RWTH AACHEN Aachener Beiträge zur Technischen Akustik

Bruno Sanches Masiero

Individualized Binaural Technology

Measurement, Equalization and Perceptual Evaluation

λογος

Individualized Binaural Technology

Measurement, Equalization and Perceptual Evaluation

Von der Fakultät für Elektrotechnik und Informationstechnik der
Rheinischen-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
genehmigte Dissertation
vorgelegt von

Mestre Engenheiro Eletricista
Bruno Sanches Masiero
aus São Paulo, Brasilien

Berichter:

Universitätsprofessor Dr. rer. nat. Michael Vorländer
Universitätsprofessor Philip Nelson, FREng

Tag der mündlichen Prüfung: 7. Dezember 2012

Bruno Sanches Masiero

Individualized Binaural Technology

Measurement, Equalization and Subjective Evaluation

Logos Verlag Berlin GmbH

λογος

Aachener Beiträge zur Technischen Akustik

Editor:
Prof. Dr. rer. nat. Michael Vorländer
Institute of Technical Acoustics
RWTH Aachen University
52056 Aachen
www.akustik.rwth-aachen.de

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de .

D 82 (Diss. RWTH Aachen University, 2012)

ISBN 978-3-8325-3274-1
ISSN 1866-3052
Vol. 13

Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
D-10243 Berlin
Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
http://www.logos-verlag.de

To

Paulo, Vera, and Livia

Abstract

In this work the importance of individualization in binaural technique is investigated. The results extend the present knowledge on the efficient measurement of individual head-related transfer functions (HRTFs) and highlight the importance of individual equalization filters in binaural reproduction, using both loudspeakers and headphones. Moreover, an integrated framework for the calculation of such equalization filters is presented.

An innovative measurement setup was developed to allow the fast acquisition of individual HRTFs. The hardware was designed to be compatible with the range extrapolation technique, which makes the description of the HRTF's distance-dependence possible. Major speedup was obtained by optimizing the multiple exponential sweep method. An individual HRTF dataset with approximately 4000 directions can be measured in less than 6 minutes with this new setup.

Crosstalk cancellation (CTC) filters are required when playing back binaural signals via loudspeakers. To allow listeners to freely move their heads, switching between multiple loudspeakers is required and the CTC filters must be constantly updated according to the tracked head position. Filter calculations are carried out in frequency-domain for speed reasons. To impose causality constraints to the regularized frequency-domain calculations, a CTC filter calculation framework was proposed, which incorporates a new approach for the multi-channel minimum-phase regularization. This framework also addresses the switching between active loudspeakers through the use of a weighted filter calculation. A sound localization test showed that individualized CTC systems provided performance similar to that from binaural listening while nonindividualized CTC systems provided significantly lower localization performance.

To deliver an authentic auditory impression without additional spectral coloration, binaural reproduction via headphones must be adequately equalized. Such equalization filters are obtained by inverting the headphone transfer function, which varies among listeners and individual fitting. To cope with these variations, a robust individual headphone equalization method was proposed. Perceptual tests showed that, in all but one of the tested situations, no audible differences between the original sound source and its binaural auditory display could be perceived.

Contents

List of Figures

List of Tables

List of Acronyms

3-AFC	three-alternative forced choice
AE	auditory event
ANOVA	analysis of variance
AP	all-pass
ATF	anatomical transfer function
BRvIEH	binaural reproduction via individually equalized headphones
CTC	crosstalk cancellation
CS	channel separation
$\widehat{\text{CS}}$	natural CS
DFT	discrete Fourier transform
DSP	digital signal processor
DTF	directional transfer function
DUT	device under test
EC	ear canal
ED	eardrums
FEC	free-air equivalent coupling
FIR	finite impulse response
FF	free-field
FFT	fast Fourier transform
HP	headphones
HpTF	headphone transfer function
HRIR	head-related impulse response
HRTF	head-related transfer function
IIR	infinite impulse response
ILD	interaural level difference
IR	impulse response
IT	interleaving

ITA	Institute of Technical Acoustics (RWTH Aachen University)
ITD	interaural time difference
LE	lateral error
LMS	least mean square
LSM	least-square minimization
LTI	linear time-invariant
MESM	multiple exponential sweep method
MIMO	multiple input, multiple output
MLS	maximum length sequence
MP	minimum-phase
OSD	optimal source distribution
OV	overlapping
PCA	principal component analysis
PDR	pressure division ratio
PE	polar error
PSD	power spectral density
QE	quadrant error
RIR	room impulse response
RM ANOVA	repeated-measures ANOVA
RMS	root-mean-square
SE	sound event
SISO	single input, single output
SH	spherical harmonic
SM	sequential method
SNR	signal-to-noise ratio
SPL	sound pressure level
TF	transfer function
VBAP	vector base amplitude panning
VR	virtual reality

List of Symbols

Mathematical Notation

$a(t)$	lowercase: signal in time-domain
$A(f)$	uppercase: signal in frequency-domain
$\boldsymbol{a}$	bold lowercase: vector in frequency-domain (unless stated otherwise)
$\boldsymbol{A}$	bold uppercase: matrix in frequency-domain (unless stated otherwise)

Mathematical Operators

$*$	convolution
$\circ$	element wise multiplication
$\lvert\cdot\rvert$	modulus of a complex number
$\lVert\cdot\rVert_2$	Euclidean-norm of a vector or matrix
$\lVert\cdot\rVert_{\boldsymbol{W}}$	weighted-norm of a vector
$(\cdot)^*$	conjugate of a complex number or adjoint operator (Hermitian transpose) of a vector or matrix
$(\cdot)^T$	transpose of a vector or matrix
$\mathrm{adj}(\cdot)$	adjugate operator of a matrix
$\det(\cdot)$	determinant of a matrix
$\mathrm{diag}(\cdot)$	diagonal matrix, whose diagonal entries are given by an input vector
$(\cdot)^+$	Moore-Penrose pseudo-inverse
mod	modulo operation
$\lceil\cdot\rceil$	ceiling operator (round up to next integer)
$(\cdot)^+$	minimum causal stable parts of a function

$(\cdot)^-$	minimum anti-causal stable parts of a function
$[\cdot]_+$	causal part of a function

Mathematical Symbols

$x(t)$	continuous time signal
T_s	sampling period
f_s	sampling frequency
$x(n)$	discrete input time vector
$y(n)$	discrete output time vector
$w(n)$	discrete time window vector
$h(n)$	impulse response of a system
$X(f)$	input spectrum
$Y(f)$	output spectrum
$X(z)$	discrete input spectrum
$H(z)$	transfer function of a system
$D(z)$	desired frequency response
$Q(z)$	equalization filter
T	linear transformation
$\delta(n)$	Kronecker delta function
$\boldsymbol{y}$	measurement vector
$\boldsymbol{A}$	transformation matrix
$\boldsymbol{x}$	vector of least square solutions
$\boldsymbol{I}$	identity matrix
$\mu(z)$	regularization parameter
$s(n)$	discrete time sine sweep
$S(z)$	sine sweep discrete spectrum
$\phi_\mathrm{sw}(n)$	sweep's phase increment
ϕ_0	sweep's starting phase.
$A_\mathrm{reg}(z)$	regularization filter
$H_\mathrm{L}(z)$	left channel HRTF

$H_{\mathrm{R}}(z)$	right channel HRTF
τ_{ikj}	travel time between the i^{th} loudspeaker to the k^{th} microphone at the j^{th} measurement position
$\delta_{\lceil i/8 \rceil}$	latency of an eight channel sound card
x_i	position of the i^{th} loudspeaker
P_{kj}	position of the k^{th} microphone at the j^{th} measurement position
r_k	distance between the k^{th} microphone and the center of the bar holding the microphones
ϕ_j	angle of the turntable at the j^{th} measurement position
d	distance between the bar holding the microphones and the stand connected to the turntable
ρ	angle between the bar holding the microphones and the stand connected to the turntable
γ	torsion of the bar hold the microphones
$\mathbf{x}$	position vector
T_{SM}	total measurement time for SM
T_{MESM}	total measurement time for MESM
τ_{sw}	length of excitation sweep
τ_{st}	length of stop margin
τ_{w}	waiting time
τ_{IR}	length of room impulse response
τ_{DUT}	length of desired impulse response
τ_{sp}	length of safety region
k	harmonic order
a_k	minimum difference between harmonic and fundamental spectrum
r_{sw}	sweep rate
Δt_k	distance between fundamental and k^{th} harmonic IR
$H_{\text{free-field}}$	free-field equalized HRTF
H_{ref}	reference transfer function
H_{diff}	diffuse field HRTF
t^i_{win}	start time of a time window

$\boldsymbol{\theta}$	direction vector
p	pressure value
m	degree of SH
n	order of SH
l	linear SH-index ($l = n^2 + n + m + 1$)
O	truncation order
c_{nm}	complex spherical expansion coefficients
ξ_{nm}	spherical expansion coefficient
h_n	spherical Hankel function of order n
P_n^m	Legendre functions of order n and degree m
$Y_n^m(\boldsymbol{\theta})$	spherical harmonic of order n and degree m for direction $\boldsymbol{\theta}$
k	wavenumber ($k = {}^{2\pi f}/_{c}$)
f	frequency
c	speed of sound
r	radial distance
ϕ	azimuth angle
θ	elevation angle
α	lateral angle
β	polar angle
$\boldsymbol{w}$	weight vector
$\boldsymbol{p}$	pressure vector
$\boldsymbol{Y}$	spherical harmonic basis functions matrix
$\boldsymbol{c}$	spherical expansion vector
$\boldsymbol{\xi}$	spherical potential vector
S	spherical surface
$H_{\mathrm{FF}}^{\mathrm{ED}}(z)$	transfer function from loudspeaker to eardrums
$H_{\mathrm{FF}}^{\mathrm{EC}}(z)$	transfer function from loudspeaker to the microphones at the entrance of the ear canal
$H_{\mathrm{HP}}^{\mathrm{ED}}(z)$	transfer function from the headphone to the eardrums
$H_{\mathrm{HP}}^{\mathrm{EC}}(z)$	transfer function from the headphone the microphones at the entrance of the ear canal.
L	number of loudspeakers

M	length of CTC filters
N	length of HRIR
$\boldsymbol{H}$	transfer matrix
$\boldsymbol{C}$	CTC matrix
$\boldsymbol{b}$	vector of binaural signals
$\boldsymbol{v}$	vector of transaural signals
$\boldsymbol{e}$	vector of ear signals
$\boldsymbol{d}$	vector of desired signals
Δ	CTC filter delay
$R(z)$	band-pass filter

1 Introduction

Spatial audio systems can, among other applications, be found in home entertainment, cinema and virtual reality systems where they are used to provide the listener with an increased sense of realism and immersion. Rumsey (2012) lists some of the spatial attributes required for sound reproduction: naturalness, source localization, distance and depth perception, envelopment and spaciousness, and apparent source width.

There are several different approaches to the creation of spatial auditory impression. Amplitude panning techniques, such as the standard stereophony or its three-dimensional extension, the "vector base amplitude panning" (Pulkki, 1997) or "ambisonics" (as originally formulated by Gerzon, 1973) are based on the psychoacoustic effects of summing localization (Blauert, 1997). These systems have the drawback that all produced phantom sources are perceived at the distance of the loudspeakers. Wave-field reconstruction methods such as "wave field synthesis" (de Vries, 1988) and "higher-order ambisonics" (Daniel, 2003) focus on completely reconstructing a desired sound field inside the reproduction space. These systems require a large number of sources to work properly and even though they are theoretically capable of rendering focused sources, their realization is severely limited in practical applications.

Lentz (2007) and Schröder et al. (2010) describe the virtual reality system at RWTH Aachen University. This system is composed of five video projection walls that severely limit the positioning of loudspeaker to be used with spatial audio generation. To ensure a spatial audio reproduction, thus increasing the immersion in the virtual environment, this system was designed to use yet another type of spatial audio reproduction method, the binaural technology.

The binaural technology can provide the listener with full three-dimensional impression, i.e. lateral position, height and distance impression, with a reduced number of transducers (Møller, 1992). Binaural technology is based on the fact that all spatial sound information perceived by humans is extracted solely from the two pressure signals captured by the listener's ears. These so-called binaural signals can be either directly recorded or they can be synthesized.

Binaural recordings are made using an artificial head or in-ear microphones (Paul, 2009). They have, however, the disadvantage that listener's head movements cannot be directly compensated for. Li and Duraiswami (2006) proposed a method to binaurally "reproduce 3D auditory scenes captured by spherical microphone arrays over headphones". Nevertheless, virtual reality applications commonly rely on binaural synthesis, which will be the focus of this work.

In the binaural synthesis, each sound source must be filtered with a head-related transfer function (HRTF) that describes the direction-dependent influence of pinna, head, and torso on the incident sound field. HRTF are, therefore, listener-dependent (Wightman and Kistler, 1989; Wenzel et al., 1993; Møller et al., 1995a), and, when using binaural technology to create an authentic spatial sound scene, individual HRTFs should be used. As a binaural synthesis is based on acoustic simulations, its performance also depends on the quality of the used acoustic model.

The first part of this work will focus on the characterization of the human listener, i.e. on the acquisition of individual HRTFs. One possibility to acquire HRTFs is to use numerical acoustic simulations, like the boundary element method (Katz, 2001) or the finite difference time-domain (Mokhtari et al., 2007), based on mesh grids obtained from e.g. magnetic resonance imaging (Mokhtari et al., 2007; Guillon et al., 2012), laser scanning (Rui et al., 2012) or photogrammetry (Fels, 2008; Dellepiane et al., 2008). However, as the outer ear contains hidden structures that contribute to the HRTF at high frequencies, the quality of such method is still not ideal. HRTFs can also be obtained through acoustic measurements. This can be done either in a direct (Bronkhorst, 1995; Møller et al., 1995a; Algazi et al., 2001; Majdak et al., 2007) or in a reciprocal manner (Zotkin et al., 2006). No matter which method is used, it is desirable that the HRTF measurement is completed in the shortest time possible, thus providing more comfort for the subject being measured and reducing measurement variability.

The presentation of near-to-head sources can greatly improve the naturalness of virtual reality systems and the synthesis of near-to-head sources requires near-field HRTFs (Brungart and Rabinowitz, 1999). Near-field HRTFs have to be either measured using a complex setup that is composed of transducers placed at increasing distances, or they can be calculated from a measurement at a single distance using the range extrapolation technique (Duraiswami et al., 2004).

The second part of this thesis will deal with the reproduction of binaural signals. The main requirement for a binaural reproduction is

that each of the listener's ears receives a distinct signal. Such a perfect channel separation can easily be achieved by reproducing binaural signals via headphones. However, headphones introduce spectral coloration and may change the acoustic impedance seen from the ear canal, influencing the naturalness of the binaural reproduction. The variation in impedance can be only controlled by choosing the correct headphone types (Møller, 1992). The coloration aspect, which is highly individual and dependent on the headphone fitting (Møller et al., 1995b), can be compensated for by using an individual equalization filter (Pralong and Carlile, 1996; Møller et al., 1995b).

If a pair of loudspeakers is used to directly reproduce a binaural signal, the sound radiated from each loudspeaker will arrive at the listener's ears, thus mixing the binaural cues contained in the binaural signal. To reestablish these cues, crosstalk must be eliminated (Bauer, 1961; Atal et al., 1966). This is achieved by using a crosstalk cancellation (CTC) filter network (Bauck and Cooper, 1993; Köring and Schmitz, 1993; Kirkeby and Nelson, 1999).

The design of CTC filters depends on the disposition of the loudspeakers in relation to the listener. If the listener is located in a limited region in the reproduction space, then the aim of the CTC filter design is to provide a wide sweet spot. This can be optimally achieved with a frequency-dependent loudspeaker distribution (Takeuchi and Nelson, 2007). On the other hand, if the listener should be allowed to freely move inside the reproduction space, then the sweet spot has to be made dynamic by constantly updating the CTC filters based on the listener's current tracked position (Gardner, 1997; Lentz, 2007). To avoid filter instability, the dynamic CTC system should switch between ideal loudspeaker configurations, dependent on the listener's position and direction (Lentz, 2006).

As CTC filters are designed based on the transfer function between loudspeakers and listener's ears, i.e. HRTFs, individualized CTC filters provide a higher channel separation than its generic counterpart (Akeroyd et al., 2007). It is, however, not yet known how a reduced channel separation influences the localization performance of a nonindividualized CTC system.

1.1 Objectives

The global aim of this work is to improve the quality of the binaural technology used in the virtual reality system at RWTH Aachen University

by means of individualization. Therefore only the single listener situation will be considered.

More specifically, the following questions will be discussed in this thesis:

- How to acquire individual HRTFs in a fast manner with full 3D information?
- How important is the individual equalization for a binaural reproduction?
- How can a binaural reproduction be equalized adequately?

1.2 Organization

First, some fundamental aspects of digital signal processing, acoustic measurement and spatial hearing, required for the better comprehension of this work, are presented in chapter 2. The characterization of human listeners for binaural synthesis, i.e. the measurement of individual HRTFs, is discussed in chapter 3, where the design of a measurement system for individual HRTFs is presented, followed by the optimization of the used excitation signals and a discussion on the conducted post-processing. The chapter concludes presenting measurement results obtained using the described system.

The focus is then put on the reproduction of binaural signals. In chapter 4, a reproduction via headphones is examined, testing the adequacy of the of two headphone types for binaural reproduction and presenting a framework for the calculation of individual equalization filters. Binaural reproduction via loudspeakers is then the focus of chapter 5, where a framework for the calculation of individual causal crosstalk cancellation filters in the frequency-domain is presented, which also incorporates the switching between active loudspeakers in 360° scenarios.

The importance of individually equalized binaural reproduction is investigated in chapter 6 by means of two perceptual tests. Section 6.1 evaluates the naturalness of individually equalized headphone binaural presentation while section 6.2 investigates the localization performance of individualized and nonindividualized crosstalk cancellation systems. At the end, chapter 7 summarizes the results and contributions presented in this thesis and possible future research is discussed.

2

Fundamental Concepts

In this chapter some fundamental aspects required for the remainder of this thesis will be briefly summarized. First, a review of the topic of *digital signal processing*, is presented. Digital signal processing is essential when dealing with acoustic measurements, such as the head-related transfer function measurement discussed in chapter 3, and filter designs, such as the headphone equalization and the crosstalk cancellation filters described in chapter 4 and chapter 5.

Both, acoustic measurements and filter design, and the acoustic holography discussed in chapter 3, can be considered as inverse problems. Therefore, some important aspects of how to solve inverse problems are discussed. Acoustic measurements are further analyzed with regard to the used excitation signal and deconvolution methods.

To conclude, a short review of the human directional hearing is presented as the perceptual evaluation presented in chapter 6 is based on it.

2.1 Digital Signal Processing

An acoustic signal can be represented as a real continuous time function $x(t)$. The signal must be discretized so that is can be *digitally* processed. The continuous signal is sampled at a regular interval T_s (reciprocal to the sampling frequency f_s). The result is a sequence of values x_n, where n is the time index. The notation $x(n)$ is used for the vector containing the values of x_n (Oppenheim and Schafer, 1989).

Tohyama and Koike (1998) state that a discrete system is "a system that transforms an input sequence $x(n)$ in an output sequence $y(n)$." Such a system can be described by its impulse response (IR) sequence $h(n)$. In this thesis all systems are considered *linear and time invariant* (LTI, Oppenheim and Schafer, 1989, p. 22). The output sequence of an LTI system is derived from the input and the system's IR by performing the convolution operation

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(n)\, h(n-k). \tag{2.1}$$

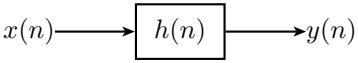


Figure 2.1: A system described by its IR $h(n)$ with input sequence $x(n)$ and output sequence $y(n)$.

The IR of most physical systems is infinitely long and can be described by an *infinite impulse response* (IIR) filter. However, these systems are commonly approximated by a *finite impulse response* (FIR) filter, which allow a more efficient calculation of the convolution operation (Oppenheim and Schafer, 1989).

A finite IR is usually obtained by truncating an infinite IR. To avoid an abrupt end to the signal, the *windowing* operation can be applied, which performs an element-wise multiplication of the sequence with a window sequence whose elements are zero (or close to zero) at the regions of unwanted components and therefore

$$x_{\text{win}}(n) = x(n) \circ w(n). \tag{2.2}$$

Oppenheim and Schafer (1989) show that complex exponential functions are eigenfunctions of LTI systems. Thus, a representation of signals as these complex exponentials, i.e. sinusoids, is very beneficial as a faster calculation of the convolution operation is possible in this domain. A representation of a signal that uses complex exponentials as basis is said to be in the *frequency-domain* while the previous representation is said to be in the *time-domain*. In the frequency-domain, a system is described by its frequency response $H(f)$. The output sequence is obtained from

$$Y(f) = H(f) \cdot X(f). \tag{2.3}$$

$Y(f)$ and $X(f)$ are the complex frequency spectrum of $y(n)$ and $x(n)$ respectively.

Time and frequency representation of a signal are closely related. If the spectrum is, for instance, real, the time representation is symmetric around the origin and noncausal. Furthermore, under certain conditions, the spectrum's magnitude and phase can also be dependent on each other. Therefore, a minimum-phase LTI system—a stable and causal system whose inverse is also stable and causal—has an amplitude and phase that are related by the *Hilbert transform* (Oppenheim and Schafer, 1989, Ch. 11).

A transformation from time into frequency-domain is obtained using the *Fourier transform*. For finite length sequences, however, the *discrete Fourier transform* (DFT) is usually preferred as efficient algorithms exist for its calculation, the so-called *fast Fourier transform* (FFT). The output of the DFT is itself a sequence of samples, equally spaced in frequency, of a signal's Fourier transform (Oppenheim and Schafer, 1989). In the following, the term frequency-domain will refer to the output of a DFT and the DFT from $x(n)$ will be denoted by $X(z)$, the discrete representation of $X(f)$.

The DFT is based on the assumption that the signal is a periodic sequence. Thus, periodicity in both time and frequency-domain is inherent to the DFT and caution is required when conducting some operations with a DFT. It is especially important to note that the linear convolution operation is redefined as a circular convolution under the DFT. This can lead to the presence of time-aliasing if the length of both signals to be convolved is not large enough (Oppenheim and Schafer, 1989). If the result of the linear convolution presents noncausal components, Using the circular convolution these components will now appear at the end of the output sequence, an effect know as *wrap around*.

The shifting operation, which can be understood as a convolution with an impulse shifted in time, must also be redefined as a circular shift. This means that when shifting a sequence by N samples, the last N samples of the sequence are simply moved to the beginning of the sequence.

2.2 Inverse Problems

All signal processing algorithms presented in this thesis can be interpreted as *inverse problems*. Inverse problems occur when a desired signal cannot be measured directly. In this case, the acquired signals are a transformation of the desired signals and an inversion of these transformations is thus required (Mammone, 1999).

