THESIS: LOCATION OPTIMIZATION OF VIRTUAL SOURCES FOR BINAURAL BEAMFORMING

Author:
Remi Storme

4682793

Supervisor:
dr.ir. R.C. Hendriks

Signal Processing Systems

Co-Supervisor:
Jordi de Vries
j Signal Processing

23rd of February 2026

To my parents, Immele & Kenneth,
my siblings, Fidel & Lolo,
my dear family & friends

CONTENTS

SUMMARY

Binaural beamforming techniques play an important role in modern hearing assistive devices and spatial audio systems, where the goal is to enhance a target sound source while preserving the natural spatial impression of the acoustic scene. Achieving effective noise reduction without distorting binaural cues such as interaural level and phase differences remains a fundamental challenge, particularly in realistic scenarios where the acoustic transfer functions of interfering sources are unknown or difficult to estimate reliably. A proposed solution [1] is the use of *virtual sources*, i.e., a predefined set of spatial directions for which acoustic transfer functions are assumed to be known and used as constraints in binaural beamforming algorithms. However, the selection of these virtual source locations is often heuristic, and their influence on binaural cue preservation and approximation accuracy is not well understood.

To address these limitations, this thesis makes two main contributions. First, a novel multi-microphone HRTF dataset is created using a dummy head equipped with six microphones per ear, resulting in a total of twelve microphones. The dataset is measured for a discrete set of loudspeaker positions and is provided in the form of raw audio recordings, impulse responses, and frequency-domain HRTFs. By substantially increasing the number of microphones per ear compared to conventional datasets, this database enables the study of binaural beamforming with increased spatial degrees of freedom. Second, using this dataset, the thesis formulates the selection of virtual source locations as a discrete combinatorial optimization problem over a finite set of candidate directions. An interaural transfer function (ITF) reconstruction error metric is employed to quantify the accuracy with which arbitrary interfering source locations can be approximated using a limited number of virtual sources.

Experimental results show that many different virtual source configurations yield similar reconstruction errors, indicating that the optimization problem is highly sensitive and characterized by a large set of near-optimal solutions. In addition, the results reveal fundamental limitations of virtual-source-based approaches when interfering sources are located far from the selected virtual source positions. Achieving reliable approximation across all directions would require a very dense set of virtual sources, and consequently a comparable number of microphones, which is impractical for realistic systems. Overall, the findings of this thesis provide new insights into the interplay between microphone array design, virtual source placement, and binaural cue preservation, and offer guidance for the design and evaluation of future binaural beamforming systems.

1

INTRODUCTION

Hearing Assistive Devices (HADs) [1–3] play an increasingly critical role in contemporary society. Hearing is fundamental for communication, social connection, education, employment, and safety. It also contributes to cognitive and emotional well-being by providing sensory stimulation. When hearing degrades, these functions are impaired, often resulting in social isolation, reduced quality of life, and cognitive decline [4, 5]. Modern HADs, of which hearing aids and cochlear implants, seen in Fig. 1.1, are primary examples, therefore aim not only to restore audibility but also to improve speech intelligibility in complex acoustic environments.

These HADs acquire acoustic signals from the environment, process them, and play back the processed signals to the user through a small loudspeaker in the ear. The processing aims to improve the hearing capabilities of the user in complex acoustic environments by enhancing the speech intelligibility of the target speaker and reducing unwanted noise, thereby reducing listener fatigue. In the case of hearing impairment, user-specific hearing loss can also be compensated for by the HAD.

Traditional hearing aids typically include only one or two microphones per ear and process each ear's signal independently (monaural systems) [6]. As a result, these systems are unable to exploit binaural spatial cues such as interaural time differences (ITDs) and interaural level differences (ILDs), which play a crucial role in spatial hearing and sound source segregation [7–9]. In normal-hearing listeners, access to these binaural cues leads to Spatial Release from Masking (SRM), i.e., an improvement in speech intelligibility when target and interfering sources are spatially separated [1]. In monaural hearing aids, however, signals are enhanced primarily based on their spectral and temporal characteristics rather than their spatial origin. This prevents the device from exploiting spatial separation between the target speaker and interfering noise sources, thereby largely eliminating SRM. Consequently, the achievable noise reduction is limited, and performance degrades severely in multi-talker scenarios where spatial separation would otherwise provide a perceptual benefit [10].

Binaural systems [1, 2, 11], on the other hand, emerged by allowing the devices at the left and right ears to communicate with each other by exchanging microphone signals in real time. By jointly processing signals from both ears, binaural systems can preserve or reconstruct binaural cues and make spatially informed decisions about noise reduction and speech enhancement, enabling better localization, externalization, and a more natural listening experience, particularly in environments with multiple sound sources.

(a) Hearing aid

(b) Cohlear implant

Figure 1.1: Hearing aid devices (HADs)

In addition to binaural processing, modern hearing aids increasingly rely on multiple microphones rather than a single or pair of microphones per device. A multi-microphone configuration provides additional spatial resolution, as each microphone captures a differently filtered version of the same acoustic scene due to its spatial position relative to the sound sources. This enables advanced spatial filtering techniques, known as beamforming, to enhance signals arriving from a desired direction while attenuating others. More microphones generally provide more degrees of freedom [1], allowing for stronger noise suppression, more stable performance across different environments, and better preservation of binaural cues. However, they also introduce challenges, such as the need for accurate acoustic modeling and robust strategies to maintain spatial naturalness.

A classical scenario in HADs is the cocktail party problem [12], where the task of understanding a target speaker in a lively, multi-talker environment becomes challenging. Human listeners rely on natural binaural hearing to focus attention on a desired sound source while suppressing irrelevant ones. This ability depends heavily on the binaural cues such as ITDs and ILDs between the ears [7]. For hearing-impaired listeners, these cues are often reduced or distorted, making it significantly harder to segregate sources and follow a conversation in a noisy environment.

The previously described cocktail party problem can be addressed using binaural beamformers, which spatially filter the incoming sound in a multi-microphone system. Beamforming allows the hearing aid to emphasize signals arriving from the direction of the target speaker while attenuating signals from other directions. Classical approaches such as the Minimum Variance Distortionless Response (MVDR) [13] and the Linearly Constraint Minimum Variance (LCMV) [14] beamformers aim to keep the target source undistorted while minimizing the overall noise power. Their binaural extensions, including the Binaural MVDR (BMVDR) and the Binaural LCMV (BLCMV) beamformers [3], apply separate constraints for the left and right ears to preserve the target's binaural cues and, in some cases, the cues of selected interferers.

However, despite their effectiveness, these methods have inherent limitations. Preserving binaural cues for multiple sources reduces the available degrees of freedom for noise reduction, resulting in a trade-off between cue preservation and interference suppression. Moreover, these beamformers rely heavily on accurate knowledge of the Acoustic Transfer Functions (ATFs) of all involved sources. In realistic listening scenarios, where interfering speakers move, appear unpredictably, or are not explicitly tracked, ATFs are difficult to estimate reliably, especially under severe noise. As a result, binaural beamformers may distort spatial cues or underperform when ATF estimates are inaccurate. This motivates the development of alternative methods that can preserve spatial perception. One promising approach is the introduction of virtual sources [1], which impose constraints at a set of artificial directions with predetermined transfer functions.

The remainder of this chapter introduces the foundational concepts required to understand the subsequent analysis. The basic principles of beamforming and spatial filtering are presented in Section 1.1. In Section 1.2, the relevant binaural cues used by the auditory system are reviewed. The framework of binaural multi-microphone noise reduction is introduced in Section 1.3, followed by an explanation of virtual sources in Section 1.4. The research questions guiding this thesis are formulated in Section 1.5. Finally, an outline of this thesis work will be provided in Section 1.6.

1.1. BEAMFORMING

Beamforming is a spatial filtering technique that exploits the spatial diversity provided by multiple microphones to selectively enhance sounds arriving from a desired direction while suppressing sounds from other directions. The *Minimum Variance Distortionless Response* (MVDR) beamformer [13] is one of the most fundamental beamforming approaches. It minimizes the output noise power subject to a distortionless response constraint for the target signal. Under ideal conditions, such as perfect knowledge of the target steering vector and a noise-only interference field, the MVDR beamformer is optimal in the sense that it achieves maximum noise reduction among all linear spatial filters satisfying the distortionless constraint. Moreover, its closed-form solution enables computationally efficient implementations.

A well-known limitation of the MVDR beamformer is that it provides no explicit control over the spatial characteristics of interfering sources, which may therefore be distorted by the filtering process. The *Linearly Constrained Minimum Variance* (LCMV) beamformer [14] generalizes the MVDR formulation by introducing multiple linear equality constraints. These constraints allow for more flexible and application-driven control of the beamformer's spatial response, for example by preserving or attenuating signals arriving from specific directions, or by imposing nulls to suppress known interferers.

Some important things to note for these two beamformers are that they rely heavily on estimates of both the cross-power spectral density matrix (CPSDM) and the acoustic transfer functions (ATFs) of the target and interfering sources [15]. When a perfect estimate of the ATFs of the target source is used, these beamformers will not distort the

target signal. Unfortunately, the estimation of the ATFs and CPSDM is very prone to errors. Another issue that arises is the real-time estimation of these ATFs and CPSDMs. It requires a lot of processing power, and existing methods are often not fast enough. Therefore, virtual sources with predetermined ATFs were introduced by [1] which is further explained in Section 1.4.

1.2. BINAURAL CUES

The auditory system of the brain utilizes binaural cues to localize a sound source. The binaural cues that are most important for the localization of a sound source consist of the *Interaural Time Differences* (ITDs) and *Interaural Level Differences* (ILDs) [16]. A third cue, *Interaural Coherence* (IC) [17], describes the similarity between the signals at the two ears and contributes primarily to the perception of source diffuseness and spatial width. Because IC is mostly relevant for distance perception and reverberant environments, it is not central to the methods studied in this thesis and will not be considered further.

The ITDs and ILDs are typically analyzed in the frequency domain, where they correspond to differences in phase and magnitude between the left- and right-ear signals, respectively. Together, these cues are encoded in the Head-Related Transfer Functions (HRTFs), which describe how sound from each direction is acoustically transformed by the head, torso, and pinnae [18].

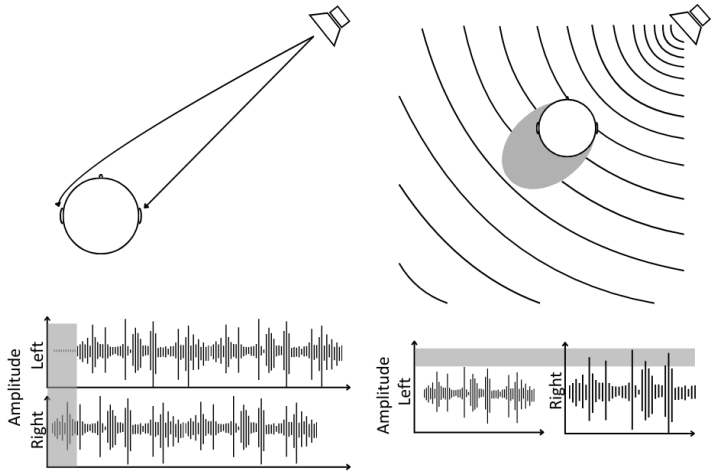


Figure 1.2: Illustration of the binaural cues: Interaural Time Differences (ITDs) on the left and Interaural Level Differences (ILDs) on the right

The ITDs and ILDs are visualized in Fig. 1.2. On the left, the time difference is shown, which mainly occurs due to the difference in arrival time between the two ears of an incoming signal from a sound source. They are most prominent for frequencies under 1.5

kHz. At higher frequencies, the short wavelength causes phase ambiguity, making ITDs less reliable for localization. On the right, the level differences are illustrated, which occur due to two main reasons: the distance between the two ears and the acoustic sound source, and the head shadowing effect, which becomes mostly apparent for high frequencies above 3 kHz. It becomes clear that neither of the binaural cues are well exploited between 1.5 kHz and 3 kHz, and therefore makes localization for this frequency range challenging [16].

1.3. MULTI-MICROPHONE BINAURAL BEAMFORMING

Binaural multi-microphone beamforming extends classical spatial filtering discussed in Section 1.1, by producing two outputs, one for the left ear and one for the right ear [1–3]. Instead of designing a single beamformer, as in the monaural case, a binaural system must design two coordinated filters that not only enhance the target but also maintain a spatially consistent relationship between the two outputs. This makes binaural extensions fundamentally more challenging than monaural beamforming.

A straightforward extension of the MVDR beamformer is the Binaural MVDR (BMVDR) [13, 19]. In this approach, two MVDR beamformers are designed separately, one for each ear, resulting from two independent optimization problems. Importantly, both beamformers operate on the full set of microphone signals from both ears. The resulting filters ensure that the target source is passed undistorted to each ear while noise and interference are suppressed, thereby preserving the binaural cues of the target signal. However, the BMVDR provides no explicit control over the binaural cues of interfering sources. Consequently, interferers may be altered differently at the two ears, leading to distorted or unstable spatial impressions.

To address these limitations, the Binaural LCMV (BLCMV) beamformer [14, 20] introduces additional linear constraints at both ears. These constraints can be used not only to preserve the target's ITDs and ILDs but also to maintain the binaural cues of a selected set of interfering speakers. This makes the BLCMV more perceptually robust: by preserving the spatial location of relevant interferers, the listener is better able to segregate sources and maintain a stable auditory scene. However, the BLCMV suffers from a fundamental limitation: each imposed linear constraint reduces the available degrees of freedom of the beamformer. Preserving the binaural cues of an interfering source typically requires multiple constraints, one per ear, which, given the limited number of microphones available in hearing aids, rapidly reduces the beamformer's remaining capability to suppress noise.

To overcome this trade-off between cue preservation and noise reduction, new binaural beamforming strategies were introduced by [1, 2]. The first of these is the Joint BLCMV (JBLCMV) beamformer, which jointly optimizes the left and right filters rather than designing them independently. This joint formulation uses significantly fewer constraints and, therefore, better balances noise reduction and cue preservation. Building on this, the relaxed JBLCMV (RJBLCMV) beamformer was also proposed by [1], where binaural

cue preservation is enforced approximately rather than exactly, freeing up even more degrees of freedom.

Despite these improvements, a major limitation remained: all previously mentioned methods rely on knowledge of the ATFs of target and interfering sources and the CPSDM. As they can not be estimated reliably in real-time scenarios, [1] introduced the virtual sources with predetermined ATFs. More on this will be explained in Section 1.4.

Existing HRTF databases typically contain only one to three microphones per ear, reflecting practical hearing-aid hardware constraints rather than the needs of research [6]. Such datasets, therefore, can only minimally be used to study how increasing the number of microphones influences spatial filtering, noise reduction, or binaural cue preservation. Because the degrees of freedom of a binaural beamformer grow with the number and placement of microphones [1, 2], evaluating the effect of larger microphone arrays is an interesting field of research.

1.4. VIRTUAL SOURCES

A major limitation of classical binaural beamforming methods is their dependence on accurate Acoustic Transfer Function (ATF) estimates for all relevant sound sources. Errors in ATF estimation can lead to target and interferer distortion, loss of binaural cue preservation, and degradation in noise-reduction performance.

To address this problem, the concept of virtual sources is introduced [1]. Instead of imposing constraints on the actual, time-varying directions of real acoustic sources, binaural constraints are defined for a predefined set of spatial directions around the listener. These directions do not correspond to physically present sound sources, but serve as fixed reference locations in space for which acoustic transfer functions are known. In practice, each virtual source is associated with a predetermined HRTF, measured in advance under controlled conditions. The virtual source directions may be chosen uniformly or strategically distributed around the listener.

A simplified illustration of the cocktail party problem is provided in Fig. 1.3 with a target source in front of the user, some interfering sources positioned at random locations around the user, and virtual sources uniformly distributed around the user. Note that all sources are assumed to be positioned in the far-field (> 1 m) with respect to the receiving microphones [21]. For the virtual sources this means their ATFs are therefore approximately distant invariant.

If the ATF of an interfering source coincides with one of the predetermined ATFs associated with the virtual sources, the binaural cues of this interferer are preserved by construction. In practice, however, interferers in the acoustic scene will generally not align exactly with any single virtual source direction. In that case, a Steering Vector Mismatch (SVM) occurs between the true ATF of the interferer and the set of available virtual-source ATFs [1].

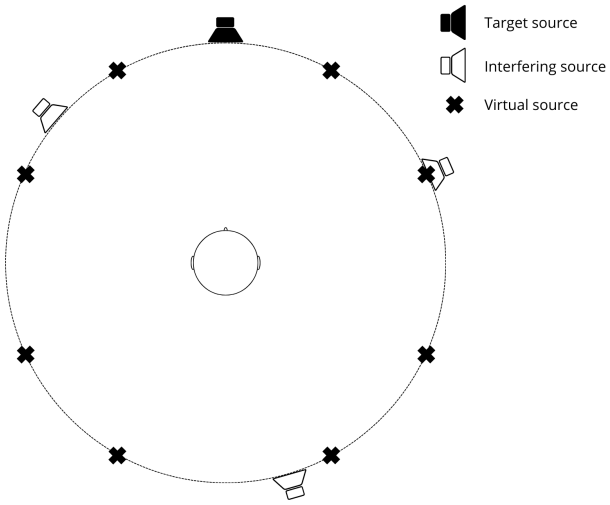


Figure 1.3: Virtual sources with predetermined ATFs uniformly distributed around the user

As argued in [1], the effect of such an interferer can then be approximated by a linear combination of multiple nearby virtual-source constraints. For example, an interfering source located between two virtual source directions in Fig. 1.3 may be partially represented by the constraints associated with both neighboring virtual sources. Increasing the number of virtual sources therefore reduces the worst-case SVM, as the angular spacing between adjacent virtual sources decreases. This observation is adopted as a hypothesis in this thesis and forms the basis for the virtual source location optimization investigated in Chapter 4.

1.5. RESEARCH QUESTIONS

The work presented in this thesis builds upon the theoretical framework of binaural multi-microphone noise reduction developed by Koutrouvelis, in particular the joint binaural LCMV beamforming approaches and the use of pre-determined acoustic transfer functions to preserve binaural cues without explicit interferer localization [1]. While this prior work establishes powerful optimization-based beamforming formulations, it largely relies on limited HRTF configurations and does not explicitly address how the choice, number, and spatial distribution of virtual source locations interact with HRTF data.

A key limitation in existing studies is not the realism of current HRTF datasets with respect to practical hearing-aid designs, but rather their suitability for systematic analysis and methodological exploration. Most publicly available HRTF datasets contain only one to three microphones per ear, which reflects the constraints of real hearing aids but restricts the available spatial degrees of freedom for studying binaural beamforming algorithms. As a result, it becomes difficult to isolate and analyze how increasing

the number and spatial distribution of microphones affects beamformer behavior, cue preservation, and robustness to steering vector mismatch.

To address this methodological gap, this thesis introduces a multi-microphone HRTF dataset measured using a dense array of microphones distributed around each ear. By providing access to a larger set of measured acoustic transfer functions, the dataset enables controlled investigations into the influence of microphone configuration and virtual source placement on binaural beamforming performance, grounded in measured data rather than acoustic models.

Within this context, the central objective of this thesis is to understand how a limited set of virtual source constraints can best approximate arbitrary interferer locations when real, multi-microphone HRTFs are available, and how this approximation affects binaural cue preservation. Rather than focusing directly on noise reduction performance, the emphasis is placed on the spatial modeling capability of the virtual-source-constrained beamformer, which forms a crucial prerequisite for effective binaural noise reduction in later stages.

MAIN RESEARCH QUESTION

This thesis is guided by the following main research question:

> ***Q1:*** *How can a limited, optimally selected set of virtual sources be used to accurately approximate binaural acoustic responses for arbitrary source directions using multi-microphone HRTF measurements?*

SUB RESEARCH QUESTIONS

To address this main question, the following sub-research questions are considered:

> ***Q1.1:*** *To what extent can binaural responses for arbitrary source directions be approximated using a limited subset of virtual sources, and how should the locations of these virtual sources be selected to minimize reconstruction error across the spatial domain?*

> ***Q1.2:*** *How can a dedicated multi-microphone HRTF measurement setup be designed and utilized to enable a systematic, measurement-driven evaluation of virtual source selection and binaural beamforming performance?*

These questions are addressed through (i) a novel multi-microphone HRTF dataset, and (ii) a systematic optimization framework for selecting virtual source locations in binaural beamforming.

Together, these research questions position the thesis as a measurement-driven extension of existing binaural beamforming theory, bridging the gap between abstract optimization formulations and the practical constraints imposed by real, multi-microphone HRTF data. Sub-research question Q1.1 specifically addresses the *virtual source location optimization problem*, investigating how accurately binaural responses for arbitrary

source directions can be approximated using a limited subset of measured HRTFs, and how the spatial locations of these virtual sources should be selected to minimize reconstruction error across the azimuth domain.

Sub-research question Q1.2 focuses on the *design and utilization of multi-microphone HRTF measurements* required to support this optimization problem. It addresses how a dedicated measurement setup can be constructed and employed to enable a systematic, measurement-driven evaluation of virtual source selection strategies and binaural beamforming performance under realistic acoustic conditions.

Addressing the location optimization problem formulated in Q1.1 inherently requires access to accurate and sufficiently dense HRTF data. Evaluating how well arbitrary source directions can be approximated using a limited set of virtual sources necessitates knowledge of the true acoustic transfer functions over a wide range of source directions. Without measured HRTFs, it is not possible to objectively quantify reconstruction errors in binaural cues or to assess the impact of virtual source placement on binaural beamforming performance.

