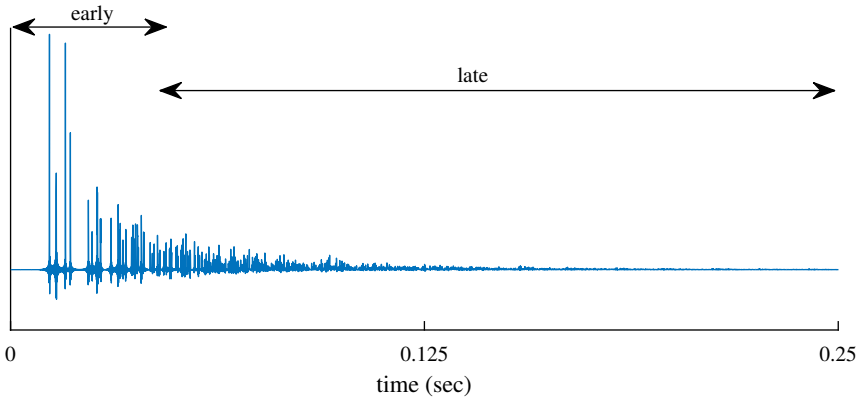

Figure 2.2: Impulse response is split into the early and late impulse response. The reverberation time in this example is $T_{60} = 0.3$ s.

a diffuse noise field [1]. The parameter that mainly characterizes the severeness of the reverberation is the reverberation time, T_{60}, which is the time it takes for the signal power to decrease by 60 dB [2]. Typically, the larger the T_{60}, the larger the portion of reverberation included in the signal. Finally, in practice, microphones have a small portion of self-noise and, thus, the signal model in 2.2 becomes

$$y_i(t) = h_i(t) * s(t) + l(t) + v(t), \quad \text{for } t_1 \leq t \leq t_2, \quad i = 1, \cdots, M, \tag{2.3}$$

where $v(t)$ is the microphone-self noise.

2.2. Multi-Microphone Signal Model in STFT Domain

The recorded microphone signals are in reallity realizations of highly non-stationary random processes. However, typically they are considered short-time stationary within a time-frame. The time-frame length depends on the application. For instance, a speech signal is considered a realization of a short-time stationary process for time-frames of about 30 ms, since the vocal tract does not change shape significantly within this time interval [3]. Apart from being non-stationary, some of the sources in the acoustic scene also change location with respect to the microphone array, i.e., they are highly non-static. However, typically, they are assumed short-time static within a time-frame. The stationarity and static assumptions are very helpful to estimate several signal statistics and parameters. Quite often, before processing, time-frames of microphone recordings are transformed to the frequency domain, and after processing, they are transformed back to the time domain. This popular representation is called the short-time Fourier transform (STFT) [3]. Pro-

cessing in the frequency domain is very convenient in spatial filtering, but also less computationally complex when we assume that different frequency bins of the signals are statistically uncorrelated. In this case, we can process each frequency-bin independently, reducing significantly the computational complexity of the filtering methods. For this reason, in this dissertation, we assume that the frequency-bins of the signals are statistically uncorrelated, and the signal model and spatial filter are computed in the frequency domain.

In the frequency domain, the impulse responses between the point source signals and the microphones are called acoustic transfer functions (ATFs). Sometimes, ATFs are normalized with respect to a reference microphone or location in space. In this case, they are typically referred to as relative ATFs. Moreover, in the case of HADs, where the head is included in the relative ATFs, the relative ATFs are typically called head-related transfer functions (HRTFs).

Consider a wireless network of N devices, where the i-th device has a microphone array of M_i microphones. All devices form a microphone array of $M = \sum_i M_i$ microphones in total. Let $\mathbf{y}_i(t,k)$ be the vector of all signals acquired by the microphones of the i-th device at the t-th time-frame and k-th frequency bin, and let $\mathbf{y}(t,k) = [\mathbf{y}_1^T(t,k), \cdots, \mathbf{y}_N^T(t,k)]^T$ be the microphone recordings of all microphones of all devices. Since the processing is performed independently per time-frequency bin, we neglect the time-frequency indices for notational convenience wherever is possible. In this dissertation, we sometimes consider $\mathbf{y}$ as a random vector and sometimes as a realization of the random vector. However, we will not use a different notation, for random variables or their realizations, but make clear from the context which of the two we are referring to. In the multi-microphone setting, we assume that the target source signal, s, is degraded by additive noise (e.g., sound coming from interfering point sources). The signal model is given by

$$\mathbf{y} = \underbrace{\mathbf{a}s}_{\mathbf{x}} + \mathbf{n}, \tag{2.4}$$

where $\mathbf{a}$ is the (relative) acoustic transfer function (ATF) vector of the target source, s is the target source signal at the original location that we want to estimate, and $\mathbf{n}$ is an additive noise component. This signal model, although very simple, is not very informative in other applications such as source separation, dereverberation and sometimes in binaural beamforming. In source separation, the point sources included in $\mathbf{n}$ need to be estimated. In dereverberation, it is typically convenient to separate the late reverberation component from $\mathbf{n}$. Therefore, we will typically use the following more informative signal model of the noisy signal $\mathbf{y}$:

$$\mathbf{y} = \underbrace{\sum_{i=1}^{r_s} \underbrace{\mathbf{a}_i s_i}_{\mathbf{x}_i}}_{\mathbf{x}} + \underbrace{\sum_{i=1}^{r_u} \mathbf{b}_i u_i + \mathbf{l} + \mathbf{v}}_{\mathbf{n}}, \tag{2.5}$$

where

- $\mathbf{x}_i = \mathbf{a}_i s_i$: is the i-th target point source signal at the microphones.

- s_i: is the i-th target point source signal at the original location.
- $\mathbf{a}_i$: is the early ATF vector of the i-th target point source signal with respect to all microphones.
- u_i: is the i-th interfering point source signal at its origin.
- $\mathbf{b}_i$: is the early ATF vector of the i-th interfering point source signal with respect to all microphones.
- $\mathbf{l}$: is the accumulated late reverberation from all source signals in the acoustic scene (including the target), which is a diffuse noise component.
- $\mathbf{v}$: is the vector of microphones' self-noises which are statistically uncorrelated with each other. Typically, they are very low in power with respect to all the other noise contributions.

The signal model in (2.5), can be also written as

$$\mathbf{y} = \underbrace{\mathbf{A}\mathbf{s}}_{\mathbf{x}} + \underbrace{\mathbf{B}\mathbf{u} + \mathbf{l} + \mathbf{v}}_{\mathbf{n}}, \tag{2.6}$$

where $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_{r_s}]$, $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_{r_u}]$, $\mathbf{s} = [s_1, \cdots, s_{r_s}]^T$, $\mathbf{u} = [u_1, \cdots, u_{r_u}]^T$.

The early ATFs are the early impulse responses in the frequency domain (see Section 2.1) and consist not only of the line-of-sight component, but also of some components due to early reflections. If we assume that all additive components in (2.5) are statistically mutually uncorrelated and have zero mean, the noisy cross-power spectral density matrix, $\mathbf{P}_\mathbf{y} = \mathrm{E}[\mathbf{y}\mathbf{y}^H]$ is given by

$$\mathbf{P}_\mathbf{y} = \underbrace{\sum_{i=1}^{r_s} \underbrace{p_{s_i}\mathbf{a}_i\mathbf{a}_i^H}_{\mathbf{P}_{\mathbf{x}_i}}}_{\mathbf{P}_\mathbf{x}} + \underbrace{\sum_{i=1}^{r_u} p_{u_i}\mathbf{b}_i\mathbf{b}_i^H + \mathbf{P}_\mathbf{l} + \mathbf{P}_\mathbf{v}}_{\mathbf{P}_\mathbf{n}}. \tag{2.7}$$

The model in (2.7) can equivalently be written as

$$\mathbf{P}_\mathbf{y} = \underbrace{\sum_{i=1}^{r_s} \underbrace{\mathbf{A}\mathbf{P}_s\mathbf{A}^H}_{\mathbf{P}_{\mathbf{x}_i}}}_{\mathbf{P}_\mathbf{x}} + \underbrace{\sum_{i=1}^{r_u} \mathbf{B}\mathbf{P}_u\mathbf{B}^H + \mathbf{P}_\mathbf{l} + \mathbf{P}_\mathbf{v}}_{\mathbf{P}_\mathbf{n}}, \tag{2.8}$$

where $\mathbf{P}_s = \mathrm{Diag}\left([p_{s_1}, \cdots, p_{s_{r_s}}]\right)$, and $\mathbf{P}_u = \mathrm{Diag}\left([p_{u_1}, \cdots, p_{u_{r_u}}]\right)$.

Depending on the application, the values of r_s, r_n change in (2.5) and (2.8). For instance, typically in (binaural) multi-microphone noise reduction $r_s = 1$, while in source separation $r_u = 0$, and in speech dereverberation $r_s = 1$ and $r_u = 0$.

2.3. Monaural Multi-microphone noise reduction

In this section, we introduce the mathematical formulation of the multi-microphone noise reduction problem. The multi-microphone noise reduction problem can be viewed as an estimation problem which aims at obtaining an accurate estimate of the target signal s through a two-step estimation procedure. The first step is to obtain the filter, and the second step is to apply the filter to the noisy measurements $\mathbf{y} \in \mathbb{C}^{M\times 1}$. This latter step can be achieved via the following operation:

$$\hat{s} = f(\mathbf{y}), \tag{2.9}$$

where $f(\cdot)$ is the filter function (also called *estimator*) which provides an *estimate*, $\hat{s}$, of the target signal based on the noisy measurement realization $\mathbf{y}$. In this dissertation, for simplicity, we focus only on linear filtering and, thus, (2.9) becomes

$$\hat{s} = \mathbf{w}^H\mathbf{y}, \tag{2.10}$$

where $\mathbf{w} \in \mathbb{C}^{M\times 1}$ is the filter which is linearly applied on $\mathbf{y}$. In the following we assume that s is deterministic while $\mathbf{n}$ is a random variable vector. The *variance* of the estimator in (2.10) is given by [4]

$$\text{var}(\hat{s}) = \text{E}\left[(\hat{s}-E(\hat{s}))^2\right] = \text{E}\left[(\mathbf{w}^H\mathbf{y}-\mathbf{w}^H\text{E}(\mathbf{y}))^2\right] = \text{E}\left[\mathbf{w}^H(\mathbf{y}-\text{E}[\mathbf{y}])(\mathbf{y}-\text{E}[\mathbf{y}])^H\mathbf{w}\right]. \tag{2.11}$$

Assuming that the noise has zero mean, and combining (2.4) and (2.11), the variance of the estimator is given by

$$\text{var}(\hat{s}) = \mathbf{w}^H\mathbf{P}_\mathbf{n}\mathbf{w}. \tag{2.12}$$

Moreover, the *bias* of the estimator is given by [4]

$$b(\hat{s}) = \text{E}(\hat{s}) - s. \tag{2.13}$$

If the bias is zero, we call the estimator *unbiased*. The primary goal of monaural multi-microphone noise reduction used in hearing-aid devices or teleconferencing is to achieve maximum intelligibility improvement. The function of intelligibility is in general complicated and not completely known, although there are many efforts on how to model and predict intelligibility [5–9]. A more convenient and mathematically easier to manipulate function is the output noise power ($\mathbf{w}^H\mathbf{P}_\mathbf{n}\mathbf{w}$) which is the variance of the linear estimator (see (2.12)). It is worth mentioning that eliminating the noise (i.e., making the variance of the linear estimator zero) is not necessarily a good condition to have good intelligibility. For instance, if we select $\mathbf{w} = \mathbf{0}$, the noise will be eliminated, but also the target signal will be eliminated, i.e., $\hat{s} = \mathbf{w}^H\mathbf{y} = 0$. It is evident that there is a trade-off between the bias and variance and we would like a filter which gives the best trade-off in terms of intelligibility. A large bias implies a large speech distortion, while a large variance implies a poor noise reduction.

Spatial filtering (see Section 2.3.1) minimizes the output noise under certain constraints, where one of these constraints intends to leave the target undistorted at the output of the filter. In contrast, spatio-temporal filtering (see Section 2.3.2) allows some target distortions, but can typically achieve more noise reduction.

2.3.1. Spatial Filtering

Two very well-known and widely-used linear spatial filters are the minimum variance distortionless response (MVDR) [10–12], and the linearly constrained minimum variance (LCMV) [12, 13] filters, where the latter is a more general filter than the MVDR. Specifically, the LCMV filter is obtained via the following optimization problem given by [12]

$$\hat{\mathbf{w}}_{\text{LCMV}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{P}}_{\mathbf{n}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \tag{2.14}$$

where $\hat{\mathbf{P}}_{\mathbf{n}}$ is an estimate of the noise CPSDM, $\mathbf{\Lambda} \in \mathbb{C}^{M \times d}$, $\mathbf{f} \in \mathbb{C}^{d \times 1}$, i.e., there are $d \leq M$ linear equality constraints. The objective function is the variance of the linear estimator (see (2.12)), while the linear constraints typically include at least the target distortionless constraint, given by

$$\mathbf{w}^H \hat{\mathbf{a}} = 1, \tag{2.15}$$

where $\hat{\mathbf{a}}$ is an estimate of the ATF vector of the target signal. The goal of the target distortionless constraint in (2.15) is to leave the target undistorted. Theoretically, if $\hat{\mathbf{a}} = \mathbf{a}$, there will be no distortions of the target signal after filtering and the LCMV filter will provide an unbiased estimator (if we assume that the noise $\mathbf{n}$ has zero mean), i.e.,

$$\mathrm{E}\left[\hat{s}\right] = \mathrm{E}\left[\mathbf{w}^H \mathbf{y}\right] = \mathrm{E}\left[\mathbf{w}^H \mathbf{a} s + \mathbf{w}^H \mathbf{n}\right] = s. \tag{2.16}$$

The remaining constraints in $\mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H$ can for instance be used as nulling constraints or additional target distortionless constraints if there are multiple target sources (i.e., $r_s > 1$). The LCMV problem in (2.14) has a closed-form solution given by [12, 13]

$$\hat{\mathbf{w}}_{\text{LCMV}} = \hat{\mathbf{P}}_{\mathbf{n}}^{-1} \mathbf{\Lambda} \left(\mathbf{\Lambda}^H \hat{\mathbf{P}}_{\mathbf{n}}^{-1} \mathbf{\Lambda} \right)^{-1} \mathbf{f}. \tag{2.17}$$

If the constraints in the LCMV problem in (2.14) consist only of the target distortionless constraint in (2.15), we have the MVDR problem which is given by [10–12]

$$\hat{\mathbf{w}}_{\text{MVDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{P}}_{\mathbf{n}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \hat{\mathbf{a}} = 1, \tag{2.18}$$

with a closed-form solution given by

$$\hat{\mathbf{w}}_{\text{MVDR}} = \frac{\hat{\mathbf{P}}_{\mathbf{n}}^{-1} \hat{\mathbf{a}}}{\hat{\mathbf{a}}^H \hat{\mathbf{P}}_{\mathbf{n}}^{-1} \hat{\mathbf{a}}}. \tag{2.19}$$

The output noise power of the MVDR is given by

$$\rho = \hat{\mathbf{w}}_{\text{MVDR}}^H \hat{\mathbf{P}}_{\mathbf{n}} \hat{\mathbf{w}}_{\text{MVDR}} = \left(\hat{\mathbf{a}}^H \hat{\mathbf{P}}_{\mathbf{n}}^{-1} \hat{\mathbf{a}} \right)^{-1}. \tag{2.20}$$

The MVDR uses only a single constraint. Therefore, it has the maximum degrees of freedom for noise reduction. For this reason, quite often in the literature, it is called the maximum SNR beamformer. Statistically, if $\hat{\mathbf{a}} = \mathbf{a}$, the MVDR filter provides

the best linear unbiased estimator (BLUE), because it minimizes the variance of the estimator under the distortionless constraint [4]. If the noise vector $\mathbf{n}$ follows a zero-mean Gaussian distribution, the MVDR filter provides also the minimum variance unbiased (MVU) estimator [4].

A commonly used alternative to the LCMV problem is the linearly constrained minimum power (LCMP) problem given by [12, 13]

$$\hat{\mathbf{w}}_{\text{LCMP}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{P}}_{\mathbf{y}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H. \tag{2.21}$$

The only difference with LCMV is that the LCMP uses in the objective function an estimate of the noisy CPSDM instead of an estimate of the noise CPSDM. It is easy to show that the LCMV and LCMP filters are equivalent under the assumption that $\hat{\mathbf{a}} = \mathbf{a}$, $\hat{\mathbf{P}}_{\mathbf{y}} = \mathbf{P}_{\mathbf{y}}$ and $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{P}_{\mathbf{n}}$ [12]. In practice, we never have perfect estimates of none of these quantities and, thus, the performance of the two spatial filters will not be equivalent [12]. Another popular spatial filter is the minimum power distortionless response (MPDR) [10–12, 14] which is a special case of the LCMP filter and is given by

$$\hat{\mathbf{w}}_{\text{MPDR}} = \frac{\hat{\mathbf{P}}_{\mathbf{y}}^{-1} \hat{\mathbf{a}}}{\hat{\mathbf{a}}^H \hat{\mathbf{P}}_{\mathbf{y}}^{-1} \hat{\mathbf{a}}}. \tag{2.22}$$

Similarly, to the relationship between the LCMP and LCMV, the MPDR is theoretically equivalent with the MVDR [10–12]. However, in practice, when $\hat{\mathbf{a}} \neq \mathbf{a}$ and $\hat{\mathbf{P}}_{\mathbf{y}} \neq \mathbf{P}_{\mathbf{y}}$ and $\hat{\mathbf{P}}_{\mathbf{n}} \neq \mathbf{P}_{\mathbf{n}}$, this is not true [10–12]. In Section 2.3.3, we explain the difference in performance between all these spatial filters in practice.

With spatial filtering, one can also estimate the target signal at a reference location different from the original location. Typically, the reference location is the location of one of the microphones in the microphone array. The microphone that is selected, is referred to as the *reference microphone*. For instance, if the first microphone is selected as the reference microphone, the distortionless constraint becomes $\mathbf{w}^H \mathbf{a} = a_1$.

2.3.2. Spatio-Temporal Filtering

Unlike spatial filters that aim at leaving the target signal undistorted, spatio-temporal filters allow some distortions on the target signal at the output of the filter. By allowing some distortions (bias) on the target signal, spatio-temporal filters typically reduce more the output noise (variance) compared to spatial filters. By assuming that s is deterministic, it is natural to seek the spatio-temporal filter which is the mean square error estimator by solving the following unconstrained optimization problem [4]:

$$\hat{\mathbf{w}}_{\text{MSE}} = \arg\min_{\mathbf{w}} \text{E}\left[\left(s - \mathbf{w}^H \mathbf{y}\right)^2\right], \tag{2.23}$$

which is equivalent to the following optimization problem:

$$\hat{\mathbf{w}}_{\text{MSE}} = \arg\min_{\mathbf{w}} \text{var}(\mathbf{w}^H \mathbf{y}) + b^2 \left(\mathbf{w}^H \mathbf{y}\right), \tag{2.24}$$

where the mean-square error objective function is now written as the sum of the variance of the estimator and the square of the bias. Unfortunately, the bias is a function of the unknown target signal s and, thus, the MSE estimator is unrealizable. A common approach to tackle this problem is to assume that s is a random variable with a certain prior distribution and try to minimize the Bayesian MSE [4]. Since we constrain the estimator to be linear the linear minimum mean-square error (LMMSE) estimator [4] is obtained. The LMMSE estimator is also referred to as multi-channel Wiener filter (MWF) [4, 15, 16]. Assuming that the noise signal is uncorrelated with the target source signal, the MWF is given by [15, 16]

$$\hat{\mathbf{w}}_{\text{MWF}} = \hat{p}_s \left(\hat{\mathbf{P}}_{\mathbf{n}} + \hat{\mathbf{P}}_{\mathbf{x}} \right)^{-1} \hat{\mathbf{a}}. \tag{2.25}$$

The LMMSE estimator in (2.25) becomes equal to the minimum mean-square error (MMSE) estimator if both the target signal and the noise signal are Gaussian distributed [4]. In many cases the Gaussian assumption is not valid. For instance, if the target signal is speech and the time-frame is not sufficiently long. In that case, the distribution is not Gaussian, but rather super-Gaussian [17]. It has been shown [16], that the MWF filter can be decomposed into an MVDR filter and a single channel Wiener filter, i.e.,

$$\hat{\mathbf{w}}_{\text{MWF}} = \frac{\hat{p}_s}{\hat{p}_s + \rho} \hat{\mathbf{w}}_{\text{MVDR}}, \tag{2.26}$$

where ρ is given in (2.20).

An alternative filter to the MWF is the speech distortion weighted MWF (SD-MWF) [18] which introduces a parameter μ which controls the trade-off between noise reduction and speech distortion. Specifically, the SD-MWF filter is given by [18]

$$\hat{\mathbf{w}}_{\text{SD-MWF}} = \hat{p}_s \left(\hat{\mathbf{P}}_{\mathbf{x}} + \mu \hat{\mathbf{P}}_{\mathbf{n}} \right)^{-1} \hat{\mathbf{a}}, \tag{2.27}$$

which is obtained from the following unconstrained optimization problem

$$\hat{\mathbf{w}}_{\text{SD-MWF}} = \arg\min_{\mathbf{w}} \text{E} \left[\left(s - \mathbf{w}^H \mathbf{y} \right)^2 + \mu \left(\mathbf{w}^H \mathbf{n} \right)^2 \right]. \tag{2.28}$$

It has been shown that the SD-MWF can be written as the concatenation of an MVDR and single-channel Wiener filter with speech distortion control [16], i.e.,

$$\hat{\mathbf{w}}_{\text{SD-MWF}} = \frac{\hat{p}_s}{\hat{p}_s + \mu\rho} \hat{\mathbf{w}}_{\text{MVDR}}. \tag{2.29}$$

It is evident from (2.29) that by setting $\mu = 1$, the SD-MWF filter becomes equivalent to the MWF filter. On the other hand, if $\mu = 0$, the MVDR spatial filter will be obtained.

2.3.3. Robustness to Relative Acoustic Transfer Function Estimation Errors

In Section 2.3.1, we reviewed several spatial filters, which all use the target distortionless constraint in (2.15). In this constraint, an estimate of the ATF of the target is used. This estimate is not perfect and, therefore, the spatial filters will cause distortions to the target signal. There are several techniques in the literature that prevent target distortions to a great extent.

As already mentioned in Section 2.3.1, the objective function of the MVDR and LCMV beamformers is the variance of the estimator in (2.12), while for the MPDR and LCMP the objective function is $\mathbf{w}^H\hat{\mathbf{P}}_\mathbf{y}\mathbf{w}$. Assume that in the objective function of the LCMP and MPDR we use a perfect estimate of the noisy CPSDM. In this case, the objective function can be decomposed as

$$\mathbf{w}^H\mathbf{P}_\mathbf{y}\mathbf{w} = \mathbf{w}^H\mathbf{P}_\mathbf{x}\mathbf{w} + \mathbf{w}^H\mathbf{P}_\mathbf{n}\mathbf{w} = p_s\mathbf{w}^H\mathbf{a}\mathbf{a}^H\mathbf{w} + \mathbf{w}^H\mathbf{P}_\mathbf{n}\mathbf{w}. \tag{2.30}$$

In addition, we assume a perfect estimate of the relative ATF of the target in the target distortionless constraint of the problem in (2.21). In this case, the term $p_s\mathbf{w}^H\mathbf{a}\mathbf{a}^H\mathbf{w}$ will be fixed in the objective function and can thus be omitted from the objective function. As a result, the LCMP will be equivalent with the LCMV. However, if we use an inaccurate estimate of the relative ATF of the target in the target distortionless constraint, the term $p_s\mathbf{w}^H\mathbf{a}\mathbf{a}^H\mathbf{w}$ will be supressed in the objective function. The amount of supression depends on the amount of the estimation error introduced in $\hat{\mathbf{a}}$ in the target distortionless constraint [10, 11]. Unlike the LCMP optimization problem, in the LCMV optimization problem the objective function does not include the term $p_s\mathbf{w}^H\mathbf{a}\mathbf{a}^H\mathbf{w}$ and, thus, even for an inaccurate ATF of the target in the distortionless constraint, the target will not be supressed by the objective function. This does not mean that the LCMV filter will remain unbiased (i.e., distortionless), but there will be no intentional supression of the target signal. The LCMV beamformer is therefore more robust to ATF estimation erros than the LCMP [12]. The same conclusions are valid for the MVDR and the MPDR beamformers, i.e., the MVDR is more robust than the MPDR [10–12].

Although the MVDR and LCMV are, in theory, more robust than the MPDR and and LCMP, in practice the challenge is to estimate the noise CPSDM, $\mathbf{P}_\mathbf{n}$. This is typically, accomplished using a voice activity detector (VAD) in the case where the target is a speech signal [19]. The VAD finds time-frequency tiles where the target is absent and there it updates the estimate of the noise CPSDM. In low SNR acoustical environments typically most VADs become unreliable [19]. Another method which estimates the noise CPSDM without the need of a VAD was presented in [20]. However, this method assumes that the target relative ATF is known and has not been tested when the ATF contains estimation errors.

Another group of robust alternatives to the LCMP and MPDR which do not depend on the noise CPSDM and do not require a VAD, are the diagonal loading methods [11, 21, 22]. The main idea of diagonal loading is that, a diagonal loading matrix is added on the noisy CPSDM in the objective function, i.e.,

$$\mathbf{w}^H(\mathbf{P}_\mathbf{y} + \epsilon\mathbf{I})\mathbf{w}. \tag{2.31}$$

By increasing the diagonal loading parameter, ϵ, the CPSDM $\mathbf{P}_{\mathbf{y}} + \epsilon\mathbf{I}$ becomes spatially more uncorrelated and the spatial filter becomes more robust to ATF errors [11, 21]. If $\epsilon \to \infty$, the MPDR beamformer becomes equivalent to the delay and sum (DS) beamformer [23] given by

$$\mathbf{w}_{\text{DS}} = \frac{\mathbf{a}}{\mathbf{a}^H\mathbf{a}}. \tag{2.32}$$

This is the beamformer that is most robust to relative ATF errors, but it does not exploit the second-order statistics of the CPSDM matrix in order to optimally suppress the noise. It is obvious that ϵ should not be huge or very small. In [22] an automatic method was proposed to find ϵ.

2.3.4. Distributed Implementations

We can conclude from Section 2.3 that all filters can be obtained after solving an optimization problem. Here we discuss how we can solve an optimization problem like the LCMV problem in a distributed fashion over multiple devices. Solving an optimization problem in a distributed fashion [24, 25] requires that the optimization problem is in some sense separable into smaller optimization problems, which can be solved at multiple devices. The local solutions should be exchanged in order to find the global solution. This is typically achieved via an iterative procedure in which the devices find a local solution and then exchange the information. This procedure is repeated several times until convergence to the global solution.

The wireless network is formed from N devices and can be viewed as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of devices (where $|\mathcal{V}| = N$), and $\mathcal{E}$ denotes the set of edges between the devices which represent the wireless links between them. Here we assume that the graph is undirected and we allow two-way communication between each neighboring pair of devices. We also denote the set of all neighbours of the i-th device as $\mathcal{N}(i)$.

Assume that we want to solve the following convex optimization problem with linear equality constraints:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} f(\mathbf{w}) \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H, \tag{2.33}$$

where $\mathbf{\Lambda} \in \mathbb{C}^{M\times d}$, and $f(\cdot)$ is a convex function which is separable with respect to the variable $\mathbf{w} \in \mathbb{C}^{M\times 1}$, i.e.,

$$f(\mathbf{w}) = \sum_{i=1}^{N} f_i(\mathbf{w}_i), \tag{2.34}$$

where

$$\mathbf{w} = \left[\mathbf{w}_1^T, \cdots, \mathbf{w}_N^T\right]^T, \quad \mathbf{w}_i \in \mathbb{C}^{M_i\times 1}. \tag{2.35}$$

Hence we have $M = \sum_{i=1}^{N} = M_i$. The linear equality constraints can always be written in a separable form as follows:

$$\sum_{i=1}^{N}\left(\mathbf{w}_i^H\mathbf{\Lambda}_i - \frac{1}{N}\mathbf{f}^H\right) = 0, \tag{2.36}$$

where

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \\ \vdots \\ \mathbf{\Lambda}_N \end{bmatrix}. \tag{2.37}$$

Thus, the convex optimization problem in (2.33) becomes

$$\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_N = \underset{\mathbf{w}_1,\ldots,\mathbf{w}_N}{\arg\min} \sum_{i=1}^{N} f_i(\mathbf{w}_i) \text{ s.t. } \sum_{i=1}^{N} \left(\mathbf{w}_i^H \mathbf{\Lambda}_i - \frac{1}{N} \mathbf{f}^H \right) = 0. \tag{2.38}$$

There are several methods within the literature that solve the problem in (2.38) in a distributed fashion. Below we review two methods: the dual decomposition method [25, 26], and the primal direction method of multipliers (PDMM) [27], respectively.

All methods reviewed below are based on the Lagrangian function and the dual function of the problem in (2.38). The Lagrangian function is given by [28]

$$L(\mathbf{w}, \boldsymbol{\lambda}) = \sum_{i=1}^{N} L_i(\mathbf{w}_i, \boldsymbol{\lambda}) = \sum_{i=1}^{N} f_i(\mathbf{w}_i) + \Re \left\{ \boldsymbol{\lambda}^H \sum_{i=1}^{N} \left(\mathbf{\Lambda}_i^H \mathbf{w}_i - \frac{1}{N} \mathbf{f} \right) \right\}, \tag{2.39}$$

which is separable in $\mathbf{w}$, but not in the dual variable $\boldsymbol{\lambda}$. Since $f(\cdot)$ is a convex function, $L(\mathbf{w}, \boldsymbol{\lambda})$ is also convex on $\mathbf{w}$ if we fix $\boldsymbol{\lambda}$. If we fix $\boldsymbol{\lambda}$, $\mathbf{w}$ is computed as

$$\mathbf{w}(\boldsymbol{\lambda}) = \underset{\mathbf{w}}{\arg\min}\, L(\mathbf{w}, \boldsymbol{\lambda}). \tag{2.40}$$

Note that only when we have the optimal $\hat{\boldsymbol{\lambda}}$ we have $\hat{\mathbf{w}} = \mathbf{w}(\hat{\boldsymbol{\lambda}})$, where $\hat{\mathbf{w}}$ is the optimizer of the problem in (2.33). Since $L(\mathbf{w}, \boldsymbol{\lambda})$ is separable in $\mathbf{w}$, all $\mathbf{w}_i$s can be computed independently as

$$\mathbf{w}_i(\boldsymbol{\lambda}) = \underset{\mathbf{w}_i}{\arg\min}\, L_i(\mathbf{w}_i, \boldsymbol{\lambda}). \tag{2.41}$$

The dual function of the problem in (2.38) is concave and is given by [28]

$$g(\boldsymbol{\lambda}) = L(\mathbf{w}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \sum_{i=1}^{N} L_i(\mathbf{w}_i(\boldsymbol{\lambda}), \boldsymbol{\lambda}). \tag{2.42}$$

Note that the dual function $g(\cdot)$ is not separable in $\boldsymbol{\lambda}$. The optimal $\hat{\boldsymbol{\lambda}}$ is computed as

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\arg\max}\, g(\boldsymbol{\lambda}). \tag{2.43}$$

Finally, since $f(\cdot)$ is convex and the Slater's condition always holds for the problem in (2.38), strong duality holds [28]. Since strong duality holds, the optimal $\hat{\mathbf{w}}_i$s are given by [28]

$$\hat{\mathbf{w}}_i = \mathbf{w}_i(\hat{\boldsymbol{\lambda}}), \text{ for } i = 1, \cdots, N. \tag{2.44}$$

Dual Decomposition Method

The dual decomposition method iteratively computes $\hat{\mathbf{w}}_i$ and $\hat{\boldsymbol{\lambda}}$ until convergence. Specifically, it consists of the following two steps:

$$\hat{\mathbf{w}}_i^{(k+1)} = \mathbf{w}_i(\hat{\boldsymbol{\lambda}}^{(k)}),\ i = 1, \cdots, N, \tag{2.45}$$

$$\hat{\boldsymbol{\lambda}}^{(k+1)} = \hat{\boldsymbol{\lambda}}^{(k)} + \alpha^{(k)} \sum_{i=1}^{N} \left(\boldsymbol{\Lambda}_i^H \hat{\mathbf{w}}_i^{(k+1)} - \frac{1}{N}\mathbf{f} \right), \tag{2.46}$$

where the second step is a gradient ascent step which tries to obtain the optimal $\hat{\boldsymbol{\lambda}}$ which maximizes the dual function $g(\cdot)$. Specifically, it finds the next dual variable $\hat{\boldsymbol{\lambda}}^{(k+1)}$ in the steepest ascent direction of $g(\cdot)$ with a step size of $\alpha^{(k)}$. Note that the second step requires all estimated $\hat{\mathbf{w}}_i^{(k+1)}$. Unlike the first step which is fully distributable, the second step is centralized. That is, all devices need to send their estimated $\hat{\mathbf{w}}_i$ to a fusion center in order to compute the $\hat{\boldsymbol{\lambda}}$ and then the fusion center has to broadcast its estimated dual variable back to the devices until convergence. This is not a robust method when the network in which the devices are connected can change over time. In this case, if the fusion center disconnects from the network, the whole system will break down.

Primal Dual Method Of Multipliers

The main problem of the dual decomposition is that its second step, which tries to solve the dual problem in (2.43), is not distributable. The problem in (2.43) can be equivalently solved via the distributed consensus optimization problem formulation as proposed in [24, 29]. The underlying idea of the equivalent distributed consensus optimization problem is that each device can have a local copy of the dual variable and the neighboring devices should try to keep them equal through pair-wise communication. That is,

$$\hat{\boldsymbol{\lambda}}_1, \cdots, \hat{\boldsymbol{\lambda}}_N = \underset{\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_N}{\arg\min} \ -\sum_{i=1}^{N} L_i(\hat{\mathbf{w}}_i(\boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i) \text{ s.t. } \boldsymbol{\lambda}_i = \boldsymbol{\lambda}_j, \forall (i,j) \in \mathcal{E}. \tag{2.47}$$

We assume that the graph of the network is connected (i.e., there is a path between all devices). Thus, all $\boldsymbol{\lambda}_i$ are constrained to be equal in (2.47), i.e., we will have $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}_i = \hat{\boldsymbol{\lambda}}_j, \forall i, j \in \mathcal{V}$. The problem in (2.47) can be solved iteratively with several distributed methods (e.g., [24, 25, 27]). Here, we will briefly review one recently proposed fast converging method referred to as the primal dual method of multipliers (PDMM) [27]. First the augmented Lagrangian of the problem in (2.47) is computed as [27]

$$\begin{aligned} \Delta(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = &-\sum_{i=1}^{N} L_i(\hat{\mathbf{w}}_i(\boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i) \\ &+ \sum_{\forall (i,j) \in \mathcal{E}} \left(\Re\left\{ \boldsymbol{\gamma}_{ij}^T (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j) \right\} + \frac{\rho}{2} ||\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j||_2^2 \right), \end{aligned} \tag{2.48}$$

where $\boldsymbol{\gamma}_{ij}$ are referred to as edge variables and $\boldsymbol{\gamma}$ is the vector with all edge variables from all devices.

Similar to the dual decomposition method, the PDMM method computes the variables $\boldsymbol{\lambda}_i$s and $\boldsymbol{\gamma}$ iteratively, by first keeping fixed the edge variables, $\boldsymbol{\gamma}$, and updating the $\boldsymbol{\lambda}_i$s via minimization of the Lagrangian in (2.48), followed by the update of the edge variables. However, from a distributed perspective, this process comes with its challenges. First of all, the function in (2.48) is not separable in $\boldsymbol{\lambda}$ because of the coupling of the variables $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ due to the quadratic term $||\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j||_2^2$. PDMM overcomes this coupling by iteratively solving the minimization problem of the function in (2.48) with respect to $\boldsymbol{\lambda}_i$ (at each node independently), by updating $\boldsymbol{\lambda}_i$ at each node based on the previous estimates of their neighboring variables $\boldsymbol{\lambda}_j, \forall j \in \mathcal{N}(i)$. This means that all devices should send their current estimates $\boldsymbol{\lambda}_i$ to their neighbors.

Once these new local estimates are found, PDMM then updates the edge variables in a similar manner to dual descent. Note that each $\boldsymbol{\gamma}_{ij}$ only requires knowledge of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ for this procedure. However, it is not obvious where the optimal $\hat{\boldsymbol{\gamma}}_{ij}$ is computed at each iteration. PDMM addresses this problem by assigning to every edge $(i,j) \in \mathcal{E}$ two *directed* edge variables, $\boldsymbol{\gamma}_{i|j}$ and $\boldsymbol{\gamma}_{j|i}$, which replace $\boldsymbol{\gamma}_{ij}$. The optimal directed edge variable $\hat{\boldsymbol{\gamma}}_{i|j}$ is computed in the i-th device, while $\hat{\boldsymbol{\gamma}}_{j|i}$ is computed in the j-th device. PDMM iterates until convergence. When PDMM converges the following two conditions hold [27]:

$$\hat{\boldsymbol{\gamma}}_{ij} = \hat{\boldsymbol{\gamma}}_{i|j} = \hat{\boldsymbol{\gamma}}_{j|i} \tag{2.49}$$

$$\hat{\boldsymbol{\lambda}}_i = \hat{\boldsymbol{\lambda}}_j, \forall (i,j) \in \mathcal{E} \Leftrightarrow \hat{\boldsymbol{\lambda}}_i = \hat{\boldsymbol{\lambda}}_j, \forall i,j \in \mathcal{V}. \tag{2.50}$$

Altogether, PDMM solves the problem in (2.47) with the following two iterative steps:

$$\hat{\boldsymbol{\lambda}}_i^{(k+1)} = \arg\min_{\boldsymbol{\lambda}_i} \Delta(\boldsymbol{\lambda}_i, \{\hat{\boldsymbol{\lambda}}_j^{(k)} : \forall j \in \mathcal{N}(i)\}, \{\hat{\boldsymbol{\gamma}}_{i|j}^{(k)} : \forall j \in \mathcal{N}(i)\}), \tag{2.51}$$

$$\hat{\boldsymbol{\gamma}}_{i|j}^{(k+1)} = \hat{\boldsymbol{\gamma}}_{j|i}^{(k)} - \rho \frac{i-j}{|i-j|}(\hat{\boldsymbol{\lambda}}_i^{(k+1)} - \hat{\boldsymbol{\lambda}}_j^{(k)}), \ \forall j \in \mathcal{N}(i). \tag{2.52}$$

The first step solves the minimization of the function in (2.48) with respect to $\boldsymbol{\lambda}_i$ using the directed edge variables $\boldsymbol{\gamma}_{j|i}^{(k)}$ and the neighboring variables $\boldsymbol{\lambda}_j^{(k)}$ from the previous iteration. The second step finds the optimal edge variables of the dual function of the problem in (2.47). Note that both steps are dependent on $\hat{\boldsymbol{\gamma}}_{j|i}^{(k)}$. In order to avoid this dependency (which results in extra communication costs), we replace $\hat{\boldsymbol{\gamma}}_{j|i}^{(k)}$ in both steps with the relation

$$\hat{\boldsymbol{\gamma}}_{j|i}^{(k)} = \hat{\boldsymbol{\gamma}}_{i|j}^{(k-1)} - \rho \frac{j-i}{|j-i|}(\hat{\boldsymbol{\lambda}}_j^{(k)} - \hat{\boldsymbol{\lambda}}_i^{(k-1)}). \tag{2.53}$$

Thus, the second step now becomes

$$\hat{\boldsymbol{\gamma}}_{i|j}^{(k+1)} = \hat{\boldsymbol{\gamma}}_{i|j}^{(k-1)} - \rho \frac{j-i}{|j-i|}\left(\hat{\boldsymbol{\lambda}}_j^{(k)} - \hat{\boldsymbol{\lambda}}_i^{(k-1)}\right) - \rho \frac{i-j}{|i-j|}\left(\hat{\boldsymbol{\lambda}}_i^{(k+1)} - \hat{\boldsymbol{\lambda}}_j^{(k)}\right), \ \forall j \in \mathcal{N}(i), \tag{2.54}$$

which is equivalent to the following:

$$\hat{\boldsymbol{\gamma}}_{i|j}^{(k+1)} = \hat{\boldsymbol{\gamma}}_{i|j}^{(k-1)} + \rho \frac{i-j}{|i-j|} \left(2\hat{\boldsymbol{\lambda}}_j^{(k)} - \hat{\boldsymbol{\lambda}}_i^{(k-1)} - \hat{\boldsymbol{\lambda}}_i^{(k+1)} \right). \tag{2.55}$$

Therefore, the two steps of PDMM do not need a fusion center and can be computed independently at each node providing that the neighboring nodes will exchange their optimal $\boldsymbol{\lambda}^{(k)}$ of the k-th iteration in order to compute the new updated variables of the $(k+1)$-th iteration. Therefore, the communication costs per-device and per-iteration is equal to the length of $\boldsymbol{\lambda}^{(k)}$ which is d. Finally, once the i-th node has computed the optimal $\hat{\boldsymbol{\lambda}}_i$, it can also compute the optimal $\hat{\mathbf{w}}_i$ via solving the problem in (2.41), i.e., $\hat{\mathbf{w}}_i = \mathbf{w}(\hat{\boldsymbol{\lambda}}_i)$.

So far, we have discussed how to solve an LCMV problem in a distributed fashion, but assuming that the objective function is separable. However, in the optimization problems of the LCMV, LCMP, MVDR and MPDR beamformers, the objective functions are not separable. In [30] the microphone signals are assumed uncorrelated and, thus, the identity matrix is used instead of the noise CPSDM matrix. Although, this is convenient for distributed implementations, it does not take into account the correlations between the noise microphone signals. This leads to performance degradation with respect to noise reduction.

2.4. Binaural Multi-Microphone Noise Reduction

As with the monaural filtering, also in binaural filtering we mainly focus on linear filtering. In monaural multi-microphone noise reduction there is only one filter applied to the noisy measurements. This provides a single estimate of the target signal either at the original location or at one reference microphone. In contrast, in the binaural setting there are two filters, $\mathbf{w}_L, \mathbf{w}_R \in \mathbb{C}^{M\times 1}$, which are both applied to the noisy measurements producing the left and right hearing-aid output signals [31]. The two output signals are estimates of the target signal at two different reference microphones, each at a different hearing aid. In this dissertation, the first and last element of each vector in (2.5) correspond to the two reference microphones. For notational convenience, we replace the indices of the first and last element of all vectors in (2.5) with L and R (e.g., $y_L = y_1$ and $y_R = y_M$). Using the signal model in (2.4), the binaural output is given by

$$\begin{bmatrix} \hat{x}_L \\ \hat{x}_R \end{bmatrix} = \begin{bmatrix} \mathbf{w}_L^H \mathbf{y} \\ \mathbf{w}_R^H \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_L^H \mathbf{a} s + \mathbf{w}_L^H \mathbf{n} \\ \mathbf{w}_R^H \mathbf{a} s + \mathbf{w}_R^H \mathbf{n} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_L^H \mathbf{a} s + \hat{e}_L \\ \mathbf{w}_R^H \mathbf{a} s + \hat{e}_R \end{bmatrix}, \tag{2.56}$$

where $\hat{e}_L = \mathbf{w}_L^H \mathbf{n}$ and $\hat{e}_R = \mathbf{w}_R^H \mathbf{n}$ are the residual noises after processing at the left and right HAs, respectively. Ideally, $\hat{e}_L$ and $\hat{e}_R$ should not only be minimized, but should also (approximately) provide the same spatial impression as the unprocessed noise n_L and n_R, respectively, to the HA user. As already explained in Chapter 1, the spatial impression is measured with binaural cues (see Section 2.4.1 for more information).

In binaural spatial filtering (see Section 2.4.2 for more information), theoretically, if there are no estimation errors on the target ATFs, the target signal at the two

reference microphones is undistorted, i.e., $\mathbf{w}_L^H\mathbf{a}s = x_L$ and $\mathbf{w}_R^H\mathbf{a}s = x_R$. This is not true in practice as the true ATFs are unknown as we have already explained before. Moreover, the binaural spatio-temporal filters (see Section 2.4.3 for more information) always provide biased estimated target signals at the two reference microphones even if perfect estimates of the noisy and target CPSDMs are available. In practice, where we cannot have perfect estimated ATFs and CPSDMs, one may expect that the spatio-temporal filtering approaches will introduce more distortions to the target signal compared to the spatial filtering approaches.

2.4.1. Binaural Cues

As explained in Chapter 1, there are three main binaural cues: the interaural level difference (ILD), the interaural phase difference (IPD), and the interaural coherence (IC). Every point source has a unique ILD and IPD per time-frequency tile before and after processing. The ILD and IPD are the mangitude and phase of the *interaural transfer function* [32]. The input and output ITF of the target source signal is given by [33]

$$\mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}} = \frac{x_L}{x_R} = \frac{a_L}{a_R}, \quad \mathrm{ITF}_{\mathbf{x}}^{\mathrm{out}} = \frac{\hat{\mathbf{w}}_L^H\mathbf{x}}{\hat{\mathbf{w}}_R^H\mathbf{x}} = \frac{\hat{\mathbf{w}}_L^H\mathbf{a}}{\hat{\mathbf{w}}_R^H\mathbf{a}}, \tag{2.57}$$

where a_L and a_R are the elements of $\mathbf{a}$ coresponding to the left and right reference microphones, respectively. The input and output IPD is the phase of the input and output ITFs, respectively [33]. That is,

$$\mathrm{IPD}_{\mathbf{x}}^{\mathrm{in}} = \angle\mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}}, \quad \mathrm{IPD}_{\mathbf{x}}^{\mathrm{out}} = \angle\mathrm{ITF}_{\mathbf{x}}^{\mathrm{out}}. \tag{2.58}$$

The input and output ILD is the squared magnitude of the input and output ITF, respectively [33]. That is,

$$\mathrm{ILD}_{\mathbf{x}}^{\mathrm{in}} = |\mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}}|^2, \quad \mathrm{ILD}_{\mathbf{x}}^{\mathrm{out}} = |\mathrm{ITF}_{\mathbf{x}}^{\mathrm{out}}|^2. \tag{2.59}$$

Similar expressions exist for all interfering point sources. The three aforementioned binaural cues are more informative about point sources. For diffuse noise fields, the IC binaural cue is more informative [34]. In the signal models in (2.6) and (2.8), the diffuse noise field is due to the late reverberation component. The input and output IC of a diffuse noise field is given by [34, 35]

$$\mathrm{IC}_{\mathbf{l}}^{\mathrm{in}} = \frac{\mathbf{e}_L^T\mathbf{P}_{\mathbf{l}}\mathbf{e}_R}{\sqrt{\mathbf{e}_L^T\mathbf{P}_{\mathbf{l}}\mathbf{e}_L}\sqrt{\mathbf{e}_R^T\mathbf{P}_{\mathbf{l}}\mathbf{e}_R}}, \quad \mathrm{IC}_{\mathbf{l}}^{\mathrm{out}} = \frac{\mathbf{w}_L^H\mathbf{P}_{\mathbf{l}}\mathbf{w}_R}{\sqrt{\mathbf{w}_L^H\mathbf{P}_{\mathbf{l}}\mathbf{w}_L}\sqrt{\mathbf{w}_R^H\mathbf{P}_{\mathbf{l}}\mathbf{w}_R}}. \tag{2.60}$$

If a binaural filter preserves the ITF after processing, it will also preserve the ILD and IPD after processing and vice versa. That is,

$$\mathrm{ITF}_{\mathbf{x}}^{\mathrm{out}} = \mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}} \iff \mathrm{ILD}_{\mathbf{x}}^{\mathrm{out}} = \mathrm{ILD}_{\mathbf{x}}^{\mathrm{in}} \ \cap \ \mathrm{IPD}_{\mathbf{x}}^{\mathrm{out}} = \mathrm{IPD}_{\mathbf{x}}^{\mathrm{in}}. \tag{2.61}$$

Usually, it is difficult for the binaural filters to exactly preserve the ITFs, ILDs, IPDs and ICs after processing. The difference between the input and the output of

a binaural cue is referred to as *binaural-cue error*. Below we provide the binaural-cue errors of the ITF, ILD, IPD and IC binaural cues based on [33, 34, 36, 37]. That is,

$$\mathrm{ITF}^{\mathrm{e}}_{\mathbf{x}} = \left|\mathrm{ITF}^{\mathrm{out}}_{\mathbf{x}} - \mathrm{ITF}^{\mathrm{in}}_{\mathbf{x}}\right| = \left|\frac{\mathbf{w}_L^H\mathbf{x}}{\mathbf{w}_R^H\mathbf{x}} - \frac{x_L}{x_R}\right|, \tag{2.62}$$

$$\mathrm{ILD}^{\mathrm{e}}_{\mathbf{x}} = \left||\mathrm{ITF}^{\mathrm{out}}_{\mathbf{x}}|^2 - |\mathrm{ITF}^{\mathrm{in}}_{\mathbf{x}}|^2\right|, \tag{2.63}$$

$$\mathrm{IPD}^{\mathrm{e}}_{\mathbf{x}} = \frac{\left|\angle\mathrm{ITF}^{\mathrm{out}}_{\mathbf{x}} - \angle\mathrm{ITF}^{\mathrm{in}}_{\mathbf{x}}\right|}{\pi},\ 0 \le \mathrm{IPD}^{\mathrm{e}}_{\mathbf{x}} \le 1, \tag{2.64}$$

$$\mathrm{IC}^{\mathrm{e}}_{\mathbf{1}} = |\mathrm{IC}^{\mathrm{out}}_{\mathbf{1}} - \mathrm{IC}^{\mathrm{in}}_{\mathbf{1}}|^2. \tag{2.65}$$

2.4.2. Binaural Spatial Filtering

Similarly to monaural spatial filtering, the LCMV framework can also be exploited in the binaural spatial filtering context. The binaural LCMV problem is given by

$$\hat{\mathbf{w}}_L, \hat{\mathbf{w}}_R = \underset{\mathbf{w}_L,\mathbf{w}_R}{\arg\min}\ \mathbf{w}_L^H\hat{\mathbf{P}}_{\mathbf{n}}\mathbf{w}_L + \mathbf{w}_R^H\hat{\mathbf{P}}_{\mathbf{n}}\mathbf{w}_R \text{ s.t. } \begin{bmatrix}\mathbf{w}_L^H & \mathbf{w}_R^H\end{bmatrix}\underbrace{\begin{bmatrix}\boldsymbol{\Lambda}_A & \boldsymbol{\Lambda}_B\end{bmatrix}}_{\boldsymbol{\Lambda}} = \underbrace{\begin{bmatrix}\mathbf{f}_A^H & \mathbf{f}_B^H\end{bmatrix}}_{\mathbf{f}^H}, \tag{2.66}$$

where the constraints are split into the two parts A, and B. The first part is dedicated to the target point source signal, while the second part is dedicated to the interfering point source signals. In the following, we discuss three methods that are based on the LCMV framework: the binaural MVDR (BMVDR), the binaural LCMV (BLCMV), and the joint binaural LCMV (JBLCMV). The A part of the constraints is the same for all methods and consists of the two distortionless constraints given by

$$\boldsymbol{\Lambda}_A = \begin{bmatrix}\hat{\mathbf{a}} & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{a}}\end{bmatrix}, \quad \mathbf{f}_A = \begin{bmatrix}\hat{a}_L^*\\ \hat{a}_R^*\end{bmatrix}, \tag{2.67}$$

where the star superscript denotes complex conjugate. If only the A part is used in the constraints, the BMVDR spatial filter is obtained [31] and is given by

$$\mathbf{w}_{\mathrm{BMVDR},L} = \mathbf{w}_{\mathrm{MVDR}}\hat{a}_L^*, \quad \mathbf{w}_{\mathrm{BMVDR},R} = \mathbf{w}_{\mathrm{MVDR}}\hat{a}_R^*. \tag{2.68}$$

Similarly, to the MVDR spatial filter, the BMVDR is the best performing binaural spatial filter in terms of noise reduction, compared to all others filters that are using additional constraints in the B part. The BMVDR preserves the directional binaural-cues of the target signal, but not of the interferers. This can easily be proved by using the input and output ITFs in (2.57). Specifically, the input and output ITFs of the target are the same, i.e.,

$$\frac{\hat{\mathbf{w}}_{\mathrm{BMVDR},L}^H\mathbf{a}}{\hat{\mathbf{w}}_{\mathrm{BMVDR},R}^H\mathbf{a}} = \frac{a_L}{a_R}. \tag{2.69}$$

However, the ITF output of the i-th interferer is the same as the ITF input of the target, i.e.,

$$\frac{\hat{\mathbf{w}}^H_{\text{BMVDR},L}\mathbf{b}_i}{\hat{\mathbf{w}}^H_{\text{BMVDR},R}\mathbf{b}_i} = \frac{a_L}{a_R}, \text{ for } i = 1, \cdots, r_u. \tag{2.70}$$

This means that the interferers will sound as coming from the target direction. The same holds for the diffuse noise component using the same arguments.

The BLCMV uses two constraints to preserve the binaural cues of the i-th interferer given by

$$\mathbf{w}^H_L\mathbf{b}_i = \eta_i\hat{b}_{iL}, \quad \mathbf{w}^H_R\mathbf{b}_i = \eta_i\hat{b}_{iR}. \tag{2.71}$$

The ITF preservation can be proved easily by using the input and output ITFs in (2.57). That is,

$$\frac{\hat{\mathbf{w}}^H_L\mathbf{b}_i}{\hat{\mathbf{w}}^H_R\mathbf{b}_i} = \frac{\eta_i\hat{b}_{iL}}{\eta_i\hat{b}_{iR}} = \frac{\hat{b}_{iL}}{\hat{b}_{iR}}, \text{ for } i = 1, \cdots, r_u. \tag{2.72}$$

The B part of the BLCMV is therefore given by

$$\mathbf{\Lambda}_B = \begin{bmatrix} \hat{\mathbf{b}}_1 & \mathbf{0} & \hat{\mathbf{b}}_2 & \mathbf{0} & \cdots & \hat{\mathbf{b}}_{r_u} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{b}}_1 & \mathbf{0} & \hat{\mathbf{b}}_2 & \cdots & \mathbf{0} & \hat{\mathbf{b}}_{r_u} \end{bmatrix} \in \mathbb{C}^{2M\times 2r_u}, \quad \mathbf{f}_B = \begin{bmatrix} \eta_1\hat{b}^*_{1L} \\ \eta_1\hat{b}^*_{1R} \\ \eta_2\hat{b}^*_{2L} \\ \eta_2\hat{b}^*_{2R} \\ \vdots \\ \eta_{r_u}\hat{b}^*_{r_uL} \\ \eta_{r_u}\hat{b}^*_{r_uR} \end{bmatrix} \in \mathbb{C}^{2r_u\times 1}, \tag{2.73}$$

where $0 \leq \eta_i \leq 1$, for $i = 1, \cdots, r_u$, controls the amount of suppression of the interfering point sources in the constraints. The BLCMV uses thus two linear equality constraints per interferer in order to a) preserve its binaural cues, b) suppress in a controlled way its power. The BLCMV has therefore $2M - 2 - 2r_u$ available degrees of freedom for reducing the objective function. As a result, the BLCMV can preserve the binaural-cues of maximally $r_u = M - 2$ interferers while still having at least 1 degree of freedom available. Binaural hearing-aid systems typically have 4 microphones and the BLCMV can thus preserve the binaural-cues of only $M - 2 = 2$ interferers. In many practical acoustic scenarios there are many more interferers and, thus, the BLCMV cannot preserve the binaural cues of all these interferers. Moreover, by trying to preserve the binaural cues of 2 interferers there will only be one degree of freedom left to suppress all the remaining interferers and diffuse noise that are not supressed by the constraints with the η parameter.

Unlike the BLCMV, the JBLCMV, which was independently proposed in [38] and in [39], uses a single constraint per-interferer, that was initially proposed in [40], to preserve its binaural cues. The constraint that the JBLCMV uses for the i-th interferer is given by

$$\text{ITF}^{\text{out}}_i = \text{ITF}^{\text{in}}_i \iff \frac{\mathbf{w}^H_L\mathbf{b}_i}{\mathbf{w}^H_R\mathbf{b}_i} = \frac{b_{iL}}{b_{iR}}. \tag{2.74}$$

This constraint however does not control the output power of the interferer as in the BLCMV. The B part of the JBLCMV is given by [38, 39]

$$\mathbf{\Lambda}_B = \begin{bmatrix} \mathbf{b}_1 b_{1R} & \mathbf{b}_2 b_{2R} & \cdots & \mathbf{b}_{r_u} b_{r_u R} \\ -\mathbf{b}_1 b_{1L} & -\mathbf{b}_2 b_{2L} & \cdots & -\mathbf{b}_{r_u} b_{r_u L} \end{bmatrix} \in \mathbb{C}^{2M \times r_u}, \quad \mathbf{f}_B = \mathbf{0} \in \mathbb{C}^{r_u \times 1}. \quad (2.75)$$

The JBLCMV has $2M - 2 - r_u$ available degrees of freedom for noise reduction, i.e., r_u more compared to BLCMV. This means that the JBLCMV can preserve the binaural cues of up to $2M - 3$ interferers and still has one degree of freedom left. In the previous example with $M = 4$, the JBLCMV will be able to preserve the binaural cues of up to 5 interferers, i.e., 3 more interferers compared to the BLCMV. Moreover, the extra degrees of freedom of the JBLCMV can be used for extra noise suppression compared to the BLCMV. Even if the BLCMV puts nulls (i.e., $\eta = 0$) to the interferers in the constraints, the overall noise level at the output of the filter will be higher compared to the JBLCMV. This is because the constraints of the BLCMV do not leave the filter free enough to maximally suppress the total noise in the objective function. In words, the more constraints, the less optimal noise reduction will be achieved.

Both the BLCMV and the JBLCMV can preserve the IPD and ILD binaural-cues which are responsible for point-source localization (see Chapter 1). However, these methods have two drawbacks. The first one is that they need estimates of the ATFs of the target and interferers, which is in general a difficult task, especially when the relative locations of the sources with respect to the head change continuously. The second drawback, is the fact that they do not preserve the interaural coherence of the diffuse noise field which leads to reduced naturalness.

A spatial filtering method that handles both aforementioned problems of the BLCMV and the JBLCMV is the BMVDR-η method [41, 42]. The BMVDR-η binaural spatial filter is given by [42]

$$\mathbf{w}_{\text{BMVDR}-\eta,L} = (1-\eta)\mathbf{w}_{\text{BMVDR},L} + \eta \mathbf{e}_L, \quad (2.76)$$
$$\mathbf{w}_{\text{BMVDR}-\eta,R} = (1-\eta)\mathbf{w}_{\text{BMVDR},R} + \eta \mathbf{e}_R, \quad (2.77)$$

where $0 \leq \eta \leq 1$ is a parameter that controls the trade-off between noise reduction and binaural-cue preservation. The larger the η, the better binaural-cue preservation and the worse noise reduction is achieved. The lower the η, the worse binaural-cue preservation and a better noise reduction is achieved. Two extreme cases are obtained for $\eta = 0$ and $\eta = 1$, where the BMVDR and unprocessed noisy scene are obtained, respectively. Note however, that unlike the JBLCMV and the BLCMV that preserve exactly (in theory) the correct locations of the sources while still achieving noise reduction, the BMVDR-η is unable to do so.

The BMVDR-IC method, proposed in [42], preserves only the interaural coherence of a diffuse noise field. Unlike with all the aforementioned methods which are using only equality constraints in (2.66), the BMVDR-IC method uses the A part

in (2.67) and an inequality constraint on the IC error measure in (2.65). That is,

$$\hat{\mathbf{w}}_L, \hat{\mathbf{w}}_R = \underset{\mathbf{w}_L, \mathbf{w}_R}{\arg\min}\ \mathbf{w}_L^H \hat{\mathbf{P}}_{\mathbf{n}} \mathbf{w}_L + \mathbf{w}_R^H \hat{\mathbf{P}}_{\mathbf{n}} \mathbf{w}_R \text{ s.t. } \begin{bmatrix} \mathbf{w}_L^H & \mathbf{w}_R^H \end{bmatrix} \mathbf{\Lambda}_A = \mathbf{f}_A^H, \quad |\mathrm{IC}_{\mathbf{l}}^{\mathrm{out}} - \mathrm{IC}_{\mathbf{l}}^{\mathrm{in}}|^2 \leq \eta, \tag{2.78}$$

where $\mathrm{IC}_{\mathbf{l}}^{\mathrm{out}}$ and $\mathrm{IC}_{\mathbf{l}}^{\mathrm{in}}$ are given in (2.60). This method is not designed to preserve the direction of point sources in the acoustic scene. Note however, that one may easily add to this method the B part of either the BLCMV or the JBLCMV in order to preserve all binaural-cues. The problem in this case will be that too many constraints are present, which will reduce the feasibility set of the optimization problem even further and, thus, reduce even further the noise reduction performance.

2.4.3. Binaural Spatio-Temporal Filtering

Similar to the monaural spatio-temporal filtering, where many filters can be expressed as a concatenation of spatial and a temporal filter, here most binaural spatio-temporal filters can be expressed as the concatenation of two spatial filters with the same post-filter. The reason why the same post-filter is used in both spatial filters, is in order to avoid harming the binaural-cues. Nevertheless, there are several binaural spatio-temporal filters that cannot be written equivalently as a spatial filter and single-channel post-filter [42]. Typically, these filters are computationally more complex because they need to adapt the entire filter vector. In contrast, the spatio-temporal filters which can split into a spatial filter and a temporal post-filter can adapt the single-channel post-filter more often while adapting the spatial filter less often, resulting in significantly lower computational complexity [42].

Similarly to the monaural SD-MWF filter in (2.29), the binaural SD-MWF (BSD-MWF) filter is given by

$$\mathbf{w}_{\text{BSD-MWF},L} = \frac{p_s}{p_s + \mu\rho} \mathbf{w}_{\text{BMVDR},L}, \quad \mathbf{w}_{\text{BSD-MWF},R} = \frac{p_s}{p_s + \mu\rho} \mathbf{w}_{\text{BMVDR},R}, \tag{2.79}$$

and is obtained from the following optimization problem:

$$\hat{\mathbf{w}}_{\text{BSD-MWF},L}, \hat{\mathbf{w}}_{\text{BSD-MWF},R} = \underset{\mathbf{w}_L, \mathbf{w}_R}{\arg\min}\ \mathrm{E}\left[\left\| \begin{bmatrix} x_L - \mathbf{w}_L^H \mathbf{y} \\ x_R - \mathbf{w}_R^H \mathbf{y} \end{bmatrix} \right\|^2 + \mu \left\| \begin{bmatrix} \mathbf{w}_L^H \mathbf{n} \\ \mathbf{w}_R^H \mathbf{n} \end{bmatrix} \right\|^2 \right]. \tag{2.80}$$

The binaural MWF (BMWF) can be obtained as special case of the BSD-MWF for $\mu = 1$, and the BMVDR for $\mu = 0$. The BSD-MWF always preserves for any value of μ the binaural cues of the target, but not of the remaining noise components.

The method proposed in [40] is a spatio-temporal version of the JBLCMV and will be referred to as BSD-MWF-ITF [40]. To the best of our knowldege this is the first method that used the equality constraint in (2.74) but only with one interferer.

The optimization problem of BSD-MWF-ITF is given by [40]

$$\hat{\mathbf{w}}_{\text{BSD-MWF-ITF},L}, \hat{\mathbf{w}}_{\text{BSD-MWF-ITF},R} = \underset{\mathbf{w}_L,\mathbf{w}_R}{\arg\min}\ \mathrm{E}\left\{\left\|\begin{bmatrix} x_L - \mathbf{w}_L^H\mathbf{y} \\ x_R - \mathbf{w}_R^H\mathbf{y}\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\mathbf{w}_L^H\mathbf{n} \\ \mathbf{w}_R^H\mathbf{n}\end{bmatrix}\right\|^2\right\}$$
$$\text{s.t.}\quad \begin{bmatrix}\mathbf{w}_L^H & \mathbf{w}_R^H\end{bmatrix}\mathbf{\Lambda}_B = \mathbf{f}_B^H, \tag{2.81}$$

where $\mathbf{\Lambda}_B$, $\mathbf{f}_B$ are exactly the same as in (2.75). This method becomes identical to the JBLCMV for $\mu = 0$. For all values $\mu > 0$, the BSD-MWF-ITF manages to preserve the binaural-cues of the interferers, but not of the target [40]. This is because the ITF output is not equal to the ITF input for the target [40].

One way to avoid the binaural-cue distortions of the target source is to first find the JBLCMV spatial filter and then apply a post-filter that will be the same for both sides. This of course does not give equivalent performance with the BSD-MWF-ITF. Another alternative is to additionally use the following constraint in the problem in (2.81):

$$\text{ITF}_{\mathbf{x}}^{\text{out}} = \text{ITF}_{\mathbf{x}}^{\text{in}} \iff \frac{\mathbf{w}_L^H\mathbf{a}}{\mathbf{w}_R^H\mathbf{a}} = \frac{a_L}{a_R}. \tag{2.82}$$

The spatio-temporal version of BMVDR-η is the BSD-MWF-η method [33, 43] and its left and right spatial filters are given by [33]

$$\begin{aligned}\mathbf{w}_{\text{BSD-MWF}-\eta,L} &= (1-\eta)\mathbf{w}_{\text{SD-MWF},L} + \eta\mathbf{e}_L, \\ \mathbf{w}_{\text{BSD-MWF}-\eta,R} &= (1-\eta)\mathbf{w}_{\text{SD-MWF},R} + \eta\mathbf{e}_R.\end{aligned} \tag{2.83}$$

With $\mu = 0$ in the filters $\mathbf{w}_{\text{BSD-MWF},L}$, $\mathbf{w}_{\text{BSD-MWF},R}$, the BSD-MWF-$\eta$ will become identical to the BMVDR-η method. It is worth mentioning that the filters in (2.83) are obtained from the following optimziation problem [33]:

$$\hat{\mathbf{w}}_{\text{BSD-MWF}-\eta,L}, \hat{\mathbf{w}}_{\text{BSD-MWF}-\eta,R} = \underset{\mathbf{w}_L,\mathbf{w}_R}{\arg\min}\ \mathrm{E}\left\{\left\|\begin{bmatrix} x_L - \mathbf{w}_L^H\mathbf{y} \\ x_R - \mathbf{w}_R^H\mathbf{y}\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\eta n_L - \mathbf{w}_L^H\mathbf{n} \\ \eta n_R - \mathbf{w}_R^H\mathbf{n}\end{bmatrix}\right\|^2\right\}. \tag{2.84}$$

Unlike all the other spatio-temporal filters introduced so far, the BSD-MWF-η does not only partially preserve the directional binaural-cues, but also the diffuse noise field. However, unlike the already introduced spatio-temporal filters, the directional binaural-cues are distorted for $\eta < 1$. The BSD-MWF-η does not depend on ATFs and, thus, it is very easy to implement with low computational complexity.