Measurements of acoustic systems play a major role throughout this thesis. As explained in section 2.3, measurements of acoustic systems are conducted using deterministic excitation signals rather than an ideal impulse. It is assumed that the excitation signal $x(n)$ is the output of a transformation T to the Kronecker delta function $\delta(n)$. Thus, the system's IR can be obtained from the system's output $y(n)$ by applying the inverse transformation T^{-1} to $y(n)$.

The equalization problem can also be considered as an inverse problem. In this case, the frequency response of the equalization filter $Q(z)$ is obtained from the difference between the desired overall response $D(z)$ and the system's equalized frequency response. The ideal equalization filter is obtained when the observed difference between the desired frequency response and the equalized frequency response is driven to zero, i.e.

$$|D(z) - Q(z)H(z)| = 0, \tag{2.4}$$

$$Q(z) = D(z)/H(z). \tag{2.5}$$

The headphone equalization filter calculation discussed in chapter 4 is a classic example of a *single input, single output* (SISO) equalization problem while the crosstalk filter design discussed in chapter 5 is an example of a *multiple input, multiple output* (MIMO) equalization problem.

Finally, the acoustic holography used for interpolation and range extrapolation in chapter 3 is also an inverse problem as the desired potential expansion coefficients cannot be directly measured and must be estimated from the pressure measurement over a spherical surface by inverting the spherical harmonic transformation.

The inverse problems cannot always be directly inverted as in eq. (2.5). This is especially true for the MIMO cases, which can have an over- or underdetermined transformation matrix. These problems can, however, be considered a minimization problem of the form

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2, \tag{2.6}$$

where $\|\cdot\|^2$ is the Euclidean norm, $\boldsymbol{y}$ is a vector containing the measured values, $\boldsymbol{A}$ is the transformation matrix, and $\boldsymbol{x}$ is the solution of the inverse problem.[1] The solution to eq. (2.6) is given by

$$\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{y} = \begin{cases} (\boldsymbol{A}^{*}\boldsymbol{A})^{-1}\boldsymbol{A}^{*}\boldsymbol{y}, & \text{if } \boldsymbol{A} \text{ is overdetermined} \\ \boldsymbol{A}^{-1}\boldsymbol{y}, & \text{if } \boldsymbol{A} \text{ is invertible} \\ \boldsymbol{A}^{*}(\boldsymbol{A}\boldsymbol{A}^{*})^{-1}\boldsymbol{y}, & \text{if } \boldsymbol{A} \text{ is underdetermined} \end{cases}, \tag{2.7}$$

where $\boldsymbol{A}^{*}$ represents the Hermitian transpose of matrix $\boldsymbol{A}$ and $\boldsymbol{A}^{+}$ its Moore-Penrose pseudo-inverse assuming that $\boldsymbol{A}^{*}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^{*}$ are not singular.

[1] Even though the norm to be minimized does not necessarily have to be the Euclidean norm, this norm will be used throughout this work, as it has an analytical solution derived from the *least-squares minimization* (LSM) when $\boldsymbol{A}$ is overdetermined and from *constraint optimization* when $\boldsymbol{A}$ is underdetermined (Nelson and Elliott, 1995, pp. 418–419).

It is often the case that the observed data contains additive noise and that the transformation matrix is not well-conditioned. This leads to an undesirable noise amplification effect that can severely affect the quality of the obtained solution (Fazi, 2010). Therefore, regularization can be introduced to mitigate the minimization problem and avoid overfitting. The most commonly used regularization scheme is the Tikhonov regularization that adds a restriction on the total energy of the solution vector.

In this case, the transformation matrix $\boldsymbol{A}$ is overdetermined, the minimization eq. (2.6) is altered to

$$\min_{\boldsymbol{x}} \left(\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu \|\boldsymbol{x}\|_2^2 \right), \tag{2.8}$$

where μ is the regularization parameter with real values in the range $0 \leq \mu \leq \infty$. μ acts as a trade-off factor between the residual error and energy of the solution. Moreover, for the equalization problems, μ also acts as an upper limit of the resulting filter's maximum gain (see appendix A).

The regularization has, however, its drawbacks. The use of regularization to solve equalization problems leads to ringing artifacts in the time-domain, as discussed by Bouchard et al. (2006) for the one channel case and by Norcross and Bouchard (2007) for the multi-channel case.

The solution of the new minimization problem eq. (2.8) is given by

$$\boldsymbol{x} = \begin{cases} \left(\boldsymbol{A}^* \boldsymbol{A} + \mu \boldsymbol{I}\right)^{-1} \boldsymbol{A}^* \boldsymbol{y}, & \text{if } \boldsymbol{A} \text{ is overdetermined} \\ \boldsymbol{A}^* \left(\boldsymbol{A}\boldsymbol{A}^* + \mu \boldsymbol{I}\right)^{-1} \boldsymbol{y}, & \text{if } \boldsymbol{A} \text{ is underdetermined} \end{cases}, \tag{2.9}$$

where $\boldsymbol{I}$ is the identity matrix.[2] The solution of the underdetermined case is thoroughly explained in appendix B.

2.3 Acoustic Measurement

Acoustic measurements focus either on acoustic signals or on acoustic systems. The measurement of acoustic signals is common in fields such as noise control, e.g. to define how "loud" an acoustic source is, or to describe how "loud" it is inside a room, or even to define how long

[2] In case $\boldsymbol{A}$ is a square matrix, any of the two solutions can be used.

a person can be exposed to a given noise. Measurements of acoustic signals can also focus on the binaural technology. In this case, the recording is carried out using either in-ear microphones or an artificial head (Paul, 2009). Binaural recordings, however, are not covered in this publication. Therefore, the term *acoustic measurement* will in the following always refer to the measurement of an acoustic system, in particular the measurement of the head-related transfer function (HRTF), as described in chapter 3.

Acoustic systems are assumed to the linear and time-invariant and can thus be described by its impulse response (IR) or by its frequency equivalent transfer function (TF).[3] If the spatial characteristic of an acoustic system is described, e.g. the directivity of a loudspeaker or a directional microphone, then a series of IRs measured at different directions will be required. The IR could be measured directly by feeding the system with an infinitely narrow impulse and recording the systems response. However, an impulse contains very little energy. When measuring under normal conditions, i.e. with background noise present, the obtained signal-to-noise ratio (SNR) will be low and many repetitions of the measurement will be necessary for an improved SNR (Müller and Massarani, 2001).

To avoid repeated measurements, the *correlation technique* is commonly applied. This technique allows measurements to be conducted with any type of excitation signals, e.g. white noise, pink noise, or music. Nevertheless, some signals may be more suitable than others as their energy is better distributed over time and, therefore, yield a better SNR. Several different excitation signals have been studied regarding their performance in terms of SNR, crest factor, and measurement duration. From all these signals, sine sweeps and pseudo-random sequences are commonly preferred as such deterministic excitation signals provide high measurement repeatability (Müller, 2008).

As discussed in section 2.2, the TF of an acoustic system is obtained by dividing the output spectrum of the system under test by the spectrum of the input signal. In time-domain, this division is equivalent to filtering the output signal with a *matched filter*, which is itself equivalent to calculating the signals' *cross-correlation* (Müller, 2008). Therefore, a class of binary signals exhibiting unity auto-correlation was commonly

[3] According to Müller (2008), "the IR can be transformed into the TF via Fourier transform (see section 2.1) and back again into the IR via the inverse Fourier transform, both are equivalent and carry the same information, which can be extracted and visualized in different ways." Thus, the terms IR and TF will be used in reciprocal form throughout this thesis.

employed in acoustic measurements. As listed by Peltonen (2000), Golay codes (Golay, 1961), Legendre sequences (Schroeder, 1979), Barker codes (Kuttruff, 2000), and *maximum length sequences* (MLS) (Borish and Angell, 1983) are examples of such sequences. The MLS technique became particularly popular as its auto-correlation can be calculated in an extremely efficient manner—in terms of computation time and memory usage—using the fast Hadamard transform. The main drawback of using pseudo-random sequences is, however, that this technique is very sensitive to time-variance and nonlinearity in the measurement chain (Müller and Massarani, 2001).

As the calculation time of IRs became less critical due to the improvement of the calculation complexity and speed of state-of-the-art personal computers, sweep measurements became increasingly popular. Sweeps offer great advantages for systems that do not fully comply with the LTI assumption, e.g. weak time variances (slow changes of the system response over time) or nonlinear transfer characteristics (harmonic distortion) (Müller and Massarani, 2001), as it can still achieve high SNR where pseudo-random sequences would succumb.

2.3.1 Exponential Sweep

Sweeps, also known in literature as chirp or swept-sine, are generally defined in time-domain as

$$s(n) = \sin(\phi_{\mathrm{sw}}(n) + \phi_0) \tag{2.10}$$

where $\phi_{\mathrm{sw}}(n)$ is the phase increment and ϕ_0 the starting phase (Müller and Massarani, 2001). It is convenient to set the starting phase to $\phi_0 = 0$ as this results in a smooth start of the signal.

The two most commonly known types of sine sweeps are the linear and the exponential sweep. They differ in terms of how $\phi_{\mathrm{sw}}(n)$ varies with time. They are aptly named as the former varies linearly in time while the later varies exponentially in time.

Exponential sweeps, when compared with linear sweeps, are known to have an advantage when it comes to non-linear systems, besides they contain higher energy at the lower frequency range, which is exactly the region where measured SNR tends to be worse. Müller and Massarani (2001) show that the nonlinear behavior a system, when measured with an exponential sweep, can be observed as anti-causal IRs for different harmonic orders k,[4] each having a length $\tau_{\mathrm{IR,k}}$ and

[4] The harmonic order k starts at 2, as $k = 1$ is the desired fundamental IR itself.

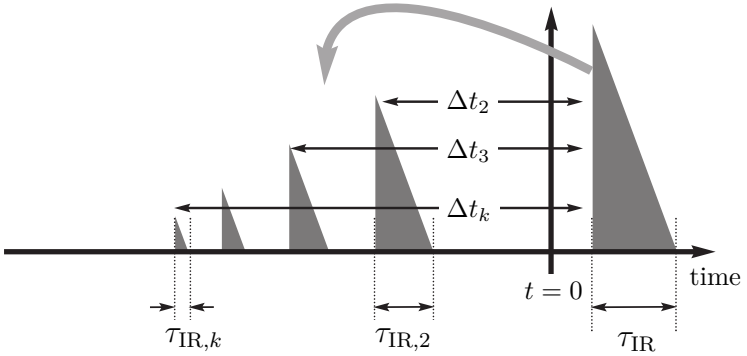


Figure 2.2: IR of a weakly nonlinear system obtained by exponential sweep measurement.

lagging Δt_k from the fundamental IR. The logarithmic magnitude of a weakly nonlinear system's IR measured with an exponential sweep is displayed schematically in fig. 2.2. Unless distortion measurements are being conducted, the fundamental IR (located to the right side in this example) is the result that we are actually interested in.

2.3.2 Regularized Deconvolution

When dividing the output spectrum by the input spectrum problems may arise for frequencies where the values of the exponential sweep $S(z)$ become very small. This would also be the case if the sweep covers only a limited bandwidth. The division of the output spectra by these small values can amplify the additive noise commonly present at the obtained system's output. This leads to undesirable artifacts in the IR.

In order to solve this problem, Farina (2007) introduced the Tikhonov regularization for sweep measurements. The regularized inverse sweep is thus obtained by applying eq. (2.9) to the excitation sweep. Multiplying the matched filter with the system's output results in the regularized transfer function

$$H_{\mathrm{reg}}(z) = \frac{Y(z)}{S(z)} \frac{1}{1 + {}^{\mu(z)}/_{|S(z)|^2}} = \frac{Y(z)}{S(z)} \cdot A_{\mathrm{reg}}(z), \tag{2.11}$$

where $A_{\mathrm{reg}}(z)$ describes the influence of the regularization as a filter. Note that the regularization parameter $\mu(z)$ is now frequency-dependent.

In this case, if the exponential sweep is band-limited in $[f_1, f_2]$, it is reasonable to keep $\mu(z) = 0$ inside this range and to set $\mu(z) \gg \max |S(z)|^2$ elsewhere. By doing so, $A_{\text{reg}}(z)$ will suppress the measurement noise outside the desired frequency range.

Note, however, that $A_{\text{reg}}(z)$ is a zero-phase band-pass filter, which means that its temporal counterpart $a_{\text{reg}}(n)$ is symmetric regarding the time axis. As described by Bouchard et al. (2006), this leads to the noncausal behavior observed in IRs obtained using the regularized deconvolution.

Minimum-phase Regularization

To solve the problem of noncausality, $A_{\text{reg}}(z)$ can be factorized into a minimum-phase (MP) regularization filter and a remaining noncausal all-pass (AP) filter (Tohyama and Koike, 1998; Bouchard et al., 2006):

$$A_{\text{reg}}(z) = A_{\text{reg,MP}}(z) \cdot A_{\text{reg,AP}}(z). \tag{2.12}$$

If $A_{\text{reg,MP}}$ is used in the deconvolution process, the resulting fundamental IRs obtained when measuring a physical (and therefore causal) system will also be causal. This is especially important for the multiple exponential sweep method presented in chapter 3. For this technique noncausal IRs are problematic as the pre-ringing from one IR can overlap with a previous IR, which leads to artifacts when the IRs are cropped out of the raw IR. The phase error introduced in the pass-band by the minimum-phase filter can be later compensated by filtering the resulting IR with the all-pass component $A_{\text{reg,AP}}(z)$, yielding back noncausal IRs.

2.4 Directional Hearing

The spatial impression perceived by human beings is based on cues imprinted on the signal arriving at the listener's ear, the so-called *binaural signal*. These cues are caused by an alteration on the incoming wave front due to reflection, diffraction, shadowing, resonance, and dispersion at the listener's body (mainly torso, head, and pinna), which means that these cues are highly individual (Wightman and Kistler, 1989; Wenzel et al., 1993; Møller et al., 1995a; Fels, 2008).

Møller et al. (1996) analyzed the localization performance of subjects with their individual and nonindividual binaural recordings. They concluded that "when compared to real life, the localization performance

was preserved with individual recordings. Nonindividual recordings resulted in an increased number of errors for the sound sources in the median plane".

Similar results have also been reported for synthesized binaural signals. Wenzel et al. (1993) analyzed the localization performance of 16 subjects presented with the HRTF of another representative subject. They concluded that "while the interaural cues to horizontal location are robust, the spectral cues considered important for resolving location along a particular cone-of-confusion are distorted by a synthesis process that uses nonindividualized HRTFs."

Middlebrooks (1999b) compared the localization performance obtained with a loudspeaker (the natural condition) and with individual and nonindividual synthesized auditory displays. He verified that "performance in the own-ear virtual (individually synthesized) condition was nearly as accurate as that in the free-field (natural listening) condition." Furthermore, he reported that "all error measures of RMS errors tended to increase with increases in the spectral difference between the listener's (directional transfer function) DTFs and the DTFs used in the localization trials".

Middlebrooks (1999b) went further and scaled the nonindividual DTFs to make it more similar to the listeners own DTF. This alteration resulted in an improved localization performance, indicating the possibility of creating "a realistic virtual synthesis of auditory space for the large number of listeners for whom it would not be practical to make individual acoustical measurements of DTFs."

The results listed above indicate the importance of individual—or at least individualized—binaural synthesis. It is important to mention that all these results were obtained with inexperienced listeners without a training period. Hofman et al. (1998) demonstrated "the existence of ongoing spatial calibration in the adult human auditory system", i.e., they showed that human listeners are able to learn how to hear with a different ear. Another very important conclusion obtained by them was that "learning the new spectral cues did not interfere with the neural representation of the original cues, as subjects could localize sounds with both normal and modified pinnae" after a sufficiently long training period. Specifically for sound localization tests, Majdak et al. (2010) showed the importance of training as subjects "learn to better localize sounds in terms of precision, bias, and quadrant error".

Specifically regarding virtual reality applications, Begault et al. (2000) argued that besides individual HRTFs, addition of head-tracking

significantly decreased quadrant confusion and the addition of reverberation significantly improved the impression of externalization.

2.4.1 Head-Related Transfer Functions

The alterations caused at the wave front arriving from any given direction can be described by a pair of filters, frequently called *head-related transfer functions* (HRTF, Blauert, 1997), but also known as *anatomical transfer function* (ATF, Hartmann, 1999). There are two transfer functions for each direction of sound incidence ($H_{\mathrm{L}}(z)$ for the left and $H_{\mathrm{R}}(z)$ for the right ear), which are combined into one HRTF.

An example of a measured HRIR is shown in fig. 2.3. This HRIR was measured with the sound source positioned at the right side of the listener. This information can be easily extracted from this plot, as the right signal is louder than the left signal, as this is attenuated by the head shadowing and it also arrives earlier than the left signal, as the acoustic path between the source and the left ear is longer than to the right ear.

The amplitude of the equivalent HRTF is shown in fig. 2.4. Once again, it is easy to verify that the source is at the listener's right side, as the right signal is louder throughout the whole human listening range. It is also possible to verify that this level difference is more pronounced in the higher frequency range.

The resonance behavior observed at fig. 2.4 can be associated to anthropometric characteristics of the human listeners. At low frequencies very low variance is observed, as for these frequencies the head and torso are small when compared to the wavelength. The shoulder reflection will create a comb filter effect, with the lowest resonance occurring in the region of 1.5 kHz and repeating every 3 kHz. The influence of the pinnae in the HRTFs is significant only to frequencies above approximately 2 kHz. A constructive resonance in the pinna results in a global maximum at approximately 5 kHz. A sharp minimum in frequencies around 9 kHz is caused by reflections in the cavum conchae back wall. As this HRTF was measured with an artificial head with no ear simulator, the typical resonance in the range of 8 kHz that occur inside the ear canal is not present.

A spherical head-related coordinate system is used (see fig. 2.5) to describe the direction of the sound incidence. The origin of this coordinate system is in the center of the head between the ears. The azimuth angle ϕ rotates counterclockwise between 0° (front direction) and

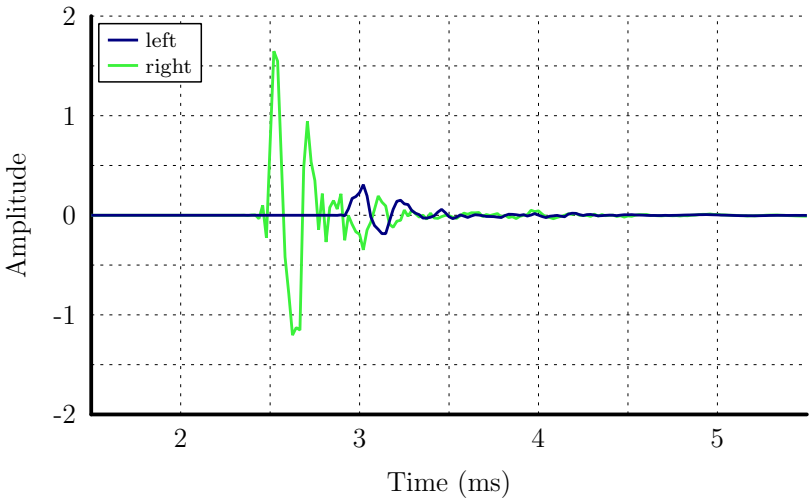


Figure 2.3: Example of head-related impulse response for sound incidence from the right.

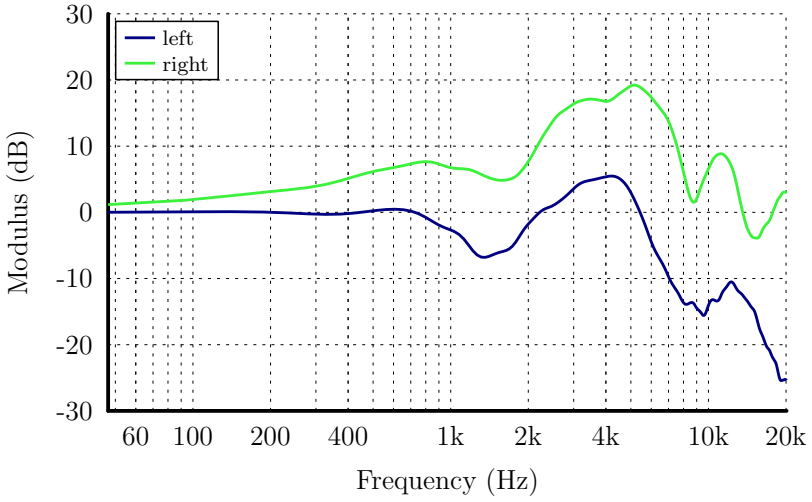


Figure 2.4: Example of head-related transfer function for sound incidence from the right.

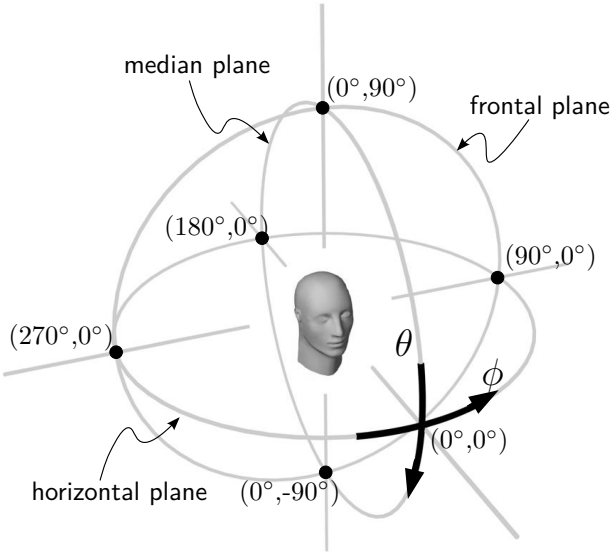


Figure 2.5: Spherical coordinate system used in the HRTF measurements. Elevation angle θ is defined in the range -90° $\leq \theta \leq$ 90° and the azimuthal angle ϕ is defined in the range 0° $\leq \phi \leq$ 360°.

360°. The elevation angle θ is defined from -90° (bottom) to 90° (top). Three planes are also defined: 1) the horizontal plane (θ=0°), 2) the frontal plane (ϕ=±90°) and 3) the median (sagittal) plane (ϕ=0°,180°). Planes parallel to the median plane are called sagittal planes.

Different definitions for the ATF can be found in relevant literature. Blauert (1997) describes three definitions of the ATF; one of them will be used throughout this thesis, the free-field HRTF.[5] This describes the transformation from the sound pressure generated by a sound source in far-field and measured at the center of the head (with the head absent) to the pressure generated by the same source at the same position and measured at the entrance of the listener's ear canals.

Another ATF definition used in this thesis is the *directional transfer function* (DTF), which removes the components common to all HRTFs, described by the diffuse-field HRTF (Middlebrooks, 1999b). The diffuse-field HRTF is computed by taking the root-mean-square (RMS) of the sound pressure at each frequency averaged across all measured HRTFs.

