Transaural Beamforming

Methods for Controllable Focused Sound Reproduction

Markus Guldenschuh

Transaural Beamforming

Methods for Controllable Focused Sound Reproduction

Diploma Thesis

at

Graz University of Technology

submitted by

Markus Guldenschuh

Institute of Electronic Music and Acoustics,
University of Music and Performing Arts Graz
A-8010 Graz, Austria

September 2009

Supervisor: Univ.-Ass. DI Dr. Alois Sontacchi
Assessor: O.Univ.-Prof. Mag.art DI Dr.tech. Robert Höldrich

Transaurales Beamforming
Methoden zur nachgeführten Schallfeldkonzentration

Diplomarbeit

an der

Technischen Universität Graz

vorgelegt von

Markus Guldenschuh

Institut für Elektronische Musik und Akustik,
Universität für Musik und darstellende Kunst Graz
A-8010 Graz, Austria

September 2009

Diese Arbeit ist in englischer Sprache verfasst.

Betreuer: Univ.-Ass. DI Dr. Alois Sontacchi
Begutachter: O.Univ.-Prof. Mag.art DI Dr.tech. Robert Höldrich

Abstract

In literature, several proposals can be found on how to synthesize or resynthesize sound fields. In general, there are two main strategies. Global approaches, like Wave Field Synthesis or Ambisonics, have been derived from physical equivalents and aim to reproduce a sound field within a whole area. Local approaches on the other hand, are designed to reproduce the sound field at the position of a user only, as for example with binaural signals via headphones.

The inherent goal of this work is to elaborate possible improvements for air traffic control communication conditions. In order to free air traffic controllers from the demanding usage of headphones, a loudspeaker-array based application is introduced. The array produces binaural signals at the ears of the user to provide spatialized audio. It therefore is a local approach of sound-source reproduction, that, however, suffers from unwanted cross talk from one binaural channel to the contralateral ear.

Different beamforming methods are investigated under the aspect of focusing quality, room excitation, dynamic adaptation and their contribution to an effective transaural rendering. It is a weighted Delay & Sum Beamformer, a Least Squares Beamformer, a Maximum Energy Difference Beamformer and a Minimum Variance Distortionless Response Beamformer. The first method proved to be very feasible for its facile implementation. Although its principle is very simple, its results are comparable to the other methods that use optimization algorithms for the sound-field manipulation.

Finally, a cross talk cancellation solution for a loudspeaker array and a binaural signal is deduced and the functionality of the overall system is undermined with measurements results.

Kurzfassung

In der Literatur gibt es zahlreiche Ansätze und Vorschläge für die Synthese und Resynthese von Schallfeldern. Dabei lassen sich primär zwei unterschiedliche Strategien unterscheiden. Globale Ansätze, wie Ambisonics oder die Wellenfeld Synthese, streben die Reproduktion eines Schallfeldes in einem größeren Areal an. Lokale Anwendungen hingegen zielen darauf ab, das Schallfeld an der definierten Position einer Hörerin zu steuern. Dazu gehört beispielsweise die Wiedergabe binauraler Signale mit Kopfhörern.

Die Motivation dieser Arbeit war, Verbesserungsmöglichkeiten für die akustische Kommunikationsschnittstelle von Fluglotsinnen zu erarbeiten. Um Fluglotsinnen vom anstrengenden, weil dauerhaftem, Gebrauch von Kopfhörern zu befreien, wird die Verwendung eines Lautsprecher Arrays vorgeschlagen. Der Lautsprecher Array soll binaurale, und damit spatialisierte, Signale an den Ohren der Nutzerin erzeugen. Es handelt sich deshalb um einen lokalen Ansatz zur Schallfeldreproduktion, der jedoch das Problem des Übersprechens von einem binauralen Kanal zum kontralateralen Ohr mit sich bringt.

In dieser Arbeit werden verschiedene Methoden zur Schallfeldkonzentration untersucht. Die Methoden werden in Bezug auf ihre Fokussierungsqualität, ihre Raumanregung, ihre dynamische Steuerbarkeit und auf die Vorbedingungen für eine korrekte transaurale Wiedergabe verglichen. Im Speziellen werden ein gewichteter Delay & Sum Beamformer, ein Least Squares Beamformer, ein Maximum Energy Difference Beamformer und ein Minimum Variance Distortionless Response Beamformer verglichen. Im Vergleich stellt sich heraus, dass der Delay & Sum Beamformer, dank seiner einfachen Implementierungsmöglichkeit, am geeignetsten für eine dynamische Echtzeitanwendung ist. Obwohl dem Delay & Sum Beamformer, im Gegensatz zu allen anderen Beamforming Ansätzen, kein Optimierungsalgorithmus zur Schallfeldproduktion unterliegt, sind seine Ergebnisse durchaus mit denen der anderen Beamforming Methoden vergleichbar.

Schlussendlich wird eine Übersprechkompensations-Lösung für einen Lautsprecher Array und ein Binauralsignal hergeleitet und die Funktionalität der Anwendung wird durch Messergebnisse gestützt.

Pledge of Integrity

I hereby certify that the work presented in this thesis is my own, that all work performed by others is appropriately declared and cited, and that no sources other than those listed were used.

Place: ____

Date: ____

Signature: ____

Eidesstattliche Erklärung

Ich versichere ehrenwörtlich, dass ich diese Arbeit selbständig verfasst habe, dass sämtliche Arbeiten von Anderen entsprechend gekennzeichnet und mit Quellenangaben versehen sind, und dass ich keine anderen als die angegebenen Quellen benutzt habe.

Ort: ____

Datum: ____

Unterschrift: ____

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Alois Sontacchi for his support and the opportunities he gave me. Thanks to Franz Zotter and Hannes Pomberger for discussion, explanations and advices, as well as to Sabrina Metz for assisting me during my measurements.

Special thanks to my family who always encouraged me in all my intentions and to Johanna for being there for me.

Markus Guldenschuh
Graz, Austria, September 2009

Credits

This work was supported in part by Eurocontrol under Research Grant Scheme - Graz, (08-120918-C). Thanks to Horst Hering for his explanations concerning the communication conditions in air-traffic control.

The application that has been developed during this thesis uses the *‘binaural Ambisonics’* patch written by Thomas Musil at the Institute of Electronic Music and a video head-tracking software named ‘*Watson’* that was developed and kindly provided by Louis-Philippe Morency.

Notation

Throughout this thesis bold lowercase letters $\mathbf{x}$ are meant to be vectors, bold uppercase letters $\mathbf{X}$ matrices and italic letters x refer to scalar values. $\mathbf{X}^T$ denotes matrix transposition, $\mathbf{X}^H$ complex conjugate transposition and $\mathbf{X}^*$ conjugation, only.

In this thesis, following physical symbols appear:

j		imaginary number $j = \sqrt{-1}$
f	[Hz]	frequency
ω	[1/s]	radial frequency $\omega = 2\pi f$
λ	[m]	wave length
c	[m/s]	speed of sound
k	[1/m]	wavenumber $k = \frac{\omega}{c}$
ν	[m/s]	sound particle velocity
p	[Pa]	sound pressure
ρ	[kg/m^3]	density
P_{ac}	[W]	acoustic power
$\vec{I}$	[W/m^2]	sound intensity

And the following abbreviations are used:

WDSB	Weighted Delay & Sum Beamformer
LSB	Least Squares Beamformer
MVDRB	Minimum Variance Distortionless Response Beamformer
MEDB	Maximum Energy Difference Beamformer
WNG	White Noise Gain
SVD	Singular Value Decomposition
XTC	Cross Talk Cancellation
HRTF	Head Related Transfer Function

Chapter 1

Introduction

1.1 Motivation

This thesis is the result of investigations that were initialized by a project supported in part by the ’Eurocontrol Research Grant’. The goal of the project is to develop an advanced communication setup for air traffic controllers. The improvements primarily should contain:

- A head set free communication setup, to free the air traffic controllers from the demanding usage of headphones.
- A fully spatialized sound to allow for an acoustically distribution of the communication partners.

It has to be considered that some dozens controllers work simultaneously in air traffic control (ATC) centers. That rises some problems, especially for the first point. If loudspeakers should be used instead of headphones, the sound excitation of the room runs the risk to be more annoying than the usage of head sets. The strategy is therefore to produce a focused sound that is always steered to the position of the user. As the user moves within a certain area, the steering has to be dynamic and adaptive. A video system is suggested to track the position and rotation of the user.

The preconditions in air traffic control deliver two relevant design criteria:

1. The bandwidth in air traffic control reaches from 300 to 2500 Hz.
2. The system should be desktop integrable and processing efficient.

Out of these criteria, the array properties (like shape, size and number of loudspeakers) and the sound focusing method have to be deduced. A desktop integrated communication setup is suggested in Fig. 1.1.

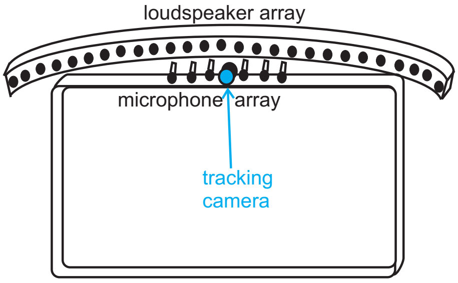


Figure 1.1: The loudspeaker- and microphone array as well as the the camera for the user tracking can be mounted over the air traffic control screen in order to yield a compact system without any head worn hardware.

Fig. 1.2 shows the block diagram of the final system. Sound spatialization is accomplished through binaural signals, which are rendered in the binaural Ambisonics system. The 2-channel binaural signals are applied to a cross-talk canceler before they are led to the beamforming stage that produces a focused sound field.

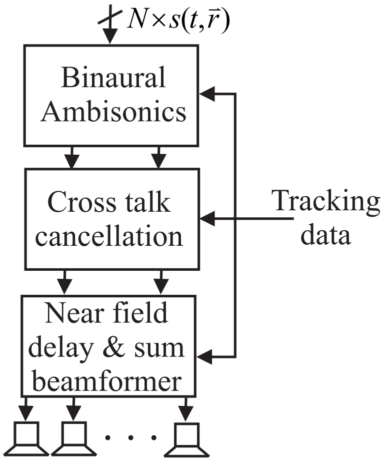


Figure 1.2: Arbitrary many sound sources can be rendered to a 2-channel binaural signal. The binaural signal runs through a cross-talk cancellation filter before it is led to the beamformer. All processing stages need to know the position and the orientation of the user.

Both terms, *beamforming* and *binaural signal* are explained in the following sections. The binaural Ambisonics system is not in the scope of this thesis as it is described by Noisternig et al. [2003].

1.2 Introduction to the Technique of Beamforming

Beamforming has been used since the late 1970s as spatial filter in sensor technologies. A first overview was given by Veen and Buckley [1988]. An array of sensors has been used to filter electromagnetic or sonic waves from a certain direction. If several sensors are in a line in the direction of the incident wave with an interelementary distance of a multiple of the wavelength, the arriving wave is in phase on every sensor. Thus, the signal can be amplified by constructive addition. In telecommunications, this principle is called smart antenna (see Fig. 1.3).

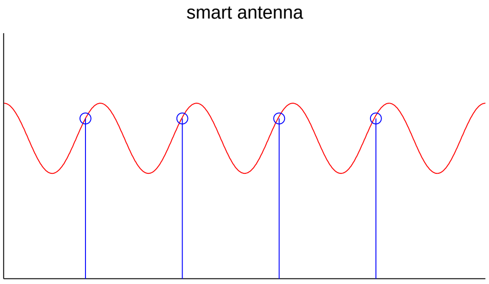


Figure 1.3: The concept of smart antennas: An array of sensors samples a wave in the distance of a wavelength.

An extensive collection of beamforming methods for microphone arrays is given by Brandstein and Ward [2001]. Due to the tight relation to loudspeaker arrays, these methods can also be applied to calculate the driving function of loudspeakers to produce a steered sound.

In this thesis, the Weighted Delay & Sum Beamformer, the Least Squares Beamformer, the Maximum Energy Difference Beamformer and the Minimum Variance Distortionless Response Beamformer are examined. The first is the physical straightforward method. It emphasizes the waves for a certain direction by simple constructive superposition without frequency dependent filtering. All other examined beamforming methods are so-called super-directive beamformers; meaning that they achieve a higher directivity than the simple Delay & Sum Beamformer by additional filtering in the frequency domain.

If a sound pressure concentration in a certain point is desired, like it is the case in this thesis, there exists another straight physical method to achieve a super-resolution. This method is called Time Reversal Mirror and it makes use of room reflections. In Yon et al. [2003], the principle of the Time Reversal Mirror is explained using a reverberating cavity that, in a first step, is closed by an array of microphones like it is shown in Fig. 1.4. At some point in this cavity a pulse is emitted. If the impulse responses recorded by the microphones are played back time-inverted,[1] the sound propagates back and focuses on the initial emitter point.

[1] Whereby the microphones are replaced by loudspeakers with identical spatial properties. (I.e. the spatial sensitivity of the microphones corresponds with the spatial radiation pattern of the loudspeakers.)

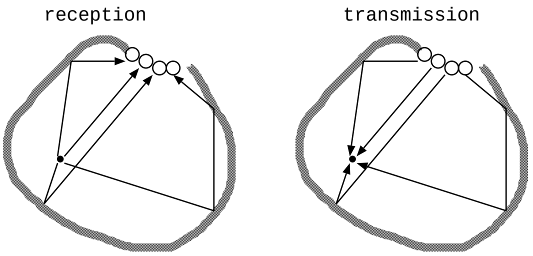


Figure 1.4: The reverberating cavity is closed by microphones in a first step. The microphones record a direct pulse with all its reflections. In a second step, the cavity is close with loudspeakers[1] that play back the microphone signals time reversed. The reflections and the direct pulse gather in the original source point where they cause a super resolution. (Adapted from Yon et al. [2003])

Fig. 1.5 shows a schematic impulse response with a direct wavefront and three discrete reflections. On the one hand, these reflections appear as pre-echos, when played back time-inverted. On the other hand, after having run through the reflection paths again,[2] they are also constructively added with the directional part. This results in a main impulse that is even stronger than the original first wavefront. An adaptive system that is able to follow a moving user needs to know all the impulse responses of the area in which the user can move. To cover one square meter only, approximately 210 impulse responses would have to be determined[3] which makes the Time Reversal Mirror impractical for a dynamic implementation.

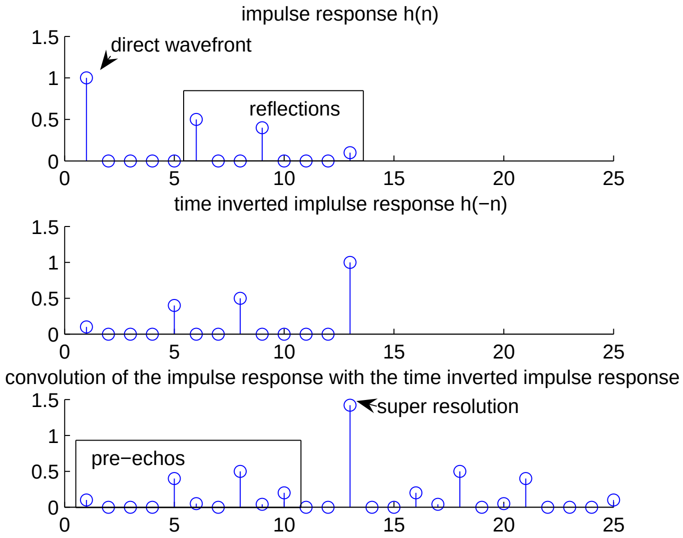


Figure 1.5: The upper figure schematizes the impulse response of one microphone and the middle figure displays its time inverted version. In acoustics, source and sink can be replaced[1]. Thus, the impulse response also counts for the path from the loudspeakers back to the original emitting point. If the time inverted impulse is sent back,[2] the reflections and the direct part superpose to a super-resolution impulse (lower figure). However, with the cost of pre-echos, caused by the time inverted signal.

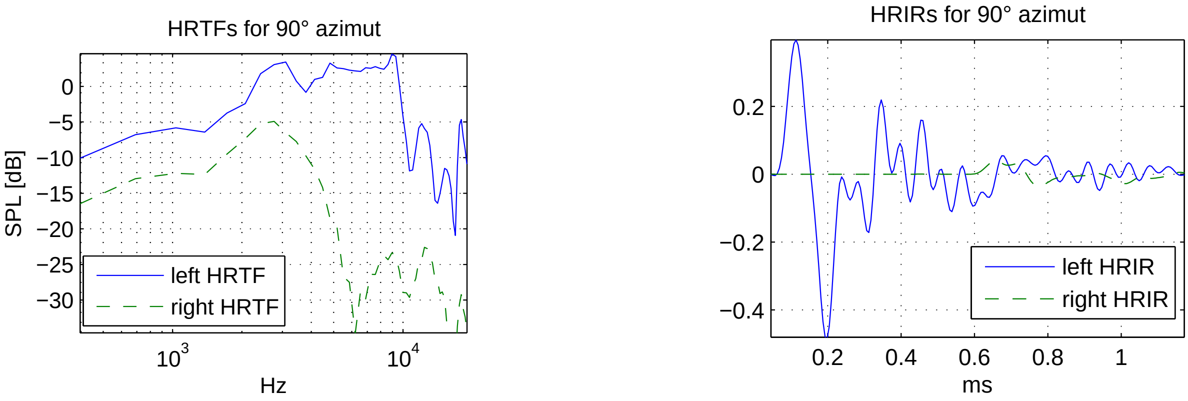
[2] I.e. a convolution with the original impulse response.

[3] The minimal spatial resolution will be a topic of Section 2.1.

1.3 Introduction to the Technique of Transaural Stereo

1.3.1 Binaural Signals

Humans with normal hearing abilities can estimate the position of sound sources due to differences between the right ear signal and the left ear signal. This pair of ear signals is called binaural signal. Binaural signals and their influence on spatial hearing were investigated by Blauert [1999]. The most relevant differences between these signals are the interaural time difference (ITD) and the interaural intensity difference (IIDs). Additional coloration is caused by reflections from the shoulders and the pinna. All these differences are integrative parts of the Head Related Transfer Functions (HRTFs). Such a pair of HRTFs is depicted in Fig. 1.6.



(a) Deep frequencies are diffracted around the head. That is why the IID is not that high for low frequencies.

(b) In the time domain the ITD can be observed. An impulse from the left arrives much sooner at the left ear.

Figure 1.6: Head Related Transfer Functions (HRTFs) and Head Related Impulse Response (HRIR) for 90° azimuth and 0° elevation.

As HRTFs depend on the shape of the ear and the torso geometry, they exhibit a unique profile for every human. Still, a good average HRTF can be achieved with measurements on dummy heads. The measurement of HRTFs is well described in Moller et al. [1995], while Algazi et al. [2001] investigated the capabilities of modeling HRTFs analytically.

HRTFs can be used to spatialize sound. If a mono source is convolved with a pair of HRTFs, the resulting binaural signal evokes a spatial impression if each channel is led directly to the ears. The most common way to perceive binaural signals is thus via headphones. If binaural signals are wanted to be played back via loudspeakers, the technique of transaural stereo has to be applied.

1.3.2 Transaural Stereo

The playback of binaural signals via loudspeakers causes an unwanted cross talk from the left binaural signal to the right ear and vice versa. With the knowledge of the HRTFs from the loudspeakers to the ears, these cross-talk paths can be attenuated. A first filter solution for cross-talk cancellation (XTC) was derived by Atal and Schroeder [1963]. Bauck and Cooper [1996] showed XTC solutions for various constellations of loudspeakers, listeners and binaural signals. A first binaural sound system for loudspeakers and tracked users has been developed by Gardner [1997]. Lentz [2006] investigated the stability of XTC filters for two loudspeakers in dependence on their opening angles. An opening angle of 90° delivered satisfying results. As a consequence, a set of four loudspeakers was placed every 90° to guarantee stable cross talk filters for a full rotation of the user.

Of course, the functionality of XTC depends on the accuracy of the HRTFs. The accuracy is determined by the closeness to the actual personal HRTFs and the precision of the tracking system. Bai et al. [2005] tried to gain robustness against lateral mismatches by applying a cross-talk network from six loudspeakers to six control points (instead of to two ear positions only).

A first transaural system with focused sound was introduced by Menzel et al. [2005] and further investigated by Laumann et al. [2008]. They used a circular array of 22 loudspeakers that produces virtual sound sources via Wave Field Synthesis (WFS). The virtual sound sources are chosen to be point sources. They are placed above the listener and render a binaural signal. Focused sound (like a point source) has the advantage that it causes less cross talk if the focus spots are set in the vicinity of the ears. Laumann et al. [2008] took the HRTFs from the virtual point sources to the user for their XTC filters. Their idea is to rotate the virtual sound sources with the user, such that the same set of XTC filters can be used for any head rotation. However, it is not clear which benefit WFS contributes to the application and why the virtual sound sources are not set into the horizontal plane of the ears.

The goal of this thesis is to investigate different sound focusing methods and their feasibility to a dynamic transaural system with a tolerance to both, lateral movements and head rotations of the user. The investigations concern the directivity of the beamforming methods (and as a consequence their room excitation), their complexity in terms of an implementation, and the precondition for a stable transaural rendering. Section 2.1 gives a theoretical overview of spatial filtering, before the different beamforming methods are described in Section 2.3 to 2.6.

Section 2.3.2 proofs the concept with measurement results and gives reason to the point source simulation of the loudspeakers. In Chapter 3, the calculation of XTC filters for 16 loudspeakers are deduced. The stability of the XTC filters is discussed and the performance results of the Transaural Beamformer are presented. Finally, Chapter 4 gives a resume of the thesis and an outlook to interesting future questions and ideas.

