Beamforming-based Acoustic Crosstalk Cancelation for Spatial Audio Presentation

Christoph Hohnerlein
319 392

Master Thesis

Submitted in partial fulfillment for the
degree of Master of Science (M.Sc.)

Audio Communication and Technology
Faculty I, Technical University of Berlin

1st Supervisor:	Prof. Sebastian Möller	Quality and Usability Lab
2nd Supervisor:	Prof. Stefan Weinzierl	Audio Communication Group
Co-Supervisor:	Dr. Jens Ahrens	Quality and Usability Lab

May 4, 2016

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Christoph Hohnerlein

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Christoph Hohnerlein

Acknowledgements

First and foremost, I would like to thank my mentor Dr. Jens Ahrens for the great amount of time and effort invested, both abroad and at home.

Furthermore, I would like to thank Hagen Wierstorf for the discussions on perception as well as Clemens Zimmer for all the help setting up the experiment. Last but not least, thanks to all the great people at CCRMA who made my research stay such a memorable experience and to all my participants for taking their time out for me on short notice.

Abstract

Beamforming can be used to tightly control the radiation pattern of speaker arrays in space of a large frequency band, including targeting and null-steering. Using a Least-Squares Frequency Invariant Beamformer (LSFIB), signal transmission to listener's ears can be simulated with high channel separation (up to 40 dB) over a frequency band of 1 kHz − 8 kHz, that is highly robust against inaccuracies in target position and reproduction setup. An eight speaker array and real-time software was built, which also adds Recursive Ambiophonic Crosstalk Elimination (RACE) to extend the working frequency band down to 250 Hz. Although the measured channel separation was lower (10 dB − 25 dB), the real array was also found to provide decent crosstalk cancelation for binaural presentation.

In a 21 subject study, this system was found to successfully deliver binaural audio material to a listener sitting 1 m in front of the array. Compared to the ground truth presentation with headphones, significantly more front-back confusions occurred, resulting in a larger absolute error ($F(1) = 91.43, p < 0.001$). Conversely, the precision of localization was significantly higher using the array when discounting front-back confusions ($F(1) = 23.56, p < 0.001$). Perceptually, only externalization was rated significantly different between headphones and the array ($t(20) = -3.983, p > 0.001$). Data suggests that for

a small group of six best-case participants, no significant differences between presentation methods were found ($F(1) = 1.91, p = 0.167$).

Deutsche Zusammenfassung

Beamforming kann verwendet werden, um die Abstrahlcharakteristik eines Lautsprecherarrays über einen breiten Frequenzbereich zu steuern. Optimiert man diese nun so, dass sich für einen Menschen vor dem Array gleichzeitig ein hoher Schalldruck an dem einen und niedriger Schalldruck an dem anderen Ohr ausbildet, besitzt man ein hohes Maß an Kontrolle über die am Ohr anliegenden Signale ohne die sonst bei Lautsprechern üblichen Übersprecheffekte. Dies ermöglicht das korrekte Abspielen von binauralem Audiomaterial, also von Ohrsignalen eines (virtuellen) Zuhörers einer räumlichen Szene.

Ein solches System aus acht Lautsprechern wurde mit Hilfe von konvexer Optimierung designed und anschließend mit guten Ergebnissen simuliert. Die Simulation betrachtete dabei die sich ausbildenden Signale einschließlich der sich ausbildenden Reflexionen an randomisierten Stellen um den Zuhörerkopf, welcher im Abstand von 1 m vor dem Array platziert wurde. Ausserdem wurde die Robustheit gegen kleine Ungenauigkeiten in Amplitude und Phase der generierten Treiberfunktion bestätigt. Insgesamt wurde Übersprechen in der Simulation mit durchschnittlich etwa 40 dB abgedämpft.

Da Beamforming (1 kHz−8 kHz) nur einen Teil der relevanten Frequenzbereichs abdeckt wurde im Bassbereich (125 Hz − 1 kHz) zusätzlich noch Recursive Ambiophonic Crosstalk Elimination (RACE) zur Kanaltrennung verwendet.

Ein solches System wurde nun real aufgebaut: Nach ausgiebiger Kalibrierung konnte die reale Übertragungsfunktion und die Kanaltrennung mit Hilfe eines Kunstkopfes ausgemessen werden. Dabei wurde eine geringere Kanaltrennung von etwa $10\,\mathrm{dB} - 25\,\mathrm{dB}$ aufgenommen.

Zuletzt wurde das Array in ein Hörexperiment mit 21 Versuchspersonen getestet, dabei sollte eine virtuelle Audioquelle auf einem Kreis um den Teilnehmer geortet werden. Als Vergleich dienten kalibrierte Kopfhörer, welche für eine solche Präsentation wegen dem inherent Geringen Übersprechen als ideal gelten. Das Abspielen von binauralen Quellen mit Hilfe des Arrays funktionierte in vielen Fällen gut, jedoch hatten viele Teilnehmern Probleme Quellen von hinten wahrzunehmen, somit war der gemessene absolute Fehler signifikant größer ($F(1) = 91.43, p < 0.001$). Dies ist wahrscheinlich auf die Überlagerung der Lokalisierung-cues des Materials durch die des Arrays, welches ausser Sicht vor den Teilnehmern platziert war, zurückzuführen. Vernachlässigt man aber solchen vorne-hinten Verwechslungen lag der Fehler unter Verwendung des Arrays signifikant unter dem der Kopfhörer ($F(1) = 23.56, p < 0.001$). Ausserdem wurde bei der Präsentation von Kopfhörern ein signifikant ($t(20) = -3.983, p > 0.001$) höheres Maß an Externalisierung wahrgenommen. Bei einer Gruppe von sechs best-case Teilnehmern wurde kein signifikanter Unterschied zwischen Lokalisierungsperformance bei Nutzung von Array oder Kopfhöhrern gemessen ($F(1) = 1.91, p = 0.167$).

Contents

List of Figures

List of Tables

Listings

Abbreviations

Notation	Description
BF	BeamForm
BRIR	Binaural Room Impulse Response
CTC	Cross-Talk Cancellation (also XTC in literature)
FBA	Front-Back Ambiguity
FSB	Filter & Sum Beamformer
HP	HeadPhones
HRIR	Head Related Impulse Response
HRTF	Head-Related Transfer Function
ILD	Interaural Level Difference
IR	Impulse Response
ITD	Interaural Time Difference
LSB	Least-Squares Beamformer
LSFIB	Least-Squares Frequency Invariant Beamformer
LTI	Linear Time-Invariant
MVDR	Minimum Variance Distortionless Response
RACE	Recursive Ambiophonic Crosstalk Elimination
SNR	Signal-Noise Ratio
WFS	Wave Field Synthesis
WNG	White Noise Gain

Nomenclature

Notation	Description
(α, β)	Azimuth α and colatitute β (see coordinate system)
$Y_n^m(\beta, \alpha)$	Complex spherical harmonic coefficients of degree n and order m
$\breve{S}_n^m(\omega)$	Expansion coefficients of sound field $S(\mathbf{x}, \omega)$
$\mathbf{x_n}$	Short hand notation for vector $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$
c	Speed of propagation. Acoustic: $c_{air} \approx 343\,\mathrm{m\,s^{-1}}$
dB	deciBell (logarithmic ratio unit): $X_{dB} = 10log_{10}\left(\frac{X}{X_0}\right)\,\mathrm{dB}$
$h_n^{(1,2)}(\cdot)$	n-th order spherical Hankel function of first and second kind
$j_n^{'}(\cdot)$	n-th order spherical Bessel function of the first kind

Chapter 1

Introduction

1.1 Aim of proposed system

The final goal of the system is to deliver binaural audio material to an untethered listener's ears using speakers, enabling the perception of virtual sound sources from any direction. This can already be accomplished using traditional headphones, which excludes a large range of situations where wearing headphones might be unacceptable such as in social interactions, work settings or in traffic.

To achieve this, two obstacles had to be overcome: First and foremost, the separation between the channels reaching the left and the right ear had to be maximized, i.e. the crosstalk needs to be suppressed. Secondly, the transmission to the ear should not add any additional changes in amplitude or phase to the original channel. Both constraints are vital as to preserve the localization cues encoded in the presented material, as well as the overall timbre.

Furthermore, the system is stable towards imprecise means of reproduction and produces decent channel separation over a large listening area as shown in simulation, measurement and user testing.

1.2 Current state of research

A common way of suppressing crosstalk between speakers is filter inversion the of transmission matrix between the ears as suggested by Cooper & Bauck (1989) and refined in Bauck & Cooper (1996). The main drawback of this method is the large instabilities of such a system, where small mistakes in path prediction can cause the system to fail. There have been many attempts of improving this scheme. Importantly, Gardner (1998) built and tested a dynamic system that keeps track of the subject and adjusts its filter inversion accordingly. Other approaches, such as the virtual sources of Menzel et al. (2005) and the beamforming of Guldenschuh et al. (2010), Guldenschuh & Sontacchi (2009), control array radiation patterns to optimally transmit binaural signals to the listeners ear.

Several approaches of crosstalk cancellation are discussed further in Section 2.4.2, while Section 2.5.3 gives an overview of various beamforming methods.

1.3 Overview of thesis

In the following Chapter 2, the basic theoretical background for the proposed speaker array will be presented. This starts with an overview of the principles of human perception of source location in Section 2.1, followed by a review of the relevant parts of system theory concerning transmission properties are discussed in Section 2.2. The final presentation of sound is reviewed in Section 2.3. Section 2.4 then takes a closer look at crosstalk and its suppression, while Section 2.5 will explore the theory of beamforming and its many applications. Section 2.6 finally quickly explains why the extensive research on sensor beamforming is applicable in speaker beamforming.

The implementational details are discussed in Chapter 3, which is split into two parts: Firstly, Section 3.1 explains the a-priori offline convex optimization to generate the beamforming weights with the necessary properties and Section

3.2, dealing with the frequency band separation and the signal flow for real-time playback capabilities.

Chapter 4 presents the validation data collected through the means of simulation in Section 4.1, including scattering at the listener's head and robustness against variances in position and weight, as well as calibration and subsequent measurements of a real array using a dummy head in Section 4.2.

For perceptual verification, the medium-sized user study is outlined in Chapter 5, which firstly states the goals of the study in Section 5.1 and then explains the setup in detail in Section 5.2. Section 5.3 presents all collected answers which are subsequently discussed in Section 5.4.

Lastly, Chapter 6 draws conclusions and gives an outlook onto possible future research.

Chapter 2

Theory

The following chapter gives a short overview over the topics and techniques used in this thesis. As this work heavily builds on the work of others, it cannot possibly reproduce the necessary theory in all thoroughness.

2.1 Spatial hearing

Spatial hearing is the ability of a listener to analyze and process a virtual scene by assigning stimuli streams to virtual positions in space. This is not only allows for the localization of a single sound source but also for the separation of multiple sources that are playing simultaneously. The three most important cues for locating both real and virtual sound sources are Interaural Level Difference (ILD), Interaural Time Difference (ITD) and changes in frequency spectrum. Furthermore, there are dynamic cues evaluated in specific situations. The physical characteristics of these cues can vary highly between people and depend on physiological factors such as structure of the pinna (outer ear) and geometry of the listener's head. The evaluation of these cues is most probably learned in child development and constantly re-trained. The following sections summarize relevant parts from Blauert (1997), Fastl & Zwicker (2007), Weinzierl (2008).

2.1.1 Interaural Level Difference (ILD)

Any source that is not positioned on the median plane, which is the vertical plane directly between the listener's ears, will invoke a difference in sound pressure reaching both ears relative to its distance from each. Humans are generally sensitive to differences of as little as 1 dB and the effect maps linearly to the level difference in dB.

2.1.2 Interaural Time Difference (ITD)

Similarly, pressure waves propagating from sources positioned outside the median plane will reach the contralateral ear slightly later. This time difference is again evaluated linearly up to a time difference of about $\tau_{ph} \approx \pm 600 \mu s$, which corresponds to the travel time across the average head ($r_{avg} \approx 10\,\mathrm{cm}$) at a propagation speed of $c_{air} \approx 343\,\mathrm{m\,s^{-1}}$. Larger differences will be perceived as a delayed copy of the original sound source.

2.1.3 Spectral cues

Consider a source straight in front of a listener's head: The only differences to an identical source straight behind the head are the changes in amplitude spectrum, which is altered due to the different paths around the head and the corresponding scattering, interference and reflection. This also applies to a pair of sources that is located above and below a listener. In fact, there is an open cone (see Figure 2.1) for any given point in space where the distance to both ears is constant, resulting in identical ILD and ITD cues.

This is especially important in up-down and front-back localization, as the user has to exclusively rely on these spectral cues for correct localization. This is a frequent problem in binaural synthesis, commonly called **front-back confusion/ambiguity**.

Figure 2.2 illustrates the ILD, ITD and coloration cues at an angle of 20°

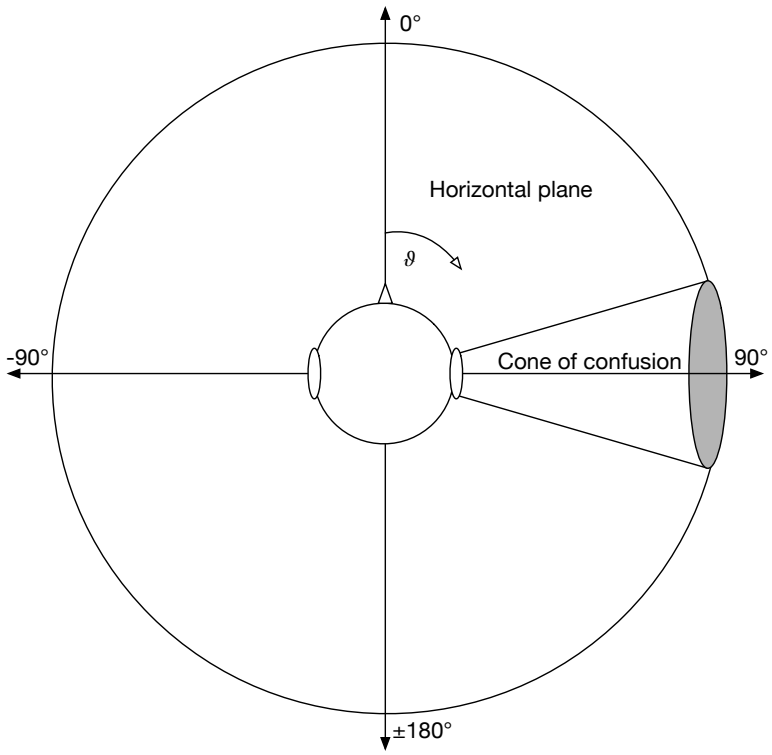


Figure 2.1 Labeling of angles used throughout this thesis with azimuth $\vartheta = 0°$ being straight ahead and positive angles denoting clockwise rotation. Additionally, the **cone of confusion** is depicted as an open cone of points that all invoke the same ILD and ITD cues.

in a set of ear signals.

More complex scenes can include reflections, movement or multiple sources and require additional psychoacoustic processes to help perceiving a realistic auditory scene. The precedence effect for example strongly anchors a source at the localization of the first arriving wave front and deprioritize following similar stimuli as reverberation or echo.

2.1.4 Frequency range of cues

Due to interaction of the corresponding wavelengths with the physical attributes of the human head, the localization cues mentioned above can only be evaluated in certain parts of the frequency spectrum. For very low frequen-

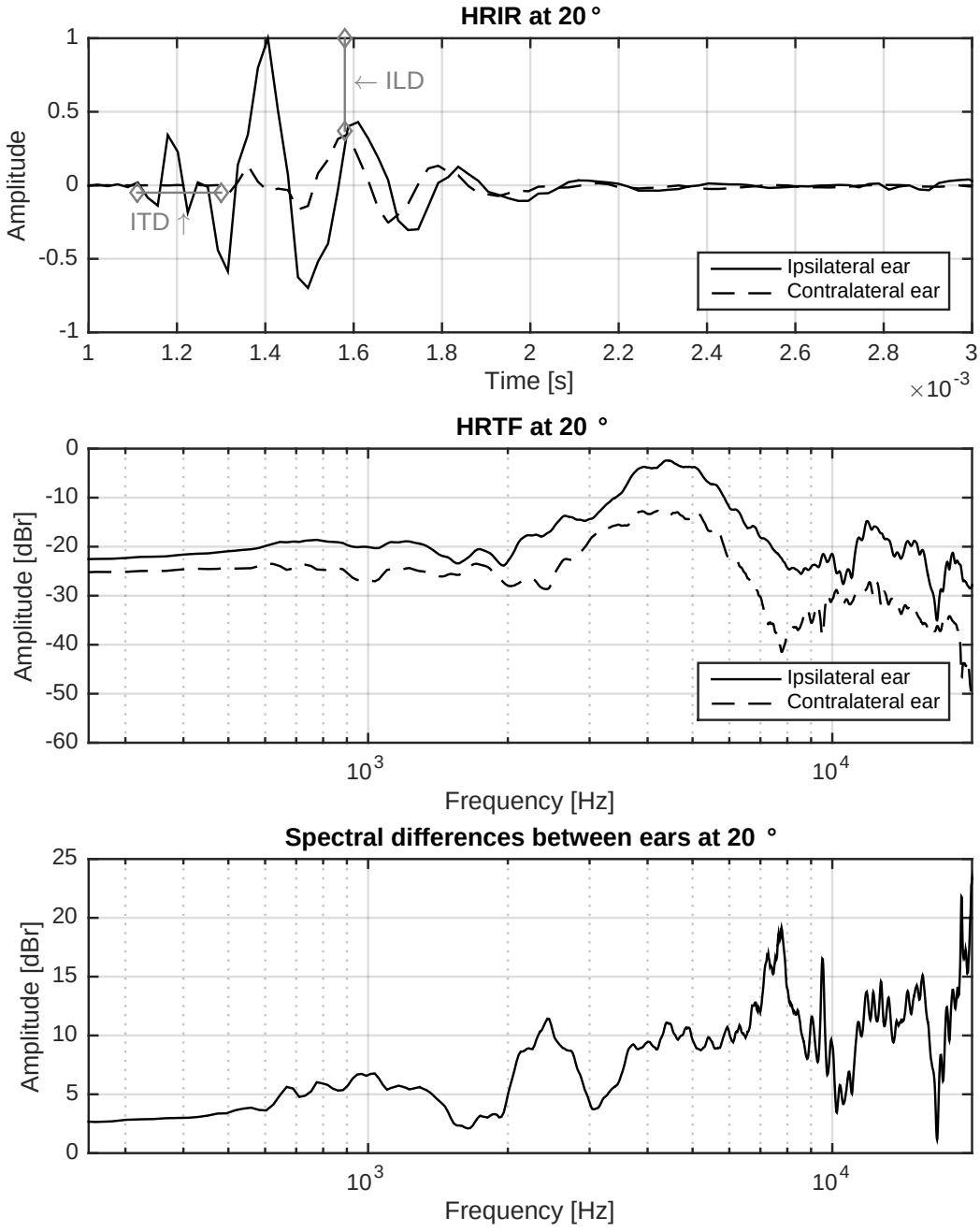


Figure 2.2 Localization cues of the ear signals invoked by a source at an angle of 20° and 1 m distance, measured with a KEMAR dummy head, see Section 2.2.2

cies **below 80 Hz**, ILD, ITD and changes in timbre are not differentiable and almost no directional information can be perceived. **Until about 800 Hz**, listeners mainly rely on interaural time differences, since half the wavelength extends past the average ear distance ($0.5 \cdot \frac{343\,\mathrm{m\,s^{-1}}}{800\,\mathrm{Hz}} = 21.4\,\mathrm{cm}$) which make differences in phase meaningful. Interaural level differences become important **above 1.6 kHz**, when shadowing of the head provides sufficient attenuation. In between, both cues are evaluated.

2.1.5 Dynamic cues

Wallach (1939) suggested that slight movements of the head are used to resolve location of static sources. This is confirmed by Wightman & Kistler (1999), where front-back confusion of binaurally synthesized sources decreased when encouraging natural head movements compared to sitting completely still. This can only be partly recreated by moving the source in relation to the listener. This is confirmed by Nykänen et al. (2013), who additionally found that added reflexions in a virtual scene help further reducing the amount of front-back confusion.

2.2 Audio transmission properties

The following section covers the most important aspects of transmitting a signal from a emitter to a receiver from the point of view of systems theory.