The spatio-temporal version of the BMVDR-IC method is the BSD-MWF-IC method proposed in [37]. This method aims to preserve the interaural coherence of the diffuse noise field. The optimization problem for this method is given by

$$\hat{\mathbf{w}}_{\text{BSD-MWF-IC},L}, \hat{\mathbf{w}}_{\text{BSD-MWF-IC},R} = \underset{\mathbf{w}_L,\mathbf{w}_R}{\arg\min}\ \mathrm{E}\left\{\left\|\begin{bmatrix} x_L - \mathbf{w}_L^H\mathbf{y} \\ x_R - \mathbf{w}_R^H\mathbf{y}\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\mathbf{w}_L^H\mathbf{n} \\ \mathbf{w}_R^H\mathbf{n}\end{bmatrix}\right\|^2\right\}$$
$$\text{s.t.}\quad |\text{IC}_{\mathbf{l}}^{\text{out}} - \text{IC}_{\mathbf{l}}^{\text{in}}|^2 \leq \eta. \tag{2.85}$$

Similar to the BMVDR-IC problem in (2.78), this is a non-convex optimization problem due to the non-convex inequality constraint. Therefore, it may have multiple local minima. Initialization plays a crucial role here.

References

[1] S. Gannot, E. Vincet, S. Markovich-Golan, and A. Ozerov, *A consolidated perspective on multi-microphone speech enhancement and source separation,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 692 (2017).

[2] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation* (Springer Science & Business Media, 2010).

[3] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice* (Prentice Hall, Upper Saddle River, NJ, 2002).

[4] S. M. Kay, *Fundamentals of statistical signal processing. Vol 1, Estimation theory* (Englewood Cliffs, NJ: Prentice-Hall PTR, 1993).

[5] N. R. French and J. C. Steinberg, *Factors governing the intelligibility of speech sounds,* J. Acoust. Soc. Amer. **19**, 90 (1947).

[6] K. D. Kryter, *Methods for the calculation and use of the articulation index,* J. Acoust. Soc. Amer. **34**, 16891697 (1962).

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, *An algorithm for intelligibility prediction of time-frequency weighted noisy speech,* IEEE Trans. Audio, Speech, Language Process. **19**, 2125 (2011).

[8] J. Jensen and C. H. Taal, *An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,* IEEE/ACM Trans. Audio, Speech, Language Process. **24**, 2009 (2016).

[9] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, *An instrumental intelligibility metric based on information theory,* IEEE Signal Process. Lett. **25**, 115 (2018).

[10] H. Cox, *Resolving power and sensitivity to mismatch of optimum array processors,* J. Acoust. Soc. Amer. **54**, 771 (1973).

[11] H. Cox, *Robust adaptive beamforming,* IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**, 1365 (1987).

[12] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing* (John Wiley & Sons, 2004).

[13] O. L. Frost III, *An algorithm for linearly constrained adaptive array processing,* Proceedings of the IEEE **60**, 926 (1972).

[14] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. IEEE **57**, 1408 (1969).

[15] B. D. Van Veen and K. M. Buckley, *Beamforming: A versatile approach to spatial filtering,* IEEE ASSP Mag. **5**, 4 (1988).

[16] M. Brandstein and D. Ward (Eds.), *Microphone arrays: signal processing techniques and applications* (Springer, 2001).

[17] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, *On optimal multichannel mean-squared error estimators for speech enhancement,* IEEE Signal Process. Lett. **16**, 885 (2009).

[18] S. Doclo and M. Moonen, *GSVD-based optimal filtering for single and multimicrophone speech enhancement,* IEEE Trans. Signal Process. **50**, 2230 (2002).

[19] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment* (John Wiley & Sons, 2006).

[20] R. C. Hendriks and T. Gerkmann, *Noise correlation matrix estimation for multi-microphone speech enhancement,* IEEE Trans. Audio, Speech, Language Process. **20**, 223 (2012).

[21] B. D. Carlson, *Covariance matrix estimation errors and diagonal loading in adaptive arrays,* **24**, 397 (1988).

[22] J. Li, P. Stoica, and Z. Wang, *On robust Capon beamforming and diagonal loading,* IEEE Trans. Signal Process. **51**, 1702 (2003).

[23] J. L. Flanagan, A. C. Surendran, and E. E. Jan, *Spatially selective sound capture for speech and audio processing,* ELSEVIER Speech Commun. **13**, 207 (1993).

[24] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation:Numerical Methods* (Prentice Hall, 1989).

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, *Distributed optimization and statistical learning via the alternating direction method of multipliers,* Foundations and Trends® in Machine learning **3**, 1 (2011).

[26] H. Everett, *Generalized lagrange multiplier method for solving problems of optimum allocation of resources,* Operations Research **11**, 399 (1963).

[27] G. Zhang and R. Heusdens, *Distributed optimization using the primal-dual method of multipliers,* **4**, 173 (2018).

[28] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2004).

[29] P. A. Forero, A. Cano, and G. B. Giannakis, *Consensus-based distributed support vector machines,* Journal of Machine Learning Research **11**, 1663 (2010).

[30] Y. Zeng and R. C. Hendriks, *Distributed delay and sum beamformer for speech enhancement via randomized gossip,* IEEE/ACM Trans. Audio, Speech, Language Process. **22**, 260 (2014).

[31] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process. Mag. **32**, 18 (2015).

[32] J. Blauert, *Spatial hearing: the psychophysics of human sound localization* (MIT press, 1997).

[33] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, *Theoretical analysis of binaural multimicrophone noise reduction techniques,* IEEE Trans. Audio, Speech, Language Process. **18**, 342 (2010).

[34] D. Marquardt, V. Hohmann, and S. Doclo, *Coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2013) pp. 8648–8652.

[35] 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. **74**, 1726 (1983).

[36] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[37] D. Marquardt, V. Hohmann, , and S. Doclo, *Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids,* IEEE/ACM Trans. Audio, Speech, Language Process. **23**, 2162 (2015).

[38] 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. **23**, 2449 (2015).

[39] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[40] 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. **23** (2015).

[41] D. Marquardt, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg (2015).

[42] 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. **26**, 1261 (2018).

[43] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

3

Relaxed Binaural LCMV Beamforming

This chapter is based on the article published as "Relaxed Binaural LCMV Beamforming", by A.I. Koutrouvelis, R.C. Hendriks, R. Heusdens and J. Jensen in IEEE/ACM Trans. Audio, Speech and Language Processing. vol. 25, no. 1, pp. 133-148, Jan. 2017.

Compared to normal-hearing people, hearing-impaired people generally have more difficulties in understanding a target talker in complex acoustic environments with multiple interfering sources. To reduce noise and improve speech comfort, single-microphone (see e.g. [1] for an overview) or multi-microphone noise reduction methods (see e.g., [2] for an overview) can be used. While the former are mostly effective in reducing listening effort, the latter are also effective in improving speech intelligibility [3]. Examples of multi-microphone noise reduction methods include the multi-channel Wiener filter (MWF) [4, 5], the minimum variance distrortionless response (MVDR) beamformer [6, 7], or, its generalization, the linearly constrained minimum variance (LCMV) beamformer [7, 8].

Traditionally, hearing aids (HAs) have been fitted *bilaterally*, i.e., the user wears a HA on each ear, and the HAs are operating essentially independently of each other. As such, the noise reduction algorithm in each HA estimates the signal of interest using only the recordings of the microphones from that specific HA [9]. Such a setup with an independent multi-microphone algorithm per ear may severely distort the binaural cues since phase and magnitude relations of the sources reaching the two ears are modified [10]. This is harmful for the naturalness of the total sound field as received by the hearing-aid user. Ideally, all sound sources (including the undesired ones) that are present after processing should still sound as if originating from the original direction. This does not only lead to a more natural perception of the acoustic environment, but can also lead to an improved intelligibility of a target talker in certain cases; more specifically, in spatial unmasking experiments [11] it has been shown that a target talker in a noisy background is significantly easier to understand when the noise sources are separated in space from the talker, as compared to the situation where talker and noise sources are co-located.

Binaural HAs are able to wirelessly exchange microphone signals between HAs. This facilitates the use of multi-microphone noise reduction methods which combine all microphone recordings from both HAs, hence allowing the usage of more microphone recordings than with the bilateral noise reduction. As such, the increased number of microphone recordings can potentially lead to better noise suppression and, thus, to a higher speech intelligibility. Moreover, by introducing proper constraints on the beamformer coefficients, binaural cue preservation of the sources can be achieved.

The LCMV method [7, 8] minimizes the output noise power under multiple linear equality constraints. One of these equality constraints is typically used to guarantee that the target source remains undistorted with respect to a certain reference location or microphone. The remaining constraints can be used for additional control on the final filter response. For example, they can be used to steer nulls in the directions of the interferers [7, 12], or to broaden the beam towards the target source in order to avoid steering vector mismatch problems [13, 14]. A special case of the LCMV method is the minimum variance distortionless response (MVDR) beamformer, which only uses the distortionless constraint of the target source [6, 7].

An alternative multi-microphone noise reduction method is the MWF [4, 5] which leads to the minimum mean square error (MMSE) estimate of the target source if the

estimator is constrained to be linear, or, the target source and the noise are assumed to be jointly Gaussian distributed [15]. However, in [16–18], it was demonstrated that speech signals in time and frequency domains tend to be super-Gaussian distributed rather than Gaussian distributed. Thus, the MWF is generally not MMSE optimal. The MWF does not include a distortionless constraint for the target source and, thus, it generally introduces speech distortion in the output [4]. Several generalizations of the MWF have been proposed, among which the speech distortion weighted MWF (SDW-MWF) [5], which introduces a parameter in the minimization procedure to control the trade-off between speech distortion and noise reduction. A well-known property of the MWF is the fact that it can be decomposed into an MVDR beamformer and a single-channel Wiener filter as a post-processor [19].

There are several binaural multi-microphone noise reduction methods known from the literature. These can be devided into two main categories [20]: a) methods based on the linearly constrained minimum variance (LCMV) framework and b) methods based on the multi-channel Wiener filter (MWF).

The binaural version of the SDW-MWF (BSDW-MWF) [21, 22] preserves the binaural cues of the target. However, it was theoretically proven that the binaural cues of the interferers collapse on the binaural cues of the target source [23] (i.e., after processing the binaural cues of the interferers become identical to the binaural cues of the target source). In [22], a variation of the BSDW-MWF (called BSDW-MWF-N) was proposed which tries to partially preserve the binaural cues of the interferers. This method inserts a portion of the unprocessed noisy signal at the reference microphones to the coresponding BSDW-MWF enhanced signals. The larger the portion of the unprocessed noisy signals, the lower the noise reduction, but the better the preservation of binaural cues of the interferers and vice versa. As such, this solution exhibits a trade-off between the preservation of binaural cues and the amount of noise reduction. In [24], a subjective evaluation of BSDW-MWF and BSDW-MWF-N shows that for a moderate input SNR indeed the subjects localized the processed interferer correctly with BSDW-MWF-N and incorrectly with BSDW-MWF. However, for a small input SNR the processed interferer was also localized correctly for BSDW-MWF. This is mainly due to the inaccurate estimates of the cross power spectral density (CPSD) matrix of the target, and due to masking effects when the processed target and processed interferer are represented to the subjects simultaneously [24]. In [25], two other variations of the BSDW-MWF were proposed. The first one is capable of preserving the binaural cues of the target and completely cancel one interferer. The second one is capable of accurately preserving the binaural cues of only one interferer, while distorting the binaural cues of the target.

Similarly to SDW-MWF, the BSDW-MWF can be decomposed into the binaural MVDR (BMVDR) beamformer and a single-channel Wiener filter [25]. The BMVDR can preserve the binaural cues of the target source, but the binaural cues of the interferers collapse to the binaural cues of the target source. In [26, 27], the binaural linearly constrained minimum variance (BLCMV) method was proposed, which achieves simultaneous noise reduction and binaural cue preservation of the target source and multiple interferers. Unlike the BMVDR, the BLCMV uses two

additional linear constraints per interferer to preserve its binaural cues. A fixed interference rejection parameter is used in combination with these constraints to control the amount of noise reduction. The BLCMV is thus capable of controlling the amount of noise reduction using two constraints per interferer. However, in hearing-aid systems with a rather limited number of microphones, the degrees of freedom (DOF) for noise reduction are exhausted quickly when increasing the number of interferers. This makes the BLCMV less suitable for this application.

In [28], a similar method to BLCMV, called optimal BLCMV (OBLCMV), was proposed which is able to achieve simultaneous noise reduction and binaural cue preservation of the target source and only one interferer. Unlike the BLCMV, the OBLCMV uses an optimal interference rejection parameter with respect to the binaural output SNR. In [29, 30] two independent works proposed the same LCMV-based method (we call it joint BLCMV (JBLCMV)) as an alternative to the BLCMV, which preserves the binaural cues of the target source and *more* than twice the number of interferers compared to the BLCMV [29]. Unlike the BLCMV, the JBLCMV requires only one linear constraint per interferer and, as a result, it has more DOF left for noise reduction. The linear constraints for the preservation of the binaural cues of the interferers have the same form as the linear constraint used in [25]. However, unlike the method in [25], the JBLCMV can preserve the binaural cues of a limited number of interferers and does not distort the binaural cues of the target source.

In this paper, we present an iterative, relaxed binaural LCMV beamforming method. Similar to the other binaural LCMV-based approaches, the proposed method strictly preserves the binaural cues of the target source. However, the proposed method is flexible to control the accuracy of binaural cue preservation of the interferers and, therefore, trade-off against additional noise reduction. This is achieved by using inequality constraints instead of the commonly used equality constraints. The task of each inequality constraint is the (approximate) preservation of the binaural cues of a single interferer in a controlled way. The proposed method is flexible to select a different value for the trade-off parameter of each interferer according to importance. The BMVDR and the JBLCMV can be seen as two extreme cases of the proposed method. On one hand, the BMVDR can achieve the best possible overall noise suppression compared to all the other aforementioned binaural LCMV-based methods, but causes full collapse of the binaural cues of the interferers towards the binaural cues of the target source. On the other hand, the JBLCMV can achieve the preservation of the maximum possible number of interferers compared to the other aforementioned binaural LCMV-based methods, but at the expense of less noise suppression. Unlike the JBLCMV and the BMVDR, the proposed method, is flexible to control the amount of noise suppression and binaural cue preservation according to the needs of the user. The relaxations used in the proposed method allow the usage of a substantially larger number of constraints for the approximate preservation of more interferers compared to all the other binaural LCMV-based methods including JBLCMV.

The remainder of this paper is organized as follows. In Section 3.1, the signal model and the notation are presented. In Section 3.2 the key idea of the binaural

beamforming is explained and several existing binaural LCMV-based methods are summarized. In Sections 3.3 and 3.4, a novel non-convex binaural beamforming problem and its iterative convex approximation are presented, respectively. In Section 3.5, the evaluation of the proposed method is provided. Finally, in Section 3.6, we draw some conclusions.

3.1. Signal Model and Notation

Assume for convenience that each of the two HAs consists of $M/2$ microphones, where M is an even number. Thus, the microphone array consists of M microphones in total. 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 k the frequency-bin index. Assume that there is only one target source and there are r interferers. The k-th frequency coefficient of the l-th frame of the j-th microphone noisy signal, $y_j(k,l)$, $j = 1, \cdots, M$, is given by

$$y_j(k,l) = \underbrace{a_j(k,l)s(k,l)}_{x_j(k,l)} + \sum_{i=1}^{r} \underbrace{b_{ij}(k,l)u_i(k,l)}_{n_{ij}(k,l)} + v_j(k,l), \tag{3.1}$$

where

- $s(k,l)$ denotes the target signal at the source location.
- $u_i(k,l)$, is the i-th interfering signal at the source location.
- $a_j(k,l)$ is the acoustic transfer function (ATF) of the target signal with respect to the j-th microphone.
- $b_{ij}(k,l)$ is the ATF of the i-th interfering signal with respect to the j-th microphone.
- $x_j(k,l)$ is the received target signal at the j-th microphone.
- $n_{ij}(k,l)$ is the i-th received interfering signal at the j-th microphone.
- $v_j(k,l)$ is additive noise at the j-th microphone.

Here we use in the signal model the ATFs for notational convinience. However, note that the ATFs can be replaced with relative acoustic transfer functions (RATF)s which can often be identified easier than the ATFs [12, 20].

In the remainder of the paper, the frequency and frame indices are neglected to simplify the notation. Using vector notation, Eq. (3.1) can be written as

$$\mathbf{y} = \mathbf{x} + \sum_{i=1}^{r} \mathbf{n}_i + \mathbf{v}, \tag{3.2}$$

where $\mathbf{y} \in \mathbb{C}^{M\times 1}$, $\mathbf{x} \in \mathbb{C}^{M\times 1}$, $\mathbf{n}_i \in \mathbb{C}^{M\times 1}$ and $\mathbf{v} \in \mathbb{C}^{M\times 1}$ are the stacked vectors of the y_j, x_j, n_{ij}, v_j (for $j = 1, \cdots, M$) components, respectively. Moreover, $\mathbf{x} = \mathbf{a}s$

and $\mathbf{n}_i = \mathbf{b}_i u_i$, where $\mathbf{a} \in \mathbb{C}^{M\times 1}$ and $\mathbf{b}_i \in \mathbb{C}^{M\times 1}$ are the stacked vectors of the a_j and b_{ij} (for $j = 1, \cdots, M$) components, respectively.

Assuming that all sources and the additive noise are mutually uncorrelated, the CPSD matrix of $\mathbf{y}$ is given by

$$\mathbf{P}_\mathbf{y} = E\left[\mathbf{y}\mathbf{y}^H\right] = \mathbf{P}_\mathbf{x} + \underbrace{\sum_{i=1}^{r} \mathbf{P}_{\mathbf{n}_i} + \mathbf{P}_\mathbf{v}}_{\mathbf{P}}, \tag{3.3}$$

3

where

- $\mathbf{P}_\mathbf{x} = E\left[\mathbf{x}\mathbf{x}^H\right] = p_s \mathbf{a}\mathbf{a}^H \in \mathbb{C}^{M\times M}$ is the CPSD matrix of $\mathbf{x}$, with $p_s = E\left[|s|^2\right]$ the power spectral density (PSD) of s.
- $\mathbf{P}_{\mathbf{n}_i} = E\left[\mathbf{n}_i\mathbf{n}_i^H\right] = p_{u_i}\mathbf{b}_i\mathbf{b}_i^H \in \mathbb{C}^{M\times M}$ is the CPSD matrix of $\mathbf{n}_i$, with $p_{u_i} = E\left[|u_i|^2\right]$ the PSD of u_i.
- $\mathbf{P}_\mathbf{v} = E\left[\mathbf{v}\mathbf{v}^H\right] \in \mathbb{C}^{M\times M}$ is the CPSD matrix of $\mathbf{v}$.
- $\mathbf{P}$ is the total CPSD matrix of all disturbances.

3.2. Binaural Beamforming

Binaural multi-microphone noise reduction methods aim at the simultaneous noise reduction and binaural cue preservation of the sources. In order to preserve the binaural cues, two different spatial filters $\hat{\mathbf{w}}_L \in \mathbb{C}^{M\times 1}$ and $\hat{\mathbf{w}}_R \in \mathbb{C}^{M\times 1}$, are applied to the left and right HA, respectively, where constraints can be used to guarantee that certain phase and magnitude relations between the left and right HA outputs are preserved. Note that both spatial filters use all microphone recordings from both HAs.

Without loss of generality, assume that the reference microphone for the left and right HA is indexed as $j = 1$ and $j = M$, respectively. In the sequel, for ease of notation, the reference terms of Eq. (3.1) use the subscripts L and R instead of $j = 1$ and $j = M$, respectively. The two enhanced output signals at the left and right HAs are then given by

$$\hat{x}_L = \hat{\mathbf{w}}_L^H \mathbf{y} \quad \text{and} \quad \hat{x}_R = \hat{\mathbf{w}}_R^H \mathbf{y}. \tag{3.4}$$

In Section 3.2.1, objective measures for the preservation of binaural cues are presented. In Sections 3.2.3—3.2.6, the BMVDR, the BLCMV, the OBLCMV, and the JBLCMV are reviewed, respectively. All reviewed methods are special cases of the general binaural LCMV (GBLCMV) framework, presented in Section 3.2.2. Finally, the basic properties of all reviewed methods are summarized in Section 3.2.7.

3.2.1. Binaural Cues

The extent to which the binaural cues of a specific source are preserved can be expressed using the input and output interaural tranfer function (ITF) [31, 32].

Often the ITF is decomposed into its magnitude, describing the interaural level differences (ILDs) and its phase, describing the interaural phase differences (IPDs). The input and output ITFs of the i-th interferer are defined as [32]

$$\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}} = \frac{n_{iL}}{n_{iR}} = \frac{b_{iL}}{b_{iR}}, \quad \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}} = \frac{\hat{\mathbf{w}}_L^H \mathbf{n}_i}{\hat{\mathbf{w}}_R^H \mathbf{n}_i} = \frac{\hat{\mathbf{w}}_L^H \mathbf{b}_i}{\hat{\mathbf{w}}_R^H \mathbf{b}_i}. \tag{3.5}$$

The input and output ILDs are defined as [32]

$$\mathrm{ILD}_{\mathbf{n}_i}^{\mathrm{in}} = |\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}}|^2, \quad \mathrm{ILD}_{\mathbf{n}_i}^{\mathrm{out}} = |\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}}|^2. \tag{3.6}$$

The input and output IPDs are given by [32]

$$\mathrm{IPD}_{\mathbf{n}_i}^{\mathrm{in}} = \angle \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}}, \quad \mathrm{IPD}_{\mathbf{n}_i}^{\mathrm{out}} = \angle \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}}. \tag{3.7}$$

Note that frequently, the IPDs are converted and measured as time delays [33], i.e., interaural time differences (ITDs). The IPDs and ILDs are the dominant cues for binaural localization for low and high frequencies, respectively [34]. Typically, the IPDs become more important for frequencies below 1 kHz, while ILDs become more important for frequencies above 3 kHz [34]. In [35] it was experimentally shown that for broadband signals, the IPDs are perceptually much more important than the ILDs for localizing a source. More specifically, it was shown that the low frequency IPDs play the most important role perceptually for correct localization. Based on this observation several proposed multi-microphone noise reduction techniques [33, 36] leave the low frequency content of the noisy measurements unprocessed, and process only the higher frequency content. Unfortunately, if a large portion of the power of the noise is concentrated at low frequencies, the noise reduction capabilities are reduced significantly. Therefore, in this paper we aim at the simultaneous preservation of binaural cues of all sources and noise reduction at all frequencies.

A binaural spatial filter, $\hat{\mathbf{w}} = [\hat{\mathbf{w}}_L^T \quad \hat{\mathbf{w}}_R^T]^T$, exactly preserves the binaural cues of the i-th interferer if $\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}} = \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}}$ [32]. Exact preservation of ITFs also implies preservation of ILDs and IPDs [32], i.e., $\mathrm{ILD}_{\mathbf{n}_i}^{\mathrm{in}} = \mathrm{ILD}_{\mathbf{n}_i}^{\mathrm{out}}$ and $\mathrm{IPD}_{\mathbf{n}_i}^{\mathrm{in}} = \mathrm{IPD}_{\mathbf{n}_i}^{\mathrm{out}}$. Non-exact preservation of binaural cues implies that there is some positive ITF error given by

$$\mathcal{E}_{\mathbf{n}_i} = |\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}} - \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}}|. \tag{3.8}$$

Moreover, non-exact presevation of binaural cues implies that there is some ILD and/or IPD errors, given by

$$\mathcal{L}_{\mathbf{n}_i} = |\mathrm{ILD}_{\mathbf{n}_i}^{\mathrm{out}} - \mathrm{ILD}_{\mathbf{n}_i}^{\mathrm{in}}|, \quad \mathcal{T}_{\mathbf{n}_i} = \frac{|\mathrm{IPD}_{\mathbf{n}_i}^{\mathrm{out}} - \mathrm{IPD}_{\mathbf{n}_i}^{\mathrm{in}}|}{\pi}, \tag{3.9}$$

where $0 \leq \mathcal{T}_{\mathbf{n}_i} \leq 1$ [32]. Eqs. (3.5), (3.6), (3.7), (3.8) and (3.9) apply also for the target source $\mathbf{x}$. As it will become obvious in the sequel, for all methods that will be discussed in this paper, the errors in Eqs. (3.8), (3.9) with respect to the target source are always zero.

As explained before, the IPD error is perceptually more important measure for binaural localization than the ILD error for broadband signals (such as speech signals contaminated by broadband noise signals), because the IPDs are perceptually more important than the ILDs for this category of signals. Moreover, the IPD error is perceptually more informative at low frequencies, while the ILD error is perceptually more informative at high frequencies.

3.2.2. General Binaural LCMV Framework

All binaural LCMV-based methods discussed in this section are based on a general binaural LCMV (GBLCMV)[1] framework which is the binaural version of the classical LCMV framework [7, 8]. The GBLCMV minimizes the sum of the left and right output noise powers under multiple linear equality constraints. That is,

$$\hat{\mathbf{w}}_{\text{GBLCMV}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\ \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H, \tag{3.10}$$

where $\hat{\mathbf{w}}_{\text{GBLCMV}} = [\hat{\mathbf{w}}^T_{\text{GBLCMV},L} \quad \hat{\mathbf{w}}^T_{\text{GBLCMV},R}]^T \in \mathbb{C}^{2M\times 1}$, $\mathbf{\Lambda} \in \mathbb{C}^{2M\times d}$ is assumed to be a full column rank matrix (i.e., $\text{rank}(\mathbf{\Lambda}) = d$), $\mathbf{f} \in \mathbb{C}^{d\times 1}$, d is the number of linear equality constraints, and

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \end{bmatrix} \in \mathbb{C}^{2M\times 2M}. \tag{3.11}$$

Similarly to the classical LCMV framework [7, 8], if $d \leq 2M$, and $\mathbf{\Lambda}$ is full column rank, the GBLCMV has a closed-form solution given by

$$\hat{\mathbf{w}}_{\text{GBLCMV}} = \begin{cases} \tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f} & \text{if } d < 2M \\ (\mathbf{\Lambda}^H)^{-1}\mathbf{f} & \text{if } d = 2M. \end{cases} \tag{3.12}$$

In GBLCMV, the total number of DOF devoted to noise reduction is $\text{DOF}_{\text{GBLCMV}} = 2M - d$. Note that in the special case where $d = 2M$, there are no DOF left for *controlled* noise reduction, i.e., $\hat{\mathbf{w}}_{\text{GBLCMV}}$ cannot reduce the objective function of the GBLCMV problem in a controlled way. Finally, if $d > 2M$, the feasible set is $\{\mathbf{w} : \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H\} = \emptyset$ and the GBLCMV problem has no solution. In conclusion, the matrix $\mathbf{\Lambda}$ has to be "tall" (i.e., $d < 2M$), to be able to simultaneously achieve controlled noise reduction and satisfy the constraints of the GBLCMV problem. The maximum number of constraints that the GBLCMV framework can handle, while achieving controlled noise reduction, is $d_{\max} = 2M - 1$, i.e., there should be always left at least one DOF for noise reduction. Generally, the more DOF (i.e., the larger $\text{DOF}_{\text{GBLCMV}}$), the more controlled noise reduction can be achieved.

The set of linear constraints of the GBLCMV framework in Eq. (3.10) can be devided into two parts,

$$\mathbf{w}^H\left[\mathbf{\Lambda}_1 \mid \mathbf{\Lambda}_2\right] = \left[\mathbf{f}_1^H \mid \mathbf{f}_2^H\right]. \tag{3.13}$$

[1] We used the word *general* in order to distinguish it from the BLCMV method [26, 27].

The first part consists of two distortionless constraints $\mathbf{w}_L^H\mathbf{a} = a_L$ and $\mathbf{w}_R^H\mathbf{a} = a_R$ which preserve the target source at the two reference microphones. This can be written compactly as

$$\mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H, \tag{3.14}$$

where

$$\mathbf{\Lambda}_1 = \begin{bmatrix}\mathbf{a} & \mathbf{0}\\ \mathbf{0} & \mathbf{a}\end{bmatrix} \in \mathbb{C}^{2M\times 2}, \quad \mathbf{f}_1 = \begin{bmatrix}a_L^*\\ a_R^*\end{bmatrix} \in \mathbb{C}^{2\times 1}.$$

All binaural methods discussed in this section are special cases of the GBLCMV framework and they share the constraints in Eq. (3.14), while the constraints $\mathbf{w}^H\mathbf{\Lambda}_2 = \mathbf{f}_2^H$ are different.

In the sequel of the paper we use the term m ($m_{\max}$) to indicate the number (maximum number) of interferers that a special case of the GBLCMV framework can preserve, while at the same time achieving controlled noise reduction. Recall that controlled noise reduction means that there is at least one DOF left for noise reduction. Moreover, $m_{\max} \leq r$ which means that some methods may be unable to preserve all simultaneously present interferers of the acoustic scene, because there are not enough available DOF.

3.2.3. BMVDR

The BMVDR beamformer [30] can be formulated using the combination of the following two beamformers

$$\hat{\mathbf{w}}_{\text{BMVDR},L} = \underset{\mathbf{w}_L\in\mathbb{C}^{M\times 1}}{\arg\min}\ \mathbf{w}_L^H\mathbf{P}\mathbf{w}_L \text{ s.t. } \mathbf{w}_L^H\mathbf{a} = a_L, \tag{3.15}$$

$$\hat{\mathbf{w}}_{\text{BMVDR},R} = \underset{\mathbf{w}_R\in\mathbb{C}^{M\times 1}}{\arg\min}\ \mathbf{w}_R^H\mathbf{P}\mathbf{w}_R \text{ s.t. } \mathbf{w}_R^H\mathbf{a} = a_R, \tag{3.16}$$

with closed-form solutions

$$\hat{\mathbf{w}}_{\text{BMVDR},L} = \frac{\mathbf{P}^{-1}\mathbf{a}a_L^*}{\mathbf{a}^H\mathbf{P}^{-1}\mathbf{a}}, \quad \hat{\mathbf{w}}_{\text{BMVDR},R} = \frac{\mathbf{P}^{-1}\mathbf{a}a_R^*}{\mathbf{a}^H\mathbf{P}^{-1}\mathbf{a}}. \tag{3.17}$$

The BMVDR is the simplest special case of the GBLCMV framework in the sense that it has the minimum number of constraints ($d = 2$) given by Eq. (3.14). Specifically, the two optimization problems in Eqs. (3.15) and (3.16) can be reformulated as the following joint optimization problem,

$$\hat{\mathbf{w}}_{\text{BMVDR}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\,\mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H, \tag{3.18}$$

where $\hat{\mathbf{w}}_{\text{BMVDR}} = [\hat{\mathbf{w}}_{\text{BMVDR},L}^T \quad \hat{\mathbf{w}}_{\text{BMVDR},R}^T]^T \in \mathbb{C}^{2M\times 1}$. Since, the BMVDR has the minimum possible number of constraints, the total number of DOF which can be devoted to noise reduction is $\text{DOF}_{\text{BMVDR}} = 2M - 2$.

The BMVDR preserves the binaural cues of the target source, but distorts the binaural cues of all the interferers [30], i.e., $m_{\max} = 0$. More specifically, after processing, the binaural cues of the interferers collapse on the binaural cues of the target source. It can be shown [30] that the binaural cues of the target source are

preserved due to the satisfaction of the two distortionless constraints of the problems in Eqs. (3.15) and (3.16). That is,

$$\mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}} = \mathrm{ITF}_{\mathbf{x}}^{\mathrm{out}} = \frac{a_L}{a_R}. \tag{3.19}$$

Therefore, the ITF error is $\mathcal{E}_{\mathbf{x},\mathrm{BMVDR}} = 0$. Furthermore, it can be shown that the binaural cues of the interferers collapse to the binaural cues of the target source [30]. More specifically, the $\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}}$ is given by

$$\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}} = \frac{b_{iL}}{b_{iR}}, \tag{3.20}$$

while $\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}}$ is given by

$$\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}} = \frac{\hat{\mathbf{w}}_{\mathrm{BMVDR},L}^H \mathbf{b}_i}{\hat{\mathbf{w}}_{\mathrm{BMVDR},R}^H \mathbf{b}_i} = \frac{\frac{\mathbf{a}^H\mathbf{P}^{-1}\mathbf{b}_i a_L}{\mathbf{a}^H\mathbf{P}^{-1}\mathbf{a}}}{\frac{\mathbf{a}^H\mathbf{P}^{-1}\mathbf{b}_i a_R}{\mathbf{a}^H\mathbf{P}^{-1}\mathbf{a}}} = \frac{a_L}{a_R} = \mathrm{ITF}_{\mathbf{x}}^{\mathrm{in}}. \tag{3.21}$$

Thus, after processing, the interferers will have the same ITF as the target source and their ITF error is given by

$$\mathcal{E}_{\mathbf{n}_i,\mathrm{BMVDR}} = \left|\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}} - \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}}\right| = \left|\frac{a_L}{a_R} - \frac{b_{iL}}{b_{iR}}\right|. \tag{3.22}$$

3.2.4. BLCMV

Another special case of the GBLCMV framework is the binaural linearly constrained minimum variance (BLCMV) beamformer [26, 27] which, unlike the BMVDR, uses additional constraints for the preservation of the binaural cues of m interferers. The left and right spatial filters of the BLCMV are given by [26, 27]

$$\begin{aligned} \hat{\mathbf{w}}_{\mathrm{BLCMV},L} = \underset{\mathbf{w}_L \in \mathbb{C}^{M\times 1}}{\arg\min}\ & \mathbf{w}_L^H \mathbf{P} \mathbf{w}_L \\ \text{s.t.}\quad & \mathbf{w}_L^H \mathbf{a} = a_L \\ & \mathbf{w}_L^H \mathbf{b}_1 = \eta_L b_{1L},\ \ldots,\ \mathbf{w}_L^H \mathbf{b}_m = \eta_L b_{mL}, \end{aligned} \tag{3.23}$$

and

$$\begin{aligned} \hat{\mathbf{w}}_{\mathrm{BLCMV},R} = \underset{\mathbf{w}_R \in \mathbb{C}^{M\times 1}}{\arg\min}\ & \mathbf{w}_R^H \mathbf{P} \mathbf{w}_R \\ \text{s.t.}\quad & \mathbf{w}_R^H \mathbf{a} = a_R \\ & \mathbf{w}_R^H \mathbf{b}_1 = \eta_R b_{1R},\ \ldots,\ \mathbf{w}_R^H \mathbf{b}_m = \eta_R b_{mR}, \end{aligned} \tag{3.24}$$

where the constraints $\mathbf{w}_L^H\mathbf{a} = a_L$ and $\mathbf{w}_R^H\mathbf{a} = a_R$ are the two common distortionless constraints used in all special cases in the GBLCMV framework, while the constraints $\mathbf{w}_L^H\mathbf{b}_i = \eta_L b_{iL}$ and $\mathbf{w}_R^H\mathbf{b}_i = \eta_R b_{iR}$, for $i = 1, \ldots, m$, aim at a) preserving the binaural cues and b) supressing the m interferers. The amount of supression is controlled via the interference rejection parameters η_L and η_R which are pre-defined

$(0 \leq \eta_L,\ \eta_R < 1)$ real-valued scalars. Binaural cue preservation is achieved only if $\eta = \eta_L = \eta_R$ [26, 28]. The two problems in Eqs. (3.23) and (3.24) can be compactly formulated as a joint optimization problem. That is,

$$\hat{\mathbf{w}}_{\text{BLCMV}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\ \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H, \tag{3.25}$$

where

$$\mathbf{\Lambda} = \begin{bmatrix}\mathbf{\Lambda}_1 & \mathbf{\Lambda}_2\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{a} & \mathbf{0} & \mathbf{b}_1 & \mathbf{0} & \cdots & \mathbf{b}_m & \mathbf{0}\\ \mathbf{0} & \mathbf{a} & \mathbf{0} & \mathbf{b}_1 & \cdots & \mathbf{0} & \mathbf{b}_m\end{bmatrix}}_{\mathbb{C}^{2M\times(d=2+2m)}},$$

and

$$\begin{aligned}\mathbf{f}^T &= \begin{bmatrix}\mathbf{f}_1^T & \mathbf{f}_2^T\end{bmatrix}\\ &= \underbrace{\begin{bmatrix}a_L^* & a_R^* & \eta_L b_{1L}^* & \eta_R b_{1R}^* & \cdots & \eta_L b_{mL}^* & \eta_R b_{mR}^*\end{bmatrix}}_{\mathbb{C}^{1\times(d=2+2m)}}.\end{aligned}$$

The available DOF for noise reduction are $\text{DOF}_{\text{BLCMV}} = 2M - d = 2M - 2m - 2$. Since $d_{\max} = 2M - 1$ (see Section 3.2.2), BLCMV can simultaneously achieve controlled noise suppression and binaural cue preservation of at most $m_{\max} = M - 2$ interferers.

The ITF errors of the target source and of the m interferers that are included in the constraints are zero, i.e., $\mathcal{E}_{\mathbf{x},\text{BLCMV}} = 0$ and $\mathcal{E}_{\mathbf{n}_i,\text{BLCMV}} = 0$, for $i = 1, \cdots, m \leq r$. However, if some interferers are not included in the constraints, their ITF error will be non-zero, i.e., $\mathcal{E}_{\mathbf{n}_i,\text{BLCMV}} > 0$, for $i = m+1, \cdots, r$.

3.2.5. OBLCMV

The OBLCMV [28] can be seen as a special case of the BLCMV (and, hence, the GBLCMV) since it solves the same optimization problem. However, it preserves the binaural cues of only one interferer (e.g., the k-th interferer) using an optimal complex-valued interference rejection parameter $\hat{\eta} = \hat{\eta}_L = \hat{\eta}_R$ with respect to the binaural output SNR. More specifically, OBLCMV solves the problem in Eq. (3.25) where $\mathbf{\Lambda}$ and $\mathbf{f}^T$, are given by [28]

$$\mathbf{\Lambda} = \begin{bmatrix}\mathbf{\Lambda}_1 & \mathbf{\Lambda}_2\end{bmatrix} = \begin{bmatrix}\mathbf{a} & \mathbf{0} & \mathbf{b}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{a} & \mathbf{0} & \mathbf{b}_k\end{bmatrix} \in \mathbb{C}^{2M\times 4},$$

$$\mathbf{f}^T = \begin{bmatrix}\mathbf{f}_1^T & \mathbf{f}_2^T\end{bmatrix} = \begin{bmatrix}a_L^* & a_R^* & \hat{\eta} b_{kL}^* & \hat{\eta} b_{kR}^*\end{bmatrix} \in \mathbb{C}^{1\times 4} \tag{3.26}$$

where $1 \leq k \leq r$. The available DOF for noise reduction are $\text{DOF}_{\text{OBLCMV}} = 2M - 4$. The ITF errors of the target source and of the k-th interferer that are included in the constraints are zero, i.e., $\mathcal{E}_{\mathbf{x},\text{OBLCMV}} = 0$ and $\mathcal{E}_{\mathbf{n}_k,\text{OBLCMV}} = 0$. However, the binaural cues of all the other $r-1$ interferers will be distorted, i.e., $\mathcal{E}_{\mathbf{n}_i,\text{BLCMV}} > 0$, for $i \in \{1, \cdots, r\} - \{k\}$.

3.2.6. JBLCMV

Recall from Section 3.2.1 that preserving binaural cues of the i-th interferer implies that the following constraint has to be satisfied

$$\mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{in}} = \mathrm{ITF}_{\mathbf{n}_i}^{\mathrm{out}} \implies \frac{\mathbf{w}_L^H \mathbf{b}_i}{\mathbf{w}_R^H \mathbf{b}_i} = \frac{b_{iL}}{b_{iR}}, \tag{3.27}$$

which can be reformulated as:

$$\mathbf{w}_L^H \mathbf{b}_i b_{iR} - \mathbf{w}_R^H \mathbf{b}_i b_{iL} = 0. \tag{3.28}$$

Compared to (O)BLCMV this unified constraint reduces the number of constraints, used for binaural cue preservation, by a factor 2. As a result, for a given number of interferers, more DOF can be devoted to noise reduction. The JBLCMV [29, 30] uses this type of equality constraints for the preservation of the binaural cues of m interferers. More specifically, the JBLCMV problem is given by

$$\hat{\mathbf{w}}_{\mathrm{JBLCMV}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\ \mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \tag{3.29}$$

where

$$\begin{aligned} \mathbf{\Lambda} &= \left[\mathbf{\Lambda}_1 \mid \mathbf{\Lambda}_2\right] \\ &= \left[\begin{array}{cc|ccc} \mathbf{a} & \mathbf{0} & \mathbf{b}_1 b_{1R} & \cdots & \mathbf{b}_m b_{mR} \\ \mathbf{0} & \mathbf{a} & -\mathbf{b}_1 b_{1L} & \cdots & -\mathbf{b}_m b_{mL} \end{array}\right] \in \mathbb{C}^{2M\times(2+m)}, \end{aligned} \tag{3.30}$$

and $\mathbf{w}_{\mathrm{JBLCMV}} = [\mathbf{w}_{\mathrm{JBLCMV},L}^T \quad \mathbf{w}_{\mathrm{JBLCMV},R}^T]^T$. Moreover,

$$\begin{aligned} \mathbf{f}^T &= \left[\mathbf{f}_1^T \mid \mathbf{f}_2^T\right] \\ &= \left[a_L^* \quad a_R^* \mid 0 \quad 0 \quad \cdots \quad 0\right] \in \mathbb{C}^{1\times(2+m)}. \end{aligned} \tag{3.31}$$

Similarly to all other special cases of the GBLCMV framework, $\mathbf{w}^H \mathbf{\Lambda_1} = \mathbf{f}_1^H$ is used for the exact binaural cue preservation of the target source, while $\mathbf{w}^H \mathbf{\Lambda_2} = \mathbf{f}_2^H$ is used for the preservation of the binaural cues of m interferers.

The JBLCMV can simultaneously achieve controlled noise reduction and binaural cue preservation of up to $m_{\max} = 2M - 3$ interferers [29]. Moreover, the DOF devoted to noise reduction is $\mathrm{DOF}_{\mathrm{JBLCMV}} = 2M - m - 2$.

3.2.7. Summary of GBLCMV methods

We summarize some of the properties of the methods discussed in Section 3.2. Table 3.1 gives an overview of two important factors: a) the maximum number of interferers' binaural cues that can be preserved while achieving controlled noise reduction $m_{\max}$, and b) the degrees of freedom (DOF) available for noise reduction. The following conclusions can be drawn from this table:

- The BMVDR has the maximum DOF, which means that it can achieve the best possible noise reduction. It preserves the binaural cues of the target source, but not the binaural cues of the interferers.

Table 3.1: Summary of a) maximum number of interferers' binaural cues that can be preserved while achieving controlled noise reduction ($m_{\max}$), and b) number of available degrees of freedom for noise reduction (DOF). All methods are special cases of the GBLCMV framework. M is the total number of microphones, and m is the number of the constrained interferers.

Method	$m_{\max}$	DOF
BMVDR [30]	0	$2M-2$
BLCMV [27]	$M-2$	$2M-2m-2$
OBLCMV [28]	1	$2M-4$
JBLCMV [29, 30]	$2M-3$	$2M-m-2$

- Unlike (O)BLCMV which uses two constraints per interferer, JBLCMV uses only one constraint per interferer. Therefore, the JBLCMV can preserve the binaural cues of more interferers, or equivalently, given the same number of interferers it has more available DOF devoted to noise reduction.

In this paper, if the number of simultaneously present interferers is $r > m_{\max}$, the extra interferers $r - m_{\max}$ are *not* included in the constraints in the GBLCMV methods, in order to always have one DOF left for controlled noise reduction.

3.3. Proposed Non-Convex Problem

In this section, we present a general optimization problem of which BMVDR and JBLCMV are special cases. More specifically, we relax the constraints on the binaural cues of the interferers, while keeping the strict equality constraints on the target source (i.e., $\mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H$). The relaxation allows to trade-off the amount of noise reduction and binaural cue preservation per interferer in a controlled way. The proposed optimization problem is defined as

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\, \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H, \underbrace{\left|\frac{\mathbf{w}_L^H\mathbf{b}_i}{\mathbf{w}_R^H\mathbf{b}_i} - \frac{b_{iL}}{b_{iR}}\right|}_{\mathcal{E}_{\mathbf{n}_i}} \leq e_i, \quad i = 1,\cdots,m. \tag{3.32}$$

The inequality constraints bound the ITF error (see Eq. (3.8)), for the interferers $i = 1,\cdots,m$ to be less than a positive trade-off parameter $e_i, i = 1,\cdots,m$. These inequality constraints will be transformed, in the sequel of this section (see Eqs. (3.34), (3.35)), in such a way that they can be viewed as relaxations of the strict equality constraints in Eq. (3.28) used in the JBLCMV method. Note that the proposed method is flexible to choose a different e_i for every interferer according to its importance. For instance, maybe certain locations are more important to be preserved than others and, therefore, a smaller e_i must be used. The trade-off

parameter, e_i, is selected as

$$e_i(c_i) = c_i \mathcal{E}_{\mathbf{n}_i,\text{BMVDR}}, \tag{3.33}$$

where $0 \leq c_i \leq 1$ controls the amount of binaural cue collapse towards the target source, and the amount of noise reduction of the i-th interferer. If $c_i = 1, \forall i$ is used in the optimization problem in Eq. (3.32), then $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\text{BMVDR}}$ which is seen as a worst case, with respect to binaural cue preservation, because there is total collapse of binaural cues of the interferers towards the binaural cues of the target source. If $c_i = 0, \forall i$ we have perfect preservation of binaural cues of the m interferers, and $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\text{JBLCMV}}$. Without any loss of generality, for notational convenience, we assume that the binaural cues of all interferers are of equal importance and, therefore, $c_i = c, \forall i$. Moreover, we keep c fixed over all frequency bins. It is worth noting that other strategies for choosing c may exist, which might lead to a better trade-off between maximum possible noise reduction and perceptual binaural cue preservation. As explained in Section 3.2.1, low frequency content is perceptually more important for binaural cue preservation than high frequency content. Thus, smaller c values for low frequencies and larger c values for higher frequencies may give a better perceptual trade-off.

The problem in Eq. (3.32) is not a convex problem and it is hard to solve. In Section 3.4 we propose a method that approximately solves the non-convex problem in an iterative way by solving at each iteration a convex problem.

3.4. Proposed Iterative Convex Problem

By doing some simple algebraic manipulations, the optimization problem in Eq. (3.32) can equivalently be written as

$$\begin{aligned} \hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\, \mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda}_1 = \mathbf{f}_1^H, \\ \frac{|\mathbf{w}_L^H \mathbf{b}_i b_{iR} - \mathbf{w}_R^H \mathbf{b}_i b_{iL}|}{|\mathbf{w}_R^H \mathbf{b}_i b_{iR}|} \leq e_i(c), \text{ for } i = 1, \cdots, m. \end{aligned} \tag{3.34}$$

Furthermore, the problem in Eq. (3.34) can be re-written as

$$\begin{aligned} \hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\, \mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda}_1 = \mathbf{f}_1^H, \\ |\mathbf{w}^H \mathbf{\Lambda}_{2,i}| \leq \underbrace{|e_i(c) \mathbf{w}_R^H \mathbf{b}_i b_{iR}|}_{f_{2,i}}, \text{ for } i = 1, \cdots, m, \end{aligned} \tag{3.35}$$

where $\mathbf{\Lambda}_{2,i}$ is the i-th column of $\mathbf{\Lambda}_2$ in Eq. (3.30).

We approximately solve the non-convex problem in Eq. (3.35) in an iterative way using $\mathbf{w}_R^H$ of the previous iteration in $f_{2,i}, i = 1, \cdots, m$. The new iterative problem is convex at each iteration and is given by

$$\begin{aligned} \hat{\mathbf{w}}_{(k)} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\, \mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda}_1 = \mathbf{f}_1^H, \\ |\mathbf{w}^H \mathbf{\Lambda}_{2,i}| \leq \underbrace{|e_i(c) \hat{\mathbf{w}}_{R,(k-1)}^H \mathbf{b}_i b_{iR}|}_{f_{2,i,(k)}}, \text{for } i = 1, \cdots, m, \end{aligned} \tag{3.36}$$

where $\hat{\mathbf{w}}_{(k)} = [\hat{\mathbf{w}}_{L,(k)}^T \quad \hat{\mathbf{w}}_{R,(k)}^T]^T$ is the estimated binaural spatial filter of the k-th iteration, which is initialized as $\hat{\mathbf{w}}_{(0)} = \hat{\mathbf{w}}_{\text{BMVDR}}$. Similarly to other existing minimum variance beamformers with inequality constraints [37, 38], the convex optimization problem in Eq. (3.36) can be equivalently written as a second order cone programming (SOCP) problem with equality and inequality constraints (see Appendix) and it can be solved efficiently with interior point methods [39].

The ITF error of the i-th interferer at the k-th iteration is given by

$$\mathcal{E}_{\mathbf{n}_i,(k)} = \left| \frac{\hat{\mathbf{w}}_{L,(k)}^H \mathbf{b}_i}{\hat{\mathbf{w}}_{R,(k)}^H \mathbf{b}_i} - \frac{b_{iL}}{b_{iR}} \right|. \tag{3.37}$$

This iterative method is stopped when all the constraints of the original problem in Eq. (3.32) are satisfied. Therefore, the stopping criterion that we use is given by

$$\mathcal{E}_{\mathbf{n}_i,(k)} \leq e_i(c), \text{ for } i = 1, \cdots, m, \tag{3.38}$$

where $e_i(c)$ is given in Eq. (3.33). Recall that $\mathbf{f}_2 = \mathbf{0}$ (i.e., $f_{2,i} = 0, \forall i$) is used in JBLCMV. Unlike JBLCMV, the proposed method uses $f_{2,i,(k)} \geq 0, \forall i$ and, therefore, the constraints dedicated for the preservation of binaural cues of the interferers are seen as relaxations of the strict equality constraints of the JBLCMV method. These relaxations enlarge the feasible set of the problem, allowing more constraints to be used compared to JBLCMV. The JBLCMV can be seen as a special case of the proposed method for $c = 0$, $f_{2,i,(1)} = 0, i = 1, \cdots m$. In this case, the relaxed constraints in the proposed method become identical to the strict constraints of the JBLCMV. Hence, the JBLCMV needs to run only one iteration of the problem in Eq. (3.36). If $c = 0$, the proposed method follows the same strategy for handling $r > m_{\max}$ simultaneously present interferers as in Section 3.2.7. However, if $c > 0$, then there is a typically large, difficult to predict $m_{\max}$[2], due to the inequality constraints and, therefore, the proposed method uses $m = r, \forall r$ constraints for the preservation of the binaural cues of all simultaneously present interferers. Finally, if $c = 1$, the proposed method does not iterate and stops immediately giving as output the initialization $\hat{\mathbf{w}}_{(0)} = \hat{\mathbf{w}}_{\text{BMVDR}}$.

The termination of the proposed iterative method *may* need a large amount of iterations because of the fixed c in Eq. (3.36). The reason for this is explained in detail in Section 3.4.1. To control the speed of termination we replace in Section 3.4.2 the fixed c in Eq. (3.36) with a decreasing parameter $\tau_{(k)}$ (initialized with $\tau_{(0)} = c$) which controls the speed of termination. In Section 3.4.3 we show under which conditions the proposed method: a) guarantees that it will find a feasible solution satisfying the stopping criterion in Eq. (3.38) in a finite number of iterations, and b) guarantees a bounded amount of binaural cue preservation and a bounded amount of noise reduction. An overview of the proposed method using the adaptive $\tau_{(k)}$ is given in Algorithm 1.

[2]The feasible set of the proposed method typically reduces by adding more inequality constraints. However it is difficult to predict after how many constraints, m, it becomes empty, i.e., what is the value of $m_{\max}$.

3.4.1. Speed of Termination

The proposed iterative method may have slow termination due to the fixed choice of c. In this section we explain the reason and in Section 3.4.2 we explain how to control the speed of termination.

Let $\Phi_{(k)}$ denote the convex feasible set in the k-th iteration of the iterative optimization problem in Eq. (3.36) given by

$$\Phi_{(k)} = \bigcap_{i=1}^{m} \left\{ \mathbf{w}_{(k)} : \mathbf{\Lambda}_1^H \mathbf{w}_{(k)} = \mathbf{f}_1, |\mathbf{w}_{(k)}^H \mathbf{\Lambda}_{2,i}| \leq f_{2,i,(k)} \right\}, \tag{3.39}$$

and $\Psi(c)$ the non-convex feasible set of the original non-convex problem of Eqs. (3.32), (3.33) given by

$$\Psi(c) = \bigcap_{i=1}^{m} \left\{ \mathbf{w} : \mathbf{\Lambda}_1^H \mathbf{w} = \mathbf{f}_1, \left| \frac{\mathbf{w}_L^H \mathbf{b}_i}{\mathbf{w}_R^H \mathbf{b}_i} - \frac{b_{iL}}{b_{iR}} \right| \leq e_i(c) \right\}, \tag{3.40}$$

where $\hat{\mathbf{w}}_{\text{JBLCMV}} \in \Psi(0)$, and $\Psi(0) \subseteq \Psi(c), 0 \leq c \leq 1$ and, therefore, $\hat{\mathbf{w}}_{\text{JBLCMV}} \in \Psi(c), 0 \leq c \leq 1$. In words, $\hat{\mathbf{w}}_{\text{JBLCMV}}$ is an element of the set $\Psi(0)$, which gives the minimum output noise power compare to the other elements of $\Psi(0)$. Note that the $\Phi_{(k)}$ changes for every next iteration, while $\Psi(c)$ is constant over time. We can think of $\Phi_{(k)}$ as a convex approximation set of $\Psi(c)$ at iteration k (see a simplistic example of the two sets in Fig. 3.1(a)).

Note that the proposed iterative method will typically try to find a solution on the boundary of $\Phi_{(k)}$. Some parts of the boundary of $\Phi_{(k)}$ will be inside or on the boundary of $\Psi(c)$, while other parts can be outside the set $\Psi(c)$. Therefore, it is possible that the estimated $\hat{\mathbf{w}}_{(k)}$ will be outside of $\Psi(c)$ (see Fig. 3.1(a) for instance). In this case, obviously, the stopping criterion is not satisfied and, therefore, the problem goes to the next iteration. In the next iteration, $\Phi_{(k+1)}$ changes and a new $\hat{\mathbf{w}}_{(k+1)}$ is estimated which can be again outside of $\Psi(c)$ (see Fig. 3.1(a) for instance). This repetition can happen many times leading to a very slow termination because the new estimate $\hat{\mathbf{w}}_{(k+1)}$ is not selected according to a binaural-cue error descent direction. To avoid this undesirable situation, we propose in Section 3.4.2 to replace the fixed c in Eq. (3.36) with an adaptive reduction parameter $\tau_{(k)}$, in order to make sure that solutions that are on the boundary of $\Phi_{(k)}$ and that are outside $\Psi(c)$ will progressively provide a reduced binaural-cue error, i.e., to move towards the direction of the interior of $\Psi(c)$ (see Fig. 3.1(b) for instance).

3.4.2. Avoiding Slow Termination

The termination of the proposed iterative method *may* need a large amount of iterations because of the fixed c in Eq. (3.36), as explained in Section 3.4.1. Therefore, the replacement of c with an adaptive reduction parameter $\tau_{(k)}$ *only* in Eq. (3.36) is useful for guaranteed termination within a *pre-selected* finite maximum number of iterations, $k_{\max}$. More specifically, the new adaptive reduction parameter that we use in Eq. (3.36) instead of c is given by

$$\tau_{(k)} = \tau_{(k-1)} - \alpha_{(k_{\max})}, \tag{3.41}$$

Algorithm 1: Proposed Iterative Method

Input: c, $k_{\max}, \mathbf{a}, \mathbf{b}_i, i = 1, \cdots, m$
Output: $\hat{\mathbf{w}}_{(k)}$
Initialisation : $\hat{\mathbf{w}}_{(0)} \leftarrow \hat{\mathbf{w}}_{\text{BMVDR}}, k \leftarrow 1, \tau_{(0)} \leftarrow c$
General comments :
{ SC stands for stopping criterion in Eq. (3.38) } .
{ SP stands for solving problem in Eq. (3.36) } .
1. **if** SC$(\hat{\mathbf{w}}_{(0)}, c) =$ **true then**
2. go to 17
3. **end if**
start iterations
4. **while** $k \leq k_{\max}$ **do**
5. **if** $k = k_{\max}$
6. $\hat{\mathbf{w}}_{(k)} \leftarrow \text{SP}\left(\hat{\mathbf{w}}_{(k-1)}, \tau_{(k)}, \mathbf{a}, \mathbf{b}_i, i = 1, \cdots, 2M-3\right)$
7. go to 17
8. **else**
9. $\hat{\mathbf{w}}_{(k)} \leftarrow \text{SP}\left(\hat{\mathbf{w}}_{(k-1)}, \tau_{(k)}, \mathbf{a}, \mathbf{b}_i, i = 1, \cdots, m\right)$
10. **end if**
11. **if** SC$(\hat{\mathbf{w}}_{(k)}, c) =$ **true then**
12. go to 17
13. **end if**
14. $k \leftarrow k + 1$
15. $\tau_{(k)} = \tau_{(k-1)} - c/k_{\max}$
16. **end while**
17. **return** $\hat{\mathbf{w}}_{(k)}$

where $\tau_{(0)} = c$ is selected according to the initial desired amount of collapse of binaural cues in the original non-convex problem in Eqs. (3.32), (3.33). The step $\alpha_{(k_{\max})}$ controls the speed of termination, and is a function of the maximum allowed number of iterations for termination given from the user, i.e.,

$$\alpha_{(k_{\max})} = \frac{c}{k_{\max}}. \tag{3.42}$$

Note that we replace c with $\tau_{(k)}$ only in Eq. (3.36) and not in the stopping criterion in Eq. (3.38). This is because, the stopping criterion is based on the fixed feasible set $\Psi(c)$ of the non-convex problem in Eq. (3.32) which should remain constant over iterations (see an example of two consecutive iterations in Fig. 3.1). Moreover, the $\tau_{(k)}$ is always non-negative, because $\tau_{(k_{\max})} = 0$. Small $k_{\max}$, speeds up the reduction of $\tau_{(k)}$ and, thus, it also speeds up the termination of the proposed method. Of course a very small $k_{\max}$ can lead to a feasible solution, $\hat{\mathbf{w}}_{(k)}$, for which $\sum_i \mathcal{E}_{\mathbf{n}_i,(k)} \ll \sum_i e_i(c)$, i.e., to be far away from the boundary of $\Psi(c)$. This means that $\hat{\mathbf{w}}_{(k)}$ provides better binaural cue preservation than the desired amount of binaural cue preservation, $e_i(c)$. As a result, there will be less noise suppression. Ideally, we would

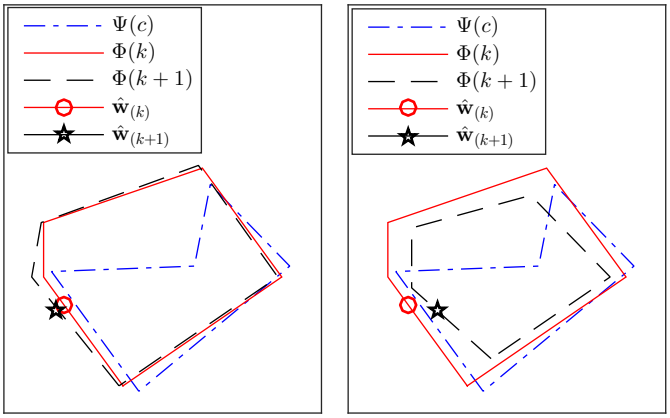

Figure 3.1: Simplistic visualization of two successive iterations (k and $k+1$) of the proposed method with (a) a fixed c, (b) a reducing $\tau(k)$. In $k+1$ iteration the stopping criterion is satisfied in (b). On the contrary, in (a) the stopping criterion is not satisfied, because $\hat{\mathbf{w}}_{(k+1)} \notin \Psi(c)$.

like to arrive as close as possible to the controlled trade-off between noise reduction and binaural cue preservation given by our initial specifications (i.e., amount of collapse). Therefore, a careful choice of $k_{\max}$ is needed in order to find a feasible solution $\hat{\mathbf{w}}_{(k)}$ that:

- achieves a total ITF error $\sum_i \mathcal{E}_{\mathbf{n}_i,(k)} \approx \sum_i e_i(c)$, i.e., to be as close as possible to the boundary of $\Psi(c)$[3].
- to terminate as fast as possible.

Of course there is a trade-off between the two goals.

3.4.3. Guarantees

In this section, we prove that the proposed iterative method using the adaptive reduction parameter in Eq. (3.41) *guarantees* termination, a bounded binaural cue preservation accuracy, and a bounded amount of noise reduction, in *at most* $k_{\max}$ iterations, for a limited number of interferers $m \leq 2M-3$. Nevertheless, our simulation experiments (see Section 3.5.3) show that our algorithm a) is capable of simultaneously achieving the same bounds for binaural cue preservation accuracy and for noise reduction of *more* interferers than $2M-3$ for $c > 0$, and b) finds a feasible solution in much fewer iterations, on average, than $k_{\max}$, for $k_{\max} = 10, 50$.

[3]Note that there may not be any element on the boundary (or in the interior) of $\Psi(c)$, which provides a total ITF error of $\sum_i e_i(c)$. The max possible total ITF error of $\Psi(c)$ may be less than $\sum_i e_i(c)$. This depends mainly on the number of constraints. Nevertheless, in general, the smaller the difference $\sum_i \mathcal{E}_{\mathbf{n}_i,(k)} - \sum_i e_i(c)$ is, the closer to the boundary of $\Psi(c)$ is the solution.

The adaptive decreasing of $\tau_{(k)}$ (see Eq. (3.41)) results in an adaptive shrinking of $\Phi_{(k)}$. Therefore, in the case where the estimated $\hat{\mathbf{w}}_{(k)}$ will be outside of $\Psi(c)$, the stopping criterion is not satisfied and, therefore, the algorithm continues with the next iteration. In the next iteration, $\Phi_{(k)}$ typically shrinks due to the decreased value of $\tau_{(k)}$ according to Eq. (3.41). The algorithm continues until there is a solution $\hat{\mathbf{w}}_{(k)} \in \Psi(c)$. Note that this does not necessarily mean that the algorithm will stop if and only if $\Phi_{(k)} \subseteq \Psi(c)$ (see e.g., Fig. 3.1(b) where the algorithm stops before $\Phi_{(k)} \subseteq \Psi(c)$). Only in the worst case scenario a solution is found when $\Phi_{(k)} \subseteq \Psi(c)$.

We show below that, for $m \leq 2M-3$, the proposed method guarantees termination within a pre-defined finite maximum number of iterations, $k_{\max}$, while achieving a bounded binaural cue preservation accuracy and a bounded amount of noise reduction. This is written more formally in Theorem 1.

Theorem 1. *If $m \leq 2M-3$, the proposed method a) will always find a solution in a finite number of iterations $k \leq k_{max}$ satisfying the stopping criterion of Eq. (3.38), and b) will always have a bounded ITF error, i.e.,*

$$0 \leq \mathcal{E}_{\mathbf{n}_i,(k)} \leq \ e_i(c), \ \textit{for } i = 1, \cdots, m, \tag{3.43}$$

and a bounded noise output power

$$\hat{\mathbf{w}}_{BMVDR}^H \tilde{\mathbf{P}} \hat{\mathbf{w}}_{BMVDR} \leq \hat{\mathbf{w}}_{(k)}^H \tilde{\mathbf{P}} \hat{\mathbf{w}}_{(k)} \leq \hat{\mathbf{w}}_{JBLCMV}^H \tilde{\mathbf{P}} \hat{\mathbf{w}}_{JBLCMV}. \tag{3.44}$$

Proof. Note that for $m \leq 2M-3$, after $k_{\max}$ iterations $\tau_{(k_{\max})} = 0$ (see Eqs. (3.41) and (3.42)) and, therefore, $\hat{\mathbf{w}}_{(k_{\max})} = \hat{\mathbf{w}}_{\text{JBLCMV}}$ because the relaxations of the proposed method in Eq. (3.36) become $\hat{\mathbf{w}}_{(k_{\max})}^H \mathbf{\Lambda}_2 = 0$, which is the same as in JBLCMV as explained in Section 3.4. Note also that $\hat{\mathbf{w}}_{\text{JBLCMV}}$ always satisfies the stopping criterion, i.e., $\hat{\mathbf{w}}_{\text{JBLCMV}} \in \Psi(c)$, for $0 \leq c \leq 1$ (see Section 3.4.1). Therefore, for $m \leq 2M-3$, the algorithm, in the worst case scenario, will terminate after $k_{\max}$ iterations. Consequently, the first part of the theorem has been proved. Thus, in the worst case scenario, the algorithm gives the solution $\hat{\mathbf{w}}_{\text{JBLCMV}}$ which results in $\mathcal{E}_{\mathbf{n}_i,(k)} = 0$ for $i = 1, \cdots, m$. Since the algorithm always terminates (i.e., satisfies the stopping criterion), the ITF error will always be $\mathcal{E}_{\mathbf{n}_i,(k)} \leq \ e_i(c)$, for $i = 1, \cdots, m$. Thus, Eq. (3.43) has been proved. Moreover, the algorithm in the worst case scenario (after $k_{\max}$) will have the noise output power $\hat{\mathbf{w}}_{\text{JBLCMV}}^H \tilde{\mathbf{P}} \hat{\mathbf{w}}_{\text{JBLCMV}}$. Finally, the noise output power cannot be less than $\hat{\mathbf{w}}_{\text{BMVDR}}^H \tilde{\mathbf{P}} \hat{\mathbf{w}}_{\text{BMVDR}}$ (because $\hat{\mathbf{w}}_{\text{BMVDR}}$ achieves the best noise reduction over all the aforementioned methods, because it has the largest feasible set). Thus, Eq. (3.44) has been proved. □

Note that, for $k = k_{\max}$ and $m > 2M-2$, $\Phi_{(k_{\max})} = \emptyset$[4]. However, for $k < k_{\max}$ and $m > 2M-2$, $\Phi_{(k)}$ may not be empty. As we will show in our experiments, indeed, usually it is not empty and, thus, we may achieve simultaneous bounded approximate binaural cue preservation and bounded noise reduction of $m > 2M-2$ interferers. This can be observed experimentally in Sections 3.5.3 and 3.5.3.