While several public HRTF datasets are available, these datasets are primarily designed for spatial audio rendering or perceptual experiments and typically contain only one to three microphones per ear. Although this reflects realistic hearing-aid hardware constraints, it limits their suitability for systematic analysis of multi-microphone binaural beamforming methods. In particular, existing datasets do not allow for controlled studies on how the number and spatial distribution of microphones influence virtual source approximation accuracy, steering vector mismatch, or the trade-off between noise reduction and binaural cue preservation.

For this reason, this thesis introduces a dedicated multi-microphone HRTF measurement dataset with a dense microphone configuration around each ear. This dataset directly addresses Q1.2 and is specifically designed to support the investigation posed in Q1.1 by enabling controlled, reproducible, and measurement-driven analyses of virtual source location optimization and binaural beamforming behavior that cannot be conducted using existing HRTF databases.

1.6. THESIS OUTLINE

Following the introduction, the thesis first establishes the theoretical and methodological foundations required for the study of binaural multi-microphone beamforming and spatial cue preservation. In Chapter 2, multi-microphone signal models are introduced in the time-frequency domain, together with a review of classical binaural beamforming techniques. Particular emphasis is placed on linearly constrained beamformers and on binaural cue descriptors such as interaural transfer functions, which later form the basis of the error metrics used in the optimization framework.

Building on this theoretical background, Chapter 3 presents a novel multi-microphone HRTF measurement dataset specifically designed to support the analysis of binaural

beamforming with increased spatial degrees of freedom. Unlike conventional HRTF databases that typically employ two microphones per ear, the dataset introduced in this thesis uses six microphones per ear, resulting in a total of twelve microphones mounted on a dummy head. Here, the measurement room analysis, measurement setup, data acquisition procedure, and post-processing pipeline are described in detail. The final outcome of this measurement campaign is a reliable dataset consisting of raw audio recordings, impulse responses, and frequency-domain HRTFs, which serves as the experimental foundation for the subsequent optimization and evaluation chapters.

Using this dataset, Chapter 4 introduces a framework for optimizing the spatial configuration of virtual source locations in binaural beamforming. The problem is formulated as a discrete combinatorial optimization task over a finite set of candidate source directions derived from the measured HRTFs (see Section 4.1). An interaural transfer function (ITF) reconstruction error metric is employed to quantify how accurately arbitrary interfering source locations can be approximated using a limited number of virtual sources, as detailed in Section 4.2. This formulation provides a systematic way to study the relationship between the number, placement, and spatial distribution of virtual sources and the resulting binaural cue preservation performance.

The proposed optimization framework is evaluated experimentally in Section 4.4. Using the multi-microphone HRTF dataset, this section analyzes the influence of different virtual source configurations on binaural reconstruction accuracy across frequencies and interferer locations. The results demonstrate that many distinct configurations yield similar error values, revealing a high sensitivity of the optimization problem to parameter choices. Furthermore, the experiments highlight fundamental limitations of virtual-source-based approaches when attempting to approximate interferers located far from the selected virtual source positions, particularly when the number of microphones and virtual sources is constrained.

Finally, Chapter 5 summarizes the main contributions of the thesis and reflects on their implications for binaural beamforming and spatial audio research. The limitations of both the proposed dataset and the virtual source optimization approach are discussed, and directions for future work are outlined.

2
BACKGROUND

This chapter provides a more in-depth overview of the methods and concepts used in this thesis. First, Section 2.1 provides a review and analysis of existing HRTF datasets. In Section 2.2, a review is done of multi-microphone binaural beamforming and noise reduction methods that were proposed by [1].

2.1. HRTF MEASUREMENTS

The mathematical model describing the acoustic path between a sound source and a listener's ear is represented by the head-related transfer function (HRTF) in the frequency domain and by the head-related impulse response (HRIR) in the time domain [6, 18]. They capture the combined acoustic effects of the head, pinnae, torso, and shoulders, and are therefore highly direction-dependent and user-specific [21]. As a result, HRTFs form the physical foundation of binaural hearing and spatial sound perception, encoding all cues required for sound source localization and externalization, including interaural time differences (ITDs), interaural level differences (ILDs), and monaural spectral properties [7].

Over the past decades, numerous HRTF datasets have been published, employing a wide range of measurement methodologies, spatial resolutions, source distances, and microphone configurations. Several studies provide comprehensive reviews and comparisons of these approaches [6, 18], highlighting their respective advantages and limitations depending on the intended application. Most existing datasets are primarily designed for perceptual spatial audio reproduction and therefore typically rely on a limited number of microphones per ear, often placed in or near the ear canal.

This section provides an overview of existing HRTF datasets and measurement methodologies, with particular attention to design choices that are relevant for binaural beamforming applications. Based on this analysis, the limitations of current datasets in the context of multi-microphone binaural beamforming are identified, motivating the development of a new HRTF dataset with an increased number of microphones per ear.

2.1.1. HEAD RELATED TRANSFER FUNCTIONS

An HRTF describes the acoustic path between a sound source at a given position in space and the signal measured at a listener's ear [18]. As illustrated in Fig. 2.1, an incoming sound wave emitted by a source is modified by the listener's anatomy before reaching the left and right ears. These modifications arise from diffraction and shadowing by

the head, reflections and resonances introduced by the pinnae, and scattering effects caused by the torso and shoulders [16, 21]. The combined effect of these interactions is commonly modeled as a linear, time-invariant, direction-dependent filter, implicitly assuming linear acoustic propagation, negligible non-linear effects at typical listening levels, and a static listener and environment during the measurement.

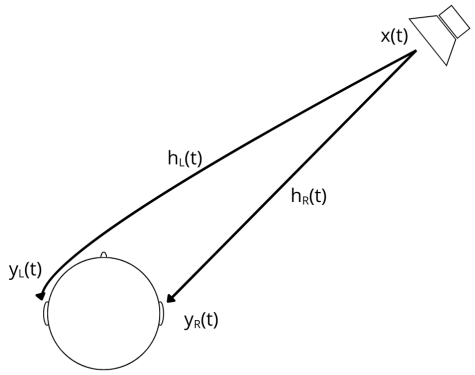


Figure 2.1: HRTF diagram

For a fixed source position, the HRTF represents the acoustic path in the frequency domain. Its time-domain representation, obtained via an inverse Fourier transform, is referred to as the HRIR [6]. The HRTF therefore characterizes how the spectral magnitude and phase of a sound are modified as a function of source direction. Since the geometry of the head and ears differs between individuals, HRTFs are inherently listener-specific. In practice, however, both individual and generic HRTFs are commonly used depending on the application and measurement constraints.

HRTFs encode the direction-dependent acoustic cues that are essential for human sound localization [16]. Differences in arrival time between the left and right ears give rise to interaural time differences (ITDs), which dominate localization at low frequencies. At higher frequencies, diffraction and head shadowing effects lead to interaural level differences (ILDs) [16]. In addition, monaural spectral cues introduced by the pinnae are encoded in the HRTFs and are crucial for resolving elevation and front-back ambiguities. Together, these cues allow the auditory system to infer the position of a sound source in three-dimensional space.

2.1.2. EXISTING HRTF DATA SETS

A wide range of HRTF datasets has been published over the past decades, differing in measurement environment, spatial resolution, source distance, receiver type, microphone placement, and post-processing methodology. Comprehensive overviews and comparisons of these datasets can be found in several review papers [6, 18], which highlight how design choices are strongly influenced by the intended application. An overview

of commonly used public datasets and their measurement configurations is provided in Table 2.1.

In general, existing datasets have been recorded with different goals in mind. For example, datasets targeting binaural rendering often prioritize dense angular sampling with one receiver microphone per ear, while datasets designed for hearing-aid and array processing research may include multiple microphones placed around the ear or head. As a consequence, datasets differ substantially in loudspeaker layout, source distance, and post-processing strategy. Most measurements are performed in anechoic or hemi-anechoic conditions in order to reduce the influence of environmental reflections and ensure reproducibility, as also reflected by the datasets listed in Table 2.1.

Another important distinction is whether HRTFs are measured in the near field ($< 1\,\mathrm{m}$) or far field ($> 1\,\mathrm{m}$). In the near field, HRTFs become strongly distance dependent due to level changes and spherical wavefront effects, whereas far-field measurements are often preferred because they allow a simplified acoustic model with approximately planar wavefronts and reduced distance sensitivity [21]. Since these design parameters directly impact which binaural cues are captured and how suitable the dataset is for multi-microphone processing, they are reviewed in more detail in the following subsections.

Table 2.1: Overview of commonly used public HRTF datasets and their measurement configurations.

Dataset	Receiver	Loudspeaker setup	# Mics/ear	Az. res.	El. res.	Distance [m]
SONICOM [22]	Human	Multiple in hoop	1	5°	$-45°$ ◊ 225° at ±12.5°	1.5
CIPIC [23]	Human & KEMAR	Multiple in hoop	1	$-80°$ ◊ 80° at ±15°	$-45°$ ◊ 230° at ±5.625	1
HAL [24]	Five different HATS models	Multiple in $\frac{1}{4}$ arc	1	5°	0° ◊ 90° at ±15°	2, 0.4
Comparative analysis [25]	Brüel & Kjær	Single	1	At 0°, 45°	At 0°, 15°	1.5
Brungart dataset [21]	KEMAR	Multiple	1	1°	/	0.125, 0.25, 0.5, 1
Denk dataset [26]	Human & HATS models	Multiple	2	7.5°	$-30°$ ◊ 90° at ±30°	2.5, 3
Bronkhorst dataset [27]	Human	Multiple in hoop	3	5°	$-45°$ ◊ 90° at ±15°	1.5
Braren dataset [28]	KEMAR	Single rotating	1	1°	1°	2
Wierstorf dataset [29]	KEMAR	Single	1	1°	/	0.5, 1, 2, 3
Oldenburg dataset [30]	Four different HATS models	Multiple in TASP	1-3	2°	2°	2

The most important properties that differentiate public datasets (see Table 2.1) are:

- **Loudspeaker configuration**
- **Sound source distance**
- **Microphone configuration**

In the next subsections, these design dimensions are compared in order to determine an appropriate measurement configuration for the aims of this thesis.

SOUND SOURCE CONFIGURATIONS

The loudspeaker configuration determines the spatial sampling of the sound field around the listener and has a direct impact on the spatial resolution of an HRTF dataset. Different datasets employ different strategies to position sound sources relative to the listener, most commonly using either single-loudspeaker or multi-loudspeaker measurement setups [6, 18].

In single-loudspeaker configurations [29], a single sound source is placed at either at a fixed position while the listener or dummy head is rotated or at a varying position while the dummy is in a fixed position to sample different source directions. This approach offers high flexibility and allows for dense angular sampling without the need for a large number of loudspeakers, as the rotation can be performed in small angular steps. However, measurement time is typically long, and the approach requires precise mechanical positioning to ensure reproducibility [6].

Multi-loudspeaker configurations, on the other hand, employ multiple sound sources distributed around the listener, often arranged on a circular arc, semi-circular arc, or spherical structure. This allows multiple source directions to be measured without rotating the listener, significantly reducing measurement time [6]. The spatial resolution in such setups is determined by the number and placement of loudspeakers and is therefore fixed by the physical construction of the array. While multi-loudspeaker systems are efficient and stable, increasing spatial resolution requires a corresponding increase in the number of loudspeakers, which adds complexity and cost.

Spatial resolution in HRTF datasets is typically specified separately for azimuth and elevation. Many datasets prioritize high azimuthal resolution, as horizontal localization is perceptually more sensitive and relies strongly on interaural time and level differences (ITDs and ILDs) [16]. In contrast, elevation perception relies primarily on monaural spectral cues and generally exhibits poorer spatial resolution. As a result, elevation resolution in HRTF datasets is often coarser, reflecting both practical measurement constraints and the reduced perceptual sensitivity to small elevation changes. High-resolution datasets may achieve angular steps of only a few degrees in both azimuth and elevation, whereas more compact datasets commonly use angular spacings of 10°–15°or more [6].

SOUND SOURCE DISTANCE

The distance between the sound source and the listener is another important parameter in the design of HRTF measurements. Far-field measurements ($> 1m$) are commonly used, as they simplify the acoustic model and are representative of many real-world listening scenarios [6, 18]. In the far field, wavefronts can be approximated as planar, and the acoustic transfer functions become largely independent of distance [21].

Near-field measurements, in contrast, capture distance-dependent effects such as changes in sound pressure level. While near-field HRTFs are relevant for applications involving sources close to the listener, they are less commonly included in large datasets due to the

increased complexity of the measurement and modeling process. Consequently, most datasets focus on far-field measurements to balance general applicability and practical feasibility.

MICROPHONE CONFIGURATIONS

Microphone configuration is a central design aspect of an HRTF dataset, as it determines how the sound field around the listener is sampled and which spatial cues are captured. Existing datasets employ a variety of microphone placements, ranging from single in-ear microphones to more complex multi-microphone arrangements distributed around the ears or head [6].

Most existing datasets employ a single microphone per ear, typically positioned at the blocked ear canal entrance, as this provides an accurate, practical, and reproducible approximation of the sound pressure at the eardrum. Single microphone-per-ear configurations are well-suited for spatial audio rendering and localization studies, but they inherently limit the spatial information available for array-based signal processing [2].

In contrast, multi-microphone configurations place multiple microphones per ear, distributed across the pinna, behind the ear, or at surrounding outer-ear locations, and are primarily motivated by hearing aid and microphone array research [26, 31]. These setups enable spatial sampling of the local sound field and provide additional degrees of freedom for beamforming and noise reduction.

Multi-microphone configurations, however, introduce increased measurement complexity. As a result, publicly available multi-microphone HRTF datasets remain relatively scarce and are typically limited to two or three microphones per ear. This small number of microphones significantly constrains the spatial degrees of freedom available for array-based processing and limits investigation of how increasing the number of microphones affects binaural beamforming performance, noise reduction capability, and binaural cue preservation.

POST-PROCESSING

After measurement, the audio data typically undergo several post-processing steps before being stored or distributed [6]. The raw microphone recordings are first deconvolved with the inverse of the excitation signal, which is commonly an exponential sine sweep, to obtain head-related impulse responses (HRIRs). To suppress the influence of room reflections and late reverberation, the resulting HRIRs are usually windowed in the time domain. Windowing limits the impulse response to the direct sound and early reflections, thereby reducing spectral coloration caused by residual room effects, particularly when measurements are performed in non-anechoic environments. Following windowing, the HRIRs are transformed to the frequency domain with an FFT to obtain HRTFs. Some datasets also apply equalization to account for microphone and loudspeaker responses.

SPATIALLY ORIENTED FORMAT FOR ACOUSTICS (SOFA)

To facilitate the exchange and reuse of HRTF data, many modern datasets adopt the Spatially Oriented Format for Acoustics (SOFA) [32]. SOFA provides a standardized framework for storing HRTFs, HRIRs, and related metadata, including source positions, listener geometry, sampling rate, and measurement conditions. By enforcing a common structure and coordinate system, SOFA enables interoperability between datasets and software tools and reduces ambiguity in the interpretation of spatial audio data.

While SOFA greatly improves data accessibility and consistency, it does not impose constraints on measurement design itself. The usefulness of a dataset for a specific application, such as multi-microphone binaural beamforming, therefore still depends primarily on the underlying measurement choices, including microphone count, spatial resolution, and source configuration.

2.1.3. LIMITATIONS OF EXISTING DATASETS

Although existing HRTF datasets provide valuable resources for spatial audio rendering and localization research, they have limitations when considered from the perspective of multi-microphone binaural beamforming. Most notably, the majority of publicly available datasets employ only one microphone per ear, and even datasets that incorporate multiple microphones are typically limited to two or three microphones per ear. While such configurations are sufficient for capturing binaural cues at a perceptual level, they provide only a limited number of degrees of freedom for array-based signal processing [1].

In binaural beamforming, the number and spatial arrangement of microphones are directly related to the achievable noise reduction [1, 2]. With only a small number of microphones per ear, it is not possible to systematically study how increasing microphone density affects beamformer performance, nor to explore trade-offs between noise suppression and the imposition of beamformer constraints such as binaural cue preservation.

These limitations motivate the development of a new HRTF dataset specifically designed for multi-microphone binaural beamforming research. By increasing the number of microphones per ear while maintaining controlled measurement conditions.

2.2. MULTI-MICROPHONE BINAURAL BEAMFORMING

Hearing-impaired people generally have more difficulties in understanding a target talker in complex acoustic environments with multiple interfering sources compared to normal hearing people [7]. To reduce noise and improve speech intelligibility, both single-microphone and multi-microphone methods can be used [1, 2]. The latter is the most effective in speech intelligibility compared to the former, as it can make use of temporal and spatial properties. Some common examples of multi-microphone noise reduction methods are the Multi-channel Wiener Filter (MWF) [33], the Minimum Variance

Distortionless Response (MVDR) beamformer [19], and its generalization, the Linearly Constrained Minimum Variance (LCMV) beamformer [20]. The focus of this thesis will be on the beamforming methods.

Traditional hearing assistive devices (HADs) have been fitted bilaterally, meaning the user wears an HAD on each ear, and they operate independently from each other. These monaural systems 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. This affects the naturalness of the sound perception. Binaural HADs, on the other hand, are able to wirelessly exchange microphone signals between HADs. This way, a higher number of microphones is available for usage compared to the bilateral HADs, which can lead to better noise suppression and therefore a higher speech intelligibility [2, 7].

This section will provide an in-depth review of the methods covered in the PhD thesis of Andreas Koutrouvelis [1]. This review includes the multi-microphone signal model, a review on the different beamformers used, and how the problem with virtual sources is handled. Some of the beamforming methods discussed are the General Binaural LCMV (GBLCMV) and the Joint BLCMV (JBLCMV) beamformers.

2.2.1. MULTI-MICROPHONE SIGNAL MODEL

It is assumed that each of the two HADs consists of a microphone array of $\frac{M}{2}$ microphones. Therefore, a total number of M microphones is used. The multi-microphone noise reduction methods that are considered in this thesis operate in the frequency domain on a frame-by-frame basis. Let l denote the frame index and k the frequency bin index. It is also assumed that there is only one target source and r interfering sources. The j-th noisy microphone signal, with $j = 1, \ldots, M$, is given by Eq. 2.1

$$y_j(k,l) = \underbrace{a_j(k,l)s(k,l)}_{x_j(k,l)} + \sum_{i=1}^{r} \underbrace{b_{ij}(k,l)u_i(k,l)}_{n_{ij}(k,l)} + v_j(k,l), \tag{2.1}$$

where the different signals are defined as follows:

- $y_j(k,l)$ denotes the j-th noisy microphone signal,
- $s(k,l)$ denotes the target signal at the source location,
- $a_j(k,l)$ represents the HRTF of the target signal with respect to the j-th microphone,
- $u_i(k,l)$ denotes the i-th interfering signal at the source location,
- $b_{ij}(k,l)$ represents the HRTF of the i-th interfering signal with respect to the j-th microphone,
- $x_j(k,l)$ denotes the received target signal at the j-th microphone,

- $n_{ij}(k,l)$ denotes the received i-th interfering signal at the j-th microphone, and
- $v_j(k,l)$ denotes additive noise at the j-th microphone.

In the rest of this thesis, the frame and frequency indices are omitted for notational convenience. This signal model can be written in vector notation as shown in Eq. 2.2

$$\mathbf{y} = \underbrace{\mathbf{a}s}_{\mathbf{x}} + \sum_{i=1}^{r} \underbrace{\mathbf{b}_i u_i}_{\mathbf{n}_i} + \mathbf{v} \tag{2.2}$$

where $\mathbf{y} \in \mathbb{C}^{M\times 1}$, $\mathbf{x} \in \mathbb{C}^{M\times 1}$, $\mathbf{n}_i \in \mathbb{C}^{M\times 1}$, $\mathbf{v} \in \mathbb{C}^{M\times 1}$, $\mathbf{a} \in \mathbb{C}^{M\times 1}$, and $\mathbf{b}_i \in \mathbb{C}^{M\times 1}$ are stacked vectors. Assuming that all sources and additive noise are uncorrelated, the Cross Power Spectral Density Matrix (CPSDM), $\mathbf{P}_y \in \mathbb{C}^{M\times M}$, of the noisy measurements $\mathbf{y}$, is given in Eq. 2.9

$$\mathbf{P}_y = E[\mathbf{y}\mathbf{y}^H] = \mathbf{P}_x + \underbrace{\sum_{i=1}^{r} \mathbf{P}_{n_i} + \mathbf{P}_v}_{\mathbf{P}} \tag{2.3}$$

where $E[\cdot]$ denotes the statistical expectation, and the CPSDMs are defined as follows:

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

2.2.2. BINAURAL BEAMFORMING

Binaural multi-microphone noise reduction methods aim at the simultaneous noise reduction and binaural cue preservation of the sources. Two estimated 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 HADs respectively. Note that both spatial filters use all microphones. In order to preserve the binaural cues, constraints can be used to guarantee that certain phase and magnitude relations between the left and right HADs output are preserved, while minimizing the noise output power. The estimated output signals of the HADs can, therefore, be written as

$$\hat{x}_L = \hat{\mathbf{w}}_L^H \mathbf{y} \quad \text{and} \quad \hat{x}_R = \hat{\mathbf{w}}_R^H \mathbf{y}. \tag{2.4}$$

for the left and right HAD, respectively. Note that the subscripts L and R refer to the reference microphone signals where each HAD has one reference microphone. This notation will be used for the rest of this thesis for convenience.

In the rest of this section, first, objective measures for binaural cue preservation are presented. Next, the various beamforming methods like BMVDR, BLCMV, and JBLCMV, which are special cases of the General BLCMV (GBLCMV) beamformer, proposed by [1] are reviewed.