2.2.1 LTI systems

All **Linear Time-Invariant (LTI) systems** exhibit two properties that make them an important abstraction for transmission processes:

Linearity The multiplication of an input $x_1(t)$ with any real scalar c_1 will result in a linear scaling of the output $y_1(t)$. This holds for multiple inputs

and corresponding scalars:

$$c_1 x_1(t) + c_2 x_2(t) + \cdots \rightarrow c_1 y_1(t) + c_2 y_2(t) + \ldots \tag{2.1}$$

Time-invariance The mapping of an input to the output of a system is not dependent on the time of transmission. Adding an arbitrary delay τ to an input $x(t)$ will produce the identical output $y(t)$ delayed by τ:

$$x(t - \tau) \rightarrow y(t - \tau) \tag{2.2}$$

2.2.2 Impulse response and transfer function

All transformations of a LTI system are encoded in its response to a Dirac impulse $\delta(t)$, which is a theoretical signal of all zeros expect at time t, where the amplitude is $+\inf$; the integral of $\delta(t)$ is 1. This response is called the **Impulse Response (IR)** $h(t)$ of that system, its Fourier transform in the frequency domain is the system's **Transfer Function** $H(\omega)$. Importantly, the transformation of such an LTI system can be applied to an arbitrary input signal $x(t)$, either in the time domain by convolving it with the impulse response $h(t)$ or in the frequency domain by multiplying with the transfer function $H(\omega)$.

This circumstance can be used to avoid the computationally expensive time-domain convolution by using the so called **fast convolution**. Here, the Fourier transforms of input ($X(\omega)$) and system ($H(\omega)$) are simply multiplied and transformed back into the time domain to obtain $y(t)$, as shown in Figure 2.3. This is faster than a single convolution due to the extremely efficient implementation of Fourier Transforms using the Fast-Fourier-Transform (FFT) algorithm, originally by Welch (1967).

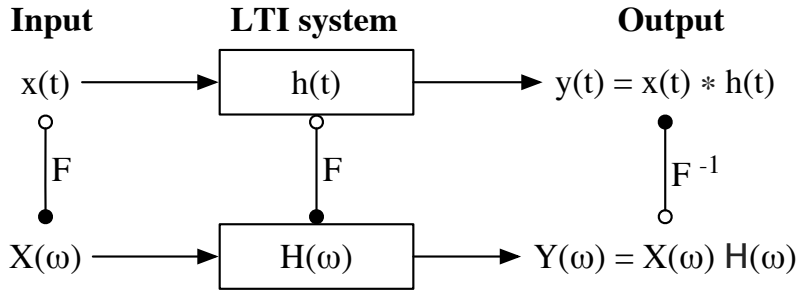


Figure 2.3 Fast convolution of an input $x(t)$ with LTI impulse response $h(t)$ using the Fourier transforms $X(\omega)$ and $H(\omega)$.

2.2.3 HRIR, HRTF and BRIR

As explained in Section 2.1, spatial hearing is highly reliant on the inter-aural differences of signals reaching a listener's ear. Furthermore, it is reasonable to consider the transmission from a sound source (loudspeaker) through a linear medium (air) to a receiver (ears) to be a linear, time-invariant system (LTI system, see Section 2.2.1). Therefore, all changes to an audio signal traveling from a certain point in space to the entrance of the ear canal can be captured in the **Head Related Impulse Response (HRIR)** or its **Head-Related Transfer Function (HRTF)** respectively. They encode all psychoacoustic cues on source localization listed in Section 2.1 as well as the properties of the transmission room.

Such transfer functions can either be recorded on dummy heads that resemble average human proportions (Gardner & Martin 1995, Møller 1992) or directly on human subjects. Ideally, one would record and store the transmission path between all points in space and both ears. For practical reasons, usually only the horizontal plane at one constant distance is considered. Since it is highly impractical to produce (and record) an actual Dirac impulse, sweeping sine measurements with subsequent deconvolution are common practice, as described in Müller & Massarani (2001).

To be independent of the room properties such as absorption and reverber-

ation, Head Related Impulse Responses (HRIRs) are commonly recorded in an anechoic setting. The acoustic properties of a room can be added later by additionally convolving with the impulse response of that room. Such a superposition of localization and room impulse response is referred to as **Binaural Room Impulse Response (BRIR)**. Alternatively, a full set of HRIRs can also be recorded directly in place, greatly increasing the recording time needed on location.

2.3 Sound reproduction

Recording of sound using notation or mechanical means predate modern audio technology by several centuries, as per McKendrick (1909) and Bouchard (2013). The evolution towards the now common record and playback techniques started in the early 20th century and proceeded rapidly: Highly developed audio systems by means of digital processing, cheap production and high quality materials are now commonplace and readily available in many form factors and a wide field of application. (Rossing 2007)

2.3.1 Traditional speaker playback

While monophonic recording and playback was the only format to save and play music until the 1960s, stereophonic systems have since been widely adopted. In recent years, a push for home cinema experience has resulted in an even higher number of speakers and the addition of specialized speakers, most commonly subwoofers. Furthermore, personal audio consumed via headphones is ubiquitous due to the rise of mobile audio players. It is therefore not surprising that the presentation of audio material is dependent on the system that is being played back on.

The common stereophonic system with two speakers forming an isosceles triangle with the listener allows for the playback of stereo material. Sources

can be artificially placed between the two speakers by adjusting the level difference between the speakers, which is referred to as panning. These so called phantom sources usually don't exhibit the correct ITD or spectral cues. Additionally, the added reflections and reverberations of the playback room affect the perception of audio material on such a stereo system. Still, humans tend to not perceive many of the mentioned problems and even prefer the sound of stereo.

2.3.2 Binaural Playback

Although very common, playing back stereo material on headphones is an incorrect mapping of input to output. This wrongly evokes source localization inside the listener's head and panning induces movement between the ears. Instead, **binaural** recordings played back on headphones are the correct mapping as explained in Bauer (1961).

Here, by recording an auditory scene through microphones placed inside a dummy head's ear canals, the actual ear signals can be transmitted to the listener's headphones, including all cues mentioned in Section 2.1. Under ideal circumstances, this will perfectly reproduce the same auditory scene including full 360° localization and externalization. For example, such a dummy head can can be seated in the front row of a sold out concert - in fact, such setup was already used as early as 1881 to present more listeners with a "first row experience" (Hertz 1981). Since then, binaural audio material has been widely used in artistic, engineering and medical contexts (Gierlich 1992).

Using the techniques described in Section 2.2, an audio source can also be imprinted with all localization cues when convolved with a HRIR, allowing arbitrary placement in auditory scene around the listener. This is called **binaural synthesis**. When using static binaural synthesis, the auditory scene will turn with the rotation on the subject's head, since the location cues stay constant. This can be alleviated by tracking the listener's position, for example

by attaching a **headtracker**. Knowing the listener's rotation in relation to the position of the audio source allows for correctly interchanging the Head-Related Transfer Functions (HRTFs), resulting in a scene that stays in place. Furthermore, small rotations of the head can help with front-back confusion, as explained in Section 2.1.5.

2.4 Crosstalk

In the following, a brief definition of crosstalk is given before examining various methods of suppressing said crosstalk.

2.4.1 Real world occurrence

In general terms, any unwanted leakage between parallel information channels is considered crosstalk. The two most common forms are electrical, caused by coupling of components on circuits or cables, and acoustical crosstalk, caused by the propagation properties of air. Although there are exceptions, crosstalk is generally considered a negative quality of a system and is tried to be avoided using isolation or suppression.

In the case of binaural material, any change of the ear signals will alter the perception in unpredictable ways. It seems clear that simply playing back binaural audio on a speaker setup will result in highly different signals reaching the listener's ears due to the inevitable crosstalk between both speakers to ear transmissions.

2.4.2 Cancellation methods

To avoid acoustic crosstalk, various methods have been tried since the 1960s:

Filter inversion

The earliest mention of crosstalk and its suppression may be found in Bauer (1961), where the mismatch between means of recording and playback is examined, which at time of writing were either binaural or stereophonic. To correct the headphone playback of material that has been recorded with a stereophonic system, a hardware circuit is suggested, which emulates the crosstalk between two channels. An inverse system, which would remove unwanted crosstalk when playing back a binaural recording on a stereophonic speaker setup is only hypothesized.

Later, when investigating the faithful reproduction of concert hall simulations in Schroeder & Atal (1963), compensation filters were used to present binaural signals to a listener inside an anechoic chamber, see Figure 2.4. As he noted himself in Schroeder (1984), suppressing crosstalk using inverse transfer functions are not very robust. They work correctly only over a very small sweet spot and might already break down due to a non-standard head shape of the listener. Of course, even slight turning of the head "destroyed the acoustic illusion".

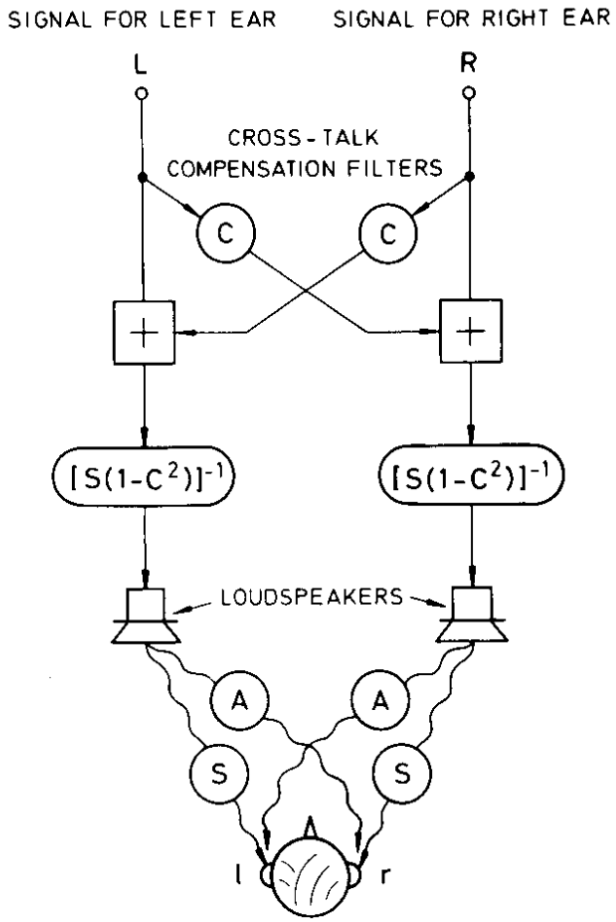


Figure 2.4 Crosstalk cancellation filter that transmit the signals of a dummy-head's ears to a listener using loudspeaker, from Schroeder (1984)

These filter inversion schemes have been refined by Damaske (1971), Mori et al. (1979), Cooper & Bauck (1989) and Bauck & Cooper (1996). Optimization of speaker position was explored by Griesinger (1989), Ward & Elko (1998, 1999), Lopez & Gonzalez (2001) and Bai et al. (2005). In his phd thesis, Gardner (1998) suggested real time modification to best fit the current listener position and then implemented and tested such as system. A great overview of various inversion methods can be found in Kaiser (2011).

Path estimation - RACE

One pragmatic and successful approach was suggested in Glasgal (2007), where the path from a speaker to the contralateral ear is estimated as a simple delayed

attenuator. Therefore, it can be cancelled out by an attenuated, delayed and *inverted* signal on the opposite channel. This cancellation signal, in turn, also needs to be compensated on the original ear, and so on. This leads to the recursive filter at the heart of the proposed Recursive Ambiophonic Crosstalk Eliminator (RACE) system.

Figure 2.5 shows the IIR filter structure of such a system. The frequency band was limited from 250 Hz to 5 kHz due to the contained localization cues carried as well as physical limitations.

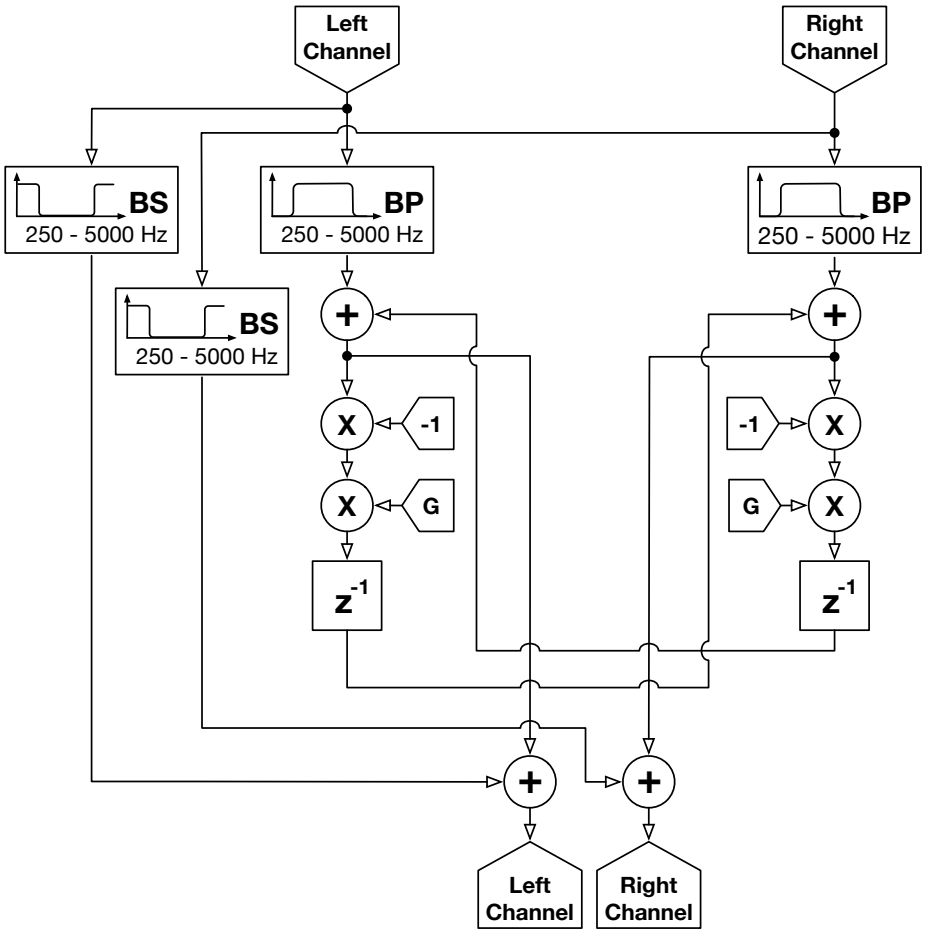


Figure 2.5 Recursive RACE filter structure after Glasgal (2007).

Array based

Other approaches leverage the optimization of speaker arrays, similar to the system of this thesis. They all generally have much improved robustness against displaced listeners and they tend to fail more gently.

Menzel et al. (2005) employed an elevated ring of speakers to synthesize two focused sources using Wave Field Synthesis (WFS) (See Ahrens et al. (2008) and Geier et al. (2010)). The system was optimized for an area of $10\,\text{cm} \times 5\,\text{cm}$, where it successfully transmitted binaural cues for several virtual sources of different location, both in azimuth as well as elevation.

Guldenschuh & Sontacchi (2009) used a beamforming array to be installed on top of a computer monitor, which aims at a person sitting 60 cm in front. Only a comparatively small bandwidth of $300\,\text{Hz} - 2500\,\text{Hz}$ was evaluated but several optimization methods were compared. This was later implemented and perceptually evaluated in an air controller setting in Guldenschuh et al. (2010).

In Ahrens et al. (2013), a larger horizontal speaker array was used to optimally leverage natural head shadowing to deliver frequency content from 2000 Hz to 9000 Hz. Furthermore, it introduces the idea of additionally using RACE to increase channel separation at lower frequencies.

Other CTC approaches

As a completely different approach, Bock & Keele Jr (1986) install a solid boundary between the subject and the speakers to measure the effects on localization and comb-filtering effects. While this certainly demonstrates a straightforward method to eliminate crosstalk, it is not very viable Figure 2.6.

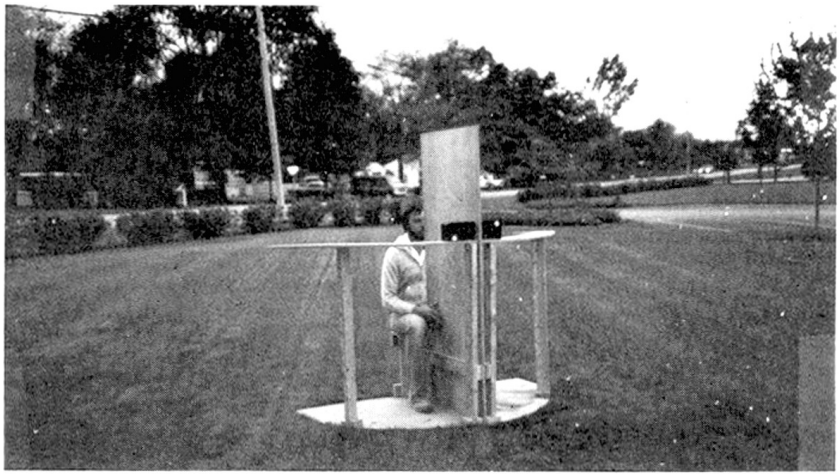


Figure 2.6 Crosstalk cancellation using a solid boundary, from Bock & Keele Jr (1986)

In Polk (1984, 2005), a hardware speaker system with additional drivers is designed, which emits the inverted stereo signal of the opposite channel to cancel the crosstalk of traditional stereo speakers. This so-called SDA system and the following surround sound strategies are actively used in sound bars and similar setups and are a registered trademark of SRS Labs Inc.

2.5 Beamforming

The following section on beamforming follows the literature in taking the perspective of an array of receiving sensors. As explained in Section 2.6, the same principles and design methods hold for speaker arrays.

Beamforming is a form of spatial filtering and is generally used to receive signals in the presence of unwanted noise, such as radiation or interfering sources, when a other modes of separation (for example time or frequency) are not available. It is heavily relied on in radar, sonar, communication, imaging, biomedical and exploration application. It can be framed as an optimization problem but the implementation heavily relies on a wide range of parameters such as signal frequency, spatial sampling distance, number of sensors, target angle, possible null angles and varying degrees of knowledge on signal or inter-

ference properties as well as adaptive qualities. From the very first application of Fourier Transform on spatially sampled data by Bartlett (1948), beamforming has a long history of refinements. The following sections should give a short introduction on the concept of beamforming but the author suggests Van Veen & Buckley (1988) for an in-depth review of various beamforming methods and Krim & Viberg (1996) as well as Van Trees (2002) for an even broader analysis of array signal processing.

2.5.1 Delay & Sum

In its simplest form, a beamformer is a single array of equally spaced sensors, of which each input may be independently delayed and weighted with a complex factor. The system output is the simple summation of all individually delayed and weighted inputs:

$$y(k) = \sum_{n=1}^{N} w_j^* x_j(k), \tag{2.3}$$

where N is the number of sensors, w_n is the $n-th$ complex weighting factor and x_n is the $n-th$ input signal. * represents the complex conjugate and is used to simplify notation.

With only one weighting factor, a beamforming system can be tuned to a certain reception pattern for a single frequency. To increase the frequency range, more frequency-dependent weights are needed, which again are multiplied with the respective incoming signal and summed to form the system's output. Now, the system is not only sampling in space (with spatially distributed sensors) but also in time with a complex filter:

$$y(k) = \sum_{n=1}^{N} \sum_{p=0}^{P-1} w_{n,p}^* x_n(k-p) \tag{2.4}$$

Figure 2.4 describes such a broadband Filter & Sum Beamformer (FSB) with a total of N sensors and P delayed complex weights for each, Figure 2.7 shows a graphical representation of such a beamformer. The sensitivity over

each angle is of course directly affected by the weights and is called the beam pattern, several of which are shown in Figure 2.8.

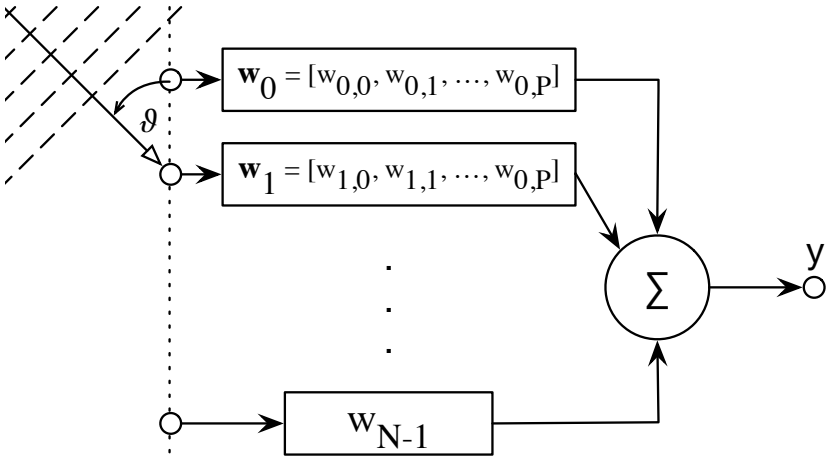


Figure 2.7 Schematic of Filter & Sum Beamformer (FSB) with N sensors and weights $\mathbf{w}$ of length P.