[4]Recall that for $m = 2M-2$ (i.e., $d = 2M$), there is a feasible solution which does not provide controlled noise reduction (see Section 3.2.2).

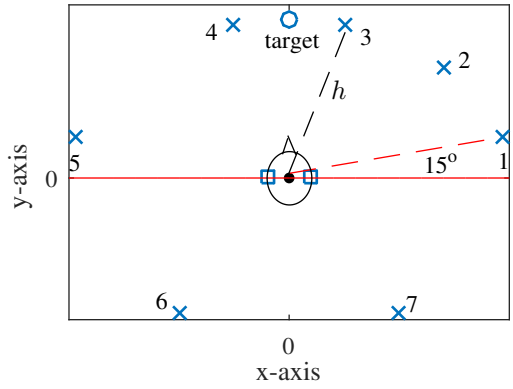

Figure 3.2: Experimental setup: □ HAs, 'o' target source, 'x' speech shaped interferers. Each source has the same distance, h, from the center of the head.

3.5. Experimental Results

In this section, the proposed method, summarized in Algorithm 1, is experimentally evaluated. In Section 3.5.1, the setup of our experiments is demonstrated. In Section 3.5.2, the performance measures are presented. In Section 3.5.3, the proposed method is compared to other LCMV-based methods with regard to binaural cue preservation and noise reduction. Moreover, we provide results with regard to the speed of the proposed method in terms of number of iterations.

3.5.1. Experiment Setup

Fig. 3.2 shows the experimental setup that we used. Two behind-the-ear (BTE) HAs, with two microphones each, are simulated and, therefore, the total number of microphones is $M = 4$. The publicly available database with the BTE impulse responses (IRs) in [40] is used to simulate the head IRs (we used the front and middle microphone for each HA). The front microphones are selected as reference microphones.

We placed all sources on a $h = 80$ cm radius circle centered at the origin $(0, 0)$ (center of head) with an elevation of 0^o degrees. The index of each interferer (denoted by 'x' marker) is indicated in Fig. 3.2. The interferers $1, 2, 3, 4, 5, 6$ and 7 are speech shaped noise realizations with the same power and are placed at 15^o, $45^o, 75^o$, 105^o, 165^o, 240^o and 300^o degrees, respectively. The target source (denoted by 'o' marker) is a speech signal in the look direction, i.e., 90^o degrees.

The duration of all sources is 60 sec. The microphone self noise at each microphone is simulated as white Gaussian noise (WGN) with $\mathbf{P_V} = \sigma^2\mathbf{I}$, where $\sigma = 3.8 * 10^{-5}$ which corresponds to an SNR of 50 dB with respect to the target signal at the left reference microphone. The noise CPSD matrices, $\mathbf{P}$, are calculated (as in Eq. (3.3)) using the ATFs of the truncated true BTE IRs, from the database, and the estimated PSDs of the sources using all available data without voice activity

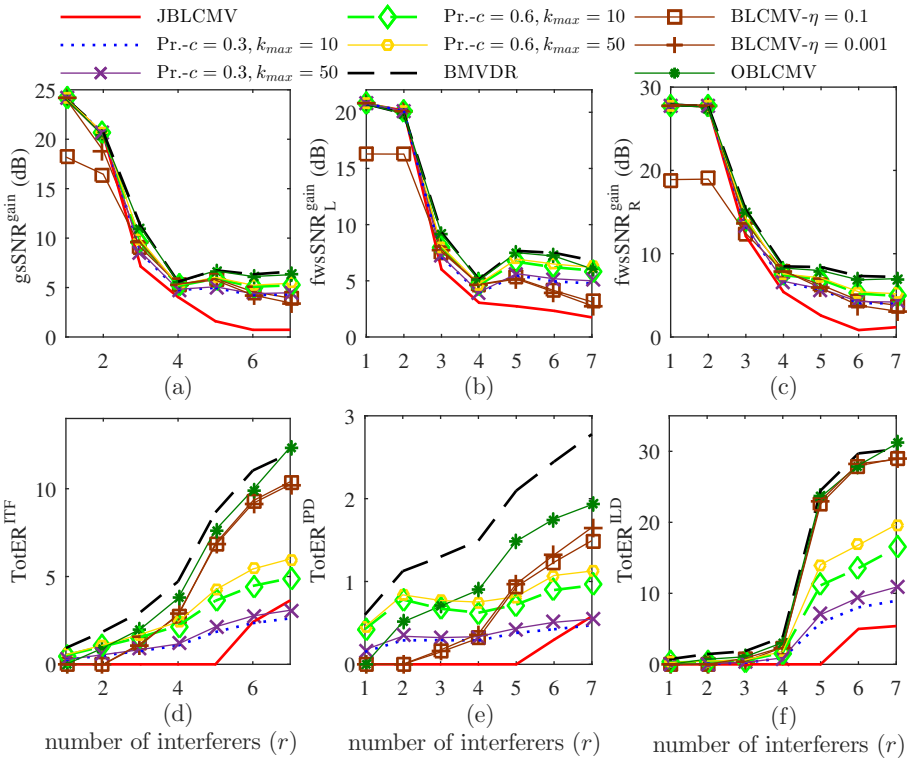

Figure 3.3: Anechoic environment: Performance of the competing methods in terms of (a,b,c) noise reduction, (d) ITF error, (e) IPD error, (f) ILD error.

detection (VAD) errors. Also, the constraints of all the aforementioned methods use the ATFs of the truncated true BTE IRs. The truncated BTE IRs length is 50 ms. The sampling frequency is $f_s = 16$ kHz. We use a simple overlap-and-add analysis/synthesis method [41] with frame length 10 ms, overlap 50% and an FFT size of 1024. The analysis/synthesis window is a square-root-Hann window. The ATFs are also computed with an FFT size of 1024. The microphone signals are computed by convolving the truncated BTE IRs with the source signals at the original locations.

3.5.2. Performance Evaluation

In this section we define the performance evaluation measures that we use to evaluate the results.

ITFs, IPDs & ILDs

Here we define four average performance measures for binaural cue preservation: the total ILD error, the total IPD error, the total ITF error, and the average ITF error ratio. As explained in Section 3.2.1, the IPD errors are perceptually more

important for frequencies below 1 kHz, and the ILD errors are perceptually more important for frequencies above 3 kHz. Thus, the evaluation of IPDs and ILDs will be done only for these frequency regions. We evaluate the total ILD and IPD errors as follows. Let $\mathcal{L}_{\mathbf{n}_i}(k,l)$ and $\mathcal{T}_{\mathbf{n}_i}(k,l)$ denote the ILD and IPD errors (for the k-th frequency bin and l-th frame), respectively, defined in Eq. (3.9). Then the total ILD and IPD errors are defined as

$$\text{TotER}^{\text{ILD}} = \sum_{i=1}^{r} \left(\frac{1}{N - k_{\text{ILD}}} \sum_{k=k_{\text{ILD}}}^{N} \left(\frac{1}{T} \sum_{l=1}^{T} \mathcal{L}_{\mathbf{n}_i}(k,l) \right) \right), \tag{3.45}$$

and

$$\text{TotER}^{\text{IPD}} = \sum_{i=1}^{r} \left(\frac{1}{k_{\text{IPD}}} \sum_{k=1}^{k_{\text{IPD}}} \left(\frac{1}{T} \sum_{l=1}^{T} \mathcal{T}_{\mathbf{n}_i}(k,l) \right) \right), \tag{3.46}$$

where N and T are the number of frequency bins and the number of frames, respectively, k_{ILD} and k_{IPD} are the first and last frequency-bin indices in the frequency regions $3-8$ kHz and $0-1$ kHz, respectively. Note that since the maximum possible value of $\mathcal{T}_{\mathbf{n}_i}(k,l)$ is 1, the maximum value of $\text{TotER}^{\text{IPD}}$ is r. Moreover, we evaluate the total ITF error given by

$$\text{TotER}^{\text{ITF}} = \sum_{i=1}^{r} \left(\frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{T} \sum_{l=1}^{T} \mathcal{E}_{\mathbf{n}_i}(k,l) \right) \right), \tag{3.47}$$

where $\mathcal{E}_{\mathbf{n}_i}$ is the ITF error defined in Eq. (3.8). Finally, we evaluate the average ITF error ratio given by

$$\text{AvER}^{\text{ITF}}(c) = \frac{1}{r} \sum_{i=1}^{r} \frac{1}{N} \sum_{k=1}^{N} \frac{1}{T} \sum_{l=1}^{T} \frac{\mathcal{E}_{\mathbf{n}_i}(k,l)}{\mathcal{E}_{\mathbf{n}_i,\text{BMVDR}}(k,l)}, \tag{3.48}$$

which measures the average amount of binaural cue collapse by comparing the ITF error of the proposed method with the ITF error of the BMVDR. Since the proposed method will always satisfy the condition $\mathcal{E}_{\mathbf{n}_i}(k,l) \leq c\mathcal{E}_{\mathbf{n}_i,\text{BMVDR}}(k,l)$ for $r \leq 2M-3$ (see Theorem 1), obviously $\text{AvER}^{\text{ITF}}(c) \leq c$ for $r \leq 2M-3$. Note that ideally the proposed method will provide a solution as close as possible to the boundary of $\Psi(c)$, i.e., $\text{AvER}^{\text{ITF}}(c) - c$ to be as small as possible (see Section 3.4.2). Moreover, for the proposed method $\text{AvER}^{\text{ITF}}(0) = 0$ and $\text{AvER}^{\text{ITF}}(1) = 1$ because for $c = 0$, $\mathcal{E}_{\mathbf{n}_i}(k,l) = 0$ (for $r \leq 2M-3$), and for $c = 1$, $\mathcal{E}_{\mathbf{n}_i}(k,l) = \mathcal{E}_{\mathbf{n}_i,\text{BMVDR}}(k,l)$.

It is worth mentioning that there are other more perceptually relevant methods (see e.g., [42, 43]) determining the ability of a user to correctly localize (before and after applying the binaural spatial filter) concurrent multiple sound sources in reverberant environments than the simple objective performance measures given in Eqs. (3.45)-(3.48). In this paper, we focus on the aforementioned simplified instrumental measures.

Note that we use the true ATFs in the constraints of the optimization problems of all competing methods. Therefore, we do not measure the corresponding error measures for the binaural cues of target source since they are always zero, because in all compared methods the distortionless constraints perfectly preserve the binaural cues of the target source.

SNR measures

We define the binaural global segmental signal-to-noise-ratio (gsSNR) gain as

$$\text{gsSNR}^{\text{gain}} = \text{gsSNR}^{\text{out}} - \text{gsSNR}^{\text{in}} \text{ dB}, \tag{3.49}$$

where the gsSNR input and output are defined as

$$\text{gsSNR}^{\text{in}} = \frac{1}{T}\sum_{l=1}^{T} \min\Big(\max\Big(\text{SNR}^{\text{in}}(l), -20\Big), 50\Big) \text{ dB}, \tag{3.50}$$

$$\text{gsSNR}^{\text{out}} = \frac{1}{T}\sum_{l=1}^{T} \min\Big(\max\Big(\text{SNR}^{\text{out}}(l), -20\Big), 50\Big) \text{ dB}, \tag{3.51}$$

respectively, where for the l-th frame, the binaural input signal-to-noise-ratio (SNR) is defined as

$$\text{SNR}^{\text{in}}(l) = 10\log_{10}\left(\frac{\sum_{k=1}^{N} \mathbf{e}^T \tilde{\mathbf{P}}_{\mathbf{x}}(k,l)\mathbf{e}}{\sum_{k=1}^{N} \mathbf{e}^T \tilde{\mathbf{P}}(k,l)\mathbf{e}}\right) \text{ dB}, \tag{3.52}$$

where $\mathbf{e}^T = [\mathbf{e}_L^T \quad \mathbf{e}_R^T]$, $\mathbf{e}_L^T = [1, 0, \cdots, 0]$ and $\mathbf{e}_R^T = [0, \cdots, 0, 1]$, $\tilde{\mathbf{P}}$ is defined in Eq. (3.11) and $\tilde{\mathbf{P}}_{\mathbf{x}}$ is similarly defined but it uses as diagonal block matrices the $\mathbf{P}_{\mathbf{x}}$ matrix. The binaural output SNR for the l-th frame, is defined as

$$\text{SNR}^{\text{out}}(l) = 10\log_{10}\left(\frac{\sum_{k=1}^{N} \mathbf{w}^H(k,l)\tilde{\mathbf{P}}_{\mathbf{x}}(k,l)\mathbf{w}(k,l)}{\sum_{k=1}^{N} \mathbf{w}^H(k,l)\tilde{\mathbf{P}}(k,l)\mathbf{w}(k,l)}\right) \text{ dB}, \tag{3.53}$$

where $\mathbf{w} = [\mathbf{w}_L^T(k,l) \quad \mathbf{w}_R^T(k,l)]^T$. Note that $\text{gsSNR}^{\text{out}}$ and gsSNR^{in} can be seen as average measures of the binaural SNR measures defined in [30]. We also use the frequency-weighted segmental SNR (fwsSNR) [44, 45] to measure the amount of noise suppression at the left and right HA. The fwsSNR gain at the left reference microphone is given by

$$\text{fwsSNR}_L^{\text{gain}} = \text{fwsSNR}_L^{\text{out}} - \text{fwsSNR}_L^{\text{in}} \text{ dB}, \tag{3.54}$$

where the input and output fwsSNR at the left reference microphone are given by [45]

$$\text{fwsSNR}_L^{\text{in}} = \frac{1}{T}\sum_{l=1}^{T} \min\left(\max\left(\sum_{j=1}^{N_{fb}} g_j \text{SNR}_{j,L}^{\text{in}}, -20\right), 50\right) \text{ dB}, \tag{3.55}$$

Table 3.2: Anechoic environment: Input noise levels for $r=1,2,3,4,5,6,7$.

Measure \ r	1	2	3	4	5	6	7
gsSNR^{in}	2.92	0.36	−0.81	−1.70	−2.88	−3.49	−3.98
$\text{fwsSNR}_L^{\text{in}}$	4.55	2.26	−0.60	−3.39	−5.77	−6.62	−6.82
$\text{fwsSNR}_R^{\text{in}}$	−2.78	−5.12	−5.90	−6.23	−6.41	−6.61	−7.14

$$\text{fwsSNR}_L^{\text{out}} = \frac{1}{T}\sum_{l=1}^{T} \min\left(\max\left(\sum_{j=1}^{N_{fb}} g_j \text{SNR}_{j,L}^{\text{out}}, -20\right), 50\right) \text{ dB}, \tag{3.56}$$

where $\text{SNR}_{j,L}^{\text{in}}$ and $\text{SNR}_{j,L}^{\text{out}}$ are the input and output SNRs, respectively, of the j-th frequency band at the left reference microphone. The SNR values of the N_{fb} frequency bands are weighted differently with weights g_j. The ranges and central frequencies of the frequency bands, and the values of $g_j, i = 1, \cdots, N_{fb}$ are selected as described in [46]. The input and output fwsSNR for the right reference microphone are defined similarly to Eqs. (3.55) and (3.56), respectively. Note that the noise-only frames are excluded from the evaluation.

3.5.3. Results

In the following experiments we evaluate the performance of the proposed and reference methods (i.e., BLCMV [27] with two different values of η, OBLCMV [28], BMVDR [30] and JBLCMV [29, 30]) as a function of the number of simultaneously present interferers, $1 \leq r \leq 7$. For instance, for $r=1$, only the interferer with index 1 is enabled while all the others are silent. For $r=2$, only the interferers with indices $1, 2$ are enabled, while the others are silent, and so on. Recall that each method has a different $m_{\max}$, except for the proposed method for $c > 0$ where $m_{\max}$ is difficult to be estimated, as explained in Section 3.4, and, therefore, m is always set to $m=r$. For each of the reference methods and the proposed method in the case of $c=0$ and if $r > m_{\max}$, we will use in the constraints only the first $m_{\max}$ interferers and the last $r - m_{\max}$ will not be preserved. For simplicity, we used the same $c = c_j$, for $j = 1, \cdots, m$, for all interferers in the proposed method. In other words, we assumed that the binaural cues of all interferers are equally important. Moreover, we selected for the adaptive change of $\tau_{(k)}$ the step parameter $\alpha(k_{\max})$ with $k_{\max} \in \{10, 50\}$. In Sections 3.5.3, 3.5.3 the simulations are carried out without taking into account room acoustics. In Section 3.5.3 the simulations are carried out by taking into account room acoustics.

SNR & Binaural Cue Preservation

In this section and in Section 3.5.3 the evaluation is undertaken in an anechoic environment. The binaural gsSNR^{in}, $\text{fwsSNR}_L^{\text{in}}$ and $\text{fwsSNR}_R^{\text{in}}$ values for $r=1,2,3,4,5,6$ and 7 are given in Table 3.2.

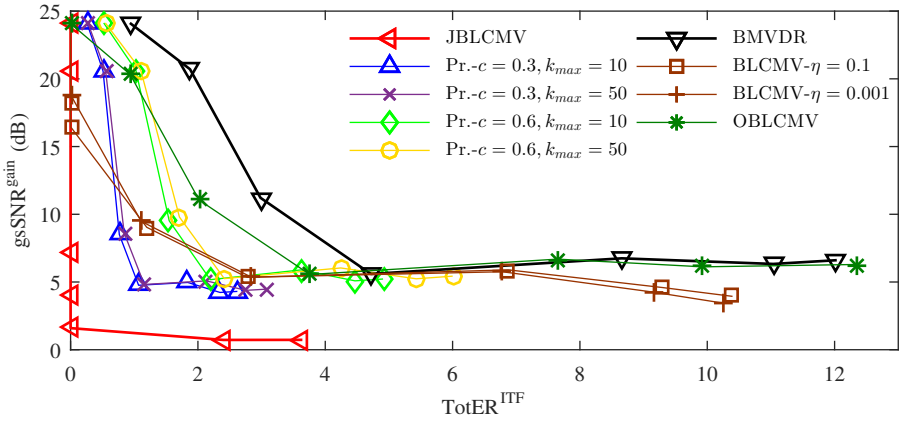

Figure 3.4: Anechoic environment: Combination of performance curves from Fig. 3.3 for the competing methods in terms of (a) noise reduction, (b) ITF error for different number of simultaneously present interferers r. The counting of r starts at the top left part of each curve.

Figs. 3.3 and 3.4 show the comparison of the proposed method (denoted by Pr.$-c=$value, $k_{\max}=$value) with the aforementioned reference methods in terms of binaural cue preservation and noise reduction. Note that BMVDR and the JBLCMV are the two extreme special cases of our method which can be denoted as Pr.$-c=1$ and Pr.$-c=0$, respectively. However, in these figures we used the original names for clarity. The performance curves are for different number of simultaneously present interferers r. As expected, the performance curves in Fig 3.3(a,d) of the proposed method always lie between the BMVDR and the JBLCMV for $m \leq 2M-3$ (see Theorem 1). Interestingly, this is also the case for $m > 2M-3$. As expected, the proposed method for $k_{\max}=50$ achieves slightly better noise reduction and worse binaural cue preservation than for $k_{\max}=10$. This is because for a larger $k_{\max}$, the proposed algorithm will provide a feasible solution closer to the boundary of $\Psi(c)$, as explained in Section 3.4.2. Fig. 3.4 is the combination of the curves of Figs. 3.3(a,d) into a single figure. Notice that the number of interferers r in this combined figure increase from $r=1$ up to $r=7$ along the curves from top-left, to bottom-right.

From Figs. 3.3(a,d), and Fig. 3.4 it is clear that, indeed the proposed method achieves a bounded noise reduction and a bounded binaural cue preservation accuracy. It is worth mentioning that a bounded performance in terms of the ITF error does not necessarily mean bounded performance in terms of ILD and IPD errors. For instance, in Fig. 3.3(e) the proposed method for $r=1,2$ with parameters $c=0.6$ and $k_{\max}=10,50$ has a larger total IPD error than the 0.6 times the total IPD error of the BMVDR. This is because, the proposed method does not bound the IPD and ILD errors separately, but their combination (i.e., the ITF error).

The BMVDR achieves the best noise reduction performance, but it does not preserve the binaural cues of the interferers. The JBLCMV accurately preserves the largest number of simultaneously present interferers and it has worse noise reduction

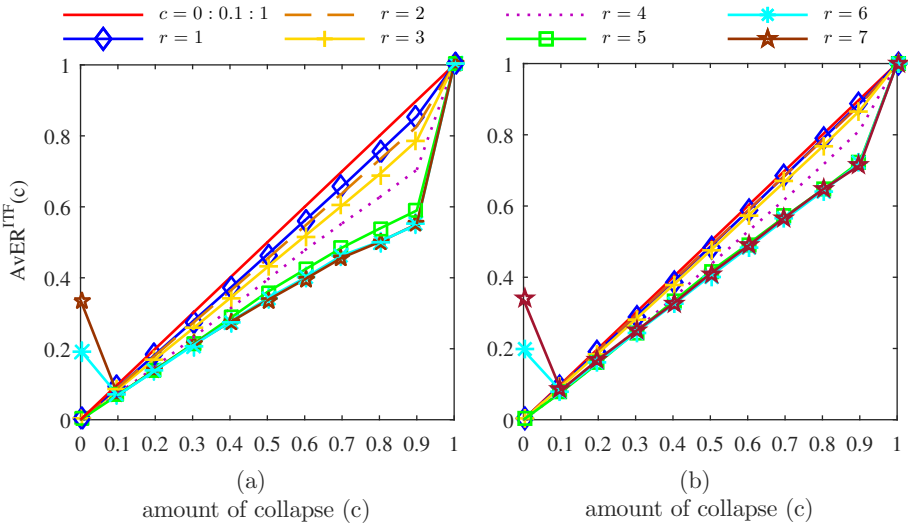

Figure 3.5: Anechoic environment: Average ITF error ratio as a function of c for $1 \leq r \leq 7$ for (a) $k_{\max} = 10$ and (b) $k_{\max} = 50$. The solid line is the c values.

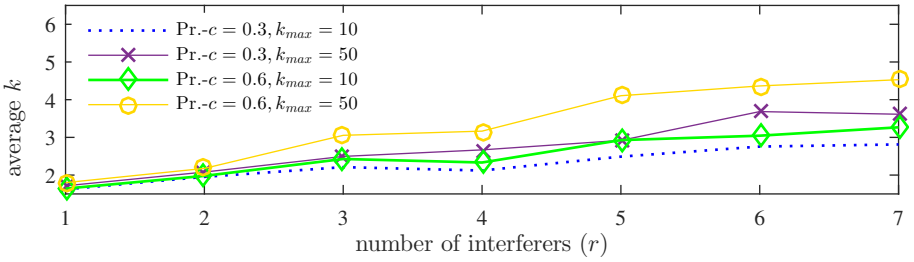

Figure 3.6: Anechoic environment: Average number of iterations as a function of simultaneously present interferers, r.

performance than all parametrizations of the proposed method. Note that $m_{\max} = 5$ for JBLCMV and, thus, the last two interferers cannot be included in the constraints and that is why the binaural cue preservation is not perfect. The OBLCMV comes second in terms of SNR performance, but it preserves the binaural cues of only one interferer.

Fig. 3.5 serves to visualize better the trade-off between fast termination and closeness to the boundary of $\Psi(c)$ (see Section 3.4.2 for details). More specifically, Fig. 3.5 shows the average ITF error ratio of the proposed method, for $k_{\max} = 10, 50$, as a function of c for different number of simultaneously present interferers r. As expected (see Section 3.5.2), $\text{AvER}^{\text{ITF}}(c) \leq c$ for $1 \leq r \leq 5$. This is also the case

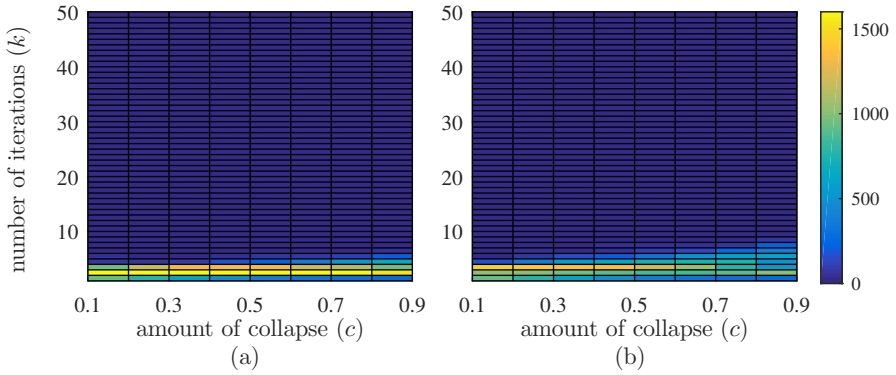

Figure 3.7: Anechoic environment: Top view of 3D histogram of number of frequency bins that have pairs (k, c) for the proposed method for (a) $k_{\max} = 10$ and (b) $k_{\max} = 50$.

for the curves for $r = 6, 7$ except for $c = 0$, as expected, because the proposed method becomes identical to the JBLCMV which can preserve the binaural cues of up to $m_{\max} = 2M - 3 = 5$ interferers while achieving controlled noise reduction. As expected, for $k_{\max} = 50$ all performance curves are closer to the boundary. In general, the larger the $m = r$, the less close the $\text{AvER}^{\text{ITF}}(c)$ of the proposed method is to c (see why in Section 3.4.2). Note that for the two extreme values $c = 0$ and $c = 1$, the proposed method becomes identical to the JBLCMV and the BMVDR, respectively. As was expected, for $c = 0$ and $r \leq 5$, $\text{AvER}^{\text{ITF}}(0) = 0$. The JBLCMV has $m_{\max} = 2M - 3 = 5$ and, therefore, for $c = 0$ and $r = 6, 7$, $\text{AvER}^{\text{ITF}}(0) > 0$. Finally, for $c = 1$, for all values of r, $\text{AvER}^{\text{ITF}}(1) = 1$ as expected.

Speed of Termination

Fig. 3.6 shows the average number of iterations (required for the proposed method to satisfy the stopping criterion) as a function of the simultaneously present interferers, r, of the four configurations of the proposed method that are tested in Figs. 3.3 and 3.4. It is clear that the proposed method terminates after 3-4 iterations on average, even for $r = 6, 7 > 2M - 3$. Note that for both tested values of $k_{\max}$, for all frames and frequency bins the proposed method terminated before reaching $k_{\max}$.

Fig. 3.7 shows a 3D histogram which depicts the statistical termination behaviour of the proposed method. Specifically, the proposed method is evaluated with different c values from 0.1 to 0.9 with a step-size 0.1. For each c value it is evaluated for all numbers of simultaneously present interferers, i.e., for $r = 1, \cdots, 7$ as in Fig 3.6. Hence, this histogram represents all gathered pair-values (c, k) of all frequency bins for all $r = 1, \cdots, 7$. The pairs (c, k) express the number of iterations (per frequency bin), k, that the proposed method need in order to terminate for a certain initial c. The z-axis, which is depicted with different colors, is the number of frequency bins that are associated with a certain pair (c, k) in the x-y axes. Again we see that, on average, after 3-4 iterations the algorithm terminates for $c = 0.1 : 0.1 : 0.9$.

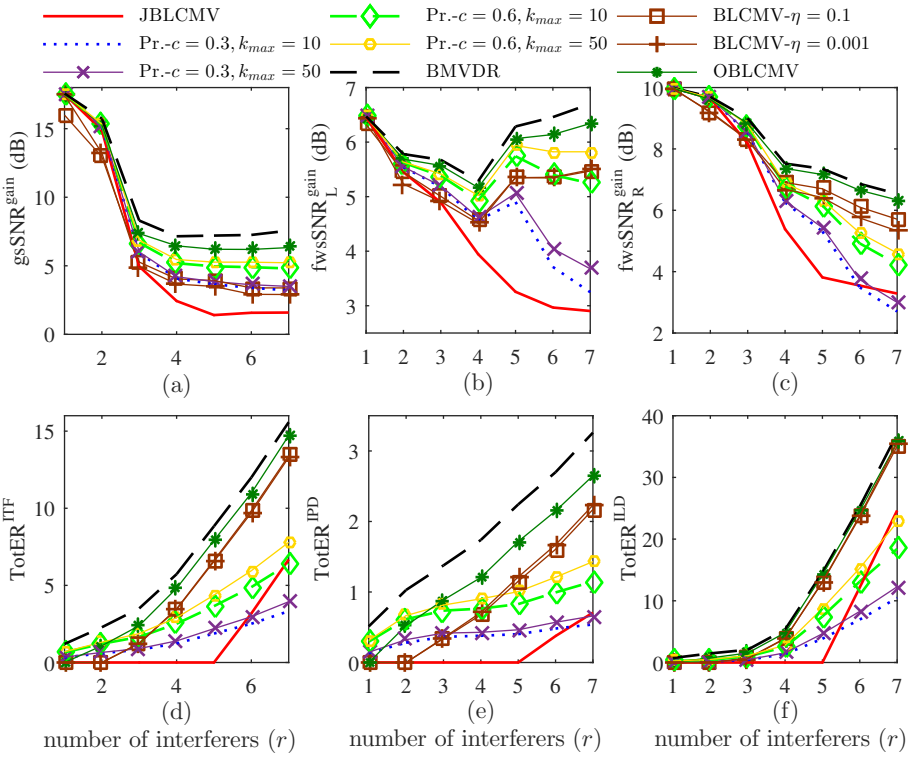

Figure 3.8: Reverberant environment (office): Performance of the competing methods in terms of (a,b,c) noise reduction, (d) ITF error, (e) IPD error, (f) ILD error.

Reverberation

Figs. 3.8, 3.9, 3.10 and 3.11 show the same experiments as in Figs. 3.3, 3.4, 3.5, and 3.6, respectively, but this time in a reverberant office environment. The same signals for the interferers and the target are used here. The reverberant BTE IRs are also taken from the database in [40]. Note that, the aforementioned database does not have the reverberant (for the office environment) BTE IRs corresponding to 240° and 300° degrees [40]. Therefore, we used the avalaible angles, 125°, 145° for the 6-th and 7-th interferer, respectively. Moreover, the sources are now placed on a $h = 100$ cm radius circle centered at the origin $(0, 0)$ (center of head) with an elavation of 0° degrees (because only this distance is available for the office environment in [40]). Similarly to the anechoic experiment, the microphone self noise at each microphone is simulated as WGN with $\mathbf{P_V} = \sigma^2\mathbf{I}$, where $\sigma = 6.1{*}10^{-5}$ which corresponds to an SNR of 50 dB with respect to the target signal at the left reference microphone. The binaural gsSNR$^{\text{in}}$, fwsSNR$^{\text{in}}_L$ and fwsSNR$^{\text{in}}_R$ values for $r = 1, 2, 3, 4, 5, 6$ and 7 are given in Table 3.3.

As it is shown in Figs. 3.8(a,d) and 3.9, again the performance of the proposed

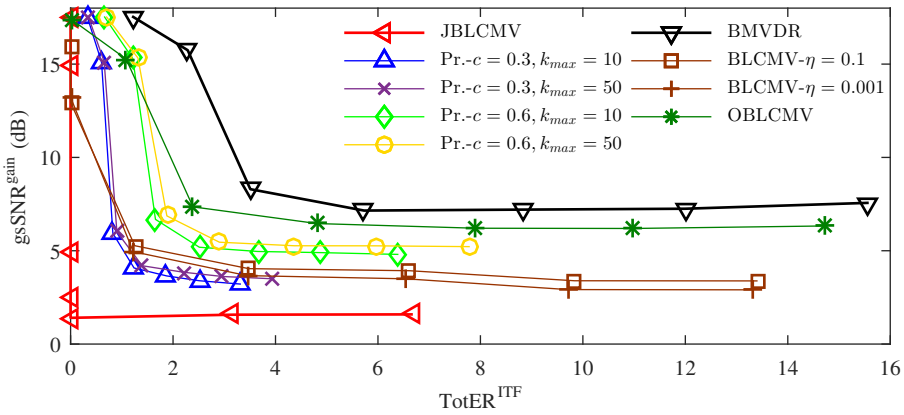

Figure 3.9: Reverberant environment (office): Combination of performance curves from Fig. 3.8 for the competing methods in terms of (a) noise reduction, (b) ITF error for different number of simultaneously present interferers r. The counting of r starts at the top left part of each curve.

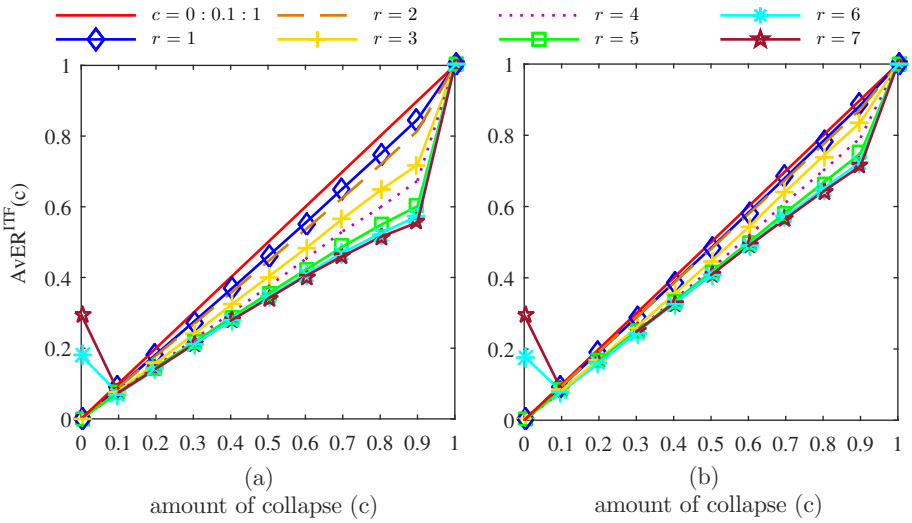

Figure 3.10: Reverberant environment (office): Average ITF error ratio as a function of c for $1 \leq r \leq 7$ for (a) $k_{\max} = 10$ and (b) $k_{\max} = 50$. The solid line is the c values.

method is bounded (see Theorem 1) even for $m > 2M - 3$. In Fig. 3.10 it is clear that the proposed method has very similar behavior as in Fig. 3.5, i.e., by increasing $k_{\max}$, the proposed method approaches closer to the boundary. Finally, in Fig. 3.11 it is shown that the speed of termination is not effected significantly

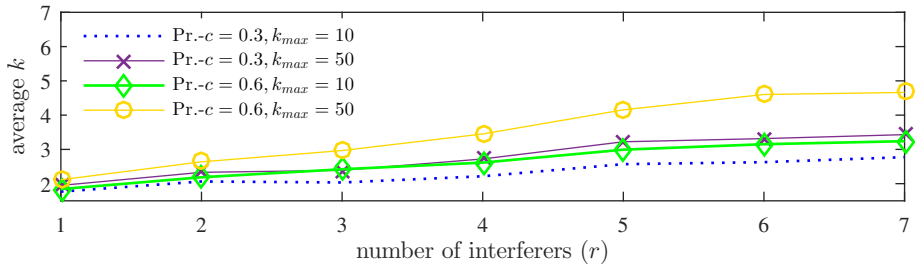

Figure 3.11: Reverberant environment (office): Average number of iterations as a function of simultaneously present interferers, r.

Table 3.3: Reverberant environment (office): Input noise levels for $r = 1, 2, 3, 4, 5, 6, 7$.

Measure \ r	1	2	3	4	5	6	7
gsSNR^{in}	−2.56	−5.35	−6.86	−7.94	−8.81	−9.48	−10.07
$\text{fwsSNR}_L^{\text{in}}$	−3.29	−4.76	−6.11	−7.43	−8.35	−8.95	−9.453
$\text{fwsSNR}_R^{\text{in}}$	−6.50	−8.02	−8.59	−8.92	−9.13	−9.32	−9.458

due to reverberation.

3.6. Conclusion

In this paper we proposed a new multi-microphone iterative binaural noise reduction method. The proposed method is capable of controlling the amount of noise reduction and the accuracy of binaural cue preservation per interferer using a robust methodology. Specifically, the inequality constraints introduced for the binaural cue preservation of the interferers, are selected in such a way that a) the total ITF error is always less or equal than a fraction of the corresponding total ITF error of the BMVDR method, and b) the achieved amount of noise reduction is larger or equal to the one achieved via JBLCMV. Therefore, the proposed method provides the flexibility to the users to parametrize the proposed method according to their needs. Moreover, the proposed method always preserves strictly the binaural cues of the target source. Although the proposed method guarantees a bounded binaural cue preservation accuracy and a bounded amount of noise reduction only for $m \leq 2M - 3$ interferers, it is experimentally demonstrated that is also capable of doing the same for more interferers and terminate in just a few iterations.

References

[1] 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* (Morgan & Claypool, 2013).

[2] M. Brandstein and D. Ward (Eds.), *Microphone arrays: signal processing techniques and applications* (Springer, 2001).

[3] K. Eneman et al., *Evaluation of signal enhancement algorithms for hearing instruments,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2008).

[4] S. Doclo and M. Moonen, *GSVD-based optimal filtering for single and multi-microphone speech enhancement,* IEEE Trans. Signal Process. **50**, 2230 (2002).

[5] A. Spriet, M. Moonen, and J. Wouters, *Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,* Signal Process. **84**, 2367 (2004).

[6] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. IEEE **57**, 1408 (1969).

[7] B. D. Van Veen and K. M. Buckley, *Beamforming: A versatile approach to spatial filtering,* IEEE ASSP Mag. **5**, 4 (1988).

[8] O. L. Frost III, *An algorithm for linearly constrained adaptive array processing,* Proceedings of the IEEE **60**, 926 (1972).

[9] J. M. Kates, *Digital hearing aids* (Plural publishing, 2008).

[10] T. Van den Bogaert, T. J. Klasen, L. Van Deun, J. Wouters, and M. Moonen, *Horizontal localization with bilateral hearing aids: without is better than with,* J. Acoust. Soc. Amer. **119**, 515 (2006).

[11] A. W. Bronkhorst, *The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,* Acta Acoustica **86**, 117 (2000).

[12] 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. **17**, 1071 (2009).

[13] H. Schmidt, A. B. Baggeroer, W. A. Kuperman, and E. K. Scheer, *Environmentally tolerant beamforming for high-resolution matched field processing: deterministic mismatch,* J. Acoust. Soc. Amer. **88** (1990).

[14] S. A. Vorobyov, *Principles of minimum variance robust adaptive beamforming design,* ELSEVIER Signal Process. **93**, 3264 (2013).

[15] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, *On optimal multi-channel mean-squared error estimators for speech enhancement,* IEEE Signal Process. Lett. **16**, 885 (2009).

[16] S. Gazor and W. Zhang, *Speech probability distribution,* IEEE Signal Process. Lett. **10**, 204 (2003).

[17] R. Martin, *Speech enhancement based on minimum mean-square error estimation and supergaussian priors,* IEEE Trans. Speech Audio Process. **13**, 845 (2005).

[18] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, *Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors,* IEEE Trans. Audio, Speech, Language Process. **15**, 1741 (2007).

[19] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment* (John Wiley & Sons, 2006).

[20] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process. Mag. **32**, 18 (2015).

[21] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2005) pp. 29–32.

[22] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[23] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, *Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions,* in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)* (2006).

[24] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, *The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids,* J. Acoust. Soc. Amer. **124**, 484 (2008).

[25] 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 Trans. Audio, Speech, Language Process. **23** (2015).

[26] E. Hadad, S. Gannot, and S. Doclo, *Binaural linearly constrained minimum variance beamformer for hearing aid applications,* in *Int. Workshop Acoustic Signal Enhancement (IWAENC)* (2012) pp. 1–4.

[27] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[28] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, *Optimal binaural lcmv beamformers for combined noise reduction and binaural cue preservation,* in *Int. Workshop Acoustic Signal Enhancement (IWAENC)* (2014) pp. 288–292.

[29] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[30] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, *Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,* IEEE Trans. Audio, Speech, Language Process. **23**, 2449 (2015).

[31] R. O. Duda, *Elevation dependence of the interaural transfer function,* in *Binaural and spatial hearing in real and virtual environments* (Mahwah, NJ: Lawrence Erlbaum, 1997) pp. 49–75.

[32] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, *Theoretical analysis of binaural multimicrophone noise reduction techniques,* IEEE Trans. Audio, Speech, Language Process. **18**, 342 (2010).

[33] 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. **5**, 529 (1997).

[34] W. M. Hartmann, *How we localize sound,* Physics Today **52**, 24 (1999).

[35] F. L. Wightman and D. J. Kistler, *The dominant role of low-frequency interaural time differences in sound localization,* J. Acoust. Soc. Amer. **91**, 1648 (1992).

[36] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, *Microphone-array hearing aids with binaural output .II. A two-microphone adaptive system,* IEEE Trans. Speech Audio Process. **5**, 543 (1997).

[37] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo, *Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem,* IEEE Trans. Signal Process. **51**, 313 (2003).

[38] R. G. Lorenz and S. P. Boyd, *Robust minimum variance beamforming,* IEEE Trans. Signal Process. **53**, 1684 (2005).

[39] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2004).

[40] H. Kayser, S. Ewert, J. Annemuller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, *Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,* EURASIP J. Advances Signal Process. **2009**, 1 (2009).

[41] J. Allen, *Short-term spectral analysis, and modification by discrete Fourier transform,* IEEE Trans. Acoust., Speech, Signal Process. **25**, 235 (1977).

[42] C. Faller and J. Merimaa, *Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,* J. Acoust. Soc. Amer. **116**, 3075 (2004).

[43] M. Dietz, S. D. Ewert, and V. Hohmann, *Auditory model based direction estimation of concurrent speakers from binaural signals,* ELSEVIER Speech Commun. **53**, 592 (2011).

[44] J. Tribolet, P. Noll, B. McDermott, and R. E. Crochiere, *A study of complexity and quality of speech waveform coders,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (1978) pp. 586–590.

[45] P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, 2013).

[46] *American National Standard Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, 1997).

4

Binaural Beamforming Using Pre-Determined Relative Acoustic Transfer Functions

This chapter is based on the article published as "Binaural Beamforming Using Pre-Determined Relative Acoustic Transfer Functions", by A.I. Koutrouvelis, R.C. Hendriks, R. Heusdens, J. Jensen and M. Guo in the Proceedings of the 25th European Signal Processing Conference (EUSIPCO), 2017.

Binaural hearing aid (HA) systems typically consist of two HAs, one at each ear, where each HA is typically equipped with multiple microphones. Allowing the HAs to collaborate, and combine their noisy microphone signals into a multi-microphone noise reduction algorithm, e.g., [1, 2], is an efficient way to achieve acoustic noise reduction. Unlike traditional monaural beamformers (BFs), e.g., [3, 4], which mainly focus on noise reduction, binaural BFs also aim to preserve the binaural cues of the sources in the acoustic scene [1]. This can be achieved through proper combination of the multi-microphone recordings of both HAs.

Many binaural BFs are based on the linearly constrained minimum variance (LCMV) framework [5]. This is due to the elegant and simple way in which constraints can be incorporated, as well as due to efficient adaptive implementations [4, 6]. The LCMV minimizes the output noise power under several linear equality constraints. In the case of binaural beamforming, these are often used to preserve the binaural cues of the present sources, while leaving the target signal undistorted at the two reference microphones. A different category of binaural noise reduction methods is based on the multi-channel Wiener filter (MWF) framework [7, 8]. The MWF-based methods [9–11] can achieve higher signal-to-noise-ratio (SNR) gains, but unlike the LCMV, they typically distort the target signal.

The binaural minimum variance distortionless response (BMVDR) BF [2] uses only two linear constraints to guarantee a distortionless response of the target at the two reference microphones. This results in binaural-cue preservation of the target source. Although this method achieves a relatively high binaural SNR gain, the price to pay is that the binaural cues of all interferers become identical to the binaural cues of the target after processing. The binaural LCMV (BLCMV) BF [12] preserves the binaural cues of the target, as well as of multiple interferers. This is achieved using two additional constraints per interferer. As a result, the degrees of freedom are exhausted fast for a small number of microphones. In contrast, the joint binaural LCMV (JBLCMV) BF [2, 13] achieves binaural-cue preservation using only one constraint per interferer. Thus, the JBLCMV can preserve the binaural cues of more interferers than the BLCMV [13].

Usually, the number of microphones per HA is relatively small, say, 2 or 3. As a result, the BLCMV and even the JBLCMV, suffer from the fact that the degrees of freedom are quickly exhausted with an increasing number of sources. This results in poor SNR gains and a small number of sources for which the binaural cues can be preserved. To overcome this problem, a relaxation of the JBLCMV method is proposed in [14]. In the current paper we refer to this method as relaxed JBLCMV (RJBLCMV). The equality constraints, used in the JBLCMV, which are meant to preserve the binaural cues of the interferers, are now replaced with inequality constraints. As a result, the binaural cues of the interferers are approximately preserved. The inequalities allow the RJBLCMV to use a larger number of constraints (and approximately preserve the binaural cues of more interferers) than other LCMV-based methods with equality constraints only, or, alternatively, to use the same number of constraints, but to trade-off binaural-cue accuracy against SNR gain.

An important limitation of all the aforementioned binaural BFs is that they

require estimates of the acoustic transfer functions (ATFs) or relative ATFs (RATFs) of the sources to form the constraints. This is rather impractical as estimation of these ATFs/RATFs is very challenging, in particular in dynamic scenarios. In this paper, we present a solution to this problem using fixed pre-determined RATFs, independent of the acoustical scenario. As a result, no tracking nor estimation of RATFs is needed. These pre-determined RATFs correspond to locations around the head. Each pre-determined RATF covers a small area in which some interferers might be present. As we use pre-determined RATFs instead of the true RATFs, steering vector mismatches (SVMs) are expected, potentially leading to a reduced preservation of the binaural cues. Increasing the number of pre-determined RATFs, however, leads to a lower expected SVM. We investigate both the JBLCMV and the RJBLCMV in the context of pre-determined RATFs, since these two methods can preserve the binaural cues of more locations than the BLCMV [13, 14]. It is to be expected that the RJBLCMV will be less sensitive to such SVMs as it, typically, allows to include much more constraints due to the introduced relaxation in binaural-cue preservation.

To guide the reader, in Section 4.1 the signal model and notation are presented. In Section 4.2 the idea of using pre-determined RATFs is introduced. In Section 4.3 the JBLCMV method is reviewed in the context of the pre-determined RATFs. In Section 4.4 we provide a useful decomposition of the JBLCMV spatial filter that explains the SVM problem due to the usage of pre-determined RATFs. In the same section, we also propose how to mitigate the SVM problem. In Section 4.5 the RJBLCMV method is reviewed in the context of the pre-determined RATFs. In Section 4.6, we evaluate the JBLCMV and RJBCMV using pre-determined RATFs. Finally, the conclusion is provided in Section 4.7.

4.1. Signal Model & Notation

Without loss of generality, let us assume that each of the two HAs has $M/2$ microphones, i.e., a total of M microphones. The processing is done in the discrete Fourier transform domain on a frame-by-frame basis, independently for each frequency bin. The noisy vector acquired from the M-microphone array for a single frequency bin is given by

$$\mathbf{y} = s\mathbf{a} + \sum_{i=1}^{r} u_i \mathbf{b}_i + \mathbf{v} \in \mathbb{C}^{M\times 1}, \tag{4.1}$$

where r is the number of interferers, $\mathbf{a}$ and $\mathbf{b}_i$ are the ATFs of the target and the i-th interferer, s and u_i are the target and the i-th interferer at the original locations, respectively, and $\mathbf{v}$ represents the background noise vector. Note that the first $M/2$ elements and the last $M/2$ elements of all vectors in Eq. (4.1) correspond to the left and right HAs, respectively. The first and last microphone of the M-microphone array are considered as the left and right reference microphones for binaural beamforming. Thus, for convenience, the first and last element of all vectors of Eq. (4.1) are indexed with subscript L and R, respectively, i.e., $\mathbf{a} = [\alpha_L, \alpha_2 ..., \alpha_{M-1}, \alpha_R]^T$ and $\mathbf{b}_i = [b_{i,L}, b_{i,2}, ..., b_{i,M-1}, b_{i,R}]^T$, etc. Each ATF is typically associated with a couple of RATFs. The RATFs of the target with respect to the left and right

reference microphones are given by $\bar{\mathbf{a}}_L = \mathbf{a}/a_L$ and $\bar{\mathbf{a}}_R = \mathbf{a}/a_R$, respectively, and for the i-th interferer $\bar{\mathbf{b}}_{iL} = \mathbf{b}_i/b_{i,L}$ and $\bar{\mathbf{b}}_{iR} = \mathbf{b}_i/b_{i,R}$.

Assuming that all sources in Eq. (4.1) are mutually uncorrelated, the cross power spectral density matrix (CPSDM) of the noisy measurements, $\mathbf{P}_\mathbf{y} \in \mathbb{C}^{M\times M}$, is given by

$$\mathbf{P}_\mathbf{y} = \mathrm{E}\left[\mathbf{y}\mathbf{y}^H\right] = p_s\mathbf{a}\mathbf{a}^H + \underbrace{\sum_{i=1}^{r} p_{u_i}\mathbf{b}_i\mathbf{b}_i^H + \mathbf{P}_\mathbf{v}}_{\mathbf{P}}, \tag{4.2}$$

where $\mathrm{E}[\cdot]$ denotes statistical expectation, $\mathbf{P}$ is the CPSDM of the total noise, p_s and p_{u_i} are the power spectral densities of the target and the i-th interferer signals, respectively, and $\mathbf{P}_\mathbf{v} = \mathrm{E}[\mathbf{v}\mathbf{v}^H]$ is the CPSDM of the background noise.

The binaural BFs consists of two spatial filters $\mathbf{w}_L, \mathbf{w}_R$ which are applied to $\mathbf{y}$, producing the outputs $x_L = \mathbf{w}_L^H\mathbf{y}$ and $x_R = \mathbf{w}_R^H\mathbf{y}$ at the left and right HAs, respectively.

4.2. Pre-Determined RATFs in Binaural Beamforming

In this section, we introduce the notion of using pre-determined RATFs in binaural beamforming. Specifically, we use m couples of pre-determined RATFs, i.e., $(\bar{\mathbf{q}}_{iL}, \bar{\mathbf{q}}_{iR})$, $i = 1, 2, \cdots, m$, where $\bar{\mathbf{q}}_{iL} = \mathbf{q}_i/q_{i,L}$ and $\bar{\mathbf{q}}_{iR} = \mathbf{q}_i/q_{i,R}$ are the pre-determined RATFs with respect to the left and right reference microphones, respectively, $\mathbf{q}_i$ is the corresponding pre-determined ATF, and $q_{i,L}$ and $q_{i,R}$ are the first and last elements of $\mathbf{q}_i$. Each pre-determined RATF couple, $(\bar{\mathbf{q}}_{iL}, \bar{\mathbf{q}}_{iR})$, corresponds to a pre-selected location in space with polar coordinates (θ_i,ϕ_i, h_i), where θ_i is the azimuth, ϕ_i the elevation, and h_i the distance from the center of the head. Note that the pre-determined RATFs are acoustic scene independent, but user dependent. Specifically, every user has its own set of anechoic head related transfer functions (HRTFs) which are used as pre-determined RATFs.

Without loss of generality, we examine the scenario where the m pre-selected locations are placed uniformly on the perimeter of a circle on the horizontal plane with radius h centered at the center of the head as shown in Fig. 4.1. As a result, we consider all azimuths equally important for binaural-cue preservation. The circle is selected to have a radius of $h > 2d^2/\lambda_{\min}$ m, where d is the distance between the two HAs and $\lambda_{\min} = 2c/F_s$, where c is the speed of sound and F_s is the sampling frequency. This is because, at this distance, the far field assumption is approximately met [15]. Consequently, the pre-determined RATFs are approximately distant invariant, i.e., there is no need to use more pre-determined RATFs for greater distances. Here, we assume that all sources are in the far-field, i.e., their distances are greater than h. A better approach, especially for nearby sources, is to have pre-determined RATFs for different elevations as well. For now we restrict ourselves to a single elevation.

If one of the m pre-determined RATF-couples a) matches with the actual RATF-couple of an interferer, i.e., $\exists j, i : (\bar{\mathbf{q}}_{jL}, \bar{\mathbf{q}}_{jR}) = (\bar{\mathbf{b}}_{iL}, \bar{\mathbf{b}}_{iR})$, and b) is included in

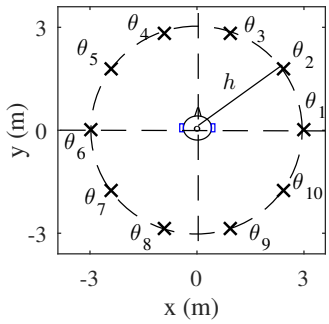

Figure 4.1: Example: $m = 10$, $h = 3$ m, 'x' markers denote locations of pre-determined RATFs around the head which is centered at the origin $(0, 0)$.

the constraints of an LCMV-based BF, the binaural cues of the interferer will be preserved. However, more interestingly (and more likely) is the case where there are interferers in the acoustic scene whose RATF-couple does not match with one of the pre-determined RATF-couples. This results in SVMs. Obviously, the expected SVM decreases when m is increased. An objective measure for the binaural-cue preservation of the i-th interferer after processing is the ITF error given by [14]

$$\mathcal{E}_i = \left| \mathrm{ITF}_i^{\mathrm{out}} - \mathrm{ITF}_i^{\mathrm{in}} \right| = \left| \frac{\mathbf{w}_L^H \mathbf{b}_i}{\mathbf{w}_R^H \mathbf{b}_i} - \frac{b_{i,L}}{b_{i,R}} \right|, \; i = 1, \cdots, r. \tag{4.3}$$

If $\mathcal{E}_i = 0$, the binaural BF preserves exactly the binaural cues of the i-th interferer. Since we use pre-determined RATFs, we also define the pre-determined ITF error which is given as

$$\mathcal{E}_i^q = \left| \frac{\mathbf{w}_L^H \mathbf{q}_i}{\mathbf{w}_R^H \mathbf{q}_i} - \frac{q_{i,L}}{q_{i,R}} \right| = \left| \frac{\mathbf{w}_L^H \bar{\mathbf{q}}_{iR}}{\mathbf{w}_R^H \bar{\mathbf{q}}_{iR}} - \bar{q}_{iR,1} \right|, \; i = 1, \cdots, m. \tag{4.4}$$

The binaural BF methods, discussed in the sequel, constrain the error $\sum_{i=1}^r \mathcal{E}_i^q$. Ideally, $\sum_{i=1}^r \mathcal{E}_i$ should be constrained as well. Constraining $\sum_{i=1}^r \mathcal{E}_i$, by constraining $\sum_{i=1}^r \mathcal{E}_i^q$, depends on a) how close the sources are to the pre-determined RATFs or, equivalently, it depends on how many pre-determined RATFs are used, and b) the number of the available degrees of freedom for noise reduction (see Section 4.4).

4.3. JBLCMV

The joint binaural LCMV (JBLCMV) spatial filter [2, 13, 14] is obtained by the following LCMV problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{C}^{2M \times 1}}{\arg\min} \underbrace{\mathbf{w}_L^H \mathbf{P} \mathbf{w}_L + \mathbf{w}_R^H \mathbf{P} \mathbf{w}_R}_{\mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w}} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \tag{4.5}$$

where $\mathbf{w} = [\mathbf{w}_L^T \quad \mathbf{w}_R^T]^T \in \mathbb{C}^{2M \times 1}$, $\mathbf{\Lambda} \in \mathbb{C}^{2M \times (2+m)}$ is assumed full column rank, $\tilde{\mathbf{P}} = \mathrm{diag}(\{\mathbf{P}, \mathbf{P}\}) \in \mathbb{C}^{2M \times 2M}$ is a block diagonal matrix, and $\mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H$ is a set

of $2+m$ linear equality constraints. The constraints aim at a) the preservation of the target at the two reference microphones, which also implies that its binaural cues are preserved, and b) the preservation of the binaural cues of m pre-selected locations, as proposed in Section 4.2. The first goal is accomplished via two linear constraints given by

$$\begin{bmatrix}\mathbf{w}_L^H & \mathbf{w}_R^H\end{bmatrix}\begin{bmatrix}\bar{\mathbf{a}}_L & \mathbf{0}\\ \mathbf{0} & \bar{\mathbf{a}}_R\end{bmatrix}=[1\quad 1]. \tag{4.6}$$

The second goal is accomplished by forcing the output ITF to be equal to the input ITF for each of the m pre-selected locations. This is accomplished using m linear constraints, i.e.,

$$\mathbf{w}_L^H\bar{\mathbf{q}}_{iL}-\mathbf{w}_R^H\bar{\mathbf{q}}_{iR}=0,\ i=1,\cdots,m. \tag{4.7}$$

Putting together all constraints we have

$$\underbrace{\begin{bmatrix}\mathbf{w}_L^H & \mathbf{w}_R^H\end{bmatrix}}_{\mathbf{w}^H}\underbrace{\begin{bmatrix}\bar{\mathbf{a}}_L & \mathbf{0} & \bar{\mathbf{q}}_{1L} & \cdots & \bar{\mathbf{q}}_{mL}\\ \mathbf{0} & \bar{\mathbf{a}}_R & -\bar{\mathbf{q}}_{1R} & \cdots & -\bar{\mathbf{q}}_{mR}\end{bmatrix}}_{\mathbf{\Lambda}\in\mathbb{C}^{2M\times 2+m}}=\underbrace{[1\ 1\ 0\cdots 0]}_{\mathbf{f}^H}. \tag{4.8}$$

Note that the available degrees of freedom for noise reduction is $2M-m-2$ and the maximum number of constraints that can be used for binaural-cue preservation of interferers/locations, while having at least one degree of freedom for noise reduction, is $2M-3$. The problem in Eq. (4.5) has a closed-form solution given by [4]

$$\hat{\mathbf{w}}=\tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f}. \tag{4.9}$$

Note that the BMVDR BF [2] is also obtained from the optimization problem in Eq. (4.5), but with $m=0$, i.e., the constraints in Eq. (4.7) are not used. It can be easily shown that the ITF error (see Eq. (4.3)) of the BMVDR is given by [14]

$$\mathcal{E}_i^{\text{BMVDR}}=\left|\frac{a_L}{a_R}-\frac{b_{iL}}{b_{iR}}\right|=\left|\bar{a}_{R,1}-\bar{b}_{iR,1}\right|,i=1,\cdots,r, \tag{4.10}$$

while the pre-determined ITF error (see Eq. (4.4)) is given by

$$\mathcal{E}_i^{q,\text{BMVDR}}=\left|\frac{a_L}{a_R}-\frac{q_{iL}}{q_{iR}}\right|=\left|\bar{a}_{R,1}-\bar{q}_{iR,1}\right|,i=1,\cdots,m. \tag{4.11}$$

4.4. SVM Problem

In this section, we provide a useful decomposition of the JBLCMV spatial filter that helps us to understand the SVM problem and how to handle it by using pre-determined RATFs. It is easy to show that if $\mathbf{\Lambda}$ and $\mathbf{f}$ in Eq. (4.8) are substituted into Eq. (4.9), the left and right spatial filters of the JBLCMV are given by

$$\hat{\mathbf{w}}_L=\rho_{L0}\mathbf{w}_{L0}+\rho_{L1}\mathbf{w}_{L1}+\cdots+\rho_{Lm}\mathbf{w}_{Lm}, \tag{4.12}$$

$$\hat{\mathbf{w}}_R=\rho_{R0}\mathbf{w}_{R0}+\rho_{R1}\mathbf{w}_{R1}+\cdots+\rho_{Rm}\mathbf{w}_{Rm}, \tag{4.13}$$

where

$$\mathbf{w}_{L0} = \frac{\mathbf{P}^{-1}\bar{\mathbf{a}}_L}{\bar{\mathbf{a}}_L^H\mathbf{P}^{-1}\bar{\mathbf{a}}_L}, \mathbf{w}_{R0} = \frac{\mathbf{P}^{-1}\bar{\mathbf{a}}_R}{\bar{\mathbf{a}}_R^H\mathbf{P}^{-1}\bar{\mathbf{a}}_R}, \tag{4.14}$$

$$\mathbf{w}_{Li} = \frac{\mathbf{P}^{-1}\bar{\mathbf{q}}_{iL}}{\bar{\mathbf{q}}_{iL}^H\mathbf{P}^{-1}\bar{\mathbf{q}}_{iL}}, \mathbf{w}_{Ri} = \frac{\mathbf{P}^{-1}\bar{\mathbf{q}}_{iR}}{\bar{\mathbf{q}}_{iR}^H\mathbf{P}^{-1}\bar{\mathbf{q}}_{iR}}, \tag{4.15}$$

and where $\rho_{Li}, \rho_{Ri}, i = 0, \cdots, m$ are functions of several generalized inner products of the form $\mathbf{z}^H\mathbf{P}^{-1}\mathbf{g}$, where $\mathbf{z}$, $\mathbf{g}$ are RATFs of the target or of the i-th pre-selected location. Note that $\mathbf{w}_{L0}$ and $\mathbf{w}_{R0}$ are the left and right MVDR BFs, of the BMVDR BF, preserving the target at the two reference microphones, while suppressing the interferers. Moreover, $\mathbf{w}_{Li}$ and $\mathbf{w}_{Ri}$ are the left and right minimum power distortionless response (MPDR) BFs [5] preserving the possible interferers close to the i-th pre-selected location at the two reference microphones while suppressing the remaining interferers which are further away. Note that we used the term MPDR for $\mathbf{w}_{Li}$ and $\mathbf{w}_{Ri}$, because we can think of $(\bar{\mathbf{q}}_{iL}, \bar{\mathbf{q}}_{iR})$ as an estimate of one or more actual RATF-couples of interferers (which are possibly close to the i-th pre-selected location) which are also present in the CPSDM $\mathbf{P}$.

It is widely known that the MPDR BF is not robust to SVMs [5, 16, 17]. The two MVDR BFs $\mathbf{w}_{L0}$ and $\mathbf{w}_{R0}$ are robust to SVMs, because the target is not present in $\mathbf{P}$ [17]. However, $\mathbf{w}_{Li}$ and $\mathbf{w}_{Ri}$ are most likely not robust to SVMs, because the interferers are present in $\mathbf{P}$, but some of the interferers might be far away from the m pre-selected locations. Therefore, the probable SVMs will most likely result in an uncontrolled amount of suppression of the r interferers from the non-robust BFs $\mathbf{w}_{Li}$ and $\mathbf{w}_{Ri}$. This will probably result in binaural-cue distortions of the interferers. Obviously, the expected SVM increase when the number of pre-determined RATFs decreases. Moreover, in [17] it was shown that the sensitivity of the MPDR BF to SVM increases when the maximum possible SNR (i.e., the SNR that can be achieved with no SVM) of the MPDR BF increases. In other words, when the number of degrees of freedom for noise reduction increases (i.e., $2M - m - 2$ increases), the SVM sensitivity, typically, increases.

4.5. RJBLCMV

The RJBLCMV [14] is a BF that relaxes some of the equality constraints and, thus, can, typically, use more constraints, for binaural-cue preservation, than the other LCMV-based methods. As a result, the RJBLCMV can preserve the binaural cues of more sources/locations than JBLCMV. The constraints in Eq. (4.8) can be partitioned as $\mathbf{w}^H[\mathbf{\Lambda}_1 \quad \mathbf{\Lambda}_2] = [\mathbf{f}_1^H \quad \mathbf{f}_2^H]$, where $\mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H$ contains the two constraints in Eq. (4.6) and $\mathbf{w}^H\mathbf{\Lambda}_2 = \mathbf{f}_2^H$ contains the pre-determined constraints in Eq. (4.7). The RJBLCMV makes use of this separation by having strict constraints with respect to the target, but inequality constraints on the m pre-selected locations,

i.e.,

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\operatorname{argmin}} \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H, \quad \left|\frac{\mathbf{w}_L^H\mathbf{q}_i}{\mathbf{w}_R^H\mathbf{q}_i} - \frac{q_{i,L}}{q_{i,R}}\right| \leq c\mathcal{E}_i^{q,\text{BMVDR}}, \quad i = 1,\cdots,m, \tag{4.16}$$

where $\mathcal{E}_i^{q,\text{BMVDR}}$ is given in Eq. (4.11), and $0 \leq c \leq 1$ is a user-defined parameter that controls the trade-off between binaural-cue accuracy and SNR gain. The maximum and minimum allowable amount of relaxation are obtained for $c = 1$ and $c = 0$, respectively. Having $0 \leq c \leq 1$ allows to relax the amount of binaural-cue preservation and trade this off with SNR gain, while guaranteeing that the amount of ITF error for a certain pre-selected location is always a proportion c below the BMVDR pre-determined ITF error. This does not necessarily imply that the ITF errors of the interferers will be a proportion c below of the BMVDR ITF error (see Eq. (4.3)). However, in Section 4.6, we experimentally show that for a large enough number of pre-determined RATFs, m, the average ITF error of the interferers is approximately a proportion c below of the BMVDR average ITF error.