BINAURAL CUES

Following [1], binaural cue preservation of a specific source can be expressed in terms of the input and output Interaural Transfer Function (ITF). The ITF is defined as the ratio between the acoustic transfer functions to the left and right ears for a given source direction and can be decomposed into a magnitude component and a phase component. The magnitude of the ITF corresponds to the Interaural Level Difference (ILD), while the phase of the ITF corresponds to the Interaural Phase Difference (IPD). The IPD describes the phase difference between the left- and right-ear signals as a function of frequency. Under the assumption that the interaural difference can be approximated by a relative time delay between the ears, the IPD is directly related to the Interaural Time Difference (ITD) through a linear phase relationship. The input and output ITFs of the i-th interfering source are given by

$$ITF_{n_i}^{in} = \frac{b_{i,L}}{b_{i,R}}, \qquad ITF_{n_i}^{out} = \frac{\hat{\mathbf{w}}_L^H \mathbf{b}_i}{\hat{\mathbf{w}}_R^H \mathbf{b}_i}, \tag{2.5}$$

and quantify how well the binaural magnitude and phase relations between the left and right ears are preserved by the beamformer. From the ITF, the ILD and IPD are obtained as

$$ILD = |ITF|^2, \qquad IPD = \angle ITF, \tag{2.6}$$

where $|\cdot|$ denotes the magnitude and $\angle(\cdot)$ denotes the phase angle of a complex-valued quantity. ILDs are most informative at higher frequencies above 3 kHz, while IPDs are most informative at lower frequencies below approximately 1 kHz [16].

A binaural beamformer $\hat{\mathbf{w}} = \begin{bmatrix}\hat{\mathbf{w}}_L^T & \hat{\mathbf{w}}_R^T\end{bmatrix}^T$ exactly preserves the binaural cues of the i-th interferer when $ITF_{n_i}^{in} = ITF_{n_i}^{out}$, in which case exact preservation of the ITF also implies exact preservation of the corresponding ILD and IPD. When binaural cue preservation is not exact, a mismatch arises between the input and output ITFs, which is quantified by the ITF error

$$\varepsilon_{n_i} = |ITF_{n_i}^{in} - ITF_{n_i}^{out}|. \tag{2.7}$$

In the remainder of this thesis, binaural cue preservation and noise reduction are evaluated over the complete frequency spectrum.

GBLCMV

All binaural LCMV-based methods discussed in this chapter are based on the General Binaural LCMV (GBLCMV) beamformer, which minimizes the sum of the left and right noise output powers under multiple linear equality constraints. Specifically, the GBLCMV beamformer is obtained by solving the constrained optimization problem

$$\hat{\mathbf{w}}_{GBLCMV} = \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min}\, \mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{\Lambda} = \mathbf{f}^H, \tag{2.8}$$

where both $\hat{\mathbf{w}}_{GBLCMV}$ and $\mathbf{w}$ are defined as $\mathbf{w} = \begin{bmatrix}\mathbf{w}_L^T & \mathbf{w}_R^T\end{bmatrix}^T \in \mathbb{C}^{2M\times 1}$, which stacks the left and right beamformer weights.

The constraint matrix $\mathbf{\Lambda} \in \mathbb{C}^{2M\times d}$ is assumed to have full column rank d, ensuring that all imposed linear constraints are mutually independent and each enforces a distinct spatial requirement on the beamformer. The vector $\mathbf{f} \in \mathbb{C}^{d\times 1}$ specifies the desired responses associated with these constraints, where d denotes the total number of linear equality constraints. The matrix $\tilde{\mathbf{P}}$ represents the block-diagonal noise covariance matrix defined as

$$\tilde{\mathbf{P}} = \begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}\end{bmatrix}, \tag{2.9}$$

with $\mathbf{P}$ denoting the CPSDM of the microphone signals. The assumption that $\mathbf{\Lambda}$ has full column rank implies that $d \le 2M$; if $d > 2M$, the constraint set becomes infeasible, and the GBLCMV problem admits no solution. When the constraint set is feasible, the optimization problem in Eq. 2.8 gives the closed-form solution

$$\hat{\mathbf{w}}_{GBLCMV} = \begin{cases}\tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}\left(\mathbf{\Lambda}^H\tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{f}, & \text{if } d < 2M,\\ (\mathbf{\Lambda}^H)^{-1}\mathbf{f}, & \text{if } d = 2M.\end{cases} \tag{2.10}$$

It can be noted that the binaural beamformer has a total of $2M$ complex degrees of freedom (DoF), corresponding to the $2M$ complex-valued filter coefficients in $\mathbf{w}$. One complex degree of freedom is equivalent to two real DoF; however, for convenience, the DoF count is expressed in the complex domain. In the GBLCMV formulation, imposing d linear equality constraints leaves $2M - d$ complex DoF available for noise reduction. In the special case where $d = 2M$, no degrees of freedom remain for noise suppression. Consequently, the constraint matrix $\mathbf{\Lambda}$ must be tall, i.e., $d < 2M$, in order to satisfy the constraints while still enabling noise reduction.

To illustrate the meaning of the linear equality constraints, they can be partitioned as

$$\mathbf{w}^H\begin{bmatrix}\mathbf{\Lambda}_1 & \mathbf{\Lambda}_2\end{bmatrix} = \begin{bmatrix}\mathbf{f}_1^H & \mathbf{f}_2^H\end{bmatrix}, \tag{2.11}$$

where the first part enforces distortionless constraints on the target source. Specifically, the constraints $\mathbf{w}_L^H\mathbf{a} = a_L$ and $\mathbf{w}_R^H\mathbf{a} = a_R$ ensure that the target source is preserved at the left and right reference microphones, respectively. These two constraints can be written compactly as

$$\mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H, \qquad \mathbf{\Lambda}_1 = \begin{bmatrix}\mathbf{a} & \mathbf{0}\\ \mathbf{0} & \mathbf{a}\end{bmatrix} \in \mathbb{C}^{2M\times 2}, \qquad \mathbf{f}_1 = \begin{bmatrix}a_L\\ a_R\end{bmatrix} \in \mathbb{C}^{2\times 1}, \tag{2.12}$$

where $\mathbf{a}$ denotes the vector HRTFs from the target source to all microphones, as defined in Section 2.2.1, and a_L and a_R are the HRTFs from the target source to the left and right reference microphones, respectively. All binaural beamforming methods discussed in

the following sections are special cases of the GBLCMV formulation in which the constraints defined by $\mathbf{\Lambda}_1$ are identical, while the additional constraint set $\mathbf{w}^H\mathbf{\Lambda}_2 = \mathbf{f}_2^H$ differs between methods.

BMVDR

The Binaural Minimum Variance Distortionless Response (BMVDR) beamformer is the simplest special case of the GBLCMV formulation, in which only the target source is preserved, and no explicit constraints are imposed on the interfering sources. In this case, the beamformer is obtained by minimizing the total noise output power subject solely to the distortionless constraints for the target source, leading to the optimization problem

$$\begin{aligned} \hat{\mathbf{w}}_{BMVDR} = \quad & \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min} \quad \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \\ & \text{s.t.} \qquad\quad \mathbf{w}^H\mathbf{\Lambda}_1 = \mathbf{f}_1^H. \end{aligned} \tag{2.13}$$

This formulation involves the minimum number of linear equality constraints among the binaural beamforming methods considered. Since the constraint set $\mathbf{\Lambda}_1$ contains $d = 2$ constraints, corresponding to the preservation of the target signal at the left and right reference microphones, the number of DoF available for noise reduction is $2M-2$. As a result, the BMVDR beamformer preserves the binaural cues of the target source while allowing the binaural cues of interfering sources to be distorted. In particular, it has been shown in [1] that the binaural cues of interferers tend to collapse toward those of the target source, leading to a perceptual co-location of target and interfering signals.

BLCMV

The Binaural Linearly Constrained Minimum Variance (BLCMV) beamformer extends the BMVDR formulation by introducing additional linear constraints to preserve the binaural cues of m interfering sources. As a result, the BLCMV beamformer preserves the binaural cues of both the target source and a selected set of m interferers. This is achieved by minimizing the total noise output power subject to distortionless constraints for the target and binaural cue preservation constraints for the interferers, which can be expressed as the constrained optimization problem

$$\begin{aligned} \hat{\mathbf{w}}_{BLCMV} = \quad & \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min} \quad \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \\ & \text{s.t.} \qquad\quad \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H. \end{aligned} \tag{2.14}$$

The complete constraint matrix $\mathbf{\Lambda}$ and the corresponding response vector $\mathbf{f}$ consist of two parts: the distortionless constraints that preserve the target source at the left and right reference microphones, and the binaural cue preservation constraints imposed on the m interfering sources. These are given by

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{a} & \mathbf{0} & \mathbf{b}_1 & \mathbf{0} & \cdots & \mathbf{b}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{a} & \mathbf{0} & \mathbf{b}_1 & \cdots & \mathbf{0} & \mathbf{b}_m \end{bmatrix} \in \mathbb{C}^{2M\times(d=2m+2)}, \tag{2.15}$$

and

$$\mathbf{f}^H = \begin{bmatrix} a_L & a_R & \eta_L b_{1,L} & \eta_R b_{1,R} & \cdots & \eta_L b_{m,L} & \eta_R b_{m,R} \end{bmatrix} \in \mathbb{C}^{1\times(d=2m+2)}. \tag{2.16}$$

Here, $\mathbf{a}$ denotes the vector of HRTFs from the target source to the M microphones, while $\mathbf{b}_i$ denotes the vector of HRTFs from the i-th interfering source to the same microphones. The scalars a_L and a_R are the HRTFs of the target source to the left and right reference microphones, respectively, and $b_{i,L}$ and $b_{i,R}$ denote the corresponding HRTFs of the i-th interferer. The parameters η_L and η_R control the extent to which the binaural cues of the interfering sources are preserved, allowing for a trade-off between interference suppression and spatial cue preservation.

Since the binaural cues of the target source and m interfering sources are preserved, two linear equality constraints are imposed per source, one for each ear. Consequently, the total number of constraints is $d = 2m+2$, which consumes the same number of DoF of the binaural beamformer. Given that the beamformer has $2M$ DoF in total, this leaves $2M-2m-2$ DoF available for noise reduction. This illustrates the fundamental trade-off in the BLCMV formulation: preserving binaural cues for an increasing number of interferers reduces the degrees of freedom that can be devoted to noise suppression.

JBLCMV

The Joint Binaural LCMV (JBLCMV) beamformer was introduced by [1, 2] as a refinement of the BLCMV approach. The key idea of this method is to preserve the binaural cues of both the target source and interfering sources while allocating more DoF to noise reduction than is possible with the BLCMV. Unlike the BLCMV, which requires two constraints per interferer, the JBLCMV formulation requires only a single constraint per interferer to preserve its binaural cues. This reduction is achieved by exploiting the fact that exact binaural cue preservation of a source is equivalent to equality of the input and output ITFs. For the i-th interferer, this condition can be expressed as

$$ITF_{n_i}^{in} = ITF_{n_i}^{out} \quad \Leftrightarrow \quad \frac{\hat{\mathbf{w}}_L^H \mathbf{b}_i}{\hat{\mathbf{w}}_R^H \mathbf{b}_i} = \frac{b_{i,L}}{b_{i,R}}. \tag{2.17}$$

By rearranging this equality, the binaural cue preservation requirement for the i-th interferer can be written as a single linear constraint,

$$\hat{\mathbf{w}}_L^H \mathbf{b}_i b_{i,R} - \hat{\mathbf{w}}_R^H \mathbf{b}_i b_{i,L} = 0 \tag{2.18}$$

$$\hat{\mathbf{w}}_L^H \overline{\mathbf{b}}_{i,L} - \hat{\mathbf{w}}_R^H \overline{\mathbf{b}}_{i,R} = 0, \tag{2.19}$$

where the normalized HRTF vectors $\overline{\mathbf{b}}_{i,L}$ and $\overline{\mathbf{b}}_{i,R}$ are defined as $\overline{\mathbf{b}}_{i,L} = \mathbf{b}_i / b_{i,L}$ and $\overline{\mathbf{b}}_{i,R} = \mathbf{b}_i / b_{i,R}$, respectively.

This joint constraint enforces equality of the input and output ITFs using a single complex constraint per interferer, thereby reducing the number of constraints required for

binaural cue preservation by about a factor of two compared to the BLCMV formulation. As a consequence, more DoF remain available for noise reduction. The resulting JBLCMV beamformer is obtained by minimizing the total noise output power subject to the joint constraint set, leading to the optimization problem

$$\begin{aligned} \hat{\mathbf{w}}_{JBLCMV} = \quad & \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min} \quad \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \\ & \text{s.t.} \qquad\quad \mathbf{w}^H\mathbf{\Lambda} = \mathbf{f}^H. \end{aligned} \tag{2.20}$$

In this formulation, the constraint matrix $\mathbf{\Lambda}$ and the corresponding response vector $\mathbf{f}$ are defined as

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 & | & \mathbf{\Lambda}_2 \end{bmatrix} \tag{2.21}$$

$$= \begin{bmatrix} \bar{\mathbf{a}}_L & \mathbf{0} & | & \bar{\mathbf{b}}_{1,L} & \cdots & \bar{\mathbf{b}}_{m,L} \\ \mathbf{0} & \bar{\mathbf{a}}_R & | & -\bar{\mathbf{b}}_{1,R} & \cdots & -\bar{\mathbf{b}}_{m,R} \end{bmatrix} \in \mathbb{C}^{2M\times(d=m+2)} \tag{2.22}$$

and

$$\mathbf{f}^H = \begin{bmatrix} \mathbf{f}_1^H & | & \mathbf{f}_2^H \end{bmatrix} \tag{2.23}$$

$$= \begin{bmatrix} 1 & 1 & | & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{1\times(d=m+2)}, \tag{2.24}$$

where $\bar{\mathbf{a}}_L = \mathbf{a}/a_L$ and $\bar{\mathbf{a}}_R = \mathbf{a}/a_R$ denote the normalized HRTFs of the target source to the left and right reference microphones, respectively.

Since the JBLCMV formulation requires only one constraint per interfering source in addition to the two target-preserving constraints, the total number of constraints is $d = m+2$. Consequently, the JBLCMV beamformer can preserve the binaural cues of up to $m = 2M-3$ interfering sources simultaneously. The number of DoF available for noise reduction is therefore given by $DoF_{JBLCMV} = 2M-m-2$, highlighting the improved trade-off between binaural cue preservation and noise suppression compared to the BLCMV method.

BEAMFORMING METHODS SUMMARY

The beamforming methods discussed above are shortly summarized in this section. Table 2.2 provides an overview of the BMVDR, the BLCMV, and the JBLCMV methods, where m_{max} is given as the maximum number of interferers that the method can constrain while still having at least one DoF left to apply noise reduction. DoF is given to be the amount of DoF a method can devote to noise reduction.

It should also be noted that, on top of the methods mentioned, [1] also came up with relaxations of the JBLCMV problem, which will not be covered in this report.

Table 2.2: Various beamforming methods with their maximum number of interferers for which they can preserve their binaural cues while still having a DoF left for noise reduction and their degrees of freedom to devote to noise reduction

Method	m_{max}	**DoF**
BMVDR	0	$2M-2$
BLCMV	$M-2$	$2M-2m-2$
JBLCMV	$2M-3$	$2M-m-2$

2.2.3. VIRTUAL SOURCES WITH PREDETERMINED HRTFS

Virtual sources with predetermined HRTFs were introduced by [1] to address the difficulty of estimating source-dependent HRTFs in real time. A set of m virtual sources is defined, each associated with a predetermined pair of normalized HRTFs $(\overline{\mathbf{q}}_{i,L}, \overline{\mathbf{q}}_{i,R})$, for $i = 1,\dots,m$. These normalized HRTFs are given by $\overline{\mathbf{q}}_{i,L} = \mathbf{q}_i / q_{i,L}$ and $\overline{\mathbf{q}}_{i,R} = \mathbf{q}_i / q_{i,R}$, and represent the HRTFs from the i-th virtual-source direction to the left and right reference microphones, respectively. The virtual sources are assumed to be located in the far field, such that the corresponding HRTFs are approximately distance-invariant and depend only on the source direction.

Assuming that r interfering sources are present in the acoustic scene, exact preservation of the binaural cues of an interferer is achieved when one of the predetermined virtual-source HRTF pairs coincides with the normalized HRTF pair of that interferer. Specifically, if for some i and j it holds that $(\overline{\mathbf{q}}_{i,L}, \overline{\mathbf{q}}_{i,R}) = (\overline{\mathbf{b}}_{j,L}, \overline{\mathbf{b}}_{j,R})$, and the corresponding virtual-source constraints are included in the beamformer design, then the binaural cues of the j-th interferer are preserved by construction.

In practice, the HRTF pairs of real interferers will generally not coincide exactly with any predetermined virtual-source HRTF pair. This mismatch gives rise to a steering vector mismatch (SVM), reflecting the discrepancy between the true interferer HRTFs and those assumed in the beamformer constraints. As a result, binaural cues are only approximately preserved. To quantify this approximation error, the virtual-source binaural cue preservation error is defined as

$$\varepsilon_{q_i} = |ITF_{q_i}^{in} - ITF_{q_i}^{out}| \tag{2.25}$$

$$= \left| \frac{\hat{\mathbf{w}}_L^H \mathbf{q}_i}{\hat{\mathbf{w}}_R^H \mathbf{q}_i} - \frac{q_{i,L}}{q_{i,R}} \right|, \quad i = 1,\dots,m. \tag{2.26}$$

This error metric does not appear as an explicit constraint in the beamformer optimization, but is used to evaluate how well the binaural cues associated with the virtual sources are preserved in the presence of steering vector mismatch.

The error metric in Eq. 2.25 is formulated directly in terms of the ITF, as this quantity jointly captures the binaural magnitude and phase relations underlying spatial perception, unlike most alternative error metrics. By measuring the absolute difference between the input and output ITFs, the metric reflects deviations in both ILDs and IPDs

within a single complex-valued measure. This aligns naturally with the JBLCMV framework, where binaural cue preservation is enforced through ITF equality constraints.

When predetermined virtual-source HRTFs are used in place of unknown interferer HRTFs, the JBLCMV beamforming problem retains the same mathematical structure, with the constraint matrix and response vector defined in terms of the virtual sources. The resulting optimization problem is given by

$$\begin{aligned} \hat{\mathbf{w}}_{JBLCMV} = \quad & \underset{\mathbf{w}\in\mathbb{C}^{2M\times 1}}{\arg\min} \quad \mathbf{w}^H\tilde{\mathbf{P}}\mathbf{w} \\ & \text{s.t.} \qquad\quad \mathbf{w}^H\Lambda = \mathbf{f}^H, \end{aligned} \tag{2.27}$$

where the constraint matrix $\mathbf{\Lambda}$ and response vector $\mathbf{f}$ are defined as

$$\mathbf{\Lambda} = \left[\mathbf{\Lambda}_1 \mid \mathbf{\Lambda}_2\right] \tag{2.28}$$

$$= \begin{bmatrix} \overline{\mathbf{a}} & \mathbf{0} \mid & \overline{\mathbf{q}}_{1,L} & \cdots & \overline{\mathbf{q}}_{m,L} \\ \mathbf{0} & \overline{\mathbf{a}} \mid & -\overline{\mathbf{q}}_{1,R} & \cdots & -\overline{\mathbf{q}}_{m,R} \end{bmatrix} \in \mathbb{C}^{2M\times(d=m+2)}, \tag{2.29}$$

and

$$\mathbf{f}^H = \left[\mathbf{f}_1^H \mid \mathbf{f}_2^H\right] \tag{2.30}$$

$$= \begin{bmatrix} 1 & 1 \mid 0 & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{1\times(d=m+2)}. \tag{2.31}$$

In this formulation, the virtual-source constraints enforce exact binaural cue preservation only at the predetermined virtual-source directions. The binaural cue preservation error in Eq. 2.25 therefore provides a quantitative measure of how well arbitrary interferer locations are approximated by the selected set of virtual sources, and forms the basis for the virtual source selection and optimization procedures investigated in later chapters.

UNIFORM VIRTUAL SOURCE PLACEMENT

In [1], virtual sources with predetermined HRTFs are typically assumed to be uniformly distributed along the perimeter of a circle in the horizontal plane. This assumption simplifies the beamformer design and guarantees symmetric coverage of the azimuth domain. However, uniform angular spacing is not necessarily optimal for binaural cue preservation or noise reduction in realistic acoustic scenes.

In practice, the directional resolution required for accurate binaural cue representation does not need to be uniform across all azimuths. Certain directions may benefit more from a higher density of virtual sources due to larger spatial gradients in ILDs and IPDs. Conversely, directions that contribute less to the overall perceptual error may require fewer constraints.

Relaxing the assumption of uniform virtual source placement and instead optimizing the spatial configuration of virtual sources therefore offers the potential for improved performance. An optimized configuration aims to minimize the expected mismatch between virtual-source HRTFs and actual interferer HRTFs, thereby improving binaural cue preservation, while simultaneously retaining as many degrees of freedom as possible for noise suppression. These considerations directly motivate the virtual source location optimization framework introduced in later chapters and are revisited in the discussion of the experimental results.

3
HRTF MEASUREMENTS

Accurate and high-resolution Head-Related Transfer Function (HRTF) measurements form the foundation of any binaural beamforming method. While many existing HRTF databases provide detailed responses for a wide range of source directions, they typically rely on only two to three microphones per ear. Such configurations are sufficient for classical binaural rendering, but they limit the ability to explore how larger microphone arrays around the ears affect spatial filtering, noise reduction performance, and localization accuracy.

In contrast, the database created in this thesis significantly expands the number of microphones used during measurement. Each ear of the dummy head is equipped with a custom multi-microphone array consisting of six microphones, resulting in a total of twelve microphones. The trade-off, on the other hand, is that the newly generated database has a significatly lower spatial resolution compared to other datasets.