Lastly, the notation can be much simplified when using bold symbols to indicate vectors to yield:

$$y(k) = \mathbf{w}^H \mathbf{x}(k), \tag{2.5}$$

where H notes the Hermitian transpose (conjugate transpose). Now, $\mathbf{w}$ describes a set of N FIR filter of length P.

2.5.2 Null Steering

A non-ideal transmission system is rarely free of unwanted noise. If the interfering source is a spatially different sender rather than general background noise, null-steering may be employed to decrease sensitivity towards particular directions. Constraints may be introduced to strongly reduce the gain for a certain angle of arrival, for the cost of increasing noise in other directions. Null-steering is a statistically optimized design pattern, as it relies on statistical properties to optimize the array response. Figure 2.8 shows such a null

steering, where in the beam pattern of the third graph, signals from $\vartheta = 40°$ are strongly rejected.

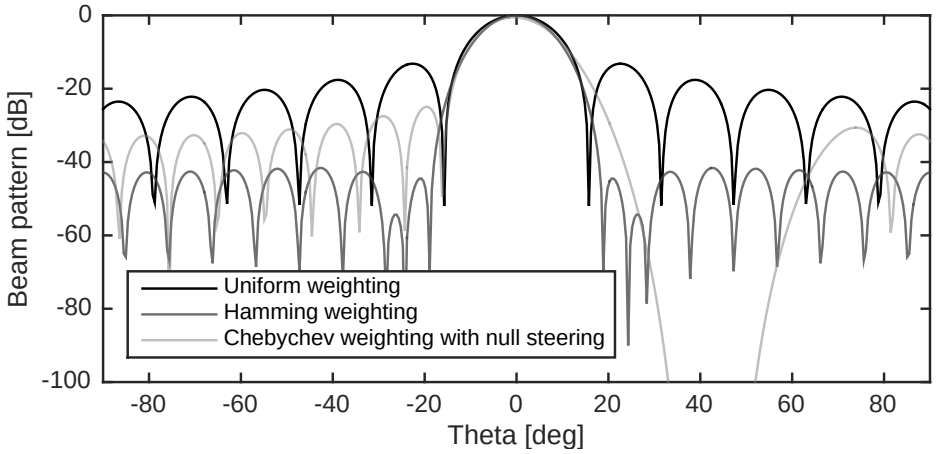


Figure 2.8 Beam patterns of uniform weighting, Hamming weighting and Chebychev weighting with null steering at $\vartheta = 40°$, after Van Trees (2002).

2.5.3 Optimization

A short review of relevant optimization strategies follows. A much more complete overview can be found in Van Veen & Buckley (1988) and Krim & Viberg (1996).

Capon and Least-Square

Capon (1969) suggested to minimize the power from other directions than the target direction, where it keeps a constant gain of 1. It uses every available degree of freedom to concentrate the received energy along one direction, outperforming conventional beamformers. This is commonly referred to as a **Minimum Variance Distorsionless Response** beamformer (also **Capon** beamformer) and has also been used to suppress crosstalk by Ma et al. (2004). Ward (2000) investigated **Least Squares optimization** of the filter coefficients for multiple head positions and the modeling delay, as previously suggested by Nelson et al. (1992) and Kazutaka et al. (1997). Frequency

invariance of these Least-Square approaches were explored by Parra (2005, 2006).

Convex optimization

Lebret & Boyd (1997) suggested exploiting the convex nature of well-defined pattern synthesis, which greatly speeds up execution time of these optimizations: Global solutions can be found with a computation time that is always small and grows gracefully with problem size. This is further explored in Mabande et al. (2012), Mabande & Kellermann (2010), Mabande et al. (2009).

2.5.4 Beamforming design as convex optimization

Following Lebret & Boyd (1997) and specifically Mabande et al. (2009), consider a linear array with N equidistant sensors. Given a certain target angle ϑ_0, there are N discrete filters $w_n(k)$ with length L that need to be optimized. The propagation delay τ_n based on a sensors distance to the center of the array d_n and the angle ϑ is given to:

$$\tau_n(\vartheta) = d_n \cos\frac{\vartheta}{c} \tag{2.6}$$

Using τ_n and the Fourier Transform of the filters $W_n(\omega)$, one can directly arrive at the response of a Filter & Sum Beamformer (FSB) to frequency ω at angle ϑ:

$$B(\omega, \vartheta) = \sum_{n=0}^{N-1} W_n(\omega) e^{-j\omega\tau_n(\vartheta)} \tag{2.7}$$

In matrix notation, this is greatly simplified to

$$\mathbf{b}(\omega) = \mathbf{G}(\omega)\mathbf{w}_f(\omega)\,, \tag{2.8}$$

with

$$\mathbf{b}(\omega) = \begin{bmatrix} B(\omega, \vartheta_0) \\ \vdots \\ B(\omega, \vartheta_{M-1}) \end{bmatrix}, \qquad \mathbf{w}_f(\omega) = \begin{bmatrix} W_0(\omega) \\ \vdots \\ W_{N-1}(\omega) \end{bmatrix},$$

$$[\mathbf{G}(\omega)]_{m,n} = \begin{bmatrix} e^{-j\omega\tau_0(\vartheta_0)}, \ldots, e^{-j\omega\tau_{N-1}(\vartheta_0)} \\ \vdots \quad , \quad \ddots \\ e^{-j\omega\tau_0(\vartheta_{M-1})}, \ldots, e^{-j\omega\tau_{N-1}(\vartheta_{M-1})} \end{bmatrix}$$

Furthermore, the array's steering vector $\mathbf{d}$ towards a target angle ϑ_{target} is defined as:

$$\mathbf{d}(\omega) = \begin{bmatrix} e^{-j\omega\tau_0(\vartheta_{target})} \\ \vdots \\ e^{-j\omega\tau_{N-1}(\vartheta_{target})} \end{bmatrix} \tag{2.9}$$

The White Noise Gain (WNG) then gives the relation of spatial selectivity in look direction ϑ_{target}:

$$A(\omega) = \frac{|\mathbf{w}_f^T(\omega)\mathbf{d}(\omega)|^2}{\mathbf{w}_f^H(\omega)\mathbf{w}_f(\omega)} \tag{2.10}$$

Least-Squares Beamformer (LSB) optimally approximate a desired response $\hat{B}(\omega, \vartheta)$, discretized into P frequencies and M angles to optimize for:

$$\hat{B}(\omega_p, \vartheta_m) \stackrel{!}{=} \sum_{n=0}^{N-1} W_n(\omega_p) e^{-j\omega_p\tau_n(\vartheta_m)} \tag{2.11}$$

or in matrix notation:

$$\hat{\mathbf{b}}(\omega_p) \stackrel{!}{=} \mathbf{G}(\omega_p)\mathbf{w}_f(\omega_p) \tag{2.12}$$

A set of filters $\mathbf{w}_f$ is to be derived by minimizing the second norm of the difference to the array response. Because this problem space is overdetermined for $M > N$ (number of angles larger than number of sensors), **convex optimization** can be used to solve:

$$\min_{\mathbf{w}_f(\omega_p)} ||\mathbf{G}(\omega_p)\mathbf{w}_f(\omega_p) - \hat{\mathbf{b}}(\omega_p)||_2^2 \tag{2.13}$$

A common constraint is **distortionlessness**, where the array response to the target direction ϑ_{target} is fixed to 1 in order to not alter the wanted signal:

$$\mathbf{w}_f^T(\omega_p)\mathbf{d}(\omega_p) = 1 \tag{2.14}$$

2.6 Helmholtz reciprocity

The following three-dimensional free-field Green's function $G_0(\cdot)$ is a solution to the inhomogeneous Helmholtz equation, which describes wave propagation (Ahrens 2012, Williams 1999):

$$G_0(\mathbf{x}, \mathbf{x}_0, \omega) = \frac{1}{4\pi} \frac{e^{-j\frac{\omega}{c}|\mathbf{x}-\mathbf{x}_0|}}{|\mathbf{x}-\mathbf{x}_0|} \tag{2.15}$$

Here, $\mathbf{x}$ is the *observation point* and $\mathbf{x}_0$ is the *source point*. Observe how both points are fully interchangeable. This means that a source - receiver pair can be swapped, which has for example also been used to efficiently capture HRTF's by Zotkin et al. (2006, 2009). For this thesis it allows the author to utilize a much larger pool of research, as many publications concerning beamforming deal with arrays of receiving sensors instead of speakers.

Chapter 3

Implementation

3.1 Generation of beamforming weights

In order to obtain the set of weights $\mathbf{w}$ for a Least-Squares Frequency Invariant Beamformer (LSFIB), convex optimization was used in an approach which relies on the ideas of Mabande et al. (2009) and the great CVX package and its documentation by Grant & Boyd (2008, 2014). Due to the nature of the toolbox, the actual optimization encompasses only a rather small set of instructions and is preceded by long and careful preparation of data.

3.1.1 Data preparation

For an optimization over M angles, N sensors and P frequencies, a matrix $[\mathbf{D}]_{M,N}$ is calculated holding the distances between each sensor and each angle. Then, $[\mathbf{G}]_{M,N,P}$ is created as $e^{-j\omega_p \mathbf{D}_{m,n}/c} = e^{-j\omega_p \tau_{m,n}}$. Two slices of this multi-dimensional matrix are of special importance:

Firstly, $\mathbf{G}(\vartheta_{target})$ represents the steering vector $\mathbf{d}$. Secondly, $\mathbf{G}(\vartheta_{\mathbf{stop}})$ represents the stop vector $\mathbf{G}_{stop}$. Additionally, a parameter null-width NW was introduced that, which specifies a broader selection of directions that are constraint for a null response $\vartheta_{stop} \pm \frac{NW}{2}$. This is important for stable crosstalk cancellation, as the optimization is forced into null constraints over a larger

angular space at the averted ear.

Finally, the ideal array response $\hat{b}$ needs to be generated. This is straightforward, as the beamformer is frequency-invariant and ideally has no response everywhere except into the target direction ϑ_{target}. This means $\hat{b}$ can be set to a vector of M zeros with a pulse of the shape `[0.2360, 0.4719, 0.7079, 0.9900, 1.0000, 0.9900, 0.7079, 0.4719, 0.2360]` at the index corresponding to ϑ_{target}.

3.1.2 Optimization

As mentioned, the CVX toolbox (Grant & Boyd 2014) was used to minimize $\min_{\mathbf{w}_f(\omega_p)} ||\mathbf{G}(\omega_p)\mathbf{w}_f(\omega_p) - \hat{\mathbf{b}}(\omega_p)||_2^2$. After many runs and subsequent simulations, it was decided to NOT employ the distortionless constraint of $\mathbf{wd} = 1$. This allows for harder constraints on the stop direction ϑ_{stop} while the resulting uneven frequency spectrum at ϑ_{target} could be fixed using equalization of the source material. Instead, the array response in the stop direction was constraint to $||\mathbf{G}_{stop}\mathbf{w}||_2^2 <= 0.01$ ($\approx -40\,\mathrm{dB}$). The full optimization statement as implemented in matlab is shown in listing 3.1.

Listing 3.1 Matlab implementation of convex optimization to generate the weights **w** using the CVX toolbox

```
w = zeros(N, P);
for f=1:P
    cvx_begin quiet
      variable wf(N) complex
      minimize( norm ( G( :,:,f ) * wf - b, 2) )
      subject to
        norm( Gstop( :,:,f ) * wf ) <= 0.01;
    cvx_end
    w(:, f) = wf;
end
```

3.1.3 Optimization results

Executing Listing 3.1 with target direction $\vartheta_{target} = -6°$ and stop direction $\vartheta_{stop} = 6°$ over a frequency range of $f_{opt} = 1\,\text{kHz} - 8\,\text{kHz}$ with $f_s = 44100$ yields $\mathbf{w}$ as a set of 8 arrays (one per speaker) holding complex coefficients in the frequency domain. Applying the inverse Fourier transform on $\mathbf{w}$ generates 8 IRs of filters that aim at the ears of a human listener located approximately 1 m in front of the array's broadside. The coefficients for frequencies outside $f_{opt} = 1\,\text{kHz}-8\,\text{kHz}$ were set to 0. Figure 3.1 shows these beamforming filters in time (top) and frequency domain (bottom). The time domain representations clearly show the familiar *sinc*-like shapes but only the frequency spectrum reveal the strong band-pass over the optimized frequencies.

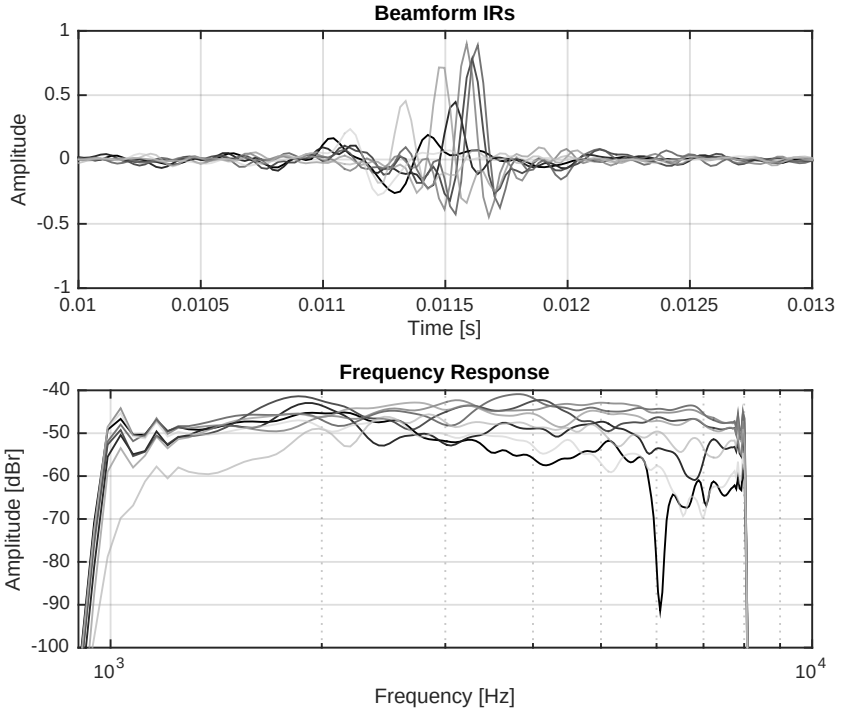


Figure 3.1 Complex beamforming weights in time (top) and frequency (bottom) domain generated for $\vartheta_{target} = -6°$, $\vartheta_{stop} = 6°$, $f_{opt} = 1\,\text{kHz} - 8\,\text{kHz}$.

The White Noise Gain (WNG) $A = \frac{|\mathbf{w}_f^T \mathbf{d}|^2}{\mathbf{w}_f^H \mathbf{w}_f}$ (see Equation 2.10) was not only calculated for the target direction but also for the stop direction, as shown in Figure 3.2.

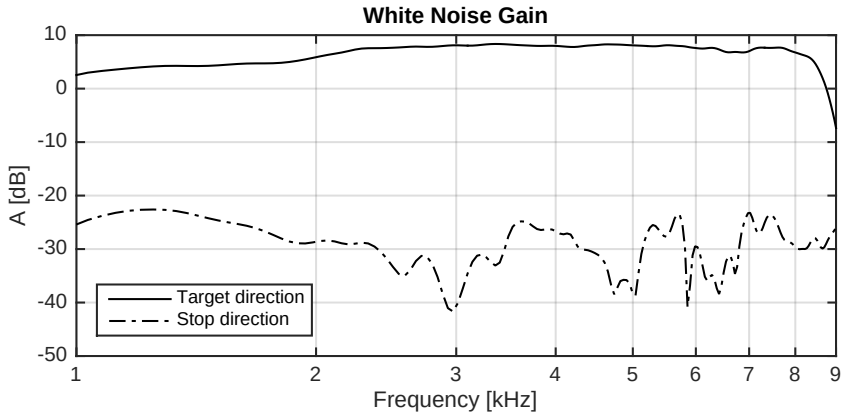


Figure 3.2 White Noise Gain (WNG) for target direction $\vartheta_{target} = -6°$ and stop direction $\vartheta_{stop} = 6°$.

A much better visualization of the arrays performance is the array response, as shown in the beam pattern in Figure 3.3. Here, the target direction $\vartheta_{target} = -6$ degree is marked by a solid line, the stop direction $\vartheta_{stop} = 6$ degree by a dashed line. The robustness against small movements of the head is visible as a continuous strip of low energy around the stop direction and a similar area of consistently high energy in the target direction. This seems to support the idea of gentle collapse when moving outside the optimal position.

A similar conclusion can be drawn from Figure 3.4, which shows the broadside emission pattern of the array in polar space. Again, the target direction $\vartheta_{target} = -6$ degree is marked by a solid line and the stop direction $\vartheta_{stop} = 6$ degree by a dashed line.

3.1.4 Time-domain windowing

The results of the convex optimization is a set $\mathbf{w}$ of transfer functions that need to be converted to stable IRs in order to be used in real-time convolution. To represent a well behaved filter in the time domain, they need to be windowed to fall to 0 at the beginning and the end of the filter. To achieve this, a raised cosine was multiplied with the IRs, as shown in Figure 3.5. Choosing long enough filters allows the contained *sinc*-functions to fall to near zero,

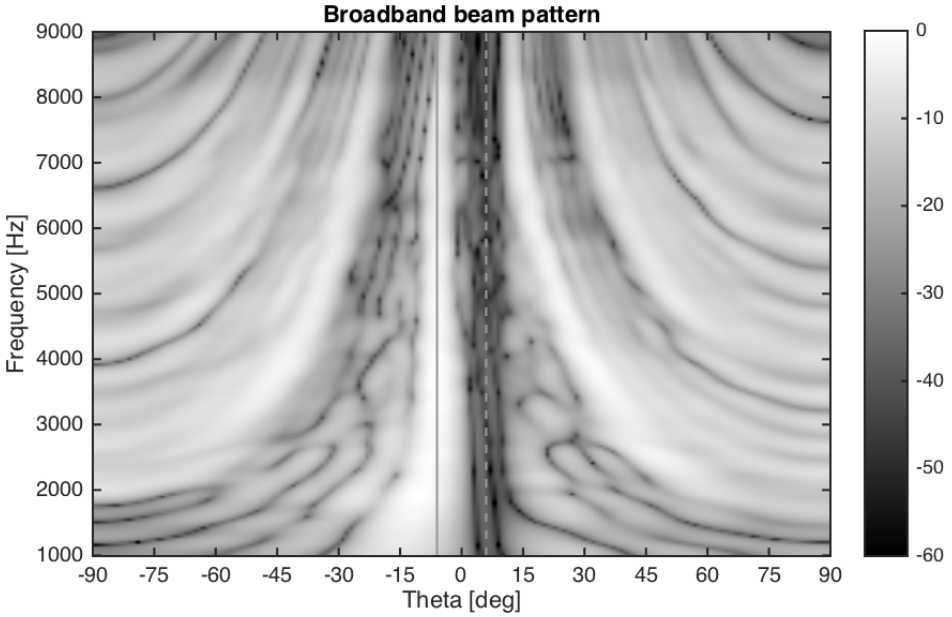


Figure 3.3 Broadband beam-pattern of 8 speaker array with 14.4 cm spacing. Target angel is $\vartheta_{target} = -6°$ (marked by sold line), stop angle is $\vartheta_{stop} = 6°$ (marked by dashed line) with a null width of 9°. Frequency range of optimization was $f_{opt} = 1\,\text{kHz} - 9\,\text{kHz}$ with $L = 1024$ points.

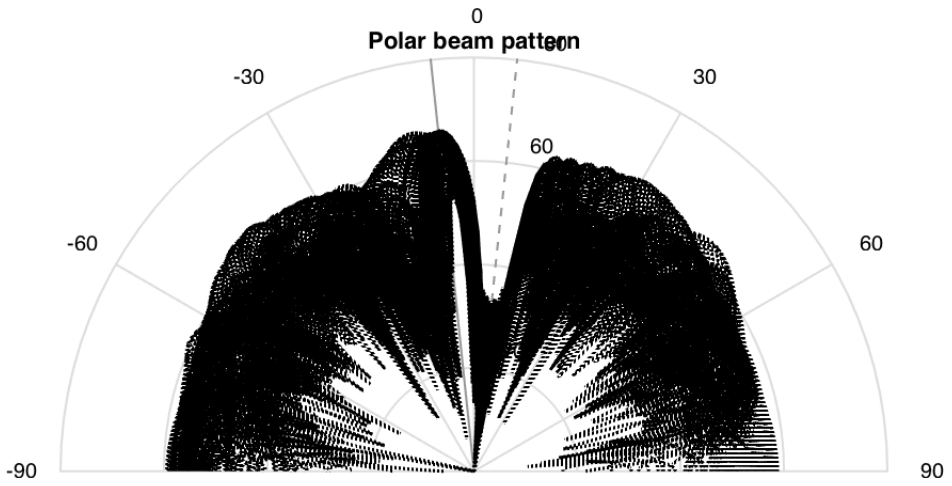


Figure 3.4 Polar beam-pattern of 8 speaker array with 14.4 cm spacing. Target angel is $\vartheta_{target} = -6°$ (marked by sold line), stop angle is $\vartheta_{stop} = 6°$ (marked by dashed line) with a null width of 9°. Frequency range of optimization was $f_{opt} = 1\,\text{kHz} - 9\,\text{kHz}$ with $L = 1024$ points.

decreasing the errors due to windowing (see Figure 3.5 bottom).