Chapter 2

Beamforming

Beamformers use an array of transducers (such as antennas, microphones or loudspeakers) to steer into a certain direction or into a certain point of a wave field. The steering in general is achieved by filtering and summing up the signals of the different array elements. In the most trivial case, the signals are simply delayed and summed (Delay & Sum Beamformer, see Fig. 2.1) such that they superpose constructively for a certain propagation direction or in a certain point. Beamforming can be done on the reception side (e.g. with microphones) or on the transmitting side. In this work, all discussions are done for beamforming with loudspeaker arrays, but due to the invertibility of acoustical paths, the following theory is valid for either case.

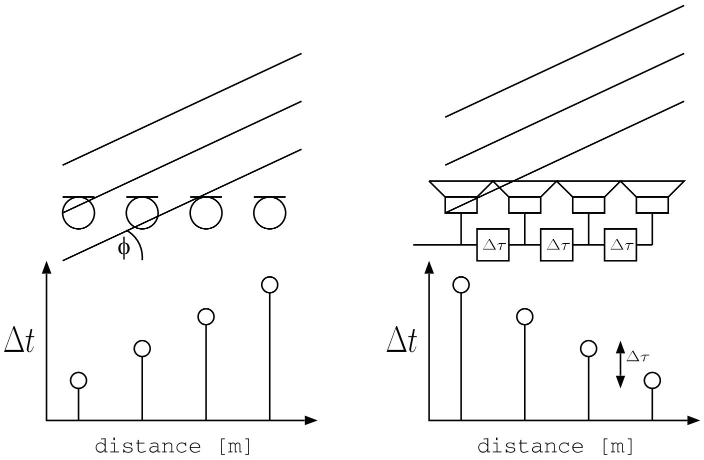


Figure 2.1: Recording and reproduction of an incident wavefront with a simple Delay & Sum Beamformer.

In order to better understand the properties and limitations of beamformers, the next section throws a glance at the spatial Fourier transform before different beamforming methods are explained and investigated in detail.

2.1 Space-Frequency Signal Processing

To deduce the properties and limitations of beamformers, let us first consider an infinite wall with in phase vibrating stripes, as it is depicted in Fig. 2.2.

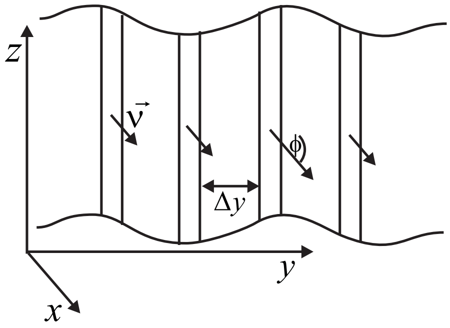


Figure 2.2: Stripes in an infinite wall that vibrate with a sound particle velocity ν into the orthogonal direction.

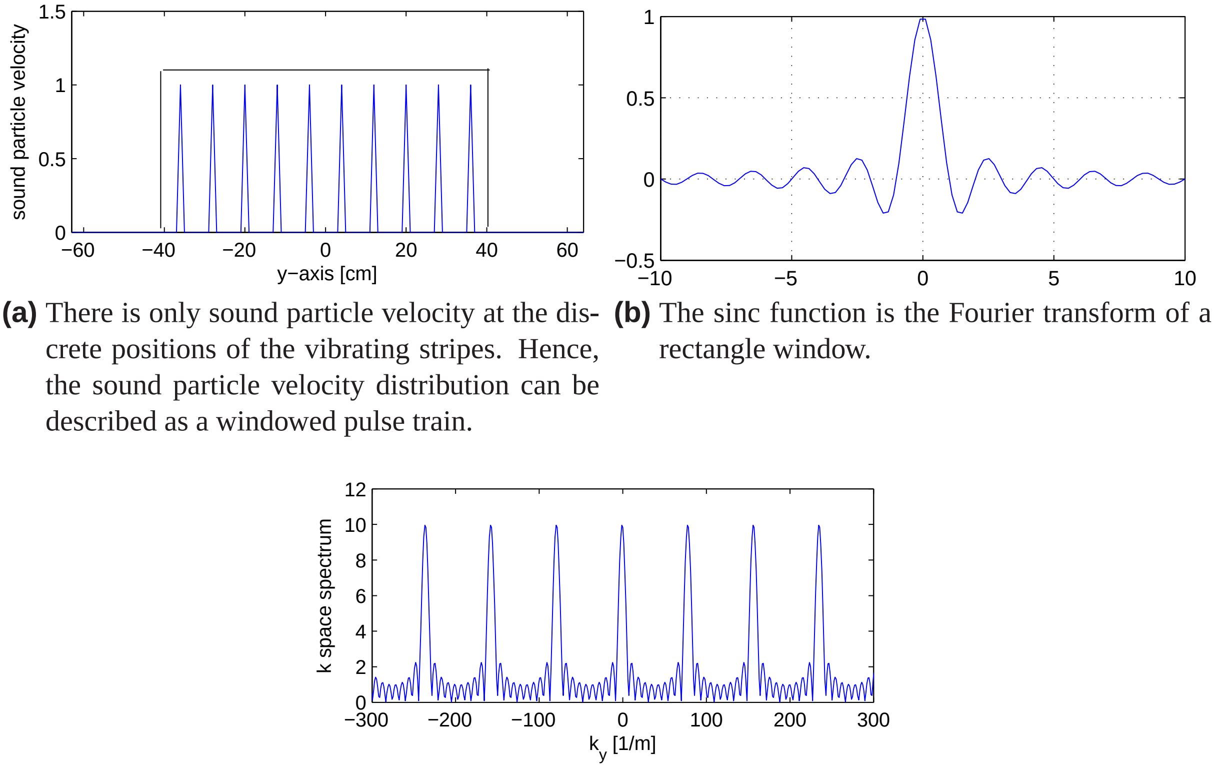
In Möser [1988] it is shown that the directivity of these vibrating stripes in the far field can be derived over the Fourier integral over the sound particle velocity ν along the y-axis. The Fourier transform yields the k-space spectrum (as in Williams [1999])

$$\Phi(k_y) = \int \nu(y) e^{-ik_y y} dy, \tag{2.1}$$

with k_y being the wavenumber in y direction.

$$k_y = \frac{\omega}{c}\sin(\phi). \tag{2.2}$$

For the beginning, the stripes are considered to be infinitesimally thin. Then, the sound particle velocity ν can be expressed as an infinite pulse train along the y-axis for there is only sound particle velocity at the positions of the stripes. As known from time-frequency processing (Oppenheim [1989]), a pulse train stays unchanged through a Fourier transform. However, real arrays of course are not infinite. The finite length of the array can be described by a windowed version of our infinite pulse train (Fig. 2.3a). The multiplication with a rectangle window corresponds to a convolution with a sinc function in the k space (Fig. 2.3c). The sinc function $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$ (also shown in Fig. 2.3b) has a main lobe around $x = 0$ and descending side lobes that approximate zero in infinity.



(a) There is only sound particle velocity at the discrete positions of the vibrating stripes. Hence, the sound particle velocity distribution can be described as a windowed pulse train.

(b) The sinc function is the Fourier transform of a rectangle window.

(c) In the k space, the pulse train stays infinite but the pulses are convolved with the sinc function because the array in **(a)** was windowed with a rectangle.

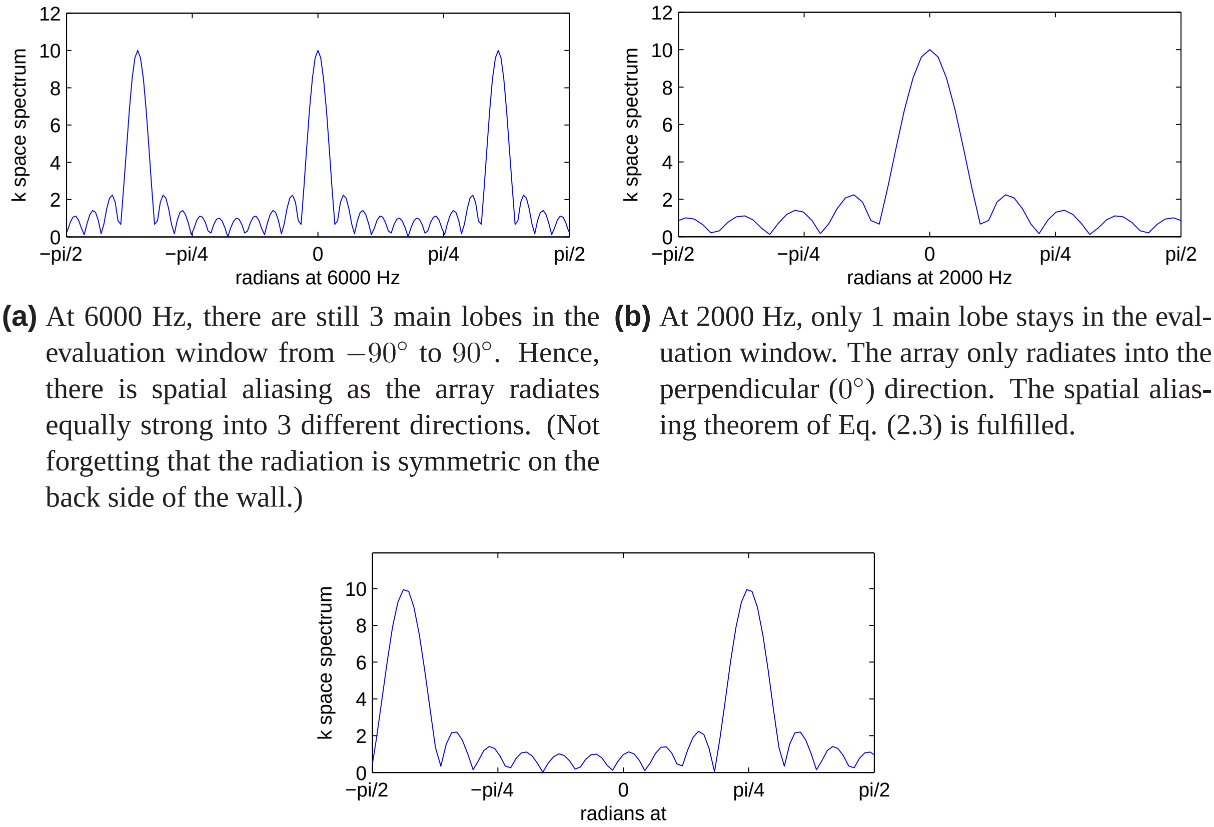
Figure 2.3: Sound particle velocity distribution of vibrating stripes along the y-axis and its spatial Fourier representation.

At a given speed of sound and a certain frequency, k_y only depends on ϕ. Thus, the k space can be evaluated for every frequency from $-90°$ to $90°$, where it can be read as a directivity diagram. The amplitudes of the k-space spectrum denote how much the array radiates into the corresponding direction. The lower the wavenumber, the less main lobes of the sinc function stay in the evaluated window. Spatial aliasing occurs as long as there is more than one main lobe in the evaluation window. Just as in time-frequency processing, the Nyquist theorem can be applied to the spatial domain, too. Spatial aliasing can be prevented as long as

$$\frac{\lambda}{2} \geq \Delta y, \tag{2.3}$$

where Δy is the distance between the vibrating stripes.

Fig. 2.4 shows the evaluation for two different frequencies. The k-space spectrum was derived from Fig. 2.3a in which $\Delta y = 8\,\text{cm}$. According to Eq. (2.3), spatial aliasing occurs at 6000 Hz but not at 2000 Hz.



(a) At 6000 Hz, there are still 3 main lobes in the evaluation window from $-90°$ to $90°$. Hence, there is spatial aliasing as the array radiates equally strong into 3 different directions. (Not forgetting that the radiation is symmetric on the back side of the wall.)

(b) At 2000 Hz, only 1 main lobe stays in the evaluation window. The array only radiates into the perpendicular ($0°$) direction. The spatial aliasing theorem of Eq. (2.3) is fulfilled.

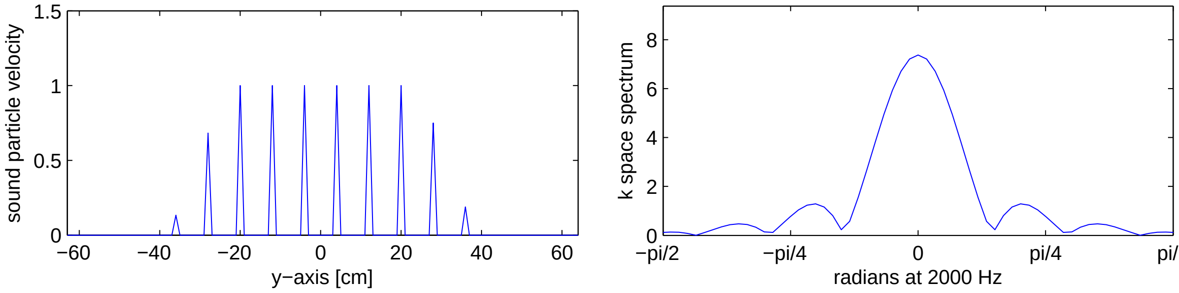
(c) First aliasing components appear, if the beam is steered to the side like in this case for a frequency slightly over the Nyquist frequency with a steering angle $\phi = -80°$.

Figure 2.4: For a particular frequency and a given speed of sound, the k-space spectrum only depends on ϕ and can be evaluated from $\frac{-\pi}{2}$ to $\frac{\pi}{2}$ or from $-90°$ to $90°$, respectively. It can then be read as a directivity diagram.

Spatial aliasing can be prevented at higher frequencies by decreasing the distance between the array elements. Again, this can be explained by signal theory as in Oppenheim [1989]: A narrower pulse train of sound particle velocity leads to a wider pulse train after the Fourier transform. Hence, there are less main lobes in one period of the k-space domain. The bandwidth in ATC reaches from 300 to 2500 Hz. This is a benefit for a loudspeaker array application, because spatial aliasing can be avoided if the distance between the loudspeakers Δy does not exceed 7 cm.

The beam is steered to an arbitrary direction ϕ through the delay times between the loudspeakers. This is also schematized in Fig. 2.1. The delay times cause a linear phase shift $e^{jk_y y\phi}$ of the sound-particle velocity that modulates the main lobe towards the steering angle ϕ. As a consequence, ambiguous main lobes might move into the evaluated directivity diagram, which is the case in Fig. 2.4c.

Comparing Fig. 2.4a and 2.4b also shows that the beam width depends on the frequency. The beam width is defined by the -3 dB decay of the beam. The smaller the wavelength, the smaller is the beam width. However, the narrowness of the beam is limited by $\frac{\lambda}{2}$ as described in Yon et al. [2003]. The second influence on the beam width is the spatial window. The spatial window determines the length of the array and how the loudspeakers are attenuated at the end positions. A shorter array (a shorter spatial window) causes a broader directivity pattern and vice versa. A smoother window (instead of the hard rectangle limitation of Fig. 2.3a) causes less side lobes, but widens up the main lobe, too, like it can be seen in Fig. 2.5.

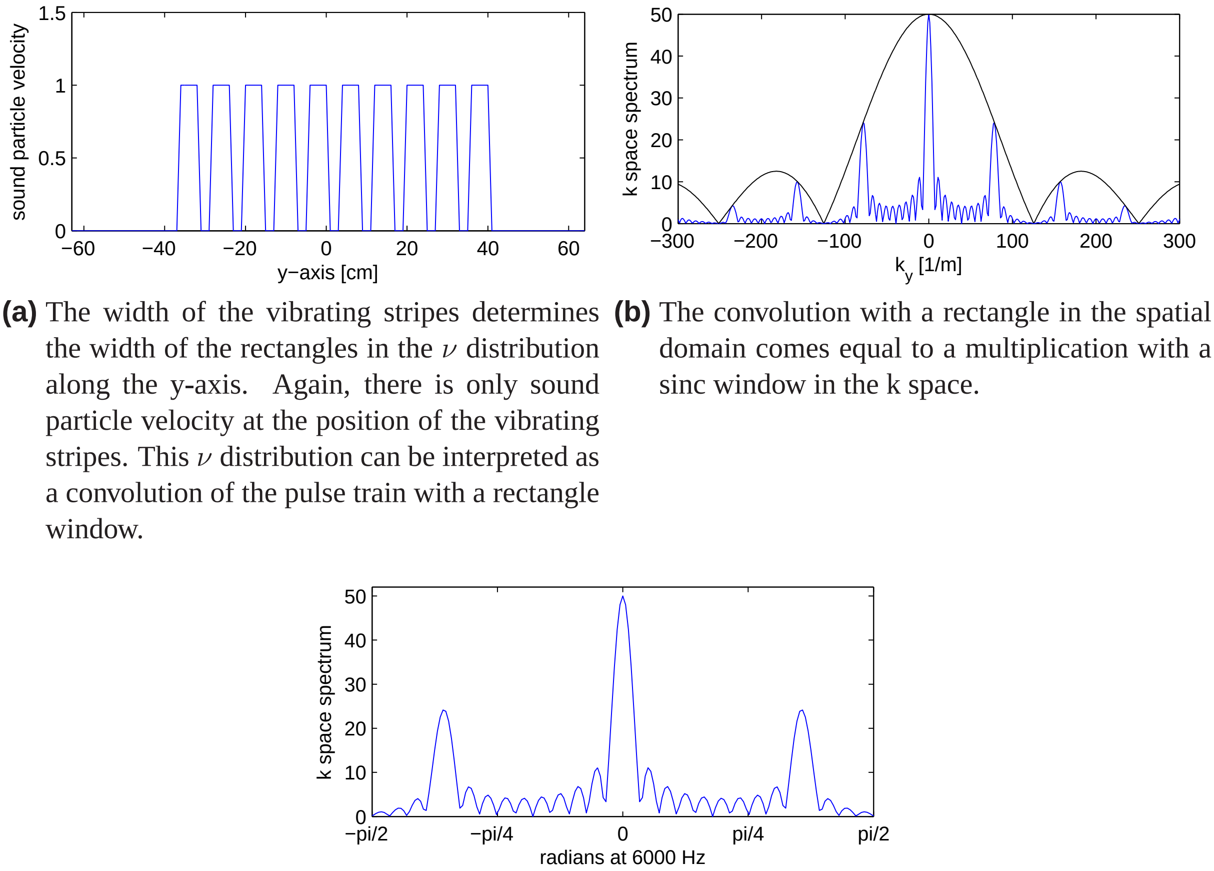


(a) The gains of the loudspeakers towards the ends of the array are smoothly reduced to zero by a Hann window.

(b) Compared to the rectangle window in Fig. 2.4b, the Hann window causes smaller side lobes but a wider main lobe in the directivity pattern.

Figure 2.5: A Hann window over the array elements compresses the side lobes but increases the width of the main lobe.

Until now, we considered the vibrating stripes to be infinitesimally thin. That is why we could model the sound particle velocity distribution as pulse train. The real stripes of Fig. 2.2 have a sound particle velocity that looks like the rectangles in Fig. 2.6a. This can be seen as a convolution of the pulse train with a rectangle window that has the width of the stripes. A convolution in the spatial domain corresponds to a multiplication in the k space, whereas a rectangle window changes to a sinc function through the Fourier transform. The result can be seen in Fig. 2.6c. The pulse in the center of the k space (at $k_y = 0$) is in the main lobe, while the other pulses will be more or less suppressed by the side lobes of the sinc function. The array looses its omnidirectionality, especially for high frequencies. This illustrates the fact, that a loudspeaker can only be considered as a point source, as long as the wavelengths are larger than the membrane diameter.



(a) The width of the vibrating stripes determines the width of the rectangles in the ν distribution along the y-axis. Again, there is only sound particle velocity at the position of the vibrating stripes. This ν distribution can be interpreted as a convolution of the pulse train with a rectangle window.

(b) The convolution with a rectangle in the spatial domain comes equal to a multiplication with a sinc window in the k space.

(c) As a consequence, the directivity of the array is not omnidirectional any more.

Figure 2.6: The influence the membrane width on the k-space spectrum and the directivity of the array.