[5] Simply called HRTF in the following.

Both the HRTF and the DTF contain, however, no distance information and can thus only be used to render sources at far-field, i.e., plane wave sources. This restriction is usually not a hindrance for virtual reality (VR) applications, especially for room acoustic simulations. However, the most impressive binaural demonstrations occur when sound sources are located in the near-field, increasing the realism of VR scenes (Lentz, 2007).

A very important recent contribution in the field of individual HRTF is the range extrapolation technique, described by Duraiswami et al. (2004) as "a way to obtain the range dependence of the HRTF from existing measurements conducted at a single range!"

2.4.2 Sound Localization

Sound localization is a very complex mechanism performed by the human brain. It is not only dependent on the directional cues contained in the binaural signal captured at the ears, but it is also intertwined with the other senses, especially vision and proprioception (Seeber, 2002).

While binaural disparities like interaural time and level differences (ITD and ILDs) play an important role for sound localization in the horizontal plane (Macpherson and Middlebrooks, 2002), monaural spectral cues are known to determine the perceived sound-source position in the sagittal planes (top/down, front/back; Blauert, 1969). In section 6.2 the localization performance will be discussed with regard to these two mechanisms.

Non-acoustical cues such as head movements, also called sounding (Blauert, 1997), are avoided by keeping the listener's head still during the signal presentation.

In fig. 2.6 a new spherical coordinate system is defined that describes the acoustic target's position on the listening experiment, based on a horizontal-polar coordinate system (Morimoto and Aokata, 1984). Again, the origin of the coordinate system is in the center of the head. The lateral angle α is defined from $-90°$ (right) to $90°$ (left). The polar angle β rotates counterclockwise between $0°$ (front) and $360°$. Every lateral angle defines a sagittal plane.

Localization in Horizontal Plane

In the horizontal plane, mainly the binaural cues (ITD and ILD) are used for the localization. The interdependency of these two cues is described by the *duplex theory* (Rayleigh, 1907).

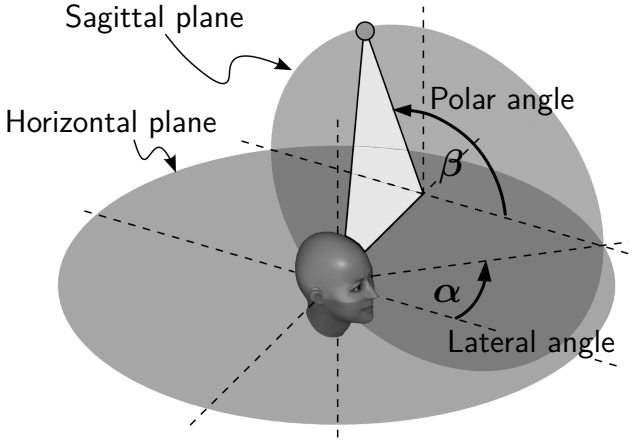


Figure 2.6: Spherical coordinate system used in the localization experiments. Lateral angle α is defined in the range -90° $\leq \alpha \leq$ 90° and the polar angle β is defined in the range 0° $\leq \beta \leq$ 360°.

Hartmann (1999) summarizes the matter of binaural cues as follows: "The physiology of the binaural system is sensitive to amplitude cues from ILDs at any frequency, but for incident plane waves, ILD cues exist physically only for frequencies above about 500 Hz. They become large and reliable for frequencies above 3000 Hz, making ILD cues most effective at high frequencies. In contrast, the binaural physiology is capable of using phase information from ITD cues only at low frequencies, below about 1500 Hz."

In section 6.2 the localization performance in the horizontal plane is analyzed with regard to the lateral error, which is the RMS of the localization error[6] in the lateral dimension (Middlebrooks, 1999b).

Localization in the Sagittal Plane

Binaural cues can only assist in the perception of lateral angles on the horizontal plane. Confusion in the sagittal plane, usually experienced as a reversal between front and back, or up and down, may occur as ITD and ILD remain (approximately) constant along the polar angles. The region where such confusions occur is called *cones of confusion* (Mills, 1972).

[6] Localization errors are calculated by subtracting the target angles from the response angles.

The hearing system relies on the monaural cues to assert the polar angle within a cone of confusion. These are direction-dependent spectral coloration of the sound due to the asymmetry of the head and especially the pinna (Middlebrooks, 1999b).

In section 6.2 the localization performance in the sagittal plane is analyzed with regard to the polar error and the quadrant error (Middlebrooks, 1999b). LE is the RMS of the localization error in the polar dimension and is used to quantify the local performance in the polar dimension. QE is the percentage of responses where the absolute polar error exceeded 90° and it is used to describe the degree of confusions.

3

Measurement of Individual Head-Related Transfer Function

Depending on the measurement setup used and the amount of directions to be measured, measurements of head-related transfer functions (HRTFs) can be very time-consuming. In contrast to artificial heads that remain static and may therefore be measured over a long period of time, human subjects have difficulty keeping their position—especially their head position—for a longer period of time. Thus, individual HRTF measurements should ideally be conducted in the shortest amount of time possible, to reduce positioning error and improve the comfort of the subject being measured.

A large number of previous work in measuring individual HRTF has been conducted by Bronkhorst (1995); Møller et al. (1995a); Algazi et al. (2001); Zotkin et al. (2006); Majdak et al. (2007), and Lentz (2007), just to name a few. Different setup strategies have been used by the different research groups: e.g. one loudspeaker being moved on an arc with a mechanical stepping device or many loudspeaker fixed on an arc or even several microphones distributed over a sphere. For the setups listed above, typical measurement duration varies from 20 min to 2.5 h to measure 1000 to 1500 HRTFs with an average angular resolution of approximately 5°.

This chapter describes a new HRTF measurement setup that was developed to allow the fast acquisition of individual HRTFs. Moreover, this setup is one of the first of its kind designed to be compatible with the range extrapolation technique. The setup consists of a circular arc and up to 40 broadband loudspeakers that can be distributed (almost) arbitrarily along the arc. By rotating the subject horizontally inside this

*Most of the results presented in this chapter have been previously published at

- Masiero; Pollow, and Fels (2011a);
- Masiero; Pollow; Dietrich, and Fels (2012);
- Dietrich; Masiero, and Vorländer (2012a);
- Pollow; Masiero; Dietrich; Fels, and Vorländer (2012b).

arc, HRTFs are measured at fixed points on a spherical grid, discretized over azimuth and elevation angles. A continuous description of the HRTF can be obtained in a post-processing stage using interpolation algorithms.

This chapter starts by introducing the requirements of the new HRTF measurement setup and describing the hardware solutions chosen to meet these criteria. Then the explanation of the optimized multiple exponential sweep method (MESM) is presented. It is used to accelerate the total measurement duration by over 90% compared with the sequential measurement method and by over 50% compared with the original MESM. Then the post-processing operations of equalization, interpolation, and range extrapolation, which are applied to the measure HRTFs, are discussed.

3.1 Hardware Design

The main aspect common to all modern HRTF measurement setups is that the measurement itself should be concluded in the shortest time possible. A method commonly used in numerical acoustics to reduce computation time is the reciprocity method, where source and receivers are exchanged to be able to simulate multiple receiver points at a single run. An HRTF measurement method based on the principle of reciprocity was proposed by Zotkin et al. (2006) using a miniature sound source placed at the entrance of the blocked ear canal and 32 microphones distributed on a spherical array of 0.7 m radius. As the excitation signal has to be played only once for each ear, this results in a very short measurement time. However, the use of a miniature sound source yields a considerably smaller signal-to-noise ratio (SNR) and restricts the measurement frequency range to frequencies above approximately 1 kHz. Zotkin et al. (2006) used a model-based extension of generic HRTFs for frequencies below this limit. If a high spatial resolution is desired, many microphones positions have to be measured, which increases the hardware costs, or measurement has to be repeated for different configurations of a smaller array, increasing measurement time.

Microphones can, in contrast to loudspeakers, be miniaturized without severe restrictions to sensitivity levels and working frequency range. A direct HRTF measurement system using two miniature microphones, ideally placed in the entrance of the blocked ear canal (Møller et al., 1995b), can provide satisfactory SNR levels. Direct

HRTF measurement setups can be divided into three categories regarding the number and configuration of physical sound sources in the setup.

- **Dense array**: with as many loudspeakers as directions to be measured.
- **Hybrid array**: with a group of loudspeakers placed on an arc, where either the arc or the subject is turned.
- **Sparse array**: only one loudspeaker is moved in every direction to be measured.

If high spatial resolution is desired, a dense setup will require a large number of hardware channels, drastically increasing the costs and complexity of the setup. This is also true for reciprocal measurements. A sparse measurement setup, on the other hand, will always be the slowest of all methods as no parallelization of the measurement procedure is possible. For example, the system built at TNO in the Netherlands needs 2.5 h to complete a measurement with 976 directions (Bronkhorst, 1995). For such setups, it is common that subjects wear a head tracking device to verify that the head position is kept constant throughout the complete measurement duration.

The hybrid array is a trade-off between speed and hardware complexity and is therefore most commonly found, e.g. the setup at the CIPIC Labs in USA that can measure an HRTF set at 1250 directions in approximately 1.5 h (Algazi et al., 2001) or the more recently constructed setup from the Acoustics Research Institute (ARI) in Austria that can measure almost 1500 directions in approximately 20 min (Majdak et al., 2007). These two setups differ in one main aspect. While in the CIPIC setup the listener sits still and the arc is rotated along the subject's interaural axis, the ARI setup has a static arc and the subject is rotated around its longitudinal axis. Because of the above-mentioned advantages, a hybrid array was chosen for the measurement arc.

Not only the array type varies in between HRTF measurement setups, but also the distance from sound sources to the listener. For example, the CIPIC's system uses an arc with 1 m radius while the system build at the Aalborg University has a radius of 1.95 m (Møller et al., 1995a), allowing subjects to be measured in standing position.

This setup should be capable of measuring both near- and far-field HRTF. One possibility would be to construct a number of arcs with different radii, as described by Lentz (2007). Another possibility is to apply the range extrapolation technique, based on the acoustical spherical holography that enables us to calculate the near-field HRTF

from the far-field measurements or vice versa. Therefore, a setup with a single arc and, thus, fixed distance was preferred.

As evanescent spherical waves will have faded away in far-field, a measurement in far-field might not contain enough information to allow the reconstruction of such waves in near-field. For this reason measurements should ideally be conducted as close as possible to the listener's head. On the other hand, to apply acoustical spherical holography, all scattering objects important for the HRTFs (the shoulders and upper torso) must be contained inside of the spherical measurement surface. If measurements were conducted with an arc of short radius, a large region in the bottom of the sampling sphere would remain uncovered, as the listener's body would be on the way of the loudspeakers. Since, Pollow et al. (2012a) could not verify that outward extrapolation was more robust than inward extrapolation and Brungart and Rabinowitz (1999) stated that "HRTFs are virtually independent of distance for sources beyond 1 m", the measurement arc was planned to have a radius of 1 m.

3.1.1 Sound Source

In acoustical spherical holography, the scattering problem is modeled as a distribution of *point sources* on a sphere containing all the scattering objects (Zotter, 2009, p. 34). When placed in the far-field, the radiation pattern of a point source can be approximated by an incident plane wave. Thus, when measuring HRTFs in far-field, regular loudspeakers can be used.

On the other hand, when measuring HRTFs at near-field, these larger loudspeakers are inappropriate as their directivity pattern diverges from that of a point source. Furthermore, two-way loudspeakers are also inadequate because the HRTFs will be blurred as their acoustic center moves from the woofer to the tweeter as the frequencies increase. Attempts have been made to produce acoustic transducers that mimic a point source. Brungart and Rabinowitz (1999) developed a source with an electrodynamic horn driver placed at the end of a long tube. Unfortunately, as this construction acts as a wave guide the frequency response shows many peaks and dips (due to wave interference) and equalization might become problematic. Qu et al. (2009) proposed the use of an electric spark as an acoustic point source. Such sources do display a flat frequency response and an almost omnidirectional radiation pattern. However the impulse generated by them is not repeatable and

the obtained SNR levels are far below the usual levels obtained using loudspeakers and correlation technique.

An ideal loudspeaker for near-field HRTF measurement setup should meet the following design criteria:

1. Broad-band reproduction
2. Omnidirectional radiation pattern
3. Low distortion artifacts due to nonlinearity

A large membrane is required to be able to generate enough sound pressure at low frequencies. In the other hand, omnidirectional radiation pattern can only be achieved if the driver's membrane is small compared with the wavelength. Thus, a small membrane is required to radiate omnidirectionally at higher frequencies. It is then clear that the choice of a single membrane size will force a trade-off between pressure level at low frequencies and omnidirectionality at high frequencies. The desired frequency range was therefore reduced from the complete audible range to the range between 300 Hz and 20 kHz. This is a reasonable restriction as, according to MØLLER et al. (1995a), HRTFs show very little individual variation below 300 Hz and its asymptotic behavior can be extrapolated (cf. section 3.3.2). Even with a relaxed frequency range restriction, only a handful of broad-band drivers are still able to meet these requirements.

Three loudspeaker drivers were analyzed regarding their frequency range, maximal sound pressure level (SPL), distortion level, and directivity. The driver with the highest maximal SPL and consequently lowest nonlinear distortion had relatively large dimensions. It was discarded due to its bundled directivity. The other two drivers showed equivalent characteristics, with lower maximal SPL, higher nonlinear distortion, and a smoother directivity pattern in the frontal direction (due to the smaller membrane diameter). Since the loudspeakers on the arc will be relatively close to the microphones, maximal SPL was not defined as a critical parameter and therefore the smaller driver with 32 mm diameter was chosen as it also allowed an easier mechanical fixation at the enclosure

To radiate in the low frequency range the chosen driver needs an enclosure, otherwise the air would just move from the front to the back of the membrane in an acoustic shortcut (reactive intensity) and no sound wave would propagate outwards (active intensity). According to the Thiele-Small parameters calculated for the chosen driver, a volume of at least 100 ml is required to allow a sound reproduction down to 300 Hz. Note that even such a small enclosure can influence the radiated sound field due to its edge diffraction. An optimization of the enclosure was carried out to minimize these effects, as described below.

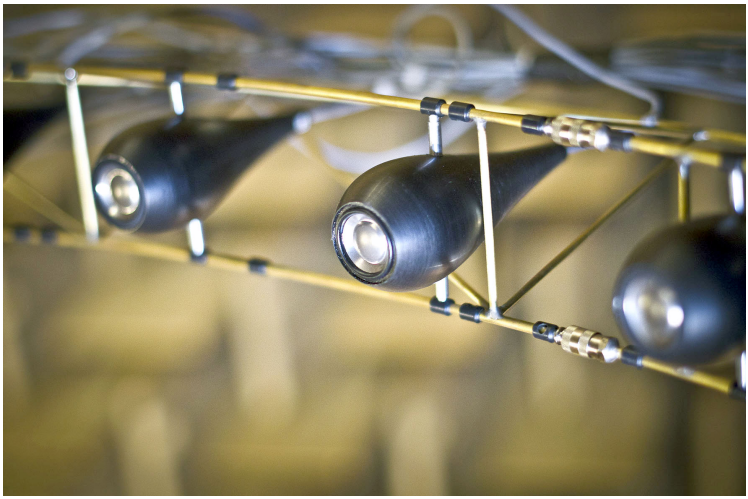


Figure 3.1: Developed drop-like loudspeaker mounted on an arc element. The perpendicular supporting truss structure allows the loudspeakers to be placed freely within delimited regions of the arc.

Enclosure Optimization

In the first step the driver's membrane velocity was measured with a laser doppler vibrometer at 154 points and these values were used as input data for the loudspeaker simulation. The vibrometry results showed the presence of eigenmodes only at very high frequencies. Three forms of enclosures were simulated: a cylinder with rounded front edge, a cylinder with both front and back edges rounded and a drop-like enclosure. All forms avoid, in varying degrees, sharp edges responsible for diffraction. Simulation results showed that, for a point on axis 1 m away from the membrane, the drop-like enclosure has the least influence in the loudspeakers frequency response and directivity (Sartor, 2010).

Furthermore, influences due to possible sound reflections by neighboring loudspeakers have to be considered. In order to verify which form yields the best results, another simulation was carried out using three identical loudspeakers placed on an arc with a 1 m radius placed 10° apart. The central loudspeaker was set as the sound source while the other two loudspeakers were left inactive, as mere diffraction bodies. Again, the drop-like enclosure showed a slightly lower influence on the radiated sound field. This form was therefore chosen for the loudspeaker enclosure, as shown in fig. 3.1. The frequency responses of

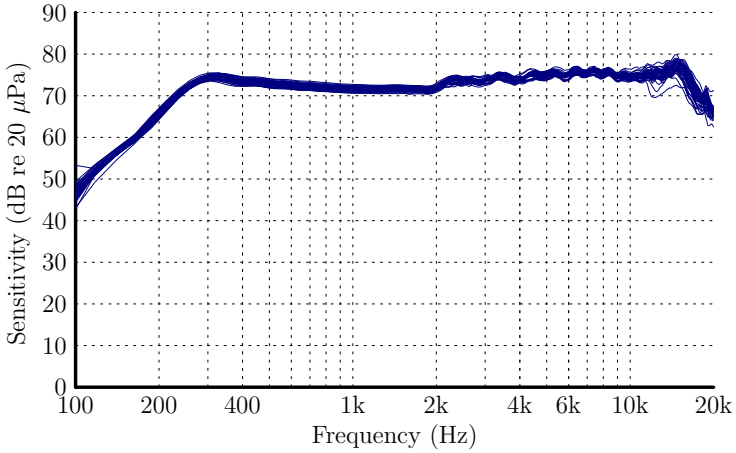


Figure 3.2: Frequency response of 40 drop-like loudspeakers measured at 1 m and 1 V.

the 40 constructed loudspeakers are shown in fig. 3.2. The constructed loudspeakers provide a relatively flat frequency response in the range from 300 Hz to 14 kHz and an acceptable SNR can be achieved up to 20 kHz. The measurement shows low variability between loudspeakers. However, an individual equalization is still required.

3.1.2 Supporting Arc & Head-Rest

The design of the arc that supports the loudspeakers also aims at minimizing the influence of the arc on the radiated sound field to avoid reflection and diffraction effects. Although they are easier to manufacture, bulky structures have a great influence on the sound field and should be avoided. On the other hand, a thin metal rod can be considered acoustically transparent if its diameter is much smaller than the wavelength of the impinging sound wave. The supporting arc was therefore designed with thin metal rods in a trellis structure to minimize disturbing scattering effects while providing sufficient stability.

As mentioned earlier in this chapter, the radius of the supporting arc was defined to be 1 m. As a person has to stay in the middle of the loudspeaker array, the use of a complete circle is not feasible. Hence, an

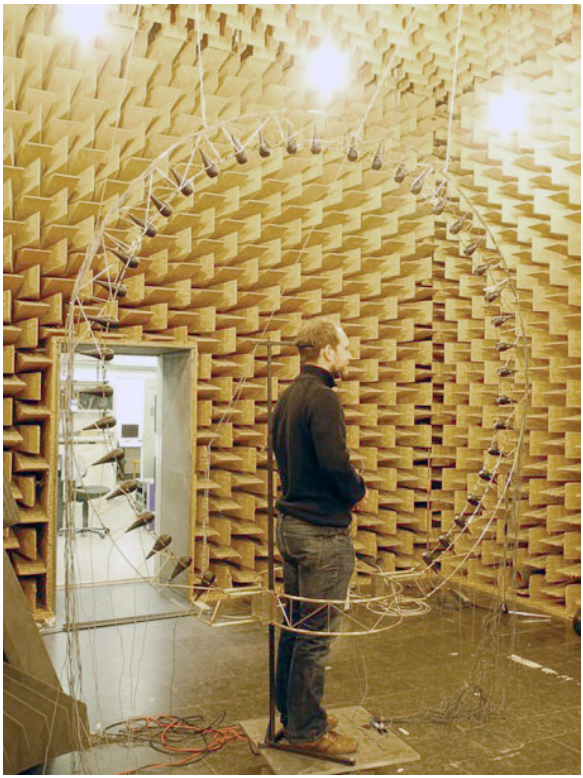


Figure 3.3: Picture of listener placed inside the developed HRTF measurement system. The metal plate where the listener stands is fixed on a turntable. The used head-rest can be seen behind the listener.

arc of 300° was chosen, allowing measurements of elevation angles from −60° to 90°.[1]

Due to its light construction, the supporting arc displays an underdamped oscillatory behavior. If the supporting arc where thus to be rotated, like in Algazi et al. (2001), a long settling time would be necessary until the arc reaches its rest position again. It was decided to keep the arc stationary and to rotate the subjects inside the arc as the settling time for the human head is assumed to be shorter than that of the arc. The subject was rotated with the help of a turntable

[1] For more details on the developing stages of the arc, please refer to (Masiero et al., 2011a).

	position variation (cm)		
	x	y	z
without head-rest	4.60	5.70	0.30
with head-rest	0.08	0.06	0.10

Table 3.1: Measurement of the head position variation for one person standing still for two minutes with and without a head-rest.

and a head-rest device was used to help the test subjects to keep still during measurement. A controllable turntable, already available at the Institute of Technical Acoustics, was used for this purpose. A head-rest was constructed with a thin metal bar fixed to the turntable at the bottom end and connected at the top end to a Y-shaped metal bar that could be adjusted in elevation and depth to be adapted to the listener's head.

Tests with a subject wearing a position tracking device showed that the natural displacement could be considerably reduced with the help of a read-rest (see table 3.1). This result is only demonstrative as it was carried out with one subject only. During the measurement, it was verified that the most critical aspect of positioning was to place the listener with its longitudinal axis matching the turntable's rotation axis. However, as long as this misplacement remains constant throughout the whole measurement, it might be possible to compensate for it at the post-processing stage (Ziegelwanger, 2012).

3.1.3 Data Acquisition and Amplifiers

In order to drive all loudspeakers independently a multi-channel measurement setup was put together. A computer is connected to two multi-channel professional sound cards that are commercially available. These were then connected via an optical interface to five commercial AD/DA converters. The converters' DA output is connected to two 20-channel, low noise and low distortion amplifiers, specially developed for this setup, with a maximum power of 10 W per channel. Two miniature microphones are directly connected to the AD input of one of the AD/DA converters. A connection diagram of the complete system is presented in fig. 3.4.