The objective of this chapter is to describe the complete process of acquiring this new, multi-microphone HRTF dataset. Section 3.1 first analyzes the measurement environment and determines the achievable signal-to-noise ratio. The measurement setup, consisting of the sound source arrangement, microphone array, and routing configuration, is explained in detail in Section 3.2. Furthermore, the process of obtaining the data and the resulting impulse responses and HRTFs are presented and discussed in Section 3.3.

3.1. ACOUSTIC ANALYSIS OF THE MEASUREMENT ROOM

To begin with the HRTF measurements, it is necessary to know more about the acoustics of the environment in which they are performed. This chapter, therefore, provides an analysis of the room acoustics as well as a measure of the SNR values that can be obtained.

First, the characteristics of the measurement room, the measurement equipment, and the measurement setup is given in Section 3.1.1. Then, an analysis will be performed on the acoustics of the room in Section 3.1.2. Finally, in Section 3.1.3, it will be explained how the SNR was obtained and calculated through measurements.

3.1.1. MEASUREMENT ROOM, EQUIPMENT, AND SETUP

The HRTF measurements will be performed in the Audio Lab of the Signal Processing (SPS) Group in the EWI Faculty of TU Delft. The audio lab can be seen in Fig. 3.1 and its

Table 3.1: Room dimensions for the measurement room and the inner dimensions of the acoustic curtain that encloses the measurement space.

Dimensions	Room	Room with curtain
Width [m]	6.68	4.38
Length [m]	7.98	6.31
Height [m]	3.09	3.09

dimensions are given in Table 3.1. Note that the audio lab is a regular office space with the given dimensions with no acoustical treatment other than a curtain that completely surrounds an inner area of the room for which the dimensions of this area are also given in Table 3.1. This curtain absorbs mid- to high-frequency frequencies from about 700 Hz onward. More on this will be explained in Section 3.1.2.

Figure 3.1: Measurement room with curtains that can be completely closed

A list of the equipment used for the acoustical analysis and the SNR measurements is given in Table 3.2. The Genelec speaker has a flat frequency response with a bandwidth of 41 Hz - 25 kHz. The AKG microphone is an omnidirectional microphone with a bandwidth of 20 Hz - 20 kHz.

Table 3.2: List of the equipment used for the measurements

Type	Brand	Model
Speaker	Genelec	8040
Microphone	AKG	C417 PP
Audio Interface	RME	Fireface UFX+
DA Converter	RME	M-16 DA

For the room acoustics and SNR measurements, a simplified measurement setup is used

in which a single microphone is placed at the center of the room, corresponding to the position of the dummy head during the subsequent HRTF measurements. A single loudspeaker is positioned at a distance of 2 m from the microphone at the same height above the ground. This configuration allows the acoustic transfer function between one sound source and one receiver to be measured, which is sufficient for characterizing the room response and determining the achievable signal-to-noise ratio. The physical layout of this setup is shown in Fig. 3.2.

The microphone signal is routed to the audio interface, while the excitation signal generated in Matlab is sent to the loudspeaker via the DA converter using an optical ADAT connection. The complete playback and recording chain, including digital routing and synchronization, follows the same signal flow principles as used for the full HRTF measurements and is illustrated in Fig. 3.12. All internal routing is handled using the RME TotalMix FX software to ensure consistent gain settings and phase-coherent recording.

Figure 3.2: Acoustical and SNR measurement setup

3.1.2. ROOM ACOUSTICS

To analyze the acoustics of the measurement room, the setup and equipment discussed in Section 3.1.1 were used together with Matlab and Room EQ Wizzard (REW) software. This section aims to get an idea of what the frequency response and decay times are of the desired frequency range, and if necessary, and within the possibilities, apply acoustic treatment to the room. Note that this analysis does not take into account all possible microphone and sound source positions but is more tailored to the application of performing HRTF measurements.

For the measurements, exponential sine sweeps at 48 kHz with a length of 5.5 s and a bandwidth of 20 Hz - 20 kHz are sent to the speaker and picked up by the microphone. Note that we are only interested in the 40 Hz - 20 kHz bandwidth as the lower limit is set by the speaker characteristics and the upper limit is set by the microphone characteristics.

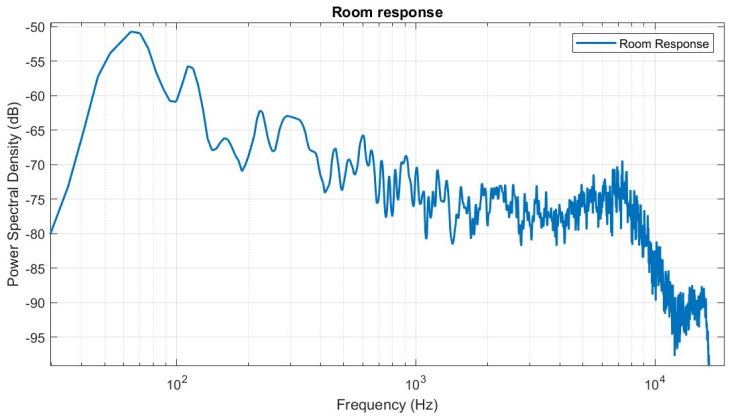


Figure 3.3: Power Spectral Density of the Room Response

The power spectral density of the measured signal can be seen in Fig. 3.3. This graph shows the frequency response of the room for the given microphone and speaker location, and shows the power present in the signal in dBFS as a function of frequency. From the spectrogram shown in Fig. 3.5a, the power present in the signal in dBFS as a function of frequency and how it varies over time is shown. This graph gives a good representation of the decay times of all frequencies. From these graphs, it is observed that three main bandwidths differ in their characteristics.

- 40 **Hz -** 200 **Hz:** In this low frequency band, it can be noticed that there are extreme peaks and valleys present at certain frequencies that also have very long decay times. They result from constructive and destructive interference, respectively, and depend on the room's dimensions and the listening and sound source positions. They are also called room modes or standing waves. This phenomenon can only be resolved by treating the room acoustically using bass traps and Helmholtz resonators. However, this falls out of the scope of this project and would become too expensive.

- 200 **Hz -** 4 **kHz:** This frequency range has a relatively evenly distributed power distribution with some small peaks and valleys (< 5 dB), which could be due to early reflections. They could be reduced with acoustical treatment of the room using absorption panels and diffusers, but again, this would fall out of the scope of this project and would be quite expensive. From the spectrogram it can be seen that within this bandwidth, decay times are constant which is desired. Note that around 450 Hz there the decay time is a lot longer meaning that some object in the room resonates at this frequency.

- 4 **kHz -** 20 **kHz:** In this frequency range, it can be noticed that a lot of power is present in the high frequencies between 4 kHz and 10 kHz. This increase in power is likely to come from the microphone response. The plot in Fig. 3.4 represents the

microphone response where it can be clearly seen that there is a 6 dB boost around 8 – 10 kHz. This response could be taken into account in the HRTF measurements.

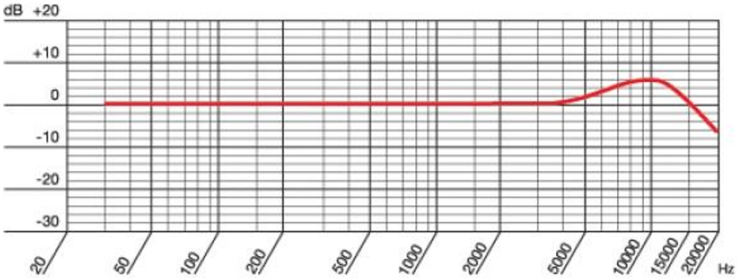


Figure 3.4: Frequency response of the AKG C417 PP microphone [34].

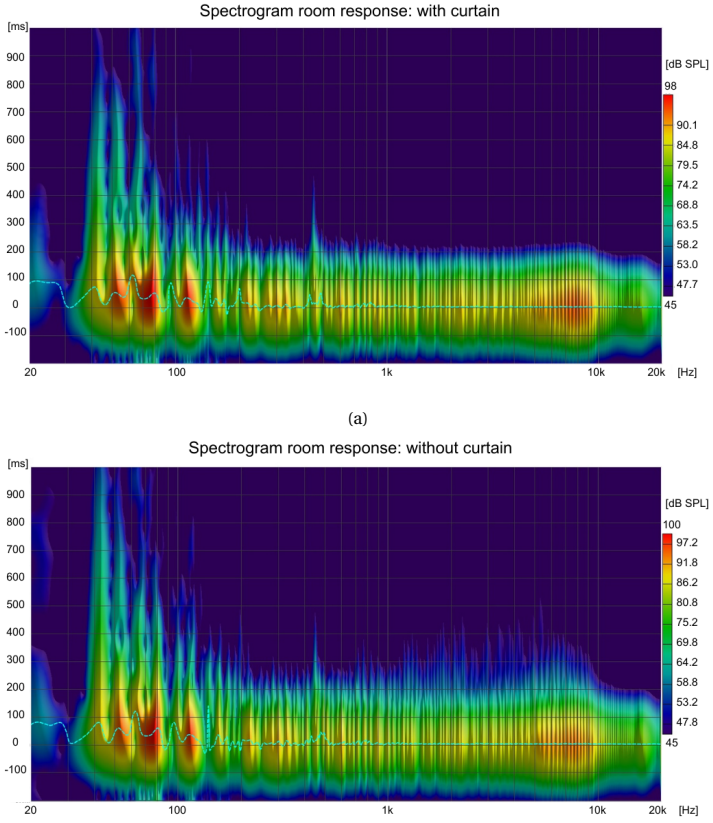


Figure 3.5: Spectrograms of the room response with (a) and without (b) acoustical curtains

Fig. 3.5 shows two spectrograms, with time and frequency represented on the horizontal and vertical axes, respectively, and signal magnitude indicated by color. The first spectrogram corresponds to the measured room response when the acoustic curtain surrounds the measurement setup, while the second shows the response with the curtain fully opened and effectively absent. A clear difference can be observed in the decay behavior: without the curtain, frequencies above 700, Hz exhibit significantly longer decay times compared to the configuration with the curtain present. This demonstrates the beneficial effect of the acoustic curtain on the room acoustics.

3.1.3. SNR MEASUREMENTS

The signal-to-noise ratio (SNR) represents the level of the desired signal relative to the level of background noise. In this section, the method used to determine the maximum achievable SNR for the measurement setup described earlier is explained.

To calculate the maximum achievable SNR, two separate recordings are performed under identical measurement conditions. First, the noise floor of the environment is recorded without any active sound source. Next, the response of the room to a logarithmic sine sweep reproduced by the loudspeaker is recorded. Using identical microphone gains and audio interface settings ensures that both recordings are directly comparable.

The SNR is computed from the ratio of the root-mean-square (RMS) values of the recorded signal and noise:

$$\mathrm{SNR}_{\mathrm{dB}} = 20\log_{10}\left(\frac{x_{RMS}}{v_{RMS}}\right), \tag{3.1}$$

where x denotes the recorded microphone signal during playback of the sweep, and v denotes the recorded noise-floor signal. The RMS value of a signal y is defined over a discrete sum of length T as

$$y_{RMS} \triangleq \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} y^2[n]}. \tag{3.2}$$

Since both recordings are evaluated over the same duration T, the normalization factor $\frac{1}{T}$ cancels out in the ratio. Moreover, by Parseval's theorem, the RMS computed in the time domain is equivalent to the RMS computed in the frequency domain, so the SNR estimate is independent of the chosen domain.

The recording process is automated using a custom Matlab script, which first records the noise floor for a duration of 5 seconds. Then, a logarithmic sine sweep from 40 Hz to 20 kHz is generated and played back through the speaker while simultaneously recording the room response through the microphone. Recordings are performed at 48 kHz. A quadratic fade-in and fade-out of 10 ms is applied to the sweep to avoid sharp transients at the beginning and end of the signal.

Fig. 3.6 shows the signals obtained during the measurement procedure in time domain and frequency domain. The orange line represents the original sweep signal, the blue

line shows the recorded signal, and the red line indicates the recorded noise floor. From these signals, the x_{RMS} and v_{RMS} values are calculated, and subsequently the SNR is determined.

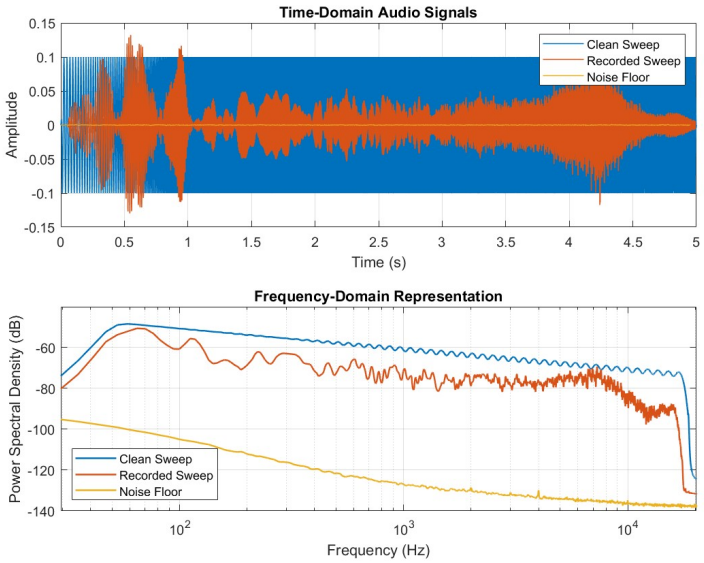


Figure 3.6: Time & frequency domain of sweep signal (blue), recorded signal (orange), and noise floor (yellow)

From the measurements, it was found that a maximum SNR of approximately $45dB$ could be achieved and was considered the practical maximum SNR achievable under comfortable listening conditions in the current measurement environment.

This SNR value is deemed sufficient for the intended HRTF measurements, where a clear distinction between the recorded signal and the background noise is essential for accurate impulse response extraction and further processing.

3.2. MEASUREMENT SETUP

The HRTF measurements were carried out in the same acoustic environment described in Section 3.1. The experimental setup in Fig. 3.7 consisted of a semi-circular arrangement of loudspeakers and a dummy head positioned at the center of the semi-circle with a multi-microphone array on both sides around the ears. A semi-circular setup only provides half the data that is required, and therefore the dummy is rotated 180° to obtain a complete set of measurements 360° around the dummy. The goal was to obtain direction-dependent head related transfer functions between each loudspeaker and each microphone positioned on the dummy to generate a new multi-microphone database of HRTF measurements.

3.2.1. SOUND SOURCE CONFIGURATION

The sound sources are arranged in a semi-circular arc around the Head Acoustics HSU III dummy head at a radius of 1.5 m, as shown in Fig. 3.7. This distance places all loudspeakers well within the acoustic far-field for the frequencies of interest, ensuring that the sound waves can be approximated as locally planar at the dummy's position. The speakers are mounted such that their acoustic centers align with the ear height of the dummy head, maintaining a constant elevation angle and isolating the measurements to the horizontal plane.

A total of twelve Genelec 8040 loudspeakers are used, each spaced 15° apart. The loudspeaker positioned directly in front of the dummy is defined as 0°. In the initial configuration (shown in yellow in Fig. 3.8), the arc spans from −90°, directly facing the left ear, to +75°, near the right ear. To obtain a complete 360° dataset, the dummy head is rotated by 180°, enabling an additional set of measurements from the rear hemisphere (shown in blue). This procedure results in a total of 24 unique source directions, uniformly distributed around the dummy head.

Figure 3.7: Sound source configuration

Although this angular resolution is coarser than that of many large-scale HRTF databases, it represents a deliberate trade-off between spatial sampling density and practical feasibility. The chosen resolution allows systematic evaluation of virtual source placement strategies and binaural beamforming behavior, while keeping the measurement effort and data dimensionality manageable.

All loudspeakers are mounted on height-adjustable stands to ensure precise vertical and horizontal alignment with the dummy head's ear canals. Each loudspeaker is aimed directly toward the geometric center of the semicircle, ensuring consistent on-axis radiation for all measurement directions. Spatial alignment was validated using a laser distance meter and angle measurement tools to minimize positional and orientational

errors across the loudspeaker array. This careful calibration ensures that the measured acoustic transfer functions primarily reflect the directional effects of the dummy head and microphone array, rather than errors in the source placement.

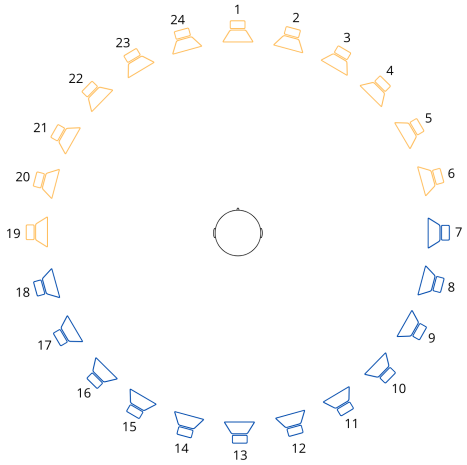


Figure 3.8: Sound source configuration

3.2.2. MICROPHONE ARRAY

A central component of the measurement setup is the custom-built multi-microphone array, designed specifically to position multiple microphones around each ear of the dummy head. Unlike most existing HRTF datasets, which typically employ a small set of 2–3 microphones distributed around the pinna, this setup uses six microphones per ear, resulting in a total of twelve microphones. Fig. 3.9 shows the 3D-model of this custom-built microphone array. While the headband is a fixed component, the arrays themselves are interchangeable. As can be seen from the 3D-model, two different arrays were proposed: a circular array, and a rectangular array. For the measurments, only the rectangular array was used on both sides of the dummy head as it has a simpler integration and easier computational design.

It is important to note that attaching an external multi-microphone array around the ears inevitably alters the acoustic scattering properties of the dummy head. As a result, the measured responses are not valid as perceptually accurate HRTFs, in the sense of representing natural human hearing. However, for the purposes of this research, focusing on beamforming performance rather than perceptual localization, the modification is inconsequential. The array and dummy head together form a composite measurement system, and only the relative transfer functions between microphones and source directions are required. Since the array geometry remains fixed throughout all measurements, the resulting dataset is fully suitable for the multi-microphone analysis carried out in this thesis.

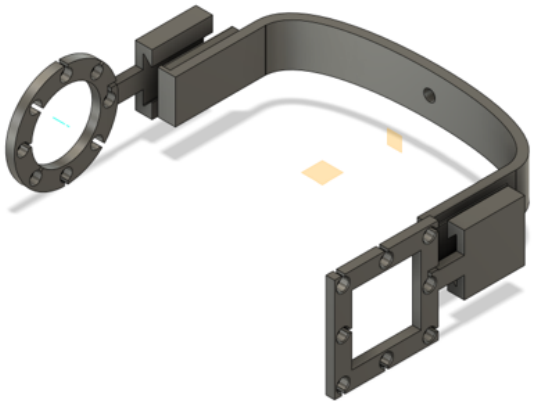


Figure 3.9: 3D model of the microphone array: circular and rectangular array

Each ear is equipped with six microphones, as illustrated in Fig. 3.10. The configuration consists of:

- Five external microphones (AKG C417 PP), mounted on the custom array.
- One built-in microphone located at the ear canal entrance of the dummy head (Head Acoustics HSU III), serving as the reference canal microphone.

This combination allows each ear to capture both the canonical in-ear signal typically used for HRTF datasets, and additional spatial samples in the near-pinna region, leading to improved spatial diversity for beamforming.

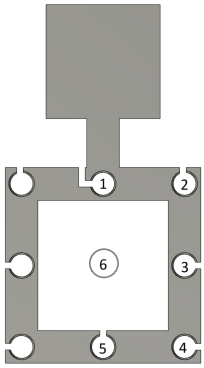


Figure 3.10: Microphone array with microphone positions

Photographs of the assembled microphone arrays mounted on the dummy head are shown in Fig. 3.11.

(a) (b)

Figure 3.11: Microphone array

3.2.3. ROUTING

The routing configuration ensures that the playback and recording chain operates reliably, with synchronized timing and consistent gain structure across all input and output channels. The complete signal flow is illustrated in Fig. 3.12 and consists of three primary components:

- **Playback path:** From Matlab to the loudspeakers (9,1,2, and 3 from Fig. 3.12)
- **Recording path:** From the microphones to Matlab (4,5,6,7,8, and 9 from Fig. 3.12)
- **Digital routing:** Internal routing using RME Totalmix FX

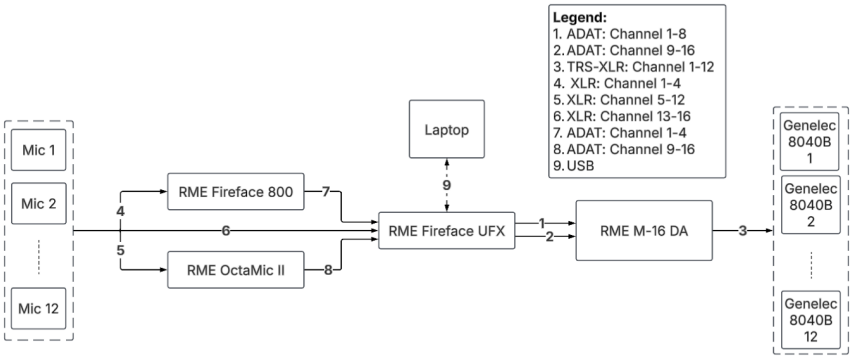


Figure 3.12: Hardware setup HRTF measurements

PLAYBACK PATH

All excitation signals used for the HRTF measurements (exponential sine sweeps) are generated in Matlab and sent to the RME Fireface UFX+ audio interface over USB. The Fireface routes the digital audio to the RME M-16 DA converter using an optical ADAT connection. Each of the 12 loudspeakers is connected to a dedicated analog output channel of the DA converter, ensuring that each source direction can be activated independently without physically reconfiguring cables.