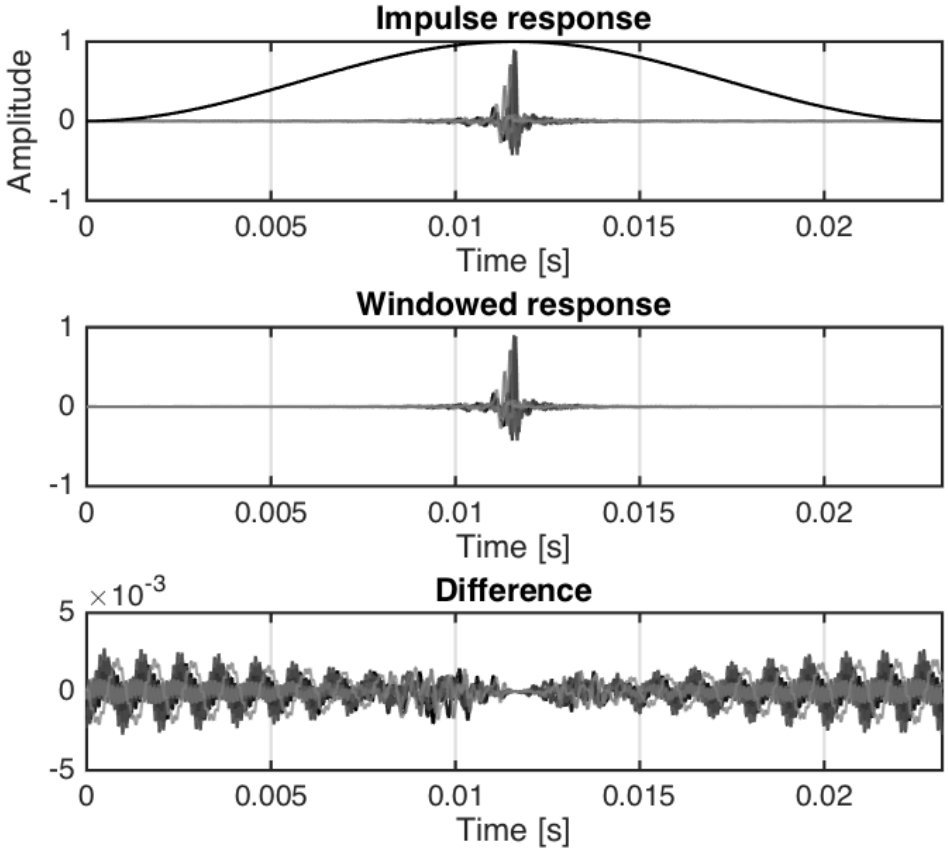


Figure 3.5 Windowing of normalized impulse responses in the time domain using a raised cosine.

The results of such windowing in the time domain is illustrated in Figure 3.6, which shows the beam pattern of the difference between beamforming with and without windowing. With a length of $L = 1024$, no significant deviations are visible.

3.2 Real-time playback

For later evaluation, a real-time system of the calculated beamforming responses was implemented using Max/MSP.

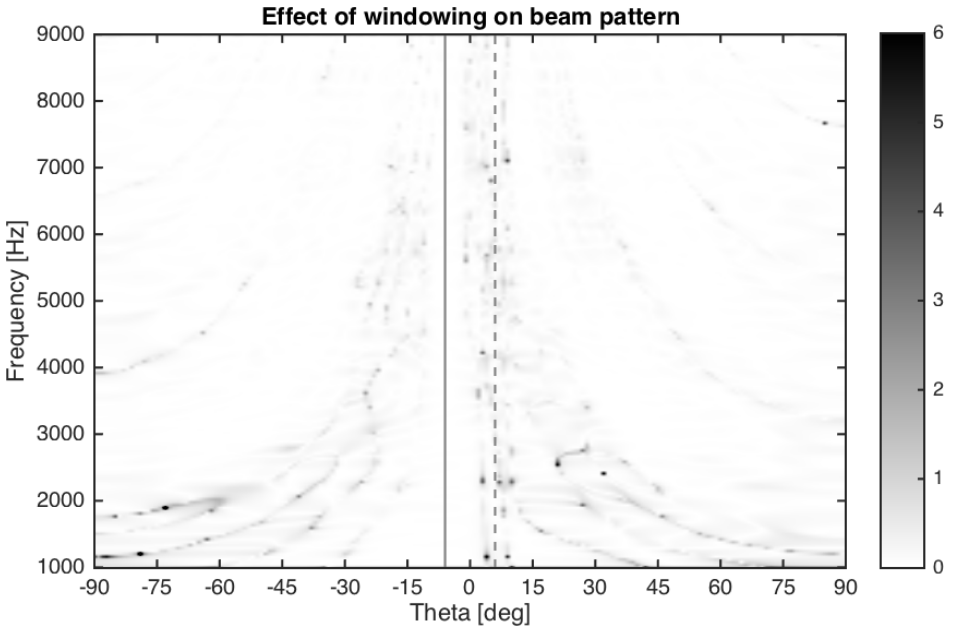


Figure 3.6 Effects of windowing on broadband beam-pattern as absolute difference in dB

3.2.1 Handling of frequency bands

Due to the expected limited frequency range of any beamforming implementation (confirmed by simulation in Section 4.1), other methods could be used simultaneously to include more of the audible spectrum. It was therefore decided to split the system into 4 frequency bands to be handled differently.

Sub-Bass (20 Hz to 250 Hz)

Frequencies this deep generally don't carry many location cues for humans, therefore this band was delegated to a single subwoofer, especially because the small speakers necessary for beamforming generally don't go this low in frequency response.

Bass (250 Hz to 1000 Hz)

Both music and speech signal will carry a lot of content in this frequency band. Both male and female voices have their fundamental frequency in this range so a decent reproduction is crucial. Unfortunately, while a decent channel

separation down to 600 Hz is theoretically possible using beamforming, this range is highly susceptible to random changes in gain or phase as seen in simulation in Section 4.1.

Instead, RACE (see Section 2.4.2) was used to suppress the crosstalk as good as possible. Initially, the parameter $\tau = \frac{\Delta d}{c} \approx \frac{0.082}{c} = 240\,\mu s$ was set due to the fixed listening position. After preliminary measurements using the KEMAR mannequin and subjective judgement, a final value of $\tau = 190\,\mu s$ was fixed. For the attenuation, an optimum was experimentally found at $G = -6\,dB$. If not a perfect cancellation method, the effect of decorrelating the ear signals helps to envelope the listener.

RACE was implemented after Figure 2.5 in Max/MSP using *gen~* and cosine interpolation between samples, see Figure 3.7.

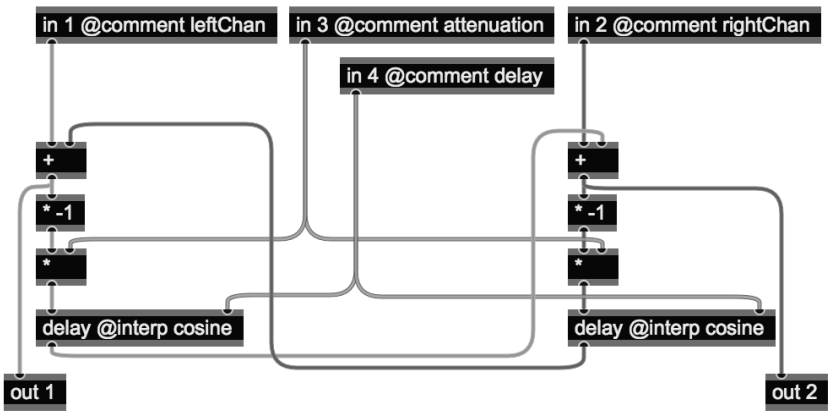


Figure 3.7 Recursive Ambiophonic Crosstalk Elimination (RACE) implementation in Max/MSP's *gen~* object.

Mid-Range (1000 Hz to 8000 Hz)

This is the frequency range where the beamformer can act comfortably and is robust against head movements and random speaker gain.

High-End (8000 Hz to 20 000 Hz)

Due to the absence of robust beamforming or general cancellation schemes for the higher frequency range, left and right channel are simply played back at the respective end of the speaker array. Shadowing of the head will generally work better as wavelengths become very small compared to the diameter of the listener's head. Luckily, little localization cues are used in this frequency range.

3.2.2 Crossover filter

To separate the frequency bands laid out in Section 3.2.1, fitting crossover filters have to be selected carefully. Generally, a crossover filter is a set of high- and low-pass filters with identical cutoff frequency, that split a signal into two frequency bands. The cutoff frequency f_C is regarded as the frequency where the filtered spectrum has dropped 3 dB below the original amplitude. Higher order filters exhibit a steeper slope with an increase of 6 dB/oct per order. Crossovers find wide use in many audio applications, for example when feeding the appropriate frequency bands to multi-driver speaker.

The most common filter is the Butterworth filter, due to its minimal ripple in the pass band and its relatively low complexity. Indeed, designing overlapping Butterworth high-pass and low-pass filters of the same cutoff frequency produces the correct frequency response. The problem arises from the asymmetric radiation pattern due to phase alignment at the crossover point, see Figure 3.8, left. While the on-axis response exhibits the desired 0 dB response, only slight vertical miss-alignment will either drop the listener into a a large cancellation hole or increase the amplitude on a peaking axis, as per Bohn (1989).

To avoid this problem, a $4th$ order Linkwitz-Riley crossover was used, which exhibits the same 0 dB amplitude on axis but has a symmetrical radiation pattern without a peaking axis. Small movements of the listeners head do not

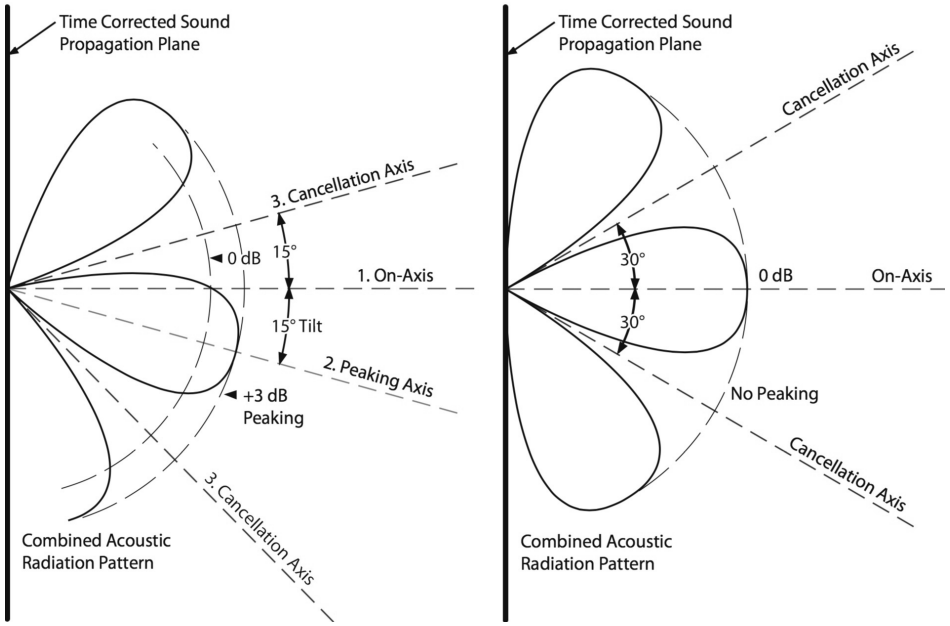


Figure 3.8 Radiation response of two different crossover filter implementations. Left: Butterworth. Right: Linkwitz-Riley. From: Bohn (1989)

result in large changes in amplitude, which is important for overall robustness. A stereo 4-band crossover with variable cutoff frequencies was implemented in Max/MSP with the filter design of the Jamoma Toolbox by Place & Lossius (2006). The correct reconstruction was validated by preliminary sweep measurements of the crossover.

Signal flow

Figure 3.9 shows the complete signal flow of the Max/MSP patch written for the system. For the HRTF convolution, a single audio source has to be convolved with both ear's HRTFs corresponding to the same direction, resulting in two output signals for the left and the right ear. Each set of HRTFs is compromised of 720 channels, spanning 360° of left and right ear signals interleaved. All convolution is done using the **timeconvolve** external by Harker (2011), which was modified to allow the selection of up to the $720-th$ channel of a buffer for convolution.

This pair of binaural signals was then fed into a 4-band stereo crossover

with a total of six 4-th order Linkwitz-Riley filter. The sub-bass and the high-end were directly mapped onto corresponding speakers, the bass-band was first routed through the RACE module and then out to the speakers. The mid-band was first convolved with the corresponding beamforming IRs, which could then be added up to obtain 8 output channels. These were then convolved with the speaker equalization filters before routed to the speakers.

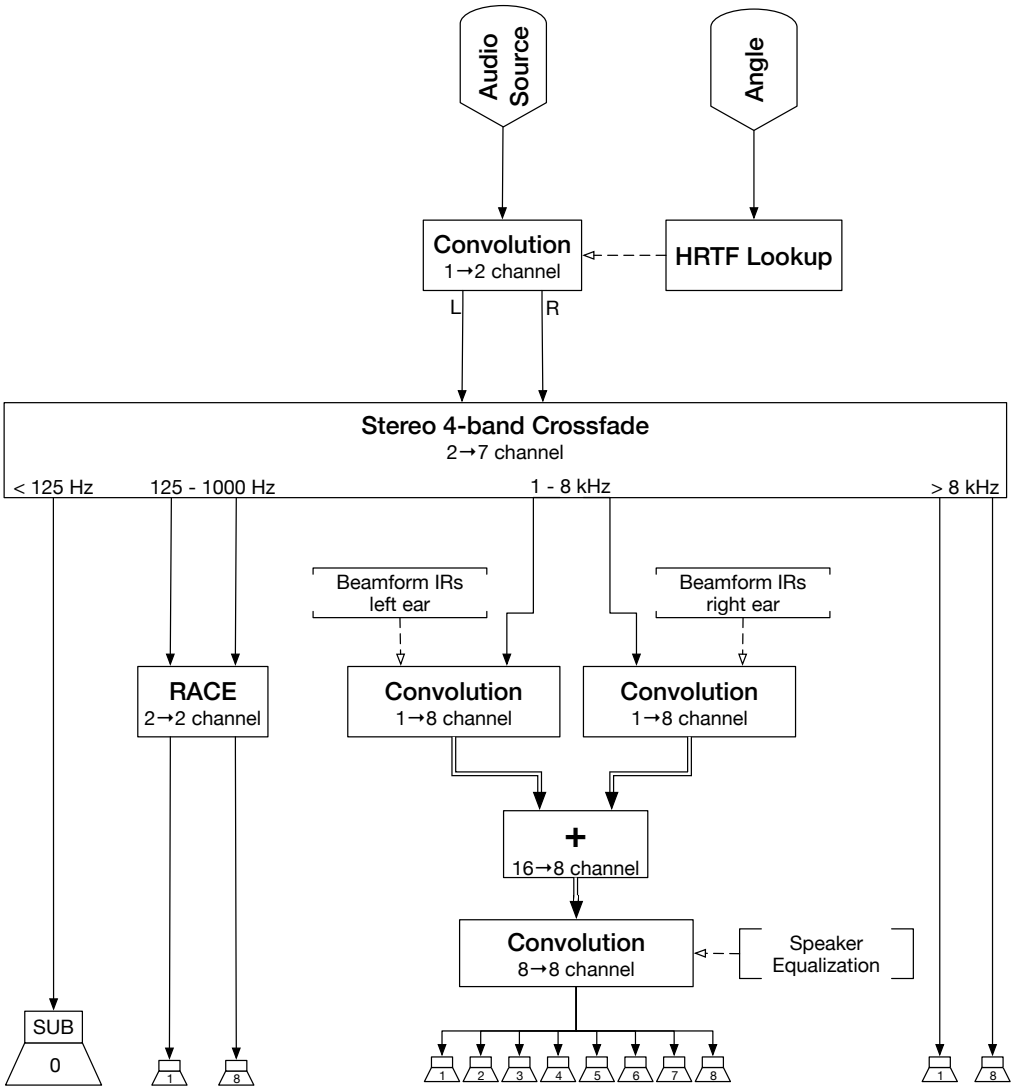


Figure 3.9 Signal flow of the implemented system

Chapter 4

Simulation and measurements

4.1 Simulation

To predict performance and ultimately feasibility, extensive simulations of the the developing sound field inside the listening area have been performed. Both mono-frequent propagation over a larger area as well as broadband behavior at specific points were evaluated.

4.1.1 Sound field

Generally, speakers may be idealized as radial point sources that behave like harmonic oscillators with radial frequency $\omega = 2\pi f$ and distance of the receiver of r:

$$p(\omega) = \frac{1}{4\pi r} e^{i \cdot \frac{\omega}{c} r} \,, \tag{4.1}$$

For a mono-frequent analysis optimized frequency ω_k, the corresponding complex factors $\mathbf{w}_k$ are used to weight each speaker. The pressure of the emerging field is the superposition of the pressure of all N speakers:

$$p(\omega_k) = \sum_{n=1}^{N} \frac{w_{k,n}}{4\pi r_n} e^{i \cdot \frac{\omega_k}{c} r_n} \tag{4.2}$$

4.1.2 Scattering

To also take the effects of the listener's head into account, the scattering of a radial, acoustically hard sphere had to be calculated, which is possible using spherical harmonics. The following is a quick outline of the necessary math and relies heavily on Ahrens (2012) and Ahrens & Spors (2011). Please refer to the nomenclature for an overview of the symbols used.

Generally, a sound field in the presence of an acoustically non-transparent object may be interpreted as the summation of the incidental field emitted by some speakers $S(\mathbf{x}, \omega)$ and the field occurring due to scattering at the object $S_{scat}(\mathbf{x}, \omega)$:

$$S_{total}(\mathbf{x}, \omega) = S(\mathbf{x}, \omega) + S_{scat}(\mathbf{x}, \omega) \tag{4.3}$$

Because the incidental field is simulated much quicker and with great precision as described in Section 4.1.1, only the emerging scattering has to be calculated using spherical harmonics. This greatly reduces the order n needed to obtain the simulated field of decent resolution.

The field of a spherical wave emitted by a speaker at point (r_s, α_s, β_s) is:

$$\breve{S}_n^m(\omega) = (-i)\frac{\omega}{c} h_n^{(2)}(\frac{\omega}{c} r_s) Y_n^{-m}(\beta_s, \alpha_s) \tag{4.4}$$

The scattered sound field is then given by:

$$\breve{S}_{n,scat}^m(\omega) = -\frac{j_n^{'}(\frac{\omega}{c}A)}{h_n^{(2)'}(\frac{\omega}{c}A)} \breve{S}_n^m(\omega) \tag{4.5}$$

The exterior field $S_e(\mathbf{x}, \omega)$ of a single source is the field expansion over degrees n and orders m:

$$S_e(\mathbf{x}, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \breve{S}_{n,e}^m(\omega) h_n^{(2)}(\frac{\omega}{c} r) Y_n^m(\beta, \alpha) \tag{4.6}$$

As we are only interested in the scattered field $S_{scat}(\mathbf{x}, \omega)$, we can directly calculate:

$$S_{scat}(\mathbf{x}, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \breve{S}^m_{n,scat}(\omega) h^{(2)}_n(\frac{\omega}{c} r) Y^m_n(\beta, \alpha) \tag{4.7}$$

The final field caused by the full array of speakers is the superposition of all scattered responses.

4.1.3 Results

Figure 4.1 shows the emerging stationary field (left side) and the pressure distribution (right side) invoked by an array situated at $[0, 1]$ at increasing frequencies. The beam was tuned to illuminate only to the left ear of the listener, simulated as a acoustically hard sphere situated at the origin. For a precise frequency spectrum at the listeners ears, please refer to Figure 4.2 or 4.3.

Especially the pressure distribution on the right allows for a great visualization of beamforming. It can be observed that even including the scattering around a head, beamforming can be regarded as a rather robust crosstalk cancellation scheme, as the targeted ear is guaranteed to receive a decent signal while the averted ear is still surrounded by very low sound pressure.