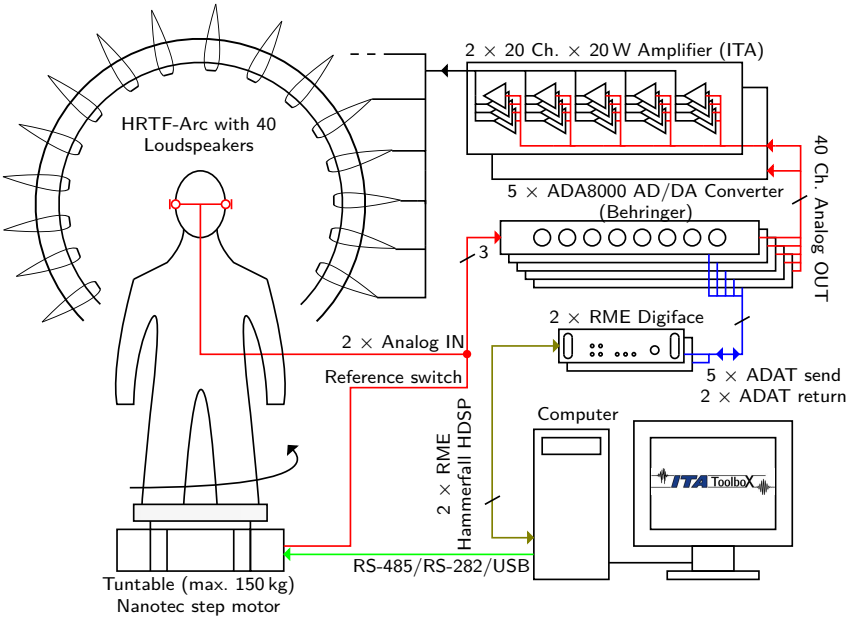


Figure 3.4: Diagram of the measurement setup. Adapted from Krechel (2012).

3.1.4 Sampling Grid

The measurement setup described in this chapter allows the loudspeakers to be placed at almost any position on the arc, only limited at the positions where perpendicular truss supporting structure is present (see fig. 3.1). As a hybrid model was chosen, where the listener is rotated in azimuth and the loudspeakers have a fixed elevation, only axisymmetric grids can be used. To improve the measurement time, the chosen grid should measure at each azimuth position as many points in elevations as possible. Examples of such spatial grids are the equiangular grid, the Gaussian grid, and the IGLOO grid (Zhang et al., 2012).

A longitude-latitude grid is any grid where points are equally distributed in T elevation angles ($\Delta\theta = 180°/T$) and U azimuth angles ($\Delta\phi = 360°/U$). An equiangular grid is formed when $\Delta\theta = \Delta\phi$. For calculations in the SH-domain (spherical harmonic domain, see section 3.3.2), the most efficient longitude-latitude grid is the Gaussian grid, whose elevation angles are defined at the roots of the Legendre polynomial for the desired order and which the inverse problem can be

efficiently solved using the weighted Hermitian of the matrix containing the spherical harmonic base functions (Zotter, 2009).

These kind of sampling schemes display a concentration of points at the poles ($\theta = 0°$ and $\theta = 180°$). To avoid this concentration and still allow a similar interpolation quality, the IGLOO method discards points in the polar region (Zhang et al., 2012), therewith achieving a greater sampling efficiency,[2] though losing the advantage of efficient calculation in the SH-domain.

The placement of the loudspeakers in the constructed setup shows some minor deviations from the exact positions of the desired sampling scheme. These deviations are caused by the structural restrictions previously mentioned. Therefore, the exact locations of the loudspeakers have to be determined so that they can be used for further data processing. A method to extract the loudspeakers' position based on acoustic measurements was described by Krechel (2012). A system consisting of two microphones mounted on an aluminum bar placed exactly 40 cm from each other is therefore used. This bar is mounted on a support arm, which is fixated at a stand, itself attached to the center of the turntable.

The transfer function between loudspeakers and the two microphones are measured for at least two azimuthal positions. The travel time τ_{ikj} between the i^{th} loudspeaker to the k^{th} microphone at the j^{th} measurement position is estimated by first convolving the measured impulse response with its minimum-phase version, then taking the Hilbert transform of the resulting convolution and finally searching for the zero-crossing in the Hilbert-transformed signal closest to the maximum value of the original convolution result.

A system of linear equations is set up using the estimated distance between the microphones and the known angle of the turntable during the measurements to obtain the exact position of the loudspeakers.

$$x_{i,\text{opt}} = \arg\min_{x_i} \sqrt{\sum_{k=1}^{2}\sum_{j=1}^{n}\sum_{i=1}^{L}\left[\|x_i - P_{kj}\| - \left(\tau_{ikj} - \frac{\delta_{\lceil i/8 \rceil}}{f_\text{s}}\right)\cdot c\right]^2} \tag{3.1}$$

where $\delta_{\lceil i/8 \rceil}$ is the latency of each sound card,[3] and c the speed of sound. x_i is the position of the loudspeaker and P_{kj} is the position of the k^{th}

[2] Zotter (2010) defines the *sampling efficiency* as the ratio between the number of points in the actual grid and the minimum number of points required to correctly represent the spherical harmonic order O, which is the maximum order obtainable with the array without spatial aliasing (please see section 3.3.2 for more details).

[3] Each sound card has eight channels and it is assumed that all channels have the same latency.

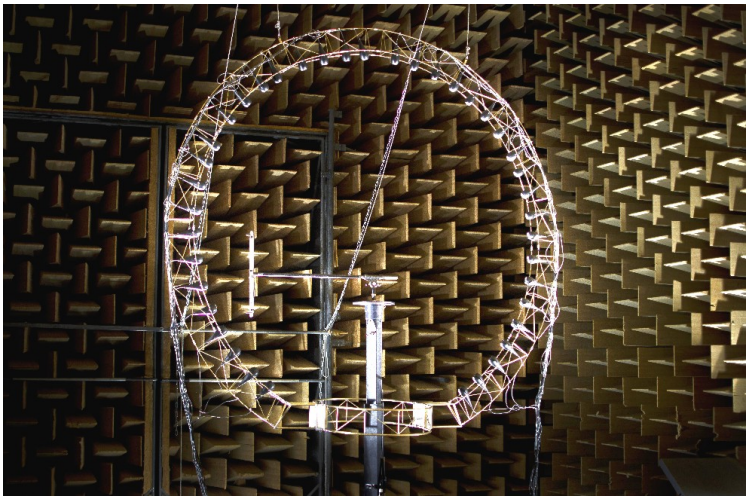


Figure 3.5: Setup developed for the measurement of the real position of the loudspeakers. Two microphones are placed at the top and bottom of the vertical bar at an exact distance of 40 cm from each other. The bar is fixated by a stand to the turntable and is turned to provide multiple measurement positions.

microphone at the j^{th} measurement position, given by equation

$$P_{kj} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \tag{3.2}$$

$$\begin{pmatrix} \sin(\gamma) \cdot r_k \cdot \cos(\phi_j) - [\sin(\rho) \cdot d + \cos(\rho) \cdot \cos(\gamma) \cdot r_k] \cdot \sin(\phi_j) \\ \sin(\gamma) \cdot r_k \cdot \sin(\phi_j) + [\sin(\rho) \cdot d + \cos(\rho) \cdot \cos(\gamma) \cdot r_k] \cdot \cos(\phi_j) \\ -\cos(\rho) \cdot d + \sin(\rho) \cdot \cos(\gamma) \cdot r_k \end{pmatrix},$$

where r_k is the distance between the k^{th} microphone and the center of the bar holding the microphones, ϕ_j is the angle of the turntable at the j^{th} measurement position, d is the distance and ρ the angle between the bar holding the microphones and the stand connected to the turntable, and γ is the torsion of the bar hold the microphones. The fact that the obtained system of equations is overdetermined is very practical as it allows not only to estimate the position of the loudspeaker, but also the orientation of the bar, its distance to the rotation axis, the latency of each sound card, and the speed of sound at the time of measurement.

3.2 Excitation Signal

The maximum speedup in the measurement time is achieved if all loudspeakers on the arc can be used simultaneously during the measurement. Xiang and Schroeder (2003) and Vanderkooy (2010) showed that parallel measurement can be performed with pseudo-random sequences, since they are mutually orthogonal. However, just as in the single channel case, measurements carried out using pseudo-random sequences are very sensitive to time-variance and nonlinearity in the measurement chain (Müller and Massarani, 2001). As discussed in section 2.3, sweeps offer greater robustness for measuring systems that do not fully comply with the LTI assumption, e.g. weak time variances (slow changes of the system response over time) or nonlinear transfer characteristics (harmonic distortion). Furthermore, with a little tweak, sweeps can also be used for a multiple parallel excitation.

Instead of mutual orthogonality, the LTI principle of superposition is now analyzed: if the system output $g(t)$ is composed by the addition of two or more filtered versions of the (time-shifted) input signal $s(t)$,

$$g(t) = \sum_i h_i(t) * s(t - \tau_i), \tag{3.3}$$

then the deconvolution of $g(t)$ by $s(t)$ will result in

$$h'(t) = \sum_i h_i(t - \tau_i). \tag{3.4}$$

As long as the input signals are adequately shifted in time, the IRs $h_i(t)$ can still be restored from $h'(t)$. Majdak et al. (2007) introduced a new fast measurement method for weakly nonlinear systems by using exponential sweeps and an optimization strategy to overcome the interference in the measurement between nonlinearities—appearing as noncausal harmonic IRs—and the system's IRs. Two different strategies to avoid this interference were proposed and combined using an optimization algorithm, with respect to either measurement time or SNR, yielding the so called *multiple exponential sweep method* (MESM).

It will be shown that, if the reverberation time of the room where the measurement is conducted is small enough, then a generalized overlapping strategy is sufficient (Dietrich et al., 2012a). Furthermore, assuming that only the first 5 ms of the measured IR contains important information for the HRIR—the rest being unwanted reflections—the sweeps can be overlapped even closer to each other, yielding an even

faster measurements with unchanged accuracy. Using an optimized version of the multiple exponential sweep technique, approximately four thousands discrete points can be measured within less than six minutes.

3.2.1 Multiple Exponential Sweep Method

The MESM proposed by Majdak et al. (2007) reduces the measurement duration significantly compared with sequential measurements using the exponential sweep method where the number of loudspeakers is high. Using the traditional sequential method (SM), the duration of a measurement made with N loudspeakers is given by

$$T_{\mathrm{SM}}(N) = N \cdot (\tau_{\mathrm{sw}} + \tau_{\mathrm{st}}), \tag{3.5}$$

where τ_{sw} is the length of the excitation sweep and τ_{st} is the stop margin, i.e. the time required to allow the system to decay after the sweep has ended.

Using the MESM, the sweeps are played back with a certain waiting time or delay τ_{w} between each subsequent sweep. Hence, sweeps of several loudspeakers might run (partly) in parallel. As a new sweep starts every τ_{w}, in the ideal case without any nonlinearities, the measurement duration with N loudspeakers is given by the sum of the waiting time of the first $N-1$ sweeps plus the length of the last sweep τ_{sw} and the required stop margin τ_{st}, thus

$$T_{\mathrm{MESM}}(N) = (N-1)\tau_{\mathrm{w}} + \tau_{\mathrm{sw}} + \tau_{\mathrm{st}}. \tag{3.6}$$

Usually the length of excitation sweeps used for HRTF measurement lies in the range from 0.2 s to 2 s, for very short to moderately long sweeps. If the measurements are conducted in suitable anechoic environments, τ_{w} is estimated to be in the range from 20 ms to 200 ms and $\tau_{\mathrm{st}} = \tau_{\mathrm{w}}$. Comparing eq. (3.5) with (3.6) and using the nominal values listed above, a theoretical speedup of 88% can be expected with the MESM when compared with the SM. The parallel measurement shows great potential for a large N, long sweeps, and a short waiting time. Hence, the minimization of this delay is of interest.

Again, in the ideal case without any nonlinearity, the smallest possible value for τ_{w} is the reverberation time τ_{IR} of the room where the measurements are conducted, as described in fig. 3.6. However, if the system is weakly nonlinear, the noncausal harmonic IRs could be superposed

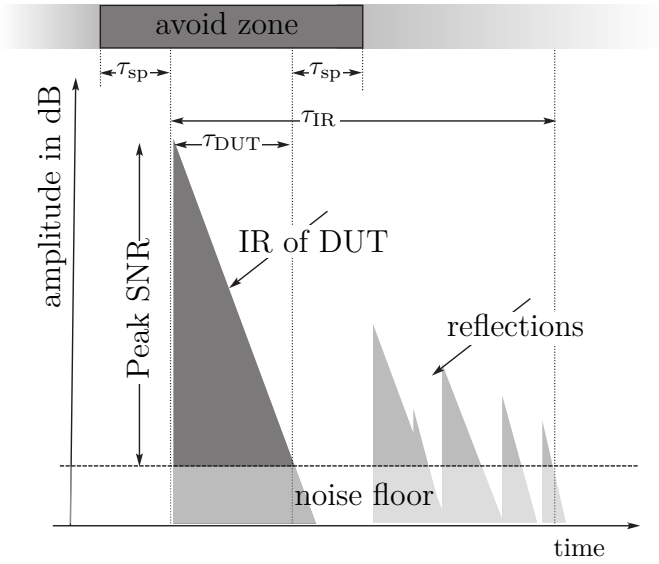
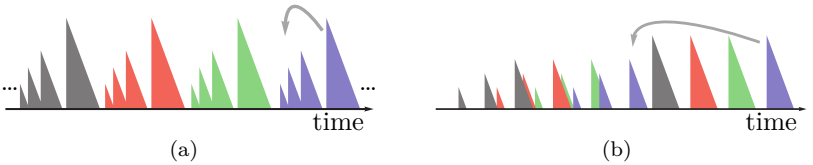


Figure 3.6: Temporal structure of an IR of a linear system measured in an anechoic environment with an exponential sweep. Reflections are caused by the measurement setup itself and other foreign objects in the room.

with the IR of interest, irrevocably corrupting the measurement. Majdak et al. (2007) suggested the use of *overlapping* (OL) and *interleaving* (IL) to avoid this from happening (see fig. 3.7).

When overlapping the harmonic IRs appear between the IRs of interest as shown in fig. 3.7(a). A drawback of the occurrence of harmonic IRs is the waiting time $\tau_{w,OL}$ that has to be increased (compared to the ideal situation)so that it does not interfere with the region of interest. Furthermore, the sweep rate can be increased to shorten the delay between fundamental and harmonic IRs. The maximum order of harmonics k_{max} present in the measurement has to be finite, and preferably small, to allow a small $\tau_{w,OL}$.

When interleaving, η IRs of interest are grouped together, placing as many fundamental IRs as possible in the time span between the first fundamental IR and its corresponding first harmonic IR, as illustrated in fig. 3.7(b). Contrary to overlapping, the sweep rate in this case should be decreased to enlarge the delay between fundamental and harmonics and thus fit more IRs of interest (i.e. fundamental IR) inside this gap.



(a) (b)

Figure 3.7: Schematic example of the temporal structure of the measurement of a weakly nonlinear system limited to three harmonic IRs obtained with (a) the overlapped IR method and (b) the interleaved IR method (with $\eta = 4$).

Majdak et al. (2007) also described how to combine both methods using the two different time delays $\tau_{\mathrm{w,IL}}$ and $\tau_{\mathrm{w,OL}}$. These methods overlap groups of interleaved sweeps, as illustrated in fig. 3.8(a). They describe that the optimum solution minimizing measurement duration is given by the interleaving waiting time

$$\tau_{\mathrm{w,IL}} = \tau_{\mathrm{IR}}, \tag{3.7}$$

that depends only on τ_{IR} and the overlapping waiting time

$$\tau_{\mathrm{w,OL}} = \Delta t_k + \eta\, \tau_{\mathrm{IR}} \tag{3.8}$$

where Δt_k is the time interval between the desired IR and the furthest harmonic IR (cf. fig. 2.2), which gives the time distance between the beginning of the first harmonic IR and the end of the last desired IR belonging to the same interleaved block. Dietrich et al. (2012a) showed that

$$\Delta t_k = \frac{\log_2(k_{\max})}{r_{\mathrm{sw}}}, \tag{3.9}$$

where r_{sw} is the sweep rate.

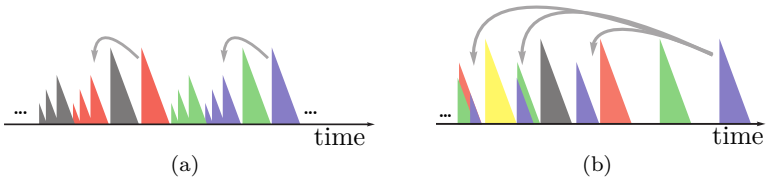
According to Weinzierl et al. (2009), the optimum value for η is given by

$$\eta = \left\lceil \frac{\tau_{\mathrm{IR}} - \tau_{\mathrm{IR},2} + {}^{1}\!/_{r_{\mathrm{sw}}}}{\tau_{\mathrm{IR}}} \right\rceil, \tag{3.10}$$

where $\tau_{\mathrm{IR,k}}$ is the length of the k^{th} harmonic RIR (see section 2.3).

The SNR and the temporal and spectral structure of the results obtained with sequential measurements and the MESM will remain the same if the following requirements are met.

1. The system is weakly nonlinear, i.e. the number of harmonic IRs present in the measurement is small.



(a) (b)

Figure 3.8: Schematic example of an IR measured with (a) the MESM as suggested by Majdak et al. (2007) with $\eta = 2$ and (b) with the optimized MESM described in this work. Note that no harmonic IR superposes the desired fundamental IRs.

2. In case nonlinearity is observed, the output level must be kept constant during the actual measurement and the calibration measurement (used to determine the number of harmonic IRs and actual length of the desired IR).
3. The smallest delay τ_{w} between two subsequent sweeps must be larger than the length of the desired IRs.
4. nonlinearity should be restricted to elements of the measurement chain where the excitation signal is not yet superposed, i.e. amplifiers and loudspeakers. Microphones and preamplifiers have to be driven in their linear range only as nonlinearity at this stage will introduce inter-modulation that might corrupt the measurement if not taken into account at the optimization stage.

3.2.2 Optimized MESM

Majdak et al. (2007) claim that their MESM provides minimal measurement duration. However, while analyzing the temporal structure of a usual HRIR measured in an anechoic environment, Dietrich observed that an additional speedup of the measurements was still possible.[4] As will be shown later in this section, if the region of interest is only a small fraction of the measured IR, a generalized overlapping strategy (with wait time $\tau_{\mathrm{w,OPT}}$) can further accelerate the measurement process.

Temporal Structure of Measured IR

It is sensible to assume that the IR of an acoustic system is causal and that its energy decays exponentially. The measurement of an HRTF is

[4] Personal communication with Pascal Dietrich in 2011.

equivalent to the measurement of directional transfer functions, where the *device under test* (DUT) is the listener's head and loudspeakers are used as sound sources.

In such measurements, the obtained IRs consist of a direct sound path plus reflection and diffraction of the listener's body, containing the desired spectral and directional information, followed by reflections due to objects or room boundaries which can be understood as unwanted artifacts. This situation is depicted in fig. 3.6. The overall length of these IRs τ_{IR} is limited by the moment all reflections cease or disappear below background noise. These reflections might still occur even in *anechoic* environment, caused either by a hard floor (hemi-anechoic chamber) or by other necessary objects in the room, e.g. lamps, support frames, doors, pedestals, etc.

It is important to notice that only the beginning of the IR, with the length τ_{DUT}, has to be protected against reflections and harmonic IRs—that might overlay the desired IR during a measurement with MESM (Dietrich et al., 2012b). Hence, an *avoid zone* around the IR of the DUT is defined by adding a safety region τ_{sp} before and after the desired response.

The definition of the avoid zone suggests that, in contrast to the MESM, the regions containing only unwanted reflections can be used to place the harmonic IRs. The overlapping method, and hence the MESM, can directly benefit from this observation by adapting eq. (3.8) to

$$\tau_{\mathrm{w,OL}} = \frac{\log_2(k_{\max})}{r_{\mathrm{sw}}} + (\eta - 1)\,\tau_{\mathrm{IR}} + \tau_{\mathrm{DUT}} + \tau_{\mathrm{sp}}. \tag{3.11}$$

Equations (3.7) and (3.10) remain unchanged, resulting in a shorter measurement duration.

Placement Strategies for Harmonic IRs

The harmonic IRs present in the measured IR do not necessarily have to be cumulated in blocks, as advocated by the MESM (Dietrich et al., 2012b). The only constraint for a valid measurement is that no harmonic IRs fall into the avoid zones, as illustrated in fig. 3.8(b).

As a practical consideration, the measurement system described in this thesis is only weakly nonlinear and can be reasonably quantified by claiming a value for total harmonic distortion below 10% for all

frequencies.[5] This results in an attenuation of all harmonics of at least 20 dB. As harmonics show the same decay rates as the fundamental IR, harmonics $\tau_{\mathrm{IR,k}}$ will always be shorter than the fundamental IR τ_{IR}. The maximum order $k_{\max}$, before the harmonics fall below background noise. The length of each harmonic IR can be obtained from a calibration measurement using sequential sweep measurements or it can be estimated from

$$\tau_{\mathrm{IR,k}} = \frac{\mathrm{SNR} - a_k}{\mathrm{SNR}} \tau_{\mathrm{IR}}, \tag{3.12}$$

where a_k is the minimum difference (in dB) between the spectrum of a harmonic k compared with the spectrum of the fundamental.

To avoid the desired IR to be corrupted by the room reflections present in the previous IR, the waiting time between sweeps must fulfill $\tau_{\mathrm{w}} \geq \tau_{\mathrm{IR}}$. Considering that all harmonic IRs must fit between two subsequent desired fundamental IRs, the waiting time constraint must be extended to satisfy

$$\tau_{\mathrm{w}} \geq \max\left(\tau_{\mathrm{DUT}} + 2\,\tau_{\mathrm{sp}} + \max(\tau_{\mathrm{IR,k}}), \tau_{\mathrm{IR}}\right) \tag{3.13}$$

Additionally, the start of each k^{th} harmonic IR must fall after the end of an avoid zone and its end must appear before the next avoid zone starts. Both constraints can be written as

$$(\Delta t_k \bmod \tau_{\mathrm{w}}) \geq \tau_{\mathrm{DUT}} + \tau_{\mathrm{sp}} \tag{3.14}$$

and

$$(\Delta t_k \bmod \tau_{\mathrm{w}}) + \tau_{\mathrm{IR,k}} \leq \tau_{\mathrm{w}} - \tau_{\mathrm{sp}}. \tag{3.15}$$

Combining the above-mentioned constraints and also substituting eq. (3.9) results in

$$\tau_{\mathrm{DUT}} + \tau_{\mathrm{sp}} \leq \left(\frac{\log_2(k)}{r_{\mathrm{sw}}} \bmod \tau_{\mathrm{w}}\right) \leq \tau_{\mathrm{w}} - \tau_{\mathrm{sp}} - \tau_{\mathrm{IR,k}}. \tag{3.16}$$

Optimization of Parameters

No analytic solution is known for finding the values $(\tau_{\mathrm{w}}, r_{\mathrm{sw}})$ that satisfy the inequalities (3.13) and (3.16) while minimizing τ_{w} and thus the

[5] Loudspeakers commonly present higher distortion levels at lower frequencies. The use of a shelving filter to suppress power at this region can reduce nonlinearity and consequently reduce the size of the harmonic IRs with the consequence of decreasing the observed SNR at this frequency range.

measurement's duration. The straightforward approach to solve this problem is to use exhaustive search (Dietrich et al., 2012a). The two-dimensional search space $(r_{\mathrm{sw}},\tau_{\mathrm{w}})$ can be normalized by the τ_{IR}. The resulting normalized search space is $(r_{\mathrm{sw}} \cdot \tau_{\mathrm{IR}},{}^{\tau_{\mathrm{w}}}/_{\tau_{\mathrm{IR}}})$ and it has the advantage that the optimization procedure becomes independent of the IR's length. The normalized size of the avoid zone is then given by

$$\alpha = \frac{\tau_{\mathrm{DUT}} + 2\tau_{\mathrm{sp}}}{\tau_{\mathrm{IR}}}. \tag{3.17}$$

It will be later shown that the smaller the value of α, the higher the chance that the new method will yield a faster measurement than the MESM. The solution is, however, dependent on the parameters k_{max} and $\tau_{\mathrm{IR,k}}$.