RECORDING PATH

All twelve microphones (six per ear) are connected in a distributed manner to the analog inputs of the RME Fireface 800, the RME OctaMic II, and the RME Fireface UFX, as only the channels with microphone preamplifiers are fit for microphone signals. Finally, all signals come together over optical ADAT to the RME Fireface UFX, which is used as main soundcard that is connected to the laptop. The gain of the mic preamps of the input channels are configured so that all incoming microphone signals have an equal gain. Each microphone is routed directly to its own Matlab input channel, ensuring that the signals remain unprocessed and phase-coherent across all channels.

DIGITAL ROUTING

The routing configuration is managed entirely within RME TotalMix FX, allowing flexible control of playback and recording paths without rewiring physical connections. As there are 12 microphone input channels and 12 loudspeaker output channels, a consistent mapping scheme between Matlab, RME TotalMix FX, and the hardware was maintained throughout the entire measurement session to guarantee that each microphone channel corresponds to the correct physical microphone position during data processing.

3.3. DATABASE

This section presents the multi-microphone HRTF database obtained using the measurement setup described in the previous sections. In total, five complete measurement sets were recorded. Each set contains:

- **Raw microphone audio recordings** for all 24 source directions and 12 microphones as a multi-channel audio file as well as a recording of the noise floor
- **Head related impulse responses (HRIRs)** obtained by deconvolution of the recorded sweeps.
- **Head related transfer functions (HRTFs)**, computed via FFT from the HRIRs.

These five datasets were used to verify the repeatability of the measurement procedure, assess the stability of the microphone routing, and analyze the influence of environmental variations such as dummy alignment, background noise, and loudspeaker drift. All datasets cover the full 360° azimuthal plane and include the responses of all twelve microphones mounted around the dummy head. The loudspeakers are positioned with a spatial resolution of 15° around the dummy at 1.5 m distance. The SNR at which the measurements are performed is 36 dB.

3.3.1. MEASUREMENT PROCESS

Each HRTF is obtained by exciting the environment with an exponential sine sweep and recording the resulting sound pressure signal at all microphones. For every loudspeaker direction, Matlab generates a 5 second exponential sine sweep with a frequency range from 40 Hz to 20 kHz, sampled at 48 kHz with a frame size of 1024 samples per frame. The effective sweep length is 5 seconds, with a short fade-in and fade-out to avoid transient clicks. The exponential sweep method is well-suited for HRTF measurements because it distributes energy equally over octaves, thereby providing high signal-to-noise ratios over the complete frequency spectrum.

For each of the 24 azimuth angles, the sweep is played through the corresponding loudspeaker while all twelve microphones record the response simultaneously. Once the measurements for the frontal semicircle are completed, the dummy head is rotated by 180 degrees and the entire procedure is repeated for the rear semicircle. In this way, a full 360° dataset is constructed without moving the loudspeaker array. All raw recordings are stored directly in Matlab as multichannel audio files, preserving sample-accurate timing across microphones.

After the recording stage, some post-processing is applied. The impulse responses are extracted by deconvolving each microphone signal with the inverse of the exponential sweep. This operation isolates the linear acoustic response of the system and effectively suppresses harmonic distortions produced by the loudspeakers. The resulting impulse responses contain the direct sound followed by early and late reflections from the room. A time window is applied to these HRIRs to limit the influence of unwanted reflections.

The frequency-domain HRTFs are obtained by applying zero-padding and a Fourier transform to the windowed impulse responses. This produces the complex frequency response for every microphone–source combination, from which both magnitude and phase information can be analysed. These HRTFs form the basis for the beamforming experiments in later chapters.

3.3.2. DATA

The resulting database consists of three primary components: the raw multichannel sweep recordings, the time-domain HRIRs obtained through deconvolution, and the HRTFs in the frequency domain. All files are organised per measurement round, ensuring that the data from the five repeated measurements are kept separate for consistency checks and statistical analysis.

The measurement data is arranged such that the first element of each data matrix corresponds to the loudspeaker configuration shown in Fig. 3.8. Specifically, the first element represents the loudspeaker positioned directly in front of the dummy head at 0°, after which the remaining loudspeakers are ordered in a clockwise direction.

The microphone indices are organized such that the first and last $\frac{M}{2}$ microphones correspond to the left and right ears, respectively. Within each ear, the final microphone index

corresponds to the in-ear canal microphone located at the entrance of the ear canal of the dummy head. These in-ear microphones are used as the reference microphones for all binaural cue extraction and subsequent analysis.

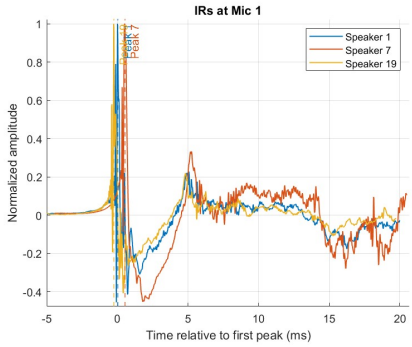
MULTI-CHANNEL AUDIO FILES

The multi-channel audio files are saved as an array of size ($N \times n \times M$), where $N = 24$ is the number of loudspeakers, $n = 240{,}000$ is the number of samples per recording, and $M = 12$ is the number of microphones. These recordings contain the full time-domain response of each microphone to the sweep signals, as well as a separate measurement of the noise floor. The recordings are stored in Matlab format, preserving sample-accurate timing across microphones, which is critical for later deconvolution and for calculating accurate interaural differences.

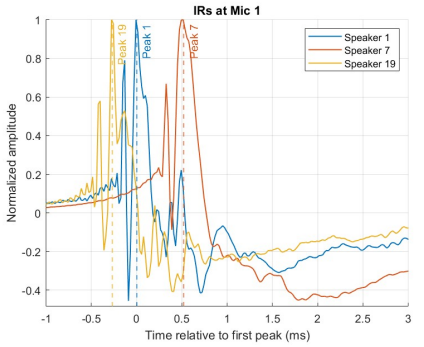
HEAD RELATED IMPULSE RESPONSES (HRIRS)

The HRIRs are obtained by deconvolving the recorded audio signals with the inverse of the exponential sweep. This isolates the linear acoustic response of the system while suppressing harmonic distortions produced by the loudspeakers. A time window is applied to the HRIRs to limit the influence of late reflections, retaining the direct sound and early reflections relevant for spatial processing. The HRIR data is provided in an array of size ($M \times N \times n$), where n is now the number of samples of the windowed HRIR.

In Fig. 3.13, the normalized HRIRs of different loudspeakers are shown for microphone 1 (left ear). The loudspeakers shown are 1, 7, and 18, located at 0°, 90°, and 270°, respectively. In Fig. 3.13a, the amplitude of the normalized HRIR is shown over time, expressed in *ms* relative to the first HRIR peak, allowing the overall structure and amplitude of the responses to be observed. A zoomed-in view around the peaks is shown in Fig. 3.13b, highlighting the relative time delays between the first arrivals from different loudspeakers. These delays are consistent with the spatial positions of the speakers, demonstrating that the measurements capture the expected directional timing differences essential for accurate binaural cue representation.



(a) Normalized HRIR at microphone 1 (left ear)



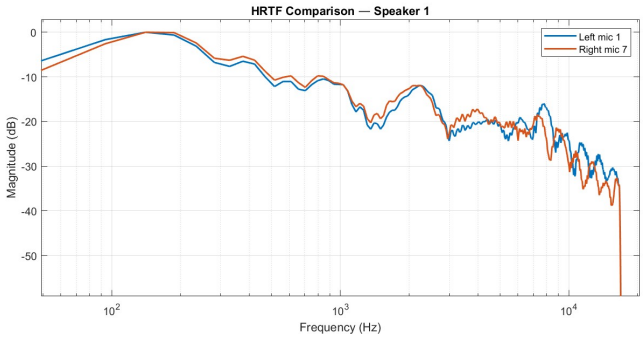
(b) Normalized HRIR peaks at microphone 1 (left ear)

Figure 3.13: Normalized HRIRs of different speakers to the left microphone

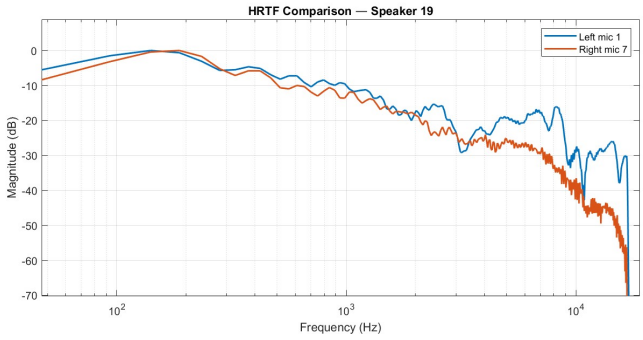
HEAD RELATED TRANSFER FUNCTIONS (HRTFS)

The HRTFs are computed from the windowed HRIRs using zero-padding and a Fourier transform, yielding complex frequency responses for each microphone–source pair. Both magnitude and phase information are preserved, providing the full binaural cue information for subsequent analysis. The HRTF data is provided in an array of size $(M \times N \times k)$, where k represents the number of frequency bins.

In Fig. 3.14, the HRTFs are plotted for two representative speakers. The first image shows the HRTFs for speaker 1 at 0° with respect to the left (mic 1) and right (mic 7) microphone, where the left and right microphones have nearly identical responses across most of the frequency spectrum, except for slight differences at high frequencies due to head shadowing and pinna effects. This symmetry is expected for a source directly in front of the head. The second image shows the HRTFs for speaker 19 at 270°, on the left side of the dummy. Here, a clear difference between left and right microphone responses is visible across a broad range of frequencies, reflecting the interaural level differences and phase differences that represent the main binaural cues for localization. These plots confirm that the measurement setup successfully captures the directional dependence of the acoustic responses, providing reliable data for binaural beamforming and spatial audio experiments.



(a) Normalized HRTFs of speaker 1 (0°) to the left and right microphone



(b) Normalized HRTFs of speaker 19 (270°) to the left and right microphone

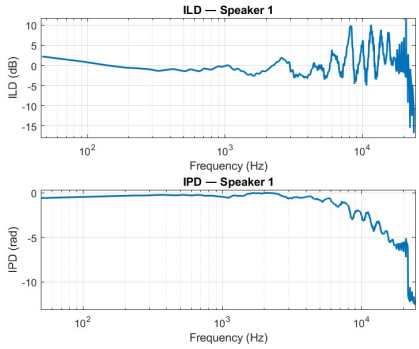
Figure 3.14: Normalized HRTFs to the left and right microphones

BINAURAL CUES

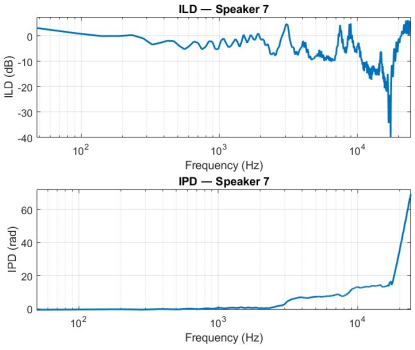
The HRTFs measured in this work encode the binaural cues that allow spatial hearing in the horizontal plane. The most relevant cues are the Interaural Level Difference (ILD), the Interaural Phase Difference (IPD), and the related Interaural Time Difference (ITD). These are computed directly from the complex HRTFs of the left and right reference microphones. The ILD and IPD are determined by Eq. 3.3.

$$ILD(f) = 20\log_{10}\left(\frac{|H_L(f)|}{|H_R(f)|}\right) \qquad IPD(f) = \angle H_R(f) - \angle H_R(f) \tag{3.3}$$

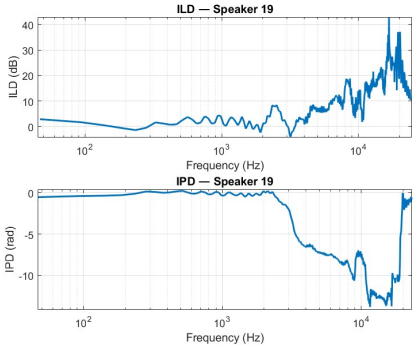
Fig. 3.15 shows the ILD and IPD of loudspeaker 1, 7, and 19 over the complete audio spectrum. It can be seen that the ILD and IPD values for frequencies under 2 kHz are close to zero while these differences are much higher at higher frequencies. For loudspeaker 1, the values of ILD and IPD are quite low as this is the loudspeaker located at 0° in front of the dummy head. These values are much higher for loudspeaker 7 and 19 as they are located at 90° and 270° respectively, which is expected.



(a) ILD and IPD of speaker 1 (0°) over the frequency spectrum



(b) ILD and IPD of speaker 7 (90°) over the frequency spectrum



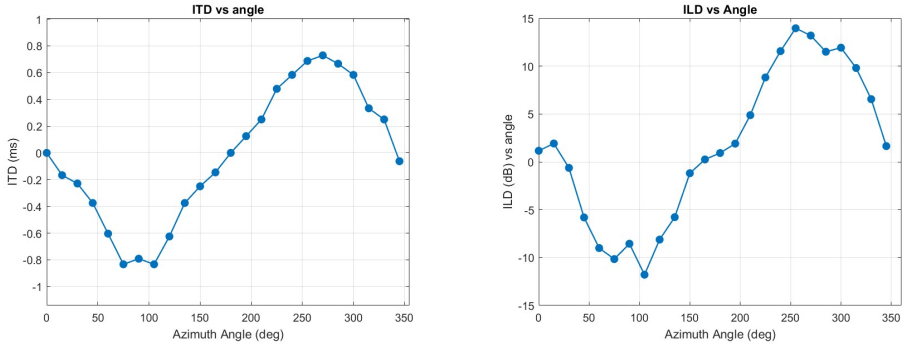
(c) ILD and IPD of speaker 19 (180°) over the frequency spectrum

Figure 3.15: Interaural Time Differences and Interaural Level Differences at speaker 1, 7, and 19

Finally, the binaural cues consisting of the ILD and ITD are plotted as a function of source azimuth in Fig. 3.16b and Fig. 3.16a, respectively. To obtain a single ILD value per direction, the ILD is averaged across frequency. This frequency-averaged ILD provides a compact summary of the overall level imbalance between the ears and is commonly used for analyzing azimuth-dependent trends in binaural cues.

The ITD is estimated in the time domain as the difference in arrival time between the peaks of the left and right ear impulse responses. This approach captures the dominant low-frequency timing cue associated with the direct sound path.

The resulting ILD and ITD curves exhibit the characteristic sinusoidal dependence on azimuth that is well documented in the literature on binaural hearing [16], with values close to zero for frontal sources and increasing magnitude toward lateral directions. Such behavior is consistent with classical models and measurements of binaural cues for human and dummy-head recordings.



(a) (b)

Figure 3.16: Binaural cues: ITD (a) and ILD (b) vs azimuth

4

LOCATION OPTIMIZATION OF VIRTUAL SOURCES

This chapter formulates virtual source selection as a discrete optimization problem. Candidate virtual source subsets are evaluated by constructing a binaural beamformer for each subset and quantifying its binaural cue reconstruction error over all possible interferer directions. The configuration minimizing the total ITF error is selected as optimal.

In many binaural beamforming approaches, such as GBLCMV or JBLCMV methods, the beamformer requires a set of acoustic transfer functions (ATFs) corresponding to the target and interfering sources. In practical scenarios, however, these ATFs are often unavailable or cannot be estimated reliably. A strategy, therefore, is to introduce a set of virtual sources, i.e., pre-defined directions around the head for which ATFs are predetermined. These virtual sources serve as constraints in the beamformer and determine the spatial structure that the beamformer can reproduce.

In [1], the locations of virtual sources are chosen using a simple uniform angular grid. While this choice is convenient, it is not clear whether it leads to optimal performance in terms of binaural cue preservation or HRTF reconstruction accuracy of the interfering sources. In this chapter, we investigate the problem of optimizing the spatial configuration of virtual sources, aiming to find a limited number of directions that best approximate arbitrary HRTFs of interfering sources through BLCMV and JBLCMV beamforming.

Rather than focusing directly on noise reduction performance, the emphasis is placed on the spatial modeling capability of the virtual-source-constrained beamformer, which forms a crucial prerequisite for effective binaural noise reduction in later stages. For this, a grid-search-based location optimization framework is proposed in Section 4.2. Using the new multi-microphone HRTF dataset obtained in Chapter 3, all possible subsets of virtual source directions are evaluated. For each candidate subset, a BLCMV and JBLCMV beamformer is constructed, and its ability to reproduce binaural cues across all directions is quantified using an interaural transfer function (ITF) error metric, introduced by Eq. 2.7. The subset that minimizes this error is selected as the optimal configuration. The new HRTF dataset allows the exploration of how the number of microphones influences the performance of the beamforming and optimization problem.

4.1. METHODOLOGY

The classical GBLCMV binaural beamforming problem is formulated as the minimization of output noise power subject to constraints that preserve the binaural cues of the target source and, where possible, selected interfering sources. These constraints require knowledge of the ATFs associated with the constrained source directions. In practice, while the target source direction is often known or can be reliably estimated, the ATFs of interfering sources are generally unknown or subject to large estimation errors [1, 2].

To address this limitation, the true interferer constraints in binaural beamformers such as the JBLCMV can be replaced with a fixed set of virtual source constraints, corresponding to predetermined directions for which ATFs are assumed to be known. Interferers located away from these directions are not constrained explicitly. Instead, in [1], their binaural cues are approximated through the spatial structure imposed by the virtual sources, which introduces a steering vector mismatch (SVM) that depends on the number and spatial distribution of the virtual sources.

The problem addressed in this chapter is, therefore, not the classical beamforming or noise reduction problem itself, but a preceding configuration problem where the locations of a fixed number of virtual sources are chosen such that the binaural cues of arbitrary interfering sources are approximated as accurately as possible. To this end, the beamformer is designed repeatedly for different candidate virtual source configurations, without evaluating actual noise reduction performance.

Specifically, for each candidate subset of virtual source directions, a JBLCMV beamformer is constructed using the target source and the selected virtual sources as constraints. This beamformer is then applied to all possible interferer directions available in the HRTF dataset. For each interferer direction, the resulting ITF is compared to the reference ITF obtained from the corresponding HRTF. The deviation between the two is quantified using an ITF error metric, which serves as a measure of binaural cue reconstruction accuracy.

By collecting this error across all interferer directions, each virtual source configuration is assigned a single performance score. The optimal configuration is defined as the subset of virtual source locations that minimizes this global ITF reconstruction error. Only after this optimal configuration has been determined can it be used meaningfully in a conventional binaural beamforming and noise reduction framework, where metrics such as noise suppression, intelligibility improvement, and cue preservation can be evaluated.

The problem considered in this chapter can therefore be viewed as a preparatory design step. The focus is not on reducing noise itself, but on determining which virtual source locations allow a binaural beamformer to best approximate the binaural cues of arbitrary interfering sources. Once this optimal configuration has been identified, it can be fixed and used in a standard binaural beamforming framework, where noise reduction and intelligibility improvements can be evaluated under realistic acoustic conditions.

4.2. GRID SEARCH METHOD

The location optimization problem is formulated as a discrete combinatorial search over a finite set of candidate virtual source directions. Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of measured sound source directions available in the HRTF dataset, and let $n_0 \in \mathcal{N}$ denote the target source direction, which is usually located at 0° in front of the user. All remaining directions $\mathcal{N}_0 = \mathcal{N} \setminus \{n_0\}$ are treated as potential virtual source locations.

For a given number of virtual sources m, the optimization problem consists of selecting a subset

$$\mathcal{V} \subset \mathcal{N}_0, \quad |\mathcal{V}| = m, \tag{4.1}$$

such that binaural cue reconstruction error for a given arbitrary interfering source is minimized over all possible interfering source locations when a JBLCMV beamformer is designed using the target source and the virtual sources in $\mathcal{V}$.

Since the search space is finite, an exhaustive grid search is performed by evaluating all $\binom{N-1}{m}$ possible subsets, where $|\mathcal{N}_0| = N - 1$. While this approach is computationally expensive, it provides a globally optimal solution for the chosen error metric and serves as a reference for future heuristic or learning-based optimization methods. The reconstruction error is quantified using an ITF error metric and is evaluated for a single interfering source over all possible spatial directions in $\mathcal{N}_0$. In this way, each candidate virtual source configuration is assessed based on its ability to preserve binaural cues for arbitrary interfering source locations.