The inequalities with the m pre-determined ATFs, $\mathbf{q}_i, i = 1,\cdots,m$, can be written in terms of pre-determined RATFs by multiplying both sides with $|q_{i,R}/q_{i,L}| = |\bar{q}_{iL,M}|$, where $\bar{q}_{iL,M}$ is the last element of $\bar{\mathbf{q}}_{iL}$. Therefore, the problem in Eq. (4.16) can be equivalently written as [14]

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\operatorname{argmin}} \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H, \quad |\mathbf{w}^H\mathbf{\Lambda}_{2,i}| \leq f_{2,i} \text{ for } i = 1,\cdots,m, \tag{4.17}$$

where $\mathbf{\Lambda}_{2,i}$ is the i-th column of $\mathbf{\Lambda}_2$, $f_{2,i} = |c\mathcal{E}_i^{q,\text{BMVDR}}\mathbf{w}_R^H\bar{\mathbf{q}}_{iR}\bar{q}_{iL,M}|$. The problem in Eq. (4.17) can be interpreted as a relaxed version of the JBLCMV. Note that if $c = 0$ in Eq. (4.17), the JBLCMV is obtained. The problem in Eq. (4.17) is non-convex and is approximately solved interatively as proposed in [14].

Unlike JBLCMV, the RJBLCMV is typically able to provide feasible solutions for $m > 2M - 2$ [14], which makes it applicable for the approximate preservation of binaural cues of more interferers/locations compared to the other strict equality constraint LCMV-based methods. Moreover, for the same number of constraints, the RJBLCMV can trade binaural-cue accuracy with improved SNR gain compared to JBLCMV [14]. As noted before, the very small number of available microphones in both HAs, limit the LCMV-based methods to preserve the binaural cues of only a very small number of pre-selected locations and, thus, the expected SVM is expected to be large. On the contrary, the RJBLCMV can typically approximately preserve much more locations and as we will see in Section 4.6, the average ITF error is smaller than with JBLCMV.

4.6. Experiments

The JBLCMV and RJBLCMV methods, using pre-determined RATFs, are evaluated in terms of noise reduction and binaural-cue preservation and compared to the

BMVDR. Noise reduction performance is measured with binaural SNR gain averaged overal all frequencies and frames (as in [14]) and the binaural cue preservation is measured with ITF error (see Eq. (4.3)) averaged over all frequencies and interferers as in [14]. In order to construct the microphone signals, and the pre-determined RATFs, we used the anechoic HRTFs from the database in [18]. The number of microphones that we used is $M = 6$, i.e., three microphones at each HA. The sampling frequency is 16 kHz, the frame length is 10 ms with an overlap of 50%, and the FFT length is 512.

The target talker is approximately in the look direction (i.e., $-5°$), with a distance of 0.8 m from the origin. We used its actual RATF (i.e., the ITF error of the target is zero for all methods) to form the distortionless constraints for all methods. The m pre-determined RATFs were selected as described in Section 4.2 with $h = 3$ m. We considered 8 simultaneously present speech shaped noise interferers with the same power as the target signal at the point that originates. Each one is randomly placed at one of the 72 possible angles of the HRTF database [18], with equal probability. The distance of the interferers from the origin is 3 m. The CPSDM $\mathbf{P}$ was computed using the true RATFs and estimated power spectral densities of the present interferers using all available data. Therefore, we examine the best possible performance of the competing methods, since no realistic estimation errors of $\mathbf{P}$ are considered. The background noise was simulated as white Gaussian noise with the same power at all microphones, with SNR = 50 dB with respect to the target signal at the left reference microphone. The maximum number of iterations for RJBLCMV were selected $k_{\max} = 10$ as in [14].

Fig. 4.2 shows the average performance for 20 different random placements of all interferers, where each random placement is accomplished as explained before. The RJBLCMV is evaluated with $m = 4, 8, 24$ pre-determined RATFs, while the JBLCMV with only $m = 4, 8$. This is because the JBLCMV can use up to $2M-3 = 9$ pre-determined RATFs. It is clear that the JBLCMV has poor performance, because it cannot use many constraints (i.e., many pre-determined RATFs) for $M = 6$, and the exptected SVM is large. On the other hand, RJBLCMV can achieve significantly better preservation of binaural cues even with just $M = 6$ microphones while still having quite a reasonable SNR gain. This is because the RJBLCMV can use much more constraints than the JBLCMV. As expected, the RJBLCMV with $c = 0.5$ has a larger SNR gain and ITF error, than with $c = 0.1$. Moreover, for $m = 24$, the RJBLCMV indeed achieves an average ITF error which is approximately c times below of the BMVDR average ITF error. Finally, both JBLCMV and RJBLCMV achieve a better binaural-cue preservation accuracy than the BMVDR for large enough m, i.e., for $m = 8, 24$.

4.7. Conclusion

A novel idea is presented for binarual-cue preservation without the need to estimate the relative acoustic transfer functions (RATFs). It is proposed to use pre-determined RATFs around the head of the hearing-aid user. The more pre-determined RATFs are used, the smaller the expected steering vector mismatch of the actual sources and the better the control of binaural-cue preservation. The

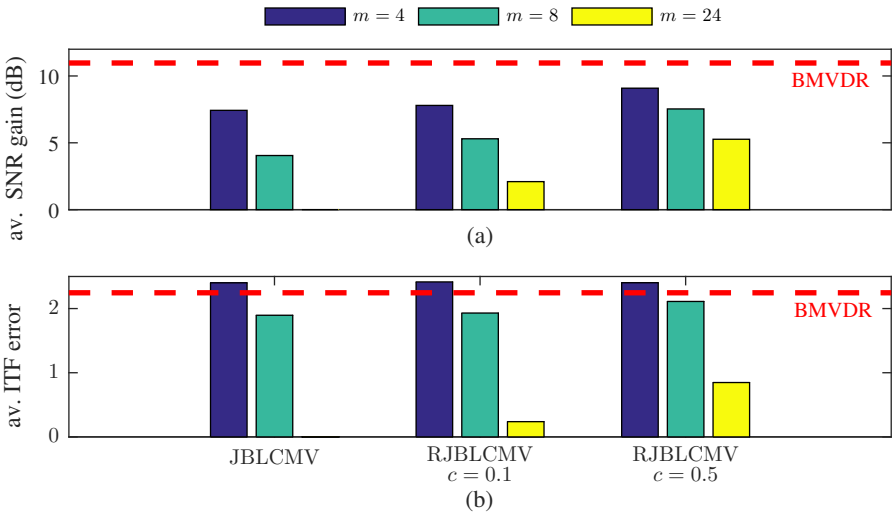

Figure 4.2: Performance of RJBLCMV (using $c = 0.1, 0.5$, and $m = 4, 8, 24$), JBLCMV (using $m = 4, 8$) and BMVDR (denoted with red dashed line) with respect to (a) average binaural SNR gain, and (b) average ITF error.

pre-determined RATFs can be used in both the RJBLCMV and the LCMV-based methods. However, it is shown that only the RJBLCMV is promising in this context, because it can use much more constraints than the other LCMV-based BFs, for a small number of microphones.

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process. Mag. **32**, 18 (2015).

[2] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, *Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,* IEEE Trans. Audio, Speech, Language Process. **23**, 2449 (2015).

[3] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. IEEE **57**, 1408 (1969).

[4] O. L. Frost III, *An algorithm for linearly constrained adaptive array processing,* Proc. of the IEEE **60**, 926 (1972).

[5] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing* (John Wiley & Sons, 2004).

[6] L. J. Griffiths and C. W. Jim, *An alternative approach to linearly constrained adaptive beamforming,* IEEE Trans. Antennas, Propag. **AP-30**, 27 (1982).

[7] S. Doclo and M. Moonen, *GSVD-based optimal filtering for single and multi-microphone speech enhancement,* IEEE Trans. Signal Process. **50**, 2230 (2002).

[8] A. Spriet, M. Moonen, and J. Wouters, *Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,* Signal Process. **84**, 2367 (2004).

[9] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[10] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, *Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions,* in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)* (2006).

[11] 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 Trans. Audio, Speech, Language Process. **23** (2015).

[12] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[13] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[14] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[15] R. Kennedy, T. Abhayapala, and D. Ward, *Broadband nearfield beamforming using a radial beampattern transformation,* IEEE Trans. Signal Process. **46**, 2147 (1998).

[16] S. A. Vorobyov, *Principles of minimum variance robust adaptive beamforming design,* ELSEVIER Signal Process. **93**, 3264 (2013).

[17] H. Cox, *Resolving power and sensitivity to mismatch of optimum array processors,* J. Acoust. Soc. Amer. **54**, 771 (1973).

[18] H. Kayser, S. Ewert, J. Annemuller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, *Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,* EURASIP J. Advances Signal Process. **2009**, 1 (2009).

5

A Convex Approximation of the Relaxed Binaural Beamforming Optimization Problem

This chapter is based on the article published as "A Convex Approximation of the Relaxed Binaural Beamforming Optimization Problem", by A.I. Koutrouvelis, R.C. Hendriks, R. Heusdens and J. Jensen in IEEE/ACM Trans. Audio, Speech and Language Processing, 2019.

Binaural beamforming (see e.g., [1] for an overview), also known as binaural spatial filtering, plays an important role in binaural hearing-aid (HA) systems [2]. Binaural beamforming is typically described as an optimization problem, where the objective is to i) minimize the output noise power, ii) preserve the target sound source at the left and right HA reference microphone, and iii) preserve the binaural cues of all sound sources after processing. The microphone array, which is typically mounted on the HA devices, has only a few microphones and, thus, there is only limited freedom (i.e., a small feasibility set) to search for a good compromise between the three aforementioned goals. Besides the challenge in finding a good trade-off among all these goals, the complexity should remain as low as possible, due to the limited computational power of the HA devices.

The binaural minimum variance distortionless response (BMVDR) beamformer (BF) [1] provides the maximum possible noise suppression among all binaural target-distortionless BFs [3]. Unfortunately, the BMVDR severely distorts the binaural-cues of the residual noise at the output of the filter. Specifically, the residual noise inherits the interaural transfer function of the target and, hence, sounds as originating from the target's direction [1]. The lack of spatial separation between the target and the noise after processing, may not only provide an unnatural impression to the user, but may also negatively effect the intelligiblity [4]. In [5, 6], the BMVDR was compared with an oracle-based (i.e., non-practically implementable) method in several noise fields (diffuse [5] and diffuse plus directional [6]). The oracle-based method has the same noise suppression as the BMVDR, but does not cause any binaural-cue distortions of the acoustic scene. The spatially correct oracle-based method achieved an improvement of about 3 dB in the 50% speech reception threshold (SRT) over the BMVDR. Therefore, there are several reasons to seek for methods that simultaneously provide the maximum possible noise suppression and binaural-cue preservation of all sources in the acoustic scene.

Several modifications of the BMVDR BF have been proposed, which can be roughly categorized into two groups. The first group consists of BFs that add or maintain a portion of the unprocessed scene at the output of the filter (see e.g., [5, 7–10]). An interesting approach, which is referred to as BMVDR-η [10], adds a portion of the unprocessed scene to the output of the BMVDR BF such that the binaural cues of the noise will be preserved in a certain extent. The second group consists of BFs, whose optimization problems have the same objective function as the BMVDR, but introduce extra equality [3, 11, 12] or inequality [13] constraints in order to preserve the binaural cues of the interferers after processing. These constraints are functions of either i) the (relative) acoustic transfer functions (R)(ATFs) of the interferers which can be estimated (see e.g., [14] for an overview), or ii) pre-determined anechoic (R)ATFs forming a grid around the head of the user as proposed in [15]. Moreover, these additional constraints in the optimization problem results in less degrees of freedom for noise reduction. With equality constraints, closed-form solutions may be derived, but the degrees of freedom can be easily exhausted when multiple interferers exist in the acoustic scene, resulting in poor noise reduction. On the other hand, inequality constraints provide more flexibility

and can approximately preserve the binaural cues of, typically, many more acoustic sources, or for the same number of acoustic sources provide a larger amount of noise reduction [13]. Unfortunately, closed-form solutions do not exist for the inequality-constrained binaural BFs and, thus, iterative methods with a larger complexity are used instead.

Recently, the relaxed binaural beamforming (RBB) optimization problem was proposed, which uses inequality constraints to preserve the binaural cues of the interfering sources [13]. The inequality constraints in the RBB are not convex, resulting in a non-convex optimization problem. In [13], a suboptimal successive convex optimization (SCO) method was proposed to approximately solve the RBB problem. In most cases, the SCO method needs to solve more than one convex optimization problem, per frequency bin, in order to converge. Convergence is achieved when all constraints of the RBB problem are satisfied. As a result, the SCO method guarantees an upper-bounded binaural-cue distortion of the interferers (as expressed by the interaural transfer function error), where the upper bound is controlled by the user.

Unfortunately, the SCO method is computationally very demanding due to its need to solve multiple convex optimization problems, per frequency bin, in order to converge. In this paper, we propose a semi-definite convex relaxation (SDCR) of the RBB optimization problem, which is significantly faster than the SCO method. This is because, the SDCR method requires to solve only one convex optimization problem per frequency bin. The main drawback of the SDCR method is that it does not guarantee user-controlled upper-bounded binaural-cue distortions as the SCO method. We solve this issue by combining the SDCR and SCO methods into a suboptimal hybrid method. The hybrid method guarantees user-controlled upper-bounded binaural-cue distortions, and still has a significantly lower computational complexity than the SCO method. Simulation experiments combined with listening tests show that both proposed methods, in most cases, provide a better trade-off between predicted intelligibility and binaural-cue preservation than the SCO method.

5.1. Signal Model and Notation

We assume that there is one target point-source signal, r point-source interferers, background noise, and two HAs with M microphones in total. The processing is accomplished per time-frequency bin independently. Neglecting time-frequency indices for brevity, the acquired M-element noisy vector in the DFT domain, for a single time-frequency bin, is given by

$$\mathbf{y} = \underbrace{s\mathbf{a}}_{\mathbf{x}} + \underbrace{\sum_{i=1}^{r} v_i \mathbf{b}_i + \mathbf{u}}_{\mathbf{n}} \in \mathbb{C}^{M\times 1}, \tag{5.1}$$

where s and v_i are the target and i-th interferer signals at the original locations; $\mathbf{a}$ and $\mathbf{b}_i$ the early acoustic transfer function (ATF) vectors of the target and i-th interferer, respectively; $\mathbf{u}$ the background noise, and $\mathbf{n}$ the total additive noise. The

background noise is due to the diffuse late reverberation from all point sources and the microphone-self noise. Assuming statistical independence between all sources, the noisy cross-power spectral density matrix (CPSDM) is given by

$$\mathbf{P}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}\mathbf{y}^H] = \mathbf{P}_{\mathbf{x}} + \mathbf{P}_{\mathbf{n}} \in \mathbb{C}^{M\times M}, \tag{5.2}$$

with $\mathbf{P}_{\mathbf{x}} = \mathrm{E}[\mathbf{x}\mathbf{x}^H] = p_s\mathbf{a}\mathbf{a}^H$ and $\mathbf{P}_{\mathbf{n}} = \mathrm{E}[\mathbf{n}\mathbf{n}^H]$ the target and noise CPSDMs, respectively, and $p_s = \mathrm{E}[|s|^2]$ the power spectral density of the target signal.

5.2. Binaural Beamforming Preliminaries

Binaural BFs consist of two spatial filters, $\mathbf{w}_L$, $\mathbf{w}_R \in \mathbb{C}^{M\times 1}$, which are both applied to the noisy measurements producing two different outputs given by

$$\begin{bmatrix}\hat{x}_L\\ \hat{x}_R\end{bmatrix} = \begin{bmatrix}\mathbf{w}_L^H\mathbf{y}\\ \mathbf{w}_R^H\mathbf{y}\end{bmatrix}, \tag{5.3}$$

where $\hat{x}_L, \hat{x}_R$ are played back by the loudspeakers of the left and right HAs, respectively. Note that the subscripts L and R are also used to refer to the two elements of the vectors in (5.1) associated with the left and right reference microphones of the binaural BF. Here, we select the first and the M-th microphones as reference microphones and, thus, $y_L = y_1$ and $y_R = y_M$. The same applies to all vectors in (5.1).

All BFs considered in this paper are target-distortionless. Their goal is not only noise supression, but also preservation of the binaural cues of all sources in the acoustic scene. In this paper, we mainly focus on preserving, after processing, the perceived direction of all point sources. Therefore, in the following, we mean directional binaural cues when we use the term binaural cues. A simple way of measuring the binaural cues of a source is via the interaural transfer function (ITF), which is a function of the ATF vector of the source [16]. The ITF of the i-th interferer before and after applying the spatial filter is given by [16]

$$\mathrm{ITF}_i^{\mathrm{in}} = \frac{b_{iL}}{b_{iR}}, \quad \mathrm{ITF}_i^{\mathrm{out}} = \frac{\mathbf{w}_L^H\mathbf{b}_i}{\mathbf{w}_R^H\mathbf{b}_i}. \tag{5.4}$$

The input and output ITF of the target is expressed similarly. Ideally, to preserve the binaural cues of the point sources, a binaural BF will produce the same ITF output as the input for all point sources. In practice, this is very difficult to achieve, when the number of interferers, r, is large and the number of microphones, M, is small [13]. As a result, most BFs will introduce some distortion to the ITF output, resulting in a non-zero ITF error given by [13]

$$\mathrm{ITF}_i^{\mathrm{e}} = \left|\mathrm{ITF}_i^{\mathrm{out}} - \mathrm{ITF}_i^{\mathrm{in}}\right| = \left|\frac{\mathbf{w}_L^H\mathbf{b}_i}{\mathbf{w}_R^H\mathbf{b}_i} - \frac{b_{iL}}{b_{iR}}\right| \geq 0. \tag{5.5}$$

5.2.1. BMVDR Beamforming

The BMVDR BF [1] achieves the maximum possible noise suppression among all binaural BFs and is obtained from the following simple optimization problem [1, 3]:

$$\begin{aligned}\hat{\mathbf{w}}_L, \hat{\mathbf{w}}_R = \underset{\mathbf{w}_L, \mathbf{w}_R}{\arg\min} & \begin{bmatrix}\mathbf{w}_L^H & \mathbf{w}_R^H\end{bmatrix} \tilde{\mathbf{P}} \begin{bmatrix}\mathbf{w}_L \\ \mathbf{w}_R\end{bmatrix} \\ \text{s.t.} \quad & \mathbf{w}_L^H \mathbf{a} = a_L \quad \mathbf{w}_R^H \mathbf{a} = a_R,\end{aligned} \tag{5.6}$$

where

$$\tilde{\mathbf{P}} = \begin{bmatrix}\mathbf{P_n} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_n}\end{bmatrix}. \tag{5.7}$$

The optimization problem in (5.6) provides closed-form solutions to the left and right spatial filters given by [1, 3]

$$\hat{\mathbf{w}}_L = \frac{\mathbf{P_n}^{-1}\mathbf{a}a_L^*}{\mathbf{a}^H\mathbf{P_n}^{-1}\mathbf{a}}, \quad \hat{\mathbf{w}}_R = \frac{\mathbf{P_n}^{-1}\mathbf{a}a_R^*}{\mathbf{a}^H\mathbf{P_n}^{-1}\mathbf{a}}. \tag{5.8}$$

It can easily be shown, that the output ITF of the i-th interferer of the BMVDR spatial filter is given by [3, 13]

$$\mathrm{ITF}_i^{\mathrm{out}} = \frac{a_L}{a_R}, \tag{5.9}$$

which is the ITF input of the target. Therefore, all interferers sound as coming from the target direction after applying the BMVDR spatial filter. The BMVDR ITF error of the i-th interferer is given by [13]

$$\mathrm{ITF}_i^{\mathrm{e,BMVDR}} = \left| \frac{a_L}{a_R} - \frac{b_{iL}}{b_{iR}} \right|. \tag{5.10}$$

5.2.2. Relaxed Binaural Beamforming

The relaxed binaural beamforming (RBB) optimization problem, introduced in [13], uses additional inequality constraints (compared to the BMVDR problem) to preserve the interferers' binaural cues. The RBB problem is given by [13]

$$\begin{aligned}\hat{\mathbf{w}}_L, \hat{\mathbf{w}}_R = \underset{\mathbf{w}_L, \mathbf{w}_R}{\arg\min} & \begin{bmatrix}\mathbf{w}_L^H & \mathbf{w}_R^H\end{bmatrix} \tilde{\mathbf{P}} \begin{bmatrix}\mathbf{w}_L \\ \mathbf{w}_R\end{bmatrix} \\ \text{s.t.} \quad & \mathbf{w}_L^H \mathbf{a} = a_L \quad \mathbf{w}_R^H \mathbf{a} = a_R, \\ & \left| \frac{\mathbf{w}_L^H \mathbf{b}_i}{\mathbf{w}_R^H \mathbf{b}_i} - \frac{b_{iL}}{b_{iR}} \right| \leq \mathcal{E}_i, \; i = 1, \cdots, m \leq r,\end{aligned} \tag{5.11}$$

where

$$\mathcal{E}_i = c_i \mathrm{ITF}_i^{\mathrm{e,BMVDR}}, \quad 0 \leq c_i \; \leq 1.$$

Note that $\mathcal{E}_i$ is c_i times the ITF error of the i-th interferer of the BMVDR BF [13]. Recall that the BMVDR causes full collapse of the binaural cues of the interferers

towards the binaural cues of the target. Therefore, the inequality constraints in (5.11) control the percentage of collapse. A small c_i implies good preservation of binaural cues of the i-th interferer, but a smaller feasibility set and, thus, less noise reduction. On the other hand, a large c_i implies worse binaural-cue preservation, but more noise reduction.

It is clear from the above that the additional inequality constraints of the RBB problem require the knowledge of the (R)ATF vectors of the interferers. In practice, interferers' (R)ATF vectors are unknown and estimation is required. Several methods for estimating RATF vectors exist (see e.g., [14] for an overview). An alternative approach is to use pre-determined ancechoic (R)ATF vectors of fixed azimuths around the head of the user, as proposed in [15]. These pre-determined (R)ATF vectors are acoustic scene independent and need to be obtained once for each user. This is useful when the (R)ATF vectors of the interferers are difficult to estimate, because e.g., the locations of the interferers relative to the head of the user are non-static. It is worth noting that by using pre-determined (R)ATF vectors, a larger number of inequality constraints, $m > r$, is typically used in (5.11). This is because we do not know where the interferers are located and we would like to cover the entire space around the head of the user.

If $c_i > 0, i = 1, \cdots, m$, the inequality constraints of the optimization problem in (5.11) are non-convex. As a result, the optimization problem in (5.11) is non-convex. In [13], a suboptimal successive convex optimization (SCO) method [13], described in Sec. 5.2.3, was proposed to approximately solve the RBB problem.

5.2.3. Successive Convex Optimization method

The successive convex optimization (SCO) method [13] approximately solves the RBB problem by solving multiple second-order cone program (SOCP) convex optimization problems per frequency bin. The SCO method converges, when all constraints of the RBB problem in (5.11) are satisfied. It has been shown that the SCO method always converges to a solution satisfying the constraints of the RBB problem if $m \leq 2M - 3$. This means that if the (R)ATF vectors of the interferers have been estimated accurately enough, the SCO method will guarantee user-controlled upper-bounded ITF error of the interferers [13]. For $m > 2M - 3$, no guarantees exist for convergence. In case the method does not converge, it stops after solving a pre-defined maximum number of convex optimization problems, $k_{\max}$. Nevertheless, for a reasonable number of inequality constraints, m, it has been experimentally shown that the SCO method always converges [13, 15]. It has been experimentally shown in [13], that for larger c_i values, the SCO method converges to solutions further away from the boundary of the inequality constraints of the RBB problem. This results in a better binaural-cue preservation and less noise reduction compared to the expected trade-off set by the user through the parameters $c_i, i = 1, \cdots, m$.

5.3. Proposed Convex Approximation Method

The proposed method is a semi-definite convex relaxation (SDCR) of the optimization problem in (5.11). First, we review two important properties that will be useful

for understanding the proposed optimization problem.

Property 1. *Any quadratic expression can be expressed as [17]*

$$\mathbf{q}^H\mathbf{Z}\mathbf{q} = tr\left(\mathbf{q}^H\mathbf{Z}\mathbf{q}\right) = tr\left(\mathbf{q}\mathbf{q}^H\mathbf{Z}\right). \tag{5.12}$$

Property 2. *We have the following equivalence relation [18]*

$$\mathbf{Z} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^H & \mathbf{C} \end{bmatrix} \succeq 0 \Leftrightarrow$$

$$\mathbf{A} \succeq 0, \quad \left(\mathbf{I} - \mathbf{A}\mathbf{A}^{\dagger}\right)\mathbf{B} = \mathbf{0}, \quad \mathbf{S}_1 \succeq 0, \tag{5.13}$$

$$\mathbf{C} \succeq 0, \quad \left(\mathbf{I} - \mathbf{C}\mathbf{C}^{\dagger}\right)\mathbf{B}^H = \mathbf{0}, \quad \mathbf{S}_2 \succeq 0, \tag{5.14}$$

with $\mathbf{S}_1 = \mathbf{C}–\mathbf{B}^H\mathbf{A}^{\dagger}\mathbf{B}$ *the generalized Schur complement of* $\mathbf{A}$ *in* $\mathbf{Z}$, $\mathbf{S}_2 = \mathbf{A}–\mathbf{B}\mathbf{C}^{\dagger}\mathbf{B}^H$ *the generalized Schur complement of* $\mathbf{C}$ *in* $\mathbf{Z}$, *and* $\mathbf{A}^{\dagger}$ *is the pseudo-inverse of* $\mathbf{A}$ *[19].*

Before, we present the proposed convex optimization problem, we first introduce an equivalent optimization problem to the problem in (5.11). That is,

$$\begin{aligned} \hat{\mathbf{w}}_L, \hat{\mathbf{w}}_R = \arg\min_{\mathbf{w}_L, \mathbf{w}_R} & \begin{bmatrix} \mathbf{w}_L^H & \mathbf{w}_R^H \end{bmatrix} \tilde{\mathbf{P}} \begin{bmatrix} \mathbf{w}_L \\ \mathbf{w}_R \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{w}_L^H\mathbf{a} = a_L \quad \mathbf{w}_R^H\mathbf{a} = a_R, \\ & \left| \frac{\mathbf{w}_L^H\mathbf{b}_i}{\mathbf{w}_R^H\mathbf{b}_i} - \frac{b_{iL}}{b_{iR}} \right|^2 \leq \mathcal{E}_i^2, \ i = 1, \cdots, m \leq r. \end{aligned} \tag{5.15}$$

By reformulating the inequality in (5.15), we obtain an equivalent quadratic constraint given by

$$\begin{aligned} & \left| \frac{\mathbf{w}_L^H\mathbf{b}_i}{\mathbf{w}_R^H\mathbf{b}_i} - \frac{b_{iL}}{b_{iR}} \right|^2 \leq \mathcal{E}_i^2 \Rightarrow \\ & \underbrace{\begin{bmatrix} \mathbf{w}_L^H & \mathbf{w}_R^H \end{bmatrix}}_{\mathbf{w}^H} \underbrace{\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^H & \mathbf{C} \end{bmatrix}}_{\mathbf{M}_i} \underbrace{\begin{bmatrix} \mathbf{w}_L \\ \mathbf{w}_R \end{bmatrix}}_{\mathbf{w}} \leq 0, \end{aligned} \tag{5.16}$$

where $\mathbf{A} = |b_{iR}|^2\mathbf{b}_i\mathbf{b}_i^H$, $\mathbf{B} = -b_{iL}^* b_{iR}\mathbf{b}_i\mathbf{b}_i^H$, $\mathbf{C} = \left(|b_{iL}|^2 - |b_{iR}|^2\mathcal{E}_i^2\right)\mathbf{b}_i\mathbf{b}_i^H$. Therefore, the optimization problem in (5.15) can be re-written as

$$\begin{aligned} \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ & \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^H \begin{bmatrix} \mathbf{a} & \mathbf{0} \\ \mathbf{0} & \mathbf{a} \end{bmatrix} = \begin{bmatrix} a_L & a_R \end{bmatrix}, \\ & \mathbf{w}^H\mathbf{M}_i\mathbf{w} \leq 0, \quad i = 1, \cdots, m. \end{aligned} \tag{5.17}$$

The matrix $\mathbf{M}_i$ is not positive semi-definite and, therefore, the quadratic inequality constraint is not convex and, hence, the optimization problem in (5.17) is not convex. The proof of non positive semi-definiteness of $\mathbf{M}_i$ uses Property 2. Specifically, note that $\mathbf{A} \succeq 0$, but $\mathbf{S}_1 = -|b_{iR}|^2\mathcal{E}_i^2\mathbf{b}_i\mathbf{b}_i^H \preceq 0$, because $\mathbf{b}_i\mathbf{b}_i^H \succeq 0$ and $-|b_{iR}|^2\mathcal{E}_i^2 \leq 0$ and, therefore, $\mathbf{M}_i$ is not positive semi-definite.

The optimization problem in (5.17) is a non-convex quadratic-constrained quadratic program (QCQP) [18, 20]. Following the methodology described in [20], we use Property 1 to re-write the optimization problem in (5.17) into the following equivalent formulation:

$$\begin{aligned}\hat{\mathbf{w}}, \hat{\mathbf{W}} = \arg\min_{\mathbf{w},\mathbf{W}} \ & \mathrm{tr}\left(\mathbf{W}\tilde{\mathbf{P}}\right) \\ \text{s.t.} \quad & \mathbf{w}^H \begin{bmatrix}\mathbf{a} & \mathbf{0} \\ \mathbf{0} & \mathbf{a}\end{bmatrix} = \begin{bmatrix}a_L & a_R\end{bmatrix}, \\ & \mathrm{tr}\left(\mathbf{W}\mathbf{M}_i\right) \leq 0, \quad i = 1, \cdots, m, \\ & \mathbf{W} = \mathbf{w}\mathbf{w}^H. \end{aligned} \tag{5.18}$$

The optimization problem in (5.18) is still not convex, but it has two differences with the problem in (5.17). The trace inequality is convex, but the new equality constraint, $\mathbf{W} = \mathbf{w}\mathbf{w}^H$ is not convex. Following [20], we apply the SDCR to the non-convex equality constraint of the problem in (5.18) and obtain the convex optimization problem given by

$$\begin{aligned}\hat{\mathbf{w}}, \hat{\mathbf{W}} = \arg\min_{\mathbf{w},\mathbf{W}} \ & \mathrm{tr}\left(\mathbf{W}\tilde{\mathbf{P}}\right) \\ \text{s.t.} \quad & \mathbf{w}^H \begin{bmatrix}\mathbf{a} & \mathbf{0} \\ \mathbf{0} & \mathbf{a}\end{bmatrix} = \begin{bmatrix}a_L & a_R\end{bmatrix}, \\ & \mathrm{tr}\left(\mathbf{W}\mathbf{M}_i\right) \leq 0, \quad i = 1, \cdots, m. \\ & \mathbf{W} \succeq \mathbf{w}\mathbf{w}^H. \end{aligned} \tag{5.19}$$

Using Property 2, the inequality constraint $\mathbf{W} \succeq \mathbf{w}\mathbf{w}^H$ can be re-written as a linear matrix inequality, and the optimization problem in (5.19) can be re-written into a standard-form semi-definite program (SDP) [20]. That is,

$$\begin{aligned}\hat{\mathbf{w}}, \hat{\mathbf{W}} = \arg\min_{\mathbf{w},\mathbf{W}} \ & \mathrm{tr}\left(\mathbf{W}\tilde{\mathbf{P}}\right) \\ \text{s.t.} \quad & \mathbf{w}^H \begin{bmatrix}\mathbf{a} & \mathbf{0} \\ \mathbf{0} & \mathbf{a}\end{bmatrix} = \begin{bmatrix}a_L & a_R\end{bmatrix}, \\ & \mathrm{tr}\left(\mathbf{W}\mathbf{M}_i\right) \leq 0, \quad i = 1, \cdots, m. \\ & \begin{bmatrix}\mathbf{W} & \mathbf{w} \\ \mathbf{w}^H & 1\end{bmatrix} \succeq 0. \end{aligned} \tag{5.20}$$

This is a convex problem, which can be solved efficiently [20]. If the solutions are on the boundary, i.e., $\hat{\mathbf{W}} = \hat{\mathbf{w}}\hat{\mathbf{w}}^H$, the minimizer, $\hat{\mathbf{w}}$, of the problem in (5.20) is

also the minimizer of the non-convex RBB problem. This means, that in the case of $\hat{\mathbf{W}} = \hat{\mathbf{w}}\hat{\mathbf{w}}^H$, the proposed problem in (5.20) is optimal and obtain solutions which satisfy the inequalities in Eqs. (5.17), (5.15) (5.11). Otherwise, if $\hat{\mathbf{W}} \succ \hat{\mathbf{w}}\hat{\mathbf{w}}^H$, the solution of the problem in (5.20) may or may not satisfy the inequalities of the RBB, which means that we lose the guarantee for user-controlled upper-bounded ITF error when the (R)ATF vectors of the interferers have been estimated accurately enough. In practice, $\hat{\mathbf{W}} = \hat{\mathbf{w}}\hat{\mathbf{w}}^H$ never occurred in our experiments and, thus, the two problems do not produce exactly the same solutions. However, we will experimentally show in Sec. 5.4 that the SDCR method always stays relatively close to the boundary of the inequality constraints of the RBB problem implying that it is a good approximation of the RBB problem.

Finally, the main advantage of the new proposed SDCR method is that it reduces significantly the computational complexity compared to the SCO method. Although SOCP problems (which are solved in the SCO method) are less computational complex than SDP [21], we will experimentally show in Sec. 5.4 that the proposed SDCR method is much less computational complex since a single convex problem is solved compared to the many more convex problems that must be solved in the SCO method per frequency bin.

5.3.1. Proposed Hybrid Method

In this section, we propose a hybrid method, which is a combination of the SDCR and the SCO methods. If the (R)ATF vectors of the interferers are estimated accurately enough, the hybrid method guarantees user-controlled upper-bounded binaural-cue distortions of the interferers as the SCO method. The proposed hybrid method is significantly faster than the SCO method and slightly slower than the SDCR method. We will experimentally show in Sec. 5.4, that the hybrid proposed method achieves solutions closer to the boundary of the inequality constraints of the RBB problem compared to the SCO method.

For a particular frequency bin, the hybrid method first solves the SDCR problem and then checks if there is a feasible solution which satisfies the inequality constraints of (5.11). If all of them are satisfied, the SDCR method will be used to approximately solve the RBB problem. Otherwise the SCO method is used to approximately solve the RBB problem in this frequency bin. Note that the SCO method always obtains a feasible solution for $m \leq 2M - 3$ (see Sec. 5.2.3) and, thus, the same holds for the hybrid method. In such a way, the hybrid method will always have a feasible solution (for $m \leq 2M-3$) which satisfies the constraints of the RBB problem, while at the same time reducing the overall computational complexity significantly. In order to avoid switching to the SCO method for just negligibly larger ITF errors than the user-controlled upper bounds $\mathcal{E}_i$, we use the following switching criterion:

$$\left| \frac{\mathbf{w}_L^H \mathbf{b}_i}{\mathbf{w}_R^H \mathbf{b}_i} - \frac{b_{iL}}{b_{iR}} \right| \leq \tilde{\mathcal{E}}_i, \ i = 1, \cdots, m, \tag{5.21}$$

Algorithm 2: Hybrid scheme

$\hat{\mathbf{w}}_1 \leftarrow$ SDCR Problem in Eq. (5.20)
if $\hat{\mathbf{w}}_1$ satisfies Eq. (5.21) **then**
  return $\hat{\mathbf{w}}_1$
else
  $\hat{\mathbf{w}}_2 \leftarrow$ SCO method [13]
  return $\hat{\mathbf{w}}_2$
end if

where $\tilde{\mathcal{E}}_i$ is a slightly increased upper bound and is given by

$$\tilde{\mathcal{E}}_i = (c_i + \epsilon)\left|\frac{a_L}{a_R} - \frac{b_{iL}}{b_{iR}}\right|, \; i = 1, \cdots, m, \tag{5.22}$$

where ϵ is very small, e.g., $0 < \epsilon < 0.1$. This modification avoids possible switching to the SCO method for negligibly larger ITF errors than the $\mathcal{E}_i$. The hybrid method is summarized in Algorithm 1.

5

5.4. Experiments

We conducted three sets of experiments. The first set (referred to as *Experiment 1*) examines the theoretical performance differences between the SCO method [13] (with $k_{\max} = 50$), the proposed SDCR method, and the proposed hybrid method (with $\epsilon = 0.05$) when the true early RATF vectors of the target and interferers are used. The second more practical set of experiments (referred to as *Experiment 2*) examines the performance of the same methods, when estimated early RATFs are used. The third practical set of experiments (referred to as *Experiment 3*) examines the performance of the same methods, when the pre-determined anechoic RATFs are used for preserving the binaural cues of the interferers (as proposed in [15]) and an estimated early RATF vector is used for preserving the binaural cues of the target. We also included in all three sets of experiments the reference methods BMVDR [1] and the BMVDR-η [5, 10]. The BMVDR-η depends on the parameter η ($0 \leq \eta \leq 1$) which controls the trade-off between noise reduction and binaural-cue preservation. Unlike the proposed methods in which a large c increases both the noise reduction performance and binaural-cue distortions, in the BMVDR-η, a large η decreases both the noise reduction performance and binaural-cue distortions.

5.4.1. Acoustic Scene Setup

The acoustic scene, considered in our experiments, was a reverberant office environment which consisted of one target female talker in the look direction (i.e., 0°), and 4 interferers, where each had the same average power at its original location, as the target signal at the original location. The first interferer was a male talker on the right-hand side of the HA user with azimuth of 80°; the second interferer was a music signal on the right-hand side of the HA user with azimuth of 50°; the third interferer was a vacuum cleaner on the left-hand side of the HA user with

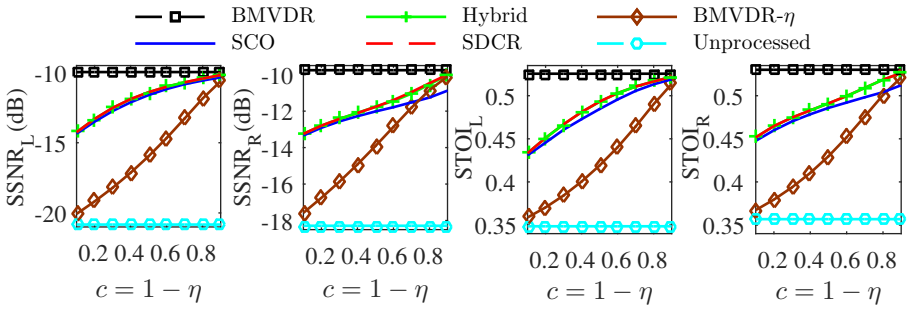

Figure 5.1: Experiment 1: Noise reduction and intelligiblity prediction performances.

azimuth $-35°$; and the fourth interferer was a high-frequency ringing mobile phone on the left-hand side with azimuth $-70°$. The microphone self-noise was set to have a 40 dB SNR at the left reference microphone, and it had the same power in all microphones.

5

5.4.2. Hearing-Aid Setup and Processing

The total number of microphones was $M = 4$; two at each HA. The sampling frequency was 16 kHz. We used the overlap-and-add processing method [22] for analyzing and synthesizing our signals. The analysis and synthesis windows were square-root Hanning windows and the overlap was 50%. The frame-length was 10 ms, i.e., 160 samples, and the FFT size was 256. The microphone signals were created using the head impulse responses (with a length of 458 ms) from the reverberant office environment from the database in [23]. Note that the true early RATF vectors were based on the first 10 ms of the impulse responses. The late reverberation was generated from the convolution of the late (after 10 ms) part of the impulse responses and the corresponding source signals.

In Experiments 2 and 3, the early RATF vector of a point source was estimated using a time-segment of 5 s in which only this point source signal (including its late reverberation) and the microphone-self noise was active. Specifically, we estimated the CPSDM and its eigenvalue decomposition and then we assigned to the early RATF vector the most significant (corresponding to the largest eigenvalue) relative eigenvector of the estimated CPSDM. The noise CPSDM was estimated using 5 seconds of a noise-only segment, where all interferers were active, but the target source was inactive.

We used the CVX toolbox [24] to solve the convex optimization problems associated with the SCO, SDCR and hybrid methods. The CVX toolbox uses an interior point method to solve the convex optimization problems [18]. We also used a common c value for all interferers in the inequality constraints, i.e., $c_i = c, \forall i$. The spatial filters of all methods were estimated only once using the same estimated noise CPSDM and, thus, they were time invariant. In the Experiment 3, for the pre-determined RATF vectors, we used the RATF vectors corresponding

to 24 pre-determined anechoic head impulse responses from the database in [23]. The pre-determined RATF vectors were associated with azimuths uniformly spaced around the head with a resolution of $360/24 = 15°$, starting from $-90°$. The pre-determined RATF vector at $0°$ was omitted from the constraints, because it was in the same direction as the RATF vector of the target. Note that the true RATF vectors of all interferers had an azimuth mismatch with the pre-determined RATF vectors' azimuths.

5.4.3. Evaluation Methodology

We measured the noise-reduction performance in terms of the segmental signal-to-noise-ratio (SSNR) only in target-presence time-regions. Let $\hat{\mathbf{X}}_L(t)$ and $\mathbf{Y}_L(t)$ denote the t-th time-frame of the estimated target and noisy signals, respectively, at the left reference microphone at the time domain, and $\mathcal{N}$ the set of the time-frames where the target is present. The SSNR at the left reference microphone is given by

$$\text{SSNR}_\text{L} = 10\log_{10}\frac{1}{|\mathcal{N}|}\sum_{t\in\mathcal{N}}\frac{||\hat{\mathbf{X}}_L(t)||_2^2}{||\hat{\mathbf{Y}}_L(t)-\hat{\mathbf{X}}_L(t)||_2^2}\ \text{dB}. \tag{5.23}$$

We also predicted intelligibility using the STOI measure [25].

We measured binaural-cue distortions with instrumental measures and a listening test. The instrumental measures were the average ITF error, interaural level difference (ILD) error and interaural phase difference (IPD) error per point source. These averages were calculated only over frequency (ommiting frequency bins with almost zero power), since we had fixed BFs over time. For the IPD error, we averaged only the frequency bins in the range of $0-1.5$ kHz, while for the ILD error, we averaged only the frequency bins in the range of $3-8$ kHz. This is because the ILDs are perceptually more important for localization above 3 kHz, while the IPDs are perceptually more important for localization below 1.5 kHz [26]. We used the expressions from [16] to compute the ILD and IPD errors for a single frequency bin.

The listening test was supplamentary to the Experiment 3 and is performed using the methodology described in [6]. Ten self-reported normal-hearing subjects participated (excluding the authors) and their age range was 26-37 years. They were asked to determine the azimuths of all point-sources in the acoustic scene when listening to signals processed by the compared methods as well as the unprocessed scene. The tested c values were 0.3 and 0.7 for the SCO, SDCR and hybrid methods. In addition to listening to the noisy and processed signals, the subjects also listened to the clean unprocessed point sources in isolation, in order to determine the reference azimuthms of the point sources. The localization errors were calculated with respect to the reference (and not the true) azimuths as in [6]. This is because we used only one set of head impulse responses from [23] to construct the binaural signals, which means that every subject will have a different reference azimuth. In this way, a significant estimation bias was removed. Two repetitions of the listening test were conducted. The reference azimuth of each source and every subject was computed as the average between the two repetitions, and the error was computed with respect to this averaged reference azimuth. The localization errors of the sources were

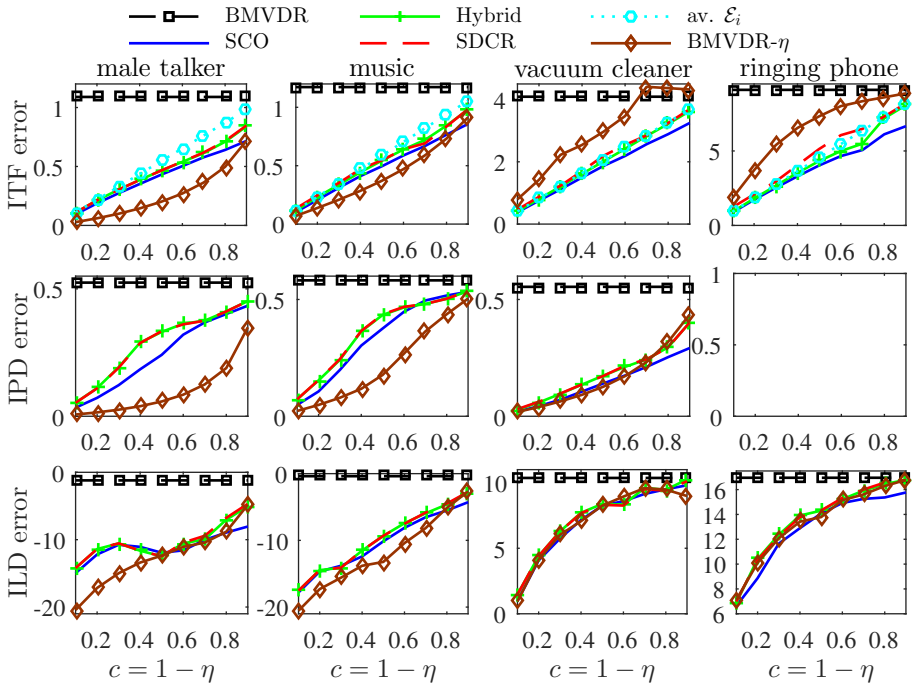

Figure 5.2: Experiment 1: Binaural-cue distortions (averaged over frequency) of interferers.

averaged over subjects and repetitions. A two-way analysis of variance (ANOVA) test [27] was performed which involves the processing method and the point source as the two factors. The ANOVA test determines i) if there are at least two of the localization error mean values significantly different for the processing method factor, ii) if there are at least two of the localization error mean values significantly different for the point source factor, iii) if there is an interaction between the two factors. Finally, multiple pairwise comparisons were undertaken through the t-test with the Bonferroni correction [27] in order determine which specific methods resulted in significantly different localization error mean values. We also measured the complexity of the compared methods in terms of the average number of convex optimization problems and average execution time per-frequency bin. Note that the BFs are fixed over time and, therefore, we do not measure varying complexity over time.

5.4.4. Experiment 1: Results with True Early RATF Vectors

In this section, the compared methods use the true early RATF vectors of the sources in the constraints. Fig. 5.1 depicts the noise reduction performance and intelligibility prediction of the unprocessed scene, the SCO, SDCR, hybrid, BMVDR,

and BMVDR-η methods at both reference microphones. The performance of SCO, SDCR and hybrid methods is measured for c values ranging from 0.1 to 0.9 with a step-size of 0.1. The performance of the BMVDR-η method is measured for η values ranging from 0.1 to 0.9 with a step-size of 0.1. In all figures, for illustration purposes, the η and c values are related as $c = 1 - \eta$. As expected, as c increases (and η decreases), the noise reduction and predicted intelligibility increase. As expected the BMVDR has the best noise reduction performance and predicted intelligibility. All methods based on the RBB problem achieve similar performances for the left reference microphone, while for the right reference microphone the SCO method achieves the worst noise reduction performance among all, especially for $c \geq 0.5$. Note that the SDCR method has almost identical performance as the hybrid method. This is because, in this example the hybrid method switched to the SCO method only a few times. Finally, the BMVDR-η method has a comparable predicted intelligibility with the proposed methods only for small η values.

Fig. 5.2 shows the binaural-cue distortions of the compared methods per interfering source. The binaural-cue distortions of the target source are always zero in Experiment 1. As expected, as c increases (and η decreases), the binaural-cue distortions increase. For the ITF errors, we also display the c times the average ITF error of the BMVDR (which is labeled as av. $\mathcal{E}_i$) in order to visualize the closeness of the estimated spatial filters at the boundary of the inequality constraints of the RBB problem. It is clear that both SDCR and hybrid methods are closer to the boundary compared to the SCO method for the same c value. Moreover, the hybrid method is for all c values (on average) below the boundary, even if we used the extended switch criterion in (5.21). On the other hand, the ITF error of the SDCR method sometimes (see ringing mobile phone) is slightly above the boundary. As explained in Sec. 5.3, this is because the SDCR method does not guarantee a user-controlled upper-bounded ITF error as the SCO or the hybrid methods do. Notably, the SCO method for large c values (e.g., $c \geq 0.6$), is not close to the boundary, while the SDCR and hybrid methods are closer to the boundary. Thus, the SDCR and hybrid methods achieve more expected binaural-cue distortions according to the trade-off parameter set by the user compared to the SCO method. Note also that the IPD error for the ringing mobile phone was not computed because it has almost zero power below 1.5 kHz.

Fig. 5.3 shows the computational complexity of the compared methods in terms of average number of convex optimization problems required to solve for convergence and average cpu time in seconds per frequency bin. The SDCR method requires to solve much less convex problems than the SCO method (especially at larger c values) and slightly less compared to the hybrid method. The hybrid method requires to solve much less convex problems than the SCO method. The fastest method among all is obviously the BMVDR-η method because it has a closed-form solution while all the other methods are iterative.

We can conclude from the above that, in most cases, both proposed methods are more optimal than the SCO method. Specifically, both proposed methods provide solutions that are closer to the expected solutions of the original RBB problem, since both proposed methods are closer to the boundary. This means that both meth-

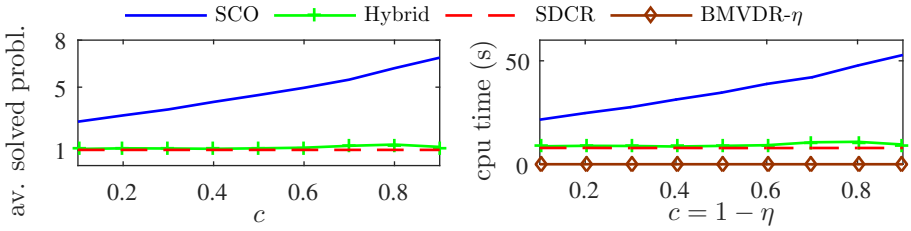

Figure 5.3: Experiment 1: Computational complexity measured as the average number of solved convex optimization problems and average computation time (seconds) per frequency bin.

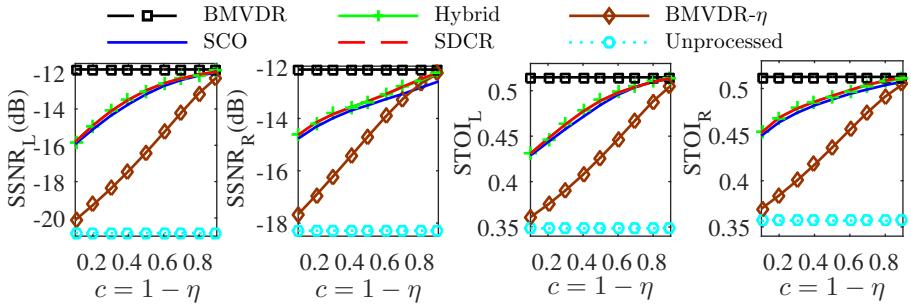

Figure 5.4: Experiment 2: Noise reduction and intelligiblity prediction performances.

ods provide a more user-controlled trade-off between noise reduction and binaural-cue preservation than the SCO method, especially in large c values. Finally both proposed methods are significantly less computationally demanding than the SCO method.

5.4.5. Experiment 2: Results with Estimated early RATF Vectors

In this section, the compared methods use estimated RATF vectors. Fig. 5.4 shows the noise reduction performance and intelligibility prediction of the compared methods which is very similar to the one in Fig. 5.1. Fig. 5.5 shows the binaural-cue distortions of the compared methods per point source (including the target source). As expected, here we have ITF errors which are sometimes above $\mathcal{E}_i$, because of the estimation errors in the RATF vectors. The computational complexity performance is omitted because is very similar to Fig. 5.3. Finally, the BMVDR-η method has a similar performance as with Experiment 1, since the only thing that has changed is the estimation error in the target RATF vector.

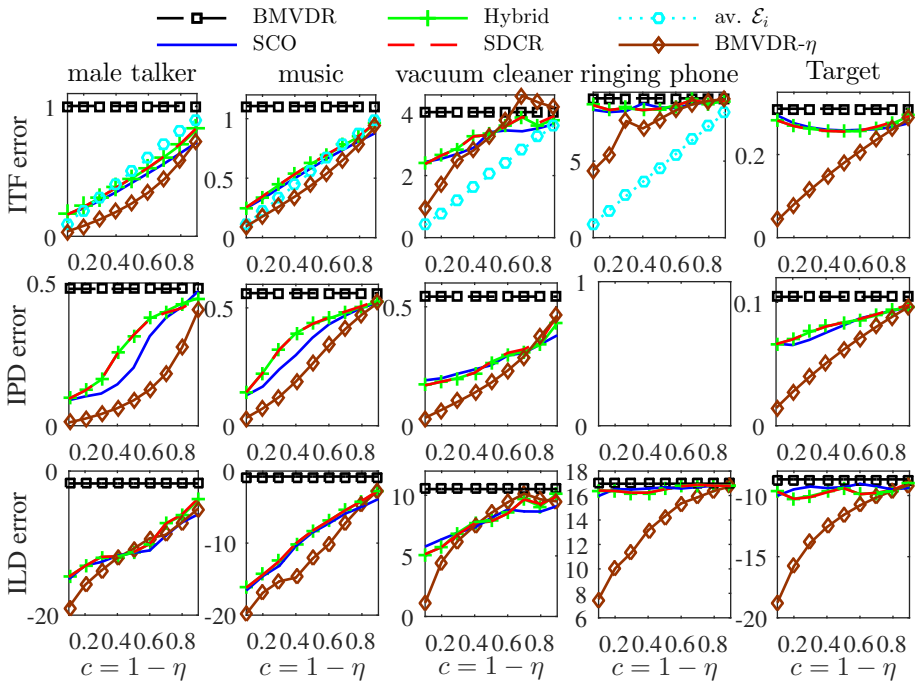

Figure 5.5: Experiment 2: Binaural-cue distortions (averaged over frequency) of point sources.

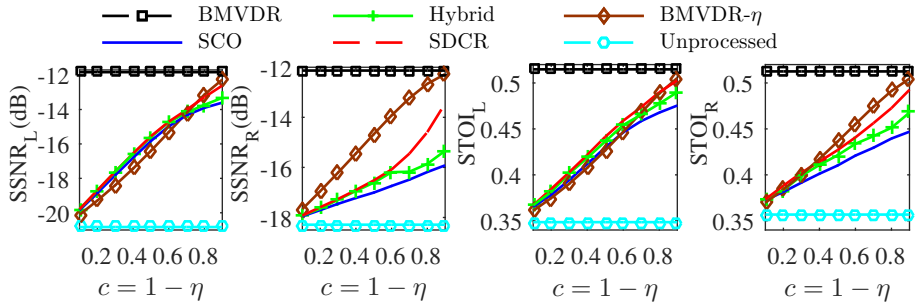

Figure 5.6: Experiment 3: Noise reduction and intelligiblity prediction performances.

5.4.6. Experiment 3: Results with Pre-Determined RATF Vectors

In this section, the SCO, SDCR and hybrid methods use the pre-determined RATF vectors for the interferers' binaural-cue preservation and an estimated early RATF vector for the target. Fig. 5.6 shows the noise reduction performance and intelligibil-

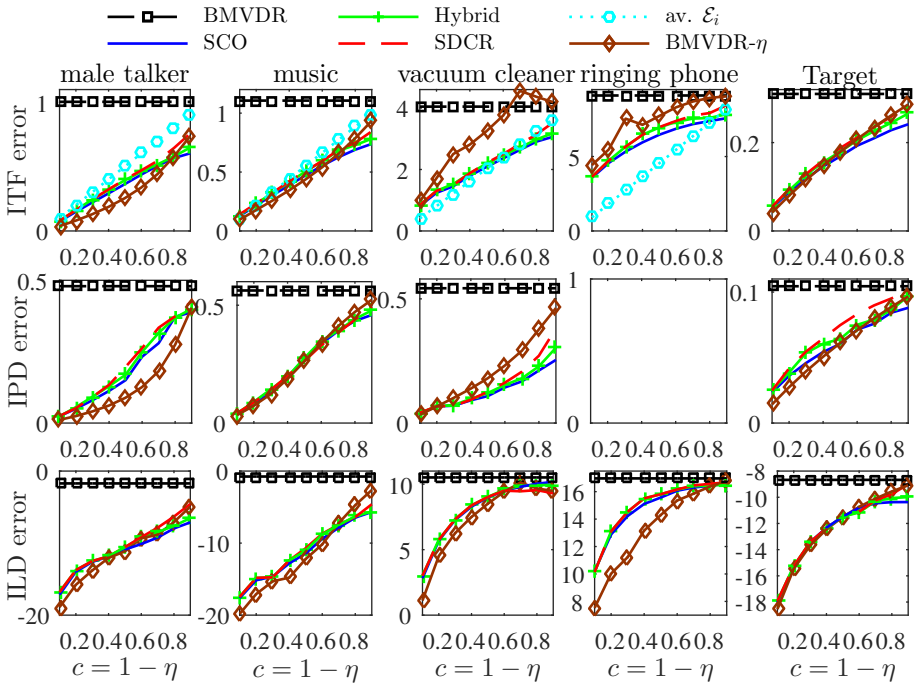

Figure 5.7: Experiment 3: Binaural-cue distortions (averaged over frequency) of point sources.

ity prediction of the compared methods. Here the gap in performance (for the same c value) between the proposed methods and the SCO method is bigger compared to the case where the true RATF vectors were used. The proposed methods (especially the SDCR method) significantly improved both noise reduction and predicted intelligibility at both reference microphones for the same c value compared to the SCO method. The reason why the performance gap between the SDCR method and the hybrid method is increased compared to Experiment 1 is because the hybrid method switched many more times to the SCO method (see Algorithm 1) in Experiment 3. In conclusion, for the same c value, both proposed methods achieved in most cases a better noise reduction and predicted intelligibility than the SCO method, especially for larger c values. The BMVDR-η method has the same performance as with the Experiment 2 and now has a comparable intelligibility improvement for all η values compared to the proposed methods.

Fig. 5.7 shows the binaural-cue distortions of the compared methods per point source (including the target source). As expected, when pre-determined RATF vectors are used, all methods do not guarantee a user-controlled upper-bounded ITF error of the interferers which will be c times the BMVDR ITF error. Therefore, all methods, in many occasions (see vacuum cleaner and ringing mobile phone), result in a larger ITF error than the average $\mathcal{E}_i$. The SCO method has the lowest binaural-cue

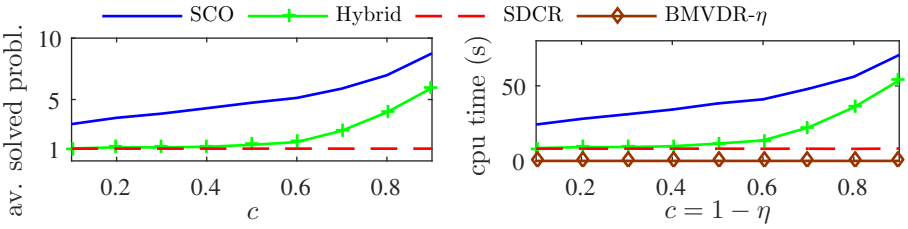

Figure 5.8: Experiment 3: Computational complexity measured as the average number of solved convex optimization problems and average computation time (seconds) per frequency bin.

distortions compared to the proposed methods because it is further away from the boundary of the inequality constraints of the RBB problem. Nevertheless, we will see later on in the listening test that the compared methods do not have significantly different binaural-cue distortions for the same c value.

In Fig. 5.8, we show the computational complexities of the compared methods. The results are similar to the results in Fig. 5.3 with the only difference that now the hybrid method does not achieve significant computational savings over the SCO method as with Experiment 1. However, the usage of the hybrid method using pre-determined RATF vectors is not critical, since no method can guarantee user-controlled upper-bounded ITF error of the interferers, unless the number of pre-determined RATF vectors is huge. However, this is not practical as it may result in non-feasible solutions or the noise reduction will be negligible.

Fig. 5.9 shows the results of the subjective localization test of Experiment 3. The examined values for the SCO, SDCR and hybrid methods are $c = 0.3, 0.7$, while for BMVDR-η we choose $\eta = 0.8$. A similar behavior as with the instrumental binaural-cue distortion measures is observed here. For a large c value we have in most cases a larger localization error. Moreover, as expected the BMVDR method has the largest localization error. Finally, the BMVDR-η method for $\eta = 0.8$ has a similar performance with the RBB-based methods for $c = 0.3$. Note that among all interferers the mobile ringing phone was the most difficult to localize for $c = 0.7$. Several users also reported difficulty in localizing the ringing phone after completing the test. We believe that this is because of the high frequency content of the ringing tone of the mobile phone and only the ILDs might have been used for localization.

Table 5.1 shows the results of the ANOVA test. We can conclude from the results that i) at least two of the mean values of the factor point source are significantly different, ii) at least two of the mean values of the factor processing method are significantly different and iii) there is a significant interaction between the two factors. Since there is a significant interaction between the two factors we have undertaken comparisons between pairs of methods for each interferer separately with several t-tests. The significance level was set to 1%. For the female talker all methods are not significantly different. For the male talker, music and vacuum cleaner, all methods are significantly different from the BMVDR method, but, surprisingly are not signif-

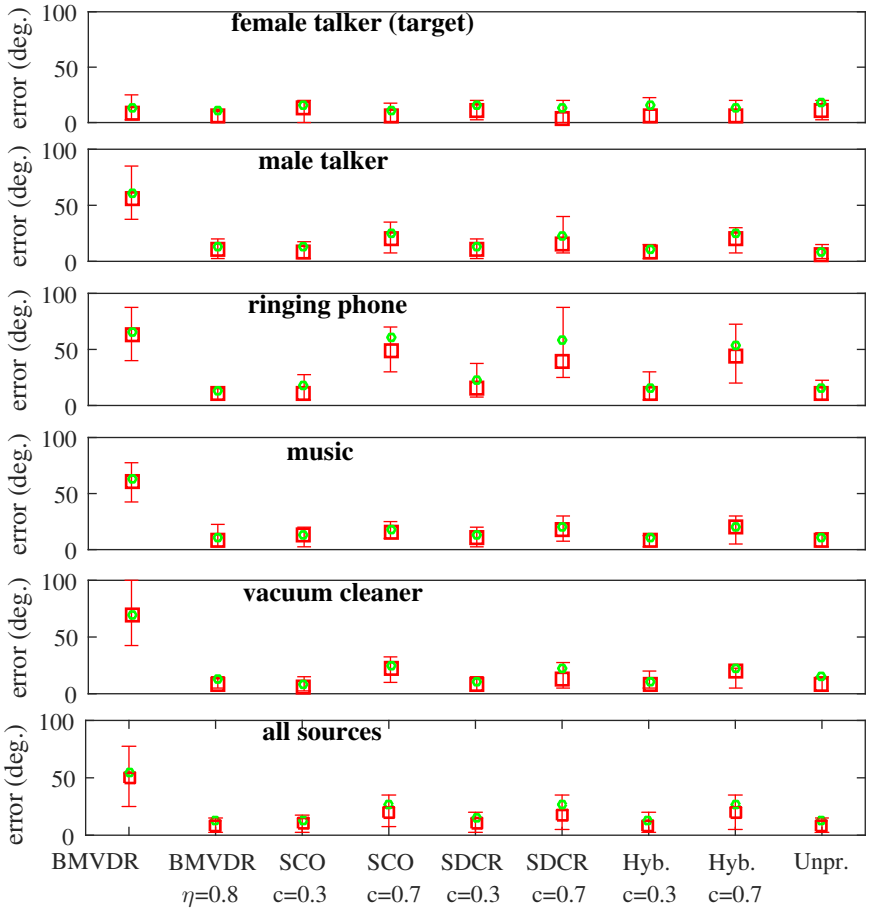

Figure 5.9: Experiment 2: Localization test measuring the localization error in degrees for all compared methods and point sources. The bottom figure is the average localization error over all point sources.

icantly different with each other. This means that even though in the instrumental measures we observed a not negligible difference in binaural-cue distortions between $c = 0.3$ and $c = 0.7$, the subjective evaluation contradicts that. For the mobile phone, the SDCR ($c = 0.3$), hybrid ($c = 0.3$), SCO ($c = 0.3$), BMVDR-η ($\eta = 0.8$) and unprocessed methods are all not significantly different, but are all significantly different with all the remaining methods. Furthermore, the SDCR ($c = 0.7$), hybrid ($c = 0.7$), SCO ($c = 0.7$) and BMVDR are not significantly different.

We can conclude from the above comparisons that the proposed methods do not cause significantly different binaural-cue distortions compared to the SCO method for the same c value and for all point sources in the acoustic scene. This means that

Table 5.1: Two-way ANOVA test with the point source and processing method as the two factors.

Source of variation	Sum Sq.	d.f.	Mean Sq.	F	p-value
Point-source(A)	165727.5	8	20715.9	39.77	8.1e-55
Proc.-method(B)	49935	4	12483.8	23.97	6.6e-19
AB	82551.8	32	2579.7	4.95	5.4e-17
Error	492196.6	945	520.8		
Total	790410.9	989			

even though we observed less binaural-cue distortions in the SCO method in Figs 5.2 and 5.7, compared to the proposed methods for the same c value, these differences are not perceptually important. However, recall that the proposed methods achieve a better noise reduction and predicted intelligibility compared to the SCO method. Thus, the proposed methods provide a better perceptual trade-off compared to the SCO method. Finally, note that the SCO, SDCR, hybrid for $c = 0.3$ and BMVDR-η for $\eta = 0.8$ methods are not statistically significantly different from the unprocessed scene for all point sources in the acoustic scene. This means that in all four methods the subjects managed (on average) to localize as good as in the unprocessed scene. However, unlike the unprocessed scene, all four methods improved noise reduction and predicted intelligibility.