Valid combinations of $(\tau_{\mathrm{w}}, r_{\mathrm{sw}})$ are shown in fig. 3.9 for an example with $k_{\mathrm{max}} = 4$, $\tau_{\mathrm{sp}} = 0\,\mathrm{s}$, $\alpha = 1$ and no decrease of the length of the harmonics: $\tau_{\mathrm{IR,k}} = \tau_{\mathrm{IR}}$. Valid combinations can always be found for high sweep rates and long delays as in this region the method is equivalent to the original overlapping method. For moderate r_{sw} the allowed τ_{w} is shorter than the overlapping method (Dietrich et al., 2012b). For very low sweep rates, the range of valid delays resulting in valid solutions becomes increasingly smaller, so that almost no valid stable solutions can be found with the numeric search algorithm used due to a finite number of discrete search points. Because of its instability, the optimized MESM should be avoided for very small r_{sw}.

Due to a strong fluctuation of the minimum delay over the sweep rate observed in fig. 3.9, it becomes evident that the sweep rate should not be fixed prior to the search. A better approach is to define a search range for the r_{sw} rate and choose the r_{sw} corresponding to the minimum value of τ_{w}. The change in SNR caused by varying r_{sw} can be neglected in most cases.

3.2.3 Numerical Comparison

As the original MESM does not take into account the temporal structure of the IR, the optimized method is compared with the original method with $\tau_{\mathrm{IR}} = \tau_{\mathrm{DUT}}$. For comparison, the maximum number of harmonics is set to $k_{\mathrm{max}} = 4$ and $\tau_{\mathrm{IR,k}} = \tau_{\mathrm{IR}}$, which can be seen as a worst case scenario for typical loudspeakers. As displayed in fig. 3.10, both methods always result in a minimum normalized delay shorter or equal to the delay obtained with just the overlapping method. The new method

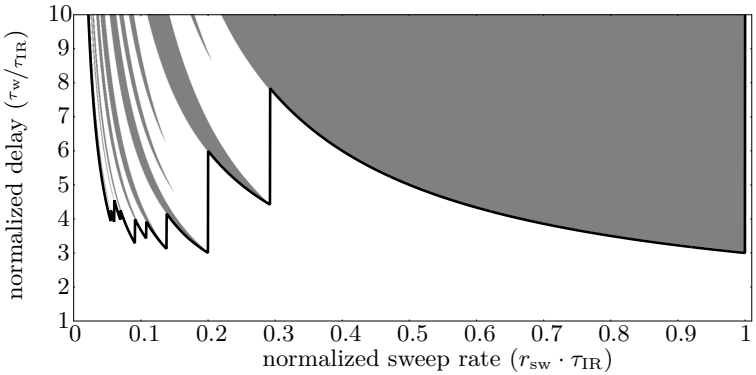


Figure 3.9: Normalized possible solutions for $k_{max} = 4$, $\alpha = 1$, $\tau_{IR,k} = \tau_{IR}$ (white: interference of harmonics with fundamentals, gray: no interference, black: minimum possible delay between sweeps)

shows slightly lower values only for sweep rates in the lower-mid ranges and for low values of α.

To conclude, a comparison is made for real values obtained using the HRTF measurement setup described in section 3.1 placed in the hemi-anechoic chamber in ITA. The desired HRIRs are very short, with an approximate duration of $\tau_{HRIR} = 4\,\text{ms}$ (Hammershøi and Møller, 2005). On the other hand, the IR of the hemi-anechoic chamber (containing reflections from the floor, supports, mounts, and doors) has a length in the order of $\tau_{IR} = 40\,\text{ms}$, thus $\alpha = 0.1$.

Sufficient SNR can be achieved with $\tau_{sw} = 1.5\,\text{s}$, which yields over 80 dB peak-to-noise ratio for the desired HRIR. The frequency range of interest is defined from 0.1 to 18 kHz. This corresponds to a sweep rate of $r_{sw} \approx 5$. The avoid zone was enlarged by $\tau_{sp} = 1\,\text{ms}$. The maximum observed harmonic order was $k_{max} = 5$ and a_k was defined in the frequency-domain: $a_2 = -35\,\text{dB}$,$a_3 = -45\,\text{dB}$, $a_4 = -40\,\text{dB}$ and $a_5 = -40\,\text{dB}$.

The best combination of sweep rate and delay found with the optimization algorithm in the region around $r_{sw} = 5$ was $r_{sw,opt} = 5.59$ and $\tau_{w,opt} = 48.095\,\text{ms}$. The new sweep has a length of 1.34 s and the theoretical change in SNR caused by the shorter excitation signal is estimated to be $\Delta\text{SNR} = 10\log_{10}\left(r_{s,opt}/r_{sw}\right) = -0.48\,\text{dB}$.

The sequential measurement of $N = 40$ loudspeakers will take 53.71 s to conclude. Using the original MESM proposed by Majdak et al. (2007), the measurement time is reduced to 7.39 s with an average waiting time

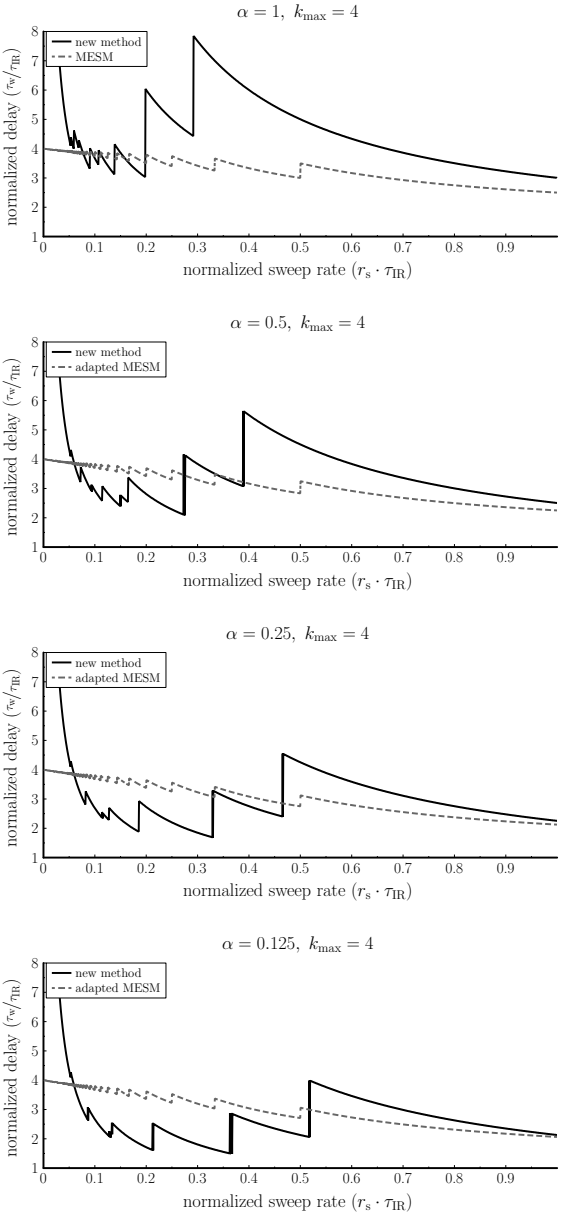


Figure 3.10: Comparison of minimum normalized delay obtained with the original and the optimized MESM for different values of α and $k_{\max} = 4$.

Method	Duration
Sequential Measurement	90 min 01 s
MESM	12 min 49 s
MESM (after eq. (3.11))	12 min 21 s
Optimized MESM (*new*)	5 min 52 s

Table 3.2: Example of the overall measurement duration for a grid with 40 positions in elevation and 100 positions in azimuth, a sweep length of 1.34 s and assuming the turn table takes 0.3 s to reach its next position.

$\overline{\tau}_{\mathrm{w,MESM}} = 155.1\,\mathrm{ms}$. Taking the time structure of the measured IR in consideration, as suggested in section 3.2.2, can further reduce the measurement time to 7.11 s with $\overline{\tau}_{\mathrm{w,MESM}} = 147.9\,\mathrm{ms}$. Finally, the optimized MESM described in this thesis will bring the measurement time down to 3.22 s.

When comparing the results obtained using all four methods in the frequency-domain, a maximum deviation of ± 0.1 dB over the entire frequency range of interest can be observed. These deviations are within the repeatability variation observed when measuring the same object with the same measurement method after reposition. Hence, the new method—in the same manner as the MESM—does not introduce noticeable errors if the previously introduced requirements are met.

The measurement duration with the newly proposed method can therefore be reduced to 6% of the time required for the sequential method. This factor can be improved when more channels are interleaved. The theoretical limit for the maximum achievable reduction is estimated from

$$\lim_{L\to\infty} \frac{T_{\mathrm{MESM}}(L)}{T_{\mathrm{ES}}(L)} = \lim_{L\to\infty} \frac{(L-1)\tau_{\mathrm{w}} + \tau_{\mathrm{sw}} + \tau_{\mathrm{st}}}{L\,(\tau_{\mathrm{sw}} + \tau_{\mathrm{st}})} = \frac{\tau_{\mathrm{w}}}{\tau_{\mathrm{sw}} + \tau_{\mathrm{st}}}. \tag{3.18}$$

The reduction would then reach 3.6 % for the parameters used in this example. The total measurement time will depend on the total number of azimuth and elevation directions that should be measured and the time it takes for the turntable to move from one to the next azimuth position. The overall measurement duration for a reasonable[6] spatial resolution with 40 positions in elevation and 100 positions in azimuth and assuming the turntable takes 0.3 s to reach its next position is given in table 3.2.

[6] In accordance with the resolution suggested by Zhang et al. (2012).

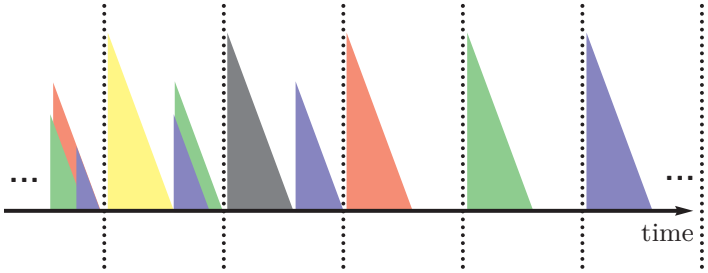


Figure 3.11: Schematic example of an IR measured with the optimized MESM described in this thesis. The dotted vertical lines represent the limits where each signal is cropped.

3.3 Post-Processing

When measuring the HRIR using any of the MESM described in the previous section, the measurement's raw output is a series of overlapped IRs, as illustrated in fig. 3.8(b). It is then necessary to extract each direction's IRs out of the raw IR. Knowing the delay time τ_w used to generate the excitation signal, the raw IR is then cropped in L signals, each has the length τ_w, as exemplified in fig. 3.11. Note that the first dotted line represents the instant $t = 0$ and the harmonic IRs on the left of it will actually appear at the end of the raw IR due to the wrap around effect (see section 2.1).

As described in section 2.3, one of the side effects of the regularized deconvolution is the occurrence of pre-ringing prior to the actual IR. Even though these ripples are constituted only by frequencies outside the desired frequency range, windowing them out might cause more harm than good. Therefore, first a minimum-phase regularized deconvolution is performed guaranteeing that no pre-ringing is introduced by the regularized spectral inversion. After cropping the IRs, each directional IR is further all-pass filtered by $A_{\text{reg,AP}}$ (described in section 2.3) to extract the effects of minimum-phase regularization.

The cropped IRs contain the desired HRIR plus room reflections and harmonic IRs from other channels. These unwanted components can be discarded by time windowing, as shown in fig. 3.12.

The start times of the IRs vary as they depend on the direction of incidence of the sound. Thus, care should be taken as to where to set the time window to prevent the desired part of the IR from extrapolating

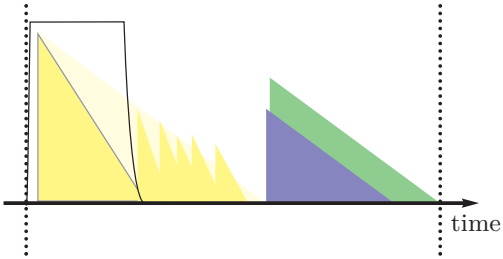


Figure 3.12: Schematic example of how unwanted room reflections and harmonic IRs are windowed out of the cropped IR.

the limits of the window. A straightforward approach is to identify the beginning of each IR using a peak detection algorithm, as for example the one described in the standard *"ISO 3382 – Measurement of room acoustic parameters – Part 1: Performance spaces"*, and to set the time window to start a few samples before the start of the actual IR. Such methods are, however, prone to uncertainty caused by noise, especially in case of contralateral HRIRs, where the SNR is intrinsically low. Ziegelwanger (2012) analyzed this effect and developed a model to robustly estimate the start time of the IRs in each direction of incidence, even for situations where the head was not adequately centered during measurement, as already commented in section 3.1.2. The length of the time window should also be adequately chosen so that all unwanted reflections are discarded.

3.3.1 Equalization

The HRIR obtained at this point is free of the influence of measurement artifact. However, it still contains the influence of the loudspeaker's and microphone's frequency response. The most common method to eliminate these influences is the *free-field equalization*, where each HRTF H is divided by the reference transfer function H_{ref}, defined as the transfer function between the same sound source placed at the same position of the HRTF measurement and the same microphone used in the measurement placed at the point corresponding to the center of the head while the subject is not present. This equalization results in the *free-field HRTF* defined by Blauert (1997) as

$$H_{\text{free-field}}(\theta, \phi, r, f) = \frac{H(\theta, \phi, r, f)}{H_{\text{ref}}(r, f)}. \tag{3.19}$$

There are, however, some practical aspects that should be observed while calculating the free-field HRTF. Like the HRIR, the reference transfer function should also have the room reflections windowed out, otherwise undesirable artifacts will occur since reflection paths are not identical in both situations and do not cancel out. Furthermore, care must be taken as the procedure described above will render the resulting HRTF noncausal for the ipsilateral directions,[7] which can have disastrous consequences for the naïve signal processing. Thus, the IRs are cropped, keeping only the range where the window was applied and the time t^i_{win} where the window starts. The cropped IR segments are then divided. As both segments contain approximately the same (small) delay, the resulting IR is not expected to be noncausal.

At this point, the low-frequency asymptote correction is applied. Hammershøi and Møller (2005) argues that at low frequencies the human head ceases to act as a scattering object for the incident sound wave, so that the ratio between the transfer function to the ears and to the reference microphone tends to 1. Acoustic data acquisition software usually disregards frequencies close to 0 Hz, nevertheless, Hammershøi and Møller (2005) show that when the interpolation is conducted by simply padding the HRIR with trailing zeros, an erroneous value at 0 Hz can have a strong influence up to the mid-frequency range. The correction is applied by simply substituting the value corresponding to 0 Hz by 1.

The delay removed prior to cropping must be reinstated. Therefore, the signal should be padded with zeros to provide a proper time shift. To avoid a sharp transition between the IR and the padded zeros, a fade-in/fade-out operation should be applied to the limits of the cropped IR. After the fading, zeros are padded to the cropped signal and subsequently, the signal is shifted by $t^i_{\text{win}} - t^{\text{ref}}_{\text{win}}$, resulting in the free-field HRTF, or respectively, free-field HRIR.

To apply the range extrapolation, a last step is required. The free-field HRTF should be multiplied by the transfer function between a point source placed at the acoustic center of the loudspeaker to an ideal receiver placed at the acoustic center of the reference microphone. The distance between these two ideal transducers can be estimated from the measured reference transfer function.

Furthermore, if the directional transfer function (DTF) is desired, the diffuse-field HRTF should be estimated from the available free-field

[7] The ipsilateral HRIRs are noncausal because the acoustic path from the source to the ipsilateral ear is shorter than the path from the source to the reference microphone.

HRTF at directions $\boldsymbol{\theta}_k$ by

$$H_{\text{diff}}(f) = \sqrt{\sum_k w_k |H(\boldsymbol{\theta}_k, f)|^2}\,, \tag{3.20}$$

where the weights w_k depend on the sampling grid that is used (Driscoll and Healy, 1994). According to Middlebrooks (1999a), the DTF is afterwards obtained by dividing the free-field HRTFs by the minimum-phase spectrum of $H_{\text{diff}}(f)$.

In an optional post-processing step the measured data can be smoothed in frequency-domain. It was shown that by smoothing the HRTFs (or DTFs) up to a certain degree, the localization accuracy will not deteriorate (Kulkarni and Colburn, 1998; Breebaart and Kohlrausch, 2001; Xie and Zhang, 2010). A spatial interpolation might also introduce a *spatial smoothing*. To the best of the author's knowledge, the psycho-acoustical effect of the HRTF spatial smoothing has not been studied yet.

3.3.2 Interpolation

As discussed in section 3.1.4, the HRTFs are measured at discrete points distributed on a spherical surface centered on the listener's head. Virtual reality applications require a smooth directional transition when synthesizing moving sources. To directly switch between neighboring HRTFs without any audible artifact, the angular distance between these HRTFs should be smaller than the minimum audible angular difference perceived by the human auditory system—depending on signal type and direction, as low as 1° (Blauert, 1997)—a very dense sampling grid would be necessary. This would require a more complex measurement setup and large data storage. Another problem that could occur is that the acoustic simulation software that is used to process the HRTFs cannot handle the sampling grid used for the measurement. A solution for both situations is to interpolate the missing HRTFs from the original data set.

Pollow et al. (2012a) divide the HRTF interpolation methods into two categories. Local interpolation methods use only the immediate neighboring HRTFs for calculations (Langendijk and Bronkhorst, 2000; Freeland et al., 2007; Lentz, 2007) while global interpolation methods use the entire HRTF set for the interpolation (Kistler and Wightman, 1992; Evans et al., 1998; Duraiswami et al. , 2004; Wang et al., 2009). The first category has the advantage that its

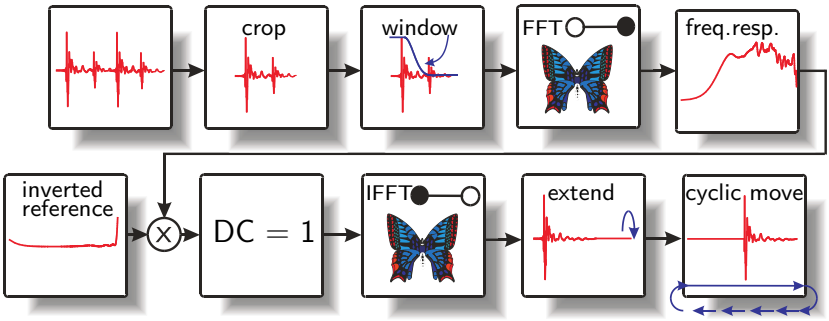


Figure 3.13: Diagram describing the post-processing stages applied to obtain a free-field equalized HRTF from the raw measurement.

calculation can be conducted in a very fast manner. However, HRTFs interpolated with the latter group tend to have a better agreement with measured data. Global interpolation methods also perform a spatial smoothing on the processed data, thus providing robustness against additive noise up to a certain level.

The two main mathematical tools used for a global interpolation in the sphere are the *principal component analysis* (PCA) and the *spherical harmonic* (SH) approximation. Evans et al. (1998) concluded that the latter method provided better interpolation results than the former technique. Moreover, Duraiswami et al. (2004) showed that using the concept of spherical holography, which in turn is based on the spherical harmonic decomposition, one can not only interpolate the HRTF over a spherical surface, but can also extrapolate its radial dependence, an advantage that no other interpolation technique can offer.

The spherical harmonics define a set of orthonormal basis over the spherical surface. They are defined (for the coordinate system defined in fig. 2.5) as

$$Y_n^m(\phi,\theta) \equiv \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} \cdot P_n^m(\sin\theta)e^{jm\phi}, \tag{3.21}$$

where P_n^m are the Legendre functions of order n and degree m. Any arbitrary function $p(\phi,\theta)$ defined on a sphere can then be expanded as

$$p(\phi,\theta) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} c_{nm}Y_n^m(\phi,\theta), \tag{3.22}$$

where c_{nm} are complex spherical expansion coefficients (Williams, 1999). These coefficients can be obtained from

$$c_{nm} = \oint_S p(\phi, \theta) Y_n^m(\phi, \theta)^* \, dS. \tag{3.23}$$

Like the continuous Fourier transform and its discrete counterpart, the DFT, spherical harmonics can also be sampled at a grid of points and still correctly describe the whole space, provided that a suitable sampling grid was used and that the data is *spatially* band limited (Zotter, 2009, ch. 4). In this case, eq. (3.23) can no longer be applied. The constant c_{nm} must be estimated from a system of linear equations composed of the equation eq. (3.22) evaluated at every sampled direction. This can be cast in a matrix form as

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix} = \begin{bmatrix} Y_1(\phi_1, \theta_1) & \cdots & Y_1(\phi_k, \theta_k) \\ Y_2(\phi_1, \theta_1) & \cdots & Y_2(\phi_k, \theta_k) \\ Y_3(\phi_1, \theta_1) & \cdots & Y_3(\phi_k, \theta_k) \\ Y_4(\phi_1, \theta_1) & \cdots & Y_4(\phi_k, \theta_k) \\ \vdots & \ddots & \vdots \\ Y_l(\phi_1, \theta_1) & \cdots & Y_l(\phi_k, \theta_k) \end{bmatrix}^T \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_l \end{bmatrix}, \tag{3.24}$$

where l is the linear SH-index defined as $l = n^2+n+m+1$. Equation (3.24) can be written in compact form as $\boldsymbol{p} = \boldsymbol{Y}\boldsymbol{c}$.