2.2 Spatial Transmission Functions

2.2.1 Free Field Green's Functions

The sound pressure of a sound source at an arbitrary observation point can be predicted with the help of Green's function. The free field Green's function relates the pressure at source point $\mathbf{r}'$ to the pressure at an observation point $\mathbf{r}$ for ideal conditions. Which are: omnidirectional point sources, no reflections, lossless medium etc. It is then the solution of the Helmholtz equation for outgoing spherical waves as in Morse [1953]:

$$G(\mathbf{r}'|\mathbf{r}|\omega) = \frac{e^{-jk|\mathbf{r}'-\mathbf{r}|}}{|\mathbf{r}'-\mathbf{r}|}. \tag{2.4}$$

The Green's functions from all loudspeaker positions of an array $\mathbf{r_1'} \dots \mathbf{r_L'}$ to one specific focus point $\mathbf{r_f}$ can be gathered to a vector

$$\mathbf{h}(\omega) = \begin{bmatrix} G(\mathbf{r_1'}|\mathbf{r_f}|\omega) & G(\mathbf{r_2'}|\mathbf{r_f}|\omega) & \cdots & G(\mathbf{r_L'}|\mathbf{r_f}|\omega) \end{bmatrix}^T. \tag{2.5}$$

The sound pressure in the focus point p_{focus} can easily be calculated if the complex weights $\mathbf{q}(\omega)$ of the loudspeakers (it is their amplitude and phase) are known

$$p_f(\omega) = \mathbf{h}^T(\omega)\,\mathbf{q}(\omega). \tag{2.6}$$

Equally, N arbitrary other field points can be considered as observation or evaluation points. The Green's functions from every source point $\mathbf{r}'_l$ to every evaluation point $\mathbf{r}_n$ are gathered in the matrix

$$\mathbf{G}(\omega) = \begin{pmatrix} G(\mathbf{r_1'}|\mathbf{r_1}|\omega) & G(\mathbf{r_2'}|\mathbf{r_1}|\omega) & \cdots & G(\mathbf{r_L'}|\mathbf{r_1}|\omega) \\ G(\mathbf{r_1'}|\mathbf{r_2}|\omega) & G(\mathbf{r_2'}|\mathbf{r_2}|\omega) & \cdots & G(\mathbf{r_L'}|\mathbf{r_2}|\omega) \\ \vdots & \vdots & \ddots & \vdots \\ G(\mathbf{r_1'}|\mathbf{r}_N|\omega) & G(\mathbf{r_2'}|\mathbf{r}_N|\omega) & \cdots & G(\mathbf{r_L'}|\mathbf{r}_N|\omega) \end{pmatrix}. \tag{2.7}$$

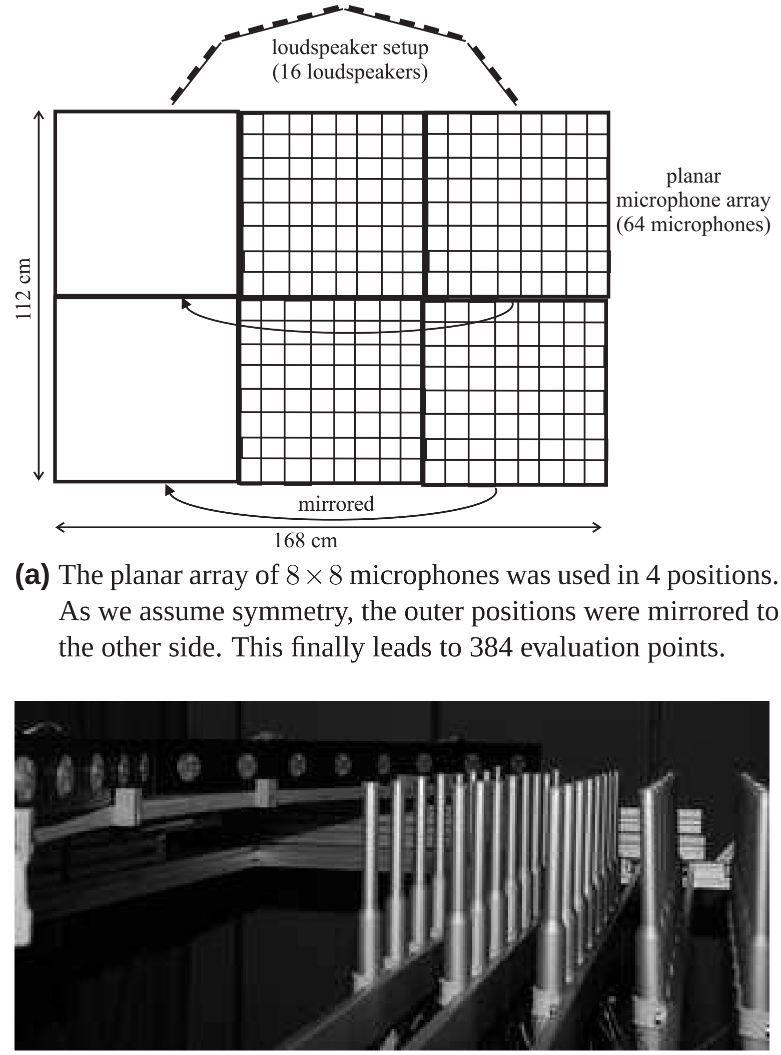
A multiplication of the source strength vector $\mathbf{q}(\omega)$ with this matrix yields the sound pressure vector $\mathbf{p}(\omega)$

$$\mathbf{p}(\omega) = \mathbf{G}(\omega)\mathbf{q}(\omega), \tag{2.8}$$

that contains the sound pressure in the N observation points $\mathbf{r}_n$. Thus the sound field of a beamformer with source strength vector $\mathbf{q}(\omega)$ can be simulated with the knowledge of matrix $\mathbf{G}(\omega)$.

2.2.2 Transfer Functions Measurement

The free field Green's functions serve as theoretical background for simulations. In order to evaluate a real beamformer, the transfer functions from an experimental loudspeaker setup to 256 field points were measured, too. The loudspeaker array should satisfy the spatial Nyquist theorem (Eq. (2.3)) and requires a width of 120 cm to cover the working area of air traffic controllers. These constraints have led to an array of 16 loudspeakers that are arranged to approximate an elliptical segment. This constellation proofed to bear focusing advantages over a straight array, as it can be read in Section 2.3.2 and in Guldenschuh et al. [2008]. A quadratic microphone array with a raster of 7×7 cm was used to prevent spatial aliasing up to 2500 Hz. The measurement setup is depicted in Fig. 2.7.



(a) The planar array of 8×8 microphones was used in 4 positions. As we assume symmetry, the outer positions were mirrored to the other side. This finally leads to 384 evaluation points.

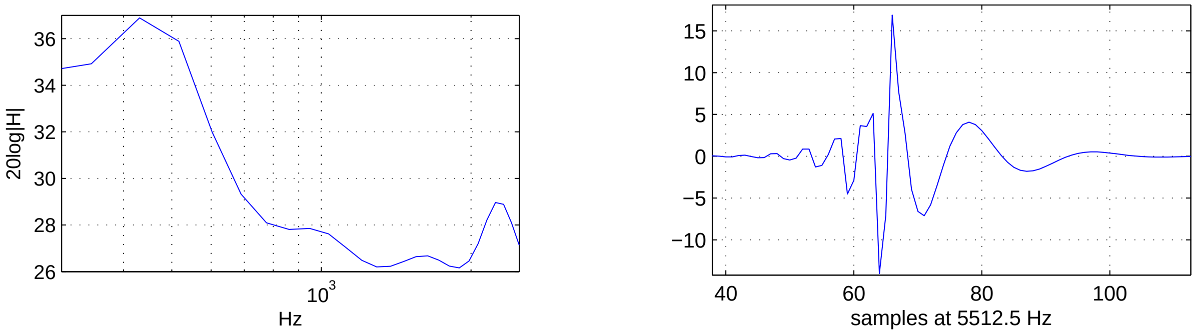
(b) Loudspeaker array with measurement microphones.

Figure 2.7: Measurement setup

The impulse responses were measured in a bandwidth from 300 to 2500 Hz with exponential sweeps as in Farina [2000]. The sweeps had a length of two seconds and were recorded with 44.1 kHz sampling frequency. In order to consider the influence of the loudspeakers and microphones, a reference measurement of all loudspeakers in one meter distance was used to equalize the sound field measurement. In the frequency domain, the equalization can be done by inverse filtering

$$H_{\text{measure.eq}} = H_{\text{measure}} H_{\text{ref}}^{-1}. \tag{2.9}$$

An equalization filter H_{ref}^{-1} is shown in Fig. 2.8.



(a) Frequency response of the inverse of H_{ref}. The used loudspeakers have a high pass characteristic. Therefore the equalization filter has to augment the basses.

(b) In the time domain, the inverse of H_{ref} has 60 taps at 5512.5 Hz sampling frequency.

Figure 2.8: Loudspeaker equalization filter in the frequency- and the time domain.

Unfortunately, early reflections from the microphone mounting device caused a comb filter that has its first notch around 2000 Hz. A sample of measured transfer functions can be seen in Fig. 2.9. Finally, the impulse responses were resampled at a 8 times lower rate and their length was reduced to 2.9 ms.

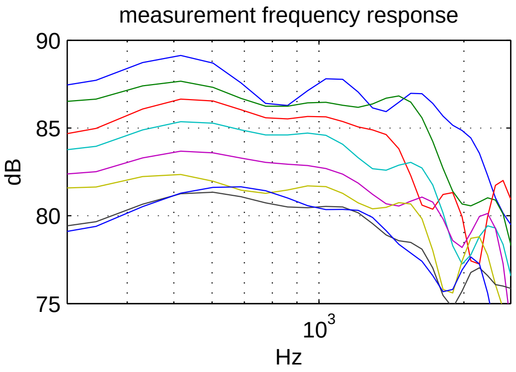


Figure 2.9: Frequency response of the first 8 loudspeakers to a central position in 90 cm distance. Early reflections from the microphone mounting device cause a first notch at around 2000 Hz.

Sound pressure levels (SPL) L_p are derived by the mean value over the squared amplitude of every frequency bin ξ within the desired bandwidth.

$$\overline{p}^2 = \frac{1}{33} \sum_{\xi=1}^{33} \|p(\xi)\|^2 \tag{2.10}$$

$$L_p = 10 \log \frac{\overline{p}^2}{p_0^2} \qquad \text{with } p_0 = 20\mu\text{Pa}. \tag{2.11}$$

In the following all sound field evaluations (also the free field simulations) were done at the sampling frequency of 5512.5 Hz with a resolution of 33 frequency bins. All sound field representations are normed by the sound pressure level at the focus point.

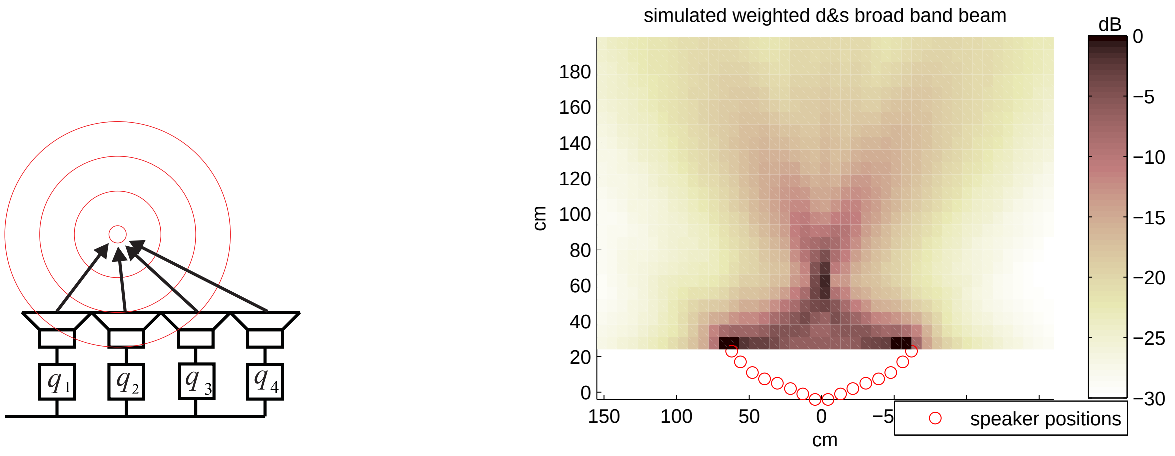
2.3 Weighted Delay & Sum Beamforming

To create a pressure concentration in a sound field, the signals of an array have to superpose constructively at a focus point. Hence, they have to reach the focus point at the same time. It follows that the source strength vector $\mathbf{q}(\omega)$ has to be the complex conjugate of the Green's functions $\mathbf{h}(\omega)$

$$\mathbf{q}(\omega) = \mathbf{h}^{*}(\omega), \tag{2.12}$$

to compensate for the phase differences of the Green's functions. The Green's functions are weighted with the reciprocal of the distance from the loudspeaker to the focus point (see Eq. (2.4)). Thus, the source strength vector $\mathbf{q}(\omega)$ is weighted with this reciprocal, too. These weights can also be interpreted as a window that suppresses the loudspeakers that are further away from the focus point. Obviously, the frequency response of real loudspeakers has to be equalized to receive unity gain in the focus point. In contrast to the beamformer in Fig. 2.1, the Weighted Delay & Sum Beamformer (WDSB) has one individual complex weight per loudspeaker, like depicted in Fig. 2.10a.

Cho and Roan [2009] pointed out that it would make sense to weight $\mathbf{q}(\omega)$ with the reciprocal of the square or the cube of the distance if the focus points are in the near field ($kr < 2$). The lowest case of our application, however, can be estimated as $kr = 2.2$ with $k = \frac{2\pi 300\,\text{Hz}}{c}$ and $r = 0.4$ m. Therefore no attention has to be paid to near field weighting. Fig. 2.10b shows a simulated beam in the bandwidth of 300-2500 Hz.



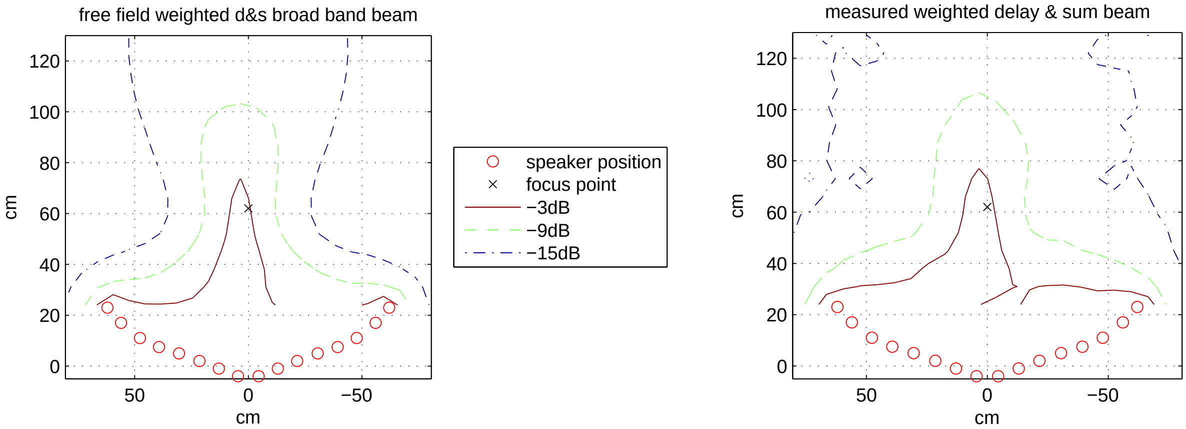
(a) Weighted Delay & Sum beamforming: Each loudspeaker has its own complex weight such that the waves superpose in the focus point.

(b) Broad-band sound pressure distribution of a WDSB.

Figure 2.10: Weighted Delay & Sum beamforming.

2.3.1 Comparison 1: Measured Beam - Simulated Beam

The beams of the ideal free field Green's functions were compared with the beams of the measured transfer functions. The first ones will be called simulated beams, while the latter ones will be called measured beams from now on.[1] SPL lines of a simulated and a measured broad-band beam are depicted in Fig. 2.11.



(a) Simulation, based on the free field Green's functions.

(b) Simulation, based on the measured transfer functions.

Figure 2.11: Comparison between the simulated and the measured beam in the close central position in the full bandwidth of 300-2500 Hz. The pressure decay is depicted in three level lines.

The comparison was done for four focus positions (marked in Fig. 2.12b), in 12 bark bands in the 384 evaluation points of Fig. 2.7a. Bark bands simulate the frequency dependent sensitivity of the ear and are therefore also called critical bands. Their bandwidths were empirically determined by Zwicker [1990]. The used bark bands are listed in Table 2.12a. The measurements shall proof the concept and justify further simulations based on the free field Green's functions.

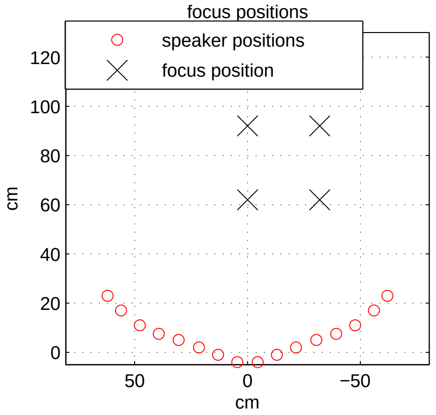
Before the SPLs of the measured and the simulated sound field can be compared, the energy of the measured sound field $\overline{\mathbf{p}}_m^2$ has to be calibrated to the total energy of the simulated sound field.

$$\overline{\mathbf{p}}_{m,\mathrm{cal}}^2 = \overline{\mathbf{p}}_m^2 \frac{\sum_{i=1}^{384} \overline{p}_{\mathrm{sim},i}^2}{\sum_{i=1}^{384} \overline{p}_{m,i}^2} \tag{2.13}$$

[1] Although the measured beams are simulated, too, but with data of the measured transfer functions.

bark	f_l [Hz]	f_u [Hz]
4	300	400
5	400	510
6	510	630
7	630	770
8	770	920
9	920	1080
10	1080	1270
11	1270	1480
12	1480	1720
13	1720	2000
14	2000	2320
15	2320	2700

(a) Lower and upper cut off frequencies of bark bands between 300 and 2700 Hz.



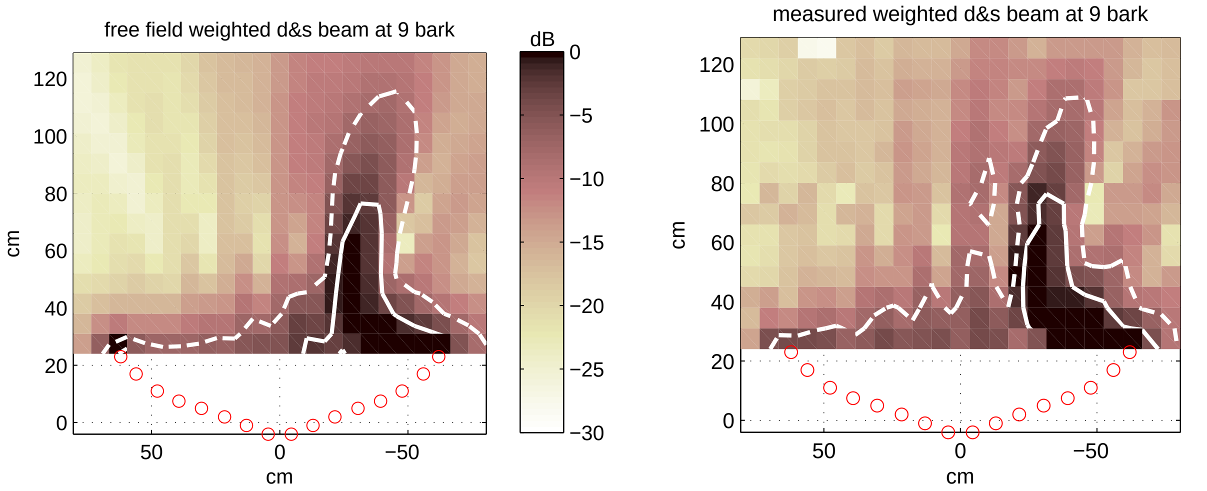
(b) Reference positions. As we assume symmetry of the sound field, all positions are chosen to be on the negative side of the x-axis.

Figure 2.12: Bark bands and focus positions for the comparison between simulation and measurement.

The comparison will exemplarily be shown at 9 bark for the close side position. The SPL distributions are shown in Fig. 2.13, and the absolute value of their differences in Fig. 2.14a. Errors outside of the focused area are less relevant, which is why a weighted difference e_w is introduced. The weights are the square root of p_{sim}, normed by the pressure in the focus point:

$$e_{w,i} = |L_{m,i} - L_{\text{sim,i}}| \sqrt{\frac{p_{\text{sim,i}}}{p_{\text{focus}}}} \tag{2.14}$$

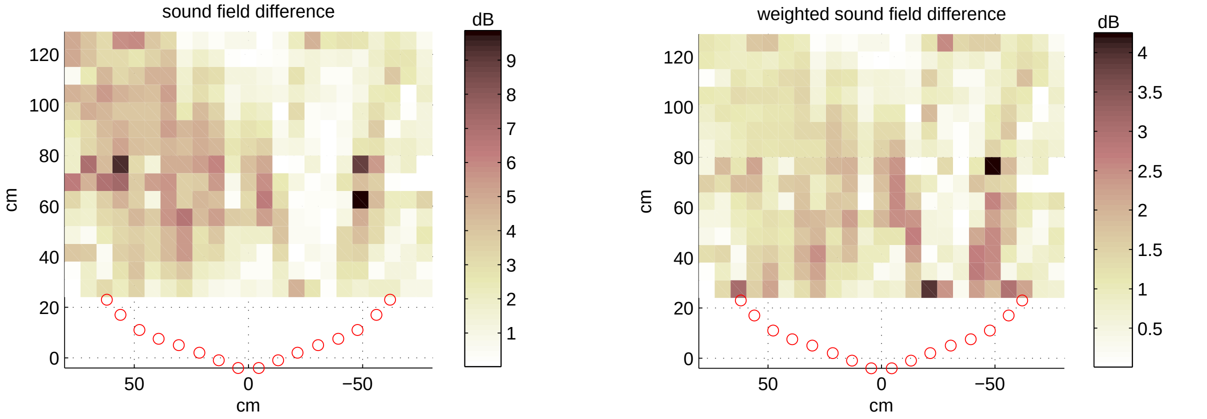
As a consequence, the error in the focus point stays the same, while errors in regions of low SPL are compressed. The weighted difference is shown in Fig. 2.14b.



(a) Simulation, based on the free field Green's functions.

(b) Simulation, based on the measured transfer functions.

Figure 2.13: Comparison between the simulated and the measured beam in the close side position in the 9$^{\mathrm{th}}$ bark band. The solid white line represents the -3 dB level line and the dashed line the -9 dB level line.



(a) SPL differences in the close side position at 9 bark. The highest difference is 9 dB.

(b) SPL differences, weighted with the square root of p_{sim}. Differences in regions of low SPL (like in the upper left corner) are suppressed. The highest difference is 4 dB.

Figure 2.14: SPL differences and weighted SPL differences between the measured and the simulated sound field in the 9$^{\mathrm{th}}$ bark band for the close side focus-position.

The mean value and the third quartile over all SPL differences and weighted differences are given in Fig. 2.15. The mean values of the weighted differences are lower than 2 dB, which is a very satisfying result. Values for the third quartile that are larger than 5 dB only occur for the unweighted difference and hence in regions of low SPL and minor interest. The measurement results give reason to the usage of the free field Green's functions, which will therefore be used for further simulations.