5.5. Conclusion

We proposed two new suboptimal methods for approximately solving the non-convex relaxed binaural beamforming (RBB) optimization problem. Both methods are significantly computationally less demanding compared to the existing successive convex optimization (SCO) method. For each frequency bin, the SCO method requires to solve many more convex optimization problems in order to converge compared to the proposed methods. Specifically, the first proposed method, which is a semi-definite convex relaxation (SDCR) of the RBB problem, solves only one convex optimization problem per frequency bin. Apart from the computational advantage, the SDCR method also achieves in most cases a better trade-off between intelligibility and binaural-cue preservation than the SCO method. However, the SDCR method does not guarantee user-controlled upper bounded ITF error when the RATF vectors of the interferers are estimated accurately enough. This problem is solved by the second proposed method, which is a hybrid combination of the SDCR and SCO methods. This method guarantees user-controlled upper-bounded ITF error, and at the same time is computationally much less demanding than the SCO method. Finally, listening tests showed that all three methods achieve not significantly different localization errors for the same amount of binaural-cue error relaxation.

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process. Mag. **32**, 18 (2015).

[2] J. M. Kates, *Digital hearing aids* (Plural publishing, 2008).

[3] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, *Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,* IEEE Trans. Audio, Speech, Language Process. **23**, 2449 (2015).

[4] A. W. Bronkhorst, *The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,* Acta Acoustica **86**, 117 (2000).

[5] D. Marquardt, *DEVELOPMENT AND EVALUATION OF PSYCHOACOUSTICALLY MOTIVATED BINAURAL NOISE REDUCTION AND CUE PRESERVATION TECHNIQUES*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg (2015).

[6] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, S. van de Par, J. Jensen, and M. Guo, *Evaluation of binaural noise reduction methods in terms of intelligibility and perceived localization,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2018).

[7] 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. **5**, 529 (1997).

[8] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, *Microphone-array hearing aids with binaural output .II. A two-microphone adaptive system,* IEEE Trans. Speech Audio Process. **5**, 543 (1997).

[9] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[10] 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. **26**, 1261 (2018).

[11] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[12] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[13] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[14] S. Gannot, E. Vincet, S. Markovich-Golan, and A. Ozerov, *A consolidated perspective on multi-microphone speech enhancement and source separation,* IEEE Trans. Audio, Speech, Language Process. **25**, 692 (2017).

[15] 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).

[16] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, *Theoretical analysis of binaural multimicrophone noise reduction techniques,* IEEE Trans. Audio, Speech, Language Process. **18**, 342 (2010).

[17] H. Anton, *Elementary linear algebra* (John Wiley & Sons, 2010).

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

[19] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed. (North Oxford Academic, Oxford, 1983).

[20] L. Vandenberghe and S. Boyd, *Semidefinite programming,* SIAM review **38**, 49 (1996).

[21] F. Alizadeh and D. Goldfarb, *Second-order cone programming,* Mathematical programming **95**, 3 (2003).

[22] J. B. Allen, *Short-term spectral analysis, and modification by discrete Fourier transform,* IEEE Trans. Acoust., Speech, Signal Process. **25**, 235 (1977).

[23] H. Kayser, S. Ewert, J. Annemuller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, *Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,* EURASIP J. Advances Signal Process. **2009**, 1 (2009).

[24] *Cvx: Matlab software for disciplined convex programming.* (2008).

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, *An algorithm for intelligibility prediction of time-frequency weighted noisy speech,* IEEE Trans. Audio, Speech, Language Process. **19**, 2125 (2011).

[26] W. M. Hartmann, *How we localize sound,* Physics Today **52**, 24 (1999).

[27] D. J. Sheskin, *Parametric and nonparametric statistical procedures,* Chapman & Hall/CRC: Boca Raton, FL (2000).

6

Binaural Speech Enhancement with Spatial Cue Preservation Utilising Simultaneous Masking

This chapter is based on the article published as "Binaural Speech Enhancement with Spatial Cue Preservation Utilising Simultaneous Masking", by A.I. Koutrouvelis, J. Jensen, M. Guo, R.C. Hendriks and R. Heusdens in the Proceedings of the 25th European Signal Processing Conference (EUSIPCO), 2017.

The rapidly increasing communication capabilities between small portable devices make the notion of binaural noise reduction (BNR) [1] increasingly tractable for wireless collaborative hearing aids (HAs) [2]. BNR methods aim at acoustic noise suppression, using the microphones from both HAs, without altering the spatial impression of the acoustic scene.

Typically, BNR methods consist of two beamformers (one at the left and one at the right HA) and, optionally, a post-filter applied to the outputs of the two beamformers for further noise suppression [1]. The BNR methods can be roughly grouped into two main categories: a) methods that require estimates of the relative acoustic transfer functions (RATFs) of all present sources (e.g., [3–7]), and b) methods which require *only* the estimated RATF of the target (e.g., [8–11]). In this paper we focus on the second category of BNR methods mainly due to the practicality of only relying on the target RATF.

The binaural minimum variance distortionless response (BMVDR) beamformer [5] consists of two MVDR beamformers [12, 13] and requires only an estimate of the noise cross-power spectral density matrix and the RATF of the target. It provides the maximum noise reduction performance within the class of binaural linearly constrained distortionless minimum variance beamformers [5, 6]. However, this is at the cost of distorting the binaural cues of the interferers [5, 6], which will coincide with the binaural cues of the target after processing [5].

The BMVDR-N method, initially proposed in [8] and further investigated in [14], combines the output of the BMVDR with a portion of the noisy unprocessed signal to preserve the binaural cues of the noise. A slightly different approach was presented in [10], referred to as the selective binaural beamformer (SBB). This method uses either the BMVDR output *or* a suppressed version of the unprocessed noisy acoustic scene, depending on whether the target or the noise is dominant in a time-frequency (TF) tile. This classification of target-dominant and noise-dominant TF tiles is accomplished using an estimate of the *input* SNR.

All aforementioned approaches have in common that they intend to preserve the spatial cues of all sources without taking the notion into account that some sources are actually inaudible *after* processing. In this paper we introduce the idea of speech enhancement with binaural cue preservation only of the sources that are audible at the output of the filter. The general advantage of this approach is that degrees of freedom which in traditional approaches are assigned to cue preservation of sources, which turn out to be inaudible after processing (and hence masked) are now released and maybe assigned to noise reduction. More specifically we apply this concept to a modification of the SBB approach. Instead of using the input SNR, we use an estimate of the BMVDR output SNRs at left and right ears [15] for the binary classification. This allows us to better control the characteristics of the noise reaching the ears of the user. Moreover, the proposed method is better aligned with masking properties than the SBB method. If the noise, after processing with the BMVDR beamformer, is inaudible in a TF tile, there is no need to preserve its binaural cues in this specific TF tile and, therefore, the maximum possible noise reduction is achieved by applying the BMVDR. On the other hand, if the noise

after processing is audible, the binaural cue distortions introduced by the BMVDR may be audible and, therefore, a scaled version of the noisy acoustic scene is used instead.

6.1. Notation and Signal Model

We assume for convenience that the two HAs have an equal number of m microphones with $M = 2m$ microphones in total. Without any loss of generality we assume that there is a single target point source and one interferering point source present in the acoustic scene. Stacking all microphone frequency-domain elements into vectors, we have the following signal model for a single TF tile

$$\mathbf{y}(t,f) = \mathbf{x}(t,f) + \underbrace{\mathbf{n}(t,f) + \mathbf{v}(t,f)}_{\mathbf{z}(t,f)} \in \mathbb{C}^{M\times 1}, \tag{6.1}$$

where $\mathbf{y}(t,f)$, $\mathbf{x}(t,f)$, $\mathbf{n}(t,f)$, $\mathbf{v}(t,f)$ and $\mathbf{z}(t,f)$ are the noisy, target, interferer, background noise and overall noise vectors for the DFT bin f and time frame t, respectively. The 1-st and the M-th microphones are selected as reference microphones[1] and the corresponding elements of all vectors in Eq. (6.1) have subscripts L and R, respectively, for notational convenience. Note that $\mathbf{x}(t,f) = \mathbf{a}(t,f)s(t,f)$ and $\mathbf{n}(t,f) = \mathbf{b}(t,f)u(t,f)$, where $\mathbf{a}(t,f)$ and $\mathbf{b}(t,f)$ are the acoustic transfer functions (ATFs) of the target and the interferer, respectively, while $s(t,f)$ and $u(t,f)$ are the target signal and interfering signal at the original positions, respectively.

The BNR methods consists of two filters $\mathbf{w}_L(t,f)$, $\mathbf{w}_R(t,f) \in \mathbb{C}^{M\times 1}$ that are applied to the noisy vector $\mathbf{y}(t,f)$, obtaining the following two outputs

$$\hat{x}_L(t,f) = \mathbf{w}_L^H(t,f)\mathbf{y}(t,f), \quad \hat{x}_R(t,f) = \mathbf{w}_R^H(t,f)\mathbf{y}(t,f),$$

where $\mathbf{w}_L(t,f)$, and $\mathbf{w}_R(t,f)$ are estimated using all microphone recordings from both HAs.

6.1.1. Binaural Spatial Information Measures

The binaural spatial information for point sources is measured in terms of the interaural level differences (ILDs) and the interaural phase differences (IPDs). The input/output ILDs and IPDs of the interferer for a single TF tile are given by[2]

$$\mathrm{IPD}_{\mathbf{n}}^{\mathrm{in}} = \angle\frac{b_L}{b_R} \quad \text{and} \quad \mathrm{IPD}_{\mathbf{n}}^{\mathrm{out}} = \angle\frac{\mathbf{w}_L^H\mathbf{b}}{\mathbf{w}_R^H\mathbf{b}}, \tag{6.2}$$

$$\mathrm{ILD}_{\mathbf{n}}^{\mathrm{in}} = \left|\frac{b_L}{b_R}\right|^2 \quad \text{and} \quad \mathrm{ILD}_{\mathbf{n}}^{\mathrm{out}} = \left|\frac{\mathbf{w}_L^H\mathbf{b}}{\mathbf{w}_R^H\mathbf{b}}\right|^2. \tag{6.3}$$

[1] The BNR methods aim at preserving the binaural cues of all sources with respect to the reference microphones.

[2] These measures/quantities as well as other measures/quantities introduced in the sequel of the paper are time-frequency varying, however for notational convenience the TF indices (t,f) in some occasions are omitted.

Similar expressions to Eqs. (6.2) and (6.3) exist for the target source. In addition, we quantify binaural spatial characteristics of the background noise in terms of the input and output magnitude square coherence (MSC) [5, 14] given by

$$\mathrm{MSC}^{\mathrm{in}} = \left| \frac{c_{LR}^{\mathrm{in}}}{\sqrt{\left(c_{LL}^{\mathrm{in}}\right)\left(c_{RR}^{\mathrm{in}}\right)}} \right|^2, \quad \mathrm{MSC}^{\mathrm{out}} = \left| \frac{c_{LR}^{\mathrm{out}}}{\sqrt{\left(c_{LL}^{\mathrm{out}}\right)\left(c_{RR}^{\mathrm{out}}\right)}} \right|^2, \tag{6.4}$$

respectively, where $c_{LR}^{\mathrm{in}} = \mathbf{e}_L^T \mathbf{P}_{\mathbf{v}} \mathbf{e}_R$, $c_{LL}^{\mathrm{in}} = \mathbf{e}_L^T \mathbf{P}_{\mathbf{v}} \mathbf{e}_L$, $c_{RR}^{\mathrm{in}} = \mathbf{e}_R^T \mathbf{P}_{\mathbf{v}} \mathbf{e}_R$, $c_{LR}^{\mathrm{out}} = \mathbf{w}_L^H \mathbf{P}_{\mathbf{v}} \mathbf{w}_R$, $c_{LL}^{\mathrm{out}} = \mathbf{w}_L^H \mathbf{P}_{\mathbf{v}} \mathbf{w}_L$, $c_{RR}^{\mathrm{out}} = \mathbf{w}_R^H \mathbf{P}_{\mathbf{v}} \mathbf{w}_R$, $\mathbf{P}_{\mathbf{v}}$ is the cross-power spectral density matrix of the background noise for a single TF tile, $\mathbf{e}_L^T = [1\ 0, \cdots, 0]$ and $\mathbf{e}_R^T = [0, \cdots, 0\ 1]$. A desired property of a BNR method is to have small MSC, IPD and ILD errors, defined as

$$\mathrm{MSC}^{\mathrm{error}}(t,f) = \left| \mathrm{MSC}^{\mathrm{out}}(t,f) - \mathrm{MSC}^{\mathrm{in}}(t,f) \right|, \tag{6.5}$$

$$\mathrm{IPD}_{\mathbf{n}}^{\mathrm{error}}(t,f) = \left| \mathrm{IPD}_{\mathbf{n}}^{\mathrm{out}}(t,f) - \mathrm{IPD}_{\mathbf{n}}^{\mathrm{in}}(t,f) \right| / \pi, \tag{6.6}$$

$$\mathrm{ILD}_{\mathbf{n}}^{\mathrm{error}}(t,f) = \left| \mathrm{ILD}_{\mathbf{n}}^{\mathrm{out}}(t,f) - \mathrm{ILD}_{\mathbf{n}}^{\mathrm{in}}(t,f) \right|. \tag{6.7}$$

It is only relevant to measure the aforementioned spatial errors of the residual noise in a TF tile, (t,f), when the residual noise is audible at the output. To reflect to which extent the processed noise is masked by the processed target we apply a weighting to the ILD, IPD and MSC errors.

The weights are computed based on the simultaneous masking principle [16] as follows. First the k-th critical band SNR (CBSNR) output with respect to the left and right reference microphones are computed. The left CBSNR is given by

$$\mathrm{CBSNR}_{k,L}(t) = \frac{\sum_{f \in \mathrm{CB}_k} \mathbf{w}_L^H(t,f) \mathbf{P}_{\mathbf{x}}(t,f) \mathbf{w}_L(t,f)}{\sum_{f \in \mathrm{CB}_k} \mathbf{w}_L^H(t,f) \mathbf{P}_{\mathbf{z}}(t,f) \mathbf{w}_L(t,f)}, \tag{6.8}$$

where CB_k denotes the index set of DFT bins corresponding to the k-th critical band, and $\mathbf{P}_{\mathbf{x}}(t,f)$ is the cross-power spectral density matrix of the target at the TF tile (t,f). A similar expression exists for the right CBSNR, $\mathrm{CBSNR}_{k,R}(t)$. Then, the weights associated with the k-th critical band are computed. Specifically, the weights for the left reference microphone are given by

$$\phi_{k,L}(t) = \begin{cases} 1, & \mathrm{CBSNR}_{k,L}(t) \leq \lambda \\ 1 - \frac{\mathrm{CBSNR}_{k,L}(t) - \lambda}{\rho - \lambda}, & \lambda < \mathrm{CBSNR}_{k,L}(t) < \rho \\ 0, & \mathrm{CBSNR}_{k,L}(t) \geq \rho \end{cases}, \tag{6.9}$$

where $\lambda = -4$ dB and $\rho = 24$ dB are the noise-tone and tone-noise masking thresholds [16]. If $\mathrm{CBSNR}_{k,L}(t) \geq 24$, the target masks completely the noise at the left reference microphone in the k-th critical band, while if $\mathrm{CBSNR}_{k,L}(t) \leq -4$, the noise completely masks the target [16]. The weights at the right reference microphone are computed as in Eq. (6.9), but using $\mathrm{CBSNR}_{k,R}(t)$ instead of $\mathrm{CBSNR}_{k,L}(t)$.

The average masking-weighted spatial information error measures for the left reference microphone are defined as

$$\text{AvMSC}_L^{\text{error}} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{N}\phi_{k,L}(t)\sum_{f\in\text{CB}_k}\text{MSC}^{\text{error}}(t,f)}{\sum_{t=1}^{T}\sum_{k=1}^{N}\sum_{f\in\text{CB}_k}\phi_{k,L}(t)},$$

$$\text{AvIPD}_L^{\text{error}} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{N}\phi_{k,L}(t)\sum_{f\in\text{CB}_k}\text{IPD}_{\mathbf{n}}^{\text{error}}(t,f)}{\sum_{t=1}^{T}\sum_{k=1}^{N}\sum_{f\in\text{CB}_k}\phi_{k,L}(t)},$$

$$\text{AvILD}_L^{\text{error}} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{N}\phi_{k,L}(t)\sum_{f\in\text{CB}_k}\text{ILD}_{\mathbf{n}}^{\text{error}}(t,f)}{\sum_{t=1}^{T}\sum_{k=1}^{N}\sum_{f\in\text{CB}_k}\phi_{k,L}(t)},$$

with T the number of time-frames and N the number of critical bands. Similar expressions exist for the right reference microphone.

6.2. Proposed Method

Similarly to the SBB method [10], the proposed method consists of two processing phases: a) the classification phase of TF tiles into target-dominant and noise-dominant, and b) the enhancement phase where the BMVDR is applied to the target dominant TF-tiles, while in the noise-dominant TF tiles a scaled (with $0 \leq g \leq 1$) version of the noisy signal is used in both HAs. Let the left and right input narrowband SNRs (NBSNRs) be given by [15]

$$\eta_L^{\text{in}} = \frac{\mathbf{e}_L^T\mathbf{P}_{\mathbf{x}}\mathbf{e}_L}{\mathbf{e}_L^T\mathbf{P}_{\mathbf{z}}\mathbf{e}_L}, \quad \eta_R^{\text{in}} = \frac{\mathbf{e}_R^T\mathbf{P}_{\mathbf{x}}\mathbf{e}_R}{\mathbf{e}_R^T\mathbf{P}_{\mathbf{z}}\mathbf{e}_R}, \tag{6.10}$$

respectively. The left and right BMVDR output NBSNRs are given by [15]

$$\eta_L^{\text{out}} = \eta_L^{\text{in}}\left(\mathbf{a}_L^H\mathbf{P}_L^{-1}\mathbf{a}_L\right), \quad \eta_R^{\text{out}} = \eta_R^{\text{in}}\left(\mathbf{a}_R^H\mathbf{P}_R^{-1}\mathbf{a}_R\right), \tag{6.11}$$

respectively, and $\mathbf{a}_L = (1/a_L)\mathbf{a}$, $\mathbf{a}_R = (1/a_R)\mathbf{a}$, $\mathbf{P}_L^{-1} = P_{\mathbf{z},(1,1)}\mathbf{P}_{\mathbf{z}}^{-1}$, and $\mathbf{P}_R^{-1} = P_{\mathbf{z},(M,M)}\mathbf{P}_{\mathbf{z}}^{-1}$, where $P_{\mathbf{z},(1,1)}$ and $P_{\mathbf{z},(M,M)}$ are the first and last diagonal elements, respectively, of $\mathbf{P}_{\mathbf{z}}$. The filters of the proposed method at the left and right HAs for a single TF tile are given by

$$\mathbf{w}_{\text{Prop.},L} = \begin{cases} \mathbf{w}_{\text{MV},L}, & \eta_L^{\text{out}} \geq \tau, \text{ and } \eta_R^{\text{out}} \geq \tau \\ g\mathbf{e}_L, & \text{otherwise} \end{cases}, \tag{6.12}$$

$$\mathbf{w}_{\text{Prop.},R} = \begin{cases} \mathbf{w}_{\text{MV},R}, & \eta_L^{\text{out}} \geq \tau, \text{ and } \eta_R^{\text{out}} \geq \tau \\ g\mathbf{e}_R, & \text{otherwise} \end{cases}, \tag{6.13}$$

with $\mathbf{w}_{\text{MV,L}}$ and $\mathbf{w}_{\text{MV,R}}$ the left and right BMVDR filters, respectively, η_L^{out} and η_R^{out} the output NBSNRs at the left and right reference microphones, respectively, and τ the threshold value which is fixed over frequency and time. The BMVDR filters are given by [5]

$$\mathbf{w}_{\text{MV},L} = \frac{\mathbf{P}_{\mathbf{z}}^{-1}\mathbf{a}a_L^*}{\mathbf{a}^H\mathbf{P}_{\mathbf{z}}^{-1}\mathbf{a}}, \quad \mathbf{w}_{\text{MV},R} = \frac{\mathbf{P}_{\mathbf{z}}^{-1}\mathbf{a}a_R^*}{\mathbf{a}^H\mathbf{P}_{\mathbf{z}}^{-1}\mathbf{a}}, \tag{6.14}$$

with $\mathbf{P_z}$ the cross-power spectral density matrix of the total noise, and a_L and a_R the two reference elements of $\mathbf{a}$.

6.2.1. Improvements of the SBB method

In our evaluation, we compare our proposed method to an improved version of the SBB method. The improvements consider two aspects. First, unlike the original SBB [10] which uses only one input NBSNR in the classification stage, our implementation of SBB uses both η_L^{in} and η_R^{in} in order to guarantee target dominance in both ears. Secondly, in the original SBB method [10], the scaling parameter g was selected as

$$g = \min\left(\frac{1}{\mathbf{w}_{\text{MV,L}}^H \mathbf{P_z} \mathbf{w}_{\text{MV,L}}}, \frac{1}{\mathbf{w}_{\text{MV,R}}^H \mathbf{P_z} \mathbf{w}_{\text{MV,R}}}\right). \tag{6.15}$$

Computing g with Eq. (6.15) might, in some situations, boost the noise. Instead, in this paper we select g as

$$g = \min\left(\sqrt{\frac{\mathbf{w}_{\text{MV,L}}^H \mathbf{P_z} \mathbf{w}_{\text{MV,L}}}{\mathbf{e}_L^T \mathbf{P_z} \mathbf{e}_L}}, \sqrt{\frac{\mathbf{w}_{\text{MV,R}}^H \mathbf{P_z} \mathbf{w}_{\text{MV,R}}}{\mathbf{e}_R^T \mathbf{P_z} \mathbf{e}_R}}\right), \tag{6.16}$$

in both the proposed and the SBB methods.

As in [10] we use an average g (computed across the noise-dominated DFT bins) for each time-frame for both the proposed and the SBB methods to mitigate coloration of the residual noise. Hence, g is time-varying but constant over frequency.

6.2.2. Basic Principle

There are two main reasons to use η_L^{out} and η_R^{out} (the proposed method) instead of η_L^{in} and η_R^{in} (the SBB method), in the classification stage. First, the main goal of the proposed method is to achieve the maximum possible noise suppression, without altering the binaural cues of the *audible* processed noise. Therefore, if the processed noise is masked by the processed target, there is no reason to preserve any binaural cues of the noise and, then, the largest possible noise reduction is achieved by using the BMVDR output.

Secondly, judging whether the noise is masked by the target is easier if this is done *after* processing (based on η_L^{out} and η_R^{out}) than *before* processing (based on η_L^{in} and η_R^{in}). This is because, after processing, the binaural cues of the noise coincides with the binaural cues of the target and one can use the monaural simultaneous masking principle described in [16]. Moreover, after processing, masking becomes independent of the spatial layout of the sources in the acoustic scene.

Based on the aforementioned two facts, the proposed method will be more robust than the SBB method to changing acoustical scenarios assuming that a fixed threshold τ is used in both methods. This will be shown in Sections 6.2.3, 6.2.4.

6.2.3. Example 1: Point Noise Source

Fig. 6.1 demonstrates the difference between the proposed method and the SBB method, for a synthetic speech shaped target source in the front (0 degrees), an

interfering speech shaped noise source to the right (-80 degrees) and a small amount of microphone self noise. Figs. 1(a) and 1(b) depict the estimated input and output NBSNRs at the left and right reference microphones, respectively. Figs. 1(c) and 1(d) show the $\text{AvIPD}_L^{\text{error}}$ and $\text{AvIPD}_R^{\text{error}}$ of the interferer vs. the output segmental SNR (SSNR) for the two methods, respectively, over a threshold value, τ, ranging from -50 dB to 50 dB with a step-size of 0.5 dB. Figs. 1(e) and 1(f) show the $\text{AvILD}_L^{\text{error}}$ and $\text{AvILD}_R^{\text{error}}$ of the interferer vs the output SSNR, respectively, for the same range of τ values. The output SSNR at the left reference microphone is defined as

$$\text{SSNR}_L^{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} 10\log_{10}\frac{||\mathbf{q}_{t,L}||_2^2}{||\hat{\mathbf{q}}_{t,L} - \mathbf{q}_{t,L}||_2^2}, \tag{6.17}$$

with $\mathbf{q}_{t,L}$ the time-frame t of the clean target signal at the left reference microphone, $\hat{\mathbf{q}}_{t,L}$ its estimate. A similar expression holds for the $\text{SSNR}_R^{\text{out}}$.

Let us examine four interesting τ values for this specific example. If $\tau > 29$ dB, both SBB and the proposed method will not achieve any noise suppression, but they will simply scale the noisy signal by g. This is because, $\eta_L^{\text{in}}, \eta_R^{\text{in}}, \eta_L^{\text{out}}, \eta_R^{\text{out}} < 29$ dB for all frequency bins. Thus, the values of the performance curves in Figs. 1(c,d,e,f) corresponding to $\tau > 29$ dB will be in the left bottom corner.

If $\tau = 22.5$ dB, most parts of the $\eta_L^{\text{out}}, \eta_R^{\text{out}}$ curves will be above $\tau = 22.5$ dB, while all the frequency bins of the curves $\eta_L^{\text{in}}, \eta_R^{\text{in}}$ will be below $\tau = 22.5$ dB. This means that the proposed method will achieve some noise reduction, while the SBB method will not suppress the noise at all. Moreover, since $\tau = 22.5$, the processed noise in all the frequency bins that correspond to $\eta_L^{\text{out}} > 22.5, \eta_R^{\text{out}} > 22.5$ will be almost inaudible and, therefore, the weighted average binaural cue errors will be approximately zero. In conclusion, a) none of the methods caused any audible binaural cue errors, b) the proposed method achieved some noise reduction, while the SBB method did not achieve any noise reduction. In Figs. 1(c,d,e,f), the performances for $\tau = 22.5$ dB are shown with a red $\square$ marker and a blue $\circ$ marker for the proposed method and the SBB method, respectively.

For $\tau = 2$ dB there will be some frequency bins (in the region 7-8 kHz) of η_L^{in}, and η_R^{in} that will be above $\tau = 2$ dB as well. The number of these frequency bins will be much less than the number of the frequency bins of $\eta_L^{\text{out}}, \eta_R^{\text{out}}$ that will be above $\tau = 2$ dB. Thus, the proposed method will achieve larger amount of noise reduction. Both methods will cause audible binaural cue errors for $\tau = 2$.

For values $\tau < -8$ dB both methods will have identical performance, i.e., both methods will apply the BMVDR beamformer to all frequency bins. This corresponds to the top right corner (marked with a black star), in Figs. 1(c,d,e,f).

It is clear that the proposed method achieves a better output SSNR than the SBB for many values of $\text{AvILD}_L^{\text{error}}$, $\text{AvILD}_R^{\text{error}}$, $\text{AvIPD}_L^{\text{error}}$ and $\text{AvIPD}_R^{\text{error}}$ errors, in this acoustic scenario.

6.2.4. Example 2: Diffuse Noise

Similarly to Fig. 6.1, Fig. 6.2 shows the difference between the proposed method and the SBB method when there is a target speech shaped source at the front (0 degrees),

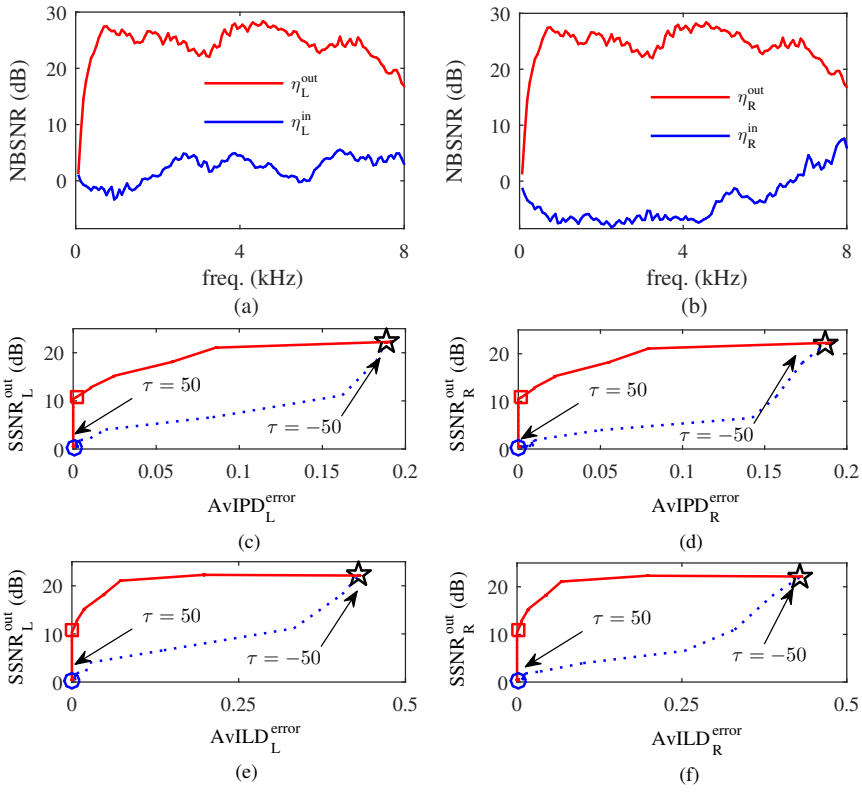

Figure 6.1: Simulation example 1 comparing the proposed (red) with the BSS (blue) method and the BMVDR (black star). For $\tau = 22.5$, the performance of the proposed and the BSS method is illustrated with a red □ marker and a blue ○ marker, respectively.

a diffuse noise field and a small amount of microphone self noise. As mentioned in Section 6.1.1, a proper measure for binaural spatial distortions in diffuse noise fields is the AvMSC$_L^{\text{error}}$ and AvMSC$_R^{\text{error}}$ errors. Therefore, in Fig. 6.2, we use the AvMSC$_L^{\text{error}}$ AvMSC$_R^{\text{error}}$ errors to show the performance difference between the two methods.

It is worth noting that in Figs. 6.2(a,b) the curves η_L^{out} and η_L^{in} have very similar structure, i.e., they are approximately vertically shifted. The same applies also for the curves η_R^{out} and η_R^{in}. This means the two methods will give more or less identical SSNR for any AvMSC error. This can be observed in Figs. 6.2(c,d), were the performance curves are very similar.

6.3. Simulations

In this section, the proposed method is compared with the SBB method [10] for $\tau = -50 : 0.5 : 50$ dB, and the BMVDR-N method [8, 14] with $N = 0 : 0.1 : 1$. The comparison is done in two different noisy acoustic scenarios. In the first

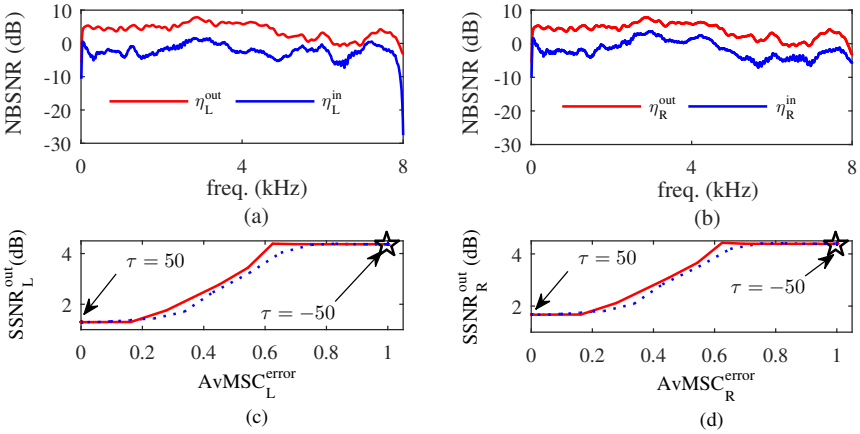

Figure 6.2: Simulation example 2 comparing the proposed (red) with the BSS (blue) method and the BMVDR (black star).

scenario the noise component is a single interferer (a male talker) on the right of the HA user (at -80 degrees). In the second scenario the noise component is diffuse noise which is created using different speech shape noise realizations from 72 different angles around the head. In both scenarios, the target is a female talker positioned in the front (i.e., 0 degrees) of the HA user, and the microphone self-noise (in all microphones) is 50 dB smaller with respect to the target signal at the left reference microphone. In both simulated scenarios, we used the anechoic head impulse responses from [17] to simulate both the point sources and the diffuse noise. The female and male talker point sources were placed 0.8 m from the head, while the point sources that are for the diffuse noise are placed 3 m from the center of the head. All simulated signals have a duration of 14 seconds in which the first 4 seconds the noise is active only. The BMVDR filters used the true $\mathbf{a}$ and an estimate of $\mathbf{P_z}$ using a perfect VAD. The $\eta_L^{\text{in}}, \eta_R^{\text{in}}, \eta_L^{\text{out}}$, and η_R^{out} are estimated using the method in [15] using a perfect VAD and the true $\mathbf{a}$. We used an overlap and add methodology for processing the signals with a frame size of 10 ms and overlap 50%. The sampling frequency is 16 kHz.

Figs. 6.3 and 6.4 show a performance comparison for the first and second simulated acoustic scenario, respectively. The gap in performance, between the SBB method and the proposed method, depends on the input/output NBSNR structure and type of the noise field as discussed in Sections 6.2.3 and 6.2.4. For the first simulated acoustical scenario, the proposed method achieves a higher noise reduction performance (as measured with SSNR) for most binaural spatial error values. This is due to the big difference of the structure of the output NBSNR compared to the structure of the input NBSNR as explained in Section 6.2.3. However, this is not the case for the second simulated scenario as expected (see Section 6.2.4), since the structure of the output NBSNR is very similar with the structure of the input NBSNR. Moreover, note that the BMVDR-N method has the worst performance over the other two methods in all acoustic scenarios for most N values.

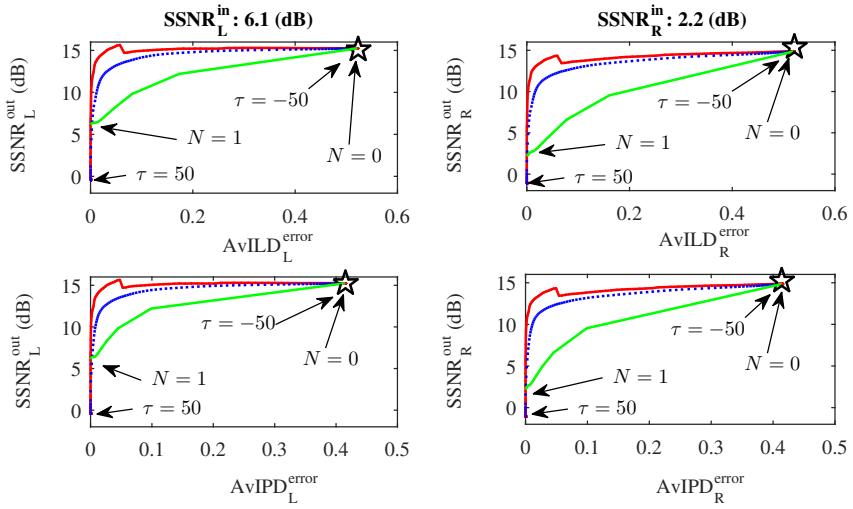

Figure 6.3: Scenario 1 comparing the proposed (red) with the BSS (blue) method, the BMVDR-N (green) and the BMVDR (black star).

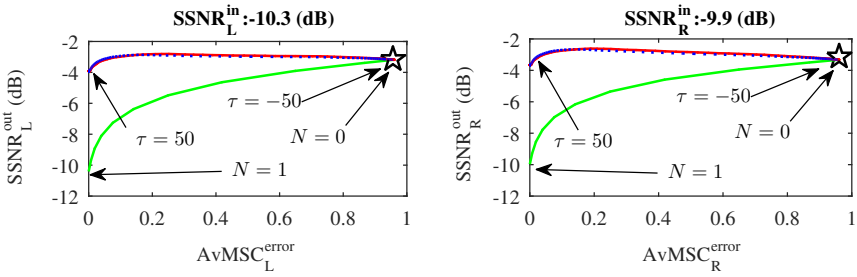

Figure 6.4: Scenario 2 comparing the proposed (red) with the BSS (blue) method, the BMVDR-N (green) and the BMVDR (black star).

6.4. Conclusion

We proposed a modified version of the selective binaural beamformer (SBB) approach. The proposed method differs from the SBB approach in the classification stage of the time-frequency (TF) tiles. It uses the output SNR for labeling the TF tiles either to target-dominant or noise-dominant. This modification is better aligned with the simultaneous masking principle. Furthermore, it was experimentally shown that in some acoustical scenarios the proposed method provides larger amount of noise reduction than BSS for the same binaural spatial distortions.

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process.

Mag. **32**, 18 (2015).

[2] J. M. Kates, *Digital hearing aids* (Plural publishing, 2008).

[3] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, *Theoretical analysis of binaural multimicrophone noise reduction techniques,* IEEE Trans. Audio, Speech, Language Process. **18**, 342 (2010).

[4] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[5] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, *Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,* IEEE Trans. Audio, Speech, Language Process. **23**, 2449 (2015).

[6] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[7] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[8] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[9] J. Thiemann, M. Müller, and S. van de Par, *A binaural hearing aid speech enhancement method maintaining spatial awareness for the user,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2014) pp. 321–325.

[10] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van der Par, *Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene,* EURASIP J. Advances Signal Process. (2016).

[11] H. As'ad, M. Bouchard, and H. Kamkar-Parsi, *Perceptually motivated binaural beamforming with cues preservation for hearing aids,* in *IEEE Canadian Conf. Electrical and Computer Engineering (CCECE)* (2016).

[12] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. IEEE **57**, 1408 (1969).

[13] B. D. Van Veen and K. M. Buckley, *Beamforming: A versatile approach to spatial filtering,* IEEE ASSP Mag. **5**, 4 (1988).

[14] D. Marquardt, *DEVELOPMENT AND EVALUATION OF PSYCHOACOUSTICALLY MOTIVATED BINAURAL NOISE REDUCTION AND CUE PRESERVATION TECHNIQUES*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg (2015).

[15] J. Jensen and M. S. Pedersen, *Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2015) pp. 5728–5732.

[16] T. Painter and A. Spanias, *Perceptual coding of digital audio,* Proceedings of the IEEE **88**, 451 (2000).

[17] H. Kayser, S. Ewert, J. Annemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, *Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,* EURASIP J. Advances Signal Process. **2009**, 1 (2009).

7

Evaluation of Binaural Noise Reduction Methods in Terms of Intelligibility and Perceived Localization

This chapter is based on the article published as "Binaural Speech Enhancement with Spatial Cue Preservation Utilising Simultaneous Masking", by A.I. Koutrouvelis, R.C. Hendriks and R. Heusdens, S. van de Paar, J. Jensen, M. Guo in the Proceedings of the 26th European Signal Processing Conference (EUSIPCO), 2018.

Binaural hearing-aid (HA) systems [1] consist of two wirelessly connected and collaborative HA devices with at least one microphone per device. In contrast, bilateral HA systems [2, 3] consist of independently working HAs. The binaural HAs can typically use a larger microphone array than bilateral HAs and, therefore, have more degrees of freedom for the beamformer. These degrees of freedom might be used to obtain a better noise reduction, or, to preserve the binaural cues of sound sources in the acoustic scene [3].

An important component in a binaural HA system is the binaural multi-microphone speech enhancement algorithm, which aims to enhance the intelligibility of the target speech signal, while at the same time to preserve the binaural cues of the acoustic scene after processing [3]. Typically, binaural multi-microphone speech enhancement methods show a trade-off between noise reduction and binaural-cue preservation. Existing binaural multi-microphone speech enhancement methods can be roughly categorized into two main groups: the spatial filtering methods (e.g., [4–8]) and the spatio-temporal filtering methods (e.g., [9–16]). The latter group typically provides a larger amount of noise reduction than spatial filtering methods, at the expense of target distortions at the output of the filter.

Only a few studies exist (e.g., [2, 15, 17, 18]) that evaluate the perceptual performance (such as intelligibility and localization) of binaural speech enhancement methods. In contrast, most studies evaluate performance using instrumental measures, e.g., predicting intelligibility (e.g., by means of STOI [19] or DBSTOI [20]) or localization accuracy (e.g., by means of interaural level and time differences errors [13], or other measures such as the ones presented in [21, 22]). Although these measures correlate well with localization and intelligibility, not all aspects of localization and intelligibility are well understood or incorporated in these measures. To understand the real trade-off between intelligibility improvement and localization accuracy, listening tests are still required.

In this paper, we evaluate two methods recently proposed in [8] and [16] by means of an intelligibility test and a localization test, and compare them with the binaural minimum variance distortionless (BMVDR) method [3]. In addition, we compare with an oracle based method [18], to get an idea of the intelligibility and perceived localization if perfect knowledge would be available. The BMVDR method provides the maximum noise reduction among all linear spatial filters, while severely distorting the binaural cues of all interferers [3]. We report the intelligibility and localization scores of self-reported normal-hearing people in several acoustic scenes.

The spatio-temporal filtering method proposed in [16] preserves the binaural cues by a binary classification of all frequency bins into target or noise-dominant bins. The classification is based on the output SNR that results by applying the BMVDR to all frequency bins. The target-dominant time-frequency bins are processed with the BMVDR, while the noise-dominant time-frequency bins are replaced with a scaled version of the corresponding unprocessed time-frequency bins.

The spatial filtering method proposed in [8] uses additional inequality constraints in the BMVDR optimization problem to preserve the binaural cues of all interferes. The inequality constraints are functions of anechoic pre-determined head-related

transfer functions (HRTFs), which are considered as known and are acoustic-scene-independent, but user-dependent [8].

Section 7.1 reviews the binaural speech enhancement methods that we evaluate. Section 7.2 shows the evaluation procedure and its results. Section 7.3 gives concluding remarks.

7.1. Overview of the Evaluated Methods

In this section, we briefly review the binaural speech enhancement methods that we evaluate in this paper. For more details, the reader is referred to the associated papers.

7.1.1. BMVDR

The BMVDR spatial filter [3] provides the maximum noise reduction compared to all the other spatial filters. It preserves the binaural cues of the target, but distorts the binaural cues of all other sound sources, and makes them identical to the target's binaural cues. The BMVDR consists of two MVDR spatial filters [23] sharing the same microphone array, but using two different reference microphones, one on each HA. The two optimization problems, associated with the two MVDR spatial filters, minimize the total output noise power under the constraints that the target signal is preserved without any distortion at the two reference microphones. As such the binaural cues of the target signal are preserved, but the binaural cues of the interferers are not, since there are no constraints for them in the optimization problems.

7.1.2. Relaxed Binaural LCMV with Pre-determined HRTFs

The relaxed binaural linearly constrained minimum variance (LCMV) with pre-determined HRTFs is a spatial filtering method introduced in [7, 8]. This method uses additional inequality constraints in the BMVDR optimization problem to preserve the binaural cues of pre-selected azimuths and/or elevations around the head [8]. The inequality constraints can be relaxed as desired using a relaxation parameter, $0 \leq c \leq 1$. The maximum amount of relaxation (i.e., $c = 1$) results in the BMVDR filter as a special case. The trade-off between noise suppression and binaural-cue preservation of this method depends not only on c, but also on the number of pre-determined HRTFs. In this paper, we only vary the c-value and not the number and locations of the pre-determined HRTFs. More specifically, we always use anechoic pre-determined HRTFs (from the database in [24]) associated with 24 uniformly spaced locations in the horizontal plane on a circle around the head with a distance of 3 m from the center of the head.

7.1.3. BMVDR with Thresholding

The BMVDR with thresholding method is a spatio-temporal filtering method introduced in [16]. First, the BMVDR filter is applied to all time-frequency bins and, next, the output narrow-band SNR, of all these enhanced time-frequency bins, is estimated. A time-frequency bin is considered target-dominant, if the output

SNR of a time-frequency bin is above a certain threshold τ. Otherwise, the time-frequency bin is considered as noise-dominant. The noise-dominant time-frequency bins are suppressed identically, so that the interaural time and level differences are not changed in order to preserve the binaural cues of the noise. In particular, if the residual noise is inaudible after applying the BMVDR method, its binaural cues need not be preserved and, therefore, maximum possible noise suppression is achieved. If the noise in some time-frequency bins dominates the target after processing, the BMVDR output is not beneficial and, thus, the BMVDR output is replaced by a scaled-down version of the unprocessed scene to suppress the interferers and preserve the binaural cues of the acoustic scene. Since this scaling reduces both the target and noise components, the target signal will be distorted.

7.1.4. Ideal Binaural Target Enhancement

This is an oracle-based method that consists of the unprocessed acoustic scene with an SNR equal to the SNR output of the BMVDR method [18]. This method achieves the same amount of noise suppression as the BMVDR while perfectly preserving the binaural cues of the complete acoustic scene.

7.2. Experiments

To evaluate the methods presented in Section 7.1, we conducted an intelligibility test, which measures the 50% speech reception threshold (SRT), and a localization test, which measures the binaural localization error of the dominant point sources in the acoustic scene. Both tests are divided into two different phases; a parameter selection phase and a testing phase. The acoustic scenes in the testing phase are different from the one in the parameter selection phase. This is done to examine the robustness in different acoustic scenes with respect to the chosen parameter settings. The main purpose of the parameter selection phase is to obtain the c and τ parameters for the relaxed binaural LCMV and the BMVDR with thresholding methods, respectively, to be used in the testing phase. The testing phase examines the performance of all methods in the remaining two acoustic scenes.

We used Beyerdynamic DT 990 PRO 250 OHM headphones for the listening tests. The average sound level of the total noise that was played via the headphones was fixed to 65 dB SPL and the target level was varied to achieve a certain SNR.

For convenience, we use the following acronyms for the compared methods in the following figures and tables: relaxed binaural LCMV (RBLCMV(c)), binaural MVDR (BMVDR), BMVDR with thresholding (BMVDR (τ)), ideal binaural target enhancement (IBTE), and unprocessed scene (UNPR).

7.2.1. Generation of Audio Signal Database

For both listening tests we created a database of unprocessed and processed 2-channel binaural signals with SNRs ranging from -28 dB to 10 dB. The unprocessed HA signals were computed using the behind-the-ear impulse response database in [24]. For the multi-microphone binaural speech enhancement methods we used the front and middle microphones from each HA to create an array of 4 microphones.

Table 7.1: Summary of acoustic scenes.

acoustic scene	point source position (degrees)			diffuse noise	recording environment	mic. noise
	female talker	male talker	music			
AC1	0	-30	90	cafeteria noise	cafeteria	yes
AC2	-30	90	-90	cafeteria noise	cafeteria	yes
AC3	0	-45	60	–	office	yes

After processing, we saved the 2-channel binaural output signals corresponding to the reference microphones.

We used as the target signal randomly selected Dutch-spoken sentences with a duration of about 2 s from a female talker, taken from the database in [25]. We padded these sentences at the beginning and at the end with extra zeros such that a length of 4 s was obtained and the spoken sentence was exactly temporally centered within the masking noise as shown in Fig. 7.1. This was done in order to avoid confusion of the listener due to simultaneous initiation of all sources.

We used four different noise types to simulate the acoustic scenes: a music signal, a randomly selected English-spoken sentence from a male talker taken from the TIMIT database [26], a diffuse cafeteria noise taken from the database in [24], and microphone-self noise. We also used three different acoustic scenes, which we denote as AC1, AC2 and AC3. Table 7.1 summarizes the acoustical sources and their locations in all acoustic scenes. Note that AC1 was used for parameter selection, while AC2 and AC3 were used for the testing phase.

The female and male talkers' signals were zero-padded to have an equal length of 4 s. For the music sound source, a 4 s fragment was extracted randomly per sentence from an approximately 5 minutes long music piece. All three noise contributions were set to have equal average power at the two reference microphones, making all disturbances equally important in the acoustic scene. The input SNR, defined as the target power with respect to the total noise power, was computed by concatenating the left and right reference microphone recordings of the target and the noise signals. The sampling frequency of all signals was set to 16 kHz.

7.2.2. Subjects

In the parameter selection phase, we used 5 native speakers of Dutch for the intelligibility test, and 5 non-native speakers of Dutch for the localization test. In the testing phase, we used 14 native speakers of Dutch for the intelligibility test, and 15 non-native speakers of Dutch for the localization test. All subjects from the parameter selection phase participated in the testing phase as well. All subjects were self-reported as normal-hearing and their age range was 20-36 years.

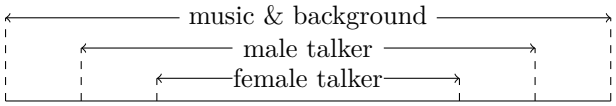

Figure 7.1: Time duration of each source signal. The background signal is a cafeteria background noise and is present only in AC1 and AC2.

7.2.3. Intelligibility Test

The target sentences (not necessarily meaningful) were part of a matrix test consisting of 5 words each, with the correct grammatical structure name, verb, number, adjective and noun. The sentences and the noise realizations were randomly selected from the database. Using a graphical user interface (GUI), the listeners had access to a 10×5 matrix with each column consisting of the 10 candidate words used to construct the sentences. The sentences were played only once to the subjects, after which they had to select from each column the word that was understood. We used the one-down-one-up adaptive staircase method [27] to find the SRT-50 scores (i.e., the SNR at which the subject scored 50% correct) for each method and subject. The adaptive track started with an initial SNR of 10 dB and an initial step-size of 8 dB. For each new reversal the step-size was halved until it became 1 dB. After this, the procedure continued until 8 more reversals were completed. Finally, the median of the last 8 reversals was computed as the SRT-50 score of every subject. Every subject had a 2-3 minutes training session before the official test, to get familiar with the GUI. Per subject, the SRT-50 was computed once for each algorithm.

7.2.4. Localization Test

In order to perform the localization test, we implemented the GUI as depicted in Fig. 7.2. There is a question on the top of the screen which asks the subject to identify the perceived direction of a specific source. The subjects were asked to listen to the algorithms by pressing the buttons on the right-hand side as many times they wanted and then identified the angle by pressing one of the circles on the image. There are 6 buttons in total on the right-hand side (for the testing phase) as shown in Fig. 7.2, because there are five competing methods and one reference signal, which is the point source in question in isolation. For the testing phase, the user pressed the 'next experiment' button $3 \times 2 \times 2 = 12$ times (i.e., there are 12 pages in total) to find the azimuths of all the point sources (3 in total) in the acoustic scenes AC2 and AC3 for two repetitions.

The algorithms were presented in a random order and in the testing phase the acoustic scenes were also presented in random order between different pages and within the same page. Moreover, the input signals to all presented algorithms had an overall SNR of -5 dB, in order to clearly hear all dominant point sources after processing. Finally, the localization errors were computed with respect to a reference signal azimuth (and not the true azimuth of the source). This is because, the HA signals were constructed using a single set of HRTFs from [24], which are different than the HRTFs of the subjects. Thus, the subjects will, typically, perceive binaural cues differently from each other. Since the localization test is to verify preservation

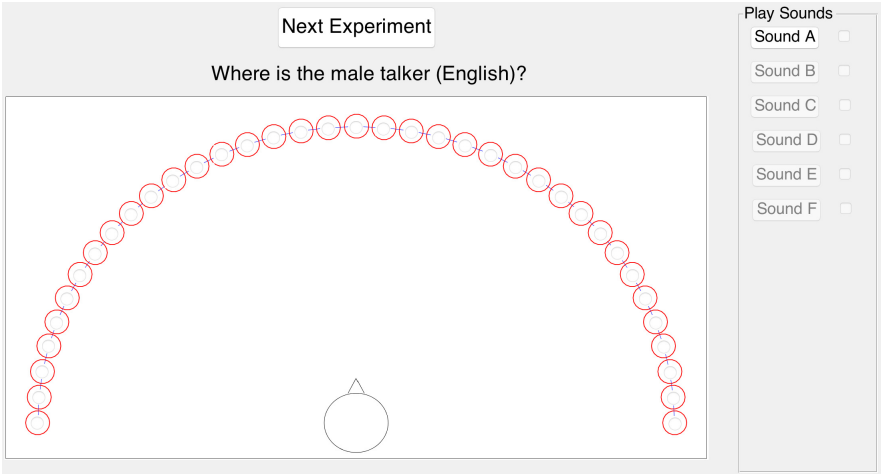

Figure 7.2: Graphical user interface of localization test.

of binaural cues, it is better to check how close the binaural cues after processing are to those before processing for each subject. Finally, since there are two repetitions, we played each reference source signal twice and calculated the average response on this as the reference location. Finally, we averaged all localization errors across all sources in the acoustic scene per algorithm.

7.2.5. Parameter Selection Phase Results

In the parameter selection phase, we compared all methods from Section 7.1 except for the ideal binaural target enhancement method. The comparisons were made only for scene AC1. For the relaxed binaural LCMV method, we tested all values of the parameter c from the set $c \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and for the BMVDR with thresholding, we tested all values of the parameter τ from the set $\tau \in \{-8, -4, 0, 4, 8\}$ dB. Fig. 7.3 shows the trade-off curves of the two methods with the SRT-50 scores on the x-axis and localization error on the y-axis, parameterized by the tested τ and c parameter. For both the SRT-50 and the localization error, the final score was calculated as the mean across different subjects. The mean localization-error scores were also computed across different sources and repetitions.

As expected (see Section 7.1.2), as c increases, the relaxed binaural LCMV method, in most cases, has an increased localization error and an increased intelligibility. The BMVDR thresholding method has a steady localization error for all tested τ values, while it provides a large intelligibility improvement for small τ values. In Fig. 7.3, two reasonably good parameter choices for the two methods are the ones with the largest intelligibility improvement and as small localization error as possible, i.e., $c = 0.7$ and $\tau = -8$ dB. We used only these two parameter choices for the testing phase (Section 7.2.6).

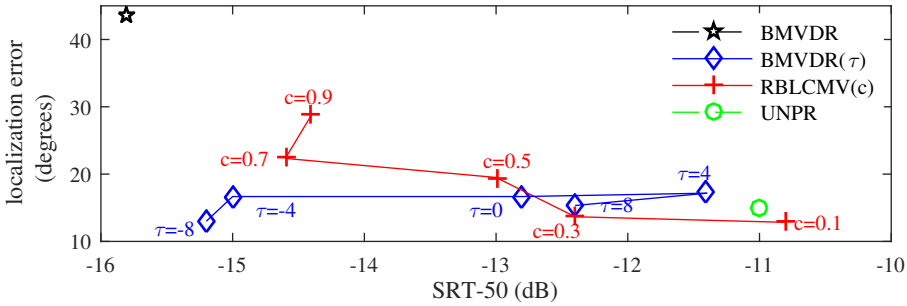

Figure 7.3: Parameter selection phase: Trade-off between localization error (degrees) and SRT-50 (dB).

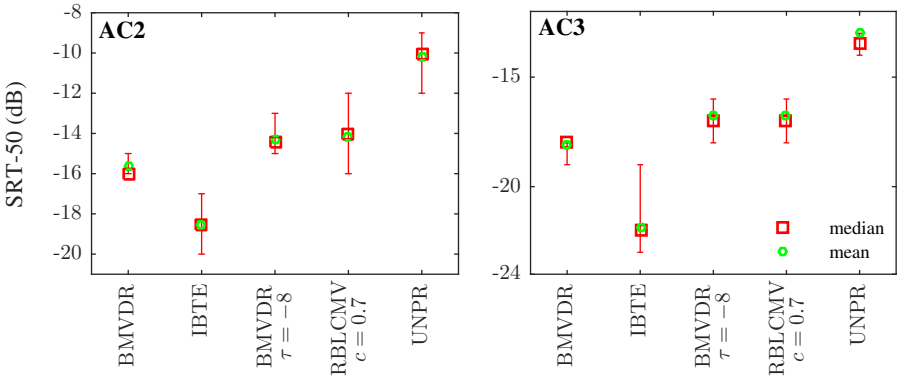

Figure 7.4: Testing phase: SRT-50 (dB) statistics.

7.2.6. Testing Phase Results

In the testing phase, we compared all methods from Section 7.1 in scenes AC2 and AC3. Fig. 7.4 shows the median and mean SRT-50, and the 0.25 and 0.75 quantiles. Fig. 7.5 shows the median and mean localization error, and the 0.25 and 0.75 quantiles. We also performed two t-tests to determine, if the compared methods are significantly different in terms of intelligibility improvement and localization error. The p-values of the intelligibility t-test are given in Table 7.2 and 7.3 for acoustic scenes AC2 and AC3, respectively. It is clear from the p-values that the relaxed binaural LCMV ($c = 0.7$) and the BMVDR thresholding ($\tau = -8$) are not significantly different from each other. The intelligibility of both proposed methods is significantly better compared to the unprocessed scene and significantly worse compared to the BMVDR. The p-values of the localization t-test are given in Table 7.4 and 7.5 for acoustic scenes AC2 and AC3, respectively. In both scenes, the proposed methods have a significantly better localization than BMVDR. Moreover, in

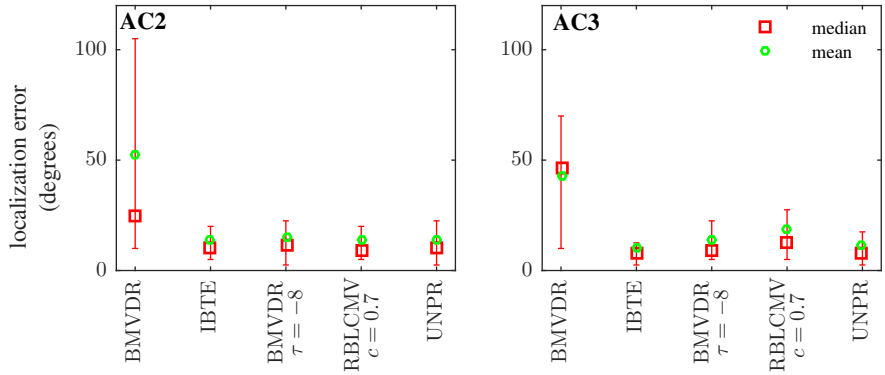

Figure 7.5: Testing phase: localization error (degrees) statistics.

Table 7.2: T-test p-values for intelligibility test in AC2.

Method	BMVDR	IBTE	BMVDR ($\tau = -8$)	RBLCMV ($c = 0.7$)	UNPR
BMVDR ($\tau = -8$)	0.0149	0	1	0.9177	0
RBLCMV ($c = 0.7$)	0.0401	0	0.9177	1	0

Table 7.3: T-test p-values for intelligibility test in AC3.

Method	BMVDR	IBTE	BMVDR ($\tau = -8$)	RBLCMV ($c = 0.7$)	UNPR
BMVDR ($\tau = -8$)	0.0259	0	1	1	0
RBLCMV ($c = 0.7$)	0.0105	0	1	1	0

scene AC2, the two proposed methods are not significantly different from the ideal target enhancement and for scene AC3 the BMVDR ($\tau = -8$) is not significantly different from the ideal binaural target enhancement or the unprocessed noisy scene. This means that the proposed methods indeed preserve the correct locations of the sources in most cases, while significantly improve the intelligibility with respect to the unprocessed scene.

7.3. Conclusion

In this paper, we perceptually evaluated two recently proposed binaural speech enhancement methods in terms of intelligibility improvement and localization error. Both methods provide a significantly better trade-off between intelligibility improvement and localization performance compared to the unprocessed scene and

Table 7.4: T-test p-values for localization test in AC2.

Method	BMVDR	IBTE	BMVDR ($\tau = -8$)	RBLCMV ($c = 0.7$)	UNPR
BMVDR ($\tau = -8$)	0	0.5645	1	0.3161	0.4153
RBLCMV ($c = 0.7$)	0	0.7800	0.3161	1	0.8673

Table 7.5: T-test p-values for localization test in AC3.

Method	BMVDR	IBTE	BMVDR ($\tau = -8$)	RBLCMV ($c = 0.7$)	UNPR
BMVDR ($\tau = -8$)	0	0.0515	1	0.0272	0.2366
RBLCMV ($c = 0.7$)	0	0.0001	0.0272	1	0.0014

the reference BMVDR method. Moreover, in most cases, the two methods are not significantly different than the ideal binaural target enhancement method in terms of localization error. Moreover, the difference between the two methods is not statistically significant in most cases.

REFERENCES

[1] J. M. Kates, *Digital hearing aids* (Plural publishing, 2008).

[2] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, *Speech enhancement with multimicrophone wiener filter techniques in mulltimicrophone binaural hearing aids,* J. Acoust. Soc. Amer. **125**, 360 (2009).

[3] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, *Multichannel signal enhancement algorithms for assisted listening devices,* IEEE Signal Process. Mag. **32**, 18 (2015).

[4] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[5] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, *Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,* IEEE Trans. Audio, Speech, Language Process. **23**, 2449 (2015).

[6] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[7] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[8] 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).

[9] A. Spriet, M. Moonen, and J. Wouters, *Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,* Signal Process. **84**, 2367 (2004).

[10] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2005) pp. 29–32.

[11] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, *Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions,* in *Int. Workshop Acoustic Echo, Noise Control (IWAENC)* (2006).

[12] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[13] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, *Theoretical analysis of binaural multimicrophone noise reduction techniques,* IEEE Trans. Audio, Speech, Language Process. **18**, 342 (2010).

[14] 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 Trans. Audio, Speech, Language Process. **23** (2015).

[15] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, *Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene,* EURASIP J. Advances Signal Process. (2016).

[16] A. I. Koutrouvelis, J. Jensen, M. Guo, R. C. Hendriks, and R. Heusdens, *Binaural speech enhancement with spatial cue preservation utilising simultaneous masking,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2017).

[17] B. Cornelis and M. Moonen, *Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel wiener filtering based noise reduction,* J. Acoust. Soc. Amer. **131**, 4743 (2012).

[18] D. Marquardt, *DEVELOPMENT AND EVALUATION OF PSYCHOACOUSTICALLY MOTIVATED BINAURAL NOISE REDUCTION AND CUE PRESERVATION TECHNIQUES*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg (2015).

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, *An algorithm for intelligibility prediction of time-frequency weighted noisy speech,* IEEE Trans. Audio, Speech, Language Process. **19**, 2125 (2011).

[20] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, *Predicting the intelligibility of noisy and nonlinearly processed binaural speech,* IEEE Trans. Audio, Speech, Language Process. **24**, 1908 (2016).

[21] C. Faller and J. Merimaa, *Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,* J. Acoust. Soc. Amer. **116**, 3075 (2004).

[22] M. Dietz, S. D. Ewert, and V. Hohmann, *Auditory model based direction estimation of concurrent speakers from binaural signals,* ELSEVIER Speech Commun. **53**, 592 (2011).

[23] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. IEEE **57**, 1408 (1969).

[24] H. Kayser, S. Ewert, J. Annemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, *Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,* EURASIP J. Advances Signal Process. **2009**, 1 (2009).

[25] R. Houben, J. Koopman, H. Luts, K. C. Wagener, A. van Wieringen, H. Verschuure, and W. A. Dreschler, *Development of a dutch matrix sentence test to assess speech intelligibility in noise,* Int. J. Audiol. **53**, 760 (2014).

[26] V. Zue, S. Seneff, and J. Glass, *Speech database development at mit: Timit and beyond,* ELSEVIER Speech Commun. **9**, 351 (1993).

[27] H. Levitt, *Transformed up-down methods in psychoacoustics,* J. Acoust. Soc. Amer. **49**, 467 (1971).

8

A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks with Arbitrary Topology

This chapter is based on the article published as "A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks With Arbitrary Topology", by A.I. Koutrouvelis, T. W. Sherson, R. Heusdens and R.C. Hendriks in IEEE/ACM Trans. Audio, Speech and Language Processing. vol. 26, no. 8, pp. 1434-1448, Jan. 2018.
Thomas W. Sherson had a significant contribution in Sections 8.3.4, 8.4.3, 8.4.4, 8.4.5, 8.4.6, 8.4.7, 8.4.8, 8.5.4.

Beamforming (see e.g., [1–3] for an overview) plays an important role in multi-microphone speech enhancement [4–7]. The aim of a beamformer is the joint suppression of interfering noise and the preservation of an unknown target signal. The increasing usage of wireless portable devices equipped with microphones and limited power supplies, makes the notion of distributed beamforming in wireless acoustic sensor networks (WASNs) attractive compared to traditional centralized implementations [8]. The last decade, there are several proposed low-complexity distributed beamformers [9–18] that mainly focus on achieving a good trade-off between noise reduction and communication cost.

Both centralized and distributed beamformers typically require an estimate of the cross-power spectral density matrix (CPSDM) of the noise/noisy measurements, and estimate(s) of the relative acoustic transfer function (RATF) vector(s) of the acoustic source(s) present in the acoustic scene. Estimation errors in these quantities result in performance degradation of beamformers. Much attention has therefore been given to the development of centralized robust beamformers which minimize the effects of RATF estimation errors (see e.g., [2, 3] for an overview). Developing robust *distributed* beamformers is more challenging than developing robust centralized beamformers, as distributed beamformers cannot afford high-complexity robust solutions. Therefore, it is desired to find very low-complexity robust distributed beamformers that achieve good performance trade-offs as described previously.

A low-complexity and easily manipulated family of beamformers are those that are calculated through linearly constrained quadratic problems such as: the minimum power distortionless responce (MPDR) beamformer [19] and its multiple constrained generalization, the linearly constrained minimum power (LCMP) beamformer [20]. Both beamformers minimize the total power of the *noisy* measurements while preserving the target. Therefore, their performance highly depends on the estimation accuracy of the RATF vector of the target source [2, 3, 21]. RATF estimation errors might result in removal of the actual target source and preservation in the direction of the wrongly estimated RATF vector.