Spherical harmonic representation is defined as an infinite summation of spherical basis functions. However, as written in eq. (3.22), in practice, the order is truncated at a maximum value O instead. Zhang et al. (2012) argues that spherical harmonics up to order $O = 46$ are required to correctly describe the spatial variation of an HRTF at 20 kHz. There are several sampling strategies on the sphere that allow more or less efficient conversions of the sampled spatial data into SH-domain for further calculations (Zotter, 2009). According to him, the most efficient method, the hyperinterpolation, requires $(O + 1)^2$ sampling points. For $O = 46$ this equals 2209 sampling points. Unfortunately, the hyperinterpolation is not an axisymmetric sampling scheme. The axisymmetric Gaussian grid requires $2(O + 1)^2$ sampling points, i.e., 4418 samples to describe the HRTFs correctly in every direction for the entire hearing range. Other axisymmetric sampling grids could reduce this number, as for example the IGLOO grid that would require only 2304 points according to Zhang et al. (2012).[8]

[8] As the IGLOO grid requires more azimuth positions with less elevation positions per azimuth the reduction in measurement points will not necessarily lead to a reduction in measurement time when using the optimized MESM.

HRTFs are interpolated by first defining the spherical expansion vector $\boldsymbol{c}$ from the measured pressure values $\boldsymbol{p}$ from the minimization problem

$$\min_{\boldsymbol{c}} \|\boldsymbol{p} - \boldsymbol{Y}\boldsymbol{c}\|_2^2. \tag{3.25}$$

Zotter (2010) shows that for special sampling grids, e.g. the Gaussian or the hyperinterpolation grids, other solutions with more efficient numerical properties exist. A generalized solution to this minimization problem can be obtained from

$$\boldsymbol{c} = \boldsymbol{Y}^{+}\boldsymbol{p}. \tag{3.26}$$

The proceeding is completed by calculating the pressure values at the new sampling points using eq. (3.22). This operation can be conducted for $\boldsymbol{p}$ described in time or in frequency-domain. It is however more intuitive to conduct this operations in frequency-domain, as frequency-dependent order truncation can be applied (cf. Pollow et al., 2012a).

It is important to remember that the spherical harmonics, and thus spherical holography, are defined only for a closed spherical surface. But the designed arc cannot provide measurements on the lower spherical cap (section 3.1.2). The missing points result in an ill-posed matrix of spherical harmonic basis functions $\boldsymbol{Y}$. Regularization should then be used to obtain a stable solution that approximates the solution to the inverse problem (Pollow et al., 2012a). Ruffini et al. (2002) describes a regularization approach based on minimizing the *surface curvature* while matching the surface to the available data. This results in smoothly interpolated values in the lower cap region where measurements were not available. The solution to this minimization problem is given by

$$\boldsymbol{c} = \left(\boldsymbol{Y}^{*}\boldsymbol{Y} + \overline{\boldsymbol{P}}\boldsymbol{B} + \mu\boldsymbol{P}\boldsymbol{B}\right)^{-1}\boldsymbol{Y}^{*}\boldsymbol{p}, \tag{3.27}$$

where $\boldsymbol{P} = \boldsymbol{Y}^{*}(\boldsymbol{Y}\boldsymbol{Y}^{*})^{-1}\boldsymbol{Y}$, $\overline{\boldsymbol{P}} = \boldsymbol{I} - \boldsymbol{P}$, μ is a regularization parameter and $\boldsymbol{B} = \mathrm{diag}(n(n+1))$, being n the order of the corresponding SH coefficient. A slightly altered version of this regularization scheme was also used by Duraiswami et al. (2004).

3.3.3 Range Extrapolation

The last step to obtain a continuous representation of the HRTFs in space is to describe its dependency on the radial distance. Lentz (2007) proposed an interpolation scheme for near-field HRTFs. His method requires, however, the measurement of a complete set of HRTFs for several distinct radial distances in the near-field.

Using the principle of reciprocity, Duraiswami et al. (2004) argued that the HRTF can be "characterized as a solution of a scattering problem". According to them, a point source placed at the entrance of the ear canal will generate pressure $p(\phi, \theta, r, f)$ at a given position $\mathbf{x}$ equivalent to the pressure that would exist at the ear if the point source was placed at $\mathbf{x}$. The pressure field at position (ϕ, θ, r) and for the wave number $k = 2\pi f/c$ (where c is the speed of sound) can be represented as

$$p(\phi, \theta, r, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \xi_{nm}(k) h_n(kr) Y_n^m(\phi, \theta), \tag{3.28}$$

where ξ_{nm} is the potential expansion coefficient, h_n the spherical Hankel functions of order n and Y_n^m the already defined spherical harmonics (Williams, 1999).

The potential expansion vector $\boldsymbol{\xi}$ can be estimated from the pressure vector $\boldsymbol{p}$ by first obtaining the spherical expansion vector $\boldsymbol{c}$, as described in the previous section, and then applying the corresponding spherical Hankel function to each element of $\boldsymbol{c}$, as follows:

$$\xi_{nm}(k) = \frac{c_{nm}(kr)}{h_n(kr)}. \tag{3.29}$$

Thus, for any given frequency f, the sound pressure field is entirely determined by the potential expansion vector $\boldsymbol{\xi}(z)$ (cf. Pollow et al., 2012a). Because the moduli from Hankel functions behave approximately as an exponentially decaying curve, problems caused by division by zero will not occur. However, the exponential growth of the spherical Hankel functions for higher orders and small arguments kr can lead to noise amplification. This effect is commonly deal with using a frequency-dependent order truncation.

As discussed in Duraiswami et al. (2004), a sufficient spatial resolution is required to capture the pressure field and the required spatial resolution is proportional to the frequency. Moreover, this method will only work if all sources are contained within a spherical surface S of a small radius and the desired interpolated/extrapolated points lie outside of S. In this case, the HRTF is obtained by applying eq. (3.28) to the new desired position.

Pollow et al. (2012a) compared the range extrapolation technique with the near-field measurements conducted by Lentz (2007) and was able to verify that this technique produced extrapolated HRTFs that matched the measured HRTFs.

3.4 Results

This section begins with a comparison of the HRTF measurement setup presented in this chapter with a previously constructed sparse-type setup. The measurements for this first comparison were made with an artificial head. As this setup was designed for the measurement of individual HRTFs, the section then concludes showing the HRTFs and HRIRs of one of the 16 individuals that have so far been measured with the new system.

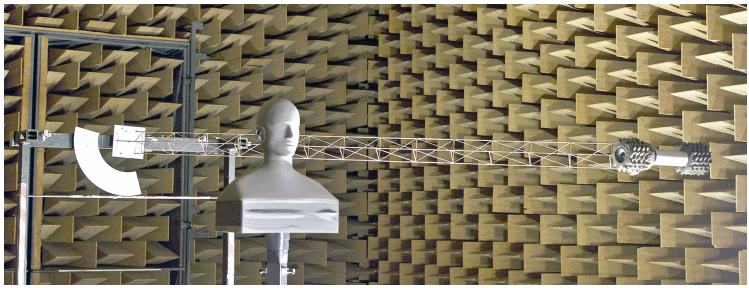
3.4.1 System Comparison

As a proof of concept the HRTFs of an artificial head were initially measured. A measurement with an artificial head offers the obvious advantage that the subject under test does not move itself during measurement and can be precisely positioned. On top of that, a comprehensive HRTF dataset of the same artificial head had already been acquired with the measurement system described by Aretz (2012) and depictured in fig. 3.14(a). A Gaussian sampling grid of order 70 with 9800 points was used, which approximately corresponds to an angular resolution of 2°. This measurement was reported to have taken around four hours to complete using sequential exponential sweeps of 16384 samples played back at a sampling rate of 44.1 kHz.[9] The measurement was conducted at a distance of 1.75 m, different than the 1 m used with the new system. Therefore, all measured HRIRs had its phase and amplitude corrected using the Green's function to the distance of 1 m, under the assumption that the HRTFs are already in far-field (Brungart and Rabinowitz, 1999).

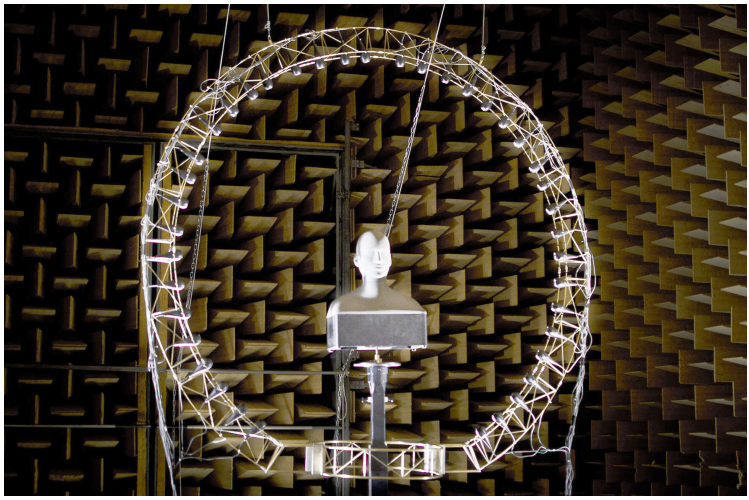
The measurement with the new system was made using a Gaussian sampling grid of order 48, however with the lowest eight elevation points missing, which results in 3840 measurement points (cf. fig. 3.14(b)). The interleaved sweeps had a length of 59094 samples, covering the frequency range from 200 Hz to 20 kHz at a sampling rate of 44.1 kHz. The waiting time between sweeps was 35 ms. The total measurement time was just short of six minutes.

The spatial, temporal and frequency characteristic of the data obtained with both systems is compared. Figures 3.15 to 3.18 display at the top balloon plots of the measured data and at the bottom the same data

[9] As this system can only measure HRTFs at one hemisphere, the measurement had to be conducted in two stages, turning the head upside-down in the second stage, what is not viable for individual HRTF measurements.



(a)



(b)

Figure 3.14: Artificial head being measured with (a) the HRTF measurement setup previously developed at the Institute of Technical Acoustics (RWTH Aachen), composed of a single loudspeaker placed at a rotating arm, and (b) the individual HRTF measurement setup presented in this thesis, with a supporting arc and 40 drop-like loudspeakers.

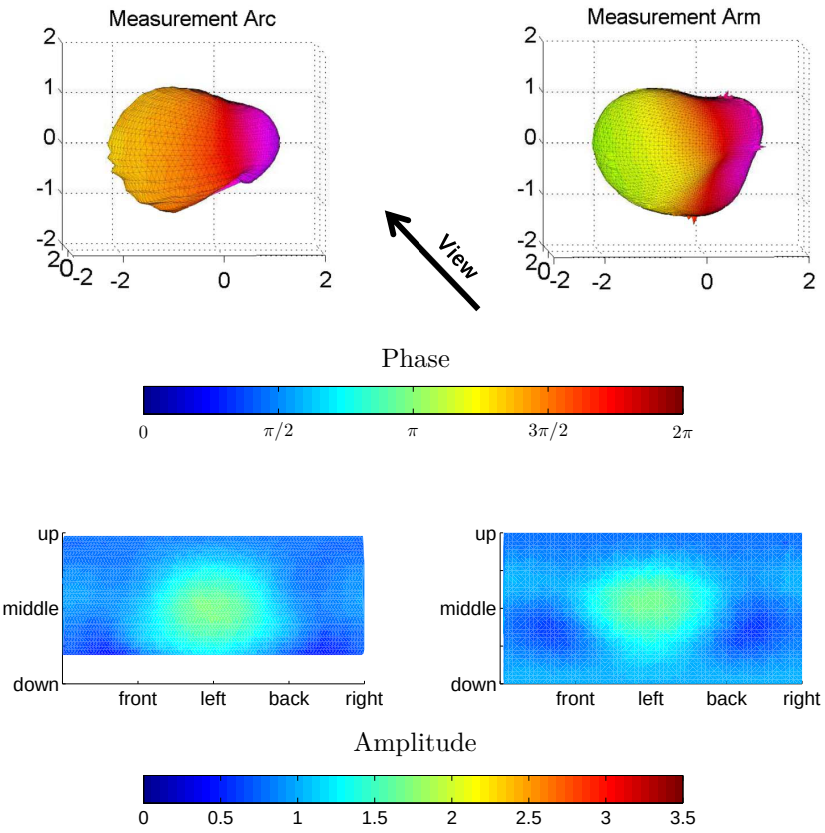


Figure 3.15: Comparison of HRTF measurement setups. The same artificial head was measured with both the measurement arc described in this chapter (left) and with the vintage measurement arm described in (Lentz, 2007) (right). The two top figures are balloon plots, where the balloon's radius represents the amplitude of the HRTF and the color its phase. The arrow indicates the head's view direction. The two bottom plots show the amplitude values of both measurements plotted in an exploded view. Plots are made for the frequency of 500 Hz.

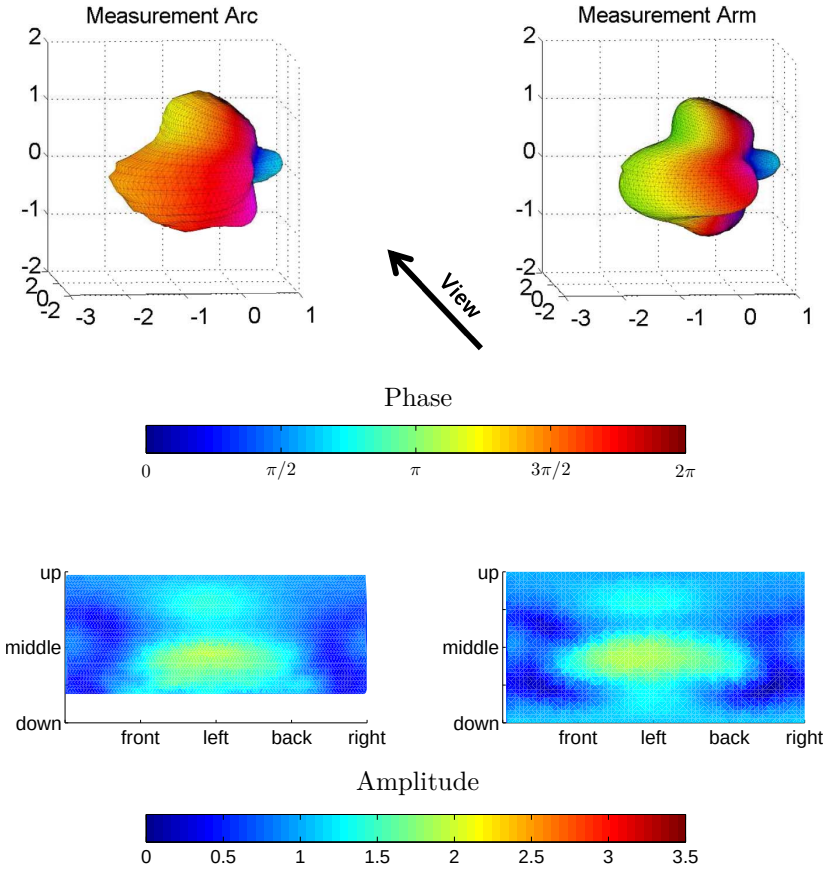


Figure 3.16: Comparison of HRTF measurement setups. The same artificial head was measured with both the measurement arc described in this chapter (left) and with the vintage measurement arm described in (Lentz, 2007) (right). The two top figures are balloon plots, where the balloon's radius represents the amplitude of the HRTF and the color its phase. The arrow indicates the head's view direction. The two bottom plots show the amplitude values of both measurements plotted in an exploded view. Plots are made for the frequency of 1000 Hz.

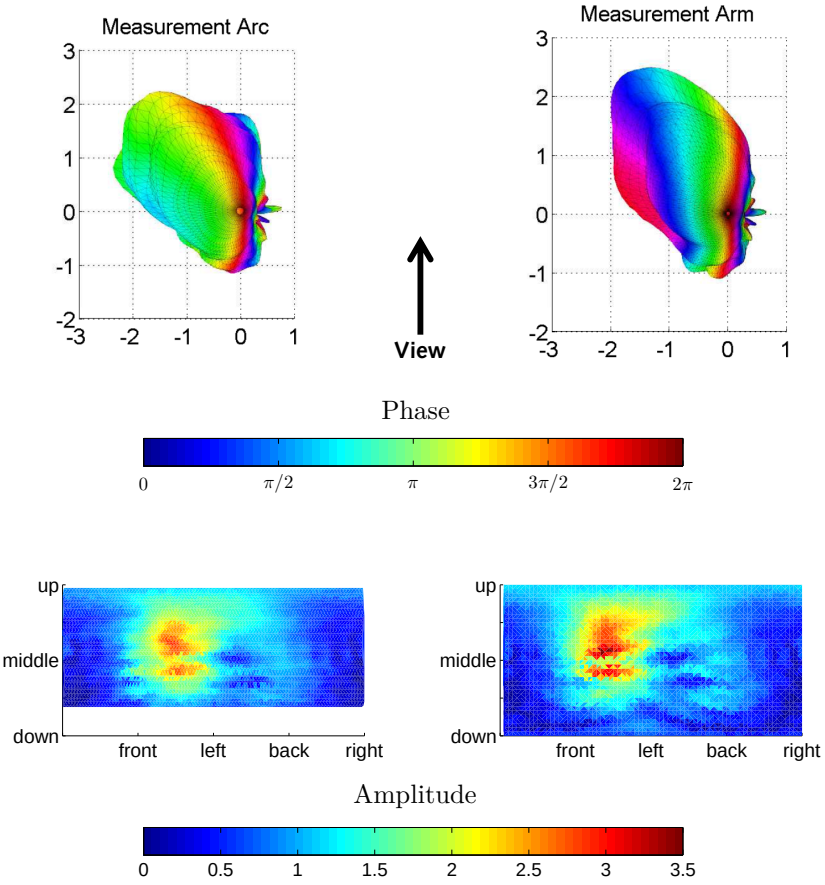


Figure 3.17: Comparison of HRTF measurement setups. The same artificial head was measured with both the measurement arc described in this chapter (left) and with the vintage measurement arm described in (Lentz, 2007) (right). The two top figures are balloon plots, where the balloon's radius represents the amplitude of the HRTF and the color its phase. The arrow indicates the head's view direction. The two bottom plots show the amplitude values of both measurements plotted in an exploded view. Plots are made for the frequency of 4000 Hz.

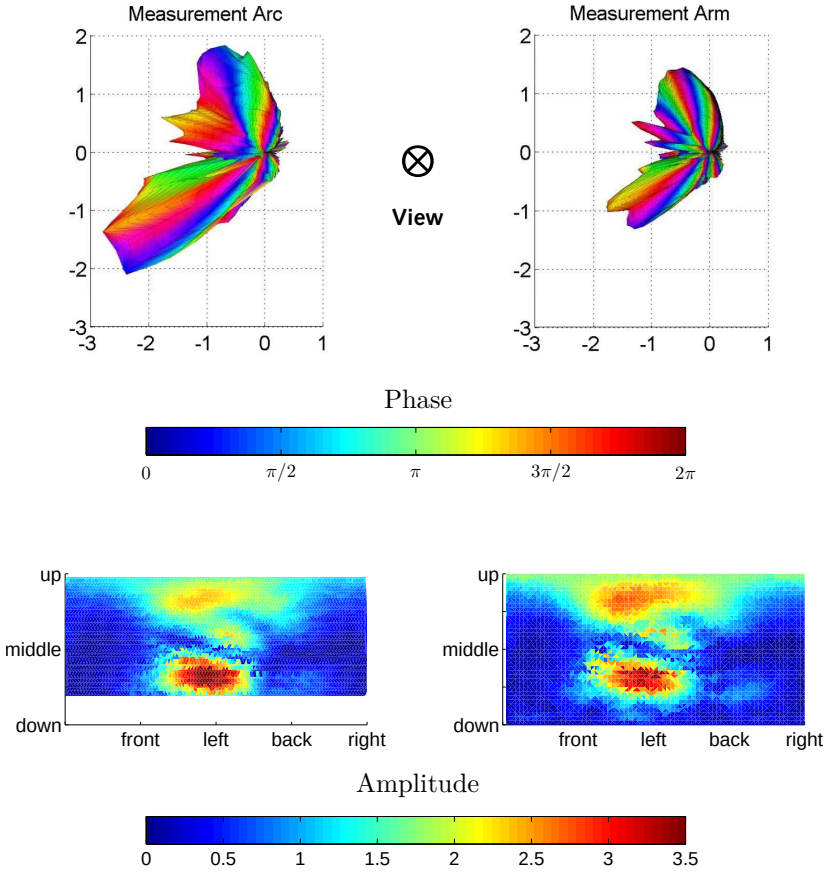


Figure 3.18: Comparison of HRTF measurement setups. The same artificial head was measured with both the measurement arc described in this chapter (left) and with the vintage measurement arm described in (Lentz, 2007) (right). The two top figures are balloon plots, where the balloon's radius represents the amplitude of the HRTF and the color its phase. The arrow indicates the head's view direction. The two bottom plots show the amplitude values of both measurements plotted in an exploded view. Plots are made for the frequency of 8000 Hz.

projected at a 2D surface. The radius of the balloon plot represents the amplitude of the HRTF and the color its phase. The arrow in between the plots shows the artificial head's view direction.

The plots at 500 Hz (fig. 3.15) make it evident the the lack of measurement data at the lower cap. To directly extract these missing values from the measured data has high uncertainty involved. For the measured region, the plots do show good similarity.

For 1 kHz, the lack of the lower cap is not as evident as for 500 Hz as now the amplitude values at this region have decreased. The edge seen in the equator of the measurement with the arm occurs because of small positioning errors as the measurement had to be done in two separate stages, each measuring one hemisphere.

The plots at 4 kHz and 8 kHz show again a good overall agreement of the data. Difference in amplitude (radius of the balloon) are observed. This are, however, compensated for when using DTFs instead of HRTFs (cf fig. 3.19). At higher frequencies it is also possible to see how the energy at the side of the contralateral ear is considerably lower than for the ipsilateral side.