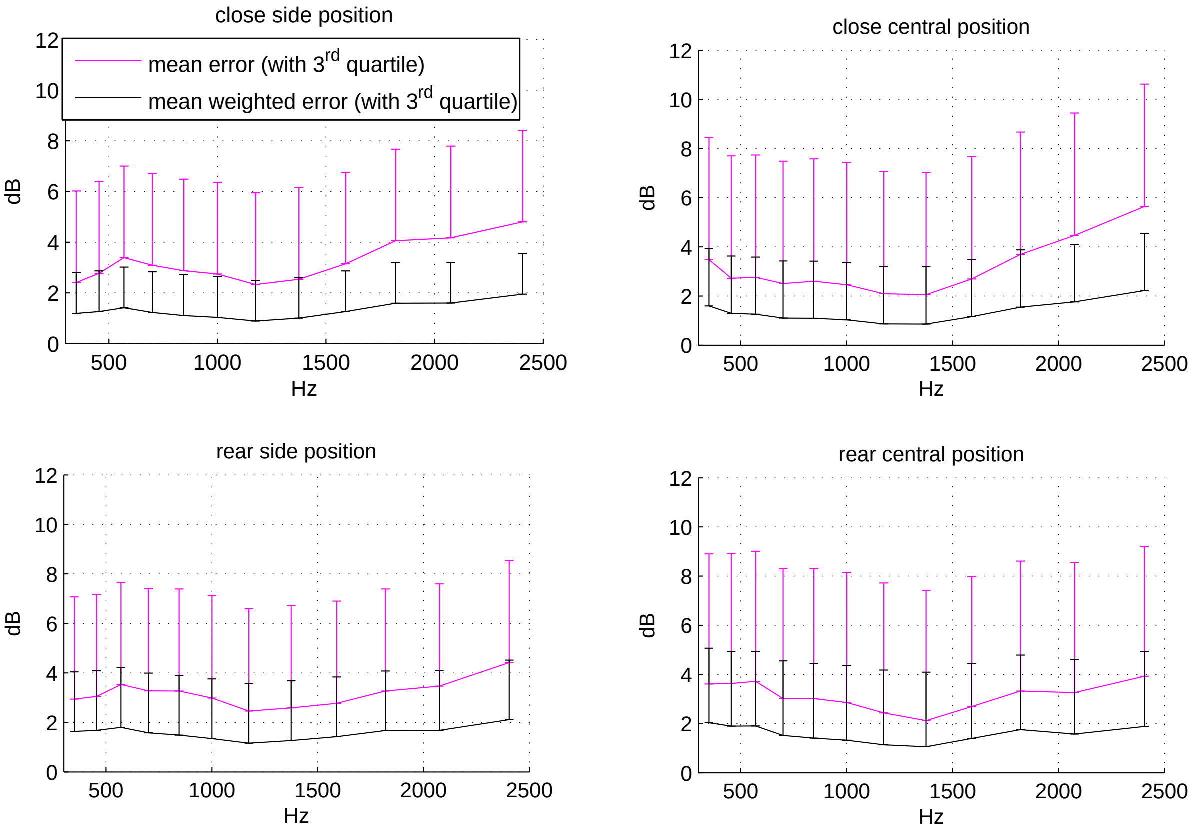


Figure 2.15: SPL difference and weighted difference between the measured and the simulated sound field for four focus positions. Differences are bigger at high frequencies, where the areas of constructive and destructive superposition are smaller. A little phase error can then cause a constructive superposition at a location where the simulation predicts a destructive superposition, or vice versa. In addition, frequencies around 2000 Hz suffer from the comb filter notch described in Fig. 2.9.

2.3.2 Comparison 2: Straight Array - Bent Array

The sound focusing properties of parabolic and elliptical reflectors are well known in room acoustics like in Fasold [1998]. In the following, the focusing properties of a bent array will be compared to the ones of a straight array. Therefore, two measures are introduced:

1. SNR$_{2D}$: For the 2D comparison, the evaluation points of Fig. 2.7a will be used. The SNR$_{2D}$ is then the relation between the sound energy in the focus point and the average sound energy in all other 383 evaluation points in dB:

$$\mathbf{SNR}_{2D} = 10 \log \left(383 \frac{\overline{p}^2_{\text{focus}}}{\sum_i \overline{p}^2_i} \right) \tag{2.15}$$

2. Spatial Rejection Ratio (SRR): As SRR, we define the difference between the sound pressure level in the focus point and the sound pressure level of the reverberant room $L_{\text{focus}} - L_r$. In Ahnert [1993] it is shown that L_r can be estimated as

$$L_r = 10 \log \frac{P_{ac}}{P_0} - 10 \log A + 6 \,\text{dB}, \tag{2.16}$$

where A is the sum of absorbing surfaces, P_0 is the reference sound power of $10^{-12}\,\text{W}$ and P_{ac} is the acoustic power of the array. P_{ac} is derived by integrating the sound intensity

$$\vec{I} = p\vec{\nu} \tag{2.17}$$

over the surface S of a sphere with radius r. If r is much greater than the array dimension l

$$r \gg l, \tag{2.18}$$

the direction of $\vec{\nu}$ approximates the normal of the spherical surface for every loudspeaker in every point of the sphere. Thus, $\vec{I}$ can be approximated as

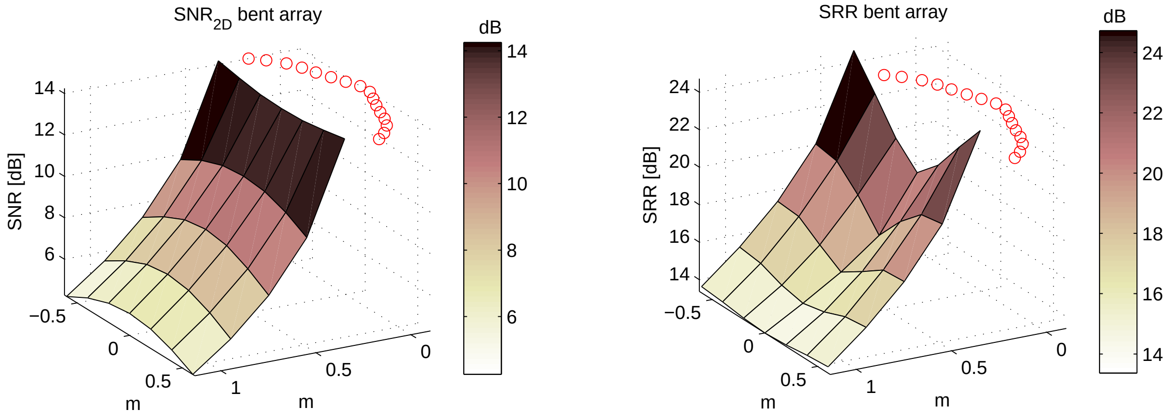
$$I = \frac{\overline{p}^2}{\rho c}. \tag{2.19}$$

As the sound intensity is determined numerically, the integral has to be transformed to a sum over N discrete points

$$P_{ac} = \frac{1}{N} \sum_{n=1}^{N} \overline{p}^2_n \frac{4\pi r^2}{\rho c}, \tag{2.20}$$

where ρc is the acoustic impedance which has 408 $\frac{\text{kg}}{\text{m}^2\text{s}}$ at 20 °C.

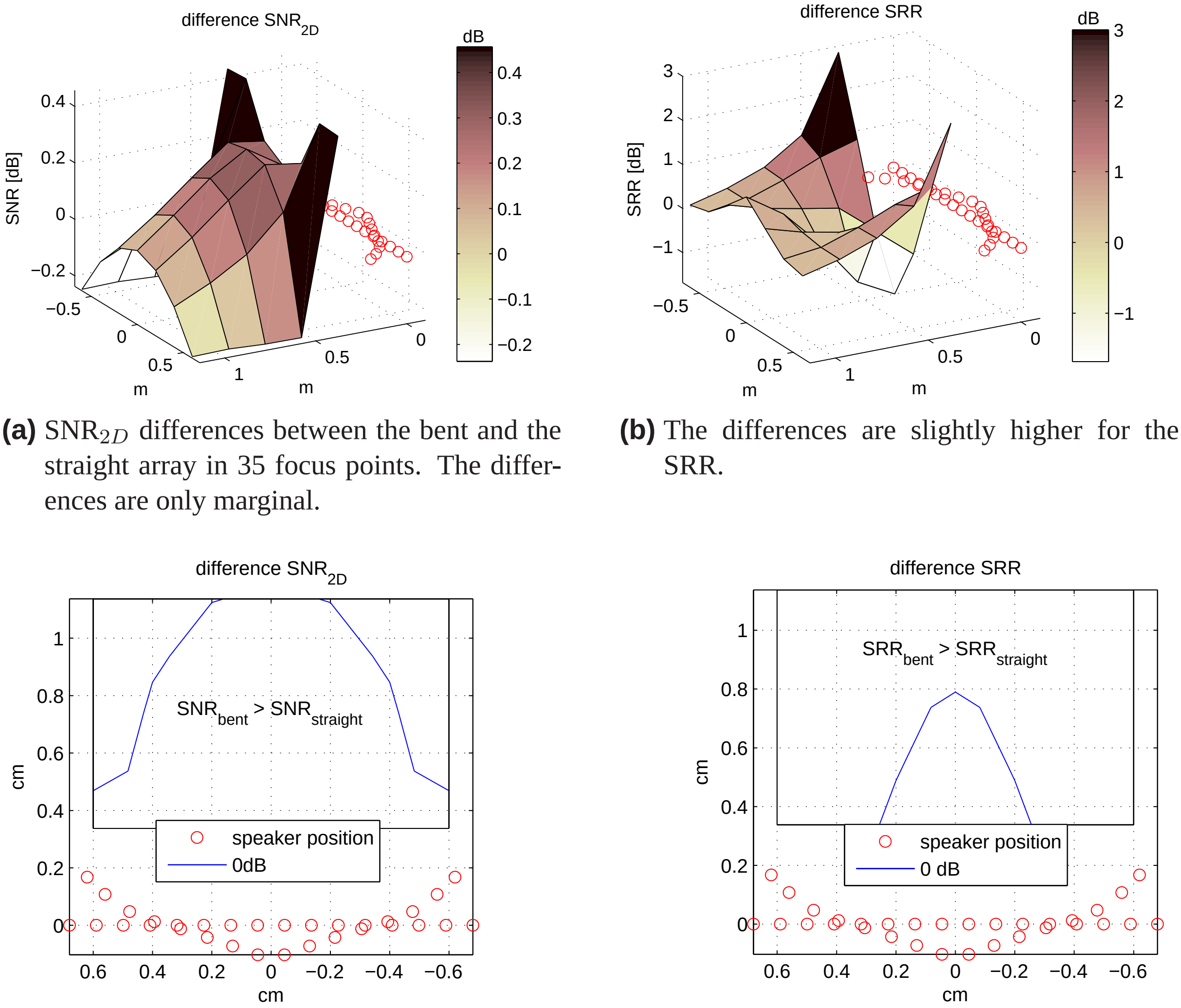
Simulations for 35 focus points are done to compare the two arrays. The centroids of the arrays are put into the coordinate origin and r is chosen to be 10 m to account for Eq. (2.18). The sphere is sampled on $N = 7482$ surface points to satisfy the spatial aliasing constraint of Eq. (2.3) on a sphere with radius 1.7 m. This radius encloses the array and the farthest focus point. The sum of reflecting surfaces A in air traffic control centers is assumed to be 200 m^2. The results for the bent array can be seen in Fig. 2.16, and the differences to the straight array (SNR$_{\mathrm{bent}}$ - SNR$_{\mathrm{straight}}$) are shown in Fig. 2.17.



(a) Obviously the horizontal sound field is stronger excited if the beam is steered further away from the array. As a consequence the SNR$_{2D}$ decreases.

(b) Also the room is stronger excited if the beam is steered further away. A remarkable notch appears for the close central focus positions. This effect is explained later on.

Figure 2.16: SNR$_{2D}$ and SRR for 35 focus positions. The array performs worse if the focus point is far away from the array.



(a) SNR$_{2D}$ differences between the bent and the straight array in 35 focus points. The differences are only marginal.

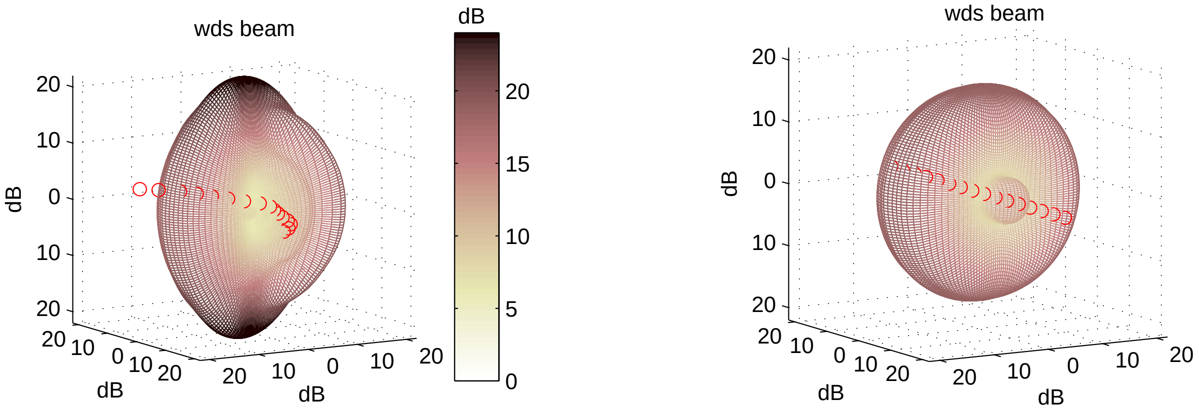
(b) The differences are slightly higher for the SRR.

(c) The 0 dB line marks the border where the bent array performs better in terms of the 2D SNR.

(d) In terms of the SRR, the bent array performs worse for close central focus positions. This effect is further investigated in Fig. 2.18.

Figure 2.17: Focusing differences of a bent and a straight array. The bent array performs slightly better for both measures. The highest differences are reached for focus points in the near corners, where the bent array benefits from its closer position.

The most striking outcome is that the bent array performs worse for close central focus points, although one would assume (also by reason of the SNR_{2D} values) that focusing works more efficient if the focus point is better surrounded by the loudspeakers. In the 2D case, this is true, but Keele [2003] pointed out that bent line arrays have a strong radiation orthogonal to their expansion plane. This effect can be seen in Fig. 2.18.



(a) The bent array has a high directivity in the vertical direction, even though the driving functions aim to steer into the horizontal plane.

(b) As expected, the directivity is concentric around the straight array.

Figure 2.18: 3D directivity of a straight and a bent array for a close central focus point in the horizontal plane.

It can be concluded that the bent array has slightly better focusing properties than a straight array. Above all, the bent array performs better for focus positions in the close side positions, where it benefits from the close loudspeakers.

Weighted Delay & Sum Beamforming is a physical straightforward beamforming method. It simply aims to focus sound in a specific point by compensating the delay times. Additional weights prevent loudspeakers from exciting the sound field too much if they are further away from the focus point. A WDSB can therefore be easily implemented as dynamic real-time application. The input signals are led into a delay line and each output channel is weighted with one multiplier. This means that all frequencies are equally weighted and, as a consequence, the frequency response at the focus point is flat. Superdirective beamformers use optimization algorithms to narrow the beam width or to enforce regions of low SPL. In general, the amplitude and phase of their driving function is frequency dependent. So their implementation requires filtering. The following 3 sections introduce and investigate such superdirective beamformers.

2.4 Least Squares Beamforming

The Least Squares (LS) algorithm finds the source strength vector that causes an output (in our case a sound field) that matches a target function with the smallest (least) squared error. Our target function $\tilde{\mathbf{p}}$ consists of p_{focus} and the sound pressure in N other field points

$$\tilde{\mathbf{p}}(\omega) = \begin{pmatrix} p_f \\ \mathbf{p} \end{pmatrix}. \tag{2.21}$$

Complementary, we combine the Green's functions from the loudspeakers to the focus point with the Green's functions to the other N field points

$$\tilde{\mathbf{G}}(\omega) = \begin{bmatrix} \mathbf{h}^T \\ \mathbf{G} \end{bmatrix}. \tag{2.22}$$

The error between the target function and the outcome of the beamformer reads as

$$\mathbf{e} = \tilde{\mathbf{p}} - \tilde{\mathbf{G}}\mathbf{q}, \tag{2.23}$$

the squared error results in the cost function

$$J(\mathbf{q}) = e^2 = \mathbf{e}^H\mathbf{e} = \tilde{\mathbf{p}}^H\tilde{\mathbf{p}} - 2\mathbf{q}^H\tilde{\mathbf{G}}^H\tilde{\mathbf{p}} + \mathbf{q}^H\tilde{\mathbf{G}}^H\tilde{\mathbf{G}}\mathbf{q}. \tag{2.24}$$

The first derivative of $J(\mathbf{q})$ describes the tangents of the error function. The tangent is flat (i.e. $J\frac{d}{d\mathbf{q}} = 0$) where $J(\mathbf{q})$ has a minimum or a maximum. As the cost function is quadratic and positive, the solution of the LS algorithm finds the minimum of the function. The derivative of J with respect to $\mathbf{q}$ is

$$-2\tilde{\mathbf{G}}^H\tilde{\mathbf{p}} + 2\tilde{\mathbf{G}}^H\tilde{\mathbf{G}}\mathbf{q} \overset{!}{=} 0, \tag{2.25}$$

and the Least Squares solution of the source strength vector is

$$\mathbf{q} = \left(\tilde{\mathbf{G}}^H\tilde{\mathbf{G}}\right)^{-1}\tilde{\mathbf{G}}^H\tilde{\mathbf{p}}. \tag{2.26}$$

From Eq. (2.23) on, the dependency on ω is omitted for reasons of compactness. In fact it is important that the LS solution has to be evaluated at every frequency. The stability of the LS filter depends on the number of considered frequency bins. However, the LS optimization for a limited number of frequency bins does not guarantee a flat frequency response in the focus point. This problem will be further discussed in Section 2.7.

An example of a Least Squares beam (LSB) is depicted in Fig. 2.19. The sound pressure is minimized within a frame around the arrangement. This frame is meant to surround the working space of an air traffic controller. The chosen target function demands unity gain at the focus point and zeros at all other control points such that

$$\mathbf{q} = (\tilde{\mathbf{G}}^H \tilde{\mathbf{G}})^{-1} \tilde{\mathbf{G}}^H \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}^T . \tag{2.27}$$

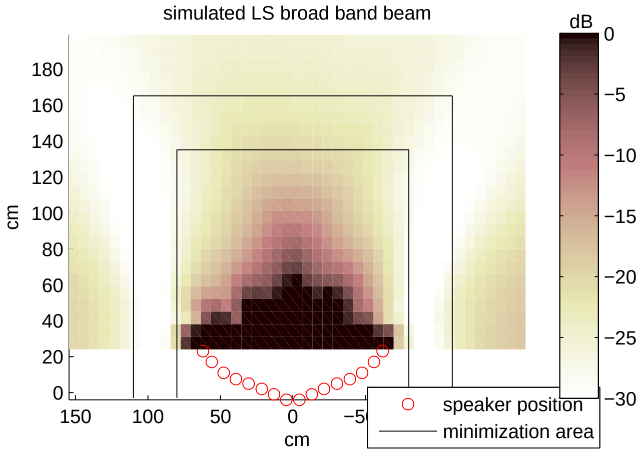


Figure 2.19: Broad-band (300-2500 Hz) sound pressure distribution of a LSB. The desired sound pressure is zero in the minimization area should be zero and has unity gain at the focus point.

For the targeted transaural application, the beams are aimed to be steered to the listener's ears. On the one hand, this minimizes the radiated sound energy and on the other hand it reduces the cross talk. In Chapter 3 it is explained that the cross talk increases with the wavelength. Therefore it would make sense to set the focus point in dependence of the frequency like illustrated in Fig. 2.20.

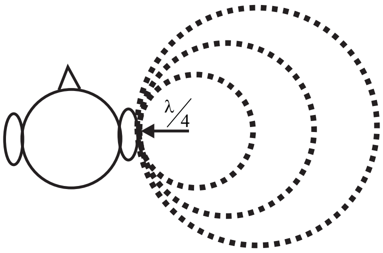


Figure 2.20: Several focus points can be located as a frequency dependent circle such that its radius to the ear approximates $\frac{\lambda}{4}$. This reduces the cross talk to the contralateral ear.

For the beam depicted in 2.21, not only the position of the focus is set frequency dependent, but also the number of considered focus points that build the focus circle. Low frequencies (that have a larger wavelength) are weighted with more focus points.

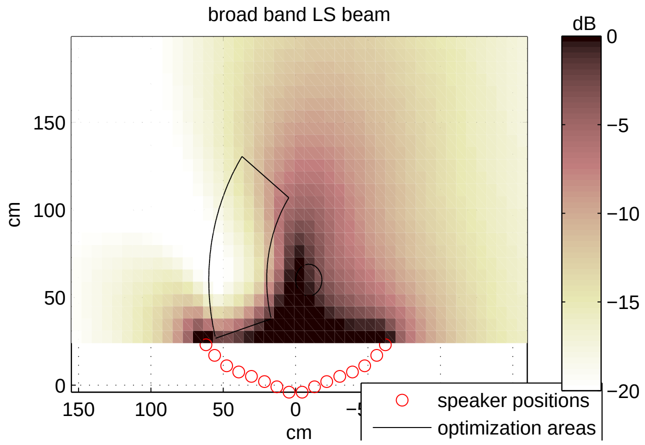


Figure 2.21: Broad-band (300-2500 Hz) sound pressure distribution of LSB with a frequency dependent focus area. For this illustration the focus area (i.e. the circle) of a middle wavelength was drawn into the pressure distribution. The sound pressure in the shaped area is desired to be zero, again.

A more detailed analysis of the LSBs will follow in Section 2.7, but it can already be anticipated that the inversion of matrix $(\mathbf{G}^H\mathbf{G})$ can cause trouble. If the ratio between the smallest and the largest singular value of $(\mathbf{G}^H\mathbf{G})$ (i.e. the condition number) becomes too large, two problems might occur:

1. Numerical round-off errors.
2. Small variations of the matrix elements (e.g. differences between the idealized and the real transfer functions) lead to large aberrations after inversion.