4.1.4 Robustness

In a real world system, many usually idealized variables may deviate substantially or even fluctuate during operation. These may include but are not limited to:

Speaker position Speaker position may vary slightly both in x and y direction, even when fixed on a rig.

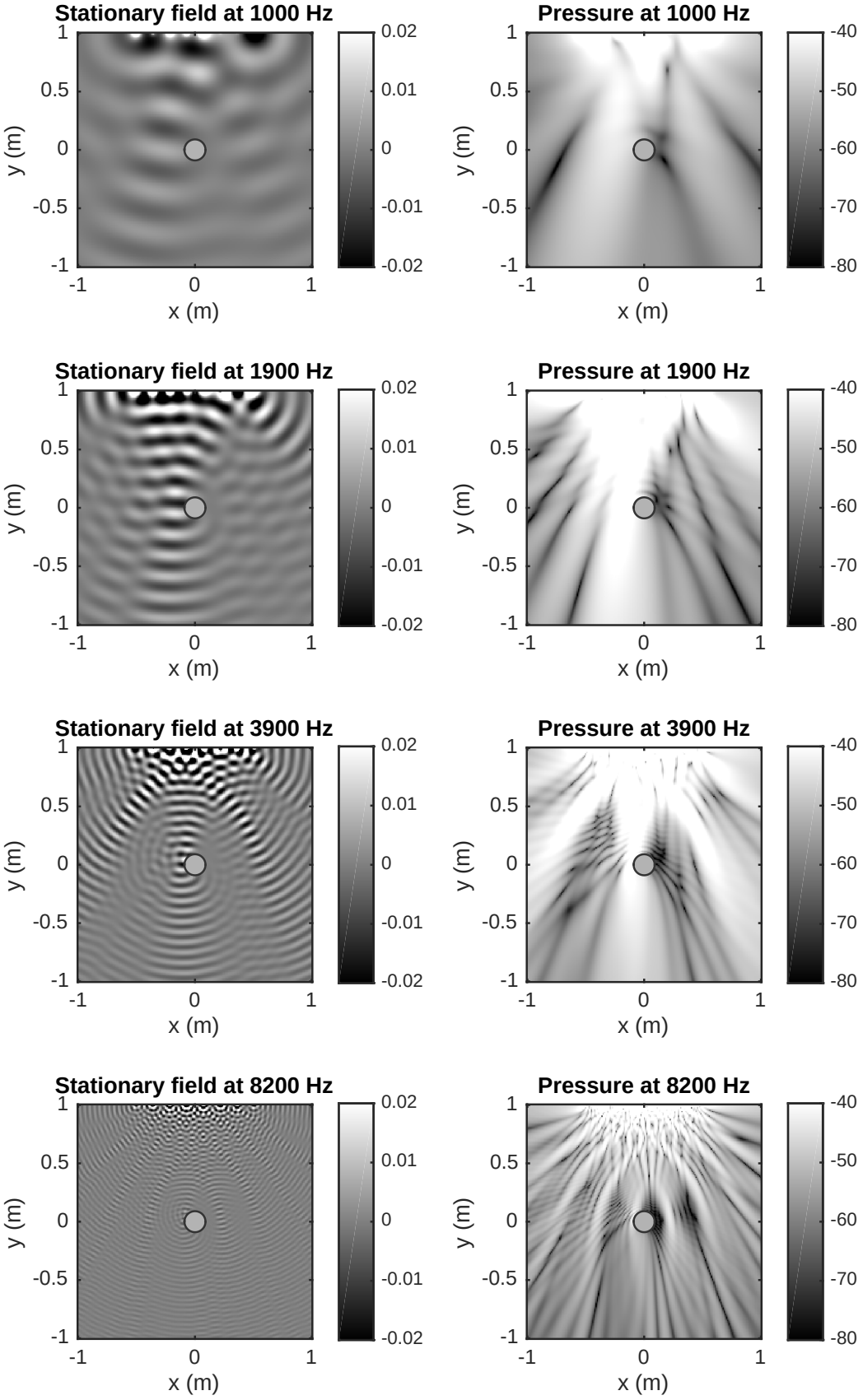


Figure 4.1 Stationary field (left) and sound pressure (right) at various frequencies over a $[2\,\mathrm{m} \times 2\,\mathrm{m}]$ area. The gray dot at $[0, 0]$ represent the listeners head, modeled as an acoustically hard sphere with a diameter of 18 cm.

Listener position Holding a static position is impossible, this includes both translation and rotation of the head. Furthermore, anthropomorphic features (such as ear position) vary between listeners.

Speaker transfer function Both magnitude and phase spectrum of each speaker will vary due to factory tolerances of electrical components and drivers.

Speaker radiation pattern Similarly, emission pattern will vary slightly due to variance in chassis construction.

Influence of the room Important acoustical properties of the room such as reverb time and early reflections may vary.

Transmission medium Although usually small in effect, temperature, humidity and air composition are not constant.

One important aspect of the proposed system in general and beamforming in particular is the robustness against small errors listed above. This is also one of the main advantages over other transaural solutions, in particular ones that employ filter inversions.

Intra-system

The simulation described in Section 4.1.1 and 4.1.2 has been rerun with random offsets in gain and phase shift on the optimal weights. This simulates the effects of non-ideal signal emission due to components in the transmission path (amplifiers, speaker position, speaker components).

Figure 4.2 shows the results of slight variations of gain ($g_{var} \in \pm 0.3\,\mathrm{dB}$) and phase ($\varphi_{var} \in \pm 0.001\frac{\omega}{c}$) to the driving function in gray ($w_{n,var} = (|(w_n)| + g_{var})e^{-j(\angle w_n + \varphi_{var})}$), the optimal response is drawn in black as reference. While the frequency response at the targeted ear stays almost completely unaffected by the added errors, the averted ear gains about $10\,\mathrm{dB} - 15\,\mathrm{dB}$, reducing the crosstalk. Still, a large gap between the ears remains with no outliers-off between 1 kHz and 9 kHz. It can therefore be assumed, that beamforming using

the available hardware (8 speakers with 14.4 cm spacing) may be employed for a frequency range from at least $f_{min} = 1000\,\text{Hz}$ up until $f_{max} = 9000\,\text{Hz}$.

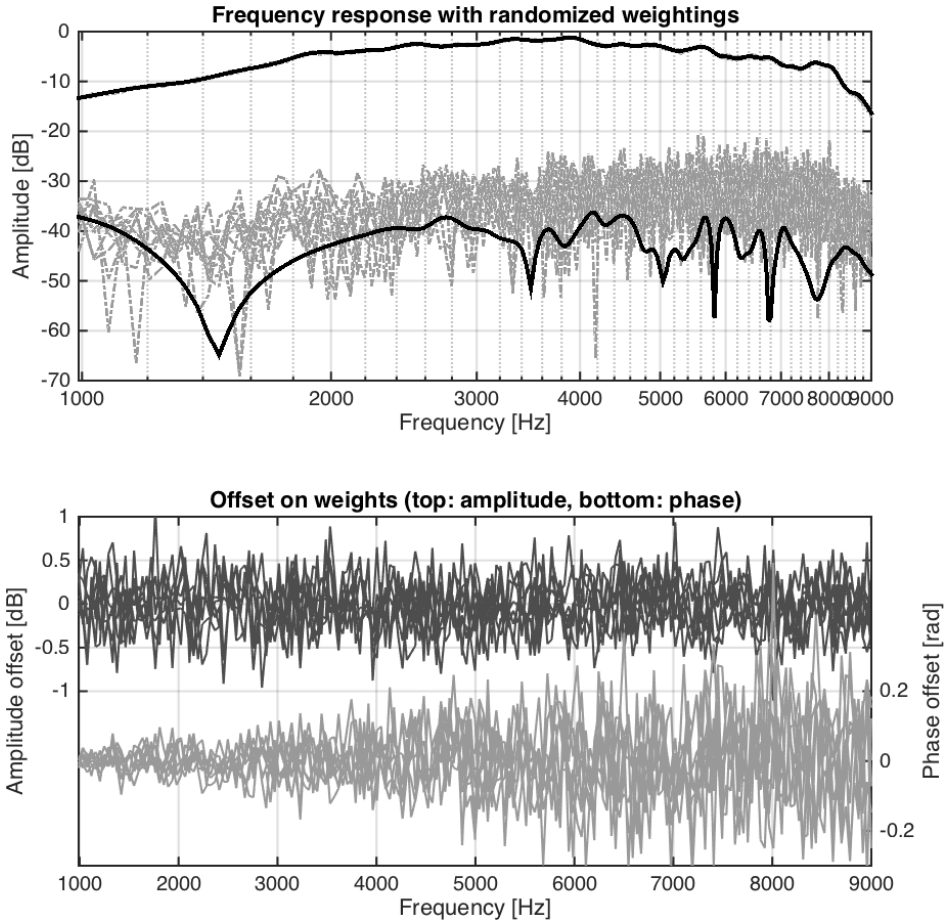


Figure 4.2 Frequency response at both ears when modifying the complex weights (top), the corresponding offsets in gain and phase are shown below.

While calibrating the array (see Section 4.2.2) helps compensating for the larger variances between speakers, it seems obvious that the performance of the array will never come close to idealized point sources used in the simulation.

Listener position variance

The robustness of the array against varying ear position (both lateral and rotational) is arguably more important than intra-system variance, as it cannot be directly compensated for. Figure 4.3 shows the frequency spectrum over a variety of ear positions around the assumed positions at the side of the sphere.

The middle graph shows the average difference between the ears. The sharp decrease of signal at the targeted ear at higher frequency strongly supports limiting the beamforming approach to a frequency range of 1 kHz − 8 kHz.

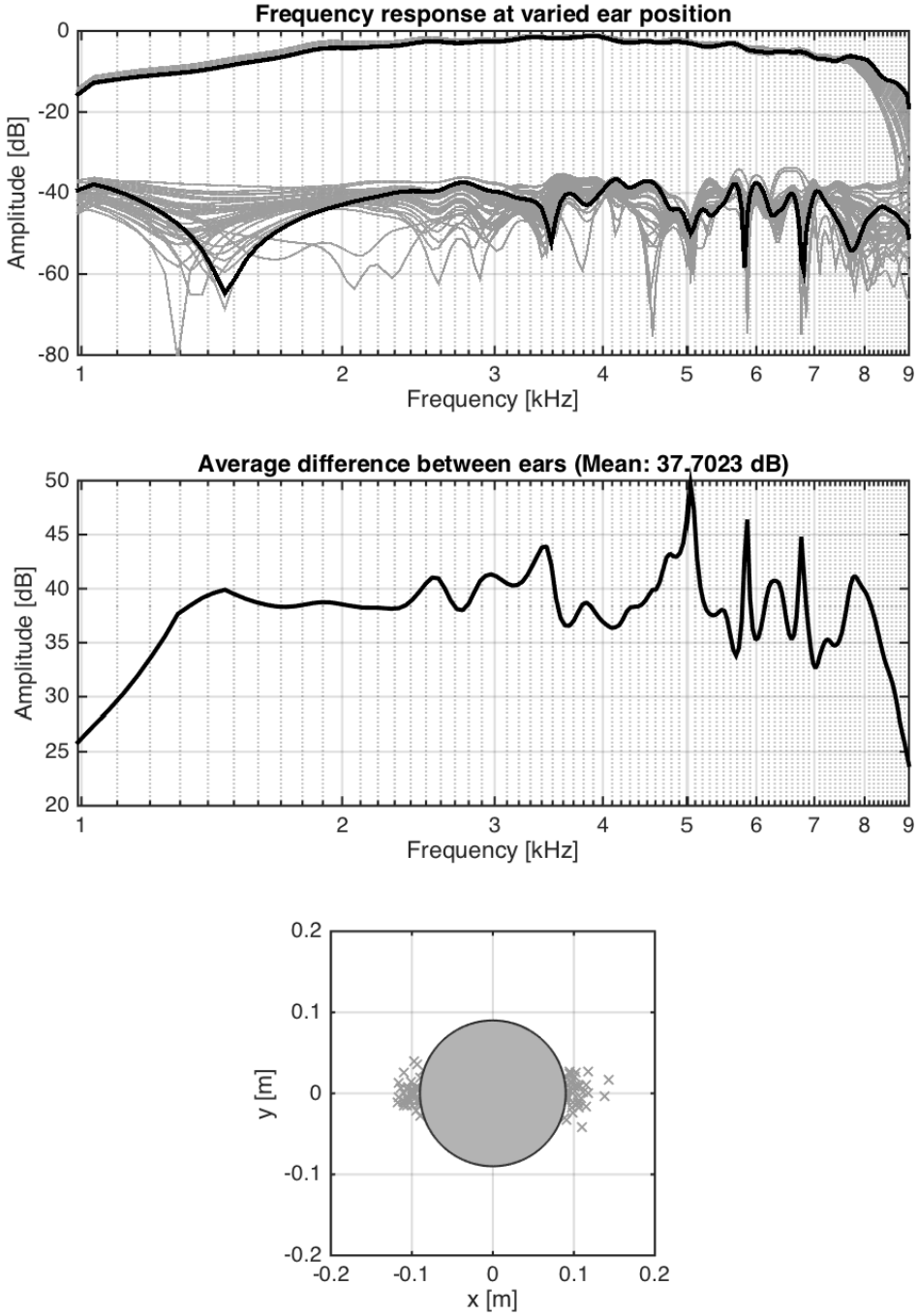


Figure 4.3 Frequency spectrum of 40 random ear positions around the assumed positions at the side of the head. Middle plot shows the average difference between ears, bottom plot shows the simulated positions.

4.2 Measurements

In order to ensure the delivery of the correct localization cues to the ears, the whole system has to be tuned for a reasonably flat amplitude response at the ears. Therefore, calibration and verification measurements were conducted using a dummy head.

4.2.1 Setup

Eight **Fostex PM0.4**, which match the previously simulated dimensions of 14.4 cm, were arranged and slightly angled so to aim at a **G.R.A.S KEMAR Hear & Torso Simulator** (large ears) in 1 m distance. The ear signals were recorded using a **Roland Quad-Capture**, which was connected to a computer via USB. From the other USB port, a **RME MADIface** was used to feed the eight speaker signals into a **RME ADI-648**, which converted the MADI signals into ADAT. These were then converted to analog using a **Ferrofish A16 MK-II** and connected to the speakers using a DB-25 snake. Figure 4.4 shows the setup used for recording the measurements.

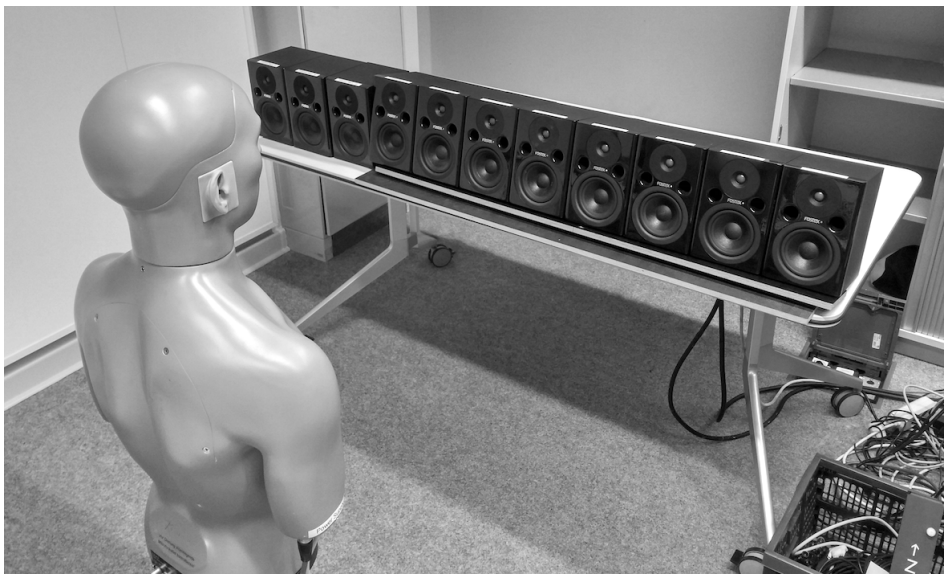


Figure 4.4 Setup used for dummy-head measurements. Only the 8 angled speakers were used.

On the computer, Max MSP was running several externals from the excellent HISSTool by Harker & Tremblay (2012) to measure the 2x2 transmission matrix of left/right channel to left/right ear. A logarithmic frequency sweep (Müller & Massarani 2001) of 30 s from 100 Hz to 20 kHz was used.

4.2.2 Array calibration

To minimize intra-system variance, each speaker was carefully measured at a distance of 1 m using a **Beyerdynamic MM-1** electret condenser measurement microphone, see Beyerdynamic (2014) for the spec-sheet. The recorded transfer functions were inverted using the HISSTools (Harker & Tremblay 2012) and could then be used as input for a convolution at each corresponding output. The time and frequency domain of the speaker equalization is shown in Figure 4.5.

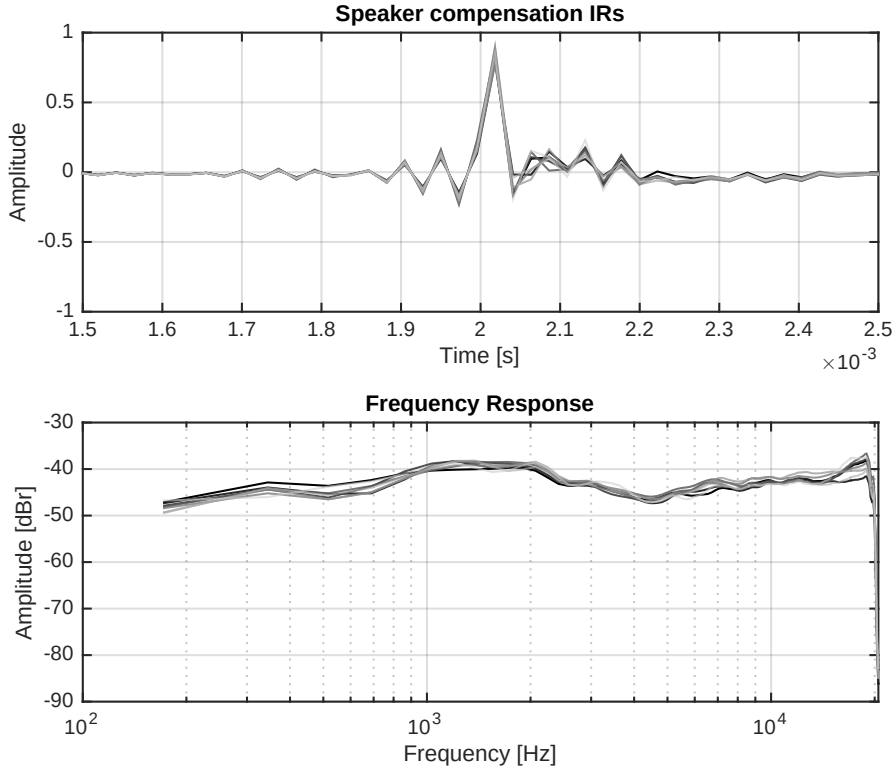


Figure 4.5 Normalized impulse responses (top) and transfer function (bottom) of inter-speaker equalization for the eight Fostex PM0.4 speakers used for validation and the user study.

4.2.3 Results

Looking at the transfer functions between the left / right channel and the left / right ear of Figure 4.6 (top), the large influence of the outer ears can be seen, as this measurement essentially produces a HRTF recording. The large boost around 4 kHz matches similar records with this KEMAR head, such as in Wierstorf et al. (2011), amplified by the non-linear frequency spectrum of the unequalized system.