4.2.1. JBLCMV BEAMFORMER

For each candidate virtual source configuration, a JBLCMV beamformer $\hat{\mathbf{w}}$ is designed independently for each frequency bin. This beamformer, defined by [1], was already shown in Section 2.2.2 but is repeated here for completeness. The beamformer minimizes the output noise power subject to binaural cue preservation constraints of the virtual sources and is obtained by solving

$$\begin{aligned} \hat{\mathbf{w}}_{JBLCMV} = \quad & \underset{\mathbf{w} \in \mathbb{C}^{2M \times 1}}{\arg\min} \quad \mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w}, \\ & \text{s.t.} \qquad \mathbf{w}^H \Lambda = \mathbf{f}^H, \end{aligned} \tag{4.2}$$

where the constraint matrix Λ and constraint vector $\mathbf{f}$ are defined in Eq. 4.3 and Eq. 4.5, respectively. The constraint matrix is given by

$$\Lambda = \begin{bmatrix} \Lambda_1 & | & \Lambda_2 \end{bmatrix}, \tag{4.3}$$

$$= \begin{bmatrix} \bar{\mathbf{a}}_L & \mathbf{0} & | & \bar{\mathbf{q}}_{1,L} & \cdots & \bar{\mathbf{q}}_{m,L} \\ \mathbf{0} & \bar{\mathbf{a}}_R & | & -\bar{\mathbf{q}}_{1,R} & \cdots & -\bar{\mathbf{q}}_{m,R} \end{bmatrix} \in \mathbb{C}^{2M \times (d=m+2)}, \tag{4.4}$$

where Λ consists of two parts, where Λ_1 enforces distortionless reproduction of the target source at both ears, and Λ_2 enforces binaural cue preservation for the selected virtual sources. The corresponding constraint vector is given by

$$\mathbf{f}^H = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & | & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{1\times(d=m+2)}. \tag{4.5}$$

The target constraints are constructed using normalized acoustic transfer functions with respect to the left and right reference microphones, given by

$$\bar{\mathbf{a}}_L = \frac{\mathbf{a}}{a_L}, \quad \bar{\mathbf{a}}_R = \frac{\mathbf{a}}{a_R} \quad \in \mathbb{C}^{M\times 1}, \tag{4.6}$$

where $\mathbf{a}$ contains the transfer functions from the target source to all microphones, and a_L and a_R denote the transfer functions to the left and right reference microphones, respectively. For each virtual source $m \in \mathcal{V}$, normalized interferer steering vectors are constructed analogously as

$$\bar{\mathbf{q}}_{m,L} = \frac{\mathbf{q}_m}{q_{m,L}}, \quad \bar{\mathbf{q}}_{m,R} = \frac{\mathbf{q}_m}{q_{m,R}} \quad \in \mathbb{C}^{M\times 1}, \tag{4.7}$$

and are incorporated into Λ_2 with opposite signs to enforce ITF preservation. The total number of constraints is therefore $d = 2 + m$. To ensure feasibility of the JBLCMV solution, the condition $d \leq 2M$ must be satisfied, where M is the number of microphones per ear. Only configurations satisfying this condition are considered during optimization. The JBLCMV beamformer admits the closed-form solution

$$\hat{\mathbf{w}}_{JBLCMV} = \begin{cases} \tilde{\mathbf{P}}^{-1}\Lambda(\Lambda^H\tilde{\mathbf{P}}^{-1}\Lambda)^{-1}\mathbf{f}, & \text{if} \quad d < 2M, \\ (\Lambda^H)^{-1}\mathbf{f}, & \text{if} \quad d = 2M, \end{cases} \tag{4.8}$$

where the estimation of $\tilde{\mathbf{P}}$ is described in the next section, and the resulting beamformer is applied in the ITF error calculation with respect to the interfering source.

CPSDM ESTIMATION

The JBLCMV beamformer requires an estimate of the cross-power spectral density matrix (CPSDM) describing the spatial second-order statistics of the noise and interfering sources. In a real acoustic scenario, this matrix must be estimated from the observed microphone signals and therefore depends on the actual positions of the interfering sources present in the environment. Consequently, the CPSDM cannot be freely designed as part of the beamformer optimization itself, but must instead be treated as part of the simulation setup.

In the virtual source optimization framework considered in this chapter, the CPSDM is therefore not associated with the virtual source locations used to define the beamformer constraints. These virtual sources exist only conceptually within the beamformer design and do not correspond to physical sound sources in a given acoustic scene. Instead, the CPSDM represents the statistics of a single real interfering source at a specific spatial location, which is varied across the set of available source directions in the HRTF dataset.

To this end, a separate CPSDM is constructed for each possible interfering source direction available in the dataset. For every candidate virtual source configuration, the beamformer is thus evaluated repeatedly, once for each possible interferer location. In the experimental setup considered here, this results in 23 CPSDMs per virtual source configuration, corresponding to the 23 non-target source directions in the horizontal plane.

The CPSDMs are derived from the raw time-domain microphone recordings corresponding to individual source directions. For each direction $n \in \mathcal{N}_0$, the multichannel microphone signal $\mathbf{x}_n(t) \in \mathbb{R}^{M\times 1}$ is segmented into overlapping frames using a Hann window and transformed to the frequency domain using a short-time Fourier transform (STFT). The FFT length is chosen to be identical to that used for the HRTF computation, ensuring alignment of the frequency bins.

Let $\mathbf{X}_n(f,t) \in \mathbb{C}^{M\times 1}$ denote the STFT vector of the microphone signals for source direction n at frequency bin f and time frame t. The CPSDM associated with an interfering source at direction n is then estimated as

$$\mathbf{P}_n(f) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_n(f,t)\mathbf{X}_n^H(f,t), \tag{4.9}$$

where T denotes the number of time frames and $(\cdot)^H$ denotes the Hermitian transpose. For notational simplicity, the frequency dependence will be omitted where possible and $\mathbf{P}_n(f)$ will be written as $\mathbf{P}_n$.

During the virtual source location optimization, the JBLCMV beamformer is designed using the target source and the selected virtual sources as constraints, while the CPSDM used in the objective function corresponds to a single interfering source realization. For an interferer located at direction $n_i \in \mathcal{N}_0$, the CPSDM employed in the beamformer design is given by

$$\mathbf{P} = \mathbf{P}_{n_i} + \mathbf{P}_v, \tag{4.10}$$

where $\mathbf{P}_v$ denotes additive spatially white noise. This construction reflects the assumption that the interferer and the background noise are uncorrelated. No aggregation over multiple interferer locations is performed within a single beamformer design. Instead, the beamformer is redesigned for each interferer direction individually.

The individual CPSDM components are scaled such that the signal-to-noise ratio (SNR) of the target source equals 40 dB and the signal-to-interference ratio (SIR) between the target source and the active interferer equals 20 dB. These values define the relative weighting between noise and interference in the CPSDM but do not affect the beamformer constraints or the ITF-based optimization criterion. Since all candidate virtual source configurations are evaluated under identical SNR and SIR assumptions for each interferer location, the resulting optimization reflects only differences in spatial modeling capability.

Finally, the binaural CPSDM used in the JBLCMV formulation is constructed as

$$\tilde{\mathbf{P}} = \begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \end{bmatrix}, \tag{4.11}$$

which is used to design a separate JBLCMV beamformer for each interfering source direction. The performance metrics reported in the subsequent sections are obtained by evaluating all candidate virtual source configurations across the full set of interferer locations, ensuring a consistent and fair comparison.

ITF ERROR METRIC

To quantify how accurately a given virtual source configuration preserves binaural cues for arbitrary interfering source directions, an interaural transfer function (ITF) error metric is employed. This choice follows directly from the binaural beamforming framework introduced by Koutrouvelis [1], in which the preservation of binaural perception cues, rather than exact signal reconstruction, is the primary design objective. Minimizing this error corresponds to preserving both interaural level and phase differences, which are the dominant cues for horizontal sound localization.

The choice of an error metric is critical, as it determines which aspects of the binaural signal are considered perceptually relevant. In the context of binaural beamforming, an appropriate metric should (i) directly reflect interaural cue preservation, (ii) be invariant to overall gain scaling, and (iii) allow meaningful comparison across source directions and frequencies. The ITF error metric satisfies these criteria by explicitly quantifying deviations in the interaural transfer ratio produced by the beamformer.

The ITF represents the ratio between the acoustic transfer functions at the left and right ears and therefore encodes the relative interaural information that underlies binaural localization, most notably interaural time differences (ITDs) and interaural level differences (ILDs) as mentioned in Section 2.2.2. Crucially, these perceptually relevant cues are invariant to absolute signal scaling and are preserved as long as the left-right transfer ratio is maintained. As a result, ITF-based measures directly assess binaural cue preservation, whereas absolute error metrics on the left or right output signals would induce cue distortion with overall gain changes.

Alternative error metrics could be considered, such as the mean-squared error between the left and right output signals, magnitude errors of individual HRTFs, or spatial correlation-based measures. However, these metrics are generally sensitive to absolute amplitude scaling and do not explicitly isolate binaural cue distortions. As a result, they may indicate large errors even when interaural level and timing cues are perceptually preserved.

In contrast, binaural beamformers such as the JBLCMV are explicitly designed to preserve interaural relationships rather than exact signal waveforms. Since the ITF captures the relative transfer between the ears, it directly reflects deviations in interaural level and phase differences, which are the dominant cues for horizontal-plane localization. By being invariant to overall gain changes, the ITF error avoids ambiguities inherent to

absolute error metrics and provides a physically meaningful measure of binaural cue preservation.

For a given interfering source located at direction $n_i \in \mathcal{N}_0$, the reference ITF is defined using the HRTF measured at the left and right reference microphones as

$$ITF_{n_i}^{\text{ref}} = \frac{b_{i,L}}{b_{i,R}}, \tag{4.12}$$

where $b_{i,L}$ and $b_{i,R}$ denote the acoustic transfer functions from source direction n_i to the left and right reference microphones, respectively. After designing the JBLCMV beamformer for a given virtual source configuration, the beamformer is applied to the same interfering source, yielding the output ITF

$$ITF_{n_i}^{\text{out}} = \frac{\hat{\mathbf{w}}_R^H \mathbf{b}_i}{\hat{\mathbf{w}}_L^H \mathbf{b}_i}, \tag{4.13}$$

where $\mathbf{b}_i \in \mathbb{C}^{M\times 1}$ denotes the vector of acoustic transfer functions from source direction n_i to the M microphones, and $\hat{\mathbf{w}}_L$ and $\hat{\mathbf{w}}_R$ denote the left and right beamformer of the JBLCMV method, respectively. The ITF reconstruction error for source direction n_i is then defined as the absolute deviation between the reference and beamformer-produced ITFs given in Eq. 4.14.

$$\varepsilon_{n_i} = |ITF_{n_i}^{out} - ITF_{n_i}^{ref}| \tag{4.14}$$

$$= \left| \frac{\hat{\mathbf{w}}_L^H \mathbf{b}_i}{\hat{\mathbf{w}}_R^H \mathbf{b}_i} - \frac{b_{i,L}}{b_{i,R}} \right| \quad \text{for} \quad i = 1, \ldots, r \tag{4.15}$$

where F denotes the number of frequency bins considered in the evaluation. This formulation explicitly measures deviations in the interaural ratio induced by the beamformer and therefore directly reflects binaural cue distortion. The error is evaluated per frequency bin and subsequently averaged across frequency to obtain a scalar error measure for each interfering source direction. To assess the overall spatial modeling capability of a virtual source configuration, the ITF error is further averaged across all possible interfering source directions,

$$\varepsilon_{\text{tot}} = \frac{1}{|\mathcal{N}_0| F} \sum_{n_i \in \mathcal{N}_0} \sum_{f=1}^{F} \varepsilon_{n_i}(f), \tag{4.16}$$

This total ITF error serves as the optimization criterion in the grid search procedure described in Section 4.2. By evaluating the error across all possible interfering source locations, the metric explicitly assesses how well a given virtual source configuration generalizes to arbitrary interferer directions. As such, the ITF error metric provides a direct and physically meaningful measure of the spatial modeling capability of the virtual-source-constrained beamformer. Therefore, the virtual source configuration with the lowest possible total error can be chosen as the optimal configuration.

4.3. EXPERIMENTAL IMPLEMENTATION

This section describes the experimental implementation of the virtual source location optimization framework introduced in this chapter. The complete procedure is summarized in Algorithm 1, which formalizes the grid-search-based evaluation of candidate virtual source configurations. The implementation follows a clear separation between beamformer design and performance evaluation: for each candidate subset of virtual source directions, JBLCMV beamformers are constructed using the target source and the selected virtual sources as constraints, while the interference statistics are varied according to the assumed location of a single real interfering source. The binaural cue preservation performance of each configuration is then evaluated exhaustively over all possible interfering source directions available in the HRTF dataset from Chapter 3. By repeating this procedure for all admissible subsets, each virtual source configuration is assigned a single performance score based on the averaged ITF reconstruction error, and the configuration yielding the minimum total error is selected as optimal. The algorithm is implemented entirely using measured acoustic transfer functions and precomputed second-order statistics, ensuring a controlled and reproducible evaluation that isolates the spatial modeling capability of the virtual-source-constrained beamformer.

The dataset contains $N = 24$ measured loudspeaker directions distributed uniformly in azimuth around the dummy head. For each direction, acoustic transfer functions were measured to $M = 12$ microphones in total, i.e., 6 microphones per ear. Throughout this chapter, the measured HRTFs are stored in a three-dimensional array

$$\mathbf{H} \in \mathbb{C}^{M \times N \times F}, \tag{4.17}$$

where $F = 513$ is the number of frequency bins. The target source direction n_0 is fixed to the frontal loudspeaker (typically 0°), corresponding to the dataset index $n_0 \in \mathcal{N}$. All remaining directions form the candidate set $\mathcal{N}_0 = \mathcal{N} \setminus \{n_0\}$, which is used both for virtual source selection and for the set of possible interfering source locations.

Two reference microphones are selected to define the binaural output and to compute reference ITFs. In the experiments, the left and right reference microphones are chosen as a fixed microphone index in the left-ear set and right-ear set, respectively (e.g., microphone 6 for the left ear and microphone 12 for the right ear in the global indexing of $\mathbf{H}$). These reference channels are used in Eq. 4.12 and Eq. 4.13 to compute ITF^{in} and ITF^{out}.

For each candidate subset $\mathcal{V} \subset \mathcal{N}_0$ of virtual source directions, the corresponding constraint matrix $\Lambda(\mathcal{V})$ and constraint vector $\mathbf{f}(\mathcal{V})$ are constructed directly from the measured acoustic transfer functions contained in the HRTF matrix $\mathbf{H}$ defined in Eq. 4.3 and Eq. 4.5. The target source constraints are formed using the normalized steering vectors $\overline{\mathbf{a}}_L$ and $\overline{\mathbf{a}}_R$ as defined in Eq. 4.6, while the virtual source constraints are obtained analogously from the normalized virtual source steering vectors $\overline{\mathbf{q}}_{m,L}$ and $\overline{\mathbf{q}}_{m,R}$ given in Eq. 4.7 for all $m \in \mathcal{V}$.

Algorithm 1 Grid-search-based virtual source location optimization

Input:
- $\mathbf{H}$: measured HRTF matrix containing acoustic transfer functions
- $\mathbf{P}_n$: CPSDMs of all interfering source directions
- n_0: index of the target source direction
- M: number of microphones
- m: number of virtual source constraints

Output:
- $\mathcal{V}^\star$: optimal subset of virtual source directions

Define:
- $\mathcal{N}$: set of all measured source directions
- $\mathcal{N}_0 = \mathcal{N} \setminus \{n_0\}$: candidate virtual and interfering source directions
- $\mathbf{P}_v = \sigma_v^2 \mathbf{I}_M$: additive white noise
- $\varepsilon(\mathcal{V})$: total ITF reconstruction error for virtual source set $\mathcal{V}$

Initialize:
- $\varepsilon_{\min} \leftarrow \infty$
- $\mathcal{V}^\star \leftarrow \emptyset$

General comments:
- { **SP** denotes solving the JBLCMV problem in Eq. 4.2. }
- { **CPSDM** is the construction of the CPSDM in Eq. 4.10 and Eq. 4.11. }
- { **Error** denotes the ITF error metric defined in Eq. 4.14. }

for all $\mathcal{V} \subset \mathcal{N}_0$ such that $|\mathcal{V}| = m$ **do**
  Construct $\Lambda(\mathcal{V})$ and $\mathbf{f}(\mathcal{V})$ from $\mathbf{H}(\mathcal{V})$
  for all $n_i \in \mathcal{N}_0$ **do**
    $\tilde{\mathbf{P}} \leftarrow \textbf{CPSDM}(\mathbf{P}_{n_i}, \mathbf{P}_v)$
    $\hat{\mathbf{w}}_{\text{JBLCMV}}^{(n_i)} \leftarrow \textbf{SP}\big(\Lambda(\mathcal{V}), \mathbf{f}(\mathcal{V}), \tilde{\mathbf{P}}\big)$
    $\varepsilon_{n_i}(\mathcal{V}) \leftarrow \textbf{Error}(\hat{\mathbf{w}}_{\text{JBLCMV}}^{(n_i)}, \mathbf{H}(n_i))$
  end for
  $\varepsilon(\mathcal{V}) \leftarrow \frac{1}{|\mathcal{N}_0|} \sum_{n_i \in \mathcal{N}_0} \varepsilon_{n_i}(\mathcal{V})$
  if $\varepsilon(\mathcal{V}) < \varepsilon_{\min}$ **then**
    $\varepsilon_{\min} \leftarrow \varepsilon(\mathcal{V})$
    $\mathcal{V}^\star \leftarrow \mathcal{V}$
  end if
end for
return $\mathcal{V}^\star$

Independently of the selected virtual source configuration, the CPSDM used in the JBLCMV design is constructed to represent the statistics of a single real interfering source at a specific direction $n_i \in \mathcal{N}_0$, together with additive spatially white noise. For each interfering source direction, a distinct binaural CPSDM $\tilde{\mathbf{P}}$ is constructed according to Eq. 4.10 and Eq. 4.11, and a corresponding JBLCMV beamformer is designed. As a result, for every candidate virtual source configuration, multiple beamformers are obtained, one for each possible interferer location, while the virtual source constraints remain fixed. This ensures that performance differences across virtual source configurations are not

influenced by assumptions about a particular interference scenario, but instead reflect robustness across all admissible interferer directions.

Using $\Lambda(\mathcal{V})$, $\mathbf{f}(\mathcal{V})$, and the CPSDM $\tilde{\mathbf{P}}$, a frequency-dependent JBLCMV beamformer is computed for each candidate subset by solving the constrained optimization problem in Eq. 4.2. The closed-form solution given in Eq. 4.8 is applied independently for each frequency bin, yielding binaural beamformer weights $\hat{\mathbf{w}}_L(f)$ and $\hat{\mathbf{w}}_R(f)$ that satisfy ($d = 2 + m$) $\leq 2M$. The resulting beamformer is therefore fully determined by the virtual source configuration.

To evaluate the spatial modeling capability of the resulting beamformer, it is subsequently applied to each possible interfering source direction $n_i \in \mathcal{N}_0$. For a given direction n_i, the beamformer output is used to compute the output interaural transfer function $ITF^{\text{out}}_{n_i}$ according to Eq. 4.13. This output ITF is then compared to the corresponding reference ITF, $ITF^{\text{ref}}_{n_i}$, obtained directly from the measured HRTFs as defined in Eq. 4.12. The deviation between these two quantities is quantified using the ITF error metric ε_{n_i} defined in Eq. 4.14, which directly measures binaural cue distortion induced by the beamformer.

For each candidate virtual source configuration $\mathcal{V}$, the ITF reconstruction error is evaluated across all interfering source directions and subsequently averaged over all interfering source directions and frequencies to obtain the total error measure $\varepsilon_{\text{tot}}(\mathcal{V})$ as introduced in Eq. 4.16. This scalar quantity summarizes how well the beamformer constrained by $\mathcal{V}$ preserves binaural cues for arbitrary interferer locations across the full spatial domain.

By exhaustively repeating this procedure for all admissible subsets $\mathcal{V}$ with $|\mathcal{V}| = m$, the grid search described in Section 4.2 assigns a unique performance score to each virtual source configuration. The optimal configuration

$$\mathcal{V}^\star = \arg \min_{\mathcal{V} \subset \mathcal{N}_0, |\mathcal{V}| = m} \varepsilon_{\text{tot}}(\mathcal{V}) \tag{4.18}$$

is finally selected as the subset that minimizes the total ITF reconstruction error. This optimal set of virtual source locations represents the configuration that best captures the spatial structure of the measured HRTF dataset within the JBLCMV framework and can subsequently be fixed for use in conventional binaural beamforming and noise reduction experiments.

4.4. EXPERIMENTAL RESULTS

This section presents the experimental results of the virtual source location optimization framework introduced in Section 4.2 and implemented in Section 4.3. The objective of these experiments is to assess how both the number and spatial configuration of virtual sources affect the ability of the JBLCMV beamformer in Eq. 4.2 to preserve binaural cues for arbitrary interfering source directions. In contrast to conventional beamforming evaluations, the focus is not on output noise reduction, but on the spatial modeling

capability induced by the constraint set $\mathcal{V}$, as quantified by the ITF reconstruction error metric. This isolates the role of the virtual source configuration itself, independent of noise suppression performance.

Prior to performing the exhaustive grid search, it was anticipated that the optimal virtual source sets $\mathcal{V}^\star$ would exhibit interpretable and potentially generalizable spatial structure. Since the virtual sources in $\mathcal{V}$ form a compact set of binaural cue constraints in the JBLCMV formulation and are optimized with respect to a global error criterion that averages across all directions and frequencies, one might expect the selected directions to act as a structured support of the ITF manifold. In particular, it was hypothesized that optimal configurations could

i. cluster around perceptually salient regions where HRTFs or ITFs vary rapidly, such as lateral directions dominated by head-shadow effects,

ii. display approximate left–right symmetry reflecting the bilateral geometry of the head and measurement setup,

iii. evolve toward a near-uniform angular spacing as the number of virtual sources m increases.

Identifying such patterns would be valuable, as it would suggest that near-optimal virtual source layouts could be predicted or parameterized, reducing the need for exhaustive combinatorial search and offering insight into which spatial regions are most influential for binaural cue reconstruction.

As will be shown in the following subsections, the experimental results only partially confirm the hypotheses stated above. While certain virtual source directions recur across independently optimized configurations, the overall solutions exhibit considerable variability when changing the number of microphones M, the number of virtual sources m, or the specific measurement realization. In many cases, distinct subsets yield nearly identical values of the total ITF error ε_{tot}, indicating that the underlying optimization problem is highly nonconvex and characterized by a large set of near-degenerate solutions. As a result, small perturbations, such as changes in array dimensionality, numerical conditioning of $\tilde{\mathbf{P}}^{-1}$, or measurement noise, can alter the identity of the optimal subset without implying a meaningful change in performance. Within this context, the repeated appearance and occasional clustering of virtual sources near the lateral directions (around $\pm 90^\circ$) is consistent with hypothesis (i) and aligns with the strong spatial HRTF variability observed in Fig. 4.4. Hypothesis (ii), however, is only weakly supported: although some configurations exhibit approximate left–right balance, the optimal solutions are often not uniquely symmetric due to the large number of near-equivalent subsets, as further quantified in Section 4.4.4. Finally, hypothesis (iii) is not clearly reflected in the results, as increasing m does not systematically lead to near-uniform angular spacing, but instead admits multiple structurally different configurations with comparable ITF reconstruction performance. This behavior motivates a deeper analysis of the error distribution across all configurations, which is addressed explicitly in Section 4.4.4 and

provides important context for interpreting the absence of a single, clearly structured optimal solution.