2.4.1 Singular Value Decomposition

Smaller condition numbers can be gained if the inverse matrix is regularized with the singular value decomposition (SVD). The SVD decomposes a $n \times m$ matrix $\mathbf{M}$ into

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{2.28}$$

where $\mathbf{U}$ is a $n \times n$ and $\mathbf{V}$ a $m \times m$ unitary matrix. $\mathbf{\Sigma}$ is a diagonal matrix with the size of $\mathbf{M}$ which contains the singular values of $\mathbf{M}$ in decreasing order on its diagonal. The previously mentioned errors can be reduced if small singular values are neglected. A commonly used criterion for the threshold of this regularization is derived from the energy of the singular values. Only the largest l singular values that reach a demanded threshold of the total energy are considered. Once the matrix is decomposed as in Eq. 2.28, the regularized inverse can easily be obtained by taking the l first rows and columns of $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$, yielding

$$\mathbf{M}_l^{-1} = \mathbf{V}_l\mathbf{\Sigma}_l^{-1}\mathbf{U}_l^H. \tag{2.29}$$

For the following simulations with SVD regularization, a threshold of 99% of the total energy has been chosen.

The next section, however, introduces a beamforming method that works without matrix inversion.

2.5 Maximum Energy Difference Beamforming

The Maximum Energy Difference Beamformer (MEDB) introduced by Shin et al. [2009] tries to maximize the difference between the sound energy in the focus point and the average sound energy in the other control points. Shin et al. [2009] define the sound energy $E(\omega)$ as

$$E(\omega) = p(\omega)^* p(\omega). \tag{2.30}$$

In the focus point E_{focus} can be predicted by

$$\begin{aligned} E_{\text{focus}} &= (\mathbf{h}^T\mathbf{q})^*\mathbf{h}^T\mathbf{q} \\ &= \mathbf{q}^H\mathbf{h}^*\mathbf{h}^T\mathbf{q}. \end{aligned} \tag{2.31}$$

The average sound intensities of the other N control points can be rewritten in the same way

$$E = \frac{1}{N}\mathbf{q}^H\mathbf{G}^H\mathbf{G}\mathbf{q}. \tag{2.32}$$

The difference of sound intensities yields

$$\begin{aligned} E_{\text{focus}} - E &= \mathbf{q}^H\mathbf{h}^*\mathbf{h}^T\mathbf{q} - \mathbf{q}^H\mathbf{G}^H\mathbf{G}\mathbf{q} \\ &= \mathbf{q}^H\left(\mathbf{h}^*\mathbf{h}^T - \mathbf{G}^H\mathbf{G}\right)\mathbf{q}. \end{aligned} \tag{2.33}$$

To obtain a meaningful cost function, this difference requires another constraint. It is set into relation with the power of the source strength vector, which is proportional to $\mathbf{q}^H\mathbf{q}$. We get

$$J(\mathbf{q}) = \frac{\mathbf{q}^H\left(\mathbf{h}^*\mathbf{h}^T - \alpha\mathbf{G}^H\mathbf{G}\right)\mathbf{q}}{\mathbf{q}^H\mathbf{q}}, \tag{2.34}$$

wherein α is a tuning factor that allows to put more or less weight on the sound energy at the focus point alone. If $\alpha = 0$, the cost function relates the energy at the focus point with the power of the source strength vector. In microphone array literature, like e.g. in Brandstein and Ward [2001], this relation is known as White Noise Gain (WNG)

$$\text{WNG}(\omega) = 10\log\left(\frac{|\mathbf{h^Tq}|^\mathbf{2}}{\mathbf{q^Hq}}\right). \tag{2.35}$$

For microphone arrays, the WNG compares the amplification of the focus point with the amplification of spatially uncorrelated noise. For loudspeaker arrays, it compares the input power with the outcome at the focus point, but is still referred to as white noise gain here.

In order to differentiate Eq. (2.34), it is rearranged to

$$\mathbf{q}^H\mathbf{q}J(\mathbf{q}) = \mathbf{q}^H \left(\mathbf{h}^*\mathbf{h}^T - \alpha\mathbf{G}^H\mathbf{G}\right)\mathbf{q}. \tag{2.36}$$

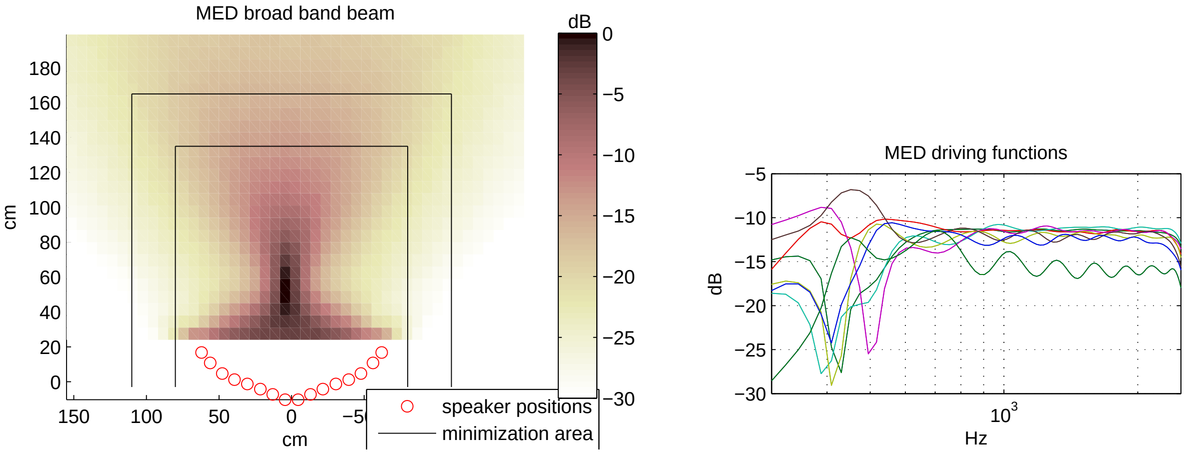
The differentiation of J with respect to $\mathbf{q}$ is

$$2\mathbf{q}J(\mathbf{q}) + \mathbf{q}^H\mathbf{q}J(\mathbf{q})\frac{d}{d\mathbf{q}} = 2\left(\mathbf{h}^*\mathbf{h}^T - \alpha\mathbf{G}^H\mathbf{G}\right)\mathbf{q}. \tag{2.37}$$

$J(\mathbf{q})$ has a maximum where $J(\mathbf{q})\frac{d}{d\mathbf{q}} = 0$. It follows that

$$\mathbf{q}J(\mathbf{q}_{\text{opt}}) = \left(\mathbf{h}^*\mathbf{h}^T - \alpha\mathbf{G}^H\mathbf{G}\right)\mathbf{q}. \tag{2.38}$$

This is a eigenvalue problem, and $\mathbf{q}_{\text{opt}}$ is the eigenvector that corresponds to the highest eigenvalue $J(\mathbf{q}_{\text{opt}})$. The MEDB is thus a beamforming algorithm that can be solved without matrix inversion. The MEDB that tries to maximize the sound energy difference between the focus point and the already known minimization area is depicted in Fig. 2.22. A value of $\alpha = 36$ brings a trade off between the WDSB and the LSB.



(a) Broad-band (300-2500 Hz) SPL distribution of a MEDB. The sound energy difference between the focus point and the frame around the arrangement is maximized.

(b) The MED driving functions for the given control points and a tuning factor $\alpha = 36$ have notch filters at low frequencies in order to reach a narrow beam width.

Figure 2.22: SPL distribution and control point disposition of a MEDB with corresponding driving functions.

2.6 Minimum Variance Distortionless Response Beamforming

The LSB and the MEDB are both optimization methods that have to be applied to several frequency bins. Still, they do not have a constraint of producing a flat frequency response at the focus point (see also Sec. 2.7). The Minimum Variance Distortionless Response Beamformer (MVDR) minimizes the squared sound pressure in the control points and enforces the source strength vector to produce unity gain at the focus point. Its constraint is

$$1 = \mathbf{h}^T \mathbf{q} = \mathbf{q}^H \mathbf{h}^*. \tag{2.39}$$

The optimization problem can be solved with the Lagrange multiplier λ. The Lagrange function $J(\mathbf{q})$ consists of the function to be minimized plus the constraint function, multiplied with λ

$$J(\mathbf{q}) = \mathbf{q}^H \mathbf{G}^H \mathbf{G} \mathbf{q} + \lambda \left(\mathbf{h}^T \mathbf{q} - 1 \right). \tag{2.40}$$

Its derivative with respect to $\mathbf{q}$ is set to zero

$$0 \overset{!}{=} 2\mathbf{q}^H \mathbf{G}^H \mathbf{G} + \lambda \mathbf{h}^T, \tag{2.41}$$

and delivers $\mathbf{q}_{\text{opt}}$, the location of the minimum of the Lagrange function

$$\mathbf{q}_{\text{opt}}^H = -\frac{\lambda}{2} \mathbf{h}^T \left(\mathbf{G}^H \mathbf{G} \right)^{-1}. \tag{2.42}$$

Inserting this solution into equation (2.39) yields $-\frac{\lambda}{2}$

$$1 = -\frac{\lambda}{2} \mathbf{h}^T \left(\mathbf{G}^H \mathbf{G} \right)^{-1} \mathbf{h}^*. \tag{2.43}$$

Hence, the optimal solution for $\mathbf{q}$ is

$$\mathbf{q}_{\text{opt}}^H = \frac{\mathbf{h}^T \left(\mathbf{G}^H \mathbf{G} \right)^{-1}}{\mathbf{h}^T \left(\mathbf{G}^H \mathbf{G} \right)^{-1} \mathbf{h}^*}. \tag{2.44}$$

A broad-band MVDR beam for the given minimization area is depicted in Fig. 2.23. Like for the LSB, matrix inversion is required to get a solution for $\mathbf{q}_{\mathrm{opt}}$. The influence of SVD regularization on the MVDR beam is investigated in the following section. Yet other possibilities for LS- and MVDR constraints can be found in Guldenschuh and Sontacchi [2009].

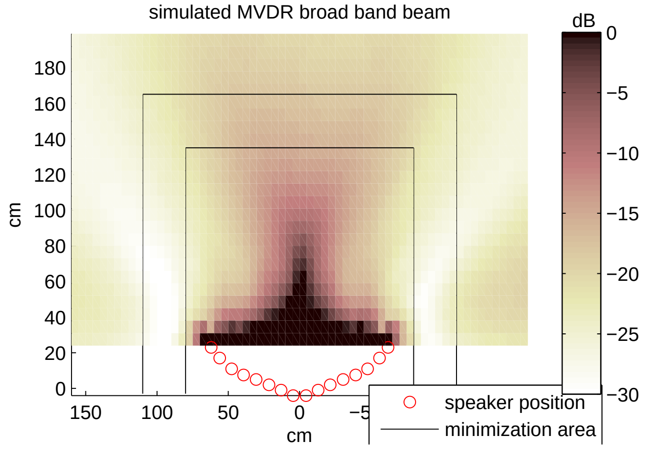


Figure 2.23: Broad band (300-2500 Hz) pressure distribution of a MVDRB. The pressure in the control points should be minimized, except for the focus point in the center, which is constrained to yield a constant gain for all frequencies.

2.7 Comparison of the Introduced Beamforming Methods

The four introduced beamforming methods will be compared in terms of their:

- Sound pressure attenuation
- Spatial Rejection Ratio
- Frequency response in the focus point
- White Noise Gain
- Filter length and,
- Condition number in case of matrix inversion

The last four points are explicitly discussed in Section 2.7.1 to 2.7.4. An overall comparison of the beamforming performances is given in Section 2.7.5. A central focus position is chosen as reference for the comparison. The problems and benefits of the different beamforming methods can be well illustrated for this reference. The minimization areas of the different methods are depicted in the previous sections. Only the LSB was applied to two different optimization constraints. In the following the LSB with the minimization frame, (as depicted in Fig. 2.19) is simply referred to as LSB, while the LSB with the frequency dependent focusing area is explicitly mentioned as $\mathrm{LSB}_{\mathrm{freq.dep}}$. The influence of the SVD regularization is only shown for the MVDRB and the LSB, but all figures and data of the $\mathrm{LSB}_{\mathrm{freq.dep.}}$ have been derived with SVD, too.

2.7.1 Condition Number

The condition numbers of $\left(\tilde{G}^H\tilde{G}\right)$ and $\left(G^H G\right)$ (i.e. for LS- and the MVDR beam, respectively) are very poor for low frequencies. Fig. 2.24 shows that the SVD regularization reduces the condition number to 25 dB. This prevents round-off errors and yields robustness against mismatches in the array geometry and the loudspeaker characteristics.

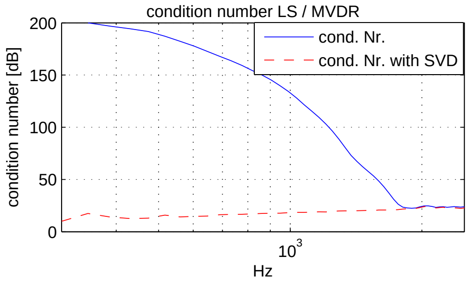
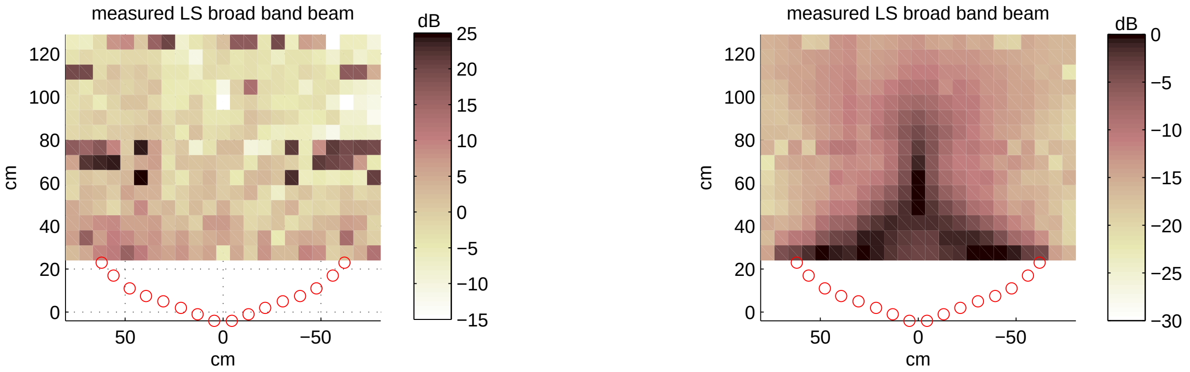


Figure 2.24: The condition numbers of the LSB and the MVDRB are nearly identical. This is why both are represented by one line. They have very low singular values at low frequencies, which are neglected by the SVD regularization.

Fig. 2.25 undermines the gain of robustness. The LS driving functions that have been derived over the (ideal) free field Green's functions are applied to the measured transfer functions. The result is a completely distorted sound field because the measured transfer functions differ from the idealized ones. However, the beam is rendered correctly if the driving functions are regularized. The specific influence of phase and amplitude differences as well as loudspeaker displacements have been investigated by Mabande and Kellermann [2007].



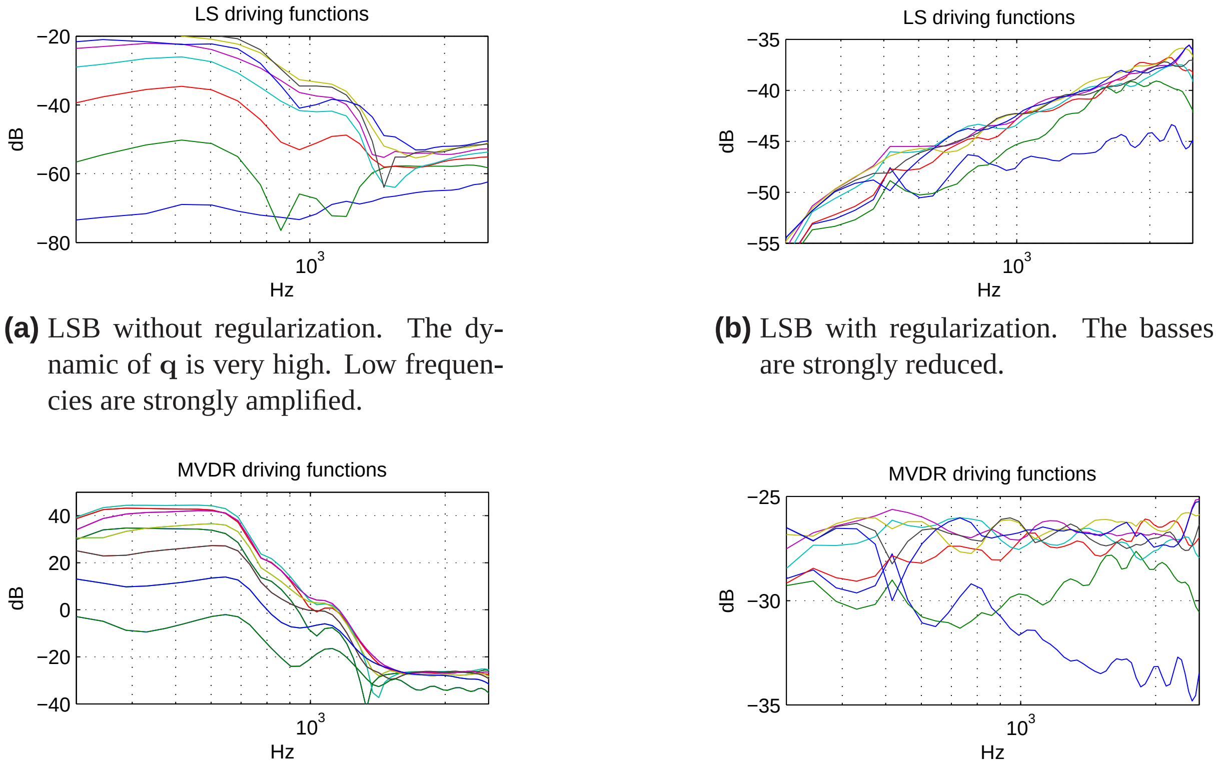
(a) Without regularization the LSB causes an undesired random sound field if the real life conditions are not exactly the same as for the simulation.

(b) The LSB with SVD regularization also works fine for imperfect hardware and mismatches in the array geometry.

Figure 2.25: The LS driving functions were applied to the measured transfer functions. The aberrations of the measurement data from the ideal transfer functions cause a waste sound field, if no regularization is applied. The MVDRB delivers equal results.

2.7.2 Driving Functions

The changes of the spectrum of the driving functions due to the SVD regularization can be seen in Fig. 2.26. The excessive bass boost is suppressed.



(a) LSB without regularization. The dynamic of **q** is very high. Low frequencies are strongly amplified.

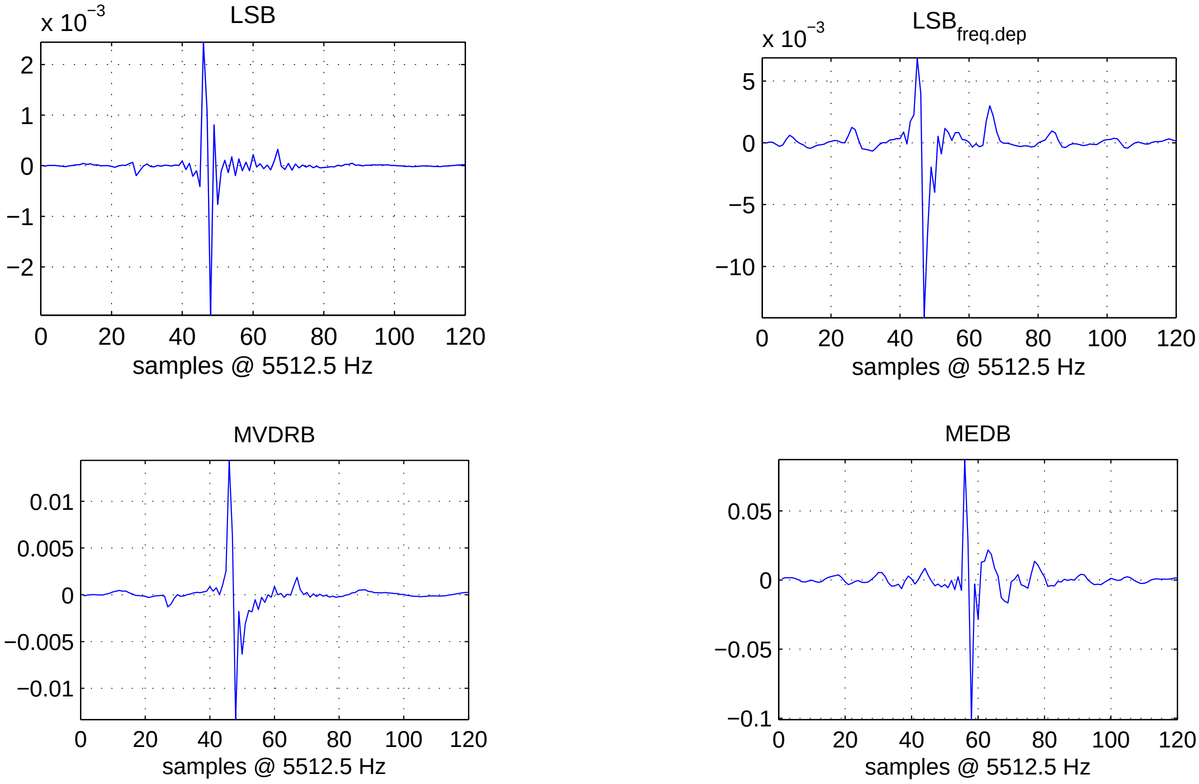
(b) LSB with regularization. The basses are strongly reduced.

(c) MVDRB without regularization. Low frequencies are strongly amplified, even more than in the case of the LSB.

(d) MVDRB with regularization. The SVD regularization causes quite flat spectra for the MVDR driving functions.

Figure 2.26: Frequency response of the LS- and MVDR driving functions.

For the implementation of a beamformer, however, it is more important to know the temporal behavior of the filters. Impulse-response examples of the driving functions for the superdirective beamformers are shown in Fig. 2.27.