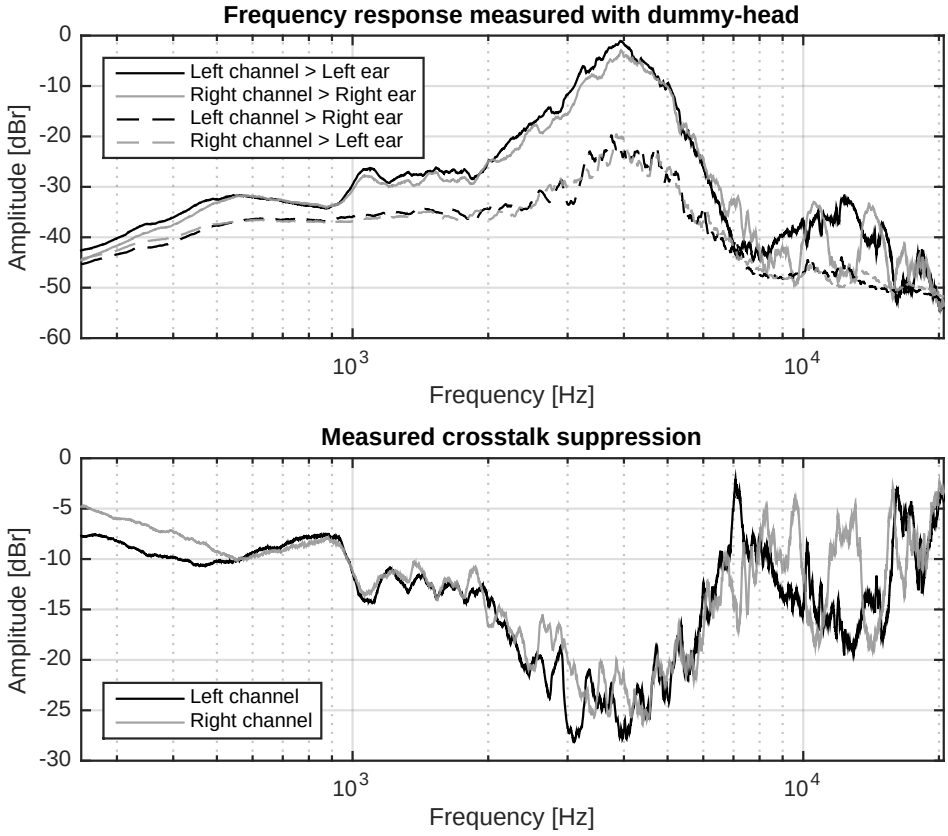


Figure 4.6 Transfer functions between left/right channel and left/right ear (top) and the crosstalk cancellation of left and right channel (bottom)

A steady separation of channels can been seen when looking at the actual crosstalk suppression as the difference between transmission and leakage (bottom). Some of the irregular pattern might be attributed to the specific shape of the ears but nevertheless, a separation of roughly 10 dB − 15 dB can be observed between 20 Hz and even 20 000 Hz. It seems that the natural

shadowing of the head provides decent suppression for the highest frequencies above 8 kHz, where no active crosstalk cancellation is active.

Lastly, the frequency response at the ears is equalized to deliver a reasonably smooth frequency spectrum to the ears as shown in Figure 4.7. This was done by first adjusting the gain between the four frequency bands of the system and then equalizing the bands itself using a combination of Max/MSP's *filtergraph* and *cascade* elements to generate large cascaded structures of biquad filters forming an EQ.

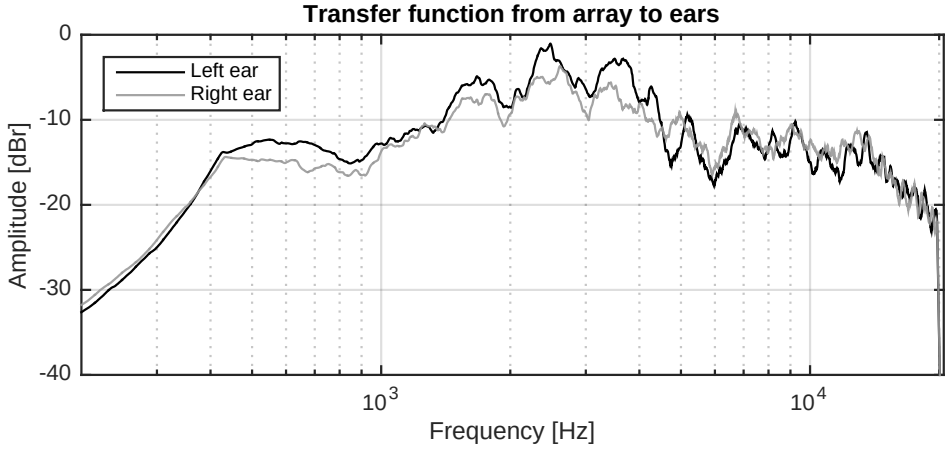


Figure 4.7 Broadband frequency spectrum delivered to the ears after calibration and equalization.

Chapter 5

User study

5.1 Goals

The main goal of the system under test is to deliver a set of HRTFs to the listener's ear to enable the use of spatial audio without the use of headphones. The performance of the array is judged by comparing the errors in localization of binaural material to delivery with calibrated headphones, which can be treated as a ground truth.

5.2 Setup

As stated in Section 5.1, participants should judge the position of sources in binaural audio presented to them via the beamforming array. The following section describe how the study was conducted.

5.2.1 Methodology

The study was split into three parts per participant as follows:

HRTF selection

A short selection process was necessary, to assign the best possible HRTF set to each subject. After some precursory tests, the approach presented in Seeber & Fastl (2003) was used to quickly reduce the number of possible HRTFs, which were then A/B compared in a Swiss-style tournament, as suggested by Iwaya (2006). A comparison between the two can be found in Grasser et al. (2014).

Firstly, each participant was presented with an identical audio source that was convolved with one of 16 HRTF sets, randomly selected of a larger collection recorded by Brinkmann et al. (2015, 2013). The participants were able to seamlessly switch between all 16 stimuli and were asked to select their four favorites that are best perceived as a source that slowly rotates around the head according to the following criteria:

- Constant planar height on ear level
- Constant distance
- Constant loudness
- No coloration
- Smooth movement

The four winning sets were then pitted against each other in A/B comparisons, where the subject had to repeatedly select the better fit according to the same criteria. A set was dismissed after loosing twice, which quickly led to a winning set that hopefully best matched the participants own HRTF.

Other selection methods such as measuring morphological features suggested by Schonstein & Katz (2010) were not tried.

Localization experiment

Front-back ambiguity is a big issue when judging source position in the horizontal plane, as all positions have a position of identical ITD and ILD cues (see

Section 2.1). Due to the missing dynamic cues of the static binaural synthesis, it was decided to instead radially move the sources about a position in the horizontal plane, following Wightman & Kistler (1999) and Nykänen et al. (2013).

The subject's task was to localize a sound source that was slowly moving around an angle $\vartheta_{test} \in [0°, \pm15°, \pm35°, \pm60°, \pm90°, \pm120°, \pm155°, \pm165°, 180°]$ and identify the corresponding circular segment. Figure 5.1 shows the graphical interface presented to the user. The binaural stimulus was presented in one session via headphones, in another session via the array, the order was randomized between subjects.

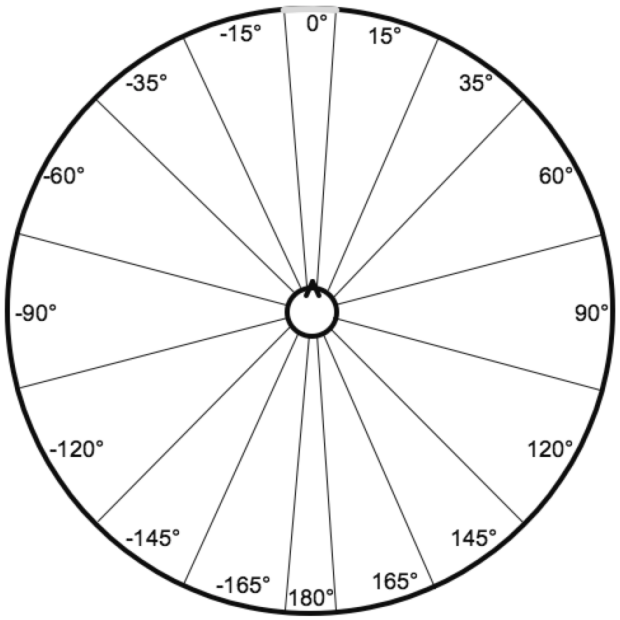


Figure 5.1 GUI of the localization experiment, the mark at the top indicate the segment the subject is currently hovering over.

Survey

Lastly, all subjects were asked to fill out a short survey on demographics (gender, age, technical background, knowledge on binaural technique, experience listening to binaural material, understanding of system before the experiment) and subjective rating on a 5-point scale with qualitative descriptions between

conditions "WITH headphones" and "WITHOUT headphones" concerning the dimensions shown in table 5.1.

Dimension	Low anchor	High anchor
Difficulty	Very hard	Very easy
Timbre	Not natural	Very natural
Externalization	Inside head	Far outside head
Spaciousness	Not at all	Very spatial

Table 5.1 Survey dimensions and scale anchors

5.2.2 Procedure

After a quick explanation and necessary paperwork, the subjects were seated in the acoustically transparent booth and the HRTF selection process with headphones was started. Before starting with the localization experiment, a short training round was conducted to familiarize the subject with task, stimulus and GUI. Condition (with/without headphones) sequence was randomized between participants. Short breaks, especially between rounds, were strongly encouraged. Lastly, the participants were asked to complete the survey.

Completion time was estimated to be 10 min (HRTF selection) + 2 (conditions) × 15 min (localization experiment) + 5 min (survey) = 45 min.

5.2.3 Equipment and layout

Eight **Fostex PM0.4** speakers were carefully placed in array formation and connected to a **RME Fireface UCX**. Three more speakers were arranged in the room as dummys. Compensated **Sennheiser HD-25** were used as headphones. All interaction and processing was done using an 13" Apple Macbook pro and Max/MSP.

The **stimulus** used was a self-produced 15 s loop of a dry guitar and drum set. It was chosen for its natural sound, it's decent use of the frequency

spectrum, the mix of transients and tonal content as well as its unobtrusiveness, making it tolerable to listen to for up to 45 min.

A booth of 2 m × 2 m was constructed using thin white sheets that were of no measurable impact. A visual marker was fixed straight ahead of the subject as the target for the 0° viewing direction. The chair was fixed over a similar marking on the floor with the laptop placed at comfortable distance in front of the subject but well below the hidden array. See Figure 5.2 for a picture of the experiment booth.

Due to the general robustness and gentle failure of the array, the subjects were not fixated but asked to generally position themselves according to both markers (wall and floor). Rotations of the head were discouraged.

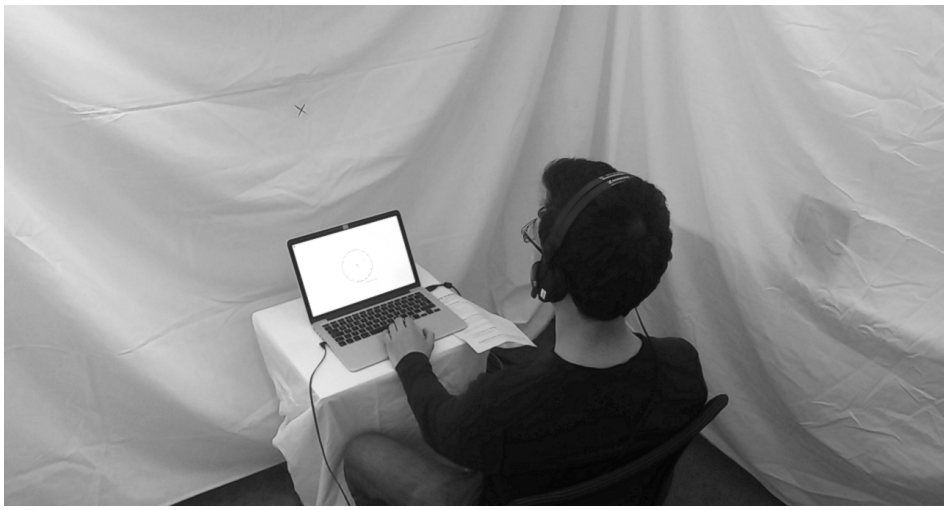


Figure 5.2 Subject seated inside booth. The visual marker to support alignment can be seen on the upper left. To the right, the shadow of a dummy speaker is visible.

All participants were seated without seeing the array but some dummy speakers on the sides of the booth were in view during entrance. To fully hide the speaker's LEDs, a second sheet of acoustically transparent curtain was fixed over the array, as can be seen in Figure 5.3. It's impact on the emitted sound field was confirmed to be negligible.

Figure 5.3 Speaker array as set up for the study behind two acoustically transparent curtain, outside the subjects view.

5.2.4 Subjects

21 subjects (19 M, 2 F), all with self-reported healthy hearing were invited. Of the 21, 9 had a background in audio engineering, 4 in general engineering and 8 had no technical background. Out of the 21 subjects, 3 had a set of personalized HRTF available which was used instead of the selection process. After preliminary examination of the data, two sets of answers were omitted due to highly implausible results.

5.3 Results

In this section, the raw results are presented without further analysis or interpretation, which can be found in Section 5.4. Throughout the following text, you may find the presentation via headphones noted as **HP** and the beamforming condition noted as **BF**.

5.3.1 Survey

The questions on difficulty and perception were recorded on an unlabeled scale $1-5$ with only the ends of the scale anchored, see Section 5.2.1. Strictly

speaking, this is an ordinal scale with only ranked measures allowed. Due to the perceived equal spacing of the numbers $1-5$, the author believes a naive approximation as an interval scale is valid. Figure 5.4 shows a full graphical overview of all given answers and table 5.2 reports some descriptives.

Dimension		Min	Max	Mean	Median	Std. Deviation
Difficulty	HP	1	5	3.143	4	1.195
	BF	1	5	3.286	4	0.9562
Timbre	HP	2	5	3.429	4	0.9783
	BF	2	4	3.143	3	0.7270
Externalization	HP	1	4	2.857	3	1.014
	BF	2	5	4.000	4	0.7746
Spaciousness	HP	1	4	3.381	4	0.8646
	BF	2	5	3.571	4	0.8106

Table 5.2 Descriptives of survey answers.

5.3.2 Localization experiment

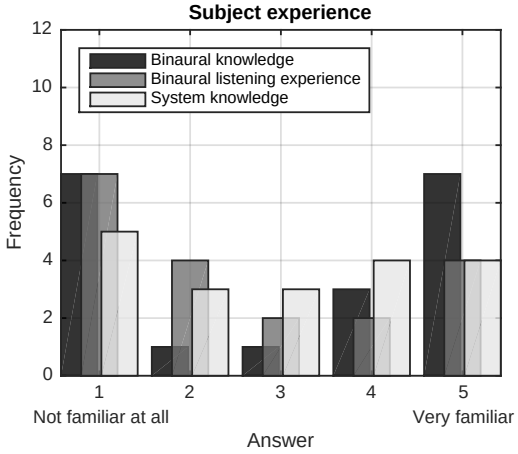
A total of 3584 answers were recorded. A normalized histogram is shown in Figure 5.5, the correct answer for each presented angle is marked with a vertical line. Figure 5.6 shows a scatter plot of the responses with the bubble size corresponding to answer frequency.

5.4 Discussion

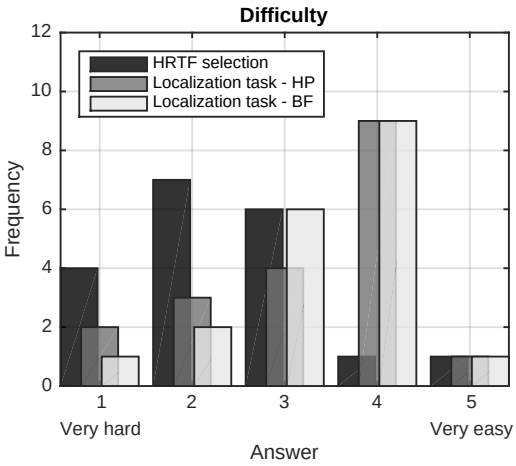
The recorded data is now further analyzed and discussed, especially for differences concerning the mode of presentation.

5.4.1 Survey

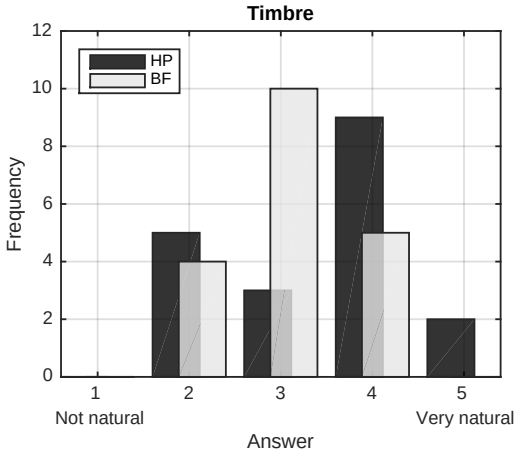
Differences in the perception of Difficulty, Timbre, Externalization and Spaciousness (Figure 5.4) were tested on significance using a paired-sample T-Test, where the null hypothesis is, that the pairwise difference between the answers



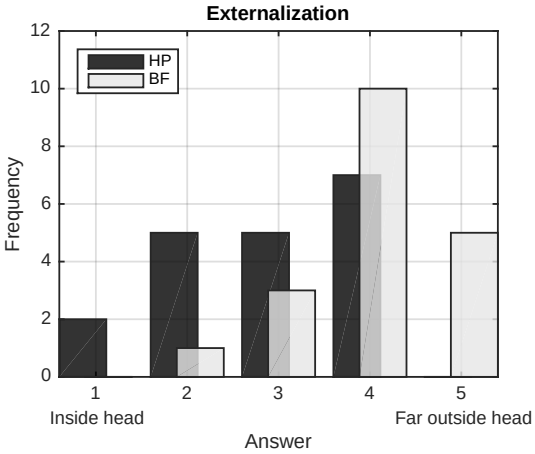
(a) Subject knowledge on binaural techniques, binaural listening and the presented system



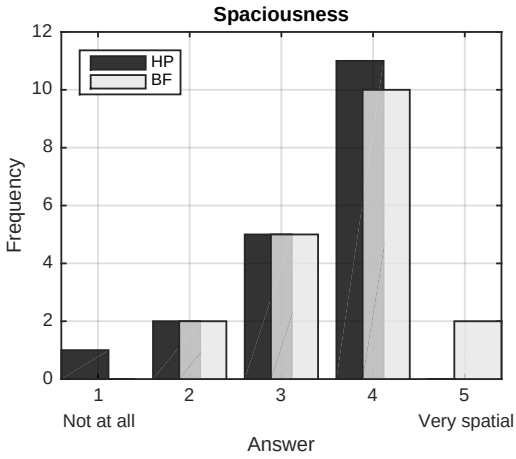
(b) Subjective difficulty rating of both localization tasks and the initial HRTF selection



(c) Perceived naturalness



(d) Perceived externalization of source



(e) Perceived spaciousness of source

Figure 5.4 Results of the electronic survey conducted after the experiment.

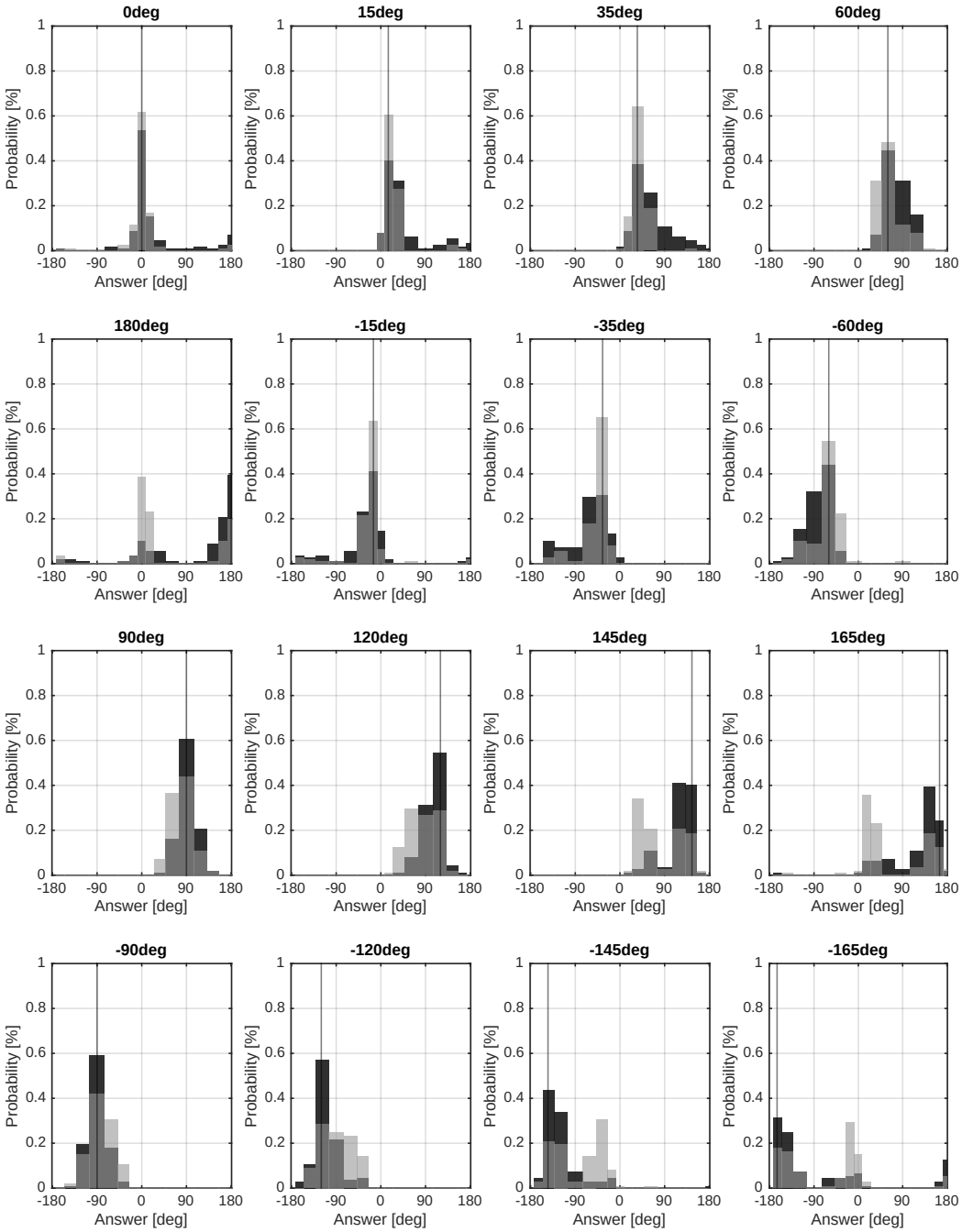


Figure 5.5 Distribution of answers for all presented angles. Dark bars depict the headphone condition, light bars the beamforming condition.

distribution has a mean equal to zero. All distributions pass the Shapiro-Wilk test of normality.