Two straightforward, low-complexity, robust alternatives to MPDR and LCMP are the minimum variance distortionless response (MVDR) beamformer [21] and the linearly constrained minimum variance (LCMV) beamformer [2], respectively. Both methods minimize the output *noise* power instead of the total noisy power, and thus require an estimate of the noise-only CPSDM. The noise CPSDM is typically estimated using a target activity detector (TAD) to identify target-free time-segments of audio. When the target is speech, this typically takes the form of a voice activity detector (see e.g., [6] for an overview). In [22], an alternative method was proposed to track the noise CPSDM also in time regions where the target is present. This method, however, highly depends on the estimation accuracy of the RATF vector of the target and its robustness to RATF estimation errors has not been tested.

Another family of low-complexity, robust alternatives to MPDR and LCMP are their diagonal loaded versions (see e.g., [23–25]). In both versions, the diagonal loading parameter, which is added to the main diagonal of the CPSDM, trades-off robustness against noise suppression. Specifically, by increasing the value of the

diagonal loading parameter, a higher robustness to RATF estimation errors and a lower noise suppression is achieved. With diagonal loading, the use of a TAD is unnecessary. To the authors' knowledge, there are no low-complexity distributed approaches for choosing the optimal diagonal loading parameter. Additionally, a constant diagonal loading parameter will not be optimal for all acoustical scenarios and all frequency bins.

From the above it becomes clear that in addition to robustness and low-cost distributed calculations, LCMV and LCMP beamformers have the additional challenge of the RATF vector estimation of the target source and possibly the interferers. There are several centralized methods for RATF vector estimation (see e.g., [7] for an overview), however, there are yet no low-complexity distributed alternatives for arbitrary network topologies. In several applications, such as teleconferencing, the sources do not change their locations significantly over time and, therefore, one may estimate the RATF vectors of the target and/or the interferers only during initialization using a centralized approach and then use these estimated RATF vectors in the distributed beamformer. The slight positional errors that will most likely occur after this initial estimation require robust distributed beamformers. Note that in this paper, we mainly focus on this type of applications, i.e., the sources that do not significantly change their locations with respect to an initial reference location.

Notably, existing distributed beamformers can be classified based on how they address the issue of forming CPSDMs in WASNs. In the first class, the CPSDMs are approximated to form distributed implementations [9–12] leading to approximately optimal performance. In the second class, the proposed beamformers obtain statistical optimality but do so at the expense of restricting the topology of the underlying WASN [13–15]. Statistically optimal beamformers which operate in unrestricted network topologies are much less common. An early example of such a beamformer is provided in [16], based on a maximum likelihood estimated LCMP beamformer. Unfortunately, this approach suffers from scaling communication costs as the number of samples used to construct the estimated CPSDM increases. In a similar vein, in [26], a distributed beamformer based on the pseudo-coherence principle was proposed. Similar to [16], the method in [26] can operate in cyclic networks. Furthermore, the authors demonstrated how the algorithm could perform near optimally with only a finite number of iterations, resulting in low transmission complexity. More recently, in [18] a topology independent distributed beamformer (i.e. able to operate in cyclic networks) was proposed. Similar in its design to [14], this method requires very limited communication between nodes while guaranteeing convergence to the optimal beamformer. However, it was also demonstrated that the rate of this convergence was slow, requiring a large number of iterations to achieve this point. In practice, with even slowly varying sound fields such a rate of convergence may be detrimental to overall performance.

In this paper, we propose a new robust distributed linearly constrained beamformer, addressing the aforementioned challenges. The optimization problem of the proposed method nulls each interferer using a linear equality constraint, reducing the full-element noise or noisy CPSDM to a *block-diagonal* form. In contrast to MVDR, MPDR, LCMV and LCMP beamformers, the proposed objective function

does not take into account correlation between different nodes in the WASN. Additionally, such an objective function is more convenient for distributed beamforming in WASNs of arbitrary topologies and significantly reduces the communication costs therein.

We show under realistic conditions, i.e., when the algorithms use non-ideally estimated RATF vectors and a non-ideal TAD, that the proposed method achieves a better predicted intelligibility than the MVDR and LCMV. The proposed method is less sensitive to RATF estimation errors, when TAD errors are not negligible, because of the block-diagonal form of the CPSDM.

The remainder of the paper is organized as follows. Section 8.1 presents the signal model. Section 8.2 reviews several methods of estimating the RATF vectors of the sources and the noisy/noise CPSDMs. Section 8.3 reviews the centralized and distributed linearly constrained beamformers. Section 8.4 presents the centralized and distributed versions of the proposed method. Section 8.5 shows the experimental results. Finally, concluding remarks are drawn in Section 8.6.

8.1. Signal Model

Consider an arbitrary undirected WASN of N nodes. Without loss of generality, we assume that the underlying network (which is potentially cyclic) is connected. Denote by $V = \{1, \cdots, N\}$ the set of node indices and by E the set of edges of the network whereby $(i, j) \in E \iff i, j \in V,\ i \neq j$ can communicate with one another. Each node κ is equipped with M_κ microphones, where $\sum_{\kappa \in V} M_\kappa = M$, thus forming an M-element microphone array. One of the M microphones is selected as the reference microphone for the beamforming purpose. The distributed beamformers presented in this paper are formulated in the short-time Fourier transform (STFT) domain on a frame-by-frame basis. The noisy DFT coefficient of the j-th ($j = 1, \cdots, M$) microphone of the k-th frequency bin of the β-th frame is given by

$$y_j(k,\beta) = \underbrace{a_j(k,\beta)s(k,\beta)}_{x_j(k,\beta)} + \sum_{i=1}^{r} \underbrace{b_{ij}(k,\beta)v_i(k,\beta)}_{n_{ij}(k,\beta)} + u_j(k,\beta) \tag{8.1}$$

with $s(k,\beta)$ and $v_i(k,\beta)$ the target source and the i-th interferer at the reference microphone, $a_j(k,\beta)$ and $b_{ij}(k,\beta)$ the RATF vectors elements of each with respect to the j-th microphone, and $x_j(k,\beta)$, $n_{ij}(k,\beta)$ and $u_j(k,\beta)$ the target source, the i-th interferer and ambient noise at the j-th microphone. Note that the reference microphone element of the RATF vectors is always equal to 1. Moreover, in the case of reverberant environments, the RATF vectors may also include a component due to early reverberation [27, 28]. Late reverberation and microphone self-noise are typically included in the ambient noise component. Note that even the late reverberation of the target has to be assigned to the ambient noise component because it reduces intelligibility [29, 30]. Thus, it should be reduced via the use of the beamformer. However, the early reflections (typically the first 50 ms [30]) are desired to be maintained because they typically contribute to intelligibility [29, 30].

Therefore, the ambient noise component is given by

$$u_j(k,\beta) = l_j^s(k,\beta) + \sum_{i=1}^{r} l_j^{v_i}(k,\beta) + c_j(k,\beta),$$

where $l_j^s(k,\beta)$ is the late reverberation component due to the target, $l_j^{v_i}(k,\beta)$ is the late reverberation component due to the i-th interferer, and $c_j(k,\beta)$ is the microphone self-noise.

In the sequel, we neglect the frame and frequency indices for the sake of brevity. Stacking all variables into vectors, Eq. (8.1) can be rewritten as

$$\mathbf{y} = \mathbf{x} + \underbrace{\sum_{i=1}^{r} \mathbf{n}_i + \mathbf{u}}_{\mathbf{n}} \in \mathbb{C}^{M\times 1}.$$

The CPSDM of $\mathbf{y}$ is given by $\mathbf{P}_\mathbf{y} = \mathrm{E}[\mathbf{y}\mathbf{y}^H]$, where $\mathrm{E}[\cdot]$ denotes statistical expectation. Assuming all sources are mutually uncorrelated, we have

$$\mathbf{P}_\mathbf{y} = \mathbf{P}_\mathbf{x} + \underbrace{\sum_{i=1}^{r} \mathbf{P}_{\mathbf{n}i} + \mathbf{P}_\mathbf{u}}_{\mathbf{P}_\mathbf{n}} \in \mathbb{C}^{M\times M}, \tag{8.2}$$

where $\mathbf{P}_\mathbf{x} = \mathrm{E}[\mathbf{x}\mathbf{x}^H] = p_s\mathbf{a}\mathbf{a}^H$ and $\mathbf{P}_{\mathbf{n}i} = \mathrm{E}[\mathbf{n}_i\mathbf{n}_i^H] = p_{v_i}\mathbf{b}_i\mathbf{b}_i^H$ are the CPSDMs of the target source and the i-th interferer at the microphones, respectively. Note that p_s and p_{v_i} are the power spectral densities of the target and the i-th interferer, respectively. Finally, the CPSDM of the ambient noise component, $\mathbf{P}_\mathbf{u}$, is given by

$$\mathbf{P}_\mathbf{u} = \mathrm{E}[\mathbf{u}\mathbf{u}^H] = \underbrace{\mathbf{P}_{\mathbf{l}s} + \sum_{i=1}^{r} \mathbf{P}_{\mathbf{l}u_i}}_{\mathbf{P}_\mathbf{l}} + \mathbf{P}_\mathbf{c} \in \mathbb{C}^{M\times M},$$

where $\mathbf{P}_\mathbf{l}$ denotes the CPSDM of the late reverberation, and $\mathbf{P}_\mathbf{c}$ the CPSDM of the microphone self-noise.

8.2. Estimation of Signal Model Parameters

The CPSDMs and the RATF vectors of the sources are unknown and have to be estimated in order to be available to the beamformers discussed in the sequel. In Sections 8.2.1 and 8.2.2, we review some existing methods for RATF vector and CPSDM estimation, respectively.

8.2.1. Estimation of RATF Vectors

In practical applications, the true RATF vectors are reverberant due to room acoustics [28, 31, 32]. Several centralized methods have been proposed to estimate these RATF vectors (see e.g., [7] for an overview). In [28], the RATF vector of the target

source is estimated by exploiting the assumption that the noise field is stationary. However, when the interferers are non-stationary, this can result in significant degradation in performance [31]. In [32] the subspaces of the target and interferers are estimated using a generalized eigenvalue decomposition (GEVD) combined with a TAD. While distributed methods have been proposed in the literature for performing GEVD-based subspace estimation in restricted network topologies (i.e., fully connected) [33], to our best knowledge, there are currently no distributed versions of the GEVD that operate in general cyclic networks.

In this work, we assume that estimates of the RATF vectors, $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}_i$, for $i = 1, \cdots, r$, are available at the initialization phase. In situations where the sources do not change their locations significantly with respect to an initial position, such as teleconferencing, the RATF vectors can be estimated (e.g., in a centralized way) during such an initialization. This will result in RATF estimation errors if the sources make some slight movements and, therefore, robust beamformers are required.

8.2.2. Estimation of CPSDMs

The LCMP and the MPDR beamformers depend on an estimate of the noisy CPSDM, $\hat{\mathbf{P}}_{\mathbf{y}}$. Typically, this estimate is computed using the sample average, which is given by

$$\hat{\mathbf{P}}_{\mathbf{y}} = \frac{1}{|L_y|} \sum_{l_y \in L_y} \mathbf{y}(l_y)\mathbf{y}^H(l_y),$$

where L_y is the set of frames of the entire time horizon and $|\cdot|$ denotes the cardinality of a set. The LCMV and the MVDR beamformers depend on an estimate of the noise CPSDM, $\hat{\mathbf{P}}_{\mathbf{n}}$. The noise CPSDM is estimated using the set of noise-only frames denoted by L_n, i.e.,

$$\hat{\mathbf{P}}_{\mathbf{n}} = \frac{1}{|L_n|} \sum_{l_n \in L_n} \mathbf{y}(l_n)\mathbf{y}^H(l_n),$$

where $|L_n| < |L_y|$. In order to obtain $\hat{\mathbf{P}}_{\mathbf{n}}$, a TAD is required to detect target presence/absence for each frame. The above two averages are updated in an online fashion, i.e., the average is updated for every frame using the average of the previous frame. This procedure becomes computationally demanding in a distributed context for two reasons. Firstly, the entire observation vector must be available at each time frame resulting in the need for data flooding. Secondly, that the storage of the entire CPSDM scales with the network size.

Estimation of the ambient noise CPSDM $\mathbf{P}_{\mathbf{u}}$ is a difficult task due to the late reverberation CPSDM $\mathbf{P}_{\mathbf{l}}$. Using a TAD it is nearly impossible to estimate $\mathbf{P}_{\mathbf{l}}$ alone. For sufficiently large rooms, the late reverberation is typically modelled as an ideal spherical isotropic noise field [7, 34].That is,

$$\hat{\mathbf{P}}_{\mathbf{l}} = \hat{p}_{\text{iso}}\mathbf{P}_{\text{iso}}, \tag{8.3}$$

where for the k-th frequency bin, the (i, j)-th element of $\mathbf{P}_{\text{iso}}$ is given by

$$\mathbf{P}_{\text{iso},i,j} = \text{sinc}\left(\frac{2\pi k f_s d_{i,j}}{\Phi c}\right), \tag{8.4}$$

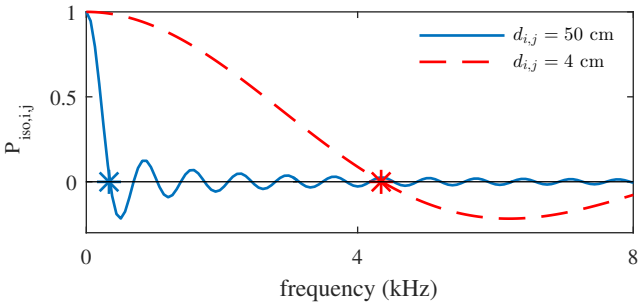

Figure 8.1: The spherically isotropic noise field correlation between two microphones i, j of distances $d_{i,j} = 4, 50$ cm and $f_s = 16$ kHz. The star marker denotes the first zero-crossing f_c.

where $d_{i,j}$ is the distance between microphones i and j, f_s is the sampling frequency, Φ is the number of frequency bins, and c is the speed of sound. The scaling $\hat{p}_{\text{iso}}$ can be estimated using several centralized methods (see e.g., [34]). To the best of our knowledge, there are no distributed methods for obtaining $\hat{p}_{\text{iso}}$.

Fig. 8.1 shows the values of the correlation function of Eq. (8.4) for various frequencies and distances $d_{i,j}$. The correlation can be roughly divided into two interesting frequency regions: one highly correlated on the left and one much less correlated on the right. The boundary between these regions occurs at the first zero-crossing given by $f_c = c/(2d_{i,j})$. It is clear that, the larger $d_{i,j}$ becomes, the smaller f_c is.

The CPSDM of the microphone self-noise, $\mathbf{P_c} = c\mathbf{I}$ (where c is the power at each microphone), can be estimated in silent frames only (i.e., neither target nor interferers are active).

8.3. Linearly Constrained Beamforming

Most linearly constrained beamformers are obtained from the following general optimization problem [1, 2, 20]

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{P} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \tag{8.5}$$

where $\mathbf{\Lambda} \in \mathbb{C}^{M \times d}$, $\mathbf{f} \in \mathbb{C}^{d \times 1}$, and $\mathbf{P} \in \mathbb{C}^{M \times M}$ is typically the CPSDM of the noise or noisy measurements. The d constraints used in the optimization problem of Eq. (8.5) include at least the distortionless constraint for the target source, i.e., $\mathbf{w}^H \mathbf{a} = 1$, and, commonly, the nulling of the interferers, $\mathbf{w}^H \mathbf{b}_i = 0$ [1, 32, 35]. If we assume that $r < M - 1$, the linearly constrained beamformer can null all interferers and still have control on the minimization of the objective function. In this case, $\mathbf{\Lambda}$ and $\mathbf{f}$ are given by

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{a} & \mathbf{b}_1 & \cdots & \mathbf{b}_r \end{bmatrix}, \text{ and } \mathbf{f} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^H. \tag{8.6}$$

It should be noted that by increasing the number of nulling constraints, the ambient output noise power may be boosted. The boost depends on the locations of the interferers [2] and the number of available degrees of freedom ($M-r-1$). However, in applications when $r \ll M-1$ this impact is much less significant. If $r < M-1$ and $\mathbf{P}$ is invertible, the optimization problem in Eq. (8.5), using the constraints in Eq. (8.6), has a closed-form solution given by [2]

$$\hat{\mathbf{w}} = \mathbf{P}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\mathbf{P}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f}.$$

When $\mathbf{P} = \mathbf{P}_{\mathbf{y}}$, the linearly constrained beamformer takes the form of the LCMP beamformer given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\mathbf{P}_{\mathbf{y}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H, \tag{8.7}$$

while if $\mathbf{P} = \mathbf{P}_{\mathbf{n}}$, the LCMV is obtained and is given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\mathbf{P}_{\mathbf{n}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H.$$

In the sequel, when we use the acronyms LCMV and LCMP we mean the LCMV and LCMP versions with the constraints given in Eq. (8.6). Another interesting linearly constrained beamformer is the one that has only the ambient noise component in the objective function [36], i.e.,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\mathbf{P}_{\mathbf{u}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H. \tag{8.8}$$

In this paper, we will refer to the linearly constrained beamformer in Eq. (8.8) as the ambient LCMV (ALCMV).

Using Eq. (8.2), the objective function of the LCMP problem, as noted in Eq. (8.7), is given by

$$\mathbf{w}^H\mathbf{P}_{\mathbf{y}}\mathbf{w} = p_s\mathbf{w}^H\mathbf{a}\mathbf{a}^H\mathbf{w} + \sum_{i=1}^{r} p_{v_i}\mathbf{w}^H\mathbf{b}_i\mathbf{b}_i^H\mathbf{w} + \mathbf{w}^H\mathbf{P}_{\mathbf{u}}\mathbf{w}.$$

Due to the included constraints in the LCMP (see Eq. (8.6)), the contributions of the early components of the sources to the objective function of Eq. (8.7) are constant. Thus, if $\hat{\mathbf{P}}_{\mathbf{y}} = \mathbf{P}_{\mathbf{y}}$, $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{P}_{\mathbf{n}}$, $\hat{\mathbf{P}}_{\mathbf{u}} = \mathbf{P}_{\mathbf{u}}$, and $\hat{\mathbf{\Lambda}} = \mathbf{\Lambda}$, the LCMP, LCMV and ALCMV beamformers are all equivalent. In practice, this never happens as there are always RATF estimation errors and CPSDM estimation errors, as explained previously.

8.3.1. RATF estimation errors

There are two interesting cases. In the first case, if $\hat{\mathbf{P}}_{\mathbf{y}} = \mathbf{P}_{\mathbf{y}}$, $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{P}_{\mathbf{n}}$, and $\hat{\mathbf{a}} = \mathbf{a}$, LCMP is equivalent to LCMV [2]. However, if $\hat{\mathbf{a}} \neq \mathbf{a}$, the LCMV beamformer (provided that $\hat{\mathbf{P}}_{\mathbf{n}}$ is accurately estimated), is more robust than the LCMP [2]. This is because LCMP will try to remove the actual target related to the RATF $\mathbf{a}$ as this is included in $\mathbf{P}_{\mathbf{y}}$, while the preservation constraint is on the wrongly

estimated $\hat{\mathbf{a}}$. However, if there are also TAD errors, $\hat{\mathbf{P}}_{\mathbf{n}}$ may also contain portions of $\mathbf{P}_{\mathbf{x}}$ and, as a result, the LCMV may also have severe performance degradation like the LCMP.

In the second case, if $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{P}_{\mathbf{n}}$, $\hat{\mathbf{P}}_{\mathbf{u}} = \mathbf{P}_{\mathbf{u}}$, and $\hat{\mathbf{b}}_i = \mathbf{b}_i$, for $i = 1, \cdots, r$, LCMV is equivalent to ALCMV. However, if any of the $\hat{\mathbf{b}}_i$'s contain estimation errors, there will be power leakage of the corresponding interferer(s), which is not controllable, neither by the objective function nor by the constraints of the ALCMV problem in Eq. (8.8). Moreover, if there are interferers whose RATF vectors have not been placed in the constraints, the ALCMV will also be unable to reduce them in a controlled way. In contrast, if $\hat{\mathbf{P}}_{\mathbf{n}}$ is estimated accurately, the LCMV will reduce these power leakages. In this case, the LCMV will most likely have a better noise reduction performance than its ALCMV counterpart.

We can conclude that the performance degradation of linearly constrained beamformers due to RATF estimation errors is mainly influenced by the selection of the CPSDM, $\mathbf{P}$, in the objective function of Eq. (8.5). A low-cost robust linearly constrained beamformer should have good performance under both RATF estimation errors and TAD errors. There are several approaches to achieve this. The most popular is via diagonal loading of $\mathbf{P}$. However, to the authors' knowledge there are no low-cost distributed approaches for optimally selecting the diagonal loading value. Another robust low-cost option is to use a fixed superdirective linearly constrained beamformer, i.e., a linearly constrained beamformer with a (semi)fixed $\mathbf{P}$ [5]. A fixed linearly constrained beamformer does not use a TAD and guarantees that there will not be any portion of $\mathbf{P}_{\mathbf{x}}$ in $\mathbf{P}$. Two interesting fixed linearly constrained beamformers are discussed in the next section.

8.3.2. Fixed Superdirective Linearly Constrained Beamformers

The fixed superdirective beamformers [5] assume a certain noise field and use in the objective function a certain coherence function like the one in Eq. (8.3). Since the early components of the interferers can be nullified using a linearly constrained beamformer, the noise field that remains is the late reverberation as explained previously in this section. Recall from Section 8.2.2, that the estimation of $\mathbf{P}_{\mathbf{u}}$ is a difficult task due to the CPSDM of the late reverberation, $\mathbf{P}_{\mathbf{l}}$. Typically, in the literature (see e.g., [5, 37, 38]) models of $\mathbf{P}_{\mathbf{l}}$ are used in beamformers instead. The most common choice is to use $\mathbf{P}_{\text{iso}}$. If one chooses $\mathbf{P} = \mathbf{P}_{\text{iso}}$, the microphone self-noise will be boosted in low frequencies [5]. Thus, a diagonal-loaded version is typically used [5, 39], i.e.,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H (p_{\text{iso}} \mathbf{P}_{\text{iso}} + \mathbf{P}_{\mathbf{c}}) \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \tag{8.9}$$

where $\mathbf{P}_{\mathbf{c}} = c\mathbf{I}$ (see Section 8.2.2). Although, the microphone-self noise power, c, typically remains constant over time, p_{iso} changes. To the best of our knowledge, there are no distributed estimation methods of the scaling coefficient p_{iso}. We call the beamformer in Eq. (8.9) as isotropic LCMV (ILCMV).

Another popular fixed linearly constrained beamformer uses in the objective

function the most simplistic option which is $\mathbf{P} = \mathbf{I}$, i.e.,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H. \tag{8.10}$$

In this paper, we will refer to this as the linearly constrained delay and sum (LCDS) beamformer. It is identical to the fixed beamformer of the generalized side-lobe canceller implementation of the LCMP beamformer (using the constraints in Eq. (8.6)) in [32]. Unlike ILCMV, the LCDS is easily distributable due to the separable nature of the objective function. This can be achieved via similar methods to those demonstrated in Section 8.4.3 and need only be performed once. Following this, the output can be computed via data aggregation or by solving a simple averaging problem, again lending itself to distributed implementations.

Similar to ALCMV, the ILCMV and LCDS beamformers cannot control power leakages due to inaccurate estimates of the interferers' RATF vectors and cannot control interferers which are not included in the constraints.

8.3.3. Other Related Linearly Constrained Beamformers

If we skip the nulling constraints and only impose the target distortionless constraint, the LCMV (LCMP) reduces to the MVDR (MPDR) [1, 19]. Similar to LCMV and LCMP, MVDR and MPDR are equivalent under the assumption that $\hat{\mathbf{P}}_{\mathbf{y}} = \mathbf{P}_{\mathbf{y}}$ and $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{P}_{\mathbf{n}}$ and $\hat{\mathbf{a}} = \mathbf{a}$ [2]. However, when $\hat{\mathbf{a}} \neq \mathbf{a}$, the MVDR is more robust to RATF estimation errors [2, 21]. A special case of the MPDR is the delay and sum (DS) beamformer [27] which replaces the noisy CPSDM with the identity matrix. The DS has worse performance compared to the MVDR (MPDR) in correlated noise fields but results in very robust performance to RATF estimation errors [21] and TAD errors.

8.3.4. Distributed Linearly Constrained Beamformers

The development of distributed beamformers has focused on adapting LCMV (LCMP) based approaches for use in WASNs. However, this adaptation has not come without additional challenges [40]. Most notable is the limited communication between devices which makes the formation of estimated CPSDMs nearly impossible without the use of a fusion center [8]. To address this, two main classes of distributed beamformers have appeared in the literature: approximately optimal variants and optimal approaches which operate in certain networks.

One such sub-optimal variant is the distributed DS beamformer introduced in [9]. Based on randomised gossip [41], this low-cost method operates in general cyclic networks but fails to exploit spatial correlation to improve noise reduction. In contrast, distributed approximations of the MVDR beamformer [10, 11] assume that disjoint nodes are uncorrelated essentially masking the true CPSDMs. While lending themselves to distributed implementations, such approaches fail to take into account the true correlations between observed signals across the network, resulting in sub-optimal performance.

By restricting the network topology, typically to be acyclic or fully connected, optimal distributed beamformers have been proposed. These algorithms [14, 15] ex-

ploit efficient data aggregation to construct global beamformers from a composition of local filters and have been shown to be iteratively optimal. However, the additional communication overhead required to maintain a constant network topology across frames can be prohibitively expensive due to unpredictable network dynamics. Furthermore, such maintenance may be impossible in the case of node failure.

It is worth mentioning that it is not the use of an acyclic network in [14, 15] itself which is limiting, but rather the need for this network to be invariant over time. In [18], this point was exploited to form a fully distributed beamformer for use in general cyclic topologies. Like [14] and [15], [18] constructs a global beamformer as a composition of local beamformers at each node. Importantly, the method by which these local beamformers are combined does not depend on the underlying network topology. This allows the network to vary between frames, overcoming the need for maintaining a fixed topology in all time instances. The method in [18] was shown to be iteratively optimal with its main drawback being a decrease in convergence rate compared to [14], requiring a larger number of frames to obtain near optimal performance.

In contrast, in [16], an optimal distributed beamformer was proposed for use in cyclic networks by exploiting the structure of estimated CPSDMs to cast LCMP beamforming as distributed consensus. However, for CPSDM estimates based on a large number of frames, the proposed algorithm's communication cost scaled poorly. In contrast to [13–15] and [18], a benefit of [16] was that the proposed implementation was frame-optimal, i.e. that it obtained the performance of an equivalent centralized implementation in each frame. The beamformer proposed in [26] exploited a similar method of distributed implementation, but exploited the pseudo coherence principle of human speech to overcome the scaling communication costs found in [16].

The approaches of both [16] and [26] made use of internal optimization schemes which require a large number of iterations per frame to obtain optimal performance. However, in [26] it was shown that near optimal performance could be obtained using only a finite number of iterations of this internal solver. Such a result raises the question whether a similar approach could be employed as a general way of reducing the transmission costs associated with cyclic beamforming methods. For the beamformers proposed in this work, this point is touched upon in Section 8.4.7.

In contrast to the methods above, the beamformers proposed in Section 8.4 are fully distributable without imposing restrictions on the underlying network topology or scaling communication costs while also being optimally computable in each frame. In this way, the proposed methods combine the strengths of existing distributed beamformers while also avoiding their various limitations.

8.4. Proposed Method

In the previous section, we have highlighted the susceptibility of several existing beamformers to RATF estimation errors and TAD errors and the challenge of deploying these algorithms in distributed contexts. Here, we propose two different linearly constrained beamformers which are efficiently distributable for arbitrary network topologies, robust to RATF estimation errors and TAD errors, while at the

same time are able to control the power leakage of the interferers.

Typically, the microphones within a node are nearby, while the microphones from different nodes are further away. Therefore, the late reverberation will be highly correlated in the first case, while in the latter less correlated (see Fig. 8.1). Therefore, providing that the nodes are sufficiently far away from each other, one may approximate the full element matrix $\mathbf{P_u}$ with the *block-diagonal* matrix $\bar{\mathbf{P}}_\mathbf{u}$ where every block corresponds to the CPSDM of the late reverberation of one node only and the microphone-self noise. Therefore, we propose the block-diagonal ALCMV (BDALCMV) which is given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \bar{\mathbf{P}}_\mathbf{u} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H. \tag{8.11}$$

Note that if every node has only one microphone, $\bar{\mathbf{P}}_\mathbf{u}$ becomes diagonal. This block-diagonalization lends itself to distributed implementations, reflecting a similar objective structure to that of the DS and LCDS beamformer.

While the proposed BDALCMV beamformer has a number of benefits from the perspective of distributed signal processing, like ALCMV, the challenge becomes the estimation of $\bar{\mathbf{P}}_\mathbf{u}$, and handling the possible power leakages of the interferers as in the case of DS, LCDS, ALCMV. Therefore, in Sections 8.4.1, and 8.4.2 we introduce two variations of the BDALCMV beamformer which do not require the estimation of $\bar{\mathbf{P}}_\mathbf{u}$ and are robust to power leakages of the interferers. Moreover, in Sections 8.4.3—8.4.7, we introduce distributed implementations of the proposed beamformers.

8.4.1. BDLCMP Beamformer

The first proposed practical variant of BDALCMV is the BDLCMP which uses in the objective function the block-diagonal noisy CPSDM, $\bar{\mathbf{P}}_\mathbf{y}$. That is,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \bar{\mathbf{P}}_\mathbf{y} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H. \tag{8.12}$$

This results in a local estimation problem, which can be carried out independently at each node without the need of a TAD. This method handles the possible power leakages due to inaccurate estimates of the interferers' RATF vectors and can suppress the interferers that are not included in the constraints.

In case of RATF estimation errors of the target source, the BDLCMP will have similar problems to the LCMP because in the block-diagonal matrices, there will be portions of the corresponding target block-diagonal CPSDMs. However, the performance degradation will not be that great as with the LCMP. This can be easily explained by considering the extreme scenario of a fully correlated noise field in which we assume that $M > r+1$, $\hat{\mathbf{P}}_\mathbf{y} = \mathbf{P}_\mathbf{y}$, $\mathbf{P_u} \approx 0$, $\hat{\mathbf{b}}_i = \mathbf{b}_i, i = 1, \cdots, r$ and $\hat{\mathbf{a}} \neq \mathbf{a}$. In this case, the optimization problem of LCMP in Eq. (8.7) will be approximately equivalent[1] to the following optimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\mathbf{P}}_\mathbf{y} \mathbf{w} \text{ s.t. } \mathbf{w}^H \tilde{\mathbf{\Lambda}} = \tilde{\mathbf{f}}^H,$$

[1] It is approximately equivalent because $\mathbf{P_u} \approx 0$. Moreover, the target RATF estimation errors should be sufficiently large.

where

$$\tilde{\mathbf{\Lambda}} = \begin{bmatrix} \hat{\mathbf{a}} & \mathbf{a} & \hat{\mathbf{b}}_1 & \cdots & \hat{\mathbf{b}}_r \end{bmatrix}, \text{ and } \tilde{\mathbf{f}}^H = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

That is, the LCMP will approximately nullify the target source. In contrast, due to the block-diagonal CPSDM, the BDLCMP will approximately nullify the target source iff $M > rN + 2r + 1$, where N is the number of nodes. Specifically, if $M > rN + 2r + 1$ is satisfied, the BDLCMP will be approximately equivalent to the following optimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \hat{\tilde{\mathbf{P}}}_{\mathbf{y}} \mathbf{w} \text{ s.t. } \mathbf{w}^H \tilde{\mathbf{\Lambda}} = \tilde{\mathbf{f}}^H,$$

where

$$\begin{aligned} \tilde{\mathbf{\Lambda}} &= \begin{bmatrix} \hat{\mathbf{a}} \; \tilde{\mathbf{a}}_1 \; \tilde{\mathbf{a}}_2 \; \cdots \; \tilde{\mathbf{a}}_N \; \hat{\mathbf{b}}_1 \cdots \hat{\mathbf{b}}_r \; \tilde{\mathbf{b}}_{11} \cdots \tilde{\mathbf{b}}_{1N} \cdots \tilde{\mathbf{b}}_{r1} \cdots \tilde{\mathbf{b}}_{rN} \end{bmatrix}, \\ \tilde{\mathbf{f}}^H &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \\ \tilde{\mathbf{a}}_i &= \begin{bmatrix} \mathbf{0} & \mathbf{a}_i & \mathbf{0} \end{bmatrix}^H, \quad \tilde{\mathbf{b}}_{ji} = \begin{bmatrix} \mathbf{0} & \mathbf{b}_{ji} & \mathbf{0} \end{bmatrix}^H \in \mathbb{C}^{M \times 1}. \end{aligned}$$

Here $\mathbf{a}_i, \mathbf{b}_{ji}$ are the elements of the RATF vector $\mathbf{a}, \mathbf{b}_j$ corresponding to node i, respectively. Note that for $M < rN + 2r + 1$ the BDLCMP will not have enough degrees of freedom to achieve $\mathbf{w}^H \tilde{\mathbf{a}}_i = 0$ $(i = 1, \cdots, N)$ and, thus, will not nullify the target signal. Thus, more microphones are needed in the BDLCMP beamformer to nullify the target signal compared to the LCMP beamformer. Hence, the BDLCMP is more robust to target RATF estimation errors compared to the LCMP for the same number of microphones M, when $M < rN+2r+1$, in this particular scenario of a fully correlated noise field. In more general noise fields, where $\mathbf{P}_{\mathbf{u}}$ is not negligible, both LCMP and BDLCMP will not nullify the target using the same finite number of microphones. However, LCMP will suppress more the target signal than the BDLCMP, because the first exploits the full-element noisy CPSDM matrix.

Fig. 8.2 shows the directivity patterns of LCMP and BDLCMP for a simple acoustic scenario with a linear microphone array separated into two nodes where each node has three microphones. The target source is at 80°, but the estimated RATF vector of the target is at 90°. The interferers and their RATF vectors are at $10°, 50°$ and 160°. All RATF vectors are anechoic in this example and there is a slight amount of microphone-self noise. It is clear from the directivity pattern in Fig. 8.2, that LCMP suppresses the target signal significantly, while BDLCMP does not.

It is worth mentioning that if $\hat{\mathbf{b}}_i \neq \mathbf{b}_i$, it easy to show (following the same steps as before) that the LCMP will typically suppress more the i-th interferer than BDLCMP, if both use the same number of microphones. This means that the power leakages of the interferers will be suppressed more with the LCMP compared to the BDLCMP. Nevertheless, we will experimentally show in Section 8.5, that the final intelligibility improvement of BDLCMP is much greater than the LCMP, because BDLCMP distorts much less the target.

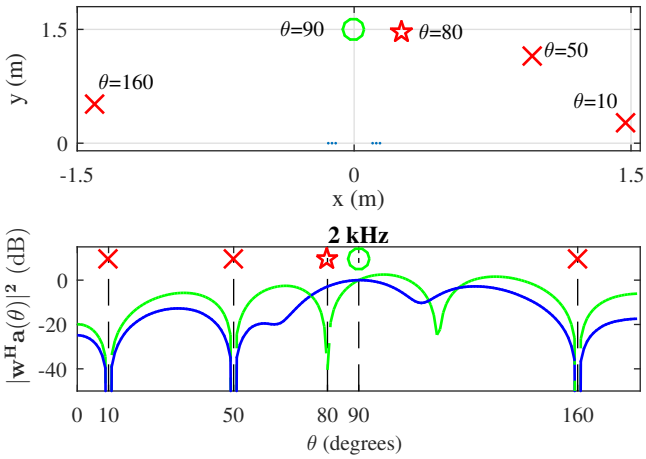

Figure 8.2: Example: three interferers (with marker 'x') and one target (with marker $\star$) at 80°. The RATF vector of the target points at 90°. The directivity pattern, $|\mathbf{w}^H\mathbf{a}(\theta)|^2$ (in dB), is computed in the range $0° \leq \theta \leq 180°$, for BDLCMP (solid line) and LCMP (dotted line), for the frequency 2 kHz.

8.4.2. BDLCMV Beamformer

To further increase the robustness of the proposed method, we introduce the BDLCMV variant which uses in the objective function the block-diagonal version of the noise CPSDM, $\bar{\mathbf{P}}_{\mathbf{n}}$. Therefore, the BDLCMV is given by

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^H\bar{\mathbf{P}}_{\mathbf{n}}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H. \tag{8.13}$$

Similar to the relationship between LCMV and LCMP, the BDLCMV typically enjoys more robustness than the BDLCMP when $\bar{\mathbf{P}}_{\mathbf{n}}$ is estimated accurately enough. However, when there are TAD errors, we will show that the performance gap reduces between the two methods. The BDLCMV also handles the possible power leakages of the interferers, and can suppress the interferers that are not included in the constraints.

If each node has only one microphone, then BDLCMV becomes diagonal. In this case, it can be viewed as a weighted version of the LCDS beamformer, and without nulling constraints, can be viewed as a weighted DS beamformer.

8.4.3. Distributed Implementation of the Proposed Method

Given a block-diagonal matrix $\bar{\mathbf{P}}$, which can be $\bar{\mathbf{P}}_{\mathbf{u}}$, $\bar{\mathbf{P}}_{\mathbf{n}}$ or $\bar{\mathbf{P}}_{\mathbf{y}}$, and a known constraint matrix $\mathbf{\Lambda}$, we now demonstrate how we can form a distributed version of the proposed methods for use in general cyclic networks by using a similar technique to that presented in [16]. Importantly, the imposed block diagonal structure of the estimated CPSDM results in a naturally separable objective function, leading to a substantial reduction in communication costs compared to those in [16]. To demon-

strate this, denote by $\mathbf{w}_\kappa$, $\mathbf{\Lambda}_\kappa$ and $\bar{\mathbf{P}}_\kappa$ the elements of $\mathbf{w}$, the rows of $\mathbf{\Lambda}$ and the block diagonal component of $\bar{\mathbf{P}}$ associated with node κ, respectively. Eqs. (8.11), (8.12) and (8.13) can therefore be rewritten as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2} \sum_{\kappa=1}^{N} \mathbf{w}_\kappa^H \bar{\mathbf{P}}_\kappa \mathbf{w}_\kappa \text{ s.t. } \sum_{\kappa=1}^{N} \mathbf{w}_\kappa^H \mathbf{\Lambda}_\kappa = \mathbf{f}^H. \tag{8.14}$$

The real-valued Lagrangian of this problem is given by

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\mu}) = \sum_{\kappa=1}^{N} \left(\frac{\mathbf{w}_\kappa^H \bar{\mathbf{P}}_\kappa \mathbf{w}_\kappa}{2} - \Re\left(\boldsymbol{\mu}^H \left(\mathbf{\Lambda}_\kappa^H \mathbf{w}_\kappa - \frac{\mathbf{f}}{N} \right) \right) \right),$$

where we have partitioned the constraint vector $\mathbf{f}$ into N equal parts, $\mathbf{f}/N$,one for each node $i \in V$. Taking complex partial derivatives [42], it follows that

$$\hat{\mathbf{w}_\kappa} = \bar{\mathbf{P}}_\kappa^{-1} \mathbf{\Lambda}_\kappa \boldsymbol{\mu}, \tag{8.15}$$

such that the corresponding dual function is thus given by

$$q(\boldsymbol{\mu}) = -\sum_{\kappa=1}^{N} \frac{\boldsymbol{\mu}^H \mathbf{\Lambda}_\kappa^H \bar{\mathbf{P}}_\kappa^{-1} \mathbf{\Lambda}_\kappa \boldsymbol{\mu}}{2} + \Re\Big(\boldsymbol{\mu}^H \mathbf{f} \Big).$$

The resulting dual optimization problem is given by

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \sum_{\kappa=1}^{N} \left(\frac{\boldsymbol{\mu}^H \mathbf{\Lambda}_\kappa^H \bar{\mathbf{P}}_\kappa^{-1} \mathbf{\Lambda}_\kappa \boldsymbol{\mu}}{2} - \Re\Big(\boldsymbol{\mu}^H \frac{\mathbf{f}}{N} \Big) \right). \tag{8.16}$$

8.4.4. Acyclic Implementation via Message Passing

We begin by demonstrating how, when the underlying network is acyclic (tree structured), the problem in Eq. (8.16) can be solved in a distributed manner. Similar to the approach introduced in [18], there is no need for this acyclic network to be constant between frames, allowing it to adapt to the time-varying connectivity of dynamic networks. This contrasts [14, 15] where the network topology must remain constant.

In the following, we consider two different approaches to compute the optimal $\boldsymbol{\mu}$ in tree structured networks. In the first approach, we exploit the fact that Eq. (8.16) can be directly solved by aggregating the sum of the local matrices $\frac{1}{2}\mathbf{\Lambda}_\kappa^H \bar{\mathbf{P}}_\kappa^{-1} \mathbf{\Lambda}_\kappa$ to a common location. In the case of acyclic networks, this aggregation can be performed efficiently with the common location forming the root node of the network. This root node can simply be a point in the network where we choose to extract the beamformer output signal.

To sketch the process of this data aggregation, we partition the set of neighbors of each node κ into two groups. The first group, denoted by $\mathcal{C}_\kappa$, represents the set of children of node κ. The second set, which is a unique node identifier, is the parent of node κ denoted by $\mathcal{P}_\kappa$. In particular, $\mathcal{P}_\kappa \cup \mathcal{C}_\kappa = \mathcal{N}(\kappa)\ \forall \kappa \in V$, where

$\mathcal{N}(\kappa) = \{\iota \mid (\kappa, \iota) \in E\}$. Note that for the root node $\mathcal{P}_\kappa = \emptyset$. These sets can be determined per frame by selecting a root node and forming a spanning tree via a breadth-first or depth-first search.

Once these sets are known, the process begins at the leaf nodes of the networks (those nodes for which $\mathcal{C}_\kappa = \emptyset$) and consists of the transmission of a message from these nodes (κ) to their parents ($\mathcal{P}_\kappa$). The aggregation messages are matrices and take the form

$$\mathbf{M}_{\kappa \to \mathcal{P}_\kappa} = \frac{\mathbf{\Lambda}_\kappa^H \bar{\mathbf{P}}_\kappa^{-1} \mathbf{\Lambda}_\kappa}{2}.$$

Of the set of remaining nodes, those nodes which have received a message from all but one of their neighbors can repeat this process (the remaining neighbor is their parent node). Their messages take a more general form given by

$$\mathbf{M}_{i \to \mathcal{P}_i} = \frac{\mathbf{\Lambda}_i^H \bar{\mathbf{P}}_i^{-1} \mathbf{\Lambda}_i}{2} + \sum_{k \in \mathcal{C}_i} \mathbf{M}_{k \to i},$$

whereby local information at each node is first combined with that from their children. This process is repeated until the root node has received messages from all its children at which point the aggregation operation is complete.

Due to their positive semidefinite structure, the transmission of each message per node comprises $\frac{1}{2}((r+1)^2 + r + 1)$ unique variables resulting in a total of $\frac{1}{2}(r^2 + 3r + 2)(N-1)$ transmitted variables for each frequency bin per frame. The optimal dual variables can then be diffused back into the network to allow the optimal beamformer weight vector to be computed at each node in parallel. This additional diffusion stage results in a further $(r+1)(N-K)$ transmitted variables where K denotes the number of leaf nodes. The beamformer output can then be computed by simply aggregating the sum $\sum_{i \in V} \mathbf{w}_i^H \mathbf{y}_i$ through the network, incurring a total cost of $(N-1)$ transmissions per frequency bin. Finally, if the estimate of $\bar{\mathbf{P}}$ does not change between frames, i.e., $\Delta\bar{\mathbf{P}} = \mathbf{0}$, the estimated weight vector need not be recomputed. An example of this occurs in noisy frames for the proposed BDLCMV method, reducing the cost of this algorithm in such frames to that of simply computing the beamformer output.

8.4.5. Cyclic Weight Vector Computation via PDMM

For more general network structures, Eq. (8.16) can be transformed to a fully distributable form. To do so, we introduce local versions of $\boldsymbol{\mu}$ at each node, denoted by $\boldsymbol{\mu}_\kappa$, and impose that $\boldsymbol{\mu}_\kappa = \boldsymbol{\mu}_\iota \; \forall \; (\kappa, \iota) \in E$. The resulting problem is given by

$$\begin{aligned} \hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \sum_{\kappa=1}^{N} \left(\frac{\boldsymbol{\mu}_\kappa^H \mathbf{\Lambda}_\kappa^H \bar{\mathbf{P}}_\kappa^{-1} \mathbf{\Lambda}_\kappa \boldsymbol{\mu}_\kappa}{2} - \Re\left(\boldsymbol{\mu}_\kappa^H \frac{\mathbf{f}}{N} \right) \right) \\ \text{s.t. } \boldsymbol{\mu}_\kappa = \boldsymbol{\mu}_\iota \quad \forall (\kappa, \iota) \in E. \end{aligned} \tag{8.17}$$

Note that at optimality, this problem is entirely equivalent to the problem in Eq. (8.16), assuming the network is connected. Due to its separable quadratic structure,

Eq. (8.17) can be solved via a wide range of existing distributed solvers [43–45]. In this work, we consider solving Eq. (8.17) using the primal dual method of multipliers (PDMM) proposed in [45].

To define the PDMM updating scheme, we begin by again considering the equivalent graph representation of the network, parameterised by node set V and edge set E. For each node κ and edge $(\kappa,\iota)\in E$, define the vectors $\boldsymbol{\mu}_{\kappa}^{(0)}=\boldsymbol{\gamma}_{\kappa,\iota}^{(0)}=\mathbf{0}\in\mathbb{C}^{r+1}$, $\forall\kappa=1,\ldots,N$, $(\kappa,\iota)\in E$ respectively. As per the PDMM algorithm in [45], the optimizers of Eq. (8.17) can then be computed by iteratively updating the dual variables $(\boldsymbol{\mu}_{\kappa})$ and directed edge variables $(\boldsymbol{\gamma}_{\kappa|\iota})$ as

$$\begin{aligned}\boldsymbol{\mu}_{\kappa}^{(t+1)}=&\left(\boldsymbol{\Lambda}_{\kappa}^{H}\bar{\mathbf{P}}_{\kappa}^{-1}\boldsymbol{\Lambda}_{\kappa}+\rho|\mathcal{N}(\kappa)|\mathbf{I}\right)^{-1}\\&\left(\frac{\mathbf{f}}{N}+\sum_{\iota\in\mathcal{N}(\kappa)}\left(\frac{\kappa-\iota}{|\kappa-\iota|}\boldsymbol{\gamma}_{\iota|\kappa}^{(t)}+\rho\boldsymbol{\mu}_{\iota}^{(t)}\right)\right)\\\boldsymbol{\gamma}_{\kappa|\iota}^{(t+1)}=&\boldsymbol{\gamma}_{\iota|\kappa}^{(t)}-\rho\frac{\kappa-\iota}{|\kappa-\iota|}\left(\boldsymbol{\mu}_{\kappa}^{(t+1)}-\boldsymbol{\mu}_{\iota}^{(t)}\right),\end{aligned}\tag{8.18}$$

where each $\rho\in(0,+\infty)$ is the step size for the iterative algorithm and t denotes the iteration index. The notation $\kappa|\iota$ is used to define the edge variable computed at node κ related to the edge $(\kappa,\iota)\in E$.

The edge based update requires the transmission of information between neighbouring nodes, as can be noted in the dependence of $\boldsymbol{\gamma}_{\kappa|\iota}^{(t+1)}$ on $\boldsymbol{\gamma}_{\iota|\kappa}^{(t)}$ and $\boldsymbol{\mu}_{\iota}^{(t)}$. As highlighted in [45] however, this only requires the transmission of the $\boldsymbol{\mu}_{\kappa}$ variables and, thus, can be performed via a broadcast transmission protocol at each node. These updates can then be iterated until a desired level of precision is achieved after which $\hat{\mathbf{w}}_j$ can be calculated locally at each node via Eq. (8.15).

Each iteration of the proposed algorithm requires the transmission of $r+1$ variables per node. In an existing optimal cyclic beamformer [16] this cost was $r+1+|L_y|$, where $|L_y|$ is the number of frames used to form a maximum likelihood estimated version of the CPSDM. The proposed method therefore requires $|L_y|$ less transmissions per iteration, resulting in a substantial saving in transmission costs.

8.4.6. Beamformer Output Computation

Once the weight vector is known, the beamformer output can then be computed via various distributed averaging techniques (see [46] for an overview). In the case of this work we again consider the use of PDMM for this task. Consider the standard distributed averaging problem given by

$$\begin{aligned}\min_{\mathbf{x}}\quad&\frac{1}{2}\sum_{\kappa=1}^{N}\|\mathbf{x}_{\kappa}-\mathbf{w}_{\kappa}^{H}\mathbf{y}_{\kappa}\|^{2}\\\text{s.t.}\quad&\mathbf{x}_{\kappa}=\mathbf{x}_{\iota}\ \forall(\kappa,\iota)\in E.\end{aligned}\tag{8.19}$$

Again, from [45], the PDMM update equations for this problem are given by

$$\mathbf{x}_\kappa^{(t+1)} = \frac{\left(\mathbf{w}_\kappa^H \mathbf{y}_\kappa + \sum_{\iota\in\mathcal{N}(\kappa)} \left(\frac{\kappa-\iota}{|\kappa-\iota|}\mathbf{z}_{\iota|\kappa}^{(t)} + \rho\mathbf{x}_\iota^{(t)}\right)\right)}{1+\rho|\mathcal{N}(\kappa)|} \tag{8.20}$$

$$\mathbf{z}_{\kappa|\iota}^{(t+1)} = \mathbf{z}_{\iota|\kappa}^{(t)} - \rho\frac{\kappa-\iota}{|\kappa-\iota|}\left(\mathbf{x}_\kappa^{(t+1)} - \mathbf{x}_\iota^{(t)}\right), \tag{8.21}$$

where $\mathbf{z}_{\kappa|\iota}$ denotes the directed edge variable owned by node κ. By iterating these updates, every node in the network can learn the average of the vector $\mathbf{w}^H\mathbf{y}$. Once the average is known, this can be scaled by a factor of N to recover the beamformer output. Alternatively, we can employ the same acyclic beamformer output computation approach as used in Sec. 8.4.4. While this removes the entirely cyclic nature of the algorithm as the tree structured network used can change in each frame, the overhead of using an acyclic network is still substantially reduced in contrast to the work of [14, 15].

8.4.7. Cyclic Beamforming with Finite Numbers of Iterations

In general distributed applications, deterministic signal processing is desirable. This point is even more pressing in the case of distributed audio processing. Thus, an unbounded requirement on the iteration count of an algorithm is cumbersome. Unfortunately, in practice, the total number of transmissions required to solve the problems in Eq. (8.17) and (8.19), via general cyclic solvers such as PDMM, is dependent not only on the choice of the solver but also on the WASN topology. As such, it is not possible to analytically bound this transmission cost for arbitrary networks. However, in the distributed beamforming method presented in [26], which also used PDMM as a solver, it was found that near optimal performance was achieved in only a limited number iterations. In this way it is expected that the number of iterations required to achieve a good level of performance is not unnecessarily large. As such we can impose a hard limit on the number of iterations performed without significantly degrading performance.

An additional observation is that, due to its dependence on a recursively averaged covariance matrix, the weight vector $\mathbf{w}$ will vary smoothly with time. With regards to the PDMM algorithm, this corresponds to the fact that both the dual and edge variables will also vary somewhat smoothly. As such, one way to improve precision even under the scenario of a finite number of iterations it to use a warm-start procedure. Defining the maximum number of iterations by $t_{\max}$, this warm-start procedure is implemented by setting

$$\boldsymbol{\mu}_\beta^{(0)} = \boldsymbol{\mu}_{\beta-1}^{(t_{\max})} \quad \text{and} \quad \boldsymbol{\gamma}_{\beta,\kappa|\iota}^{(0)} = \boldsymbol{\gamma}_{\beta-1,\kappa|\iota}^{(t_{\max})}, \tag{8.22}$$

where the additional subscript denotes the frame index β. In the case of a constant CPSDM estimate this procedure allows the finite iterations in multiple frames to be used to solve the same problem i.e. a higher precision weight vector can be achieved.

Table 8.1: Transmission costs of distributed beamformers in dynamic sound fields. N denotes the number of nodes, K denotes the number of leaf nodes, r denotes the number of interferers, and $t_{\max}$ denotes the maximum number of iterations.

Beamformer Weight Vector Computation	
Algorithm	Transmissions per frame & frequency bin
BDLCMV/BDLCMP (Cyclic)	$t_{\max}(r+1)N$
BDLCMV/BDLCMP (Acyclic)	$\frac{1}{2}(r^2+3r+2)(N-1)+(r+1)(N-K)$
BDLCMV (Acyclic $\Delta\bar{\mathbf{P}}=\mathbf{0}$)	0
DLCMV (Acyclic) [14]	$(2N-1-K)$
DGSC (Acyclic) [15]	$(2N-1-K)+(r+1)(N-K)$
TI-DANSE (Cyclic) [18]	$(2N-1-K)(r+1)$
Beamformer Output Computation	
Algorithm	Transmissions per frame & frequency bin
Cyclic	$t_{\max}N$
Acyclic	$N-1$

In the case of slowly varying weight vectors, this allows the algorithm to track the optimal weight vector while still only incurring a finite iteration cost per frame.

A warm-start procedure cannot be used in the case of the beamformer output computation as it varies rapidly between frames. However, only a finite number of iterations are required per frame to achieve near-optimal performance. Thus, an iteration limit can be imposed to achieve a fully cyclic implementation. The performance of this iteration-limited output computation and the warm-started weight vector computation introduced above are demonstrated in Sec. 8.5.4.

8.4.8. Comparing the Transmission Costs of Different Beamformer Implementations

Table 8.1 includes the transmission costs of the distributed implementations of the BDLCMV/BDLCMP algorithm proposed in this paper. It is worth noting that these transmission costs do not include the additional overhead associated with those algorithms which exploit a TAD or the costs of forming a spanning tree. However, due to the per frequency bin nature of the algorithm, these costs are assumed to be far lower than those associated with running the algorithm.

From Table 8.1, our proposed acyclic implementation appears to require a notable increase in total transmission cost when we allow $\bar{\mathbf{P}}$ to vary. However unlike existing approaches, it does so while ensuring we exactly solve the problem in each frame. In contrast, the alternative methods listed require multiple frames to reach optimality [47]. As such, the proposed acyclic approach offers a competitive advantage as it exactly attains the performance of a centralized implementation in each frame while incurring a fixed transmission cost. In contrast, the iterative nature of

DLCMV, DGSC and TI-DANSE means that they require multiple frames to achieve the same precision, essentially scaling their effective transmission costs.

The proposed cyclic implementation of BDLCMV/BDLCMP, like other existing approaches within the literature [14, 15] allows for a tradeoff between per-frame optimality and communication overhead. Importantly, when combined with the warm-start procedure introduced in Eq. (8.22), this allows for near-optimal performance while reducing the total transmission overhead per frame. In particular, in Sec. 8.5.4 we will demonstrate the effect of combining this warm-start procedure with a single iteration, that is $t_{\max} = 1$. In this case, a negligible decrease in performance is achieved while incurring a transmission cost in line with existing acyclic distributed beamformers.

Finally, by providing two methods of beamformer output computation, we allow designers to implement a fully cyclic beamforming algorithm if they desire. Perhaps more attractive though is a hybrid style approach, similar to that used in [18], which combines cyclic weight vector computation with an acyclic output computation stage. This takes advantage of the transmission savings of both approaches while, as the acyclic topology can vary between frames, removes the need for acyclic network management in contrast to [14, 15].

8.5. Experimental Results

We compare the performance of the proposed beamformers (except of the BDALCMV, where an estimate of $\bar{\mathbf{P}}_{\mathbf{u}}$ is difficult to obtain), and six existing centralized beamformers (the MPDR, MVDR, LCMP, LCMV, LCDS and DS) in terms of noise suppression, predicted intelligibility improvement, robustness to RATF estimation errors and TAD errors. Table 8.2 summarizes the compared linearly constrained beamformers. Note that the ALCMV and ILCMV are not included in the comparisons since there are no distributed estimation methods of p_{iso}. Note that the MPDR, MVDR, LCMP, LCMV, LCDS and DS are distributable under the distributed LCMV (DLCMV) [14], as well as the distributed DS beamformer proposed in [9]. Specifically, we examine the performance of centralized implementations of the aforementioned beamformers to which their distributed counterparts converge [14].

8.5.1. Experiment Setup

The simulations are conducted in a simulated reverberant environment with reverberation times $\mathrm{T}_{60} = 0.2$ s and $\mathrm{T}_{60} = 0.5$ s using the image method [48]. A box-shaped room with dimensions $6 \times 4 \times 3$ is selected for the reverberant environment. The configuration of the nodes and acoustic sources are depicted in Fig. 8.3. We considered an example scenario where a number of people are sitting around a table with a set of mobile phones on the table, each equipped with multiple microphones. In this case, $N = 5$ nodes were placed on a virtual surface (with no physical properties) and four sources were placed around the surface. Each node was equipped with 3 microphones forming a uniform linear array with an inter-microphone distance of 2 cm. This resulted in a total of $M = 15$ microphones. Three of the four

Table 8.2: Summary of compared linearly constrained beamformers which are all special cases of the optimization problem in Eq. (8.5). Note that $\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$ is the constraints in Eq. (8.6).

Method	$\mathbf{P}$	Constraints	Target activity detection
MPDR	$\mathbf{P_y}$	$\mathbf{w}^H\mathbf{a} = 1$	no
MVDR	$\mathbf{P_n}$	$\mathbf{w}^H\mathbf{a} = 1$	yes
DS	$\mathbf{I}$	$\mathbf{w}^H\mathbf{a} = 1$	no
LCMP	$\mathbf{P_y}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	no
LCMV	$\mathbf{P_n}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	yes
LCDS	$\mathbf{I}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	no
BDLCMP	$\bar{\mathbf{P}}_\mathbf{y}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	no
BDLCMV	$\bar{\mathbf{P}}_\mathbf{n}$	$\mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H$	yes

sources were interferering talkers (2 female and 1 male) with the remainder being the target source (a male talker). Each signal had a simulated duration of 30 s and was sampled at $f_s = 16$ kHz. The power of each interferer at its original position was set to be approximately equal to the power of the target source at its original position (i.e., a 0 dB SNR). The impulse responses between microphones and sources were computed using the toolbox in [49], with length 200 ms. The closest microphone to the target was selected as the reference microphone (see Fig. 8.3). The microphone-self noise was white Gaussian noise with 40 dB SNR with respect to the target signal at the reference microphone.

As can be noted in Fig. 8.3, the distance between any two nodes was quite big (i.e., the distance between the closest microphone-pair, where the two microphones belonged to two different nodes, was at least 0.5091 m). Thus, the ambient noise was approximately spatially uncorrelated between different nodes. As explained in Section 8.1, the late reverberation, which is the main contribution in the ambient noise component, becomes approximately uncorrelated between two microphones with distance d above a certain threshold $f_c = c/(2d)$. Here, the distance of the closest microphone-pair where the microphones belong to two different nodes is 0.5091 m corresponding to $f_c = 333.9$ Hz (if $c = 340$ m/s). Note that the correlation between any other microphone-pair with microphones in different nodes will have even smaller f_c.

On the other hand, the late reverberation for microphones within a node is highly correlated. The distance between two consecutive microphones is $d = 0.02$ m and, resulting in $f_c = 8.5$ kHz, which is greater than $f_s/2 = 8$ kHz.

8.5.2. Processing

STFT frame-based beamforming was performed using an overlap and save (OLS) procedure [50]. We used a rectangular analysis window with length $2L_{\text{fr}} = 50$ ms,

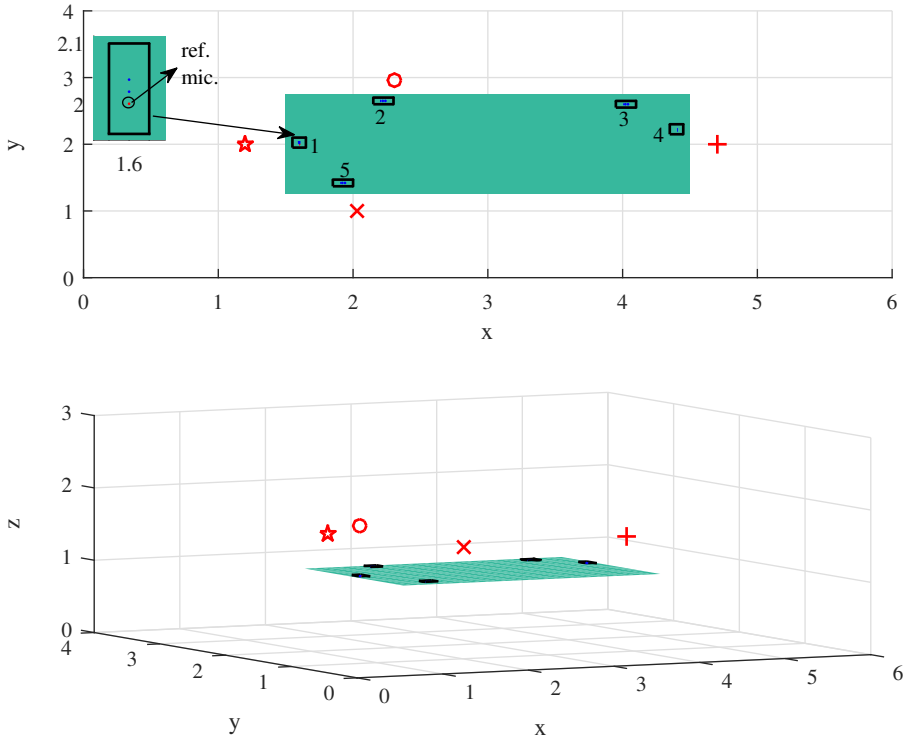

Figure 8.3: Experimental setup from two different angles: three interferers (two female talkers with markers '+' and 'x' and one male talker with marker 'o'), one target (a male talker with marker $\star$), and five nodes, with three microphones each, sitting on the virtual surface. The height of the virtual surface is 1 m.

where $L_{\text{fr}} = 25$ ms is the length of the current frame. Thus, the early-reverberant RATF vectors of the sources are associated with an impulse response of length 50 ms. The analysis window was applied on the current frame and the previous frame in order to a) mitigate circular convolution problems, and b) to be able to handle large phase shifts in the constraints due to the large microphone array aperture. The FFT length is $\Phi = 1024$.

In order to achieve a smoother processing than standard OLS, the analysis window was shifted by $L_{\text{fr}}/2$ samples[2]. A Hann window (synthesis window) was then applied, with length L_{fr}, on the last L_{fr} processed samples. Finally, the last $L_{\text{fr}}/2$ processed samples were saved in order to add them to the corresponding samples of the next windowed segment.

The CPSDMs, for the k-th frequency bin and β-th analysis segment, were estimated via recursive averaging as described in Section 8.2.2. Note that the

[2]The standard OLS procedure usually shifts the analysis window by L_{fr}.

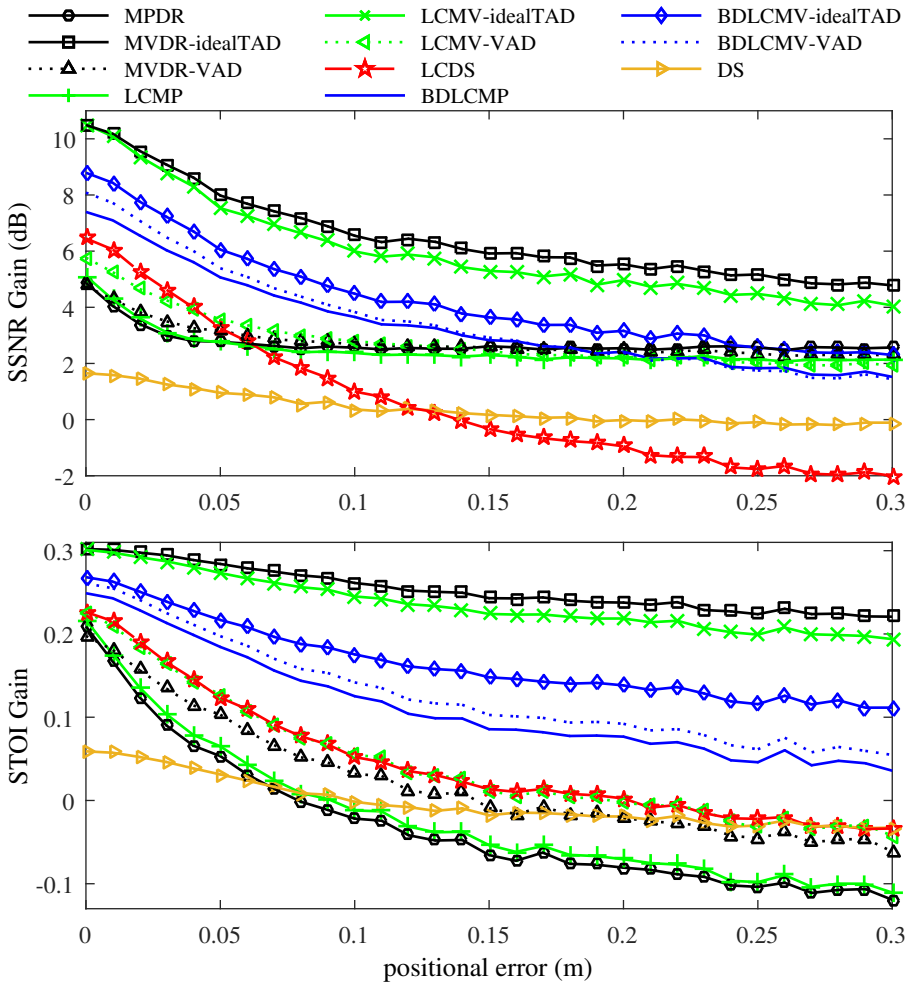

Figure 8.4: Reverberation time $T_{60} = 0.2$ s: Comparison of the beamformers in Table 8.2 as a function of positional error between training and testing positions. The methods that depend on a TAD are computed using an ideal TAD and the state-of-the-art voice activity detector (VAD) proposed in [51].

block-diagonal CPSDMs were recursively averaged locally at each node. The noise CPSDM and the block-diagonal noise CPSDM were estimated using an ideal TAD and a non-ideal state-of-the-art voice activity detector proposed in [51]. For simplicity, the TAD decision is based only on the reference microphone signal.