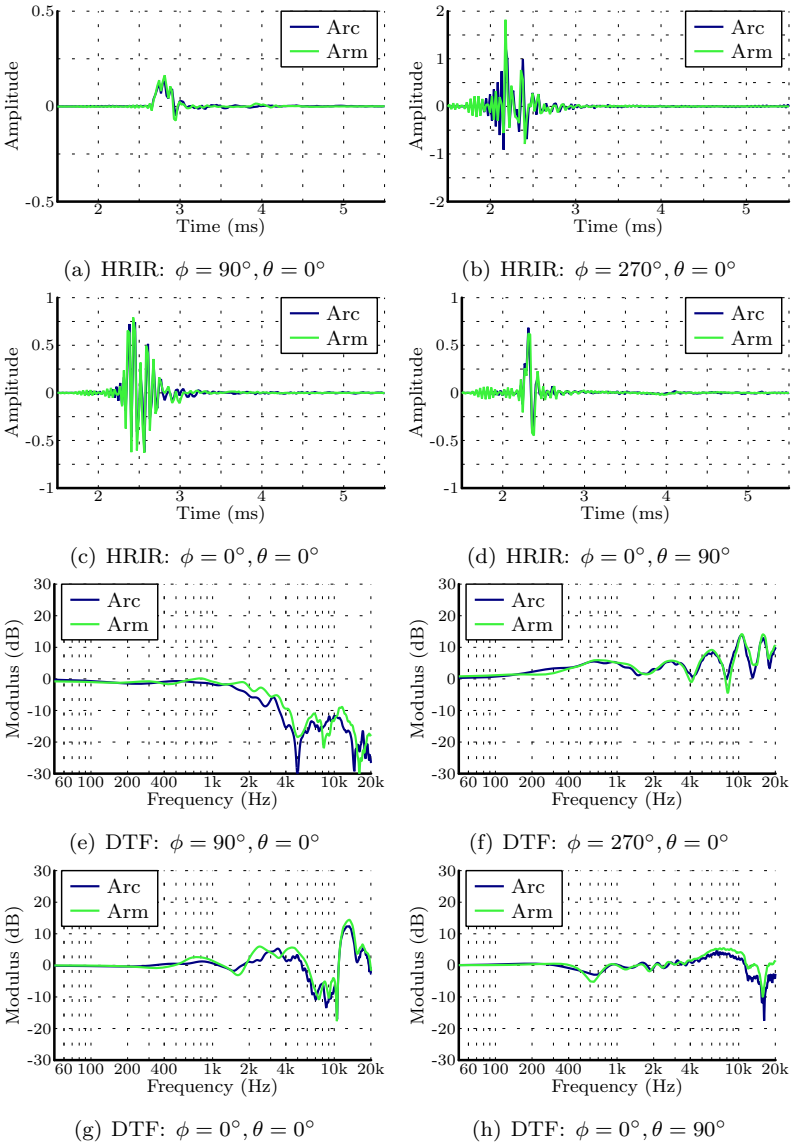
The phase information encodes the distance information. As the ears are shifted in relation to the center of the head, a phase variation is observed. The different phase behavior seen at these plots is caused by the fact that, at each measurement, the artificial head was not identically positioned in relation to the systems' center. The main spatial features are, however, similar throughout the whole measured frequency range when considering only the amplitude values.

Also the frequency- and time-domain characteristics of the HRTFs measured with both systems were compared and showed good agreement. Examples for four directions are shown in fig. 3.19.

In the frequency-domain, only little deviation in the higher frequencies is clearly noticeable. In the time-domain, the main difference is observed in the pre-ringing, caused by the regularized deconvolution of the signals. These artifacts can be windowed out without compromising the quality of the HRIRs.

3.4.2 Individual Measurement

This system was designed specially for the measurement of individual HRTFs. So far, 16 listeners have been measured with he system. However,



(a) HRIR: $\phi = 90°, \theta = 0°$

(b) HRIR: $\phi = 270°, \theta = 0°$

(c) HRIR: $\phi = 0°, \theta = 0°$

(d) HRIR: $\phi = 0°, \theta = 90°$

(e) DTF: $\phi = 90°, \theta = 0°$

(f) DTF: $\phi = 270°, \theta = 0°$

(g) DTF: $\phi = 0°, \theta = 0°$

(h) DTF: $\phi = 0°, \theta = 90°$

Figure 3.19: Comparison of HRTF measurement setups. The same artificial head was measured with both the measurement arc described in this chapter (Arc) and with the vintage measurement arm described in (Lentz, 2007) (Arm). HRIRs of four different exemplary directions (a-d) and their equivalent DTFs (e-h) show reduced variability between the measurement setups.

no localization test could be conducted to evaluate the perceptual quality of the acquired HRTFs.

The data from one of these listeners is shown below. The excitation signal was the same used for the measurement of the artificial head, described in the previous section. The spectrogram of the raw signal acquired at the left ear for one azimuthal position is shown in fig. 3.20. The 40 interleaved sweeps can be clearly seen. The last 20 sweeps have lower amplitude at higher frequencies, as could be expected once they originated from the contralateral side.

The deconvolved impulse responses can be seen in fig. 3.21. The variation in amplitude for the ipsi- and contralateral sides can be clearly seen. The small impulses present at the end of the signal are the harmonic impulse response of the first measured directions. The SNR, or better said, the peak-to-noise ratio is of approximately 80 dB when the source is directly in front of the ear and decreases to approximately 50 dB for the contralateral side. This obtained SNR is expected to be sufficient to provide high quality binaural synthesis.

The time and frequency-domain response for three positions in the median plane and one position at extreme lateral angle are shown in fig. 3.22, respectively.

3.5 Discussion

HRTF measurements were traditionally conducted in far-field, restricting the auralization to distant sources. A series of new measurements at shorter distance is required to simulate near-field effects. To avoid the need for extra measurements, the range extrapolation technique is used. It provides a spatially continuous representation of the HRTFs by using a reciprocal formulation of the modal components of an outgoing spherical wave. This results in a setup-independent and compact description of individual HRTFs, allowing the evaluation of any binaural transfer functions at any point in near- or far-field, though with some limitations due to noise and numerical instability.

The HRTF measurement setup described in this chapter was designed to meet the requirements of the reciprocal acoustic holography. This method also assumes the excitation source to be an acoustic point source. Therefore, the loudspeakers used in this setup were designed to have (approximately) an omnidirectional directivity in the entire range of

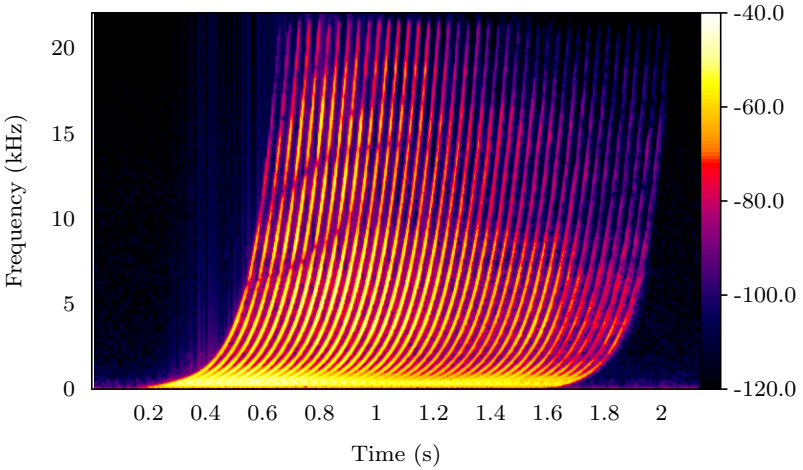


Figure 3.20: Spectrogram from the multiple exponential sweep signal acquired at the left ear of a subject for one azimuthal measurement position. The color scale is given in decibels relative to 1.

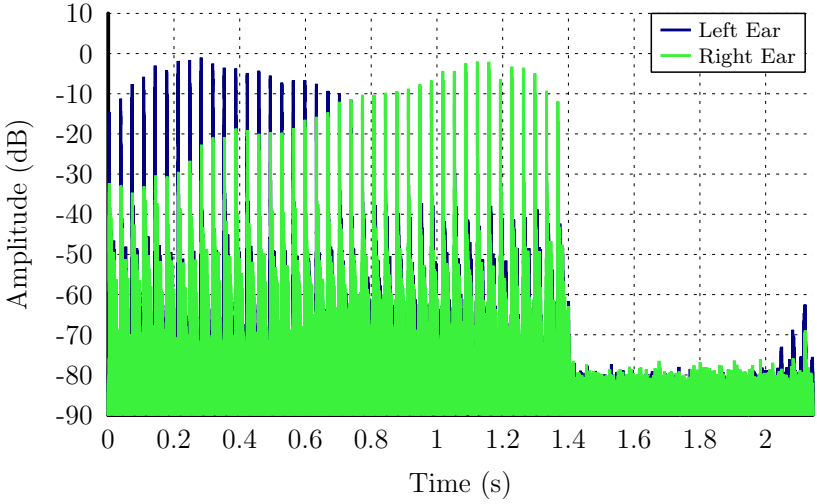
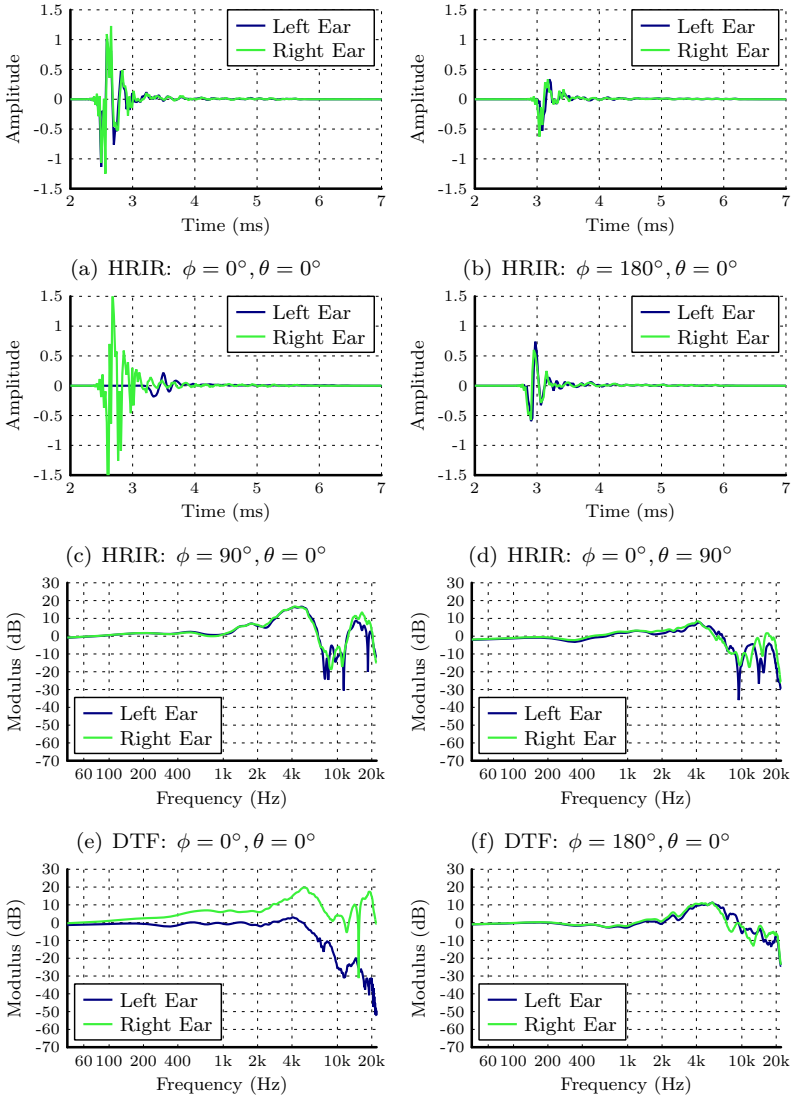


Figure 3.21: The deconvolved impulse responses obtained from the signal depictured in fig. 3.20. The peak-to-noise ratio is of approximately 80 dB for sources at the ipsilateral side and as low as 50 dB for sources at the contralateral side.



(a) HRIR: $\phi = 0°, \theta = 0°$

(b) HRIR: $\phi = 180°, \theta = 0°$

(c) HRIR: $\phi = 90°, \theta = 0°$

(d) HRIR: $\phi = 0°, \theta = 90°$

(e) DTF: $\phi = 0°, \theta = 0°$

(f) DTF: $\phi = 180°, \theta = 0°$

(g) DTF: $\phi = 90°, \theta = 0°$

(h) DTF: $\phi = 0°, \theta = 90°$

Figure 3.22: Individual HRIRs (a-d) and their equivalent individual DTF (e-h), obtained using the new HRTF measurement system presented in this chapter.

application. A small broadband loudspeaker chassis with reduced nonlinear behavior was chosen. During measurements it was verified that this chassis presented time variance which could not be compensated for and influenced the resulting HRTFs. Thus, it is recommended for new similar setups to choose loudspeakers that are not only omnidirectional, but that are also time invariant, i.e., their sensitivity does not change during the operation. Loudspeakers can, however, have a nonlinear behavior as this effect can be compensated for by performing the exponential sweep measurement.

According to acoustic holography theory, it is advisable to conduct measurements as close as possible to the scattering objects, i.e. the head and torso, and extrapolate the HRTF outwards. On that account the radius of the measurement arc was chosen to be 1 m. Results published at a later stage by Pollow et al. (2012a) did not show any difference between inward and outward extrapolation of HRTF measurements. They were, however, able to show that the acoustic holography method is much more exact than other published methods for the HRTF near-field compensation.

Acoustic holography assumes that all scattering objects are contained inside a spherical surface S of small radius and that all sound sources are located outside S. For this reason, the supporting arc was constructed using a thin metal rod truss structure, which is supposed to be acoustically transparent, and the form of the loudspeakers was chosen to minimize reflections. During measurements, though, it was observed that unwanted reflections from neighboring loudspeakers and structure-borne sound from the arc were still present. Therefore, the design of the arc and the loudspeakers should be reviewed. The loudspeakers should be integrated into a rigid arc. Moreover, it was also verified that the mechanical stability of the chosen trellis structure was not sufficient.

The new setup was, however, not only used for adequate range extrapolation measurements. The other purpose of this new setup was to measure a sufficient number of points on the sphere to adequately describe the HRTF in the *shortest time* possible. This objective was achieved by optimizing the excitation signal used for the measurement. Majdak et al. (2007) originally proposed to overlap the exponential sweep signals in order to reduce measurement time without compromising the obtained SNR. This thesis extends their multiple exponential sweep method (MESM) in two ways: 1) by relaxing the overlapping requirements and 2) by making better use of the HRIR's temporal structure. The original requirement for signal overlapping was that all harmonic IRs should decay below background noise level before the previous desired IR occurs.

This requirement was relaxed so that no harmonic IRs shall fall into a predefined avoid zone containing the desired IR. Observing the temporal structure of an HRIR measured in a typical (hemi-)anechoic chamber, one notes that only the initial portion of the IR describes the HRIR, being the rest of the signal composed by reflections due to objects in the room or to room boundaries. As these unwanted reflections must be windowed out, these regions can be used as place-holders for harmonic IRs of other directions. These reflections are responsible for a longer reverberation time of the chamber, which in turn limits the waiting time between subsequent sweeps. The size of the hemi-anechoic chamber used with this setup was *reduced* by building a wall of absorbing material near the arc, thus reducing the size of the chamber's IR and consequently speeding up the measurement. Altogether, these optimization allowed the HRTF measurement in 3840 directions in less than six minutes,[10] which would take approximately 12 min with the original MESM and over 1.5 h using the sequential method.

The post-processing of the raw IR is executed in accordance with the latest research found in the literature, as e.g. in Hammershøi and Møller (2005). One should only keep in mind that the HRTFs were no longer measured under far-field assumption and therefore do not allow the distance effect to be compensated for simply by the Green's function, as was the case with the free-field HRTF. To serve as input data for spherical holography, the HRTFs should be multiplied by the transfer function of a point source placed at the acoustic center of the corresponding loudspeaker to a point receiver placed at the origin of the head coordinate system.

Results of measurements showed the presence of a spatial ripple in the resulting HRTFs. The effect of these ripples when calculating the spherical wave spectrum are negligible in amplitude, but have a strong effect in the phase, as described in (Krechel, 2012). These ripples are caused either by reflection artifacts still present in the measurement or by the fact that the subject's rotation axis, when placed over the turntable, cannot be precisely aligned with the arc's symmetry axis. In the first case, ripples are eliminated by adequate time windowing. In the second case, the proposed way to mitigate this effect is to place the reference microphone at the center of the turntable and conduct a regular measurement rotating it, just like the subjects are, and using the transfer function obtained for each direction to equalize the HRTF of the correspondent azimuthal direction.

[10] Being the fastest HRTF measurement setup known to the author.

To find the spherical expansion vector required for interpolation, it is extremely important to know the position of the loudspeakers as precisely as possible. Therefore, a calibration system was developed using only the turntable and two microphones placed at a known distance from each other. This system gives not only the exact position of the loudspeakers, but also the orientation of the bar holding the microphones, the distance to the rotation axis, the latency of the sound card and the speed of sound at the time of measurement (Krechel, 2012).

A method to accomplish interpolation and range extrapolation of the HRTF based on acoustic spherical holography was described. This method assumes the HRTF has a continuous and smooth distribution in space, described by a finite number of spherical harmonics. Ideally, the number of available sampling points should be sufficient to unambiguously represent all needed spherical harmonic basis functions. This restriction is hardly achievable in practice and a small amount of spatial aliasing is to be expected. Zhang et al. (2012) argues that spherical harmonics up to order $O = 46$ are required to correctly describe the spatial variation of an HRTF at 20 kHz. Pollow et al. (2012c) show that displacing the origin of the spherical coordinate system to the acoustic center of reciprocal HRTF pressure field allows a more compact description of the HRTF with fewer spherical coefficients (therefore, with a lower maximum order). Consequently, fewer measurement points are also required. A potential vector $\boldsymbol{\xi}$ for the head coordinate system can be obtained from the potential vector $\boldsymbol{\xi}'$, with origin placed at the entrance of the ear canal, by the translation operation in the spherical coordinates, thoroughly described in (Zotter, 2009, pp. 36-50).

To avoid the effects of spatial aliasing—and also improve the overall measurement quality—efforts have been made to develop a method that can measure the HRTF on a continuous surface or at least along a circle (Ajdler et al., 2007; Fukudome et al., 2007; Enzner, 2009). These techniques are, however, not as robust to nonlinearity as the correlation measurement technique using the exponential sweep described in section 3.2. Krechel (2012) described an approach to dynamically acquire the HRTF along a circle using the correlation technique, allowing a further considerable speedup in comparison to the sequential measurement technique. In this method, the listener is continuously turned while the excitation signals are being played. Even though this continuous movement breaks the assumption of time invariance implicit to the correlation technique, it is plausible to assume that the system is “almost” invariant at each small time interval, while one frequency is

been played. A post-processing step is then necessary to compensate for this continuous movement.

The recently developed *Compressive Sampling* theory was also studied as a way to reduce the number of required sampling points and to avoid spatial aliasing (Masiero and Pollow, 2010). Compressive sampling proposes a new framework on how to effectively sample information with a reduced number of sensors. The main idea behind this concept is that if the information to be sampled can be sparsely described in a space that is incoherent to the measurement space, then this information can be restored by ℓ_1 minimization. Unfortunately, compressive sampling could not be applied to the HRTFs using the spherical harmonic basis functions, as especially high frequency HRTFs cannot be considered sparse in the SH-domain. A set of basis functions extracted from the "principal component analysis" of a group of individual HRTFs, similar to the basis described in (Kistler and Wightman, 1992), but spanning the whole sphere, might be a good candidate for an incoherent representation domain. Other possible basis would be spherical wavelets (Freeden and Windheuser, 1997) or the Slepian functions (Slepian, 1964).

4

Binaural Reproduction using Headphones

The reproduction of binaural signals via headphones is straightforward as headphones are able to deliver each binaural channel independently to each ear. The headphone reproduction does add spectral coloration to the reproduced sound, but at first glance it seems as if this effect can easily be mitigated by an equalization filter. There are, however, some difficulties involved in the design of such a filter, which is obtained from the inverse of the *headphone transfer function* (HpTF). First, the HpTF varies for each listener. Therefore headphone equalization filters must be shaped individually. Second, at high frequencies the HpTF is strongly dependent on the headphone fitting and therefore the equalization filter should be robust to (small) fitting variations. Third, HpTF are usually not minimum-phase, i.e. they contain all-pass components that when inverted result in a noncausal equalization filter.

The HpTF and HRTF are commonly measured with a microphone placed at the entrance of the ear canal. However, a correct binaural reproduction occurs when the sound pressure at the listener's eardrums is ideally matched. Møller (1992) proposed a measurement technique to verify whether a given set of headphones is able to provide an authentic binaural reproduction. This technique is applied in this chapter to verify the adequacy of the used headphones.

Measured HpTFs were evaluated with regard to the inter-subject variability and intra-subject variability to the headphone fitting. For frequencies up to 4 kHz a low variability was observed in both cases. Above this frequency , standing waves start to build up inside the cavity and thus the resulting pressure at the listeners' eardrums becomes strongly dependent on the geometry of the listener's ear and on the headphone fitting (Schmidt, 2009). This high variability for subjects

[†]Part of the results presented in this chapter have been previously published in

- Masiero and Fels (2011b);
- Masiero and Fels (2011a);
- Fels and Masiero (2011).

corroborate for an individual equalization. To reduce the variability between fittings, the listeners should fit the headphones themselves at the most comfortable position.

This chapter starts by characterizing adequate headphones and microphones for an authentic binaural reproduction. Then the influence of headphone fitting on the measured individual HpTF is analyzed. Furthermore, an individual headphone equalization technique, perceptually robust to small variations in headphone fitting, is presented. The chapter concludes with a discussion of the obtained results.

4.1 Headphone Type

According to Møller (1992), ideal binaural reproduction via headphones is obtained if the headphone listening condition is equal to the free-field listening condition. He shows that for this to be true, the acoustical impedance seen from the ear canal should be the same for the two conditions. To verify if a headphone fulfills this requirement, he defined the *pressure division ration* (PDR) as

$$\mathrm{PDR}(z) \equiv \frac{H_{\mathrm{FF}}^{\mathrm{ED}}(z) / H_{\mathrm{FF}}^{\mathrm{EC}}(z)}{H_{\mathrm{HP}}^{\mathrm{ED}}(z) / H_{\mathrm{HP}}^{\mathrm{EC}}(z)} \tag{4.1}$$

where $H_{\mathrm{FF}}^{\mathrm{ED}}(z)$ is the transfer function from a free-field sound source (FF) to the listener's eardrums (ED), $H_{\mathrm{FF}}^{\mathrm{EC}}(z)$ is the transfer function from a free-field sound source to the microphones at the entrance of the ear canal (EC), $H_{\mathrm{HP}}^{\mathrm{ED}}(z)$ is the transfer function from the headphones (HP) to the listener's eardrums, and $H_{\mathrm{HP}}^{\mathrm{EC}}(z)$ is the transfer function from the headphones to the microphones at the entrance of the listener's ear canal (cf. fig. 4.1).

The idea behind the PDR is to verify if equalized headphones, when playing a binaural signal, can generate the same sound pressure at the listener's eardrum that would be generated by the original free-field sound source. This will only occur if $\mathrm{PDR}(z) = 1$. Møller et al. (1995b) showed that all headphones tested by them fulfilled this criterion for frequencies below 2 kHz and that only a few fulfilled this criterion for frequencies between 2 and 7 kHz. They named the headphones belonging to the second group as *free-air equivalent coupling* (FEC) headphones as they "do not disturb the radiation impedance as seen from the ear" (Møller, 1992). The methodology used by them did not allow a reliable

measurement of the PDR for frequencies above 7 kHz. Later investigations confirmed that FEC headphones cause the smallest variation on the impedance seen from the ear canal when compared to the free-field listening condition (Kleber and Vorländer, 2001; Cruzado, 2002).