	Student T-Test			Shapiro-Wilk Normality	
Dimension (HP - BF)	t	df	p	W	p
Difficulty	−0.513	20	0.614	0.921	0.093
Timbre	1	20	0.329	0.944	0.259
Externalization	**−3.983**	20	**< .001**	0.934	0.163
Spaciousness	−0.677	20	0.506	0.920	0.086

Table 5.3 T-Test between the answer distributions for headphone and beamforming presentation. Bold dimensions are statistically significant with 95% confidence level.

As can be seen in table 5.3, only **externalization** was perceived significantly different ($p < 0.001$) between headphones ($M = 2.857$, $SD = 1.014$) and beamforming ($M = 4$, $SD = 0.7746$). This should come at no surprise, as this is easily explained by the additional cues added by playing through the array in a real room. To increase the comparability between headphones and array presentation, a BRIR with similar qualities to the experiment cabin could be added to the material played back on headphones.

5.4.2 Localization experiment

Figure 5.6 shows a scatter plot of answers over a density distribution of these answers. Due to the strong quantization of the data, a small amount of noise is added for better visualization in the background density distributions.

An ideal system with ideal subject would lead to a direct mapping of presented (x-Axis) to answered (y-Axis) angle, resulting in a diagonal line from the bottom left to the top right. Apart from obvious deviations, front-back confusion manifests itself in "branches" that split of the main diagonal at ±90°. In fact, both systems exhibit this behavior with different severity: Using headphones, a roughly equal amount of front-back confusion can be observed, while the beamforming presentation clearly shows less front-back confusion in the front (branches that fold towards 0°) and more back-front confusion in

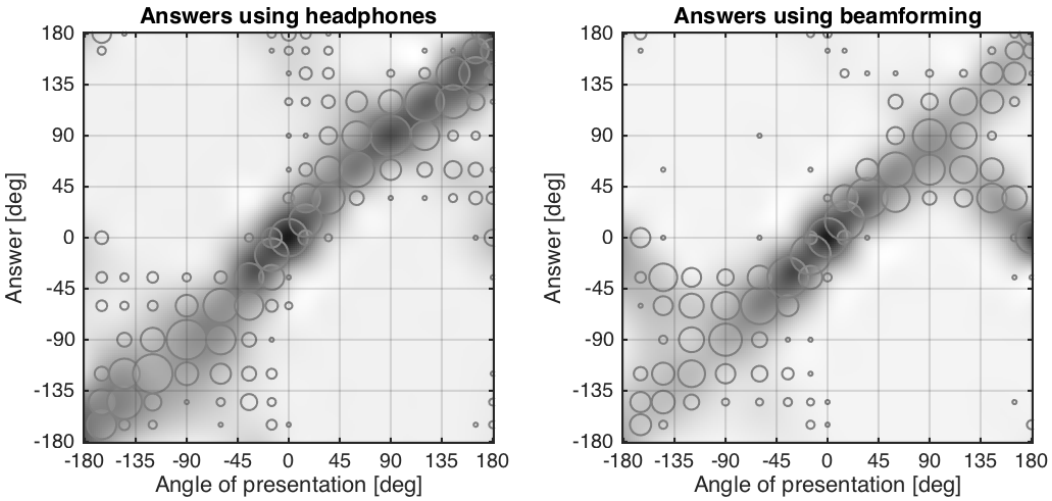


Figure 5.6 Scatter plot of answers over density plot, bubble size corresponds to answer frequency.

the back. This fits the comments of several subjects, who reported little to no stimuli from the back half when localizing the source using the array.

This also becomes evident when looking at Figure 5.7, which shows the distribution of absolute error over all answers split by presentation mode with error bars marking the 95% confidence interval. Looking at the headphone presentation, a similar increase in absolute error can be observed for the front and the back while the error is minimal at ±90°. With the array, errors in the front half up to ±90° are consistently low, after which they sharply increase.

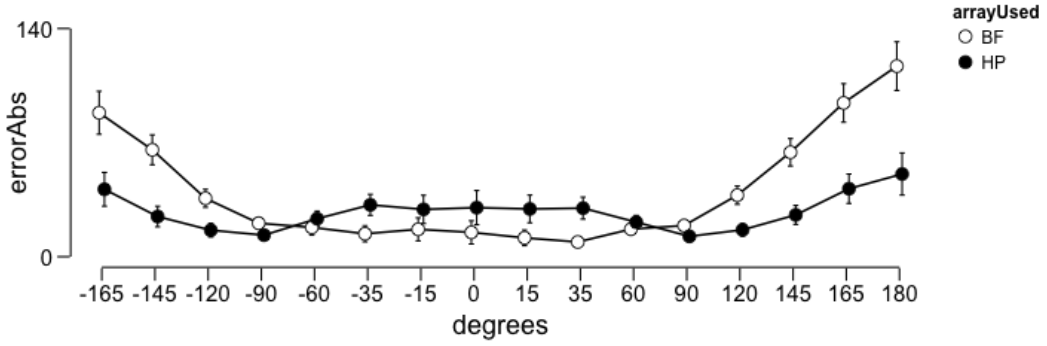


Figure 5.7 Absolute error over answers for both modes of presentation, error bars mark 95% confidence interval.

This is validated by plotting the absolute error in polar form as shown in Figure 5.8 (here again with a small amount of noise added for better visual-

ization). A 1-way ANOVA confirms that the number of front-back ambiguity was significantly different ($F(1) = 91.43$, $p < 0.001$) for the two experiment conditions (WITH/WITHOUT headphones), Table 5.4 shows the distribution. This is further explored in the following Section 5.4.2.

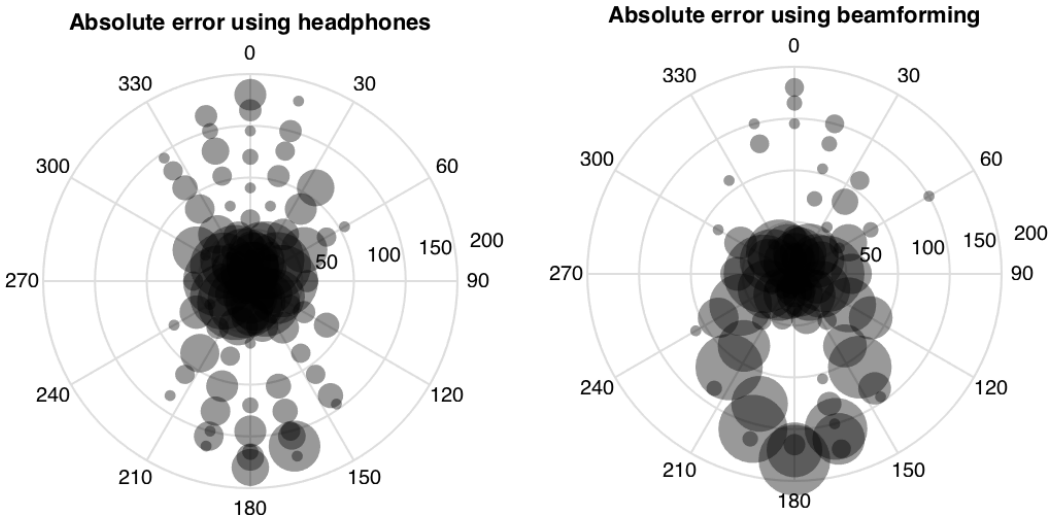


Figure 5.8 Polar scatter plot of absolute error at each angle, bubble size corresponds to error frequency.

Table 5.4 Rate of front-back ambiguity with both modes of presentation

	Number	FBA	Percentage
Headphones	1792	241	13.5 %
Beamforming	1792	466	26.0 %
Total	3584	797	19.7 %

Front-back ambiguity

Front-back confusion is common in binaural presentation as the difference is solely encoded in the coloration cues. This strongly punishes a rather minor error when considering only absolute deviation from the presented angle, as ITD and ILD can suggest a mirrored source location up to 180° rotated away, (see Section 2.1).

To test how the errors in localization compare without front-back ambiguity, an additional FBA error was calculated, which is identical to absolute

error but instead compares the answer to the mirrored angle of presentation, if a front-back confusion is detected. For example, answering 10° when presented with a 180° stimulus would count as an absolute error of 170° but only as a 10° FBA error. It's distribution is shown in Figure 5.9. Now, beamforming seems to perform slightly better when excluding front-back confusion.

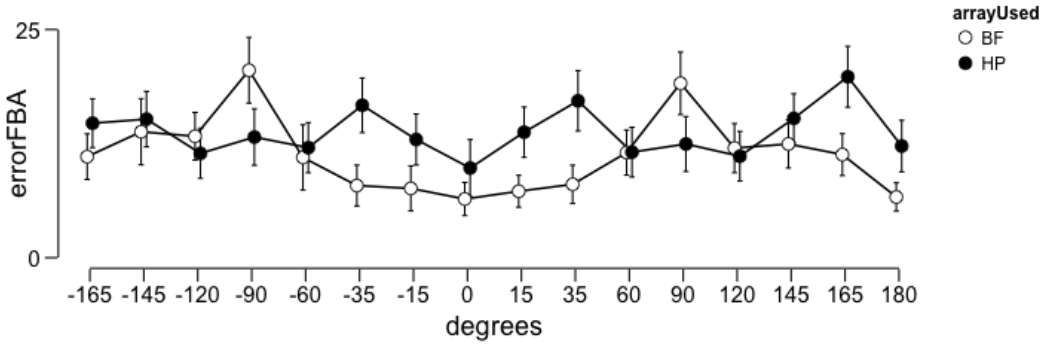
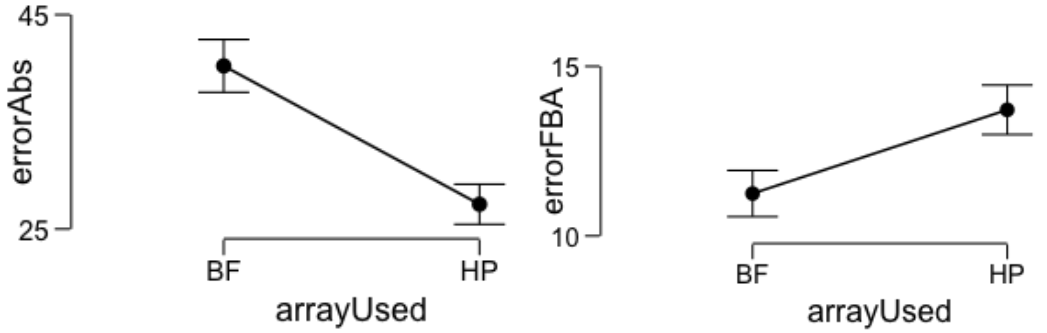


Figure 5.9 Relative error (without front-back confusion) over answers for both modes of presentation, error bars mark 95% confidence interval.

Figure 5.10 clearly shows how the absolute error combined over all directions is significantly higher when using the array. When not counting front-back confusion this is turned around and the beamforming approach has significantly lower errors compared to headphones ($F(1) = 23.56, p < 0.001$).



(a) Effect of the presentation on the absolute error.

(b) Effect of the presentation on the error without front-back ambiguity.

Figure 5.10 Effects on error of not penalizing front-back confusion.

Table 5.5 shows some descriptives of both errors.

This means that apart from ensuring decent listening position, the proper selection of the HRTF is much more critical for presenting reliable sources

from the array-averted side when using beamforming. While presenting audio sources from the back clearly worked for some participants, a larger group of subjects had various degrees of problems perceiving any sources coming from the back.

	Mean [°]	Median [°]	Std. Dev [°]
answers	7.702	15	91.11
errors	-3.548	0	67.93
errors abs	33.76	20	47.69
errors (w/o FBA)	12.48	0	15.28

Table 5.5 Descriptives of localization experiment.

5.4.3 Best case

It is always hard to strictly compare performance for system highly reliant on the perceptual cues of non-individual HRTFs. While on-par performance to headphones could not be achieved in terms of front-back ambiguity, the recorded data suggests a subgroup of 6 participants for whom the array clearly worked in presenting binaural stimuli. As obvious from their answers (see Figure A3 for the corresponding histogram), this best case group of subjects barely experienced any front-back confusion and consistently matched or beat headphones in localization precision.

Figure 5.12 shows the distribution of absolute errors over all directions for the both conditions, Figure 5.12 a scatter plot in polar space. While a trend towards better localization using the speaker array might be assumed, the differences between presentation mode are not significant ($F(1) = 1.91$, $p = 0.167$).

Perception wise, some subjects mention that some sources weren't immediately perceived from the back but the location were rather learned due to coloration. Nevertheless, figures 5.11, 5.12 and 5.13 confirm that the beamforming condition was very comparable for these subjects.

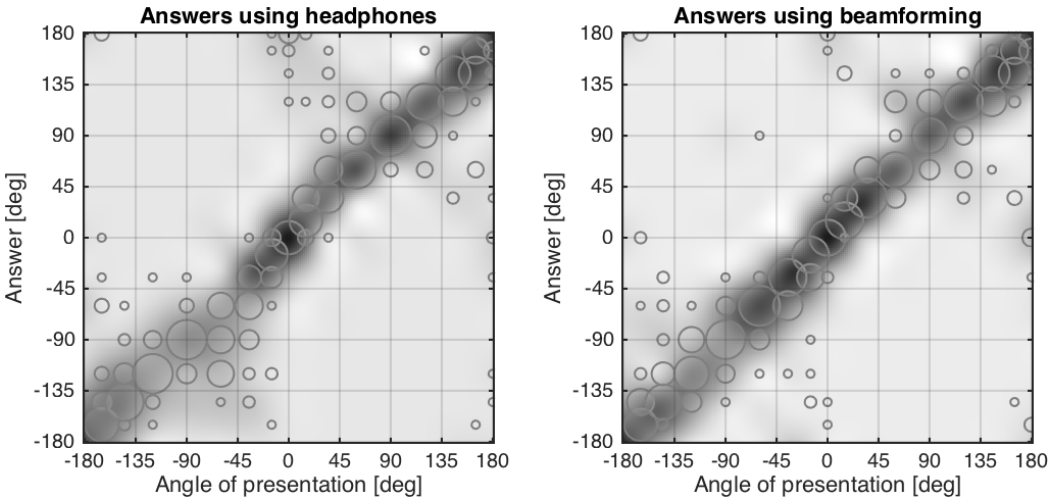


Figure 5.11 Scatter plot of answers over density plot for best case subjects, bubble size corresponds to answer frequency.

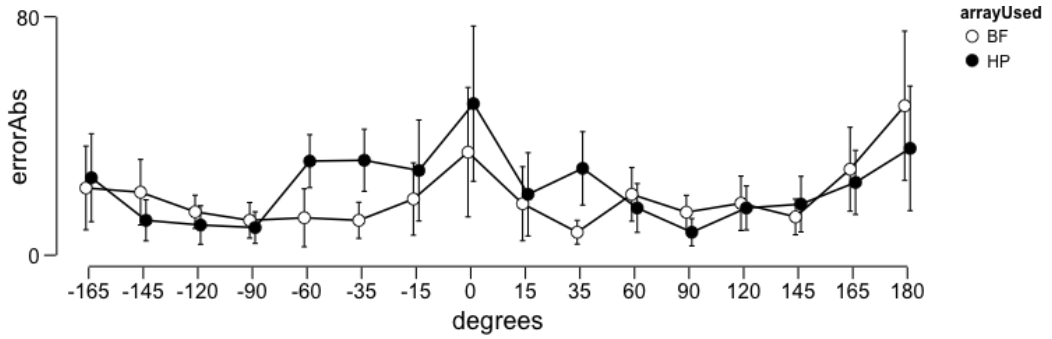


Figure 5.12 Absolute error over answers for both modes of presentation for best case subjects, error bars mark 95% confidence interval.

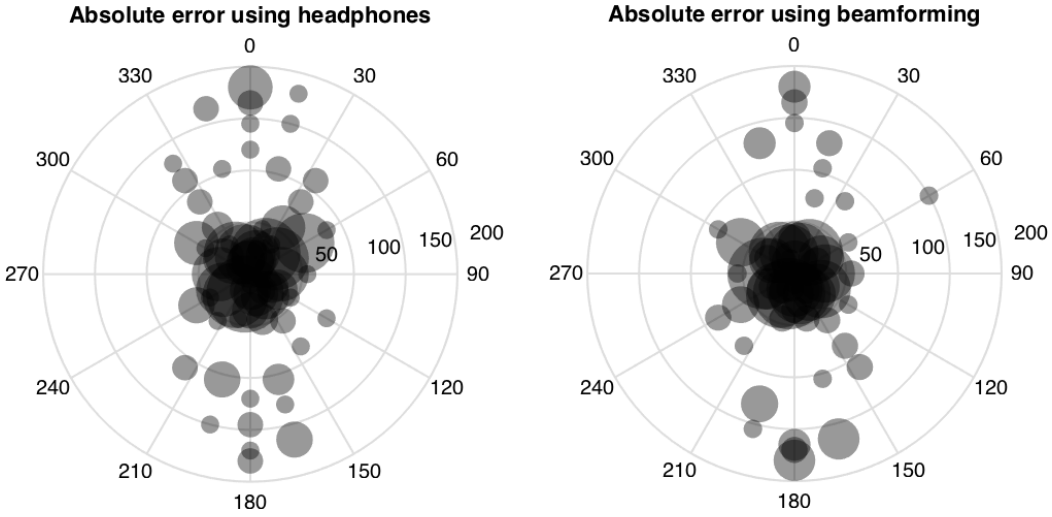


Figure 5.13 Polar scatter plot of absolute error at each angle for best case subjects, bubble size corresponds to error frequency.

5.4.4 Other observations

Interestingly, informal tests seem to show that front-back confusion becomes less of a problem when placing the array BEHIND the listener. It seems that the cues for sources in the front were strong enough to not be overpowered, while cues for the back were helped through the natural localization due to inaccuracies, room response etc. It was decided to not pursue this further due to time constraints.

Chapter 6

Summary

6.1 Conclusions

In this thesis, the feasibility of using a beamforming-based approach for transaural presentation and the resulting advantages over traditional crosstalk cancellation methods can be confirmed. Simulating an eight-speaker Least-Squares Frequency Invariant Beamformer (LSFIB) array, an average channel separation of $30\,\mathrm{dB} - 40\,\mathrm{dB}$ was obtained over a frequency range $1\,\mathrm{kHz} - 8\,\mathrm{kHz}$, which was robust against modifications of driving function and listener displacement. A low budget speaker array of the same dimensions was built and extended to employ Recursive Ambiophonic Crosstalk Elimination (RACE) at lower frequencies. With this array, a channel separation of $10\,\mathrm{dB} - 20\,\mathrm{dB}$ could be measured over a wider range of $250\,\mathrm{Hz} - 20\,\mathrm{kHz}$ using a dummy head.

In a 21 subject study, this system was found to successfully deliver binaural audio material to a listener sitting 1 m in front of the array. Compared to the ground truth presentation with headphones, significantly more front-back confusions occurred, resulting in a larger absolute error ($F(1) = 91.43, p < 0.001$). Conversely, the precision of localization was significantly higher using the array when discounting front-back confusions ($F(1) = 23.56, p < 0.001$). Perceptually, externalization was rated significantly lower on headphones compared to

the array presentation ($t(20) = -3.983, p > 0.001$). Data suggests that for a small group of six best-case participants, no significant differences between presentation methods were found ($F(1) = 1.91, p = 0.167$).