4.4.1. OPTIMAL VIRTUAL SOURCE CONFIGURATIONS

This subsection presents the optimal virtual source configurations $\mathcal{V}^\star$ obtained from the exhaustive grid search for different values of the number of virtual sources m and the number of microphones M. For each pair (M, m), the optimal configuration is determined independently by minimizing the total ITF reconstruction error $\varepsilon_{\text{tot}}(\mathcal{V})$ over all admissible subsets $\mathcal{V} \subset \mathcal{N}_0$ with $|\mathcal{V}| = m$, as defined in Section 4.2.

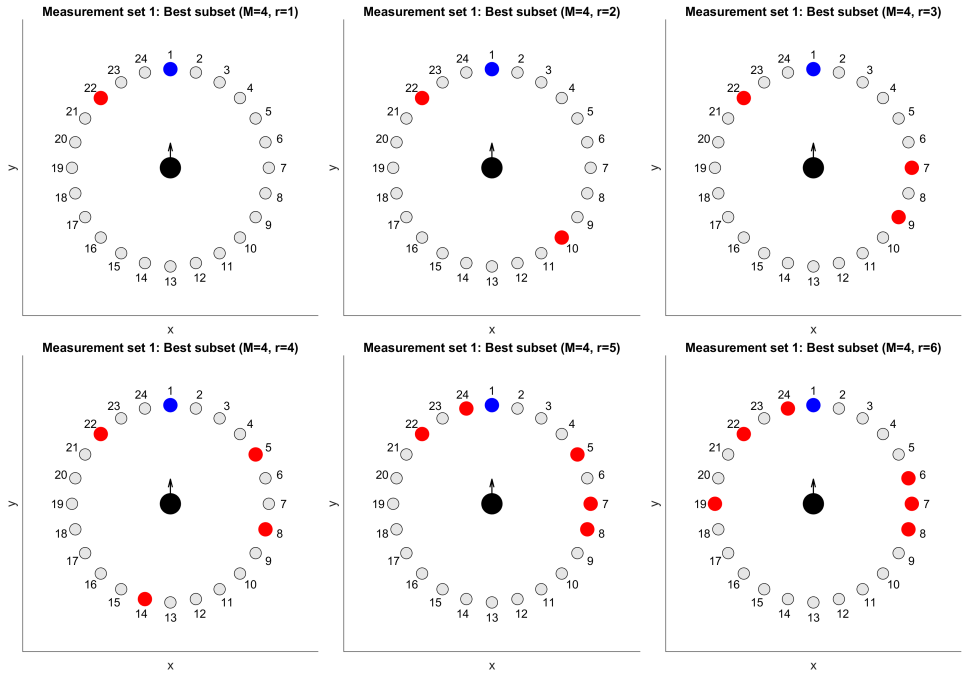


Figure 4.1: Optimal virtual source configurations for varying number of virtual sources $m = [1,2,3,4,5,6]$ and fixed number of microphones $M = 4$ with the target source shown in blue and the selected virtual sources shown in red.

Fig. 4.1 shows the optimal virtual source configurations for a fixed number of microphones ($M = 4$) and varying number of virtual sources $m = [1,\ldots,6]$. Each panel displays the target source direction together with the corresponding optimal set $\mathcal{V}^\star$ for the given value of m. While the number of selected directions increases with m by definition, the spatial arrangement of the selected directions does not exhibit an immediately obvious or uniform pattern. In particular, the selected virtual sources are not evenly distributed over azimuth, nor do they follow a simple symmetric or regularly spaced configuration.

At the same time, a closer inspection of Fig. 4.1 reveals that certain directions recur across multiple independently optimized configurations. For example, specific direc-

tions such as indices 22 and 7 appear in several optimal subsets for different values of r, whereas other directions are selected less consistently or only for specific configurations. It can also be deduced from the plots that on there are often clusters being formed around the 90° point corresponding to position index 7. This observation suggests that some candidate directions contribute more robustly to reducing the global ITF reconstruction error than others under the considered optimization criterion, while a range of alternative directions can be exchanged without substantially affecting performance.

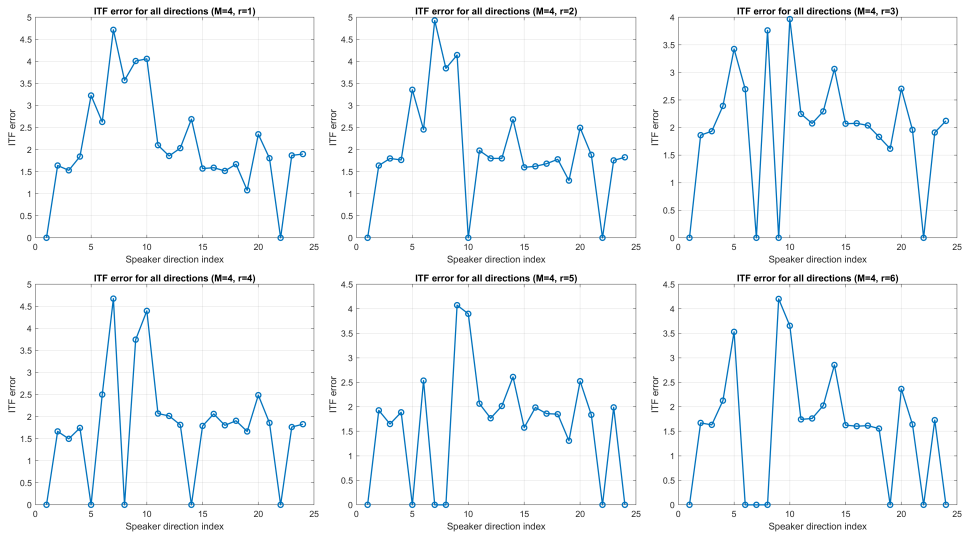


Figure 4.2: Total ITF error over the possible interfering source directions for varying number of virtual sources $m = [1,2,3,4,5,6]$ and fixed number of microphones $M = 4$

The corresponding ITF error profiles across all possible interfering source directions are shown in Fig. 4.2. As expected, the ITF error is zero at the target direction and at the selected virtual source directions, reflecting that the corresponding JBLCMV constraints are satisfied exactly. Beyond these constrained directions, the ITF error varies across the azimuth without exhibiting a clear or consistent spatial pattern that directly mirrors the locations of the virtual sources. This indicates that there is no direct correlation between the distance of the interfering source location and the virtual source positions and that the influence of the virtual source constraints on the ITF reconstruction error is distributed globally across directions, rather than being localized around the selected constraint positions.

The influence of the number of microphones on the optimal configurations is illustrated in Fig. 4.3, which shows the optimal virtual source sets for a fixed number of virtual sources ($m = 4$) and increasing microphone counts $M = [4,6,8,10,12]$. Although all configurations contain the same number of virtual sources, the selected directions differ across values of M. Similar to before, there are certain directions that keep recurring.

This indicates that the identity of the optimal virtual source set depends on the array dimensionality, even when the optimization criterion and candidate direction set remain unchanged.

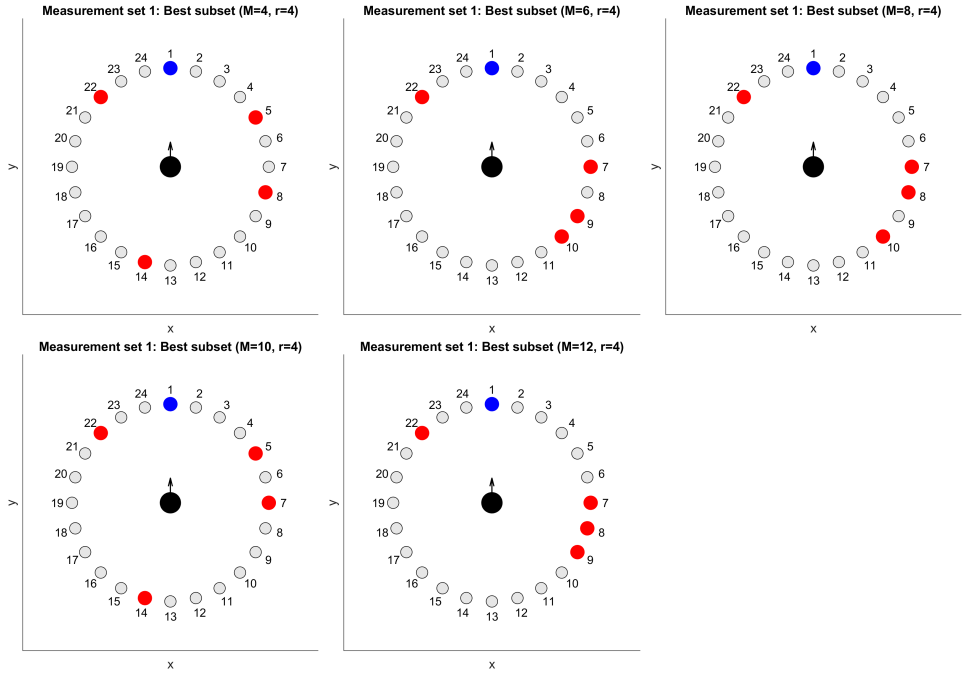


Figure 4.3: Optimal virtual source configurations for varying number of microphones $M = [4, 6, 8, 10, 12]$ and fixed number virtual sources $m = 4$ with the target source shown in blue and the selected virtual sources shown in red.

Finally, Fig. 4.4 provides a visualization of the spatial variability of the measured HRTFs by showing the average difference between neighboring source directions, averaged over all frequencies. For two neighboring directions n and $n+1$, the HRTF difference is defined as

$$\Delta H(n) = \frac{1}{F} \sum_{f=1}^{F} \left\| \mathbf{H}(f, n+1) - \mathbf{H}(f, n) \right\|_2, \tag{4.19}$$

where $\mathbf{H}(f, n) \in \mathbb{C}^{M \times 1}$ denotes the HRTF vector at frequency bin f for direction n, and $\| \cdot \|_2$ denotes the Euclidean norm across microphones. This measure captures how rapidly the acoustic transfer functions change as a function of source direction.

The figure shows that the HRTF differences are substantially larger at lateral positions than at frontal or rear directions. In particular, pronounced peaks are observed around $\pm 90°$, corresponding to directions where head-shadow effects induce strong interaural

level and phase variations. These regions of high spatial HRTF variability provide useful context for interpreting the virtual source configurations shown earlier. Specifically, the repeated selection and clustering of virtual source directions near these lateral angles may be related to the fact that small angular deviations in these regions lead to relatively large changes in binaural cues.

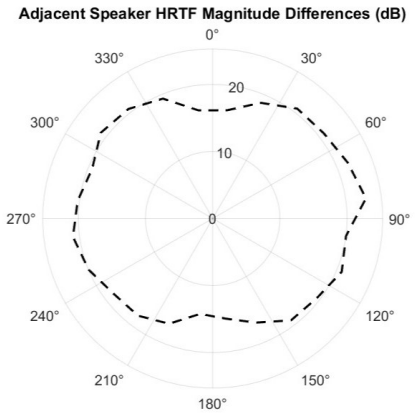


Figure 4.4: HRTF differences averaged over all frequencies

At this stage, this observation is intended as contextual rather than causal. The optimization procedure does not explicitly incorporate any measure of local HRTF variability, and the relationship between HRTF differences and virtual source selection is therefore not enforced but emerges implicitly through the global ITF-based error criterion. A more detailed analysis of this relationship is deferred to later sections.

4.4.2. EFFECT OF THE NUMBER OF MICROPHONES

The influence of the number of microphones on the achievable ITF reconstruction performance is summarized in Fig. 4.5, which shows the minimum total ITF error obtained for each microphone count. Overall, the curve exhibits a clear decreasing trend: as more microphones are included in the JBLCMV design, the best achievable value of ε_{tot} decreases. This behavior is consistent with the constrained optimization in Eq. 4.2, since a larger set of microphone signals increases the number of spatial degrees of freedom available to satisfy the binaural constraints while minimizing the quadratic form $\mathbf{w}^H \tilde{\mathbf{P}} \mathbf{w}$.

A key observation, however, is that the improvement saturates beyond approximately $M = 6$ microphones. While increasing M from smaller values yields a noticeable reduction in $\varepsilon_{\text{tot}}(\mathcal{V})$, further increases above 6 provide only marginal gains. In other words, the curve flattens for $M \geq 6$, indicating diminishing returns in ITF reconstruction accuracy as additional microphones are added. This stagnation suggests that, beyond this point, the remaining error is dominated less by a lack of spatial degrees of freedom and more by limitations of the virtual-source constraint model itself.

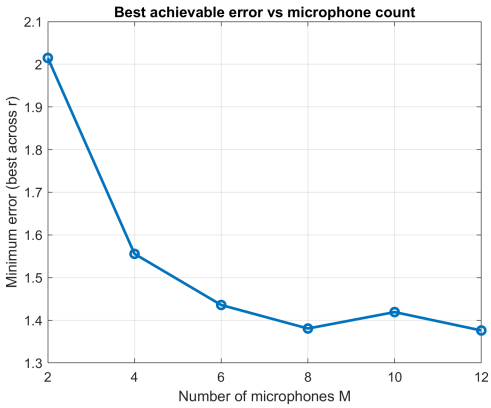


Figure 4.5: Total ITF error as a function of the number of microphones

From a practical perspective, this result is important: within the considered dataset, SNR/SIR assumptions, and ITF-based optimization criterion, using more than approximately six microphones does not substantially improve the spatial modeling capability of the virtual-source-constrained JBLCMV beamformer. This indicates that a microphone count around $M = 6$ can already achieve near-optimal performance for the proposed framework, which is encouraging for real-world implementations where array size, device form factor, power consumption, and calibration effort place strict constraints on the number of microphones that can be deployed. Nevertheless, increasing the number of microphones still increases the available degrees of freedom in the beamformer design, which can be exploited either to accommodate a larger number of constraints (e.g., more virtual sources) or to achieve stronger noise suppression while maintaining the binaural cue preservation constraints. In that sense, additional microphones may yield benefits that are not fully reflected by the ITF reconstruction error metric alone.

4.4.3. EFFECT OF THE NUMBER OF VIRTUAL SOURCES

The effect of the number of virtual sources on the achievable ITF reconstruction performance is summarized in Fig. 4.6, which shows the minimum total ITF error obtained for each value of m for several fixed microphone counts M. Across all considered microphone counts, the curves exhibit a clear and approximately linear decreasing trend: increasing the number of active virtual sources consistently reduces the best achievable value of ε_{tot}. This behavior is expected, since enlarging the constraint set $\mathcal{V}$ increases the number of source directions for which binaural cue preservation is enforced. In particular, the JBLCMV solution in Eq. 4.8 is computed under the constraint

$$\mathbf{w}^H \Lambda(\mathcal{V}) = \mathbf{f}^H, \tag{4.20}$$

so that adding virtual sources (i.e., increasing $|\mathcal{V}| = m$) directly increases the number of ITF-preservation constraints that must be satisfied. As a result, fewer interferer directions remain fully unconstrained, and the resulting beamformer tends to approximate

the reference ITFs more accurately when averaged over all directions and frequencies as in Eq. 4.16.

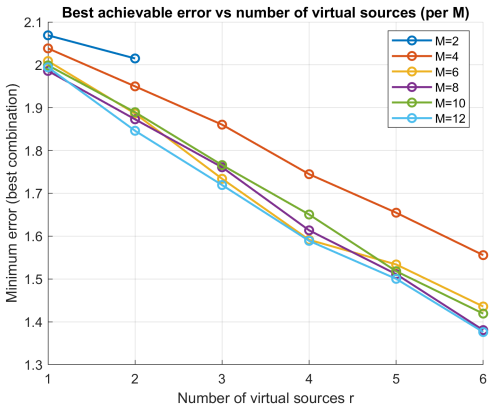


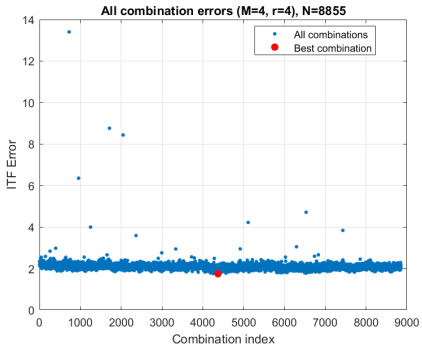
Figure 4.6: Total ITF error as a function of the number of active virtual sources

The observed monotonic decrease is also consistent with the hypothesis made by [1]: when more virtual sources are used, the steering-vector mismatch for arbitrary interferer locations is reduced, because the constrained directions provide a denser set of anchor points that shapes the spatial behavior of the beamformer. In other words, increasing m increases the spatial resolution of the virtual-source constraint model, which manifests in a lower global ITF reconstruction error. The near-linear character of the decrease in Fig. 4.6 suggests that, within the tested range of m, each additional virtual source contributes a roughly comparable reduction in the averaged error measure, rather than producing only isolated improvements for a small subset of interferer directions.

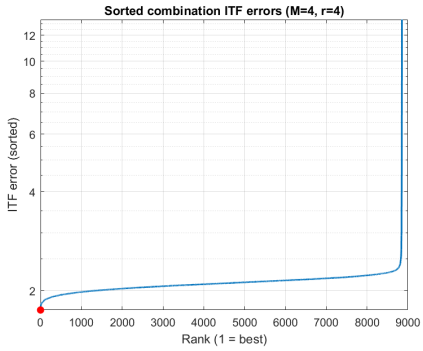
Finally, note that increasing m also increases the total number of constraints $d = 2 + m$ and therefore consumes more degrees of freedom in the beamformer design for a fixed microphone count M. This introduces an inherent trade-off: while larger m improves binaural cue modeling (lower ε_{tot}), it also leaves fewer degrees of freedom available for other objectives such as noise reduction in a full beamforming evaluation. In the present experiments, however, the metric in Eq. 4.16 isolates the cue-reconstruction aspect, and the results in Fig. 4.6 confirm that, under this criterion, using more virtual sources systematically improves ITF reconstruction performance.

4.4.4. CONFIGURATION ERRORS

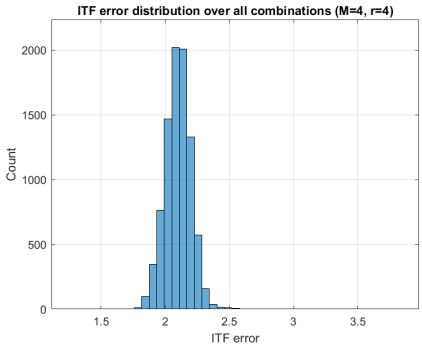
While the previous subsections focus on the best-performing virtual source configurations, this subsection examines the distribution of ITF reconstruction errors across all possible configurations for a representative case with $M = 4$ microphones and $m = 4$ virtual sources. The purpose of this analysis is to characterize the structure of the optimization landscape and to assess how distinctive the optimal configuration is relative to the remaining candidates.



(a) ITF error for all possible subsets $\mathcal{V}$



(b) ITF error sorted from smallest to largest for all possible subsets of $\mathcal{V}$



(c) ITF error distribution of all possible subsets $\mathcal{V}$

Figure 4.7: ITF error analysis with $M = 4$ and $m = 4$

Fig. 4.7a shows the total ITF reconstruction error $\varepsilon_{\text{tot}}(\mathcal{V})$ for all admissible subsets $\mathcal{V} \subset \mathcal{N}_0$ with $|\mathcal{V}| = m$. A first observation is that the majority of configurations yield error values within a relatively narrow range. Apart from a limited number of configurations with clearly elevated error, most subsets perform comparably in terms of the averaged ITF metric defined in Eq. 4.16. This already suggests that the optimization problem does not exhibit a single, sharply isolated optimum, but rather a broad set of configurations with similar performance.

This observation becomes more explicit in Fig. 4.7b, where the same error values are sorted in ascending order. The curve exhibits a steep decrease for the best-performing configurations, followed by an extended, gently sloping region in which a large number of configurations achieve near-identical error values. Only at the upper end of the curve does the error increase sharply again, corresponding to a small set of poorly performing outlier configurations. The relatively flat region indicates that many distinct virtual source subsets are effectively interchangeable with respect to the ITF reconstruction criterion, since small changes in $\mathcal{V}$ lead to only marginal differences in ε_{tot}.

The histogram shown in Fig. 4.7c provides a complementary view of the same phenomenon. The error distribution is strongly concentrated around a central value, with a toward higher errors. This confirms that most virtual source configurations cluster around similar performance levels, while only a few configurations lead to substantially degraded ITF reconstruction.

Taken together, these results highlight that the virtual source location optimization problem is characterized by a large set of near-optimal solutions. As a consequence, the identity of the optimal configuration $\mathcal{V}^\star$ can be sensitive to changes in experimental conditions, such as the number of microphones M, the assumed SNR/SIR levels, numerical conditioning of $\tilde{\mathbf{P}}^{-1}$, or measurement variability in the HRTFs, even when the resulting performance differences remain negligible. This explains why the optimal configurations observed in previous subsections vary across parameter settings.

4.4.5. LOCALITY OF ACOUSTIC TRANSFER FUNCTIONS

An additional perspective on the virtual-source framework can be obtained by viewing acoustic transfer functions as samples of an underlying low-dimensional acoustic manifold. In [35], this concept is studied for relative transfer functions (RTFs), defined as ratios of acoustic transfer functions (ATFs). These ratios are closely related to the interaural transfer functions (ITFs) used in this thesis, as both describe relative information rather than absolute signal levels.