The RATF vectors were estimated once using additional 2 s recordings per source. Specifically, each talker spoke alone for 2 s, while all the others were silent. The CPSDM matrices of each talker were computed as described in Section 8.2.2

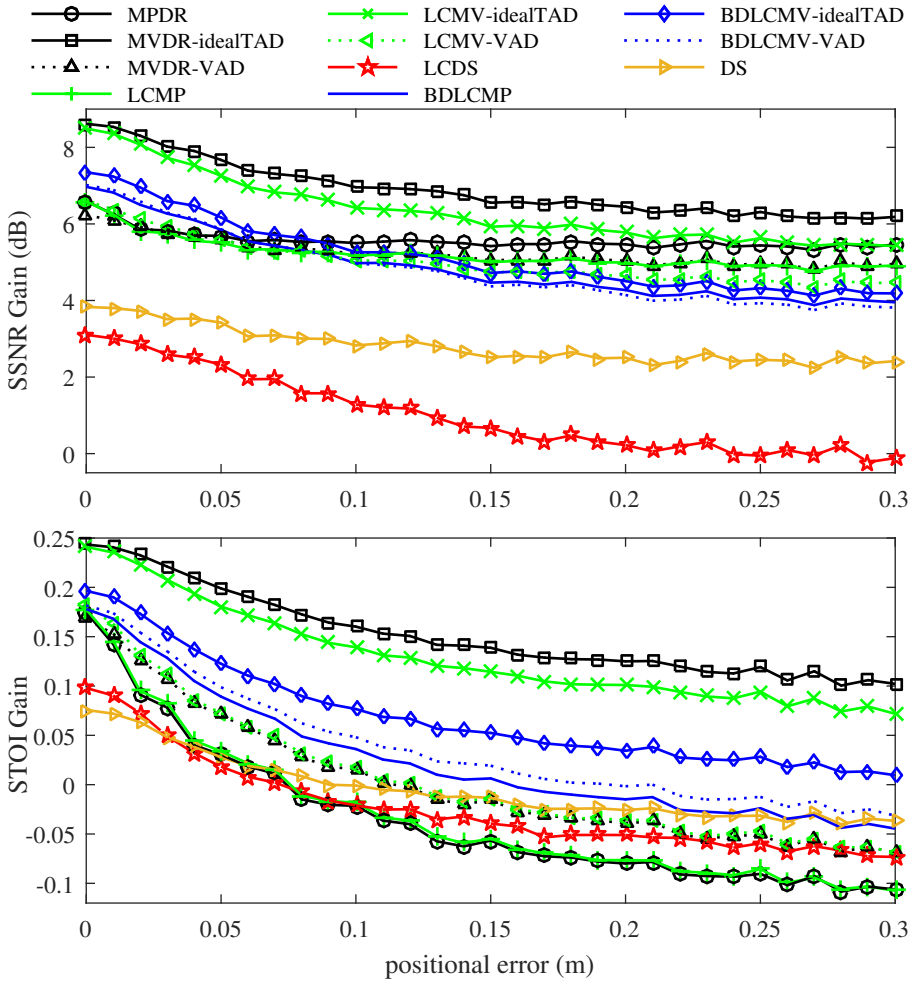

Figure 8.5: Reverberation time $T_{60} = 0.5$ s: Comparison of the beamformers in Table 8.2 as a function of positional error between training and testing positions. The methods that depend on a TAD are computed using an ideal TAD and the state-of-the-art voice activity detector (VAD) proposed in [51].

and the dominant relative eigenvector from each CPSDM was selected as an estimate of the RATF vector for each source[3]. These initial positions of the talkers, in which the RATF vectors were estimated, will be referred to as training positions and were nearby to the testing positions depicted in Fig. 8.3. Therefore, the RATF estimation errors of all sources can be modeled as a function of positional error

[3]If there is a noise component which is always active, such as an air-condition, a more accurate method of estimating the RATF of the talkers is by using the GEVD approach [32].

between the training positions and the testing positions.

8.5.3. Robustness to RATF estimation errors

Figs. 8.4 and 8.5 show the performance of the aforementioned beamformers in terms of segmental-signal-to-noise-ratio (SSNR) gain and the short-time objective intelligibility measure (STOI) [52] gain as a function of positional error for $T_{60} = 0.2$ s and $T_{60} = 0.5$ s, respectively. Note that the noise that is computed in the SSNR consists of the interferers, background, and target distortion noise. The erroneous training locations were uniformly distributed over a sphere centered around the true source locations having a radius ranging from $0 - 0.30$ m in 0.01 m steps. For every value of positional error, the average performance of 20 different setups was measured. Each setup used the same source signals at the same testing locations as shown in Fig. 8.3. However, a different set of initial training positions, computed as mentioned previously, were used in each setup. Likewise, different realizations of the microphone-self noise were also used in each setup.

It is clear that the proposed beamformers are more robust for the combination of large positional and TAD errors. Specifically, the BDLCMV and the BDLCMP provide significantly better predicted intelligibility improvement compared to all the other methods using a non-ideal TAD or not using a TAD. The BDLCMV with the non-ideal TAD is slightly better than the BDLCMP. Thus, in this particular scenario a TAD is not necessary for the proposed method, since it will create errors and the performance advantage will be small. Note that for $T_{60} = 0.5$ s and for large positional errors, the proposed methods achieve worse noise reduction, but better intelligibility improvement, than the other methods. As explained in Section 8.4, this is because the proposed beamformers distort the target signal much less than the other beamformers.

The LCMV using the non-ideal TAD is much more robust than the LCMP and gives much higher predicted intelligibility improvement. It is worth noting that for $T_{60} = 0.2$ s the fixed LCDS has almost the same predicted intelligibility improvement as the LCMV. This makes the usage of the LCMV beamformer, in this particular acoustic scenario, obsolete in the distributed context since LCDS has significantly lower communication costs. On the other hand, for $T_{60} = 0.5$ s the performance of LCDS deteriorates significantly and becomes also worse compared to the DS beamformer. Moreover, the MVDR using a non-ideal TAD has almost the same predicted intelligibility improvement with the LCMV using the non-ideal TAD for $T_{60} = 0.5$ s.

In conclusion, for those simulations using a non-ideal TAD, the proposed methods are the most robust out of those considered. Moreover, the proposed method incurs lower communication costs, as explained in Section 8.4, making it a strong candidate for distributed beamforming.

8.5.4. Limiting Iterations per Frame for PDMM Based BDLCMP/BDLCMV

We now compare the impact of a finite iteration cap on the optimality of both the computed beamformer weight vector and beamformer output signal. For these

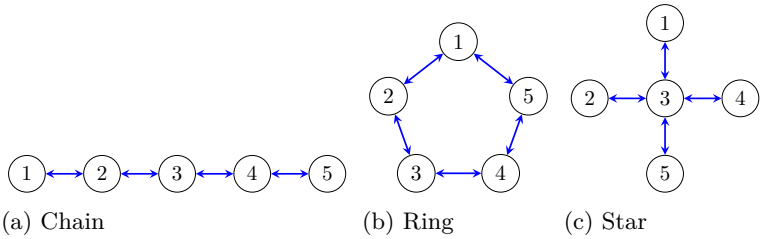

Figure 8.6: Chain, Ring and Star topologies for the considered five node network.

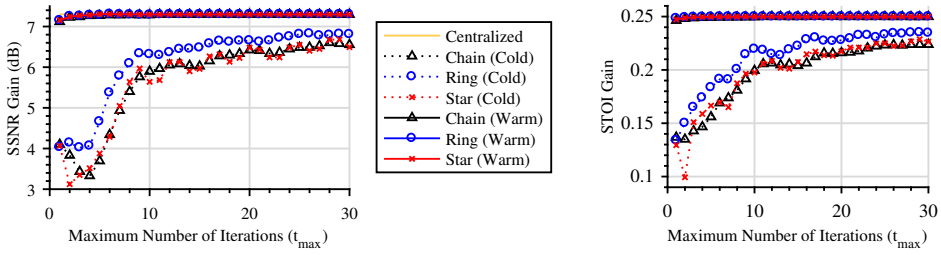

Figure 8.7: Comparing the effect of a finite iteration limit on PDMM beamformer weight vector computation. Cold-start (cold) and warm-start (warm) scenarios are considered with the beamformer output being computed exactly via acyclic data aggregation.

simulations, the same setup, as introduced in Sec. 8.5.1, was used. The case of BDLCMP with no RATF estimation errors was considered where by the centralized beamformers used previously were substituted with their cyclic counterparts introduced in Sec. 8.4.5. For these simulations, three standard network configurations (a chain, a ring and a star network) were considered to highlight the impact network topology can play on convergence. Examples of these three network topologies are included below in Figures 8.6a, 8.6b, 8.6c respectively. A step size of $\rho = \frac{1}{2}$ was heuristically selected for all simulations. With a more refined selection of this parameter, we expect that faster convergence could be achieved.

Fig. 8.7 shows a comparison of convergence rates of both cold and warm-started beamformer weight vector computation for the three networks considered. As expected, while all three methods require many iterations (> 30) to achieve reasonable weight vector estimation, when combined with a warm-start procedure, even a single iteration per frame achieves near optimal gains in both STOI and SSNR. Thus, for slowly varying CPSDM estimates, the cyclic BDLCMP/BDLCMV approach offers an opportunity to dramatically reduce transmission costs while maintaining near optimal performance. Furthermore, the effectiveness of this warm-start does not seem to vary significantly with network topology.

For beamformer output computation, as demonstrated in Fig. 8.8, the story is similar. While the dynamic nature of the beamformer output does not facilitate a warm-start procedure, the simplicity of the problem means that within 10 iterations or so a near optimal beamformer output is computed.

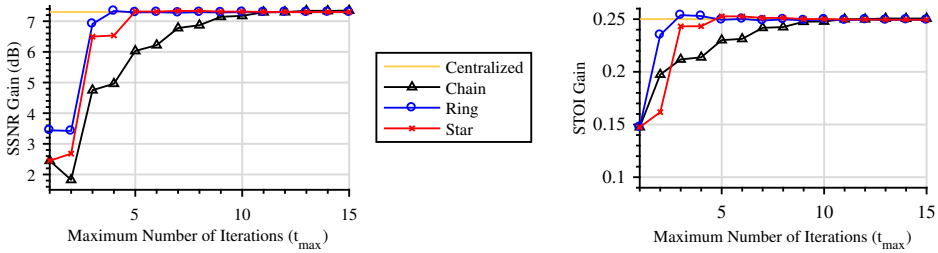

Figure 8.8: Comparing the effect of a finite iteration limit on PDMM beamformer output computation. For each of the networks considered the beamformer weight vector is computed exactly via acyclic data aggregation.

Unlike the beamformer weight vector computation, here we can more clearly observe the effect of network topology on convergence. In particular, the chain network, which has a larger diameter than either the ring or the star network, requires roughly twice the number of iterations to approach optimal convergence. This point is consistent with the fact that an even length chain network has twice the diameter of a ring network of the same size. However, this may be able to be remedied with more careful step size selection.

8.6. Conclusion

In this paper, we proposed a new distributed linearly constrained beamformer, which provides increased robustness to TAD and RATF estimation errors compared to traditional LCMV-based beamformers. Moreover, the proposed approach is immediately distributable due to its use of a block-diagonal CPSDM. Unlike most competing distributed beamformers, the proposed method can be applied in arbitrary network topologies, while at the same time having much lower communication costs in comparison to competing cyclic approaches and comparable costs to acyclic ones. Furthermore, the general nature of the distributed algorithm facilitates a trade off between transmission costs and per-frame optimality allowing it to be tailored to the needs of a particular application.

References

[1] B. D. Van Veen and K. M. Buckley, *Beamforming: A versatile approach to spatial filtering,* IEEE ASSP Mag. **5**, 4 (1988).

[2] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing* (John Wiley & Sons, 2004).

[3] S. A. Vorobyov, *Principles of minimum variance robust adaptive beamforming design,* ELSEVIER Signal Process. **93**, 3264 (2013).

[4] J. Benesty, M. M. Sondhi, and Y. Huang (Eds), *Springer handbook of speech processing* (Springer, 2008).

[5] M. Brandstein and D. Ward (Eds.), *Microphone arrays: signal processing techniques and applications* (Springer, 2001).

[6] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment* (John Wiley & Sons, 2006).

[7] S. Gannot, E. Vincet, S. Markovich-Golan, and A. Ozerov, *A consolidated perspective on multi-microphone speech enhancement and source separation,* IEEE Trans. Audio, Speech, Language Process. **25**, 692 (2017).

[8] A. Bertrand, *Applications and trends in wireless acoustic sensor networks: A signal processing perspective,* in *18th IEEE Symp. on Comm. and Vehicular Tech.* (2011) pp. 1–6.

[9] Y. Zeng and R. C. Hendriks, *Distributed delay and sum beamformer for speech enhancement via randomized gossip,* IEEE Trans. Audio, Speech, Language Process. **22**, 260 (2014).

[10] 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 Signal Enhancement (IWAENC)* (2012) pp. 1–4.

[11] M. O'Connor and W. B. Kleijn, *Diffusion-based distributed MVDR beamformer,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2014) pp. 810–814.

[12] 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).

[13] A. Bertrand and M. Moonen, *Distributed node-specific LCMV beamforming in wireless sensor networks,* IEEE Trans. Signal Process. **60**, 233 (2012).

[14] A. Bertrand and M. Moonen, *Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission,* IEEE Trans. Signal Process. **61**, 3447 (2013).

[15] S. Markovich, S. Gannot, and I. Cohen, *Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks,* IEEE Trans. Audio, Speech, Language Process. **21**, 343 (2013).

[16] 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).

[17] S. Doclo, M. Moonen, T. V. den Bogaert, and J. Wouters, *Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids,* IEEE Trans. Audio, Speech, Language Process. **17**, 38 (2009).

[18] J. Szurley, A. Bertrand, and M. Moonen, *Topology-independent distributed adaptive node-specific signal estimation in wireless sensor networks,* IEEE Trans. Signal and Info. Process. Over Networks **3**, 130 (2017).

[19] J. Capon, *High-resolution frequency-wavenumber spectrum analysis,* Proc. of the IEEE **57**, 1408 (1969).

[20] O. L. Frost III, *An algorithm for linearly constrained adaptive array processing,* Proc. of the IEEE **60**, 926 (1972).

[21] H. Cox, *Resolving power and sensitivity to mismatch of optimum array processors,* J. Acoust. Soc. Amer. **54**, 771 (1973).

[22] R. C. Hendriks and T. Gerkmann, *Noise correlation matrix estimation for multi-microphone speech enhancement,* IEEE Trans. Audio, Speech, Language Process. **20**, 223 (2012).

[23] H. Cox, *Robust adaptive beamforming,* IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**, 1365 (1987).

[24] B. D. Carlson, *Covariance matrix estimation errors and diagonal loading in adaptive arrays,* IEEE Trans. Aerosp. Electron. Systems **24**, 397 (1988).

[25] J. Li, P. Stoica, and Z. Wang, *On robust Capon beamforming and diagonal loading,* IEEE Trans. Signal Process. **51**, 1702 (2003).

[26] 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)* (IEEE, 2016) pp. 1088–1092.

[27] J. L. Flanagan, A. C. Surendran, and E. E. Jan, *Spatially selective sound capture for speech and audio processing,* ELSEVIER Speech Commun. **13**, 207 (1993).

[28] S. Gannot, D. Burshtein, and E. Weinstein, *Signal enhancement using beamforming and nonstationarity with applications to speech,* IEEE Trans. Signal Process. , 1614 (2001).

[29] J. S. Bradley, *Predictors of speech intelligibility in rooms,* J. Acoust. Soc. Amer. **80**, 837 (1986).

[30] J. S. Bradley and H. Sato, *On the importance of early reflections for speech in rooms,* J. Acoust. Soc. Amer. **113**, 3233 (2003).

[31] S. Gannot and I. Cohen, *Speech enhancement based on the general transfer function GSC and postfiltering,* IEEE Trans. Speech Audio Process. , 561 (2004).

[32] 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. , 1071 (2009).

[33] 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. **106**, 209 (2015).

[34] S. Braun and E. A. P. Habets, *Dereverberation in noisy environments using reference signals and a maximum likelihood estimator,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2013).

[35] M. Souden, J. Benesty, and S. Affes, *A study of the LCMV and MVDR noise reduction filters,* IEEE Trans. Signal Process. **58**, 4925 (2010).

[36] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[37] J. Bitzer, K. U. Simmer, and K. Kammeyer, *Theoretical noise reduction limits of the generalized sidelobe canceler (GSC) for speech enhancement,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vol. 5 (1999) pp. 2965–2968.

[38] I. A. McCowan and H. Bourlard, *Microphone array post-filter based on noise field coherence,* IEEE Trans. Audio, Speech, Language Process. **11**, 709 (2003).

[39] E. N. Gilbert and S. P. Morgan, *Optimum design of directive antenna arrays subject to random variations,* Bell Labs Technical Journal **34**, 637 (1955).

[40] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, *Instrumenting the world with wireless sensor networks,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vol. 4 (2001) pp. 2033–2036.

[41] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms,* IEEE Trans. on Information Theory **52**, 2508 (2006).

[42] D. Brandwood, *A complex gradient operator and its application in adaptive array theory,* IEE Proc. Pts. F and H **130**, 11 (1983).

[43] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization,* IEEE Trans. Automatic Control **54**, 48 (2009).

[44] 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 **3**, 1 (2011).

[45] G. Zhang and R. Heusdens, *Distributed optimization using the primal-dual method of multipliers,* IEEE Trans. Signal and Info. Process. Over Networks (2017).

[46] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, *On distributed averaging algorithms and quantization effects,* IEEE Trans. Automatic Control. **54**, 2506 (2009).

[47] S. Markovich, A. Bertrand, M. Moonen, and S. Gannot, *Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks,* ELSEVIER Signal Process. **107**, 4 (2015).

[48] J. B. Allen and D. A. Berkley, *Image method for efficiently simulating small-room acoustics,* J. Acoust. Soc. Amer. **65**, 943 (1979).

[49] E. A. P. Habets, *Room impulse response generator,* https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator/ (2010).

[50] J. J. Shynk, *Frequency-domain and multirate adaptive filtering,* IEEE Signal Process. Mag. **9**, 14 (1992).

[51] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, *Voice activity detection: Merging source and filter-based information,* IEEE Signal Process. Lett. **23**, 252 (2016).

[52] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, *An algorithm for intelligibility prediction of time-frequency weighted noisy speech,* IEEE Trans. Audio, Speech, Language Process. **19**, 2125 (2011).

9

Joint Estimation of the Multi-Microphone Signal Model Parameters

The worst form of inequality is to try to make unequal things equal.

Aristotle

This chapter is based on the article submitted as "Joint Estimation of the Multi-Microphone Signal Model Parameters", by A.I. Koutrouvelis, R.C. Hendriks, R. Heusdens and J. Jensen in IEEE/ACM Trans. Audio, Speech and Language Processing.

Microphone arrays (see e.g., [1] for an overview) are used extensively in many applications, such as source separation [2–6], multi-microphone noise reduction [1, 7–13], dereverberation [14–19], sound source localization [20–23], and room geometry estimation [24, 25]. All the aforementioned applications are based on a similar multi-microphone signal model, typically depending on the following parameters: i) the early relative acoustic transfer functions (RATFs) of the sources with respect to the microphones; ii) the power spectral densities (PSDs) of the early components of the sources, iii) the PSD of the late reverberation, and, iv) the PSDs of the microphone-self noise. Other parameters, like the target cross power spectral density matrix (CPSDM), the noise CPSDM, source locations and room geometry information, can be inferred from (combinations of) the above mentioned parameters. Often, none of these parameters are known *a priori*, while estimation is challenging. Often, only a subset of the parameters is estimated, see e.g., [14–17, 19, 26–30], typically requiring rather strict assumptions with respect to stationarity and/or knowledge of the remaining parameters.

In [15, 17] the target source PSD and the late reverberation PSD are jointly estimated assuming that the early RATFs of the target with respect to all microphones are known and all the remaining noise components (e.g., interferers) are stationary in time intervals typically much longer than a time frame. In [19, 26, 31], it was shown that the method in [15, 17] may lead to inaccurate estimates of the late reverberation PSD, when the early RATFs of the target include estimation errors. In [19, 26], a more accurate estimator for the late reverberation PSD was proposed, independent of early RATF estimation errors.

The methods proposed in [27, 28] do not assume that some noise components are stationary like in [17], but assume that the total noise CPSDM has a constant [27] or slow-varying [28] structure over time (i.e., it can be written as an unknown scaling parameter multiplied with a constant spatial structure matrix). This may not be realistic in practical acoustical scenarios, where different interfering sources change their power and location across time more rapidly and with different patterns. Moreover, these methods do not separate the late reverberation from the other noise components and only differentiate between the target source PSD and the overall noise PSD. As in [17], these methods assume that the early RATFs of the target are known. In [28], the structure of the noise CPSDM is estimated only in target-absent time-frequency tiles using a voice activity detector (VAD), which may lead to erroneous estimates if the spatial structure matrix of the noise changes during target-presence.

In [30], the early RATFs and the PSDs of all sources are estimated using the expectation maximization (EM) method [32]. This method assumes that only one source is active per time-frequency tile and the noise CPSDM (excluding the contributions of the interfering point sources) is estimated assuming it is time-invariant. Due to the time-varying nature of the late reverberation (included in the noise CPSDM), this assumption is often violated. This method does not estimate the time-varying PSD of the late reverberation, neither the PSDs of the microphone-self noise.

While the aforementioned methods focus on estimation of just one or several of the required model parameters, the method presented in [4] jointly estimates the early RATFs of the sources, the PSDs of the sources and the PSDs of the microphone-self noise. Unlike [30], the method in [4] does not assume single source activity per time-frequency tile and, thus, it is applicable to more general acoustic scenarios. The method in [4] is based on the non-orthogonal joint-diagonalization of the noisy CPSDMs. This method unfortunately does not guarantee non-negative estimated PSDs and, thus, the obtained target CPSDM may not be positive semidefinite resulting in performance degradation. Moreover, this approach does not estimate the PSD of the late reverberation. In conclusion, most methods only focus on the estimation of a subset of the required model parameters and/or rely on assumptions which may be invalid and/or impractical.

In this paper, we propose a method which jointly estimates all the aforementioned parameters of the multi-microphone signal model. The proposed method is based on confirmatory factor analysis (CFA) [33–36] and on the non-orthogonal joint-diagonalization principle introduced in [4]. The combination of these two theories and the adjustment to the multi-microphone case gives us a robust method, which is applicable for temporally and spatially non-stationary sources. The proposed method uses linear constraints to reduce the feasibility set of the parameter space and thus increase robustness. Moreover, the proposed method guarantees positive estimated PSDs and, thus, positive semidefinite target and noise CPSDMs. Although generally applicable, in this manuscript, we will compare the performance of the proposed method with state-of-the-art approaches in the context of source separation and dereverberation.

The remaining of the paper is organized as follows. In Sec. 9.1, the signal model, notation and used assumptions are introduced. In Sec. 9.2, we review the CFA theory and its relation to the non-orthogonal joint diagonalization principle. In Sec. 9.3, the proposed method is introduced. In Sec. 9.4, we introduce several constraints to increase the robustness of the proposed method. In Sec. 9.5, we discuss about the implementation and practicality of the proposed method. In Sec. 9.6, we conduct experiments in several multi-microphone applications using the proposed method and existing state-of-the-art approaches. In Sec. 9.7, we draw conclusions.

9.1. Preliminaries

9.1.1. Notation

We use lower-case letters for scalars, bold-face lower-case letters for vectors, and bold-face upper-case letters for matrices. A matrix $\mathbf{A}$ can be expressed as $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_m]$, where $\mathbf{a}_i$ is its i-th column. The elements of a matrix $\mathbf{A}$ are denoted as a_{ij}. We use the operand $\mathrm{tr}(\cdot)$ to denote the trace of a matrix, $\mathrm{E}[\cdot]$ to denote the expected value of a random variable, $\mathrm{diag}(\mathbf{A}) = [a_{11}, \cdots, a_{mm}]^T$ to denote the vector formed from the diagonal of a matrix $\mathbf{A} \in \mathbb{C}^{m \times m}$, and $||\cdot||_F^2$ to denote the Frobenius norm of a matrix. We use $\mathrm{Diag}(\mathbf{v})$ to form a square diagonal matrix with diagonal $\mathbf{v}$. A hermitian positive semi-definite matrix is expressed as $\mathbf{A} \succeq 0$,

where $\mathbf{A} = \mathbf{A}^H$ and its eigenvalues are real non-negative. The cardinality of a set is denoted as $|\cdot|$. The minimum element of a vector $\mathbf{v}$ is obtained via the operation $\min(\mathbf{v})$.

9.1.2. Signal Model

Consider an M-element microphone array of arbitrary structure within a possibly reverberant enclosure, in which there are r acoustic point sources (target and interfering sources). The i-th microphone signal (in the short-time Fourier transform (STFT) domain) is modeled as

$$y_i(t,k) = \sum_{j=1}^{r} e_{ij}(t,k) + \sum_{j=1}^{r} l_{ij}(t,k) + v_i(t,k), \tag{9.1}$$

where k is the frequency-bin index; t the time-frame index; e_{ij} and l_{ij} the early and late components of the j-th point source, respectively; and v_i denotes the microphone self-noise. The early components include the line of sight and a few initial strong reflections. The late components describe the effect of the remaining reflections and are usually referred to as late reverberation. The j-th early component is given by

$$e_{ij}(t,k) = a_{ij}(\beta,k)s_j(t,k), \tag{9.2}$$

where $a_{ij}(\beta,k)$ is the corresponding RATF with respect to the i-th microphone, $s_j(t,k)$ the j-th point-source at the reference microphone, β is the index of a *time-segment*, which is a collection of *time-frames*. That is, we assume that the source signal can vary faster (from time-frame to time-frame) than the early RATFs, which are assumed to be constant over multiple time-frames (which we call a time-segment). By stacking all microphone recordings into vectors, the multi-microphone signal model is given by

$$\mathbf{y}(t,k) = \sum_{j=1}^{r} \underbrace{\mathbf{a}_j(\beta,k)s_j(t,k)}_{\mathbf{e}_j(t,k)} + \underbrace{\sum_{j=1}^{r} \mathbf{l}_j(t,k)}_{\mathbf{l}(t,k)} + \mathbf{v}(k) \in \mathbb{C}^{M\times 1}, \tag{9.3}$$

where $\mathbf{y}(t,k) = [y_1(t,k), \cdots, y_M(t,k)]^T$ and all the other vectors can be similarly represented. If we assume that all sources in (9.3) are mutually uncorrelated and stationary within a time-frame, the signal model of the CPSDM of the noisy recordings is given by

$$\mathbf{P}_{\mathbf{y}}(t,k) = \sum_{j=1}^{r} \mathbf{P}_{\mathbf{e}_j}(t,k) + \mathbf{P}_{\mathbf{l}}(t,k) + \mathbf{P}_{\mathbf{v}}(k) \in \mathbb{C}^{M\times M}, \tag{9.4}$$

where $\mathbf{P}_{\mathbf{e}_j} = p_j(t,k)\mathbf{a}_j(\beta,k)\mathbf{a}_j^H(\beta,k)$, $p_j = E[|s_j(t,k)|^2]$ is the PSD of the j-th source at the reference microphone, $\mathbf{P}_{\mathbf{l}}(t,k)$ the CPSDM of the late reverberation and $\mathbf{P}_{\mathbf{v}}(k)$ is a diagonal matrix, which has as its diagonal elements the PSDs of the microphone-self noise. Note that $p_j(t,k)$ and $\mathbf{P}_{\mathbf{l}}(t,k)$ are time-frame varying, while

the microphone-self noise PSDs are typically time-invariant. The CPSDM model in (9.4) can be re-written as

$$\mathbf{P}_{\mathbf{y}}(t,k) = \mathbf{P}_{\mathbf{e}}(t,k) + \mathbf{P}_{\mathbf{l}}(t,k) + \mathbf{P}_{\mathbf{v}}(k), \tag{9.5}$$

where $\mathbf{P}_{\mathbf{e}}(t,k) = \mathbf{A}(\beta,k)\mathbf{P}(t,k)\mathbf{A}^H(\beta,k)$ and $\mathbf{A}(\beta,k) \in \mathbb{C}^{M\times r}$ is commonly referred to as mixing matrix and has as its columns the early RATFs of the sources. As we work with RATFs, the row of $\mathbf{A}(\beta,k)$ corresponding to the reference microphone is equal to a vector with only ones. Moreover, $\mathbf{P}(t,k)$ is a diagonal matrix, where $\text{diag}\left(\mathbf{P}(t,k)\right) = \left[p_1(t,k),\cdots,p_r(t,k)\right]^T$.

9.1.3. Late Reverberation Model

A commonly used assumption (adopted in this paper) is that the late reverberation CPSDM has a known spatial structure, $\mathbf{\Phi}(k)$, which is time-invariant but varying over frequency [14, 17]. Under the constant spatial-structure assumption, $\mathbf{P}_{\mathbf{l}}(t,k)$ is modeled as [14, 17]

$$\mathbf{P}_{\mathbf{l}}(t,k) = \gamma(t,k)\mathbf{\Phi}(k), \tag{9.6}$$

with $\gamma(t,k)$ the PSD of the late reverberation which is unknown and needs to be estimated. By combining (9.5), and (9.6), we obtain the final CPSDM model given by

$$\mathbf{P}_{\mathbf{y}}(t,k) = \mathbf{P}_{\mathbf{e}}(t,k) + \gamma(t,k)\mathbf{\Phi}(k) + \mathbf{P}_{\mathbf{v}}(k). \tag{9.7}$$

There are several existing methods [15, 17–19, 26] to estimate $\gamma(t,k)$ under the assumption that $\mathbf{\Phi}(k)$ is known. There are mainly two methodologies of obtaining $\mathbf{\Phi}(k)$. The first is to use many pre-calculated impulse responses measured around the array as in [7]. The second is to use a model which is based on the fact that the off-diagonal elements of $\mathbf{\Phi}(k)$ depend on the distance between every microphone pair. The distances between any two microphone pairs is described by the symmetric microphone-distance matrix $\mathbf{D}$ with elements d_{ij} which is the distance between microphones i and j. Two commonly used models for the spatial structure are the cylindrical and spherical isotropic noise fields [10, 37]. The cylindrical isotropic noise field is accurate for rooms where the ceiling and the floor are more absorbing than the walls. These models are accurate for sufficiently large rooms [10].

9.1.4. Estimation of CPSDMs Using Sub-Frames

The estimation of $\mathbf{P}_{\mathbf{y}}(t,k)$, is achieved using overlapping multiple *sub-frames*. The set of all used sub-frames within the t-th time-frame is denoted by Θ_t, and the number of used sub-frames is $|\Theta_t|$. We assume that the noisy microphone signals within a time-frame are stationary and, thus, we can estimate the noisy CPSDM using the sample CPSDM, i.e.,

$$\hat{\mathbf{P}}_{\mathbf{y}}(t,k) = \frac{1}{|\Theta_t|}\sum_{\theta\in\Theta_t}\mathbf{y}_\theta(t,k)\mathbf{y}_\theta^H(t,k), \tag{9.8}$$

with θ the sub-frame index. Fig. 9.1 summarizes how we split time using sub-frames, time-frames and time-segments.

9.1.5. Problem Formulation

The goal of this paper is to jointly estimate the parameters $\mathbf{A}(\beta,k)$, $\mathbf{P}(t,k)$, $\gamma(t,k)$, and $\mathbf{P}_{\mathbf{v}}(k)$ for the β-th time-segment of the signal model in (9.7) by only having estimates of the noisy CPSDM matrices $\hat{\mathbf{P}}_{\mathbf{y}}(t,k)$ for all time frames belonging to the β-th time-segment and possibly having an estimate $\hat{\mathbf{\Phi}}(k)$ and/or $\hat{\mathbf{D}}$. From now on, we will neglect time-frequency indices to simplify notation wherever is necessary.

9.2. Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) [33, 34, 36] aims at estimating the parameters of the following CPSDM model:

$$\mathbf{P}_{\mathbf{y}} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \mathbf{P}_{\mathbf{v}} \in \mathbb{C}^{M\times M}, \tag{9.9}$$

where $\mathbf{P}_{\mathbf{v}} = \text{Diag}([q_1,\cdots,q_M]^T)$ and $\mathbf{P} \succeq 0$. In CFA, some of the elements in $\mathbf{A}$ and $\mathbf{P}$ are fixed such that the remaining variables are uniquely identifiable (see below). More specifically, let Υ and $\mathcal{K}$ denote the sets of the selected row-column index-pairs of the matrices $\mathbf{A}$ and $\mathbf{P}$, respectively, where their elements are fixed to some known constants $\tilde{a}_{ij}$, and $\tilde{p}_{kr}$.

There are several existing CFA methods (see e.g. [36], for an overview). Most of these are special cases of the following general CFA problem

$$\begin{aligned}
\hat{\mathbf{A}},\hat{\mathbf{P}},\hat{\mathbf{P}}_{\mathbf{v}} = \underset{\mathbf{A},\mathbf{P},\mathbf{P}_{\mathbf{v}}}{\arg\min}\ & F(\hat{\mathbf{P}}_{\mathbf{y}},\mathbf{P}_{\mathbf{y}}) \\
\text{s.t.}\quad & \mathbf{P}_{\mathbf{y}} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \mathbf{P}_{\mathbf{v}}, \\
& \mathbf{P}_{\mathbf{v}} = \text{Diag}([q_1,\cdots,q_M]^T), \\
& q_i \geq 0,\ i=1,\cdots,M, \\
& \mathbf{P} \succeq 0, \\
& a_{ij} = \tilde{a}_{ij},\ \forall (i,j)\in\Upsilon, \\
& p_{kr} = \tilde{p}_{kr},\ \forall (k,r)\in\mathcal{K},
\end{aligned} \tag{9.10}$$

with $F(\hat{\mathbf{P}}_{\mathbf{y}},\mathbf{P}_{\mathbf{y}})$ a cost function, which is typically one of the following cost functions: maximum likelihood (ML), least squares (LS), or generalized least squares (GLS). That is,

$$F(\hat{\mathbf{P}}_{\mathbf{y}},\mathbf{P}_{\mathbf{y}}) = \begin{cases} \text{(ML): } \log|\mathbf{P}_{\mathbf{y}}| + \text{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right), & [34], \\ \text{(LS): } \frac{1}{2}||\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}||_F^2, & [36, 38], \\ \text{(GLS): } \frac{1}{2}||\hat{\mathbf{P}}_{\mathbf{y}}^{-\frac{1}{2}}(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}})\hat{\mathbf{P}}_{\mathbf{y}}^{-\frac{1}{2}}||_F^2, & [39], \end{cases} \tag{9.11}$$

where $\mathbf{P}_{\mathbf{y}}$ is given in (9.9). Notice, that the problem in (9.10) is not convex (due to the non-convex terms $\mathbf{A}\mathbf{P}\mathbf{A}^H$) and may have multiple local minima.

There are two necessary conditions for the parameters of the CPSDM model in (9.9) to be uniquely identifiable[1]. The *first identifiability condition* states that the

[1] We say that the parameters of a function are uniquely identifiable if there is one-to-one relationship between the parameters and the function value.

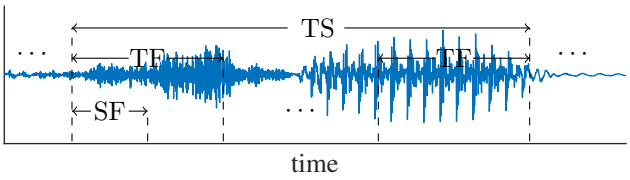

Figure 9.1: Splitting time into time-segments (TS), time-frames (TF), and sub-frames (SF).

number of equations should be larger than the number of unknowns [36, 40]. Since $\hat{\mathbf{P}}_{\mathbf{y}} \succeq 0$, there are $M(M+1)/2$ known values, while there are $Mr - |\Upsilon|$ unknowns due to $\mathbf{A}$, $r(r+1)/2 - |\mathcal{K}|$ unknowns due to $\mathbf{P}$ (because $\mathbf{P} \succeq 0$), and M unknowns due to $\mathbf{P}_{\mathbf{v}}$ (because $\mathbf{P}_{\mathbf{v}}$ is diagonal). Therefore, the first identifiability condition is given by [40]

$$\frac{M(M+1)}{2} \geq Mr + \frac{r(r+1)}{2} - |\Upsilon| - |\mathcal{K}| + M. \tag{9.12}$$

The identifiability condition in (9.12) is not sufficient for guaranting unique identifiability [36]. Specifically, for any arbitary non-singular matrix $\mathbf{T} \in \mathbb{C}^{r \times r}$, we have $\mathbf{P}_{\mathbf{y}}(\mathbf{A}, \mathbf{P}, \mathbf{P}_{\mathbf{v}}) = \mathbf{P}_{\mathbf{y}}(\mathbf{A}\mathbf{T}^{-1}, \mathbf{T}\mathbf{P}\mathbf{T}^{H}, \mathbf{P}_{\mathbf{v}})$ and, therefore [34]

$$F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{A}, \mathbf{P}, \mathbf{P}_{\mathbf{v}}) = F(\hat{\mathbf{P}}_{\mathbf{y}}, \underbrace{\mathbf{A}\mathbf{T}^{-1}}_{\tilde{\mathbf{A}}}, \underbrace{\mathbf{T}\mathbf{P}\mathbf{T}^{H}}_{\tilde{\mathbf{P}}}, \mathbf{P}_{\mathbf{v}}). \tag{9.13}$$

This means that there are infinitly many optimal solutions $(\tilde{\mathbf{A}}, \tilde{\mathbf{P}} \succeq 0)$ of the problem in (9.10). Since there are r^2 variables in $\mathbf{T}$, the *second identifiability condition* of the CPSDM model in (9.9) states that we need to fix at least r^2 of the parameters in $\mathbf{A}$ and $\mathbf{P}$ [34, 40], i.e.,

$$|\Upsilon| + |\mathcal{K}| \geq r^2. \tag{9.14}$$

This second condition is necessary but not sufficient, since we need to fix the proper parameters and not just any r^2 parameters [34, 40] such that $\mathbf{T} = \mathbf{I}$. For a general full-element $\mathbf{P}$, a recipe on how to select the r^2 constraints in order to achieve unique identifiability is provided in [34].

9.2.1. Simultaneous CFA (SCFA) in Multiple Time-Frames

The β-th time-segment consists of the following $|\mathcal{B}_\beta|$ time-frames: $t = \beta|\mathcal{B}_\beta| + 1, \cdots, (\beta+1)|\mathcal{B}_\beta|$, where $\mathcal{B}_\beta$ is the set of the time-frames in the β-th time-segment. For ease of notation, we can alternatively re-write this as $\forall t \in \mathcal{B}_\beta$. The problem in (9.10) considered $|\mathcal{B}_\beta| = 1$ time-frame. Now we assume that we estimate $\hat{\mathbf{P}}_{\mathbf{y}}(t)$ for $|\mathcal{B}_\beta| \geq 1$ time-frames in the β-th time-segment. We also assume that $\forall (t_i, t_j) \in \mathcal{B}_\beta, \hat{\mathbf{P}}_{\mathbf{y}}(t_i) \neq \hat{\mathbf{P}}_{\mathbf{y}}(t_j)$, if $i \neq j$. Recall that the mixing matrix $\mathbf{A}$ is assumed to be static within a time-segment. Moreover, $\mathbf{P}_{\mathbf{v}}$ is time-invariant and, thus, shared among different time-frames within the same time-segment. One can exploit these two facts in order to increase the ratio between the number of equations and the number of unknown parameters [33, 35] and thus satisfy the first and

second identifiability conditions with less microphones. This can be done by solving the following general simultaneous CFA (SCFA) problem [35]

$$\begin{aligned}\hat{\mathbf{A}}, \{\hat{\mathbf{P}}(t)\}, \hat{\mathbf{P}}_{\mathbf{v}} = & \underset{\mathbf{A}, \{\mathbf{P}(t)\}, \mathbf{P}_{\mathbf{v}}}{\arg\min} \sum_{\forall \tau \in \mathcal{B}_\beta} F(\hat{\mathbf{P}}_{\mathbf{y}}(\tau), \mathbf{P}_{\mathbf{y}}(\tau)) \\ \text{s.t.} \quad & \mathbf{P}_{\mathbf{y}}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H + \mathbf{P}_{\mathbf{v}}, \ \forall t \in \mathcal{B}_\beta, \\ & \mathbf{P}_{\mathbf{v}} = \text{Diag}([q_1, \cdots, q_M]^T), \\ & q_i \geq 0, \ i = 1, \cdots, M, \\ & \mathbf{P}(t) \succeq 0, \forall t \in \mathcal{B}_\beta, \\ & a_{ij} = \tilde{a}_{ij}, \ \forall (i,j) \in \Upsilon, \\ & p_{kr}(t) = \tilde{p}_{kr}(t), \ \forall (k,r) \in \mathcal{K}_t, \ \forall t \in \mathcal{B}_\beta. \end{aligned} \tag{9.15}$$

The CFA problem in (9.10) is a special case of SCFA, when we select $|\mathcal{B}_\beta| = 1$. The first identifiability condition for the SCFA problem becomes

$$|\mathcal{B}_\beta| \frac{M(M+1)}{2} \geq Mr + |\mathcal{B}_\beta| \frac{r(r+1)}{2} - |\Upsilon| - \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| + M. \tag{9.16}$$

We conclude from (9.12) and (9.16) that the SCFA problem (for $|\mathcal{B}_\beta| > 1$) needs less microphones compared to the problem in (9.10) to satisfy the first identifiability condition, assuming both problems have the same number of sources. Moreover, the second identifiability condtion in the SCFA problem becomes

$$|\Upsilon| + \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| \geq r^2. \tag{9.17}$$

From (9.14) and (9.17), we conclude that the SCFA problem (for $|\mathcal{B}_\beta| > 1$) satisfies easier the second identifiability condition compared to the problem in (9.10), if both problems have the same number of sources and microphones.

9.2.2. Special Case (S)CFA: $\mathbf{P}(t)$ is Diagonal

A special case of (S)CFA, which is more suitable for the application at hand, is when $\mathbf{P}(t)$, $\forall t \in \mathcal{B}_\beta$ are constrained to be diagonal due to the signal model in (9.5). We refer to this special case as the diagonal (S)CFA problem. By constraining $\mathbf{P}(t)$ to be diagonal, the total number of fixed parameters in $\mathbf{A}, \mathbf{P}(t), \forall t \in \mathcal{B}_\beta$ is

$$|\Upsilon| + \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| = |\Upsilon| + |\mathcal{B}_\beta| (\frac{r^2}{2} - \frac{r}{2}). \tag{9.18}$$

It has been shown in [41, 42] that in this case, and for $r > 1$, the class of the only possible $\mathbf{T}$ is $\mathbf{T} = \mathbf{\Pi}\mathbf{S}$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{S}$ is a scaling matrix, if the following condition is satisfied

$$2\kappa_{\mathbf{A}} + \kappa_{\mathbf{Z}} \geq 2(r+1), \tag{9.19}$$

where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_{|\mathcal{B}_\beta|} \end{bmatrix}, \quad \mathbf{z}_t = \operatorname{diag}\left(\mathbf{P}(t)\right), t \in \mathcal{B}_\beta, \tag{9.20}$$

and $\kappa_{\mathbf{A}}, \kappa_{\mathbf{Z}}$ are the Kruskal-ranks [41] of the matrices $\mathbf{A}$ and $\mathbf{Z}$, respectively. We conclude, that if (9.16) is satisfied, and there are at least r^2 fixed variables in $\mathbf{A}$ and $\mathbf{P}(t), \forall t \in \mathcal{B}_\beta$, and the condition in (9.19) is satisfied, then the parameters of (9.9) (for $\mathbf{P}(t)$ diagonal) will be uniquely identifiable up to a possible scaling and/or permutation.

9.2.3. Diagonal SCFA vs Non-Orthogonal Joint Diagonalization

The diagonal SCFA problem in Sec. 9.2.2 is very similar to the joint diagonalization method in [4] apart from the two positive semidefinite constraints that avoid improper solutions, and which are lacking in [4]. Finally, it is worth mentioning that the method proposed in [4] solves the scaling ambiguity by setting $a_{ii} = 1$ (corresponding to a varying reference microphone per-source), which means r fixed elements in $\mathbf{A}$, i.e., $|\Upsilon| = r$. Therefore, in [4], the total number of fixed parameters in $\mathbf{A}, \mathbf{P}(t), \forall t \in \mathcal{B}_\beta$ is given by

$$|\Upsilon| + \sum_{\forall t \in \mathcal{B}_\beta} |\mathcal{K}_t| = r + |\mathcal{B}_\beta|(\frac{r^2}{2} - \frac{r}{2}). \tag{9.21}$$

By combining (9.21) and (9.17), the second identifiability condition becomes

$$r + |\mathcal{B}_\beta|(\frac{r^2}{2} - \frac{r}{2}) \geq r^2. \tag{9.22}$$

Note that for $r \geq 2$, if $|\mathcal{B}_\beta| \geq 2$, the second identifiability condition is always satisfied, but the permutation ambiguity still exists and needs extra steps to be resolved [4]. However, for $r = 1$, the second identifiability condition is satisfied for $|\mathcal{B}_\beta| \geq 1$ and there are no permutation ambiguities. By combining (9.21), and (9.16), the first identifiability condition for the diagonal SCFA with $|\Upsilon| = r$ becomes

$$|\mathcal{B}_\beta| \frac{M(M+1)}{2} \geq Mr + |\mathcal{B}_\beta| r - r + M. \tag{9.23}$$

9.3. Proposed Diagonal SCFA Problems

In this section, we will propose two methods based on the diagonal SCFA problem from Sec. 9.2.2 to estimate the different signal model parameters in (9.7). Unlike the diagonal SCFA problem and the non-orthogonal joint diagonalization method in [4], the first proposed method also estimates the late reverberation PSD. The second proposed method skips the estimation of the late reverberation PSD and thus is more similar to the diagonal SCFA problem and the non-orthogonal joint diagonalization method in [4]. Since we are using the early RATFs as columns of $\mathbf{A}$, we fix all the elements of the ρ-th row of $\mathbf{A}$ equal to 1, where ρ is the reference microphone index. Thus, unlike the method proposed in [4], which uses a varying

reference microphone (i.e., $a_{ii} = 1$), we use a single reference microphone (i.e., $a_{\rho j} = 1$).

Although our proposed constraints $a_{\rho j} = 1$ will resolve the scaling ambiguity (described in Sec 9.2.2), the permutation ambiguity (described in Sec 9.2.2) still exists and needs extra steps to be resolved. In this paper, we do not focus on this problem and we assume that we know the perfect permutation matrix per time-frequency tile. The interested reader can find more information on how to solve permutation ambiguities in [4–6]. An exception occurs in the context of dereverberation where, typically, a single point source (i.e., $r = 1$) exists and, therefore, a single fixed parameter in $\mathbf{A}$ is sufficient to solve both the permutation and scaling ambiguities.

9.3.1. Proposed Basic Diagonal SCFA Problem

The proposed basic diagonal SCFA problem is based on the signal model in (9.7), which takes into account the late reverberation. Here we assume that we have computed a priori $\hat{\mathbf{\Phi}}$. The proposed diagonal SCFA problem is given by

$$\begin{aligned}
\hat{\mathbf{A}}, \{\hat{\mathbf{P}}(t)\}, \hat{\mathbf{P}}_{\mathbf{v}}, \{\hat{\gamma}(t)\} = \underset{\substack{\mathbf{A}, \{\mathbf{P}(t)\}, \\ \mathbf{P}_{\mathbf{v}}, \{\gamma(t)\}}}{\arg\min} & \sum_{\forall \tau \in \mathcal{B}_\beta} F(\hat{\mathbf{P}}_{\mathbf{y}}(\tau), \mathbf{P}_{\mathbf{y}}(\tau)) \\
\text{s.t.} \quad & \mathbf{P}_{\mathbf{y}}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H + \gamma(t)\hat{\mathbf{\Phi}} + \mathbf{P}_{\mathbf{v}}, \ \forall t \in \mathcal{B}_\beta \\
& \mathbf{P}_{\mathbf{v}} = \text{Diag}([q_1, \cdots, q_M]^T), \\
& q_i \geq 0, \ i = 1, \cdots, M, \\
& \mathbf{P}(t) = \text{Diag}([p_1(t), \cdots, p_r(t)]^T), \ \forall t \in \mathcal{B}_\beta, \\
& p_j(t) \geq 0, \ \forall t \in \mathcal{B}_\beta, \ j = 1, \cdots, r, \\
& \gamma(t) \geq 0, \ \forall t \in \mathcal{B}_\beta, \\
& a_{\rho j} = 1, \ \text{for } j = 1, \cdots, r.
\end{aligned} \tag{9.24}$$

We will refer to the problem in (9.24) as the SCFA_{rev} problem. The objective function of the SCFA_{rev} problem depends on $\gamma(t)$. This means that we have $|\mathcal{B}_\beta|$ additional unknowns in (9.23). Thus, the first identifiability condition becomes

$$|\mathcal{B}_\beta| \frac{M(M+1)}{2} \geq Mr + |\mathcal{B}_\beta| r - r + |\mathcal{B}_\beta| + M. \tag{9.25}$$

A simplified version of the SCFA_{rev} problem is obtained when the reverberation parameter γ is omitted. This problem therefore uses the signal model of (9.9) instead of (9.7). We will refer to this simplified problem as the $\text{SCFA}_{\text{no-rev}}$ problem. The only differences between the $\text{SCFA}_{\text{no-rev}}$ and the method proposed [4], is that in the $\text{SCFA}_{\text{no-rev}}$ we use a fixed reference microphone and positivity constraints for the PSDs.

Since, we have r fixed parameters in $\mathbf{A}$ corresponding to the reference microphone, in both proposed methods, the total number of fixed parameters in $\mathbf{A}$ and $\mathbf{P}(t), \forall t \in \mathcal{B}_\beta$ is the same as in (9.21). The second identifiability condition of all proposed methods is therefore the same as in (9.22).

9.3.2. SCFA_{REV} versus $\text{SCFA}_{\text{NO-REV}}$

Although the SCFA_{rev} method typically fits a more accurate signal model to the noisy measurements compared to the $\text{SCFA}_{\text{no-rev}}$ method, it does not necessarily guarantee a better performance over the $\text{SCFA}_{\text{no-rev}}$ method. In other words, the *model-mismatch* error is not the only critical factor in achieving good performance. Another important factor is how *over-determined* is the system of equations to be solved is, i.e., what is the ratio of knowns and unknowns. With respect to the over-determination factor, the $\text{SCFA}_{\text{no-rev}}$ method is more efficient, since it has less parameters to estimate, if $\mathcal{B}_\beta$ is the same in both methods. Consequently, the problem boils down to how much is the model-mismatch error and the over-determination. Thus, it is natural to expect that for not highly reverberant environments, the $\text{SCFA}_{\text{no-rev}}$ method may perform better than the SCFA_{rev} method, while for highly reverberant environments the inverse may hold.

9.4. Robust Estimation of Parameters

In Secs. 9.4.1—9.4.5, we propose additional constraints in order to increase the robustness of the initial versions of the two diagonal SCFA problems proposed in Sec. 9.3. The robustness is needed in order to overcome CPSDM estimation errors and model-mismatch errors. We use linear inequality constraints (mainly simple box constraints) on the parameters to be estimated. These constraints limit the feasibility set of the parameters to be estimated and avoid unreasonable values.

A less efficient alternative procedure to increase robustness would be to solve the proposed problems with a multi-start optimization technique such that a good local optimum will be obtained. Note that this procedure is more computational demanding and also (without the box constraints) does not guarantee estimated parameters that belong in a meaningful region of values.

9.4.1. Constraining the Summation of PSDs

If the model in (9.7) perfectly describes the acoustic scene, the sum of the PSDs of the point sources, late reverberation, and microphone self-noise at the reference microphone equals $p^{\mathbf{y}}_{\rho\rho}$ (where ρ is the reference microphone index and $p^{\mathbf{y}}_{\rho\rho}$ is the (ρ,ρ) element of $\mathbf{P}_{\mathbf{y}}$). That is,

$$||\text{diag}\,(\mathbf{P})\,||_1 + \gamma\phi_{\rho\rho} + q_\rho = p^{\mathbf{y}}_{\rho\rho}, \tag{9.26}$$

where $\phi_{\rho\rho}$ is the ρ-th diagonal element of $\mathbf{\Phi}$. In practice, the model is not perfect and we do not know $p^{\mathbf{y}}_{\rho\rho}$, but an estimate $\hat{p}^{\mathbf{y}}_{\rho\rho}$. Therefore, a box constraint is used, instead of an equality constraint. That is,

$$0 \leq ||\text{diag}\,(\mathbf{P})\,||_1 + \gamma\hat{\phi}_{\rho\rho} + q_\rho \leq \delta_1\hat{p}^{\mathbf{y}}_{\rho\rho}, \tag{9.27}$$

where δ_1 is a constant which controls the underestimation or overestimation of the PSDs. This box constraint can be used to improve the robustness of the SCFA_{rev} problem, but cannot be used by the $\text{SCFA}_{\text{no-rev}}$ problem, since it does not estimate γ. A less tight box constraint that can be used for both $\text{SCFA}_{\text{no-rev}}$, SCFA_{rev} problems is

$$0 \leq ||\text{diag}\,(\mathbf{P})\,||_1 \leq \delta_2\hat{p}^{\mathbf{y}}_{\rho\rho}. \tag{9.28}$$

One may see the inequality in (9.28) as a sparsity constraint, natural in audio and speech processing as the number of the active sound sources is small (typically much smaller than the maximum number of sources, r, existing in the acoustic scene) for a singe time-frequency tile. In this case, δ_2 controls the sparsity. A low δ_2 implies large sparsity, while a large δ_2 implies low sparsity. The sparsity is over frequency and time.

9.4.2. Box Constraints for the Early RATFs

Extra robustness can be achieved if the elements of the early RATFs are box-constrained as follows:

$$\Re(l_{ij}) \leq \Re(a_{ij}) \leq \Re(u_{ij}),\ \Im(l_{ij}) \leq \Im(a_{ij}) \leq \Im(u_{ij}), \tag{9.29}$$

where u_{ij}, l_{ij} are some complex-valued upper and lower bounds, respectively[2]. We select the values of u_{ij}, l_{ij} based on relative Green functions. Let us denote with $\mathbf{f}_j \in \mathbb{R}^{3\times 1}$ the location of the j-th source, with $\mathbf{m}_i$ the location of the i-th microphone, and with $d_{ij} = ||\mathbf{f}_j - \mathbf{m}_i||_2$ the distance between the j-th source and i-th microphone. The anechoic ATF (direct path only) at the frequency-bin k between the j-th source i-th microphone is given by [43]

$$\tilde{a}_{ij}(k) = \frac{1}{4\pi d_{ij}} \exp\left(\frac{j2\pi f_s k}{K}\frac{d_{ij}}{c}\right), \tag{9.30}$$

where K is the FFT length, c is the speed of sound, and d_{ij}/c is the time of arrival (TOA) of the j-th source to the i-th microphone. The corresponding anechoic relative ATF with respect to the reference microphone ρ is given by

$$a_{ij}(k) = \frac{\tilde{a}_{ij}(k)}{\tilde{a}_{\rho j}(k)} = \frac{d_{\rho j}}{d_{ij}} \exp\left(\frac{j2\pi f_s k}{K}\frac{\left(d_{ij} - d_{\rho j}\right)}{c}\right), \tag{9.31}$$

where $\left(d_{ij} - d_{\rho j}\right)/c$ is the time difference of arrival (TDOA) of the j-th source between microphones i and ρ. What becomes clear from (9.31) is that the anechoic relative ATF depends only on the two unknown parameters $d_{ij}, d_{\rho j}$. The upper and lower bounds of the real part of (9.31) can be written compactly using the following box inequality

$$-\frac{d_{\rho j}}{d_{ij}} \leq \Re\left(a_{ij}(k)\right) \leq \frac{d_{\rho j}}{d_{ij}}, \tag{9.32}$$

and similarly for the imaginary part of $a_{ij}(k)$.

Among all the points on the circle with any constant radius and center the middle point between microphones with indices i and ρ, the inequality in (9.32) becomes maximally relaxed for the maximum possible $d_{\rho j}$ and minimum possible d_{ij}, i.e., when the ratio $d_{\rho j}/d_{ij}$ becomes maximum. This happens when the j-th source is

[2] An alternative method would be to constrain $||a_{ij}||$ with real lower and upper bounds but that would lead to a non-linear inequality constraint and, thus, a more complicated implementation.

in the endfire direction of the two microphones and closest to i-th microphone. In this case we have $d_{\rho j} = d_{\rho i} + d_{ij}$ and, therefore, (9.32) becomes

$$-\frac{d_{\rho i} + d_{ij}}{d_{ij}} \leq \Re\left(a_{ij}(k)\right) \leq \frac{d_{\rho i} + d_{ij}}{d_{ij}}. \tag{9.33}$$

The imaginary part of $a_{ij}(k)$ is constrained similarly to (9.33). In the inequality in (9.33), the parameters $d_{\rho i}, d_{ij}$ are unknown. Now, we try to relax this inequality and find ways that are independent of these unknown parameters.

Note that the quantity $|d_{ij} - d_{\rho j}|/c$ should not be allowed to be greater than the sub-frame length in seconds, i.e., N/f_s, where N is the sub-frame length in samples. If it is greater than N/f_s, the signal model given in (9.7) is invalid, i.e., the CPSDM of the j-th point source cannot be written as a rank-1 matrix, because it will not be fully correlated between microphones i, ρ. Therefore, we have

$$\frac{|d_{ij} - d_{\rho j}|}{c} \leq \frac{N}{f_s} \iff |d_{ij} - d_{\rho j}| \leq \frac{Nc}{f_s}. \tag{9.34}$$

Note that the inequality in (9.34) should also hold in the endfire direction of the two microphones, which means

$$d_{\rho i} \leq \frac{Nc}{f_s}. \tag{9.35}$$

The inequality in (9.33) is maximally relaxed for the maximum possible $d_{\rho i}$ and the minimum possible d_{ij}. The maximum allowable $d_{\rho i}$ is given by (9.35). Moreover, another practical observation is that the sources cannot be in the same location as the microphones. Therefore, we have

$$d_{ij} \geq \lambda, \tag{9.36}$$

where λ is a very small distance (e.g., 0.01 m). Therefore, the maximum range of the real part of the relative anechoic ATF is given by

$$-\frac{\frac{Nc}{f_s} + \lambda}{\lambda} \leq \Re\left(a_{ij}(k)\right) \leq \frac{\frac{Nc}{f_s} + \lambda}{\lambda}. \tag{9.37}$$

The imaginary part of $a_{ij}(k)$ is constrained similar to (9.37). The above inequality is based on anechoic free-field RATFs. In practice, we have early RATFs which include early echoes and/or directivity patterns which means that we might want to make the box constraint in (9.37) less tight.

9.4.3. Tight Box Constraints for the Early RATFs based on $\hat{\mathbf{D}}$

In Sec. 9.4.2 we proposed the box constraints in (9.37) based on practical considerations without knowing the distance between sensors or between sources and sensors. In this section we assume that we have an estimate of the distance matrix (see Sec. 9.1.3), $\hat{\mathbf{D}}$. Consequently we know $\hat{d}_{\rho i}$ and, therefore, we can make the box

constraint in (9.37) even tighter. Specifically, the inequality in (9.33) is maximally relaxed as follows

$$-\frac{\hat{d}_{\rho i}+\lambda}{\lambda} \leq \Re\left(a_{ij}(k)\right) \leq \frac{\hat{d}_{\rho i}+\lambda}{\lambda}. \tag{9.38}$$

The imaginary part of $a_{ij}(k)$ is constrained similar to (9.38).

9.4.4. Box Constraints for the Late Reverberation PSD

In this section, we take into consideration the late reverberation. We can be almost certain that the following box constraint is satisfied:

$$0 \leq \gamma(t,k)\min\left(\mathrm{diag}(\hat{\mathbf{\Phi}})\right) \leq \min\left[\mathrm{diag}\left(\hat{\mathbf{P}}_{\mathbf{y}}(t,k)\right)\right]. \tag{9.39}$$

This box constraint is only applicable in the $\mathrm{SCFA}_{\mathrm{rev}}$ problem. The box-constraint in (9.39) prevents large overestimation errors which may result in speech intelligibility reduction in noise reduction applications [18, 44].

9.4.5. All microphones have the same microphone-self noise PSD

Here we examine the special case where $\mathbf{P}_{\mathbf{v}}(k) = q(k)\mathbf{I}$, i.e., all microphones have the same self-noise PSD. Moreover, since the microphone self-noise is stationary, we can be almost certain that the following box-constraint holds

$$0 \leq q(k) \leq \min_{\forall t \in \mathcal{B}_\beta}\left(\min\left[\mathrm{diag}\left(\hat{\mathbf{P}}_{\mathbf{y}}(t)\right)\right]\right). \tag{9.40}$$

Similar to the constraint in (9.39), the constraint in (9.40) avoids large overestimation errors.

By having a common self-noise PSD for all microphones, the number of parameters are reduced by $M-1$, since we have only one microphone-self noise PSD for all microphones. Hence, the first identifiability condition for the $\mathrm{SCFA}_{\text{no-rev}}$ problem is now given by

$$|\mathcal{B}_\beta|\frac{M(M+1)}{2} \geq Mr + |\mathcal{B}_\beta|r - r + 1. \tag{9.41}$$

Similarly, the first identifiability condition for the $\mathrm{SCFA}_{\mathrm{rev}}$ problem is now given by

$$|\mathcal{B}_\beta|\frac{M(M+1)}{2} \geq Mr + |\mathcal{B}_\beta|r - r + |\mathcal{B}_\beta| + 1. \tag{9.42}$$

9.5. Practical Considerations

In this section, we discuss practical problems regarding the choice of several parameters of the proposed methods and implementation aspects. Although, we have already explained the problem of over-determination in Sec. 9.3.2, in Sec 9.5.1, we discuss additional ways of achieving over-determination. In Sec. 9.5.2, we discuss about some limitations of the proposed methods. Finally, in Secs. 9.5.3 and 9.5.4, we discuss how to implement the proposed methods.

9.5.1. Over-determination Considerations

Increasing the ratio of the number of equations over the number of unknowns obviously fits better the CPSDM model to the measurements under the assumption that the model is accurate enough and the early RATFs do not change within a time-segment. There are two main approaches to increase the ratio of the number of equations over the number of unknowns. The first approach is to reduce the number of the parameters to be estimated while fixing the number of equations as already explained in Sec. 9.3.2. In addition to the explanation provided in 9.3.2, we could also reduce the number of parameters by source counting per time-frequency tile and adapt r. However, this is out of the scope of the present paper and here we assume that we have a constant r in the entire time-frequency horizon which is the maximum possible r. The second approach is to increase the number of time-frames $|\mathcal{B}_\beta|$ in a time-segment and/or the number of microphones M. Increasing $|\mathcal{B}_\beta|$ is not practical, because typically, the acoustic sources are moving. Thus, $|\mathcal{B}_\beta|$ should not be too small but also not too large. Note that $|\mathcal{B}_\beta|$ is also effected by the time-frame length denoted by $\mathcal{T}$. If $\mathcal{T}$ is small we can use a larger $|\mathcal{B}_\beta|$, while if $\mathcal{T}$ is large, we should use a small $|\mathcal{B}_\beta|$ in order to be able to also track moving sources. However, if we select $\mathcal{T}$ to be very small, the number of sub-frames will be smaller and consequently the estimation error in $\hat{\mathbf{P}}_{\mathbf{y}}$ will be large and will cause performance degradation.

9.5.2. Limitations of the Proposed Methods

From the identifiability conditions in (9.23), (9.25), (9.41) and (9.42) for fixed $|\mathcal{B}_\beta|$ and r, we can obtain the minimum number of microphones needed to satisfy these inequalities. Alternatively, for a fixed M and r we can obtain the minimum number of time-frames $|\mathcal{B}_\beta|$ needed to satisfy these inequalities. Finally, for a fixed M and $|\mathcal{B}_\beta|$ we can find the maximum number of sources r for which we can identify their parameters (early RATFs and PSDs). Let $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ the minimum number of microphones satisfying the identifiability conditions in (9.23), (9.25), (9.41) and (9.42), respectively. Moreover, let $\mathcal{J}_1$, $\mathcal{J}_2$, $\mathcal{J}_3$ and $\mathcal{J}_4$ the minimum number of time-frames satisfying the identifiability conditions in (9.23), (9.25), (9.41) and (9.42), respectively. In addition, let $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$ the maximum number of sources satisfying the identifiability conditions in (9.23), (9.25), (9.41) and (9.42), respectively. The following inequalities can be easily proved:

$$
\begin{array}{llll}
\mathcal{M}_3 \leq \mathcal{M}_4, & \mathcal{M}_1 \leq \mathcal{M}_2, & \mathcal{M}_4 \leq \mathcal{M}_2, & \mathcal{M}_3 \leq \mathcal{M}_1, \\
\mathcal{J}_3 \leq \mathcal{J}_4, & \mathcal{J}_1 \leq \mathcal{J}_2, & \mathcal{J}_4 \leq \mathcal{J}_2, & \mathcal{J}_3 \leq \mathcal{J}_1, \\
\mathcal{R}_3 \geq \mathcal{R}_4, & \mathcal{R}_1 \geq \mathcal{R}_2, & \mathcal{R}_4 \geq \mathcal{R}_2, & \mathcal{R}_3 \geq \mathcal{R}_1.
\end{array}
$$

9.5.3. Online Implementation Using Warm-Start

The estimation of the parameters is carried out for all time-frames within one time-segment. Subsequently, in order to have low latency, we shift the time-segment one time-frame. For the $|\mathcal{B}_\beta| - 1$ time-frames in the current time-segment that overlap with the time-frames in the previous time-segment, the parameters are initialized using the estimates from the corresponding $|\mathcal{B}_\beta| - 1$ time-frames in the previous

time-segment. The parameters of the most recent time-frame are initialized by selecting a value that is drawn from a uniform distribution with boundaries corresponding to the lower and upper bound of the corresponding box constraint. Only for the first time-segment, the early RATFs are initialized with the r most dominant relative eigenvectors from the averaged CPSDM over all time-frames of the first time-segment.