(c) The LS- and MVDR driving functions have a fast decay of the impulse response.

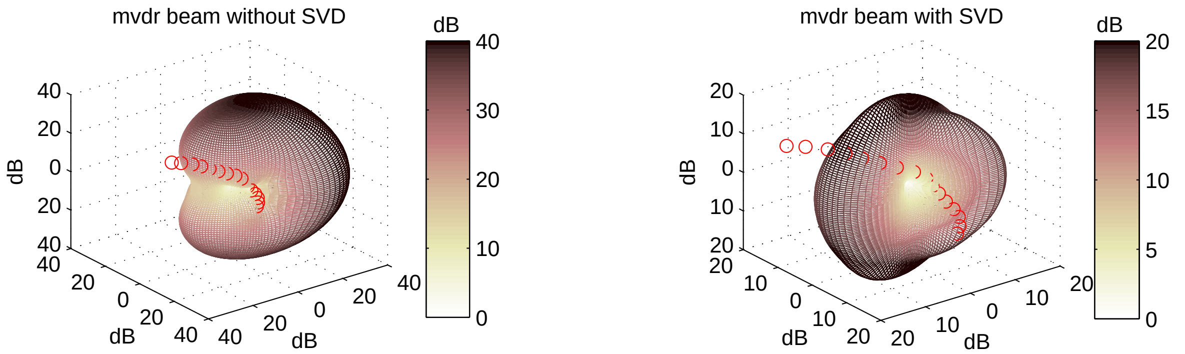
(d) The LS$_{\mathrm{freq.dep}}$- and MED driving functions have strong ripples before and after the impulse.

Figure 2.27: Example of driving functions for the superdirective beamformers in the time domain.

The LS- and the MVDR driving functions decay smoothly and do not have a lot of prepulses, while the LS$_{\mathrm{freq.dep}}$- and MED driving functions have strong ripples before and after the impulse. In order to get a comparable value for the required length of the filters, the number of samples that have got 98 % of the impulse-response energy are considered as filter length. 1 % of the total energy lies before and after the considered samples, respectively. The comparison of the filter lengths follows in Table 2.1.

2.7.3 White Noise Gain

The WNG relates the SPL at the focus point to the energy of the loudspeaker signals and is therefore a measure of the beamformer's efficiency. Firstly, the influence of the SVD regularization is investigated again. The LS- and MVDR beams produce a low sound pressure in the horizontal minimization area, but radiate extensively into every other direction if no regularization is applied. Fig. 2.28 shows the 3D directivity of an MVDR beam and the improvement due to SVD regularization.

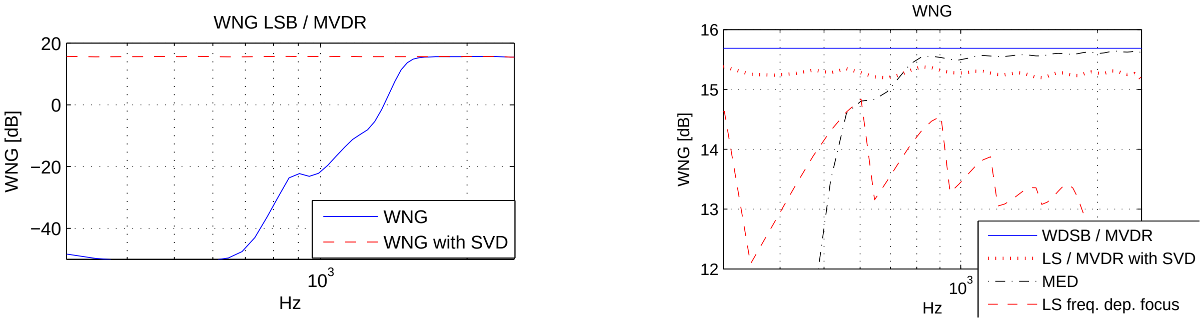


(a) Directivity of an MVDR beam without regularization. The array radiates much more energy into the back side of the array than into the focus point.

(b) Directivity of a MVDR beam with SVD regularization. The array radiates mainly into the focus point. The SPLs in the back of the array are at least 20 dB below the SPLs of the unregularized beam.

Figure 2.28: 3D directivity of a MVDR beam. Without regularization, the acoustic power of the MVDR beam is about 20 dB higher. Still, it is impressing how the sound pressure in the minimization area is suppressed.

Fig. 2.29a shows the improvement of the WNG due to the regularization. At low frequencies it is increased by almost 70 dB. This makes the LS- and the MVDR beam almost as efficient as the Weighted Delay & Sum beam that has the best possible WNG (as in Brandstein and Ward [2001]). The comparison between the WNGs of the different beamformers can be seen in Fig. 2.29b. The MED Beamformer and the LSB$_{\mathrm{freq.dep.}}$ have the worst WNG.



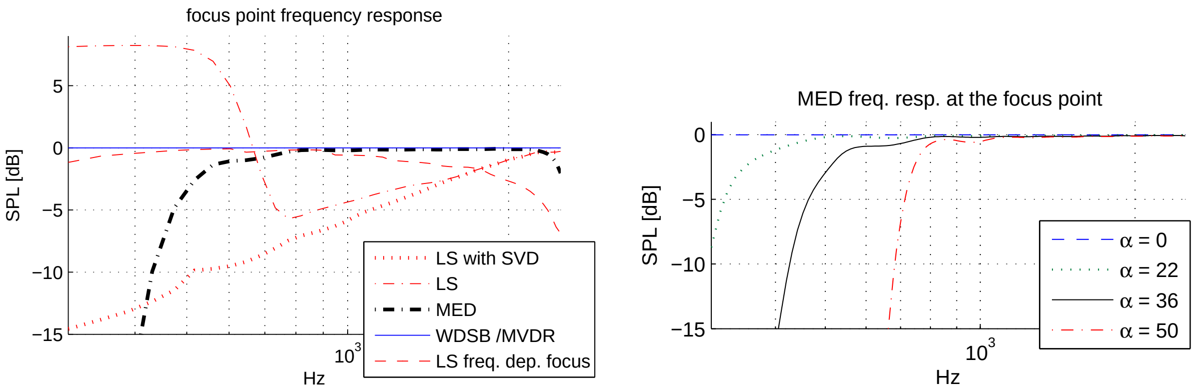
(a) The White Noise Gain of the LS and MVDR beam are almost identical, again. Thus, both are represented by one line. SVD regularization improves the WNG, especially for low frequencies.

(b) The Weighted Delay & Sum beam has an optimal WNG. It requires the least input energy to derive the demanded SPL at the focus point. The regularized LS- and MVDR beam show almost the same performance, while the MED beam decreases to -15 dB for low frequencies.

Figure 2.29: White Noise Gain

2.7.4 Frequency Response at the Focus Point

The WNG and the frequency response of the driving functions determine the frequency response at the focus point. This can be clearly observed for the regularized LSB. The gain of its driving functions (Fig. 2.26b) rises with the frequency, but the WNG is flat. As a consequence, the frequency response at the focus point rises with the frequency, too. The frequency responses of all beamformers are depicted in Fig. 2.30a.



(a) All beamforming methods, except for the LSB and the MEDB cause a flat frequency response at the focus point. It is remarkable that the $\mathrm{LSB}_{\mathrm{freq.dep.}}$ causes a quite flat spectrum, too.

(b) MED frequency response at the focus point. The high pass cut-off frequency increases with the tuning factors α.

Figure 2.30: Frequency responses at the focus point.

It is remarkable that the $\mathrm{LSB}_{\mathrm{freq.dep.}}$ produces a quite flat spectrum at the focus point. In contrast to the LSB, the $\mathrm{LSB}_{\mathrm{freq.dep.}}$ does not attenuate the basses. This is because more focus points are considered for low frequencies (as explained in Fig. 2.20). Therefore, the $\mathrm{LSB}_{\mathrm{freq.dep.}}$ puts more effort into reaching the unity gain at low frequencies.

The frequency response yielded by the MED beam depends on the tuning factor α. In general, it has a high-pass characteristic. Fig. 2.30b shows that higher values of α cause higher cut-off frequencies. A tuning factor of $\alpha = 0$ optimizes the WNG and hence yields the same result as the Weighted Delay & Sum Beamformer.

2.7.5 Overall Comparison

Table 2.1 gives an overview of the comparison. The LSB and the MVDR are only considered with regularization. Apart from the average WNG and the filter length, the Spatial Rejection Ratio as defined in Section 2.3.2 is stated. Except for the $\text{LSB}_{\text{freq.dep.}}$, all optimization methods aimed to produce a low sound pressure in a frame around the user space. $\text{SPL}_{\text{frame}}$ denotes the average sound pressure level in this frame. The $\text{LSB}_{\text{freq.dep.}}$ was introduced to produce a low sound pressure in an interaural distance of the focus point. A point, 20 cm next to the focus point in parallel to the x axis, was chosen as reference for SPL_{20cm}. This reference point is also depicted in Fig. 2.31.

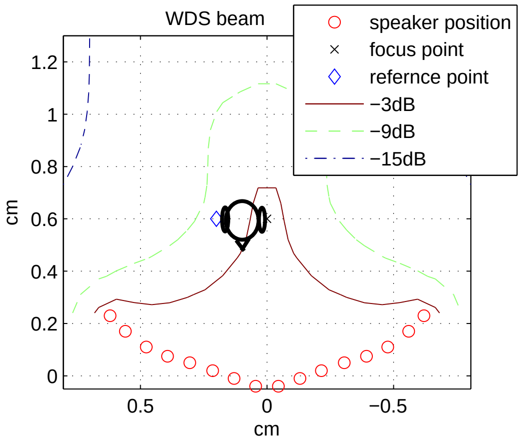


Figure 2.31: Reference point for SPL_{20cm}. If the beam is steered to the ear of a user, this point is meant to mark the position of the contralateral ear.

The SPL attenuation of the beams in three bands are shown in Appendix A.

As last measure, the variation of the frequency response in the focus point is introduced. It is the standard deviation σ of the SPL at the focus point in dB over frequency.

Table 2.1 shows that the LS- and the MED beams have low average SPLs in the minimization areas and a good SRR. However the simple Weighted Delay & Sum beam has comparable results and the advantage that it causes a flat frequency response at the focus point. Besides, it can easily be realized with a delay line and one multiplication only. Its sound pressure in the focus point is 18 dB higher than the SPL of a reverberant room. In Moore [1995], this difference is stated to be above the masking level. Thus, it can be assumed that coworkers in the same room are not disturbed by the loudspeaker signals. The $\mathrm{LSB}_{\mathrm{freq.dep.}}$ and the MVDRB do not really perform better than the Weighted Delay & Sum Beamformer, but are more complex in their realization. Finally, it can be concluded that the WDSB is the most feasible beamformer for a dynamic application. Its implementation is very processing efficient and its results are very satisfying. Consequently the WDSB is used for further investigations on transaural stereo with focused sound. These investigations follow in the next chapter.

Table 2.1: Comparison of the beamforming methods. The LSB and the MED have very good results for the SPL attenuation and the SRR, but they have a strong variation of the frequency response at the focus point. The MVDRB and the $\mathrm{LSB}_{\mathrm{freq.dep.}}$ do not really perform better than the WDSB. Considering its efficient implementation, the WDSB is the most feasible beamforming method for the given conditions.

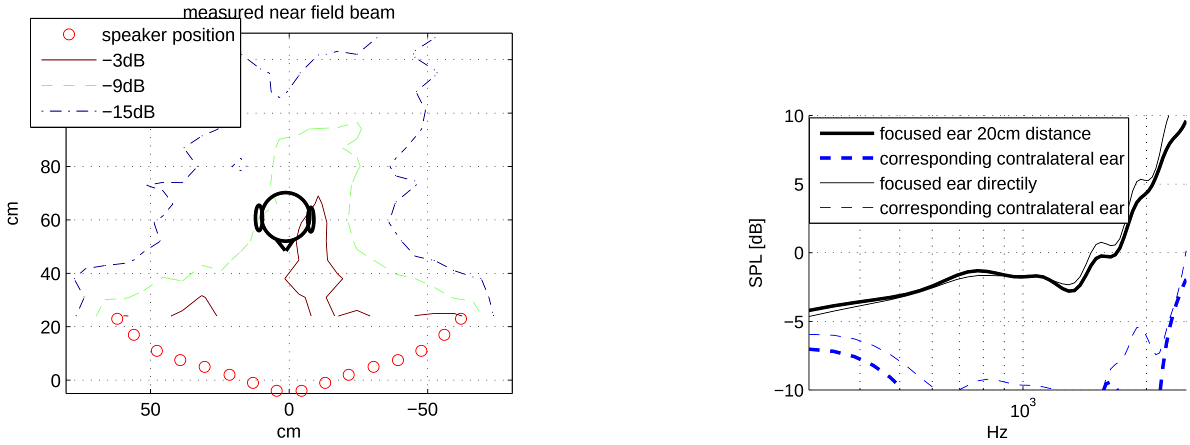
	WDSB	MVDR	$\mathrm{LSB}_{\mathrm{freq.dep.}}$	MED	LSB
filter length (taps @ 5512.5 Hz)	delay line + multiplication	39	69	99	20
average WNG	15.7 dB	15.5 dB	13.4 dB	10 dB	15.5 dB
SRR	17.5 dB	17.5 dB	15.7 dB	18 dB	19 dB
$\mathrm{SPL}_{\mathrm{frame}}$	-17.5 dB	-18 dB	-13 dB	-19 dB	-20 dB
$\mathrm{SPL}_{\mathrm{20cm}}$	-11 dB	-11 dB	-13 dB	-14 dB	-15 dB
σ	0 dB	0 dB	1.5 dB	6 dB	4 dB

Chapter 3

Transaural Beamforming

3.1 Concept of Transaural Stereo for a Loudspeaker Array

The beamformers (described in the previous chapter) minimize the total energy that excites the sound field. They also cause a natural channel separation (between the left ear signal and the right ear signal), if the beams are steered to the ears of the user. As stated in the introduction, this channel separation is important for a transaural stereo application. Fig. 3.1a shows the SPL contours of a measured WDSB. The head symbol marks the listener position. The broad-band sound pressure at the position of the contralateral ear is already attenuated by 9 dB. The channel separation over frequency was measured with a dummy head. Fig. 3.1b shows the channel separation of a beam to the eardrum and a beam with a focus 20 cm outside of the ear. Low frequencies have a larger beam width and, therefore, cause a higher cross talk. The beam that is focused 20 cm outside of the ear causes less cross talk at low frequencies, with hardly any loss of channel separation at high frequencies.

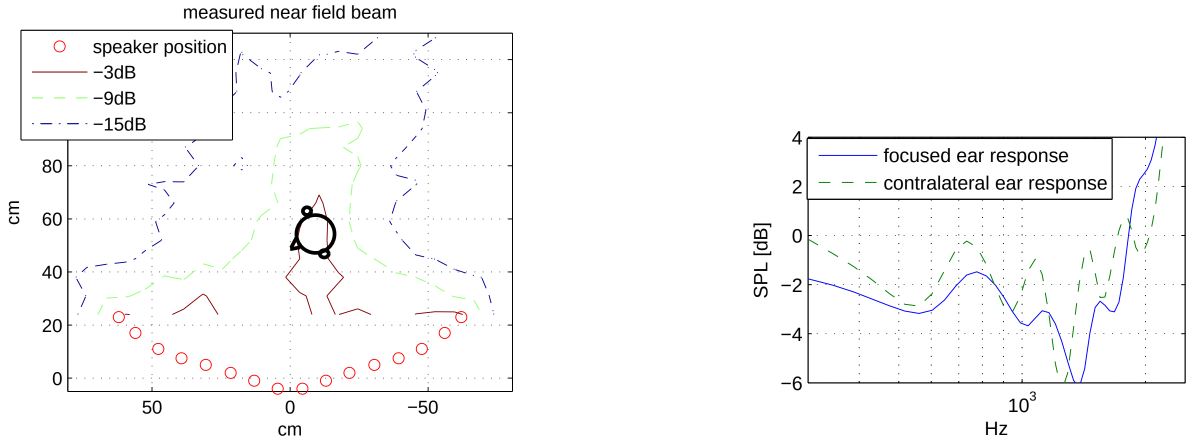


(a) The broad band sound pressure decay at the position of the contralateral ear is -9 dB.

(b) SPL of the cross-talk over frequency. The beam width decreases with the frequency. Therefore the cross-talk decreases, too. The channel separation of low frequencies is 1 dB better, if the beam is focused 20 cm away from the ear.

Figure 3.1: Channel separation due to a WDS beam that is steered to one ear.

If the ear is turned away from the array, the results are deteriorated. This particular constellation is depicted in Fig. 3.2 and the corresponding cross talk over frequency in Fig. 3.2b.



(a) No channel separation due to beamforming can be assumed if the contralateral ear is facing the loudspeaker array.

(b) The cross talk is even higher than the beam at the rear focused ear itself.

Figure 3.2: There is no channel separation, if the focused ear is turned away from the loudspeaker array.

For either case a cross talk canceler (XTC) has to be applied, for two reasons:

1. To reduce the cross talk (above all, for the low frequencies).
2. To equalize the frequency response at the focused ear.

The second point is not really included in the term cross-talk canceler, but it is equally important for a transaural stereo application. The binaural signals have to reach the ears without alteration. Therefore, the XTC has to cause a flat frequency response at the entrance to the ear channel.

The relation between the binaural input signals L and R and the signals at the eardrum E_l and E_r is given in Eq. (3.1). All variables express quantities in the frequency domain. The dependency on ω, however, is omitted for the sake of compactness. L and R are split into 2×16 loudspeaker signals in the beamforming stage (q_{ji}). These 2×16 loudspeakers signals reach the ears over 16×2 Head Related Transfer Functions H_{ij}

$$\begin{pmatrix} E_l \\ E_r \end{pmatrix} = \begin{pmatrix} H_{1,l} & H_{2,l} & \cdots & H_{16,l} \\ H_{1,r} & H_{2,r} & \cdots & H_{16,r} \end{pmatrix} \begin{pmatrix} q_{l,1} & q_{r,1} \\ q_{l,2} & q_{r,2} \\ \vdots & \vdots \\ q_{l,16} & q_{r,16} \end{pmatrix} \begin{pmatrix} L \\ R \end{pmatrix}. \tag{3.1}$$

These transfer paths are also schematized in Fig. 3.3.

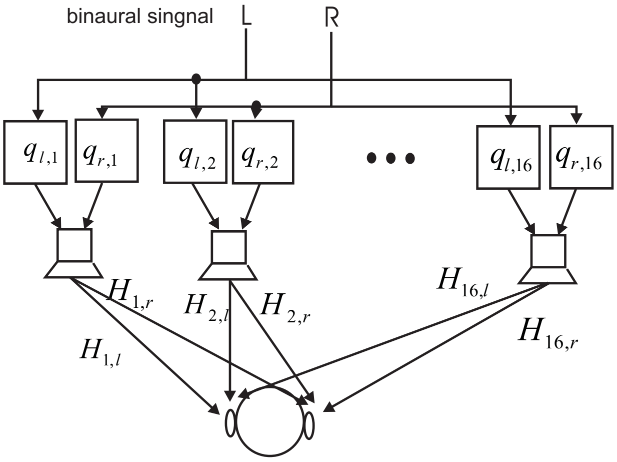


Figure 3.3: The binaural signals are delayed and weighted by the complex factors $q_{i,j}$ and reach the ears via the HRTFs $H_{j,i}$. The overall transfer functions are derived by superimposing these weighted HRTFs.

With the definition of a composite transfer matrix

$$\mathbf{T} = \begin{pmatrix} T_{ll} & T_{rl} \\ T_{lr} & T_{rr} \end{pmatrix} = \begin{pmatrix} H_{1,l} & H_{2,l} & \cdots & H_{16,l} \\ H_{1,r} & H_{2,r} & \cdots & H_{16,r} \end{pmatrix} \begin{pmatrix} q_{l,1} & q_{r,1} \\ q_{l,2} & q_{r,2} \\ \vdots & \vdots \\ q_{l,16} & q_{r,16} \end{pmatrix}, \tag{3.2}$$

equation (3.1) simplifies to

$$\begin{pmatrix} E_l \\ E_r \end{pmatrix} = \mathbf{T} \begin{pmatrix} L \\ R \end{pmatrix}. \tag{3.3}$$

The XTC matrix $\mathbf{C}$ is applied to the input signals L and R

$$\begin{pmatrix} \hat{E}_l \\ \hat{E}_r \end{pmatrix} = \mathbf{TC} \begin{pmatrix} L \\ R \end{pmatrix}, \tag{3.4}$$

in order to achieve the desired ideal transmission

$$\mathbf{TC} = \mathbf{I}, \tag{3.5}$$

where $\mathbf{I}$ is the identity matrix. Thus,

$$\mathbf{C} = \mathbf{T}^{-1}. \tag{3.6}$$

The calculation of the XTC matrix bears 2 problems:

1. The transfer matrix $\mathbf{T}$ has to be identified.
2. This transfer function matrix has to be inverted.

ad 1.

The transfer matrix is derived over the multiplication of the beamforming weights with the HRTFs from all loudspeakers to the left and right ear, respectively. The beamforming weights can be calculated with knowledge about the position of the user and the loudspeaker array, as it is deduced in Chapter 2. The HRTFs are derived from a data base. For the following simulations and measurements, a data base of 36 HRTFs in the horizontal plane has been used. The HRTFs have been measured with a dummy head in 10° steps. Positions between the measured grid are linearly interpolated in phase and amplitude. To take different distances into account, suitable delays and gains are applied on the interpolated HRTFs.

ad 2.