Although RTFs and ITFs are represented as high-dimensional, frequency-dependent feature vectors, their dominant latent variable in a fixed environment is the source position. As a result, acoustic responses lie on a nonlinear manifold of much lower intrinsic dimension [35]. A central question addressed in [35] is whether distances between such feature vectors reflect physical angular proximity. To this end, several distance measures are analyzed, including Euclidean distance, PCA-based distance, and diffusion distance.

A key result in [35] is the strong locality of linear distance measures. Euclidean distances between RTFs were found to vary monotonically with source angle only within a narrow neighborhood of approximately $\Delta \approx 4^\circ$, while PCA-based distances extended this range to about $\Delta \approx 6^\circ$. Beyond these ranges, distances no longer reliably encode angular proximity, indicating that the acoustic manifold is only locally linear and becomes globally nonlinear. In contrast, diffusion distances preserve a meaningful relationship over much larger angular ranges by explicitly exploiting manifold structure.

This observation directly impacts the virtual-source-based JBLCMV framework studied in this chapter. In this framework, binaural cue preservation is enforced exactly at a finite set of virtual source directions $\mathcal{V}$. For unconstrained interferer directions, the method implicitly assumes that ITFs remain close to the reference ITFs. However, the ITF reconstruction error used in this work is a Euclidean distance between complex-valued ITFs, and therefore inherits the same locality limitations observed for Euclidean RTF distances in [35]. As a consequence, a virtual source can only provide a meaningful approximation

for interferers located within a few degrees of its position. For interferers farther away, the ITF error becomes largely insensitive to angular distance and is dominated by the global nonlinear structure of the acoustic manifold.

This fundamentally limits the original motivation of the virtual-source concept introduced by [1] under a Euclidean ITF error criterion, which aims to approximate the binaural cues of arbitrary interfering sources using a small set of predetermined HRTFs. To reliably approximate arbitrary interferer locations under a Euclidean ITF error metric, virtual sources would need to densely cover the azimuth with a spacing below approximately 4°. Over a full 360° range, this would require more than 90 virtual sources. Since the number of JBLCMV constraints grows linearly with the number of virtual sources, such a configuration would require a comparable number of microphones to satisfy $d = 2 + m \leq 2M$, which is infeasible in practice.

The limitations imposed by this locality are further strengthened by the 15° angular resolution of the HRTF dataset used in this chapter. This resolution is substantially coarser than the angular scale at which Euclidean ITF distances are meaningful. As a result, even neighboring measured directions already lie outside the locality regime, preventing the optimization from capturing local manifold behavior. This provides a principled explanation for the flat error landscape and large set of near-degenerate solutions observed in Section 4.4.4, as well as the sensitivity of the optimal configuration $\mathcal{V}^\star$ to changes in parameters despite negligible performance differences.

In summary, the distance-manifold analysis indicates that the combination of the virtual-source framework with a Euclidean ITF error metric constitutes an inherently local approximation strategy. While ITF constraints enforce correct interaural ratios at the selected directions, reliable generalization to arbitrary interferer locations would only be achievable with unrealistically dense virtual source spacing and correspondingly large microphone arrays. This insight clarifies the observed flat error landscape and the presence of many near-degenerate solutions under the chosen metric.

Importantly, however, this conclusion is tied to the specific choice of error measure. The auditory system does not perceive complex-valued ITFs through a Euclidean distance in the full frequency domain, but rather through perceptual cues such as band-dependent ITDs and ILDs, with frequency-dependent sensitivity and reduced reliability in certain regions. It therefore remains possible that virtual-source-based constraint sets could yield meaningful perceptual cue preservation when evaluated using a more perceptually grounded metric, for example by using band-limited and cue-weighted deviations in ILD/IPD (or ITD) rather than a broadband Euclidean ITF distance. Exploring such perceptually motivated metrics is outside the scope of the present chapter, but it provides a clear direction for future work and may reconcile the virtual-source concept with perceptual performance even when Euclidean ITF distances exhibit strong locality.

5

Conclusion and Future Work

This thesis investigated the problem of virtual source location optimization for binaural beamforming, motivated by the constraints and uncertainties inherent to practical hearing assistive devices and related spatial audio systems. As discussed in Chapter 1 and Chapter 2, such systems must operate with a limited number of microphones, under unknown and time-varying acoustic conditions, while preserving binaural cues that are essential for spatial perception, such as interaural level and phase differences. In this context, the use of virtual sources with predetermined acoustic transfer functions (ATFs) has emerged as a promising strategy to impose spatial structure on binaural beamformers without requiring explicit localization or real-time ATF estimation of interfering sources.

Building on the binaural beamforming framework originally proposed by [1] and reviewed in Section 2.2, this thesis aimed to deepen the understanding of how such virtual sources approximate arbitrary interfering sound fields, and to what extent their spatial configuration determines binaural cue preservation. Rather than assuming a fixed or heuristic placement of virtual sources, the central question addressed in this work was whether optimal virtual source configurations can be identified, how robust such configurations are, and which fundamental limitations arise from the virtual-source-based method itself.

The main research question of this thesis (Q1) asked how a limited, optimally selected set of virtual sources can be used to accurately approximate binaural acoustic responses for arbitrary source directions using multi-microphone HRTF measurements. To address this question, the work was structured around two complementary sub-research questions: Q1.1, which focuses on the optimization of virtual source locations, and Q1.2, which concerns the design and use of multi-microphone HRTF measurements required to support such an optimization in a controlled and reproducible manner.

To answer these questions, this thesis combined a measurement-driven experimental foundation with a systematic optimization framework. A new multi-microphone HRTF dataset was created to enable controlled and reproducible evaluation, and an exhaustive location optimization strategy was developed to analyze the behavior of virtual source configurations across a range of parameters. The main contributions, their implications, and directions for future research are summarized below.

5.1. CONTRIBUTION I: MULTI-MICROPHONE HRTF DATASET

The first major contribution of this thesis is the development of a new multi-microphone HRTF dataset, described in detail in Chapter 3. This contribution directly addresses sub-research question Q1.2 by providing a dedicated measurement foundation that enables systematic, measurement-driven evaluation of virtual source selection and binaural beamforming performance. In contrast to most publicly available HRTF databases reviewed in Section 2.1, which typically employ at most two or three microphones per ear and are primarily intended for binaural rendering or localization studies, the dataset generated in this work was explicitly designed to support multi-microphone binaural beamforming research.

Each ear of the dummy head is equipped with a custom array of six microphones, resulting in a total of twelve microphones distributed around the head. This configuration significantly increases the available spatial degrees of freedom compared to existing datasets and enables systematic investigation of how microphone count, placement, and selection affect the trade-off between binaural cue preservation and noise suppression. By providing additional degrees of freedom, the dataset allows more flexibility in imposing beamforming constraints, either to preserve more sources without distortion or to allocate more degrees of freedom to noise reduction, under a fixed and well-defined acoustic geometry. The measurements were performed using a controlled loudspeaker setup with discrete source directions, as described in Section 3.2, ensuring reproducibility across experiments.

Table 5.1: Key properties of the multi-microphone HRTF dataset introduced in this thesis.

Property	**Specification**
Receiver	Head Acoustics HSU III
Microphones per ear	6
Elevation	Horizontal (azimuth only)
Azimuth coverage	360°
Azimuth resolution	15°
Number of source directions	24
Source distance	1.5 m
Measurement repetitions	5 independent sets
Provided data	Audio recordings, HRIRs, HRTFs
Sampling rate	48 kHz
SNR	36 dB

The essential properties of the resulting multi-microphone HRTF dataset are summarized in Table 5.1. The dataset is provided at multiple processing levels. For each measurement set, the raw recorded audio signals, the derived impulse responses (IRs), and the corresponding frequency-domain HRTFs are all available. Multiple independent measurement sets were recorded, enabling repeatability checks and robustness analysis. Together with the accompanying signal processing and evaluation code developed in this thesis, the complete experimental pipeline, from raw measurements to binaural

beamforming evaluation, is fully reproducible. As such, the dataset not only underpins the experimental results presented in this work (see Section 4.4), but also constitutes a reusable research resource for future studies on binaural beamforming, spatial filtering, and HRTF-based modeling.

5.2. CONTRIBUTION II: VIRTUAL SOURCE LOCATION OPTIMIZATION FRAMEWORK

The second main contribution of this thesis addresses sub-research question Q1.1 by developing a systematic framework for evaluating and optimizing virtual source locations for binaural beamforming. The goal of this contribution is to quantify to what extent arbitrary binaural responses can be approximated using a limited subset of measured HRTFs interpreted as virtual sources, and to investigate how the spatial placement of these virtual sources influences binaural cue preservation. The complete methodology is described in Chapter 4 and is built directly on the multi-microphone HRTF dataset introduced in this work.

Virtual source selection is formulated as a discrete combinatorial optimization problem over subsets of candidate source directions. For each candidate configuration, a binaural beamformer is constructed using the selected virtual sources as constraints, after which its performance is evaluated for all possible interfering source directions present in the dataset. To ensure a comprehensive and unbiased analysis, an exhaustive grid-search strategy is employed (Section 4.2), in which all admissible virtual source subsets of a given size are evaluated.

Performance is quantified using an interaural transfer function (ITF) reconstruction error, defined in Section 4.2.1. This metric measures the Euclidean distance between the complex-valued ITFs of the beamformer output and the corresponding reference HRTFs, thereby jointly capturing deviations in interaural level and interaural phase across frequency. As such, the metric directly reflects binaural cue distortion and provides a physically interpretable measure of approximation accuracy.

The experimental results presented in Section 4.4 show that, while virtual sources can indeed be used to approximate binaural responses for arbitrary source directions, the resulting optimization landscape does not exhibit a single, well-defined optimum. Instead, as demonstrated by the configuration error analysis in Section 4.4.4, a large number of distinct virtual source configurations yield nearly identical total ITF errors. The corresponding error distributions are characterized by a broad plateau of near-equivalent solutions and a small number of clear outliers. Importantly, the apparent sensitivity of the optimal configuration to modeling parameters such as the number of virtual sources, assumed SNR/SIR, or microphone configuration is therefore not the result of numerical instability, but rather reflects a fundamental property of the underlying approximation problem.

A key explanation for this behavior is provided by the spatial locality of the virtual source

approximation. As analyzed in Section 4.4.5, accurate ITF reconstruction is primarily achieved when an interfering source lies within a small angular neighborhood (typically below 4°) of one of the selected virtual source directions. Within this local region, multiple virtual source configurations can approximate a given interferer equally well, leading to similar reconstruction errors. As the angular distance to the nearest virtual source increases, the ITF reconstruction error grows approximately linearly up to this range and becomes strongly non-linear beyond it. This demonstrates that, under the ITF-based error metric, virtual sources provide a local approximation of the acoustic transfer functions rather than a globally valid spatial representation.

An important consequence of this locality is that many distinct virtual source configurations achieve comparable overall performance by covering similar regions of the spatial domain, even when their exact directional placements differ. Achieving uniformly low ITF error across all directions would therefore require a very dense virtual source grid. However, increasing the number of virtual sources directly increases the number of beamforming constraints, rapidly exhausting the available degrees of freedom and limiting noise reduction capability. This reveals a fundamental trade-off between spatial coverage, approximation accuracy, and practical feasibility, and answers Q1.1: while virtual source locations can be optimized to reduce binaural cue distortion, the resulting configurations are inherently non-unique and sensitive to system parameters.

5.3. IMPLICATIONS AND LIMITATIONS

Taken together, the results of Chapter 4 clarify both the potential and the inherent limitations of virtual-source-based binaural beamforming. While virtual sources provide an elegant mechanism to impose spatial structure without explicit interferer localization, their effectiveness is fundamentally constrained by spatial locality and the limited number of available degrees of freedom.

These findings place important context around earlier work [1] that assumes uniformly spaced virtual source grids. The present results show that uniform spacing is not necessarily optimal, but also that attempting to optimize virtual source locations does not yield robust, transferable solutions under the given conditions. Beyond a certain point, increasing spatial resolution leads to diminishing returns and increased sensitivity rather than consistent performance gains.

Several limitations of the present work should also be acknowledged. First, the optimization framework relies on exhaustive combinatorial search (Section 4.2), which is computationally expensive and restricts the analysis to offline evaluation. While appropriate for uncovering fundamental behavior, this approach is not directly applicable to real-time or adaptive systems.

Second, the ITF-based error metric, while closely linked to binaural cue preservation, does not fully capture perceptual effects such as externalization, localization blur, or listener-dependent variability. As a result, configurations with similar ITF errors may still differ perceptually.

Finally, all experiments were conducted using a single dummy-head geometry under far-field conditions, as described in Chapter 3. Inter-subject variability, head movements, near-field sources, and dynamic acoustic scenes were not considered, although these factors are expected to play an important role in practical hearing assistive devices.

5.4. DIRECTIONS FOR FUTURE RESEARCH

The results presented in this thesis highlight several directions for future research that can further clarify the behavior of virtual-source-based binaural beamforming and address the limitations identified in Chapter 4. A first and particularly important direction concerns the choice of performance metric. The ITF-based reconstruction error employed in this work is closely linked to binaural cue preservation, but its inherently local nature was shown to play a central role in the observed sensitivity of configuration. Exploring alternative error metrics that are less sensitive to local angular mismatches may therefore lead to more robust and interpretable optimization outcomes. In particular, diffuse-field or diffuse-distance-based error metrics [35], which integrate spatial deviations over a broader angular region, may reduce locality effects and provide a more global measure of approximation quality.

A second promising direction is the development of optimization strategies with lower computational complexity. While the exhaustive grid-search approach used in this thesis was essential for characterizing the structure of the optimization landscape, it is computationally expensive and limited to offline analysis. Future work could investigate greedy selection methods, sequential forward selection, convex relaxations, or stochastic optimization techniques to approximate the performance of exhaustive search at a fraction of the computational cost. Such methods may also provide additional insight into which virtual source directions contribute most strongly to overall performance.

In addition, further analysis is needed to better understand why certain optimization results deviate from intuitive expectations. For instance, the sensitivity of optimal configurations to small parameter changes and the emergence of clustered virtual source placements suggest that interactions between the beamformer constraints, the HRTF spatial structure, and the error metric play a critical role. Studying these interactions more explicitly,e.g., by analyzing the angular gradients of ITF errors, the conditioning of the constraint matrices, or the distribution of residual degrees of freedom, could help disentangle algorithmic effects from fundamental modeling limitations.

Another relevant extension would be to investigate how the conclusions drawn in this work generalize across different acoustic conditions and listener geometries. Incorporating additional dummy-head measurements, personalized HRTFs, or near-field source configurations could reveal how robust the observed locality effects and configuration sensitivities are with respect to inter-subject variability and source distance. Similarly, extending the analysis to dynamic acoustic scenes or time-varying interferer distributions would provide further insight into the applicability of virtual-source-based methods in realistic hearing assistive devices.

Finally, integrating perceptually motivated evaluation criteria alongside signal-based metrics could help bridge the gap between analytical optimization and subjective listening experience. Combining virtual source optimization with perceptual models of localization, externalization, or speech intelligibility may enable the design of binaural beamformers that are not only optimal under a mathematical criterion, but also perceptually robust and practically relevant.

5.5. CONCLUDING REMARKS

In conclusion, this thesis demonstrates that virtual source location optimization is a subtle yet fundamentally constrained problem in binaural beamforming with predetermined acoustic transfer functions. By combining a purpose-built multi-microphone HRTF dataset with a systematic, measurement-driven optimization framework, this work clarifies both why virtual sources can be effective and why their optimization does not yield robust, globally optimal solutions. These insights contribute to a deeper understanding of spatial constraint design in binaural signal processing and provide a solid foundation for future research in this field.

BIBLIOGRAPHY

[1] A. Koutrouvelis, "Multi-Microphone Noise Reduction for Hearing Assistive Devices", en, PhD thesis (Delft University of Technology, 2018).

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones", en, IEEE Signal Processing Magazine **32**, 18–30 (2015).

[3] J. Zhang, "A Parametric Unconstrained Binaural Beamformer Based Noise Reduction and Spatial Cue Preservation for Hearing-Assistive Devices", en, in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (June 2021), pp. 791–795.

[4] S. Fortunato, F. Forli, V. Guglielmi, E. De Corso, G. Paludetti, S. Berrettini, and A. FetonI, "Ipoacusia e declino cognitivo: revisione della letteratura", en, Acta Otorhinolaryngologica Italica **36**, 155–166 (2016).

[5] F. R. Lin, K. Yaffe, J. Xia, Q.-L. Xue, T. B. Harris, E. Purchase-Helzner, S. Satterfield, H. N. Ayonayon, L. Ferrucci, E. M. Simonsick, and Health ABC Study Group, "Hearing loss and cognitive decline in older adults", eng, JAMA internal medicine **173**, 293–299 (2013).

[6] S. Li and J. Peissig, "Measurement of Head-Related Transfer Functions: A Review", en, Applied Sciences **10**, 5014 (2020).

[7] P. Derleth, E. Georganti, M. Latzel, G. Courtois, M. Hofbauer, J. Raether, and V. Kuehnel, "Binaural Signal Processing in Hearing Aids", en, Seminars in Hearing **42**, 206–223 (2021).

[8] J. Cubick, J. M. Buchholz, V. Best, M. Lavandier, and T. Dau, "Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners", The Journal of the Acoustical Society of America **144**, 2896–2905 (2018).

[9] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise", The Journal of the Acoustical Society of America, 10.1121/1.395906 (1988).

[10] T. Vicente, J. M. Buchholz, and M. Lavandier, "Modelling binaural unmasking and the intelligibility of speech in noise and reverberation for normal-hearing and hearing-impaired listeners", en, The Journal of the Acoustical Society of America **150**, 3275–3287 (2021).

[11] A. W. Bronkhorst and R. Plomp, "Binaural speech intelligibility in noise for hearing-impaired listeners", The Journal of the Acoustical Society of America, 10.1121/1.398697 (1989).

[12] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears", en, The Journal of the Acoustical Society of America **25**, 975–979 (1953).

[13] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone Subset Selection for MVDR Beamformer Based Noise Reduction", IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**, 550–563 (2018).

[14] J. Zhang, A. I. Koutrouvelis, R. Heusdens, and R. C. Hendriks, "Distributed Rate-Constrained LCMV Beamforming", en, IEEE Signal Processing Letters **26**, 675–679 (2019).

[15] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust Joint Estimation of Multimicrophone Signal Model Parameters", en, IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**, 1136–1150 (2019).

[16] W. M. Hartmann, "Localization and Lateralization of Sound", en, in *Binaural Hearing: With 93 Illustrations*, edited by R. Y. Litovsky, M. J. Goupell, R. R. Fay, and A. N. Popper (Springer International Publishing, Cham, 2021), pp. 9–45.

[17] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence", en, The Journal of the Acoustical Society of America **116**, 3075–3089 (2004).

[18] V. Bruschi, L. Grossi, N. Dourou, A. Quattrini, A. Vancheri, T. Leidi, and S. Cecchi, "A Review on Head-Related Transfer Function Generation for Spatial Audio", Applied Sciences **14**, 11242 (2024).

[19] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical Analysis of Binaural Transfer Function MVDR Beamformers with Interference Cue Preservation Constraints", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 10.1109/TASLP.2015.2486381 (2015).

[20] E. Hadad, S. Doclo, and S. Gannot, "The Binaural LCMV Beamformer and its Performance Analysis", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 10.1109/TASLP.2016.2514496 (2016).

[21] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions", en, The Journal of the Acoustical Society of America **106**, 1465–1479 (1999).

[22] I. Engel, R. Daugintis, T. Vicente, A. O. T. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The SONICOM HRTF Dataset", en, Journal of the Audio Engineering Society **71**, 241–253 (2023).

[23] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database", en, in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575) (2001), pp. 99–102.

[24] A. Vidal, P. Herzog, C. Lambourg, and J. Chatron, "HRTF measurements of five dummy heads at two distances", en, in 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA) (Sept. 2021), pp. 1–8.

[25] V. Bruschi, A. Terenzi, N. A. Dourou, S. Spinsante, and S. Cecchi, "Comparative Analysis of HRTFs Measurement Using In-Ear Microphones", en, Sensors **23**, 6016 (2023).

[26] F. Denk, S. M. A. Ernst, S. D. Ewert, and B. Kollmeier, "Adapting Hearing Devices to the Individual Ear Acoustics: Database and Target Response Correction Functions for Various Device Styles", en, Trends in Hearing **22**, 2331216518779313 (2018).

[27] A. W. Bronkhorst and J. A. Verhave, "A Microphone-Array-Based System for Restoring Sound Localization with Occluded Ears", en,

[28] H. S. Braren and J. Fels, *A High-Resolution Individual 3D Adult Head and Torso Model for HRTF Simulation and Validation: HRTF Measurement*, en, 2020.

[29] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances", en, (2011).

[30] J. Thiemann and S. Van De Par, "A multiple model high-resolution head-related impulse response database for aided and unaided ears", en, EURASIP Journal on Advances in Signal Processing **2019**, 9 (2019).

[31] D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera, "Design of microphone arrays for hearing aids optimized to unknown subjects", Signal Processing **93**, 3239–3250 (2013).

[32] *SOFA (Spatially Oriented Format for Acoustics) - Sofaconventions*.

[33] B. Cornelis, S. Doclo, T. Van Dan Bogaert, M. Moonen, and J. Wouters, "Theoretical Analysis of Binaural Multimicrophone Noise Reduction Techniques", en, IEEE Transactions on Audio, Speech, and Language Processing **18**, 342–355 (2010).

[34] AKG, *AKG C417 Polar Patterns*, Datasheet, 2012.

[35] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A Study on Manifolds of Acoustic Responses", en, in *Latent Variable Analysis and Signal Separation*, Vol. 9237, edited by E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Series Title: Lecture Notes in Computer Science (Springer International Publishing, Cham, 2015), pp. 203–210.