9.5.4. Solver

The non-convex optimization problems that we proposed can be solved with various existing solvers within the literature such as [45–48]. In this paper, we used the standard MATLAB optimization toobox to solve the optimization problems which implements a combination of the methods in [46–48]. These methods require first and sometimes second-order derivatives of the objective function. The first-order derivatives of the objective functions in (9.11) with respect to most parameters have been obtained already in [4, 34–36] without taking into account the late reverberation PSD. Thus, here we provide only the first-order derivatives with respect to the late reverberation PSD parameter. We have

$$\text{ML: } \frac{\partial F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P}_{\mathbf{y}})}{\partial \gamma} = \text{tr}\left(\mathbf{P}_{\mathbf{y}}^{-1}\left(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}\right)\mathbf{P}_{\mathbf{y}}^{-1}\hat{\mathbf{\Phi}}\right), \tag{9.43}$$

$$\text{LS: } \frac{\partial F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P}_{\mathbf{y}})}{\partial \gamma} = \text{tr}\left(\left(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}\right)\hat{\mathbf{\Phi}}\right), \tag{9.44}$$

$$\text{GLS: } \frac{\partial F(\hat{\mathbf{P}}_{\mathbf{y}}, \mathbf{P}_{\mathbf{y}})}{\partial \gamma} = \text{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}^{-1}\left(\mathbf{P}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{y}}\right)\hat{\mathbf{P}}_{\mathbf{y}}^{-1}\hat{\mathbf{\Phi}}\right). \tag{9.45}$$

For the second-order derivatives, we used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximated Hessian [36].

9.6. Experiments

In this section, we show the performance of the proposed methods in the context of two multi-microphone applications. The first application is dereverberation of a single point source ($r = 1$). The second application is source separation combined with dereverberation examined in an acoustic scene with $r = 3$ point sources. In this paper, we use the perfect permutation matrix for all compared methods in the source separation experiments. For these experiments we selected the maximum-likelihood objective function in (9.11). The values of the parameters that we selected for both applications are summarized in Table 9.1. All methods based on the diagonal SCFA methodology are implemented using the online implementation in Sec. 9.5.3. The acoustic scene we consider for the source separation example is depicted in Fig. 9.2. The acoustic scene we consider for the dereverberation example is similar with the only difference that the music signal and male talker sources (see Fig. 9.2) are not present. The room dimensions are $7 \times 5 \times 4$ m. The reverberation time for the dereverberation application is selected $T_{60} = 1$ s, while for the source separation, $T_{60} = 0.2$ and 0.6 s. The microphone signals have a duration of 50 s and the

Table 9.1: Summary of parameters used in the experiments.

Parameter	Definition	Value
M	number of microphones	4
K	FFT length	256
$\mathcal{T}$	time-frame length	2000 samples (0.125 s)
N	sub-frame length	200 samples (0.0125 s)
ov_N	overlapping of sub-frames	75%
$\hat{\mathbf{\Phi}}$	spatial coherence matrix	spherical isotropic model
ρ	reference microphone index	1
δ_1	controls overestimation underestimation	1.2
δ_2	controls sparsity	1
c	speed of sound	343m/s
λ	minimum possible source-microphone distance	1 cm
f_s	sampling frequency	16 kHz
q	mic. self noise PSD	$9 * 10^{-6}$

duration of the impulse responses used to construct the microphone signals is 0.5 s. The microphone signals were constructed using the image method [43]. The microphone array is circular with a consecutive microphone distance of 2 cm. The reference microphone is the right-top microphone in Fig. 9.2. Moreover, we assume that the microphone-self noise has the same PSD at all microphones. Finally, it is worth mentioning that the early part of a room impulse response (see Sec. 9.1.2) is of the same length as the sub-frame length.

9.6.1. PERFORMANCE EVALUATION

We will perform two types of performance evaluations in both applications. The first one measures the error of the estimated parameters, while the second one measures the performance by using the estimated parameters in a source estimation algorithm and measure instrumental intelligibility and sound quality of the estimated source waveforms. We measure the average PSD errors of the sources, the average PSD error of the late reverberation, and the average PSD error of the microphone-self noise using the following three measures [49]:

$$E_s = \frac{10}{C(K/2+1)r} \sum_{t=1}^{C} \sum_{k=1}^{K/2+1} \sum_{j=1}^{r} \left| \log \frac{p_j(t,k)}{\hat{p}_j(t,k)} \right| \text{ (dB)}, \tag{9.46}$$

$$E_l = \frac{10}{C(K/2+1)r} \sum_{t=1}^{C} \sum_{k=1}^{K/2+1} \left| \log \frac{\gamma(t,k)}{\hat{\gamma}(t,k)} \right| \text{ (dB)}, \tag{9.47}$$

$$E_v = \frac{10}{C(K/2+1)r} \sum_{t=1}^{C} \sum_{k=1}^{K/2+1} \left| \log \frac{q(t,k)}{\hat{q}(t,k)} \right| \text{ (dB)}. \tag{9.48}$$

We also compute the underestimates (denoted as above with superscript un) and overestimates (denoted as above with superscript ov) of the above averages as in [44] since a large overestimation error in the noise PSDs and a large underestimation error in the target PSD typically results in large target source distortions in the context of a noise reduction framework [44]. On the other hand, a large underestimation error in the noise PSDs may result in musical noise [44]. We also evaluate the average early RATF estimation error using the Hermitian angle measure [50] given by

$$E_A = \frac{1}{rV} \sum_{j=1}^{r} \sum_{\beta=1}^{V} \text{acos}\left(\frac{|\mathbf{a}_j^H(\beta,k)\hat{\mathbf{a}}_j(\beta,k)|}{||\mathbf{a}_j^H(\beta,k)||_2 ||\hat{\mathbf{a}}_j(\beta,k)||_2} \right) \text{ (rad)}. \tag{9.49}$$

If the PSD of a source in a frequency-bin is negligible for all time-frames within a time-segment, the estimated PSD and RATF of this source at that time-frequency tile are skipped from the above averages.

To evaluate the intelligibility and quality of the j-th target source signal, the estimated parameters are used to construct a multi-channel Wiener filter (MWF) as a concatenation of a single-channel Wiener filter (SWF) and a minimum variance distortionless response (MVDR) beamformer [1]. That is,

$$\hat{\mathbf{w}}_j = \frac{\hat{p}_j}{\hat{p}_j + \hat{\mathbf{w}}_{j,\text{MVDR}}^H \hat{\mathbf{P}}_{j,\mathbf{n}} \hat{\mathbf{w}}_{j,\text{MVDR}}} \hat{\mathbf{w}}_{j,\text{MVDR}}, \tag{9.50}$$

and

$$\hat{\mathbf{w}}_{j,\text{MVDR}} = \frac{\hat{\mathbf{P}}_{j,\mathbf{n}}^{-1} \hat{\mathbf{a}}_j}{\hat{\mathbf{a}}_j^H \hat{\mathbf{P}}_{j,\mathbf{n}}^{-1} \hat{\mathbf{a}}_j}, \tag{9.51}$$

where

$$\hat{\mathbf{P}}_{j,\mathbf{n}} = \sum_{\forall i \neq j} \hat{p}_i \hat{\mathbf{a}}_i \hat{\mathbf{a}}_i^H + \hat{\gamma} \mathbf{\Phi} + \hat{q} \mathbf{I}. \tag{9.52}$$

The noise reduction of the j-th source is evaluated using the segmental-signal-to-noise-ratio (SSNR) for the j-th source only in sub-frames where the j-th source is active after which we average the SSNRs over all sources. Moreover, for speech sources, we measure the predicted intelligibility with the SIIB measure [51, 52] and average SIIB over all speech sources.

9.6.2. Reference State-of-the-Art Dereverberation and Parameter Estimation Methods

For the dereverberation we first estimate the PSD of the late reverberation using the method proposed in [19, 26]. Specifically, we first compute the Cholesky decomposition $\hat{\mathbf{\Phi}} = \mathbf{L}_{\mathbf{\Phi}} \mathbf{L}_{\mathbf{\Phi}}^H$ after which we compute the whitened estimated noisy CPSDM

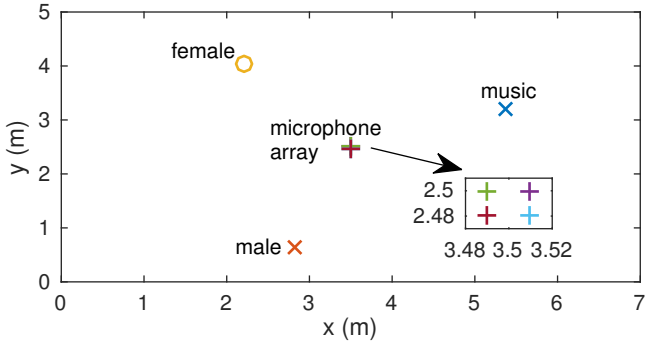

Figure 9.2: Acoustic scene with $r = 3$ sources and $M = 4$ microphones.

as

$$\mathbf{P_{w1}} = \mathbf{L}_{\mathbf{\Phi}}^{-1} \hat{\mathbf{P}}_{\mathbf{y}} (\mathbf{L}_{\mathbf{\Phi}}^{H})^{-1}. \tag{9.53}$$

Next, we compute the eigenvalue decomposition $\mathbf{P_{w1}} = \mathbf{V}\mathbf{R}\mathbf{V}^H$, where the diagonal entries of $\mathbf{R}$ are sorted in descending order. The PSD of the late reverberation is then computed as

$$\hat{\gamma} = \frac{1}{M-1} \sum_{i=2}^{M} \mathbf{R}_{ii}. \tag{9.54}$$

Having an estimate of the late reverberation, we compute the noise CPSDM matrix as $\hat{\mathbf{P}}_{\mathbf{n}} = \hat{\gamma}\hat{\mathbf{\Phi}} + \mathbf{P_v}$ and use it to estimate the early RATF and PSD of the target in the sequel.

We estimate the early RATF of the target using the method proposed in [8, 53]. We first compute the Cholesky decomposition $\hat{\mathbf{P}}_{\mathbf{n}} = \mathbf{L_n}\mathbf{L}_{\mathbf{n}}^H$. We then compute the whitened estimated noisy CPSDM as $\mathbf{P_{w2}} = \mathbf{L}_{\mathbf{n}}^{-1} \hat{\mathbf{P}}_{\mathbf{y}} (\mathbf{L}_{\mathbf{n}}^H)^{-1}$. Next, we compute the eigenvalue decomposition $\mathbf{P_{w2}} = \mathbf{V}\mathbf{R}\mathbf{V}^H$, where the diagonal entries of $\mathbf{R}$ are sorted in descending order. We compute the early RATF as

$$\hat{\mathbf{a}} = \frac{\mathbf{L_n}\mathbf{V}_1}{\mathbf{e}_1^T \mathbf{L_n}\mathbf{V}_1}, \tag{9.55}$$

with $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$. We improve even further the accuracy of the estimated RATF by estimating the RATFs of all time frames within each time-segment and then use the average of these as the RATF estimate. Finally, the target PSD is estimated as proposed in [15, 28], i.e.,

$$\hat{p} = \hat{\mathbf{w}}_{\mathrm{MVDR}}^H \left(\hat{\mathbf{P}}_{\mathbf{y}} - \hat{\mathbf{P}}_{\mathbf{n}} \right) \hat{\mathbf{w}}_{\mathrm{MVDR}}, \tag{9.56}$$

where $\hat{\mathbf{w}}_{\mathrm{MVDR}}$ is given in (9.51).

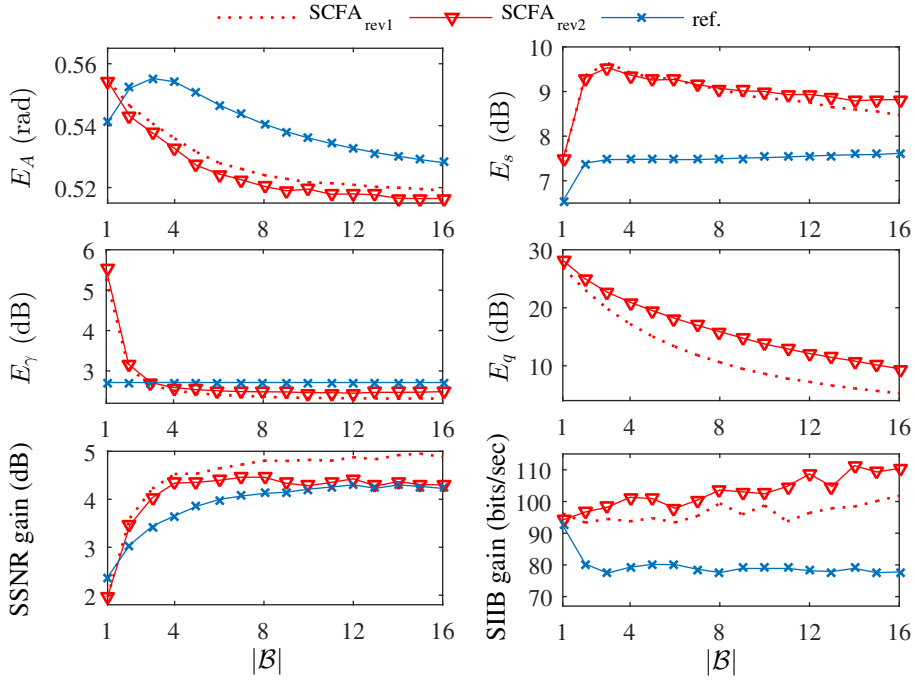

Figure 9.3: Dereverberation results: The proposed methods are denoted by $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$. The ref. is the reference method reviewed in Sec. 9.6.2.

9.6.3. Dereverberation

We compare two different versions of the proposed SCFA_{rev} problem referred to as $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$. Unlike the $\text{SCFA}_{\text{no-rev}}$ problem, the SCFA_{rev} problem also estimates the late reverberation PSD and thus is more appropriate in the context of dereverberation. Both versions use the box constraint for the γ parameter in (9.39) and the box constraint of the early RATF in (9.38). Moreover, since we assume that the microphones-self noise PSDs are all equal, both versions will use the box constraint in (9.40). Both methods use the true distance matrix $\hat{\mathbf{D}} = \mathbf{D}$. The $\text{SCFA}_{\text{rev1}}$ uses the linear inequality in (9.27), while the $\text{SCFA}_{\text{rev2}}$ does not use a constraint for the sum of PSDs. We also include in the comparisons the state-of-the-art approach described in Sec. 9.6.2 (denoted as ref.). The reference method does not estimate the microphone-self noise PSD and we assume for the reference method that we have a perfect estimate, i.e., $\mathbf{P}_{\mathbf{v}} = q\mathbf{I}$. We consider a single target source without interfering signals so that the signal model in (9.7) reduces to

$$\mathbf{P}_{\mathbf{y}} = p_1 \mathbf{a}_1 \mathbf{a}_1^H + \underbrace{\gamma \mathbf{\Phi} + q\mathbf{I}}_{\mathbf{P}_{\mathbf{n}}}. \tag{9.57}$$

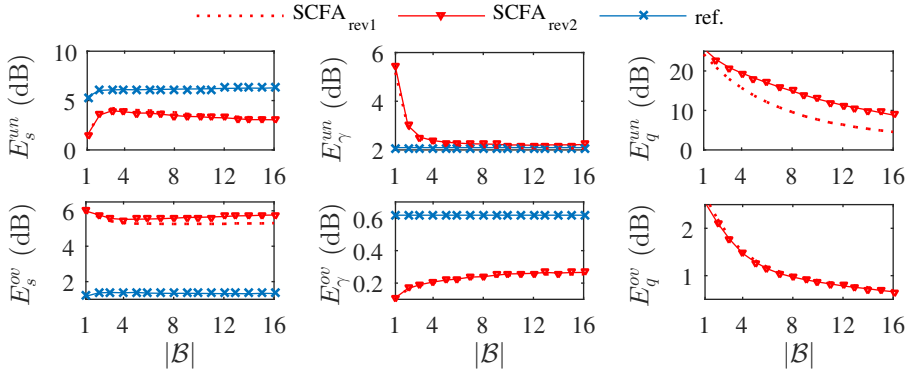

Figure 9.4: Underestimates (with superscript un) and overestimates (with superscript ov): The proposed methods are denoted by $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$. The ref. is the reference method described in Sec. 9.6.2.

After having estimated all the model parameters for the proposed and reference methods, the estimated parameters are used within the MWF given in (9.50), which is applied to the reverberant target source in order to enhance it.

Fig. 9.3 shows the results of the compared methods. It is clear that in almost all evaluation criteria both proposed methods are significantly outperforming the reference method, except for the overall source PSD error E_s. However, the proposed methods have all larger intelligibility gain and better noise reduction performance compared to the reference method for $|\mathcal{B}_\beta| \geq 2$. Fig. 9.4 shows the underestimates and overestimates for the PSDs. It is clear that although the overall PSD error E_s is lower for the reference method, the proposed method has a lower underestimation error for the target, E_s^{un}, and a lower overestimation for the noise, E_γ^{ov}, which means less distortions to the target signal and therefore increased intelligibility.

9.6.4. SOURCE SEPARATION

We consider $r = 3$ source signals. In this acoustic scenario, the signal model is given by

$$\mathbf{P}_\mathbf{y} = \mathbf{P}_\mathbf{e} + \gamma\mathbf{\Phi} + q\mathbf{I}. \tag{9.58}$$

First we estimate the signal model parameters. We examine the performance of the proposed $\text{SCFA}_{\text{no-rev}}$ method and the proposed methods $\text{SCFA}_{\text{no-rev1}}$, $\text{SCFA}_{\text{no-rev2}}$, $\text{SCFA}_{\text{rev1}}$, $\text{SCFA}_{\text{rev2}}$. Unlike the methods $\text{SCFA}_{\text{rev1}}$, $\text{SCFA}_{\text{rev2}}$, the methods $\text{SCFA}_{\text{no-rev1}}$ and $\text{SCFA}_{\text{no-rev2}}$ are based on the $\text{SCFA}_{\text{no-rev}}$ problem. The $\text{SCFA}_{\text{no-rev2}}$ method uses the box constraints in (9.28), (9.38) (which assumes full knowledge of $\hat{\mathbf{D}} = \mathbf{D}$), and (9.40). We also use the method $\text{SCFA}_{\text{no-rev1}}$ where the only difference with $\text{SCFA}_{\text{no-rev2}}$ is that $\text{SCFA}_{\text{no-rev1}}$ uses the RATF box constraint in (9.37) which does not depend on $\hat{\mathbf{D}}$. For the reference method, we use the method proposed in [4] (denoted as m. Parra), modified such that is as much aligned as possible with the

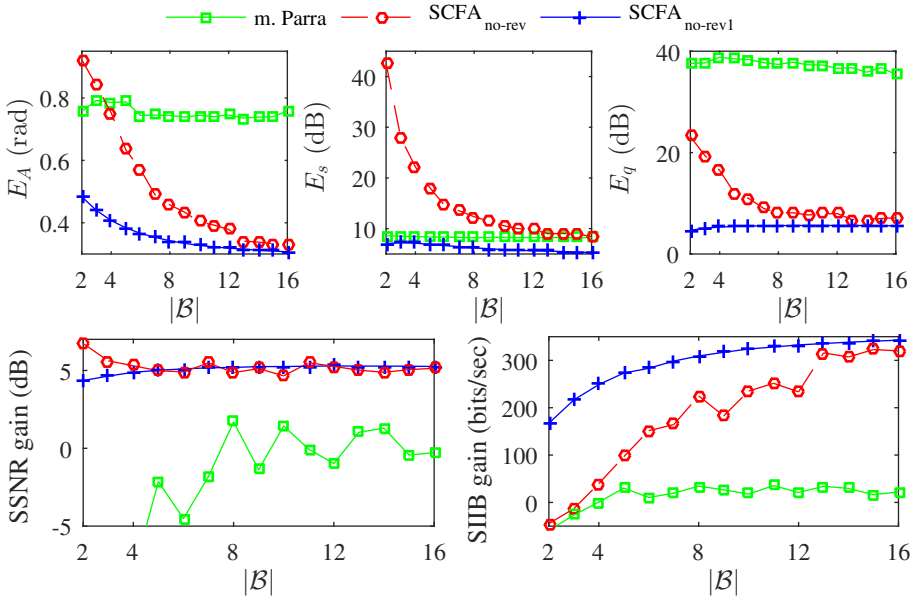

Figure 9.5: Source separation results for $T_{60} = 0.2$ s: Comparison of m. Parra method and the proposed blind methods $\text{SCFA}_{\text{no-rev}}$ and $\text{SCFA}_{\text{no-rev1}}$.

proposed methods. Specifically, we solved the optimization problem of the reference method differently compared to [4]. Unlike [4] which uses the constraints $a_{ii} = 1$, we set the reference microphone row of $\mathbf{A}$ equal to the unity vector, as we did in all proposed methods. In addition, instead of the LS objective function used in [4], we used the ML objective function as with the proposed methods. We also used the same solver (see Sec. 9.5.4) for all compared methods. Note that the authors in [4] have solved the iterative problem using first-order derivatives only, while here we also use an approximation of the Hessian. Finally, the extracted parameters for both the reference and proposed methods are combined with the MWF in (9.50) where for each different source signal we use a different MWF $\hat{\mathbf{w}}_i$.

9

Low reverberation time: $T_{60} = 0.2s$

In order to have a clear visualization of the performance differences, we group the comparisons in two figures. Fig. 9.5 compares all blind methods that do not depend on $\hat{\mathbf{D}}$ or $\hat{\mathbf{\Phi}}$, i.e., $\text{SCFA}_{\text{no-rev}}$, $\text{SCFA}_{\text{no-rev1}}$ and the reference method (referred to as m. Parra). Recall that the only difference between the $\text{SCFA}_{\text{no-rev}}$ method and the m. Parra is the positivity constraints for the PSDs. It is clear that using these positivity constraints improves performance significantly. Note also that the usage of extra inequality constraints from $\text{SCFA}_{\text{no-rev1}}$ is beneficial for improving the performance even more significantly.

In Fig. 9.6, we compare the best-performing $\text{SCFA}_{\text{no-rev1}}$ method of Fig. 9.5

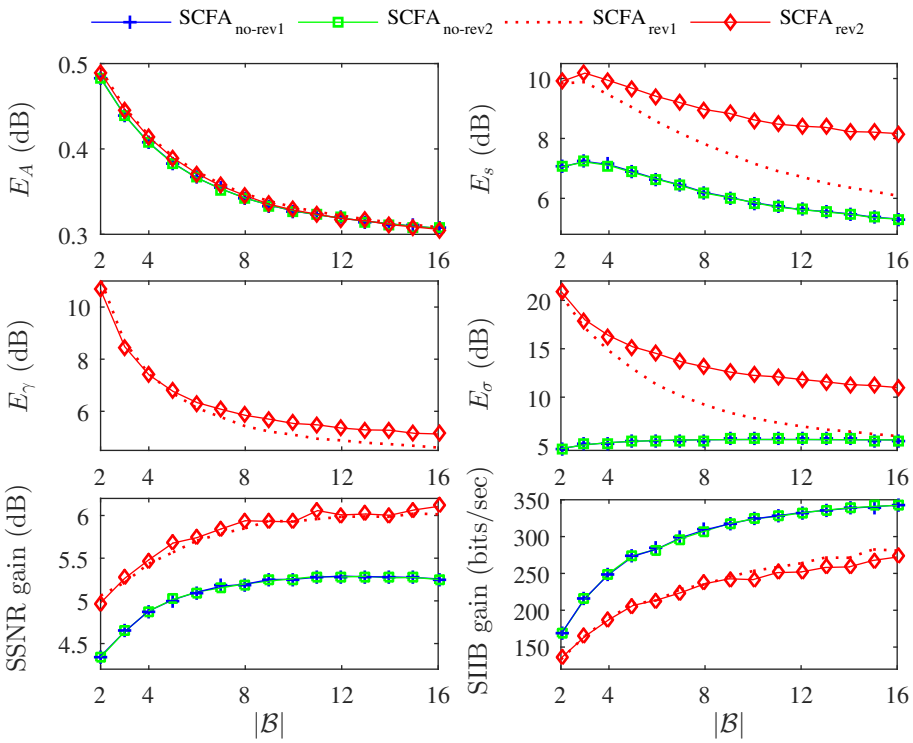

Figure 9.6: Source separation results for $T_{60} = 0.2$ s: Comparison of the proposed $\text{SCFA}_{\text{no-rev2}}$, $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$ methods which assume knowledge of $\mathbf{D}$, and the proposed blind method denoted by $\text{SCFA}_{\text{no-rev1}}$.

with $\text{SCFA}_{\text{no-rev2}}$, $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$. The problems that estimate the late reverberation parameter γ have worse estimation accuracy for the PSD of the sources and microphone-self noise and worse predicted intelligibility improvement compared to the rest of the proposed methods. This is mainly due to the low reverberation time ($T_{60} = 0.2$ s) and the large number of parameters of $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$ as argued in Sec. 9.3.2. However, both $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$ achieve a better noise reduction performance than the other methods. Finally, it is worth noticing that the $\text{SCFA}_{\text{no-rev1}}$ has almost identical performance with the $\text{SCFA}_{\text{rev2}}$ method which used the extra information of $\hat{\mathbf{D}} = \mathbf{D}$.

LARGE REVERBERATION TIME: $T_{60} = 0.6s$

In Figs. 9.7 and 9.8, we compare the same methods as in Fig. 9.5 and 9.6, respectively, but with $T_{60} = 0.6$. Here we observe that the methods which estimate γ become more accurate in RATF estimation, since now the contribution of late reverberation is significant (see the explanation in Sec. 9.3.2). Moreover, when

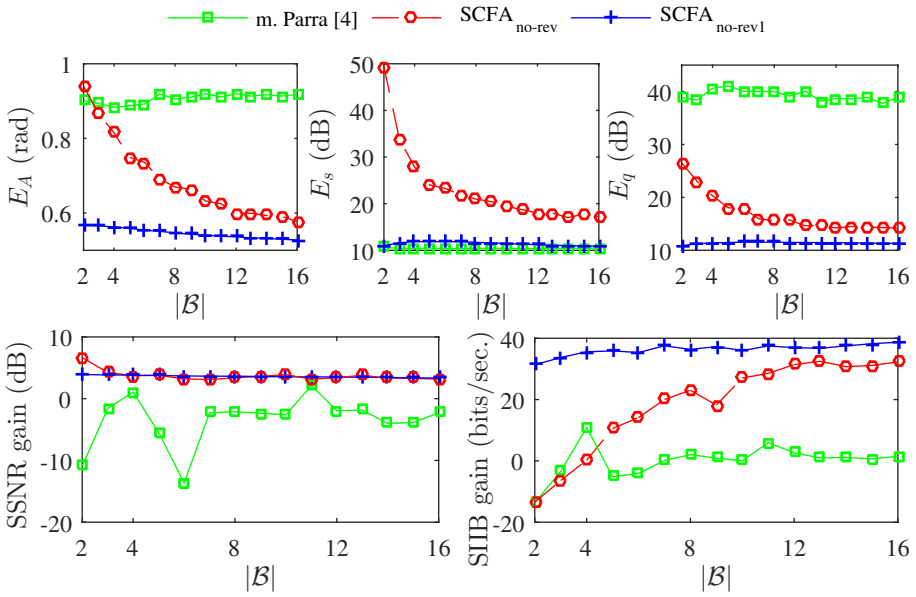

Figure 9.7: Source separation results for $T_{60} = 0.6$ s: Comparison of m. Parra method and the proposed blind methods $\text{SCFA}_{\text{no-rev}}$ and $\text{SCFA}_{\text{no-rev1}}$.

the number of time-frames per time-segment $|\mathcal{B}_\beta|$ increases significantly the methods $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$ have the same predicted intelligibility improvement compared to the other proposed methods but have a much better noise reduction performance.

In conclusion, we observe that in both applications the proposed approaches have shown remarkable robustness in highly reverberant environments. The box constraints that we used indeed provided estimates that are useful in both examined applications. Specifically, the box constraints avoided large overestimation errors in the late reverberation and microphone-self noise PSDs and large underestimation errors for the point sources PSDs. As a result the sources were not distorted significantly and combined with the good noise reduction performance we achieved large predicted intelligibility gains compared to the reference methods.

9.7. Conclusion

In this paper, we proposed several methods based on the combination of confirmatory factor analysis and non-orthogonal joint diagonalization principles for estimating jointly several parameters of the multi-microphone signal model. The proposed methods achieved, in most cases, a better parameter estimation accuracy and a better performance in the context of dereverberation and source separation compared to existing state-of-the-art approaches. The inequality constraints introduced to

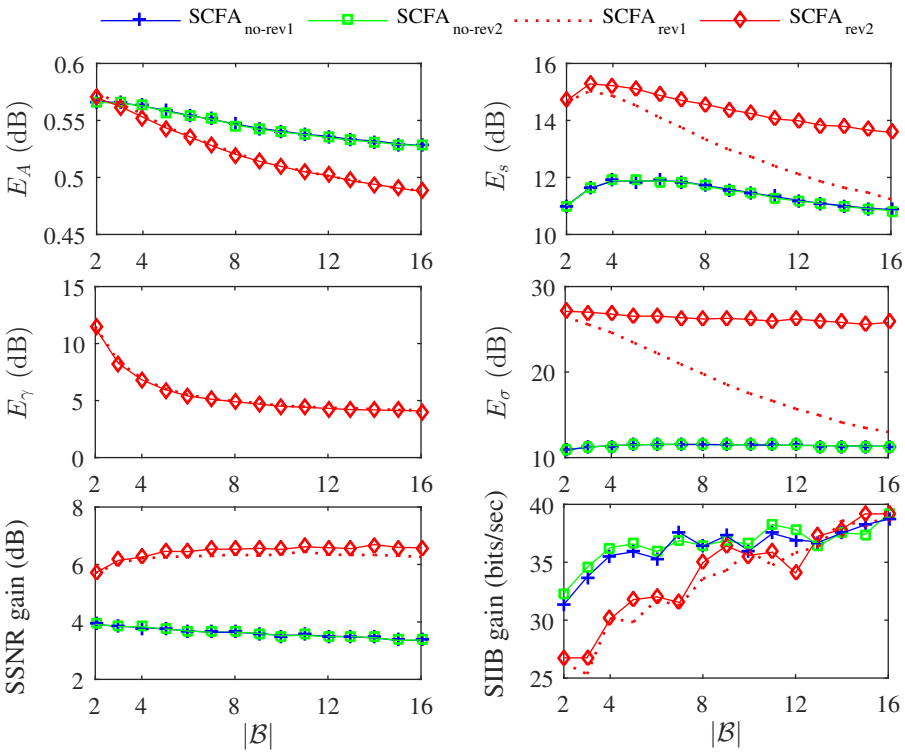

Figure 9.8: Source separation results for $T_{60} = 0.6$ s: Comparison of the proposed $\text{SCFA}_{\text{no-rev2}}$, $\text{SCFA}_{\text{rev1}}$ and $\text{SCFA}_{\text{rev2}}$ methods which assume knowledge of $\mathbf{D}$, and the proposed blind method denoted by $\text{SCFA}_{\text{no-rev1}}$.

limit the feasibility set in the proposed methods resulted in increased robustness in highly reverberant environments in both applications.

9

References

[1] M. Brandstein and D. Ward (Eds.), *Microphone arrays: signal processing techniques and applications* (Springer, 2001).

[2] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and E. Moulines, *A blind source separation technique using second-order statistics,* IEEE Trans. Audio, Speech, Language Process. **45**, 434 (1997).

[3] J. F. Cardoso, *Blind signal separation: statistical principles,* Proc. of the IEEE **86**, 2009 (1998).

[4] L. Parra and C. Spence, *Convolutive blind separation of non-stationary sources,* IEEE Trans. Audio, Speech, Language Process. **8**, 320 (2000).

[5] R. M. H. Sawada, S. Araki, and S. Makino, *Frequency-domain blind source separation of many speech signals using near-field and far-field models,* EURASIP J. Applied Signal Process. **2006**, 1 (2006).

[6] D. Nion, K. Mokios, N. D. Sidiropoulos, and A. Potamianos, *Batch and adaptive parafac-based blind separation of convolutive speech mixtures,* IEEE Trans. Audio, Speech, Language Process. **18**, 1193 (2010).

[7] T. Lotter and P. Vary, *Dual-channel speech enhancement by superdirective beamforming,* EURASIP J. Applied Signal Process. **2006**, 1 (2006).

[8] 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. , 1071 (2009).

[9] 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. **22**, 785 (2014).

[10] S. Gannot, E. Vincet, S. Markovich-Golan, and A. Ozerov, *A consolidated perspective on multi-microphone speech enhancement and source separation,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 692 (2017).

[11] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, *Relaxed binaural LCMV beamforming,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 137 (2017).

[12] 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. **26**, 1434 (2018).

[13] 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**, 550 (2018).

[14] S. Braun and E. A. P. Habets, *Dereverberation in noisy environments using reference signals and a maximum likelihood estimator,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2013).

[15] A. Kuklasinski, S. Doclo, S. H. Jensen, and J. Jensen, *Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2014) pp. 61–65.

[16] S. Braun and E. A. P. Habets, *A multichannel diffuse power estimator for dereverberation in the presence of multiple sources,* EURASIP J. Audio, Speech, and Music Process. **2015**, 34 (2015).

[17] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, *Maximum likelihood psd estimation for speech enhancement in reverberation and noise,* IEEE/ACM Trans. Audio, Speech, Language Process. **24**, 1599 (2016).

[18] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, *Evaluation and comparison of late reverberation power spectral density estimators,* IEEE/ACM Trans. Audio, Speech, Language Process. **26**, 1056 (2018).

[19] I. Kodrasi and S. Doclo, *Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation,* IEEE/ACM Trans. Audio, Speech, Language Process. **26**, 1106 (2018).

[20] 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. **21**, 2193 (2013).

[21] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, *Auto-localization in ad-hoc microphone arrays,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2013) pp. 106–110.

[22] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, *Localizing multiple audio sources in a wireless acoustic sensor network,* ELSEVIER Signal Process. **107**, 54 (2015).

[23] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, *Informed sound source localization using relative transfer functions for hearing aid applications,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 611 (2017).

[24] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, *Inference of room geometry from acoustic impulse responses,* IEEE Trans. Audio, Speech, Language Process. **20**, 2683 (2012).

[25] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, *Acoustic echoes reveal room shape,* Proc. of the National Academy of Sciences **110**, 12186 (2013).

[26] I. Kodrasi and S. Doclo, *Late reverberant power spectral density estimation based on eigenvalue decomposition,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2017) pp. 611–615.

[27] U. Kjems and J. Jensen, *Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement,* in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)* (2012) pp. 295 – 299.

[28] J. Jensen and M. S. Pedersen, *Analysis of beamformer directed single-channel noise reduction system for hearing aid applications,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2015) pp. 5728 – 5732.

[29] R. C. Hendriks and T. Gerkmann, *Noise correlation matrix estimation for multi-microphone speech enhancement,* IEEE Trans. Audio, Speech, Language Process. **20**, 223 (2012).

[30] B. Schwartz, S. Gannot, and E. A. P. Habets, *Two model-based EM algorithms for blind source separation in noisy environments,* IEEE/ACM Trans. Audio, Speech, Language Process. **25**, 2209 (2017).

[31] A. Kuklasinski and J. Jensen, *Multichannel wiener filters in binaural and bilateral hearing aidsspeech intelligibility improvement and robustness to doa errors,* J. of the Audio Engineering Society **65**, 8 (2017).

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm,* J. Royal Statist. Soc. B **39**, 1 (1977).

[33] D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method* (London Butterworths, 1963).

[34] K. G. Jöreskog, *A general approach to confirmatory maximum likelihood factor analysis,* Psychometrika **34**, 183 (1969).

[35] K. G. Jöreskog, *Simultaneous factor analysis in several populations,* Psychometrika **36**, 409 (1971).

[36] S. A. Mulaik, *Foundations of factor analysis* (CRC press, 2009).

[37] H. Kuttruff, *Room acoustics* (CRC Press).

[38] K. G. Jöreskog, *Factoring the multitest-multioccasion correlation matrix, ETS Research Bulletin Series*, (1969).

[39] K. G. Jöreskog, *Factor analysis by generalized least squares,* Psychometrika **37**, 243 (1972).

[40] K. G. Jöreskog and D. N. Lawley, *New methods in maximum likelihood factor analysis,* British J. Math. Statist. Psycol. **21**, 85 (1968).

[41] J. B. Kruskal, *Three-way arrays: Rank and uniqueness of trilinear decompositions with application to arithmetic complexity and statistics,* Linear Alg. Appl. **18**, 95 (1977).

[42] L. D. Lathauwer, *Blind identification of underdetermined mixtures by simultaneous matrix diagonalization,* IEEE Trans. Signal Process. **56**, 1096 (2008).

[43] J. B. Allen and D. A. Berkley, *Image method for efficiently simulating small-room acoustics,* J. Acoust. Soc. Amer. **65**, 943 (1979).

[44] 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. **20**, 1383 (2012).

[45] D. P. Bertsekas, *Projected newton methods for optimization problems with simple constraints,* SIAM J. Control and Optim. **20**, 221 (1982).

[46] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming,* SIAM J. on Optim. **9**, 877 (1999).

[47] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming,* Mathematical Programming **89**, 149 (2000).

[48] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, *An interior algorithm for nonlinear optimization that combines line search and trust region steps,* Mathematical programming **107**, 391 (2006).

[49] R. C. Hendriks, J. Jensen, and R. Heusdens, *Dft domain subspace based noise tracking for speech enhancement,* in *ISCA Interspeech* (2007) pp. 830 – 833.

[50] 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.

[51] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, *An instrumental intelligibility metric based on information theory,* IEEE Signal Process. Lett. **25**, 115 (2018).

[52] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, *An evaluation of intrusive instrumental intelligibility metrics,* IEEE/ACM Trans. Audio, Speech, Language Process. **26**, 2153 (2018).

[53] S. Markovich 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.

10

Conclusions and Future Research

In this chapter, we draw the conclusions of this dissertation and we highlight several open problems, which are worth investigating. In addition, we suggest how to approach these problems in future research.

10.1. Conclusion

The primary goal of hearing assistive devices (HADs) is to improve intelligibility and, thus, to make communication easier. To do so, multi-microphone noise reduction algorithms are typically employed in HADs to preserve a target source from a specific direction while suppressing all other undesired sources. This is achieved by combining multiple signals after having properly changed their phase and magnitude. As binaural localization is also based on the magnitude and phase relationships between the two ears, such modifications may result in binaural-cue distortions. Binaural-cue distortions on the interferers and the target source result in an unnatural impression of the acoustic environment. However, it will also i) reduce intelligibility by negatively effecting the binaural release from masking [1, 2], ii) may put in risk the user's life as localization of point sources is important in many daily life situations, e.g., in traffic situations. Moreover, noise reduction may not only harm the spatial impression of the sources in the acoustic scene, but may also introduce distortions to the target signal which may effect intelligibility.

It becomes evident from the above that there is a need on developing (binaural) multi-microphone noise reduction algorithms which aim at the following goals:

Goal 1 : minimization of noise at the output of the filter.

Goal 2 : minimization of target distortions at the output of the filter.

Goal 3 : minimization of the binaural-cue distortions at the output of the filter.

Since intelligibility improvement depends on how these goals have been achieved simultaneously, (binaural) multi-microphone noise reduction methods implicitly aim at intelligibility improvement as well. The amount of intelligibility improvement depends on how much weight we put on each of the aforementioned goals. The maximization of intelligibility can be seen as a multi-objective optimization problem where the objectives are the three aforementioned goals. The objectives are weighted differently leading to a trade-off between the three objectives. The values of the weights that maximize intelligibility are unknown in general. Moreover, (binaural) multi-microphone noise reduction methods depend on several acoustic scene dependent parameters such as the ATFs of the sources, the powers of the sources, etc. An inaccurate estimation of these parameters implies reduced performance with respect to the three aforementioned goals. Three research questions that were presented in Chapter 1, and which have been addressed in the current dissertation are

Q1: *Can we find binaural multi-microphone noise reduction methods that can (approximately) preserve the binaural cues of all sources in the acoustic scene while at the same time improve intelligibility significantly?*

A1: To address this research question, we have proposed two binaural multi-microphone noise reduction methods (a spatial and a spatio-temporal filtering method). Both methods achieve a significantly larger intelligibility compared to the unprocessed acoustic scene, and they achieve non-significant perceived localization distortions compared to the unprocessed scene. A more in-depth conclusion is drawn in Section 10.1.1.

Q2: *Can we develop multi-microphone noise reduction methods that are robust to estimation errors on the microphone signals parameters?*

A2: To address this research question, we have proposed two beamformers that are robust against ATF estimation errors. Apart from being robust, the proposed beamformers have also an efficient distributed implementation on general cyclic networks. A more detailed conclusion is drawn in Section 10.1.2

Q3: *Can we accurately estimate the parameters from the acoustic scene?*

A3: We have proposed an accurate method that jointly estimates several important parameters of the acoustic scene. More specifically, it estimates the ATFs and power spectral densities (PSDs) of the sources, the PSD of the late reverberation and the PSD of the microphone-self noise. The conclusions are drawn in Section 10.1.3.

10.1.1. Proposed Binaural Multi-Microphone Noise Reduction Methods

In Chapter 3, we proposed an optimization problem that we referred to as the relaxed binaural beamformer (RBB). It minimizes the output noise power while approximately preserving the directional binaural cues of the sources in the acoustic

scene. Unlike existing methods in the literature that use strict equality constraints to preserve the directional binaural-cues [3–7], the RBB problem uses inequality constraints instead. As a result, the RBB problem has a larger feasibility set which implies i) more degrees of freedom for noise reduction, and ii) more freedom to preserve the directional binaural cues of more sources.

One of the limitations of the RBB problem is its non-convex nature due to its non-convex inequality constraints. Initially, we proposed in Chapter 3, a sub-optimal successive convex optimization method to approximately solve the RBB problem. This method solves multiple convex optimization problems per time-frequency bin resulting in large computational costs. To tackle this computational problem we proposed in Chapter 5 an alternative sub-optimal method to solve the RBB problem, which is based on the semi-definite relaxation principle [8]. The proposed method requires to solve only one convex optimization problem per time-frequency bin. This does not only results in a much faster implementation, but also, in most cases, to more optimal performance compared to the initial approach.

The RBB problem requires, similar as existing binaural multi-microphone noise reduction methods in the literature [3–7], estimates of the ATFs of the sources to be preserved. Tracking these ATFs is a very challenging task due to the continuous movements of the head of the user and the sources. To tackle this problem, we proposed in Chapter 4 a method based on the use of many pre-determined ATFs around the head of the user essentially covering a grid of the entire space. These pre-determined ATFs are fixed and, thus, require no tracking. The performance of the spatial filter highly depends on the number of the pre-determined ATFs. That is, the more pre-determined ATFs we use, the less degrees of freedom are left for noise reduction. To compensate this problem, we can increase the amount of relaxation on the inequality constraints, such that we will be able to increase the number of constraints. As a consequence, the grid will have sufficient resolution.

While in Chapters 3, 4 and 6 we examined binaural spatial filtering methods, in Chapter 5, we proposed a binaural spatio-temporal filtering method, referred to as the BMVDR-thresholding method, which approximately preserves both the directional and diffuse binaural cues of the acoustic scene without the need to estimate the ATFs of the interfering sources. BMVDR-thresholding is based on the idea of spectral masking. It tries to achieve the best possible noise reduction performance using the BMVDR method only at those time-frequency tiles where the residual noise after processing is inaudible. If the residual noise is audible at the output of the filter, a scaled-down version of the noisy acoustic scene is used instead, such that the binaural cues of the audible residual noise will be preserved. The selection is based on a user-defined hard threshold applied on the narrowband output SNR of the BMVDR spatial filter. This threshold essentially controls the trade-off between noise reduction and binaural-cue preservation. We have experimentally shown, using instrumental measures, that the BMVDR-thresholding achieves a more optimal trade-off between noise reduction and binaural-cue preservation compared to the BMVDR-η method proposed in [9–11] and reviewed in Chapter 2.

In Chapter 7, we conducted a subjective evaluation of our two proposed methods (the RBB method using pre-determined ATFs from Chapter 4 and the BMVDR-

thresholding from Chapter 6 where the subjects were normal-hearing people. The conclusions of our results can be summarized in the next bullet points:

- If we optimize the trade-off parameters of the RBB and BMVDR-thresholding methods for a specific acoustic scene, the latter method provides the best trade-off between noise reduction and binaural-cue preservation.
- For fixed trade-off parameters which have been optimized in a different acoustic scene than the acoustic scene under investigation, both methods provide a very similar performance. This is because the BMVDR-thresholding depends on the output SNR value which is acoustic-scene dependent. Thus, an optimal value of its trade-off parameter in one acoustic scene is not necessarily optimal in another scene. On the other hand, the RBB method does not depend on the output SNR, but on fixed locations around the head of the user and as such has a more predictable performance compared to the BMVDR-thresholding method.
- Both proposed methods achieve a significantly better intelligibility compared to the unprocessed scene while at the same time manage not to introduce significant localization distortions after processing.

10.1.2. Proposed Robust Multi-Microphone Noise Reduction Methods

In Chapter 8, we proposed two beamformers, namely block-diagonal LCMV (BDLCMV) beamforming, and block-diagonal LCMP (BDLCMV) beamforming, which are robust against ATF estimation errors. At the same time they are low-complexity distributed implementable when multiple devices collaborate in a wireless acoustic sensor network. The distributed calculations also play an important role in robustness, since they do not depend on a single device which plays the role of the fusion center.

The main idea of BDLCMV and BDLCMP is that by nulling the interferers in the acoustic scene, the remaining noise is the diffuse late reverberation. This is approximately uncorrelated between microphones of different devices and partially correlated between microphones of the same device. A reasonable approximation of the noisy and noise CPSDMs are their block diagonal versions, where each block corresponds to the microphones of a single device. The proposed block-diagonal structure of the noise and noisy CPSDM results in increased robustness against ATF estimation errors. At the same time, the objective function of the optimization problem is now fully separable and efficient distributed implementations can be used.

We have shown that the minimization of the objective function of the BDLCMV and BDLCMP methods cannot harm the target signal as severely as the objective functions of the LCMV and LCMP methods, respectively, in case of ATF and CPSDM estimation errors. The main reason is the block-diagonal structure of the CPSDM in the objective function. Unlike diagonal loading techniques [12, 13] which need to adapt their diagonal loading parameter per time-frequency bin, the proposed method has a fixed strategy which provides a much simpler implementation.

The proposed approaches can be solved with several distributed implementations exploiting its fully separable optimization problem. We have shown that by using the primal direction method of multipliers (PDMM) method [14] we can have a fully distributable implementation on arbitrarily structured networks including cyclic networks while at the same time being frame-optimal. That is, unlike several other existing methods [15], which need many frames to reach an equivalent performance to the corresponding centralized implementation, we achieve an equivalent performance to the centralized counterpart at each frame while having a small amount of communication costs. The per-frame convergence of the proposed method to the performance of the centralized counterpart requires only a very small number of iterations which leads to very small amount of communication costs between the devices.

10.1.3. Proposed Signal-Model Parameter Estimation Methods

In many applications such as source separation, binaural noise reduction and dereverberation, we need to estimate some or all of the parameters of the signal model in (2.8). The parameters included in this model are i) the ATFs of the sources, ii) the PSDs of the sources, iii) the PSD of the late reverberation, and iv) the PSDs of the microphones-self noise. Having estimates of these parameters allow us to construct parametric versions of the noise and target CPSDMs without the need of a target activity detector to detect target presence or absence.

In Chapter 9, we proposed a method that jointly estimates all the aforementioned parameters of the multi-microphone signal model. The proposed method is based on the combination of two theories; the confirmatory factor analysis theory [16–18] and the non-orthogonal joint diagonalization theory [19]. Unlike the standard formulations of the confirmatory factor analysis and the non-orthogonal joint diagonalization, we also introduced several linear inequality constraints to the parameters to be estimated. This increases the robustness of our method against modeling and estimation errors.

We experimentally showed that in most cases we achieve a large improvement in estimation accuracy of the parameters compared to other existing state-of-the-art approaches [19–21]. Moreover, by using the estimated parameters from our approach we obtained significant gains in predicted intelligibility improvement and noise reduction in the context of source separation and dereverberation compared to other existing state-of-the-art approaches [19, 21, 22].

10.2. Open Problems and Suggestions for Future Research

In this section, we list some important open problems which are worth further investigation. We also provide suggestions on how to approach these problems.

Perceptually-based trade-offs in binaural multi-microphone noise reduction

The BMVDR-thresholding method, proposed in Chapter 6, classifies the time-frequency tiles based on the audibility of the residual noise at the output of the BMVDR filter. After the classification stage, it decides either to leave the BMVDR processed signals as they are, or to replace them with a scaled version of the noisy acoustic scene. This method tries to maintain only the binaural-cues of the audible time-frequency tiles of the residual noise using a hard-threshold based on the the output narrow-band SNR of the BMVDR. This threshold is acoustic scene dependent and, thus, a fixed threshold will not always perform as expected. Possible future extensions could include a more sophisticated perception based classification mechanism of the time-frequency tiles which will be acoustic scene independent.

Perceptual considerations can be also applied in the RBB method presented in Chapters 3, 4 and 5. The parameters that control the amount of relaxation of the inequality constraints can vary over frequency, time and spatial direction. As we discussed in Chapter 1, the frequencies in the range 1.5 to 3 kHz are less important for localization. At these frequencies a larger amount of relaxation can be introduced. Also, in the directions where we can use visual cues, we may allow more binaural-cue distortions, as it can be hypothesized that the visual cues might partly compensate for a small error in spatial sound location. Finally, the degrees of freedom can be exploited more efficiently if the RBB method uses only constraints on the audible interfering sources per time-frequency tile.

In conclusion, perceptual trade-offs aim at exploiting the available degrees of freedom more efficiently and are more in line with the auditory system. Potentially, this will lead to a larger intelligibility improvement and smaller *perceived* binaural-cue distortions compared to the trade-offs proposed so far.

From the results of the proposed methods we can draw the conclusion that the BMVDR is always slightly better in noise reduction and in intelligibility improvement compared to the RBB and BMVDR-thresholding methods. Perceptual trade-offs may be the solution to this problem. By exploting the binaural realiase of masking [1, 2] we can potentially achieve a better intelligibility improvement compared to the BMVDR. Another way to achieve a larger intelligibility improvement compared to the BMVDR is to apply post-filters at the output of the left and right spatial filters which will further reduce the noise. This is something that has already been investigated in the context of other spatial filters in the literature (see e.g., [11]). The results are indeed satisfactory.

10

Fast Algorithms for Relaxed Binaural Beamforming Problem

The RBB problem is a computationally demanding non-convex problem. We have proposed a sub-optimal method by solving one convex optimization problem per-time frequency tile. Nevertheless, solving a convex optimization problem requires several iterations. In hearing aids the required complexity will be still prohibitive. As a result we need adaptive implementations that spread iterations across time-frames instead of the same time-frame. This will have some performance degradation with respect to noise reduction, but the question is how much?

Distributable ATF Estimation of Multiple Sources

The work in Chapter 8 is easily distributable, and, robust against ATF estimation errors. However, the method assumes that the ATFs of the sources are known. Ideally, the ATFs should also be estimated in distributed fashion. Possible future extensions include a distributed implementation of the ATF estimation method presented in Chapter 9.

Room Geometry Estimation and Tracking of Multiple Sources

ATFs include information of both the location of the sources and the locations of the walls due to the early reflections which are captured in the ATFs [23]. It would be interesting to investigate how accurately we can solve the problems of room geometry estimation and tracking of multiple sources using the method proposed in Chapter 9.

References

[1] H. Levitt and L. R. Rabiner, *Binaural release from masking for speech and gain in intelligibility,* J. Acoust. Soc. Amer. **42**, 601 (1967).

[2] A. W. Bronkhorst, *The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,* Acta Acoustica **86**, 117 (2000).

[3] 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 Trans. Audio, Speech, Language Process. **23** (2015).

[4] E. Hadad, S. Doclo, and S. Gannot, *The binaural LCMV beamformer and its performance analysis,* IEEE Trans. Audio, Speech, Language Process. **24**, 543 (2016).

[5] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, *Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints,* IEEE Trans. Audio, Speech, Language Process. **23**, 2449 (2015).

[6] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, *Improved multi-microphone noise reduction preserving binaural cues,* in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)* (2016).

[7] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, *Optimal binaural lcmv beamformers for combined noise reduction and binaural cue preservation,* in *Int. Workshop Acoustic Signal Enhancement (IWAENC)* (2014) pp. 288–292.

[8] L. Vandenberghe and S. Boyd, *Semidefinite programming,* SIAM review **38**, 49 (1996).

[9] T. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, *Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues,* IEEE Trans. Signal Process. **55**, 1579 (2007).

[10] D. Marquardt, *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*, Ph.D. thesis, Carl von Ossietzky Universität Oldenburg (2015).

[11] D. Marquardt and S. Doclo, *Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering,* IEEE Trans. Audio, Speech, Language Process. **26**, 1261 (2018).

[12] B. D. Carlson, *Covariance matrix estimation errors and diagonal loading in adaptive arrays,* **24**, 397 (1988).

[13] J. Li, P. Stoica, and Z. Wang, *On robust Capon beamforming and diagonal loading,* IEEE Trans. Signal Process. **51**, 1702 (2003).

[14] G. Zhang and R. Heusdens, *Distributed optimization using the primal-dual method of multipliers,* **4**, 173 (2018).

[15] S. Markovich, A. Bertrand, M. Moonen, and S. Gannot, *Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks,* ELSEVIER Signal Process. **107**, 4 (2015).

[16] K. G. Jöreskog and D. N. Lawley, *New methods in maximum likelihood factor analysis,* British J. Math. Statist. Psycol. **21**, 85 (1968).

[17] K. G. Jöreskog, *A general approach to confirmatory maximum likelihood factor analysis,* **34**, 183 (1969).

[18] K. G. Jöreskog, *Simultaneous factor analysis in several populations,* **36**, 409 (1971).

[19] L. Parra and C. Spence, *Convolutive blind separation of non-stationary sources,* IEEE Trans. Audio, Speech, Language Process. **8**, 320 (2000).

[20] 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. **17**, 1071 (2009).

[21] I. Kodrasi and S. Doclo, *Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation,* IEEE Trans. Audio, Speech, Language Process. **26**, 1106 (2018).

[22] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, *Maximum likelihood psd estimation for speech enhancement in reverberation and noise,* IEEE Trans. Audio, Speech, Language Process. **24**, 1599 (2016).

[23] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, *Inference of room geometry from acoustic impulse responses,* IEEE Trans. Audio, Speech, Language Process. **20**, 2683 (2012).

A

Appendix

In this section, we show how the optimization problem in Eq. (3.36) can be equivalently written as a second order cone programming (SOCP) problem. For convenience, we reformulate the optimization problem in Eq. (3.36) using RATFs instead of ATFs. The left and right RATFs of the i-th interferer are $\bar{\mathbf{b}}_{i,L} = (1/b_{iL})\mathbf{b}_i$ and $\bar{\mathbf{b}}_{i,R} = (1/b_{iR})\mathbf{b}_i$, respectively, while the left and right RATFs of the target are $\bar{\mathbf{a}}_L = (1/a_L)\mathbf{a}$ and $\bar{\mathbf{a}}_R = (1/a_R)\mathbf{a}$, respectively. It is easy to show that the constraints of the optimization problem in Eq. (3.36) can be equivalently written as

$$\underbrace{\begin{bmatrix} \bar{\mathbf{a}}_L^H & \mathbf{0}^H \\ \mathbf{0}^H & \bar{\mathbf{a}}_R^H \end{bmatrix}}_{\mathbf{\Phi}_1^H} \mathbf{w} = \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{\mathbf{q}_1}, \tag{A.1}$$

$$\left|\mathbf{\Phi}_{2,i}^H \mathbf{w}\right| \leq \underbrace{\left|\tau_{(k)}\zeta \bar{\mathbf{b}}_{i,R}^H \hat{\mathbf{w}}_{R,(k-1)}\right|}_{\mathbf{q}_{2,i}}, i = 1, \cdots, m, \tag{A.2}$$

where $\zeta = |\bar{a}_{R,1}^* \bar{b}_{i,L,M}^* - 1|$ (with $\bar{a}_{R,1}^*$ the first element of $\bar{\mathbf{a}}_R^H$ and $\bar{b}_{i,L,M}^*$ is the last element of $\bar{\mathbf{b}}_{i,L}$) and $\mathbf{\Phi}_{2,i}$ is the i-th column of the matrix $\mathbf{\Phi}_2$ given by

$$\mathbf{\Phi}_2 = \begin{bmatrix} \bar{\mathbf{b}}_{1L}, \cdots, \bar{\mathbf{b}}_{mL} \\ -\bar{\mathbf{b}}_{1R}, \cdots, -\bar{\mathbf{b}}_{mR} \end{bmatrix}. \tag{A.3}$$

Similar to [1, 2], we convert the complex vectors and matrices to real-valued ones, i.e.,

$$\breve{\mathbf{w}} = \begin{bmatrix} \breve{\mathbf{w}}_L \\ \breve{\mathbf{w}}_R \end{bmatrix}, \breve{\mathbf{w}}_L = \begin{bmatrix} \mathrm{Re}\{\mathbf{w}_L\} \\ \mathrm{Im}\{\mathbf{w}_L\} \end{bmatrix}, \breve{\mathbf{w}}_R = \begin{bmatrix} \mathrm{Re}\{\mathbf{w}_R\} \\ \mathrm{Im}\{\mathbf{w}_R\} \end{bmatrix}, \tag{A.4}$$

$$\breve{\mathbf{a}}_L = \begin{bmatrix} \mathrm{Re}\{\bar{\mathbf{a}}_L\} \\ \mathrm{Im}\{\bar{\mathbf{a}}_L\} \end{bmatrix}, \breve{\mathbf{a}}_R = \begin{bmatrix} \mathrm{Re}\{\bar{\mathbf{a}}_R\} \\ \mathrm{Im}\{\bar{\mathbf{a}}_R\} \end{bmatrix} \tag{A.5}$$

A

$$\breve{\mathbf{a}}_L = \begin{bmatrix} -\mathrm{Im}\{\bar{\mathbf{a}}_L\} \\ \mathrm{Re}\{\bar{\mathbf{a}}_L\} \end{bmatrix}, \breve{\mathbf{a}}_R = \begin{bmatrix} -\mathrm{Im}\{\bar{\mathbf{a}}_R\} \\ \mathrm{Re}\{\bar{\mathbf{a}}_R\} \end{bmatrix} \tag{A.6}$$

$$\breve{\mathbf{b}}_{iL} = \begin{bmatrix} \mathrm{Re}\{\bar{\mathbf{b}}_{iL}\} \\ \mathrm{Im}\{\bar{\mathbf{b}}_{iL}\} \end{bmatrix}, \breve{\mathbf{b}}_{iR} = \begin{bmatrix} \mathrm{Re}\{\bar{\mathbf{b}}_{iR}\} \\ \mathrm{Im}\{\bar{\mathbf{b}}_{iR}\} \end{bmatrix}, \tag{A.7}$$

$$\breve{\mathbf{b}}_{iL} = \begin{bmatrix} -\mathrm{Im}\{\bar{\mathbf{b}}_{iL}\} \\ \mathrm{Re}\{\bar{\mathbf{b}}_{iL}\} \end{bmatrix}, \breve{\mathbf{b}}_{iR} = \begin{bmatrix} -\mathrm{Im}\{\bar{\mathbf{b}}_{iR}\} \\ \mathrm{Re}\{\bar{\mathbf{b}}_{iR}\} \end{bmatrix}, \tag{A.8}$$

$$\breve{\mathbf{P}} = \begin{bmatrix} \mathrm{Re}\{\mathbf{P}\} & -\mathrm{Im}\{\mathbf{P}\} \\ \mathrm{Im}\{\mathbf{P}\} & \mathrm{Re}\{\mathbf{P}\} \end{bmatrix}, \tilde{\breve{\mathbf{P}}} = \begin{bmatrix} \breve{\mathbf{P}} & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{P}} \end{bmatrix}, \tag{A.9}$$

$$\breve{\mathbf{\Phi}}_1 = \begin{bmatrix} \breve{\mathbf{a}}_L & \mathbf{0} & \breve{\mathbf{a}}_L & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{a}}_R & \mathbf{0} & \breve{\mathbf{a}}_R \end{bmatrix}, \tag{A.10}$$

$$\breve{\mathbf{\Phi}}_2 = \begin{bmatrix} \breve{\mathbf{b}}_{1L}, \cdots, \breve{\mathbf{b}}_{mL} \\ -\breve{\mathbf{b}}_{1R}, \cdots, -\breve{\mathbf{b}}_{mR} \end{bmatrix}, \breve{\mathbf{\Phi}}_2 = \begin{bmatrix} \breve{\mathbf{b}}_{1L}, \cdots, \breve{\mathbf{b}}_{mL} \\ -\breve{\mathbf{b}}_{1R}, \cdots, -\breve{\mathbf{b}}_{mR} \end{bmatrix}. \tag{A.11}$$

Note that $\mathbf{w}^T \tilde{\breve{\mathbf{P}}} \mathbf{w} = ||\tilde{\breve{\mathbf{P}}}^{1/2} \mathbf{w}||_2^2$, where $\tilde{\breve{\mathbf{P}}}^{1/2}$ is the principal square root of $\tilde{\breve{\mathbf{P}}}$. The convex optimization problem in Eq. (3.36) can be equivalently written as

$$\begin{aligned} \hat{\breve{\mathbf{w}}}_{(k)} = & \arg \min_{t, \breve{\mathbf{w}}} t \text{ s.t. } \breve{\mathbf{w}}^T \breve{\mathbf{\Phi}}_1 = \breve{\mathbf{q}}_1^T, \\ & ||\tilde{\breve{\mathbf{P}}}^{1/2} \breve{\mathbf{w}}||_2 \leq t, \\ & \left\| \begin{bmatrix} \breve{\mathbf{\Phi}}_{2,i}^T \\ \breve{\mathbf{\Phi}}_{2,i}^T \end{bmatrix} \breve{\mathbf{w}} \right\|_2 \leq q_{2,i,(k)}, \text{ for } i = 1, \cdots, m, \end{aligned} \tag{A.12}$$

where $\breve{\mathbf{q}}_1^T = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}$, $\breve{\mathbf{\Phi}}_{2,i}$ is the i-th column of $\breve{\mathbf{\Phi}}_2$, and $\breve{\mathbf{\Phi}}_{2,i}$ is the i-th column of $\breve{\mathbf{\Phi}}_2$. Note that the problem in Eq. (A.12) is a standard-form SOCP problem [3].

REFERENCES

[1] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo, *Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem,* IEEE Trans. Signal Process. **51**, 313 (2003).

[2] R. G. Lorenz and S. P. Boyd, *Robust minimum variance beamforming,* IEEE Trans. Signal Process. **53**, 1684 (2005).

[3] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2004).

Acknowledgements

I would like to express my appreciation and gratitude to all those people that have helped and supported me during my PhD studies. First of all, I would like to express my heartfelt appreciation and gratitude to my supervisors, Prof. Richard Hendriks and Prof. Richard Heusdens, who gave me the opportunity to work in a very interesting topic with an excellent group from which I learned a lot during the past 4 years. Your suggestions, criticism, personal guidance and trust were invaluable in the culmination of this work. I would also like to thank you for your great help in completing this thesis.

Many thanks to Prof. Jesper Jensen and Dr. Meng Guo from Oticon A/S who gave me valuable feedback over the past four years and hosted me in Oticon A/S for three months which was a very valuable experience. Thanks a lot for your hospitality, I had a very nice time in Denmark. Many thanks also to Prof. Steven van de Par for the collaboration we had and the valuable knowledge I obtained from him.

I would also like to thank the remaining members of the circuits and systems (CAS) group at Delft University of Technology for the nice time we had all these years. It was a very friendly atmosphere with very interesting and fruitful discussions from which I learned a lot. Special thanks to the head of the CAS group, Prof. Alle-Jan van der Veen who was always helpful and a great advisor to me. I would also like to thank Prof. Geert Leus for giving me valuable feedback for my thesis.

Finally, I wish to thank my parents Ioannis and Eleni, my sister Fani, my girlfriend Theodora and my good friends for their continuous, vital support and encouragement all these years that I was studying at Delft University of Technology. Your love and support helped me a lot to achieve this difficult task. I will be always grateful for your help!

Curriculum Vitæ

Andreas I. Koutrouvelis was born in 1988 in Patras, Greece. He received the B.Sc. degree in computer science from the University of Crete, Greece, in 2011 and the M.Sc. degree in Electrical Engineering from Delft University of Technology (TU-Delft), the Netherlands, in 2014. From February 2012 to July 2012, he was a research intern at Philips Research, Eindhoven, the Netherlands and from October 2014 to December 2014 he was researcher in the Circuits and Systems Group (CAS) in TU-Delft. Since, January 2015 he is pursuing the Ph.D. degree in TU-Delft (CAS). In 2017, he visited Oticon A/S in Copenhagen, Denmark. His research interests include acoustic signal processing, noise reduction and dereverberation, optimization theory, factor analysis, and speech analysis.