Matrix inversion is problematic if the determinant $\det(\mathbf{T})$ is close to zero. This can be prevented if a bias β is added to the determinant. This regularization may be frequency dependent like it is shown by Kirkeby and Nelson [1999]. They suggested

$$\det(\mathbf{T})^{-1} = \frac{\det(\mathbf{T})^{H}}{\|\det(\mathbf{T})\|^{2} + \beta \, \|H\|^{2}}, \tag{3.7}$$

wherein H is a filter that determines the frequencies on which β works. The absolute values of $\det(\mathbf{T})$ have been investigated for 384 head positions with 0° and 30° head rotation each. The head positions correspond to the measurement points of Fig. 2.7a. The lowest values of the determinant always appeared at 300 Hz where the cross talk path is equally strong as the direct path. However, the determinant does not fall under an absolute value of 0.09, which of course is no numerical problem for inversion. Still, the determinant can lead to dynamical problems if the ratio between the highest and the lowest value within the used bandwidth becomes to large. The beamformer, in general, makes sure that the direct paths (T_{ll} and T_{rr}) have higher amplitudes than the off-diagonal elements. In the worst cases (as depicted in Fig. 3.2), the cross-talk response can be equally loud as the direct paths. Therefore, the dynamic range of T_{ll} and T_{rr} needs to be investigated only. Fig. 3.4 shows the level lines within which a certain dynamic level is not exceeded.

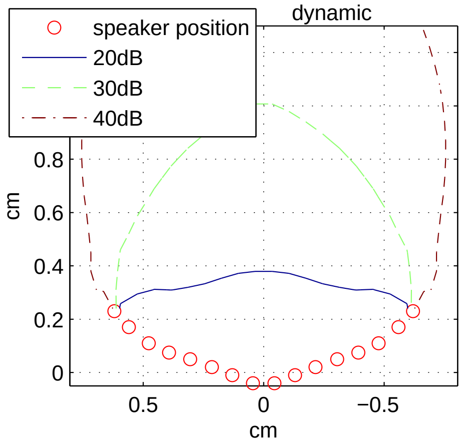


Figure 3.4: The dynamic range does not exceed 30 dB in the major part of the working area. This is a value that ordinary sound equipment should cope with.

The frequency response is bounded within a range of 40 dB across the whole working area. If the sound equipment cannot cope with this dynamic range, the regularization factor β needs to take values greater than zero. The smallest value of $\det(\mathbf{T})$ was $\sim \frac{1}{10}$. Choosing, e.g., a β that is a factor of 10 higher ($\beta = 1$), reduces the dynamic of the basses by 20 dB; of course with the cost of a high pass characteristic in the ear signal.

3.2 Cross-Talk Cancellation Results

The influence of the XTC filter on the sound field is shown in Fig. 3.5. Besides the expected beam, a strong SPL attenuation at an interaural distance can be observed.

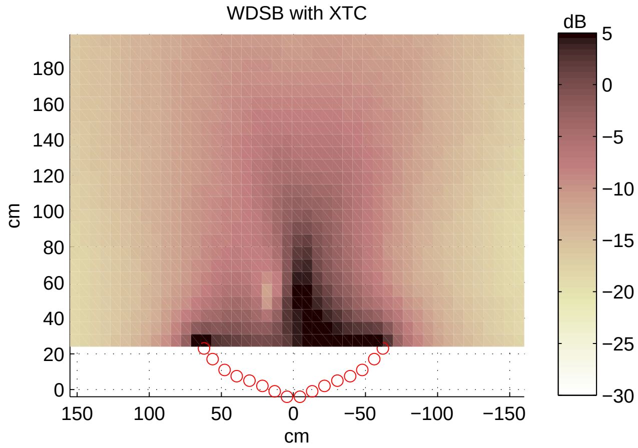
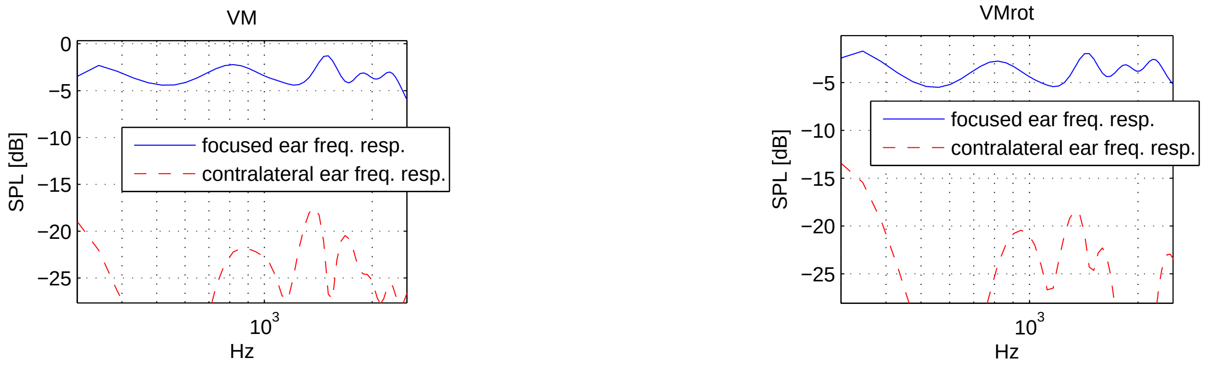


Figure 3.5: The sound field of the Transaural Beamformer shows a strong SPL attenuation at an interaural distance from the focused point.

Concerning the transfer functions, three factors are important to render a correct transaural signal:

1. A flat frequency response at the focused ear.
2. A high channel separation.
3. A uniform group delay of the transmitted energy. The transfer function to the focused ear should not only have a flat frequency response, but also a pulse like temporal behavior.

Fig. 3.6 shows the improvement on the situations depicted in Fig. 3.1 and 3.2 due to the cross-talk canceler. The frequency response has a ripple of less than 5 dB and the channel separation is larger than 10 dB for all frequencies.

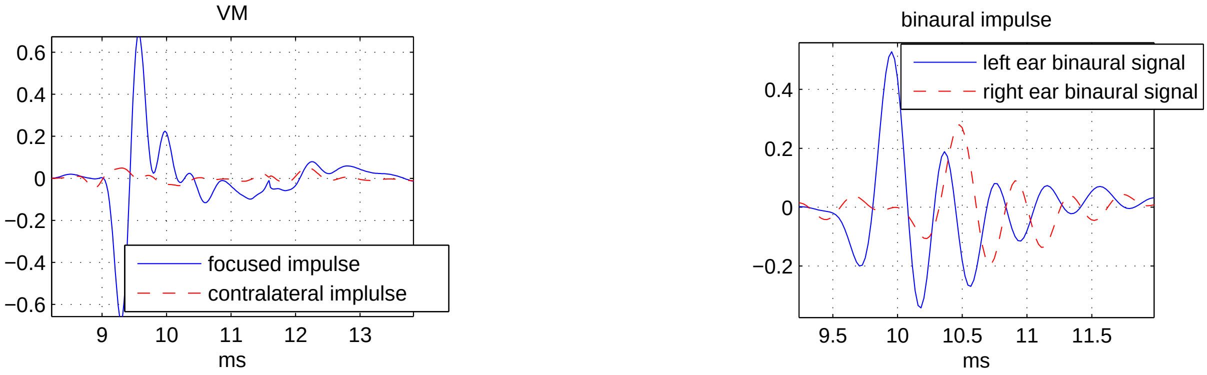


(a) Central position, the nose points to the array.

(b) Central position, 30° head rotation.

Figure 3.6: Frequency responses after XTC. Two effects can be observed: Firstly, the channel separation is much larger, also for the low frequencies and secondly, the frequency response at the focused ear is equalized.

For both cases, the temporal behavior (shown in Fig. 3.7a) is pulse like and mainly deteriorated by the band pass filter from 300 to 2500 Hz.



(a) For the focused ear, the temporal response of the XTC is a pulse. Its dispersion mainly comes due to the applied band-pass filter. At the contralateral ear the impulse response is smeared out and compressed.

(b) A binaural impulse from the left (90°) was fed to the transaural beamformer for a central head position. There is a small pre-echo in the right ear binaural signal at 9.5 ms.

Figure 3.7: Temporal behavior of the XTC.

The frequency responses of the transaural beamformer were measured for four head positions with 0° and 30° rotation each. The positions are marked in Fig. 3.8. The variations in the frequency responses are smaller than 6 dB and the channel separation is at least 10 dB for all positions. The amplitudes of all frequency responses are shown in Appendix B. The temporal behaviors equal the one presented in Fig. 3.7. These results are satisfactory, however, they vary with the head position and rotation.

A quality measure is introduced to assess the dependency on these positions and rotations. As the temporal behavior looks equally good for all measurements, the quality measure Q_1 only takes the channel separation and the ripple in the frequency response into account. The channel separation $\mathrm{SPL}_{\mathrm{dif}}$ is simply given by the average amplitude difference in dB. The ripple σ^2_{SPL} is defined as the variance of the amplitude over frequency. The average channel separation is 14 dB in the worst case and 22 dB in the best case. The variance varies between 1 and 3 dB^2. The two quantities are scaled to match their range and are added to

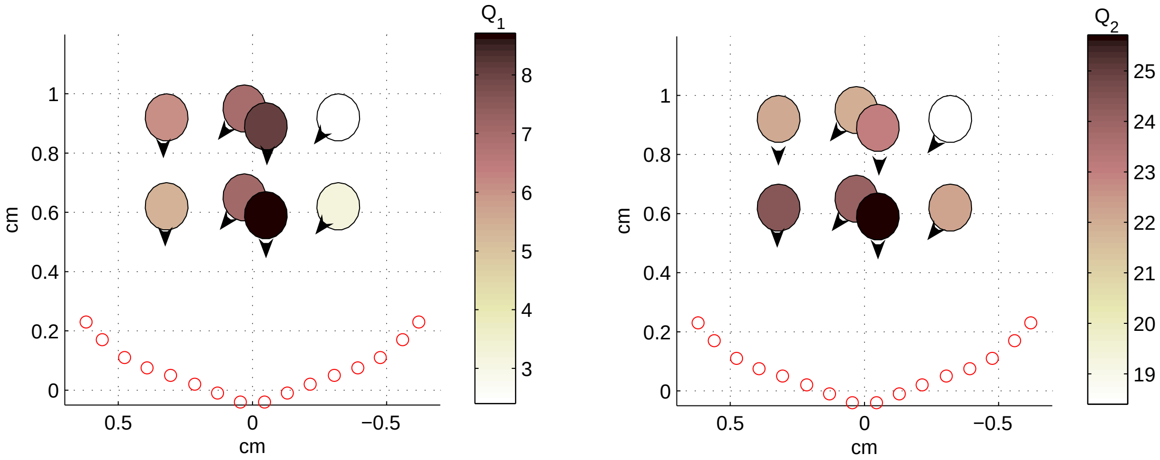
$$Q_1 = \frac{1}{2}\mathrm{SPL}_{\mathrm{dif}} - 2\,\sigma^2_{\mathrm{SPL}}. \tag{3.8}$$

Of course, σ^2_{SPL} is desired to be small; as a consequence it is subtracted from $\mathrm{SPL}_{\mathrm{dif}}$.

Until now, only the correct transaural rendering has been considered for the quality factor Q_1. The excitation of the room, however, should also be taken into account, as stated in the introduction. The Spatial Rejection Ratio of the Transaural Beamformer varies between 11 and 15 dB for the given head positions and has therefore the same rage as the two already considered (scaled) properties. The quality measure Q_2 includes the room excitation and is defined as

$$Q_2 = \frac{1}{2}\mathrm{SPL}_{\mathrm{dif}} - 2\,\sigma^2_{\mathrm{SPL}} + \mathrm{SRR}. \tag{3.9}$$

The results of Q_1 and Q_2 are shown in Fig. 3.8a. It can be concluded that the quality of the transaural beamformer decreases with the distance from the array, with the degree of head rotation and with the distance from the symmetry axis.



(a) XTC quality. Central positions and 0° head rotations have the best preconditions for a correct transaural rendering.

(b) Transaural beamforming quality. The room excitation is included in the measure. Beams to the close side positions cause little room excitation (see also Fig. 2.16b), that is why the close side head positions also perform better in terms of the transaural beamforming quality Q_2.

Figure 3.8: Evaluated head positions and rotations. Please note that the head positions on the symmetry axis are coincident for both rotations. They are only displaced for the representation. The quality of the transaural beamformer decreases with the distance from the array, with the degree of head rotation and with the distance from the symmetry axis.

Chapter 4

Resume and Outlook

4.1 Resume

Binaural signals evoke a spatial sound-impression if they are transmitted to the ears directly. They are used in various virtual- or augmented reality applications, in which they are mostly played back via headphones. Transaural stereo is a method that preprocesses binaural signals in order to play them back over loudspeakers. The preprocessing is necessary because the transmission paths from the loudspeakers to the ears would cause a coloration of the binaural signals. Especially the cross talk from the left binaural signal to the right ear and vice versa leads to a grave alteration of the binaural signal which impairs the spatial impression.

In this thesis, a transaural stereo application with focused sound is introduced. The focusing bears two advantages. Firstly, it prevents a strong sound excitation of the room, and secondly, it achieves a better channel separation if the beams are steered to the ears. A focused sound can be accomplished with an array of loudspeakers. The physical straightforward method of sound focusing is a Weighted Delay & Sum Beamformer (WDSB). It delays the loudspeaker signals such that they coincide at a focus point where they are added constructively. The geometrical properties (i.e. the loudspeaker positions and their distance to the focus point) suffice to calculate the delay times and the weights of the WDSB. The Weighted Delay & Sum Beamformer can easily be implemented with a delay line and one multiplication per loudspeaker channel.

Superdirective beamformer like the Least Squares- (LS), the Maximum Energy Difference- (MED) and the Minimum Variance Distortionless Response (MVDR) Beamformer have additional filters and are traditionally used to produce a smaller beam width (i.e an improved directivity). They use optimization algorithms to render the sound field for a given constraint. This constraint, however, does not necessarily have to be a small beam width. A low sound pressure in any point of a sound field can be forced. The LS and the MVDR algorithm require matrix inversion. It turned out that the matrices that are to be inverted are ill conditioned for most constraints. Regularization, using singular value decomposition (SVD), facilitate applicable LS and MVDR solutions. The MED algorithm leads to a eigenvalue problem which can be solved without matrix inversion. However, it demands filters that are five times as long as a LS- or MVDR filter to yield comparable results.

All mentioned beamforming methods were investigated in terms of their spatial pressure decay, their excitation of the room, their complexity and their frequency response at the focus point. Especially, they are compared with respect to their contribution to an effective transaural rendering. The LS- and the MED Beamformer show very good results of directivity and sound pressure attenuation, but they do not produce a flat frequency response at the focus point. In contrast, the MVDR Beamformer has the constraint of producing unity gain at the focus point. Also the LS Beamformer can be tuned to produce an equalized spectrum at the focus point. However, they do not perform much better than the simple Weighted Delay & Sum Beamformer.

Finally, it can be concluded that the WDSB is the most feasible focusing method for the given constraints. Mainly, it benefits from its low processing load and its optimal input / output relation. Subsequently, the WDSB has been applied to a loudspeaker array with 16 elements for which the cross-talk cancellation matrix is deduced. Two kinds of measurements proof the concept and the functionality of the system. Firstly, the sound field has been measured with a microphone array in order to compare it with the simulated sound fields, and secondly, the cross-talk canceler was evaluated with dummy-head measurements. A channel separation of at least 10 dB could be gained (even for head rotations which are difficult to handle) and the variations in the frequency response stay below 5 dB.

4.2 Outlook

A prototype of the proposed Transaural Beamformer shall be evaluated at the Eurocontrol Experimental Center (EEC) in Brétigny-sur-Orge, France. The prospective evaluation concerns technical as well as psychological matters. The technical questions are: Does the Transaural Beamformer disturb neighboring air traffic controllers; and, in how far does it increase the noise level in the air traffic control center? The psychological aspects concern the concentration of air traffic controllers and its consequence on the error prevention. Commonly, air traffic controllers use stereo headphones whereas one channel is designated for the pilots of the air crafts and the other one for controllers of neighboring air-space sectors. In contrast, the Transaural Beamformer produces spatialized sound, which makes the audio impression much more natural than this conventional stereo separation. In the evaluation, it should be investigated if the increase of comfort also leads to an increase of concentration and if the spatial separation of the communication partners can reduce communication mistakes.

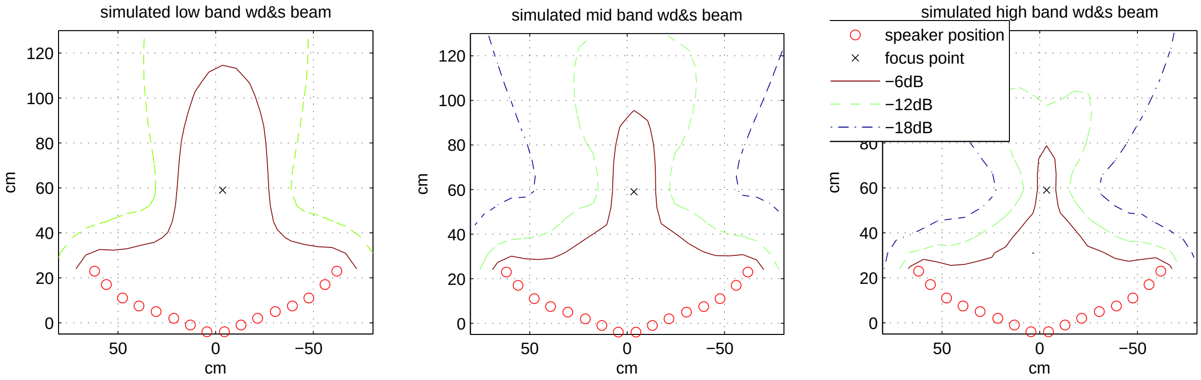
Listening test at the Institute of Electronic Music and Acoustics will evaluate the perception of the transaural signals. In particular, it will be investigated if the virtual sound sources are located correctly and if their spatial distribution helps to distinguish between one from another.

Air traffic control communication suffers from a low bandwidth of 300 to 2500 Hz and additive transmission noise. The intelligibility of speech can be improved by artificial bandwidth extension as proposed by Schäfer [2008] as well as by multi-band compression (Domínguez [2009]). Both methods can easily be integrated into the system of the Transaural Beamformer. The Transaural Beamformer itself benefits from the low bandwidth as it reduces the risk of spatial aliasing. It also has an positive influence on the cross-talk canceler. The cross-talk canceler is based on Head Related Transfers Functions (HRTFs). In general, these HRTFs are unique for every human as they depend on the geometry of ear, head and torso. The individuality of this geometries, however, has above all influence on higher frequencies. On the one hand, it can therefore be assumed, that the HRTFs, measured with the dummy head, will be good enough to provide accurate spatialized audio. On the other hand it raises the questions, if the HRTFs could not be modeled analytically. This would reduce the memory allocation of the system, as the HRTF data base could be omitted. Finally, further beamforming methods can be investigated, like for example the MVDR Beamformer with White Noise Gain constraint that was proposed by Mabande et al. [2009].

Appendix A

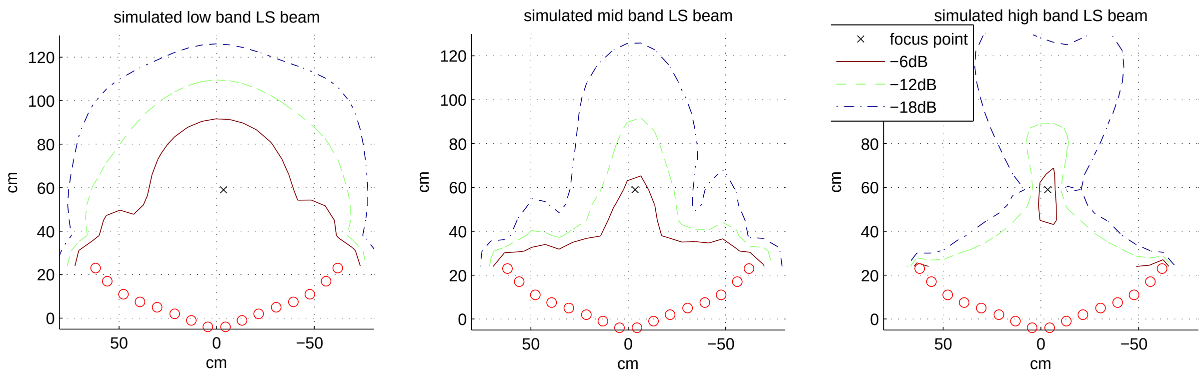
Comparison of the SPL Distribution of the Different Beamformers

The pressure attenuation of the beamformers are shown in 3 bands in Fig. A.1 to A.7. The cut-off frequencies are logarithmically spaced to consider the sensitivity of the ear. It can be seen that the LSB and the MVDRB loose their steep decay towards the back end at low frequencies if SVD regularization is applied. As it is shown in Fig.2.24, the SVD has the strongest impact on these low frequencies. It can be concluded that it is easier for the optimization methods to cause a narrow beam than a beam that has to decay towards the back end.



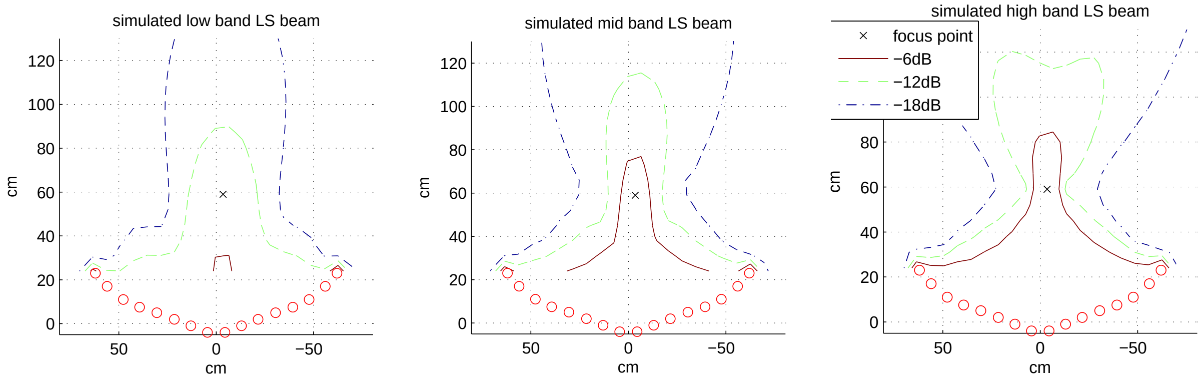
(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.1: Simulated WDSB in 3 bands. The beam width decreases with the frequency.