Ultimately, the presentation of binaural audio without headphones is absolutely magical. With comparatively small demands for hardware, space and computing power (especially compared to sound field methods such as Wave Field Synthesis (WFS) or High-Order Ambisonics (HOA)), realistic spatial audio presentation could be achieved for a single subject, if a good fit for their HRTF could be found. Still, the array is not yet ready for adoption in its current state but uncovered several interesting interesting leads that will be discussed in the following section.

6.2 Future work

As explained in Section 2.3.2, non-dynamic binaural synthesis results in a auditory scene that rotates with the head of the listener, prevent the use of head movement to judge localization. Employing a headtracker would alleviate this problem but would again require fixing a device to the listeners head. Optical tracking, such as the Microsoft Kinect, could be used to extract tracking information for a dynamic binaural synthesis. Gardner (1998) confirms a large reduction in front-back reduction when using dynamic binaural synthesis.

There's a second benefit from tracking the user - despite the much increased robustness compared to other approaches, moving the listener too far from the optimized position will result in unpredictable perception, especially due to the somewhat unconstrained array behavior outside the optimized angles. This could partly be amended by tracking the absolute position of a user inside the listening area and adjusting the array response (along with RACE parameters) accordingly. Ideally, due to the slow computation time, a large set of optimized driving functions would have to be pre-calculated, covering all feasible listening positions. The system then simply needs to know the

listeners position in space and swap out the current set for one with better fit. Again, improvements in location judgements by the listener due to dynamic steering was confirmed by Gardner (1998).

Another interesting avenue of research is the accidental discovery of much decreased front-back ambiguity when placing the array behind the user. For still unknown reasons, it seems that the cues of sources facing the user (which are in this case averted from the array) are still strong enough to be located correctly, while sources behind the user have the added localization cues of the room. This could greatly reduce the amount of back-front confusion. Furthermore, there is one more spatial dimension (up-down) that has not yet been tested at all.

Lastly, one could could improve the physical aspects of the array or its calibration, although it is unclear how much there is to profit - it seems implausible to ever come close to the results of simulations. This can include smaller speaker (and therefore smaller speaker spacing) or more linear speaker. Furthermore, speaker with highly optimized radiation characteristic such as coaxial speakers could be used to minimize issues of intereference.

References

Ahrens, J. (2012), *Analytic Methods of Sound Field Synthesis*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Ahrens, J., Rabenstein, R. & Spors, S. (2008), The Theory of Wave Field Synthesis Revisited, *in* 'Audio Engineering Society Convention 124'.

Ahrens, J. & Spors, S. (2011), On the scattering of synthetic sound fields, *in* 'Audio Engineering Society Convention 130', Audio Engineering Society.

Ahrens, J., Thomas, M. R. & Tashev, I. (2013), Gentle acoustic crosstalk cancelation using the spectral division method and Ambiophonics, *in* 'Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on', IEEE, pp. 1–4.

Bai, M. R., Tung, C.-W. & Lee, C.-C. (2005), 'Optimal design of loudspeaker arrays for robust cross-talk cancellation using the Taguchi method and the genetic algorithm', *The Journal of the Acoustical Society of America* **1**17(5), 2802.

Bartlett, M. S. (1948), 'Smoothing periodograms from time series with continuous spectra', *Nature* **1**61(4096), 686–687.

Bauck, J. & Cooper, D. H. (1996), 'Generalized transaural stereo and applications', *Journal of the Audio Engineering Society* **4**4(9), 683–705.

Bauer, B. B. (1961), 'Stereophonic earphones and binaural loudspeakers', *Journal of the Audio Engineering Society* **9**(2), 148–151.

Beyerdynamic (2014), 'Mm-1: Electret condenser measurement microphone datasheet'. http://north-america.beyerdynamic.com/shop/media/datenblaetter/DAT_MM1_EN_A3.pdf.

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

Bock, T. M. & Keele Jr, D. B. (1986), The Effects of Interaural Crosstalk on Stereo Reproduction and Minimizing Interaural Crosstalk in Nearfield Monitoring by the Use of a Physical Barrier: Part 1, *in* 'Audio Engineering Society Convention 81', Audio Engineering Society.

Bohn, D. (1989), *Linkwitz-Riley Crossovers: A Primer*, RaneNote.

Bouchard, J. (2013), Older technological sound recording mediums., PhD thesis.

Brinkmann, F., Lindau, A., Müller-Trapet, M., Vorländer, M. & Weinzierl, S. (2015), Cross-validation of measured and modeled head-related transfer functions, *in* '41. Jahrestagung für Akustik', Deutsche Gessellschaft für Akustik, Nürnberg.

Brinkmann, F., Lindau, A., Weinzierl, S., Geissler, G. & van de Par, S. (2013), A high resolution head-related transfer function database including different orientations of head above the torso, *in* 'Proceedings of the AIA-DAGA 2013 Conference on Acoustics'.

Capon, J. (1969), 'High-resolution frequency-wavenumber spectrum analysis', *Proceedings of the IEEE* **57**(8), 1408–1418.

Cooper, D. H. & Bauck, J. L. (1989), 'Prospects for Transaural Recording', *Journal of the Audio Engineering Society* **37**(1/2), 3–19.

Damaske, P. (1971), 'Head-Related Two-Channel Stereophony with Loudspeaker Reproduction', *The Journal of the Acoustical Society of America* **50**(4B), 1109–1115.

Fastl, H. & Zwicker, E. (2007), *Psychoacoustics: facts and models*, number 22 *in* 'Springer series in information sciences', 3rd. ed edn, Springer, Berlin ; New York.

Gardner, W. G. (1998), *3-D audio using loudspeakers*, Springer Science & Business Media.

Gardner, W. G. & Martin, K. D. (1995), 'HRTF measurements of a KEMAR', *The Journal of the Acoustical Society of America* **9**7(6), 3907–3908.

Geier, M., Wierstorf, H., Ahrens, J., Wechsung, I., Raake, A. & Spors, S. (2010), Perceptual Evaluation of Focused Sources in Wave Field Synthesis, *in* 'Audio Engineering Society Convention 128'.

Gierlich, H. W. (1992), 'The application of binaural technology', *Applied Acoustics* **3**6(3), 219 – 243.

Glasgal, R. (2007), 360 localization via 4. x race processing, *in* 'Audio Engineering Society Convention 123', Audio Engineering Society.

Grant, M. & Boyd, S. (2008), Graph implementations for nonsmooth convex programs, *in* V. Blondel, S. Boyd & H. Kimura, eds, 'Recent Advances in Learning and Control', Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, pp. 95–110. `http://stanford.edu/~boyd/graph_dcp.html`.

Grant, M. & Boyd, S. (2014), 'CVX: Matlab software for disciplined convex programming, version 2.1', `http://cvxr.com/cvx`.

Grasser, T., Rothbucher, M. & Diepold, K. (2014), *Auswahlverfahren für HRTFs zur 3D Sound Synthese*.

Griesinger, D. (1989), 'Equalization and Spatial Equalization of Dummy-Head Recordings for Loudspeaker Reproduction', *J. Audio Eng. Soc* **3**7(1/2), 20–29.

Guldenschuh, M., Shaw, C. & Sontacchi, A. (2010), Evaluation of a Transaural Beamformer, Nizza.

Guldenschuh, M. & Sontacchi, A. (2009), Transaural stereo in a beamforming approach, *in* 'Proc. DAFx', Vol. 9, pp. 1–6.

Harker, A. (2011), 'Aharker externals'.

Harker, A. & Tremblay, P. A. (2012), The HISSTools impulse response toolbox: Convolution for the masses, *in* 'Proceedings of the International Computer

Music Conference', The International Computer Music Association, pp. 148–155.

Hertz, B. F. (1981), '100 Years with Stereo-The Beginning', *J. Audio Eng. Soc* **2**9(5), 368–370, 372.

Iwaya, Y. (2006), 'Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears', *Acoustical science and technology* **2**7(6), 340–343.

Kaiser, F. (2011), Transaural Audio-The reproduction of binaural signals over loudspeakers, PhD thesis, Diploma Thesis, Universität für Musik und darstellende Kunst Graz/Institut für Elekronische Musik und Akustik/IRCAM, March 2011.

Kazutaka, A., Asano, F. & Suzuki, Y. (1997), 'Sound field reproduction by controlling the transfer functions from the source to multiple points in close proximity', *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences* **8**0(3), 574–581.

Krim, H. & Viberg, M. (1996), 'Two decades of array signal processing research: the parametric approach', *IEEE Signal Processing Magazine* **1**3(4), 67–94.

Lebret, H. & Boyd, S. (1997), 'Antenna array pattern synthesis via convex optimization', *Signal Processing, IEEE Transactions on* **4**5(3), 526–532.

Lopez, J. J. & Gonzalez, A. (2001), 'Experimental evaluation of cross-talk cancellation regarding loudspeakers' angle of listening', *IEEE Signal Processing Letters* **8**(1), 13–15.

Ma, W.-K., Ching, P.-C. & Vo, B.-N. (2004), 'Crosstalk resilient interference cancellation in microphone arrays using Capon beamforming', *IEEE Transactions on Speech and Audio Processing* **1**2(5), 468–477.

Mabande, E., Buerger, M. & Kellermann, W. (2012), Design of robust polynomial beamformers for symmetric arrays, *in* '2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 1–4.

Mabande, E. & Kellermann, W. (2010), Design of robust polynomial beamformers as a convex optimization problem, *in* 'Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)'.

Mabande, E., Schad, A. & Kellermann, W. (2009), Design of robust superdirective beamformers as a convex optimization problem, *in* 'Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on', IEEE, pp. 77–80.

McKendrick, J. G. (1909), 'The gramophone as a phonautograph', *Nature* **8**0, 188–191.

Menzel, D., Wittek, H., Theile, G., Fastl, H. & others (2005), The binaural sky: A virtual headphone for binaural room synthesis, *in* 'International Tonmeister Symposium'.

Møller, H. (1992), 'Fundamentals of binaural technology', *Applied acoustics* **3**6(3), 171–218.

Mori, T., Fujiki, G., Takahashi, N. & Maruyama, F. (1979), 'Precision Sound-Image-Localization Technique Utilizing Multitrack Tape Masters', *J. Audio Eng. Soc* **2**7(1/2), 32–38.

Müller, S. & Massarani, P. (2001), 'Transfer-Function Measurement with Sweeps'.

Nelson, P. A., Hamada, H. & Elliott, S. J. (1992), 'Adaptive inverse filters for stereophonic sound reproduction', *Signal Processing, IEEE Transactions on* **4**0(7), 1621–1632.

Nykänen, A., Zedigh, A. & Mohlin, P. (2013), Effects on localization performance from moving the sources in binaural reproductions, *in* 'INTER-NOISE and NOISE-CON Congress and Conference Proceedings', Vol. 247, Institute of Noise Control Engineering, pp. 4023–4031.

Parra, L. C. (2005), Least squares frequency-invariant beamforming, *in* 'Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on', IEEE, pp. 102–105.

Parra, L. C. (2006), 'Steerable frequency-invariant beamforming for arbitrary arrays', *The Journal of the Acoustical Society of America* **1**19(6), 3839.

Place, T. & Lossius, T. (2006), Jamoma: A modular standard for structuring patches in max, *in* 'Proceedings of the International Computer Music Conference', pp. 143–146.

Polk, M. (1984), 'Sda speakers: Designed in stereo', *AUDIO* **6**8(6), 32–41.

Polk, M. S. (2005), 'SDA™ Surround Technology White Paper', *Polk Audio, Nov* .

Rossing, T. D., ed. (2007), *Springer handbook of acoustics*, Springer, New York, N.Y.

Schonstein, D. & Katz, B. F. (2010), HRTF selection for binaural synthesis from a database using morphological parameters, *in* 'International Congress on Acoustics (ICA)'.

Schroeder, M. & Atal, B. (1963), 'Computer simulation of sound transmission in rooms', *Proceedings of the IEEE* **5**1(3), 536–537.

Schroeder, M. R. (1984), 'Progress in Architectural Acoustics and Artificial Reverberation: Concert Hall Acoustics and Number Theory', *J. Audio Eng. Soc* **3**2(4), 194–203.

Seeber, B. U. & Fastl, H. (2003), 'Subjective selection of non-individual head-related transfer functions'.

Van Trees, H. L. (2002), *Optimum array processing: part IV of detection, estimation, and modulation*, Wiley, New York.

Van Veen, B. & Buckley, K. (1988), 'Beamforming: a versatile approach to spatial filtering', *IEEE ASSP Magazine* **5**(2), 4–24.

Wallach, H. (1939), 'On Sound Localization', *The Journal of the Acoustical Society of America* **1**0(4), 270–274.

Ward, D. B. (2000), 'Joint least squares optimization for robust acoustic crosstalk cancellation', *IEEE Transactions on Speech and Audio Processing* **8**(2), 211–215.

Ward, D. B. & Elko, G. (1998), Optimum loudspeaker spacing for robust crosstalk cancellation, *in* 'Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998', Vol. 6, pp. 3541–3544 vol.6.

Ward, D. B. & Elko, G. W. (1999), 'Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation', *Signal Processing Letters, IEEE*

6(5), 106–108.

Weinzierl, S., ed. (2008), *Handbuch der Audiotechnik*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Welch, P. D. (1967), 'The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms', *IEEE Transactions on audio and electroacoustics* **1**5(2), 70–73.

Wierstorf, H., Geier, M. & Spors, S. (2011), A free database of head related impulse response measurements in the horizontal plane with multiple distances, *in* 'Audio Engineering Society Convention 130', Audio Engineering Society.

Wightman, F. L. & Kistler, D. J. (1999), 'Resolution of front–back ambiguity in spatial hearing by listener and source movement', *The Journal of the Acoustical Society of America* **1**05(5), 2841–2853.

Williams, E. G. (1999), *Fourier acoustics: sound radiation and nearfield acoustical holography*, Academic press.

Zotkin, D. N., Duraiswami, R., Grassi, E. & Gumerov, N. A. (2006), 'Fast head-related transfer function measurement via reciprocity', *The Journal of the Acoustical Society of America* **1**20(4), 2202–2215.

Zotkin, D. N., Duraiswami, R. & Gumerov, N. A. (2009), Regularized HRTF fitting using spherical harmonics, IEEE, pp. 257–260.

Appendix

The first half of this thesis was spent at the CCRMA institute at Stanford university, where large parts of the implementation and simulation completed. Figure 5.2 shows the two speaker models used for the initial implementation: After extensive tests on a large array of **Adam A3x** speakers (various configurations between 8 and 16 were tried), a more portable array of smaller dimensions consisting of 8 **Meyersound MM-4XP** (white) was used for verification and demos in conjunction with a **MOTU 16A**.

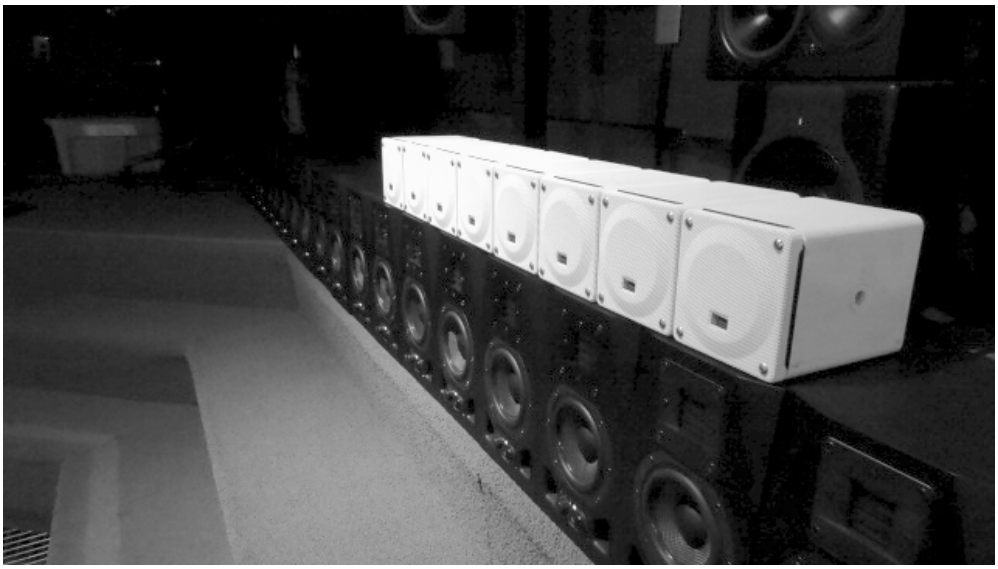


Figure A1 Two speaker arrays used at the CCRMA institute: a 8-speaker Meyer Sound MXP MM-4XP (white) and a larger array consisting of several Adam A3x.

Figure A2 shows a test session inside CCRMA's listening room, were both arrays were set up to be compared.

Figure A2 One of the many test sessions at CCRMA's listening room.

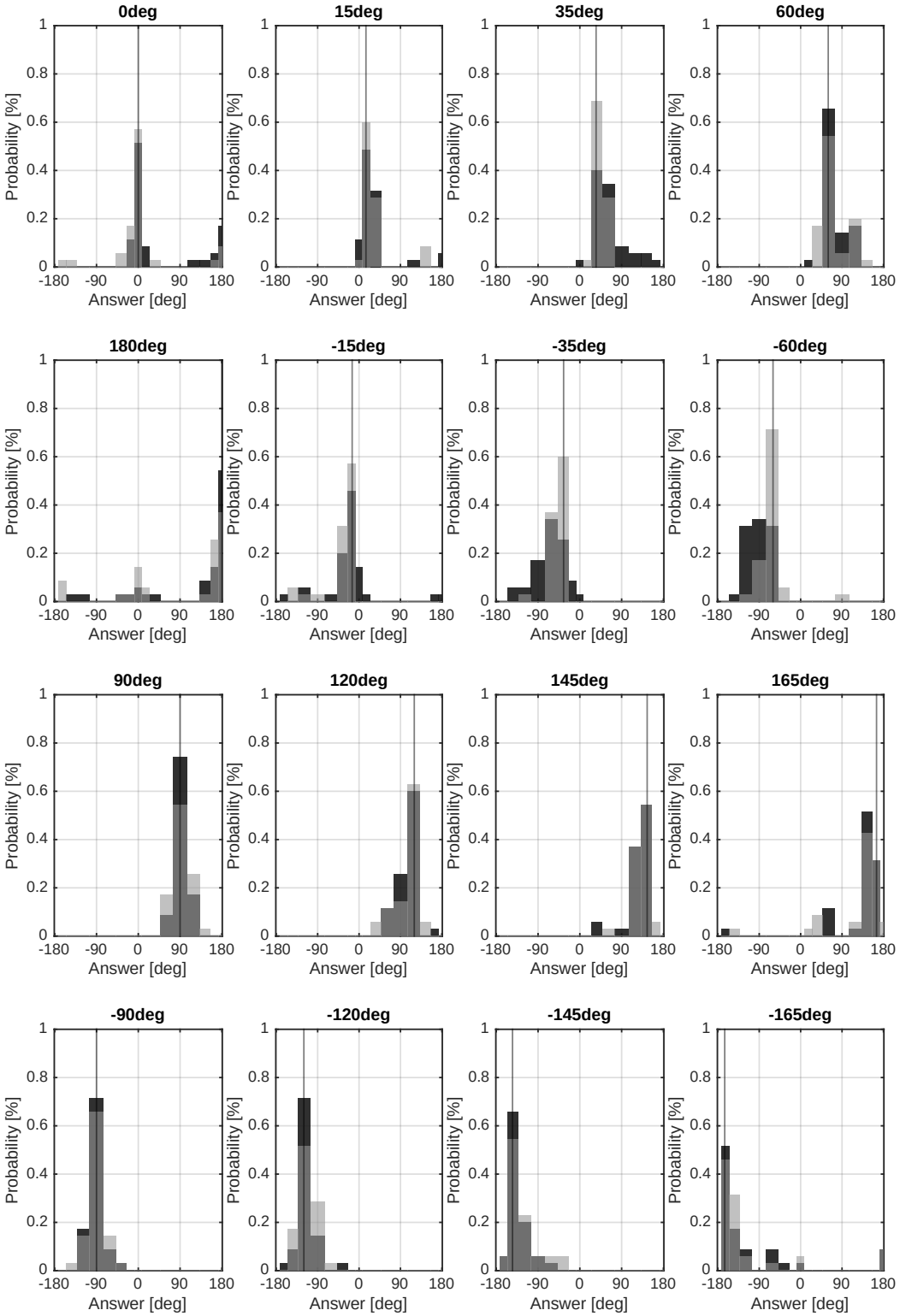


Figure A3 Distribution of answers for all presented angles for best case subjects. Dark bars depict the headphone condition, light bars the beamforming condition.