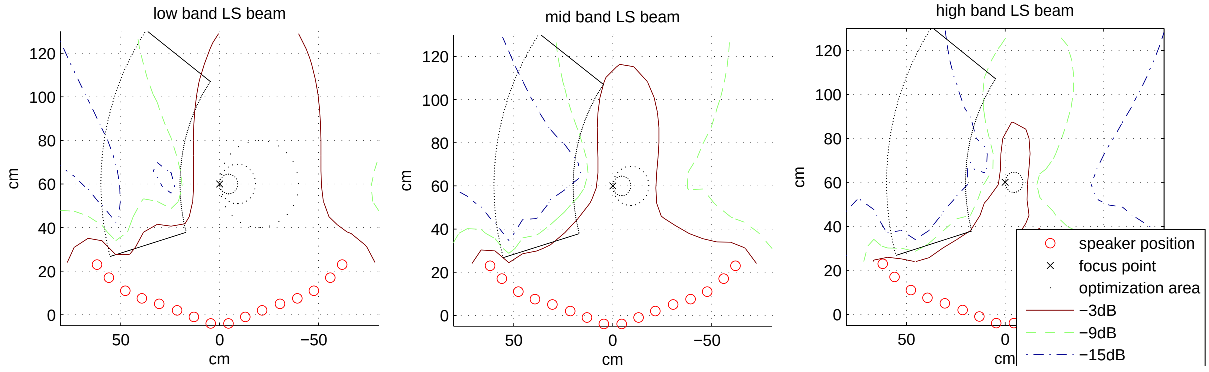
(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.2: Least squares beam. The beam has a steep decay towards the back.



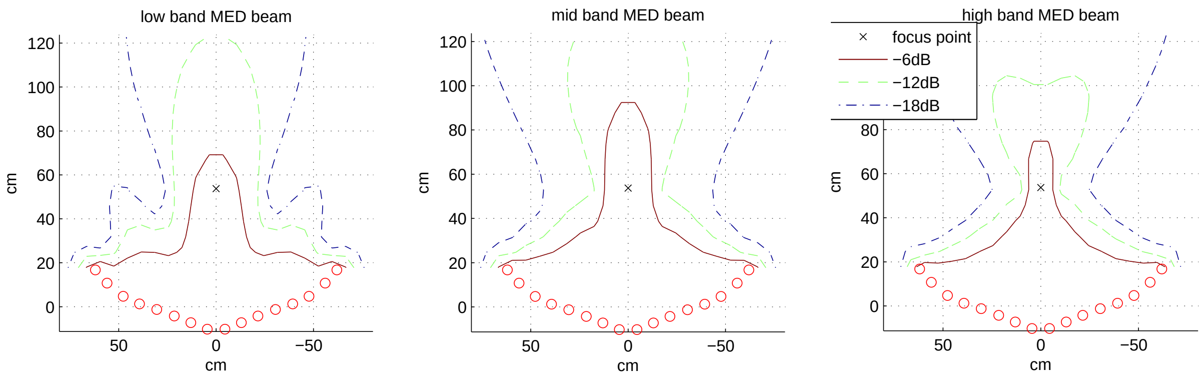
(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.3: LSB with SVD regularization. There is a shift of weights to the high frequencies. The beam is still narrow, but does not decay as fast towards the rear end as the LSB without regularization.



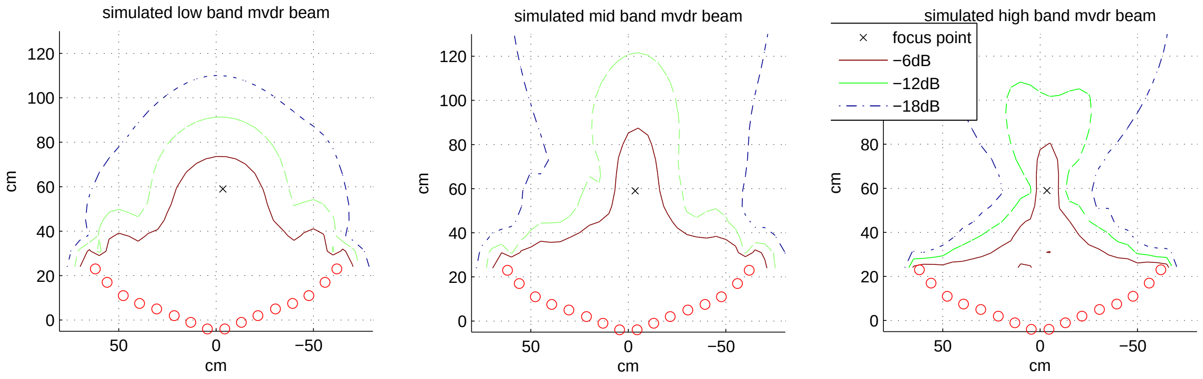
(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.4: LSB with frequency dependent focus area. The size of the focus area decreases with the frequency. Low frequencies are therefore stronger weighted.



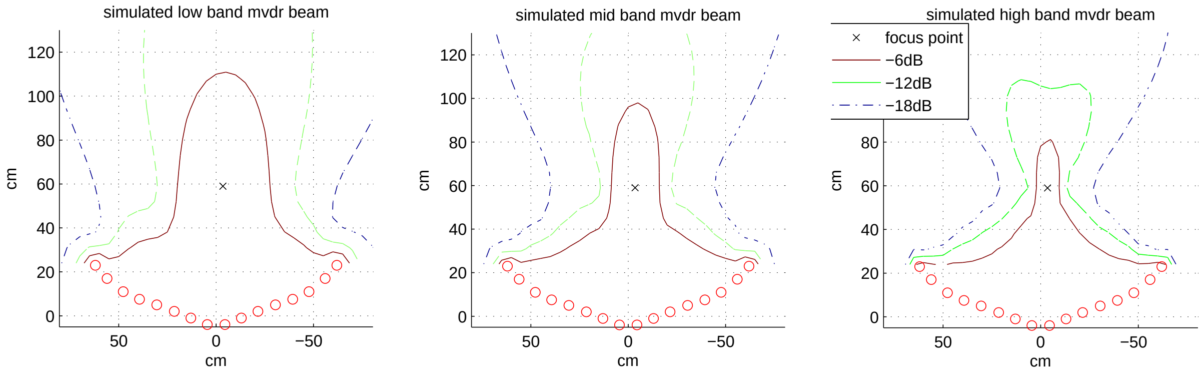
(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.5: MED beam. The tuning factor α was set to get a trade off between the WDSB and the LSB. There is a suppression of low frequencies, but not as strong as for the LSB with SVD regularization.



(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.6: MVDR beam. Like the LSB, the MVDRB has a steep decay towards the rear end at low frequencies. In the upper frequency bands, it does not perform as well as the LSB, because it has the constraint of unity gain at the focus point.



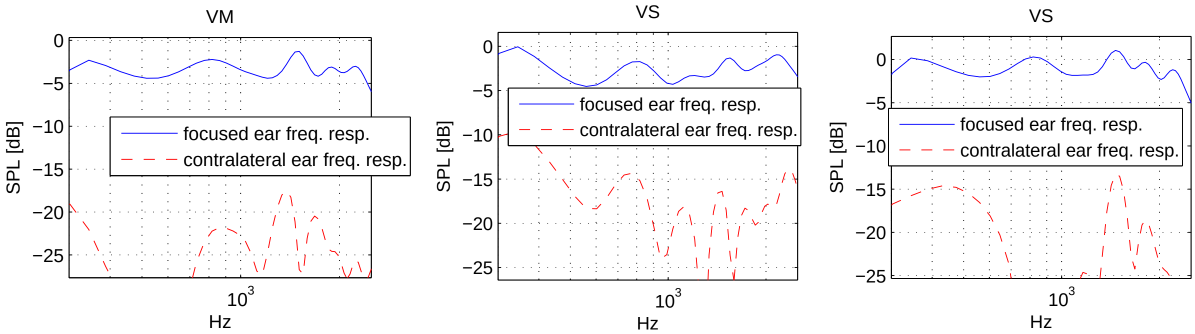
(a) 300-610 Hz **(b)** 610-1230 Hz **(c)** 1230-2500 Hz

Figure A.7: MVDR with SVD regularization. The beam looses its steep decay towards the end at low frequencies, but it is still narrower than the WDSB.

Appendix B

Frequency Responses of the Cross-Talk Canceler

The cross talk of the Transaural Beamformer was measured in four positions with $0°$ and $30°$ head rotation each. These positions are marked in Fig.3.8. The frequency responses of these measurements are presented in the following.

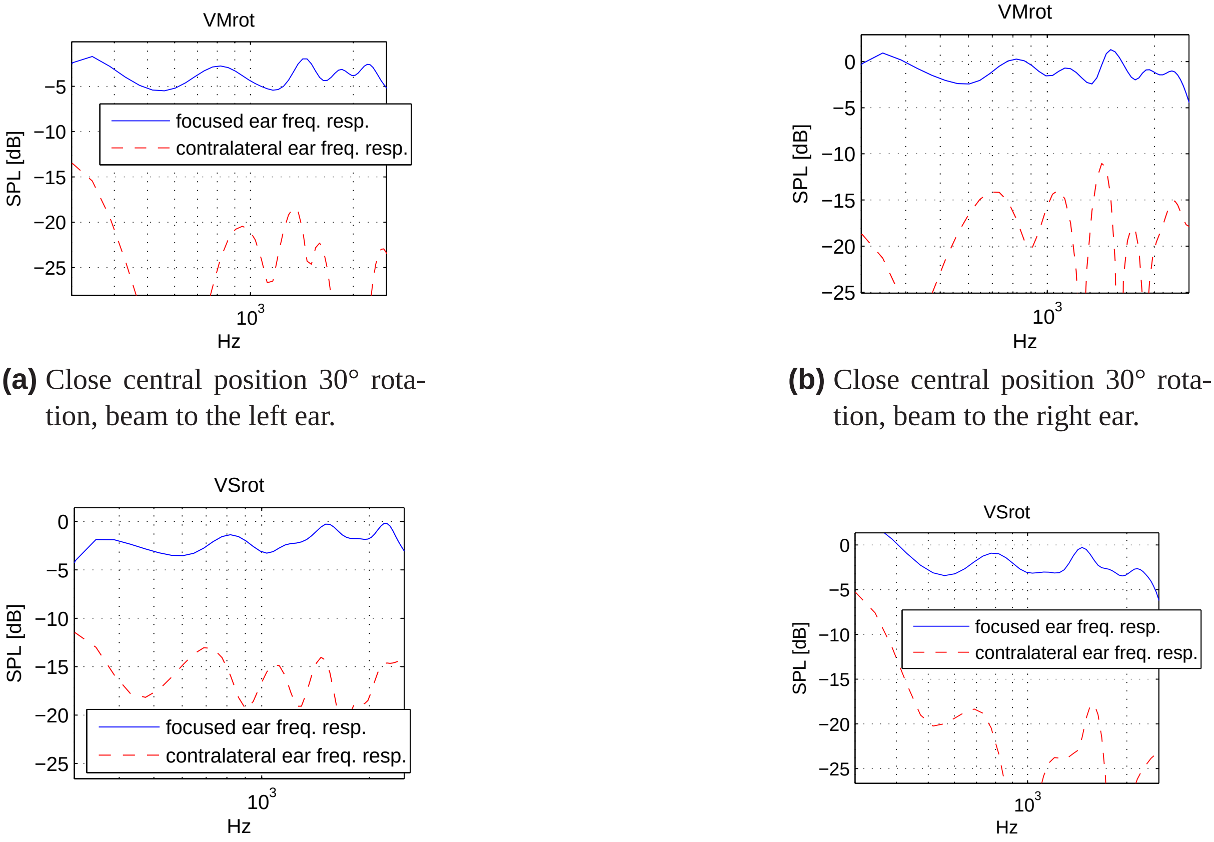


(a) Close central position.

(b) Close side position, beam to the left ear.

(c) Close side position, beam to the right ear.

Figure B.1: XTC frequency responses for the close head positions and 0° head rotation.



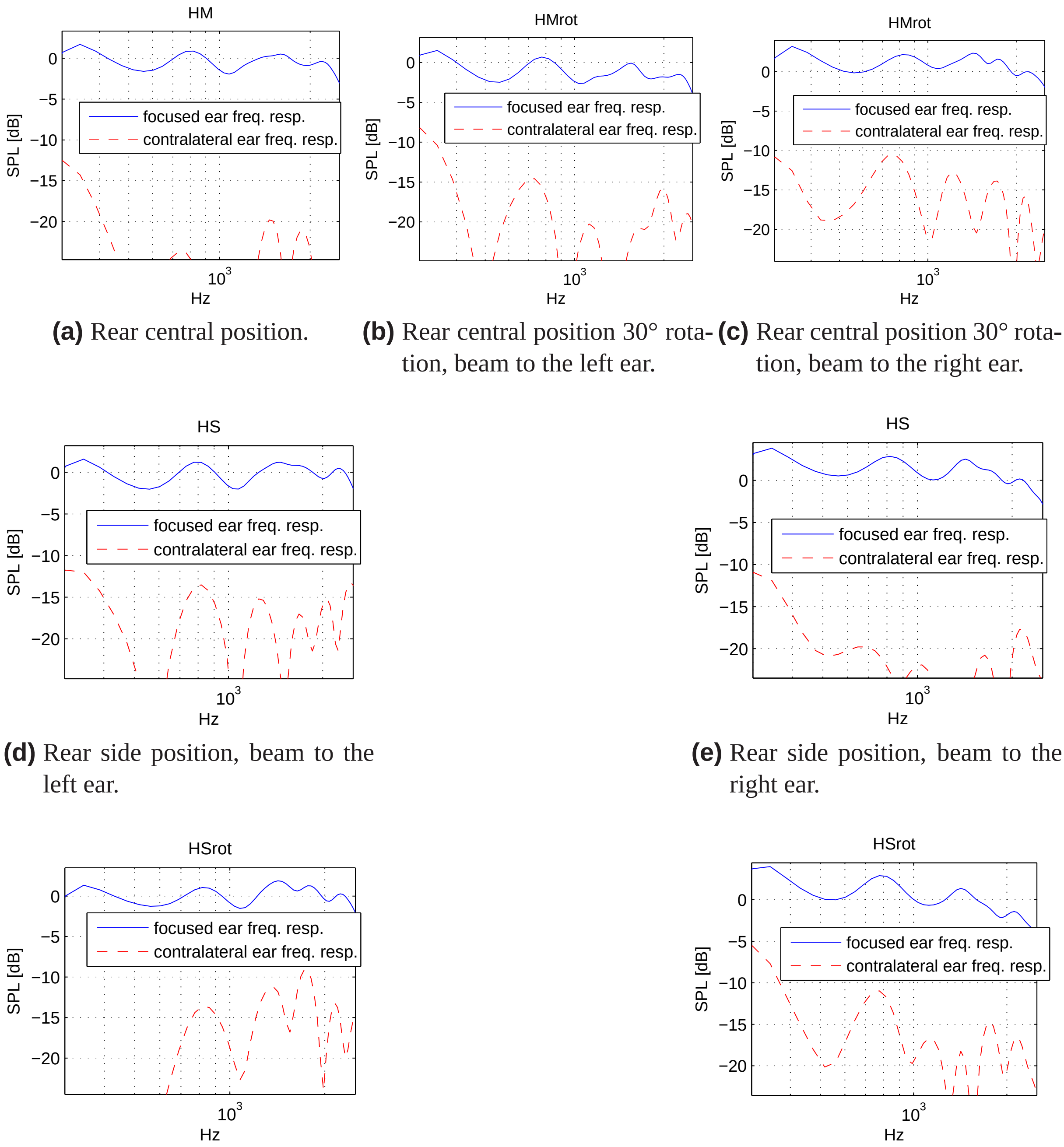
(a) Close central position 30° rotation, beam to the left ear.

(b) Close central position 30° rotation, beam to the right ear.

(c) Close side position 30° rotation, beam to the left ear.

(d) Close side position 30° rotation, beam to the right ear.

Figure B.2: XTC frequency responses for the close head positions with 30° head rotation.



(a) Rear central position.

(b) Rear central position 30° rotation, beam to the left ear.

(c) Rear central position 30° rotation, beam to the right ear.

(d) Rear side position, beam to the left ear.

(e) Rear side position, beam to the right ear.

(f) Rear side position 30° rotation, beam to the left ear.

(g) Rear side position 30° rotation, beam to the right ear.

Figure B.3: XTC frequency responses for the rear head positions.

Bibliography

Ahnert, W. (1993). *Beschallungstechnik; Grundlagen und Praxis*. Stuttgart [u.a.].

Algazi, V. R., Avendano, C., and Duda, R. O. (2001). Estimation of a spherical-head model from anthropometry. *Journal of the Audio Engineering Society.*, 49:472–479. ID: 208781859.

Atal, B. S. and Schroeder, M. R. (1963). Computer simulation of sound transmission in rooms. In *IEEE Conv. Record*, pages 150–155.

Bai, M. R., Tung, C.-W., and Lee, C.-C. (2005). Optimal design of loudspeaker arrays for robust cross-talk canellation using the taguchi method and the genetic algorithm. *J. Audio Eng. Soc.*, 117(5):2802–2813.

Bauck, J. and Cooper, D. H. (1996). Generalized transaural stereo and applications. *J. Audio Eng. Soc.*, 44(9):683–705.

Blauert, J. (1999; 1999). *Spatial hearing; The psychophysics of human sound localization*. MIT Pr., Cambridge, Mass. [u.a.].

Brandstein, M. and Ward, D. B., editors (2001). *Microphone arrays : signal processing techniques and applications*. Springer, Berlin ; London.

Cho, Y. T. and Roan, M. J. (2009). Adaptive near-field beamforming techniques for sound source imaging. *The Journal of the Acoustical Society of America.*, 125(2):944. ID: 302066175.

Domínguez, A. C. (2009). Pre-processing of speech signals for noisy and band-limited channels. Master's thesis, School of Electrical Engineering, Kungliga Tekniska Högskolan, Stockholm, Sweden.

Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th AES Convention*, pages 18–22.

Fasold, W. (1998). *Schallschutz und Raumakustik in der Praxis; Planungsbeispiele und konstruktive Lösungen*. Verl. für Bauwesen.

Gardner, W. G. (1997). Head tracked 3-d audio using loudspeakers. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA.

Guldenschuh, M. and Sontacchi, A. (2009). Transaural stereo in a beamforming approach. In *DAFx-09*.

Guldenschuh, M., Sontacchi, A., Zotter, F., and Höldrich, R. (2008). Principles and considerations to controllable focused sound source reproduction. In *7th Eurocontrol INO Workshop*.

Keele, D. B. (2003). Full-sphere sound field of constant-beamwidth transducer (cbt) loudspeaker line arrays. *JOURNAL- AUDIO ENGINEERING SOCIETY*, 51:611–624. ID: 208876013.

Kirkeby, O. and Nelson, P. A. (1999). Digital filter design for inversion problems in sound reproduction. *JOURNAL- AUDIO ENGINEERING SOCIETY*, 47(7/8):583–595. ID: 210604884.

Laumann, K., Theile, G., and Fastl, H. (2008). A virtual headphone based on wave field synthesis. In *Acoustics 08 Paris*, pages 3593–3597, Paris, France.

Lentz, T. (2006). Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments. *J. Audio Eng. Soc.*, 54(4):283–295.

Mabande, E. and Kellermann, W. (2007). Towards superdirective beamforming with loudspeaker arrays. In *Conf. Rec. International Congress on Acoustics*, Madrid, Spain.

Mabande, E., Schad, A., and Kellerman, W. (2009). Robust superdirectional beamforming for hands-free speech capture in cars. In *NAG/DAGA 2009*, Rotterdam, NL.

Menzel, D., Wittek, H., Theile, G., and Fastl, H. (2005). Binaural sky: A virtual headphone for binaural room synthesis. In *Tonmeistersymposium 2005*, Hohenkammer, Germany.

Moller, H., Sorensen, M. F., Hammershoi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society.*, 43(5):300. ID: 87542869.

Moore, B. C. (1995). *Hearing*, chapter Frequency Anlaysis and Masking, pages 161–206. Academic Press, San Diego, USA / London, UK.

Morse, P. M. (1953). *Methods of theoretical physics*. McGraw-Hill Book Co, New York ; London.

Möser, M. (1988). *Analyse und Synthese akustischer Spektren*. Berlin [u.a.].

Noisternig, M., Sontacchi, A., Musil, T., and Höldrich, R. (2003). A 3d ambisonics based binaural sound reproduction system. In *24th international AES Conference: Multichannel Audio*.

Oppenheim, A. V. (1989). *Discrete-time signal processing*. Prentice-Hall, Englewood Cliffs, NJ [u.a.].

Schäfer, F. (2008). Artificial bandwidth extension of narrowband speech. Master's thesis, Graz University of Technology, Graz, Austria.

Shin, M., Lee, S. Q., Kim, D., Wang, S., Park, K. H., Fazi, F. M., Nelson, P. A., and Seo, J. (2009). Maximization of acoustic energy difference between two spaces. *to be published*.

Veen, B. D. V. and Buckley, K. M. (1988). Beamforming: a versatile approach to spatial filtering. *ASSP Magazine, IEEE*, 5; 5(2):4–24. ID: 1.

Williams, E. G. (1999). *Fourier acoustics; Sound radiation and nearfield acoustical holography*. Academic Press, London [u.a.].

Yon, S., Tanter, M., and Fink, M. (2003). Sound focusing in rooms: The time-reversal approach. *The Journal of the Acoustical Society of America.*, 113(3):1533. ID: 96597221.

Zwicker, E. (1990). *Psychoacoustics; Facts and models*. Berlin [u.a.].