Two candidate headphones—one having an electrodynamic transducer (Sennheiser HD-600) and the other having an electrostatic transducer (Stax SRλ), both of open-type—were tested regarding their PDR by measuring all four TFs defined in eq. (4.1) using an artificial head equipped with an IEC 711 ear simulator, i.e. with an artificial ear canal. For frequencies below 10 kHz both measured headphones complied with the FEC criterion, agreeing with the results presented by Völk (2011b). The electrodynamic headphones were chosen for further tests.

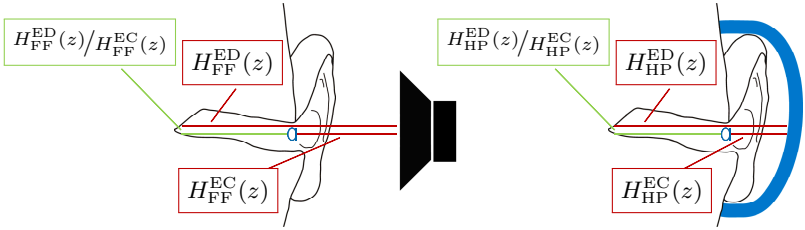
The PDR may also vary due to the headphone fitting and the type of microphones that are used. Oberem (2012) investigated both aspects. She conducted repeated measurements with the same setup described above, replacing the headphones and microphones at each new measurement. The results show that the miniature microphone placed with an ear plug at the entrance of the ear canal (also called acoustic *meatus*) yields the best results, as can be seen in fig. 4.2. Therefore, further measurements were conducted using miniature microphones placed at the entrance of the blocked ear canal.

4.2 Variability of Headphone Fitting

The variability of headphone transfer function (HpTF)[1] due to the fitting has been extensively investigated (Toole, 1984; Møller, 1992; Cruzado, 2002; Völk, 2011a). Furthermore, Paquier and Koehl (2010) confirmed the importance of headphone variations as they were able to verify that these spectral differences are audible.

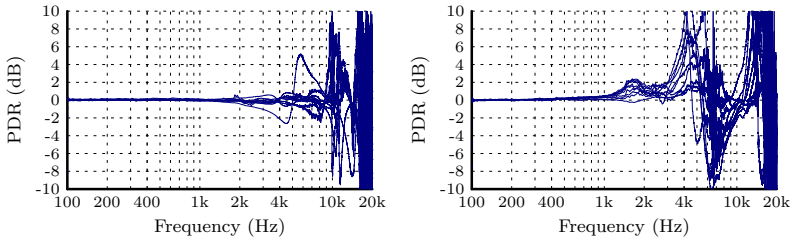
A series of individual HpTF measurements was carried out with 15 listeners. Generally speaking, it can be observed that interindividual HpTFs have a low variability up to approximately 4 kHz. In this frequency region, where the headphone works as an acoustic cavity, just a constant level variation is observed, caused by variable leakage, as described by Toole (1984). For higher frequencies, resonances are observed which vary with headphone fitting and the geometry of the listeners' ears. This higher variation occurs for two reasons:

[1] All HpTF plots have the y-axis displayed in dB relative to 1 Pa/V and only results of the left ear are displayed.



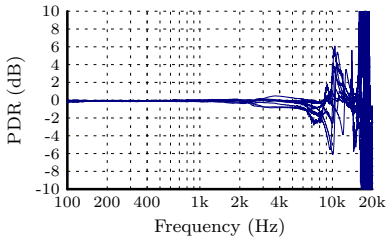
(a) (b)

Figure 4.1: Sketch of transfer paths required for the PDR calculation. (a) The free-air condition, measured with a loudspeaker in free-field and (b) the headphone condition.

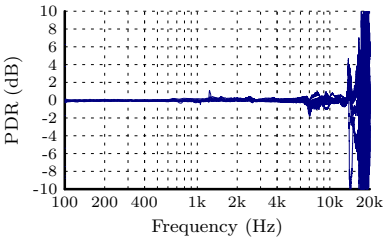


(a) Short Probe

(b) Long Probe

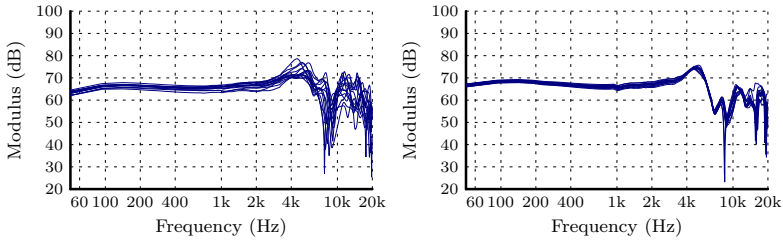


(c) Open Dome



(d) Blocked Meatus

Figure 4.2: PDR measured with an electrodynamic open headphone (Sennheiser HD-600) and with (a) a short probe microphone, (b) a long probe microphone, (c) a miniature microphone in an open dome, and (d) a miniature microphone in an ear plug blocking the meatus.



(a) Interindividual (b) Intraindividual

Figure 4.3: HpTF measured at the left ear with the Sennheiser HD-600 headphone for (a) fifteen different subjects and (b) one single subject with fifteen repetitions (at each new measurement the subject replaced the headphone for a comfortable fit). The good agreement at the intraindividual measurement was also observed with other subjects.

1. Because at this frequency range standing waves start to build up inside the headphone's cavity.

2. Because of the size of the external ear structures (Schmidt, 2009, p. 84)—meaning that in this region the HpTF behavior is highly individual, as can be seen in fig. 4.3(a).

Each individual repeated the HpTF measurement 15 times. For every new measurement the listener was instructed to place the headphone at its most comfortable fit. The result from the interindividual measurement for one exemplary listener can be seen in fig. 4.3(b) and confirmed that low measurement variability can be achieved if listeners are allowed to fit the headphone themselves at a comfortable fit. Masiero and Fels (2011b) showed that if listeners are instructed to place the headphones at extreme positions and not only at comfortable positions, the variability of the measured HpTF increases especially in the frequency range above 4 kHz.

4.3 Robust Individual Equalization

The variability results described in the last section agree with the results presented by Hammershøi and Møller (2005), who claim that "it can be seen that the variations between measurements are much less than

variations between subjects. The low variation in the repeated measurements means that an individual headphone (equalization) filter can be reliably designed. The high variation in transfer functions across subjects means that it probably should be designed individually". Therefore, to achieve a robust equalization, the equalization filter is constructed from the average of several individual HpTF measurements,[2] always completely removing the headphones in between measurements.

Headphone repositioning will cause variations that itself affect the equalized frequency response and appear as peaks or dips in the higher frequency range. Bucklein (1981) conducted speech intelligibility tests and showed that human listeners are more sensitive to spectrum irregularities in form of peaks than to equivalent valleys. Assuming that this behavior extends also to spatial perception, headphone equalization filters should also avoid the occurrence of resonance peaks in the equalized response. This can be achieved by applying a notch smoothing algorithm to the amplitude spectrum (cf. Müller, 1999, p. 192), which first smooths the entire frequency response and then compares it with the original function. At regions where this difference is higher than a given threshold, a cross-fading is made, thus locally smoothing the original function. Optionally, a softer smoothing can be done throughout the whole frequency spectrum afterwards.

Regarding the filters' overall gain, ideally, the equalization filter should not alter the loudness of the reproduced signals; but loudness measurements are dependent on the type of signal being used. For broadband signals, if the overall sound pressure level is kept constant, negligible variation on the loudness values should be observed. Therefore, the smoothed average HpTF is normalized by the root mean square value of the frequencies showing low variability, i.e., below 4 kHz. The applied weight should be the average of the RMS from both ears to allow the proper equalization of the interaural level difference.

As with any other equalization filter, care must be taken at frequencies outside the roll-off frequencies as correction at these regions may lead to very large gains that can produce undesired nonlinearity in the equalized response. Likewise, to equalize a headphone at low frequencies (below approximately 100 Hz) a very long FIR equalization filter is required. Since these equalization filters are aimed for use in real time virtual reality systems, it is of interest that these filters are kept short in order to avoid extra latency. As the low frequency range does

[2] Spectral average should be obtained independently for the amplitude spectrum and the group delay to avoid the unwanted phase canceling effect.

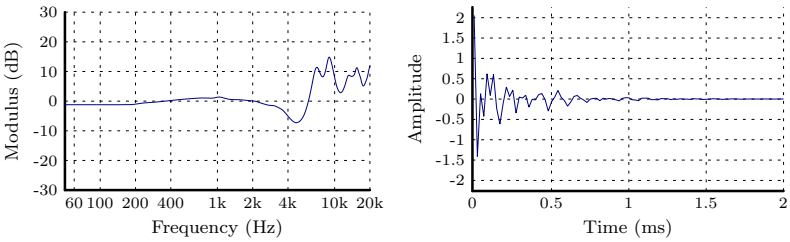


Figure 4.4: Individual headphone equalization filter calculated from the average of seven HpTFs. A notch smoothing algorithm was applied followed by a 1/6 octave smoothing (a). The time response (b) is obtained from the minimum-phase spectrum of (a).

not contribute to localization and the HpTF variation due to individual fitting in this frequency range is very low, this frequency region can be left untouched. This is done by substituting the frequencies below the first observed maximum of the HpTF with a constant line of the same value as the amplitude of the first maximum.

The last step is to invert the HpTF. Minnaar et al. (1999) discuss that HpTFs generally contain all-pass components that, when inverted, will drive the equalization filter to be noncausal. This effect could be compensated for by inserting a delay in the equalization filter or by equalizing only the minimum-phase component of the HpTFs. With the headphone fitting variation, the first option might lead to the compensation of a nonexistent all-pass section while the second option will not correct the present all-pass sections. Minnaar et al. (1999) suggest "that it will be more safe not to equalize for an all-pass that is there than to equalize for an all-pass that is not there." Therewith, only the magnitude spectrum of the smoothed average HpTF was inverted and the equivalent minimum-phase spectrum was obtained using the Hilbert transform, thus producing a causal and compact equalization filter.

4.4 Results

Figure 4.4 shows an example of an individualized headphone equalization filter. This filter was calculated from the average of seven individually measured HpTF magnitude spectra, each with a new headphone fit-

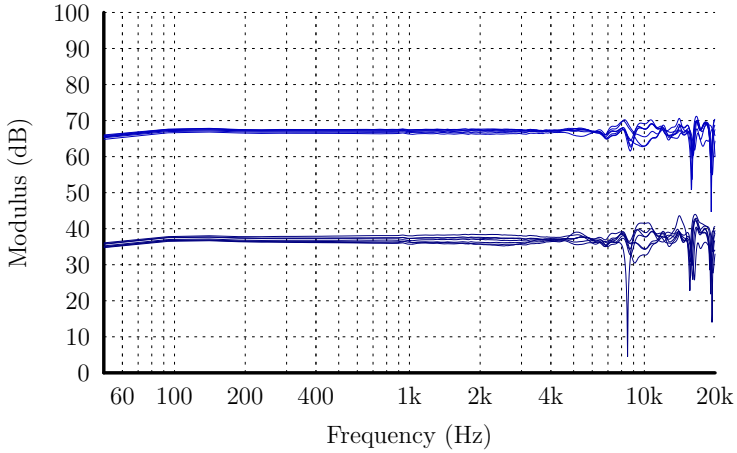


Figure 4.5: Equalization response for an individual headphone equalization filter. Seven HpTFs were averaged to generate the filter. The upper curves are obtained by multiplying the equalization filter with the same HpTFs used for the calculation. The lower curves are obtained by multiplying the equalization filter with the seven other HpTFs measured for the same listener. These curves are shifted by -30 dB for clarity.

ting. A notch smoothing algorithm was applied followed by a 1/6 octave logarithmic smoothing. The low frequency correction is truncated at approximately 200 Hz and the time response is obtained from the minimum-phase spectrum.

Applying this filter to the original HpTFs will result in a considerably flat equalized headphone response, except for spectrum dips at high frequencies, as depicted in the upper curves of fig. 4.5. This is, however, not a realistic situation as these HpTFs are the same ones used for the filter calculation. A realistic situation can be observed if the equalization filter is applied to a different set of HpTFs from the same listener. In this case, the variability increases slightly, as can be seen in the lower curves of fig. 4.5.

4.5 Discussion

Headphone Transfer Functions (HpTFs) were measured using an artificial head and individual subjects, confirming that for low and middle

frequencies only small level variations are present while for the high frequencies very individual resonance patterns are found. Low measurement variability is achieved if subjects are allowed to fit the headphone themselves at the most comfortable position. This underlines the fact that individual equalization should be used when possible.

A robust equalization filter design is proposed, inverting the average of the magnitude spectra of several individually measured HpTFs. A notch smoothing algorithm is applied to avoid high peaks at the spectrum of the equalization filter. The peaks in the HpTF—dips in the equalization filter—are not altered as the human hearing system is more sensitive to irregularities in form of peaks than to irregularities in form of dips. To avoid a noncausal equalization filter, only the minimum-phase component of the averaged HpTF is inverted. Furthermore, the headphone is not equalized for very low frequencies to keep the FIR equalization filter short.

The effectiveness of this filter calculation strategy is evaluated in chapter 6. It was verified that adequate individual equalization provide realistic binaural reproduction, in the sense that the listeners cannot differentiate between the real binaural signal generated by a loudspeaker and the virtual binaural signal synthesized with the individual HRTF and played back via individually equalized headphones.

If, however, an individual equalization cannot be carried out, Larcher et al. (1998) suggest the use of diffuse-field-compensated headphones with no additional equalization and diffuse-field equalized binaural signals. These signals can be synthesized by using DTFs instead of HRTFs.

5

Binaural Reproduction using Loudspeakers

If loudspeakers are used to reproduce a binaural signal, left and right signal will arrive mixed together at the listener's left and right ear, thus destroying the binaural cues and the spatial impression. To reestablish these cues *crosstalk cancellation* (CTC) filters, presented in section 5.1, are used to generate (from the input binaural signal) transaural signals to be fed to the loudspeakers which should interact to reproduce the binaural signal at the listener's ears with sufficient channel separation (defined in section 5.2).

The crosstalk cancellation is achieved by means of constructive and destructive wave interference. At some frequencies CTC filters may display elevated gains that might require the loudspeakers to reproduce very high sound pressures; only to be later (partially) canceled at the listener's ears (Takeuchi and Nelson, 2007). To avoid clipping and distortion at these frequencies, the overall gain of the CTC filter has to be reduced, causing the dynamic range of the reproduced binaural signal to shrink. According to Nelson and Rose (2006), these frequencies with extreme high energy will result in a poorly damped ringing behavior in the time-domain. This behavior will also occur in the spatial-domain, leading to a very narrow region with adequate binaural reproduction—the so called *sweet spot*.

The sweet spot can be enlarged by using a loudspeaker array with the high-frequency (tweeter) sources placed close to each other and the low-frequency sources (woofer) placed opposite to each other (Bauck and Cooper, 1996; Takeuchi and Nelson, 2007). However, to apply this loudspeaker placement strategy to an immersive virtual reality (VR) system, where the listener is constantly moving and consequently the sweet spot is also constantly shifting, would not be viable as a very large number of high frequency transducers distributed all around the

[‡]The results presented in this chapter are an extension of the results published in

- Masiero; Fels, and Vorländer (2011b);
- Masiero and Vorländer (2012).

reproduction space would be required. Therefore, all loudspeakers in this chapter are assumed to be broad-band sources.

CTC filters can be realized in various ways: either analog or digital, FIR or IIR, with or without room equalization. This thesis focuses on CTC filters for immersive virtual reality applications, which must be constantly updated to compensate for the listener's motion, should be as short as possible to save calculation power, and should introduce the shortest possible latency in the system to allow fast reaction to listener movements. For such applications, the use of digital FIR filters is preferred as it allows efficient filter updates with fast filter calculation (Lentz, 2007).

Digital CTC filters can be calculated either in time or in frequency-domain. Calculations in time-domain produce strictly causal filters. On the other hand, frequency-domain calculations are computationally more efficient, but may result in noncausal filters (see discussion in section 5.3). Apart from that, there is no substantial difference between the results achieved using the two methods (Parodi, 2008). In section 5.3 a general framework is introduced for the calculation of digital CTC filters with causality constraints in the frequency-domain.

Causal filters could also be achieved using a minimum-phase version of the HRTFs, as did by Gardner (1997). Parodi (2008) compared the performance of this generic CTC filter proposed by Gardner (1997) with time and frequency-domain least-mean-square approximations and concluded that the generic CTC filters provided reduced channel separation performance. The generic CTC filter calculation proposed by Gardner (1997) can also only be applied to two loudspeaker CTC filter. The framework presented in this chapter can be applied to generate CTC filters for an unlimited number of loudspeakers.

Kirkeby et al. (1998b) showed that CTC filters have infinitely long impulse responses (IRs) even when they are derived from a set of finite head-related transfer functions (HRTFs). They proposed the use of Tikhonov regularization to control undesirably large peaks in the frequency response of the CTC filters. As these peaks are responsible for weakly damped ringing behavior in time-domain, the use of regularization also reduces the length of the CTC filters. The side-effect of regularization is the appearance of unwanted noncausal artifacts in both the CTC filters and the resulting ear signals. Taking into account the results of minimum-phase regularization, the presented framework is extended to force the filters resulting from the regularized inverse problem to be causal.

To avoid filter instability as users rotate their head inside an immersive VR environment, Lentz (2006) suggested the use of four loudspeak-

ers from which only two loudspeakers are active at a time, depending on the orientation of the listener's head. In section 5.4 an improved solution for the filter switching strategy is presented which integrates spatial fading in the filter design stage and makes it possible to smoothly switch between active loudspeakers.

The CTC filter calculation framework presented in this chapter is based on the knowledge of the transfer-path between loudspeakers and listener's ears, i.e., the HRTF. However, no restriction is made if these HRTFs have to be individually measured. The influence of individualized HRTFs in localization performance with CTC systems will be later discussed in section 6.2.

In this chapter some results will be discussed in regard to the obtained channel separation, defined in section 5.2, which is commonly referred to in literature as a quality predictor for CTC filters. In section 6.2 the relationship between channel separation and localization performance of a CTC system is investigated.

5.1 CTC Reproduction System

Figure 5.1 shows the setup of a CTC system with two loudspeakers. The transmission path from the loudspeakers to the listener's left and right eardrums can be written in the frequency-domain as

$$E_{\mathrm{L}}(z) = H_{1\mathrm{L}}(z)V_1(z) + H_{2\mathrm{L}}(z)V_2(z), \tag{5.1a}$$
$$E_{\mathrm{R}}(z) = H_{1\mathrm{R}}(z)V_1(z) + H_{2\mathrm{R}}(z)V_2(z), \tag{5.1b}$$

where $E_{\mathrm{L}}(z)$ and $E_{\mathrm{R}}(z)$ are the signals at the listener's ears, $V_1(z)$ and $V_2(z)$ are the signals fed to the loudspeakers and $H_{nL}(z)$ and $H_{nR}(z)$ represent the acoustic path from the n^{th} loudspeaker to the left and right ears, respectively.

Equation eq. (5.1), can be written in matrix formulation as

$$\begin{bmatrix} E_{\mathrm{L}}(z) \\ E_{\mathrm{R}}(z) \end{bmatrix} = \begin{bmatrix} H_{1\mathrm{L}}(z) & H_{2\mathrm{L}}(z) \\ H_{1\mathrm{R}}(z) & H_{2\mathrm{R}}(z) \end{bmatrix} \cdot \begin{bmatrix} V_1(z) \\ V_2(z) \end{bmatrix} \tag{5.2}$$

or

$$\boldsymbol{e} = \boldsymbol{H}\boldsymbol{v}, \tag{5.3}$$

where elements of $\boldsymbol{e}$ are the signals at the listener's ears, elements of $\boldsymbol{H}$ (called *acoustic transfer matrix*) describe the acoustic propagation

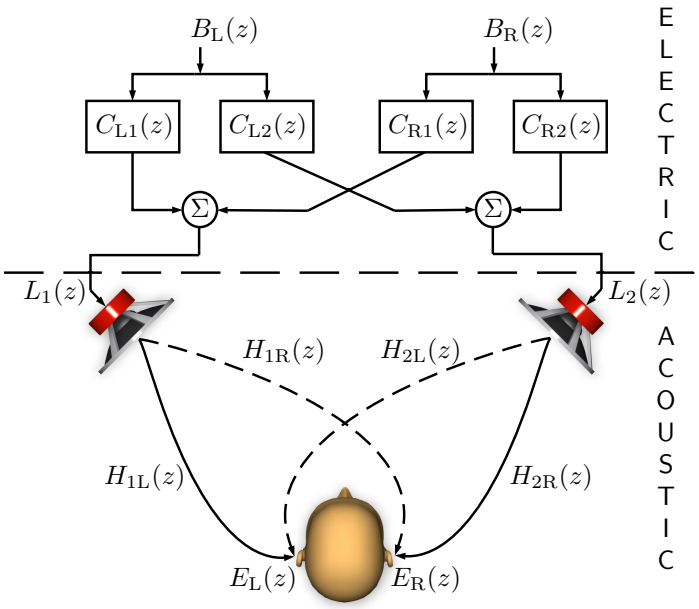


Figure 5.1: Diagram of a binaural reproduction system using loudspeakers, i.e., a crosstalk cancellation (CTC) system. The CTC filters are shown in the upper part and the acoustic paths are shown in the lower part of the figure. The solid and dashed lines show the direct and the crosstalk paths, respectively.

paths, and the elements of $\boldsymbol{v}$ are the loudspeaker signals, all in frequency-domain.[1]

The crosstalk paths can be canceled out using an adequate filter structure. This should be always placed between the input binaural signal and the loudspeakers (see fig. 5.1), and can be represented as matrix $\boldsymbol{C}$, the so-called *crosstalk cancellation matrix*, such that

$$\boldsymbol{v} = \boldsymbol{C}\boldsymbol{b}, \tag{5.4}$$

where the elements of $\boldsymbol{b}$ are the left and right binaural signals to be presented, resulting in the complete transmission path

$$\boldsymbol{e} = \boldsymbol{H}\boldsymbol{C}\boldsymbol{b}. \tag{5.5}$$

[1] All equations presented in this section are in frequency-domain. They can be recast, however, in a time-domain representation, as discussed in section 5.3.

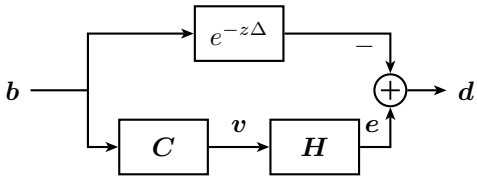


Figure 5.2: The crosstalk cancellation problem displayed as a block diagram.

A correct binaural reproduction is achieved when, apart from a time delay, the binaural signal $\boldsymbol{b}$ is exactly reproduced at the listener's ears. As discussed in section 2.2, this problem can be studied as a minimization problem, viz. the minimization of the reproduction error

$$\boldsymbol{d} = \left(\boldsymbol{e} - \boldsymbol{b} \cdot e^{-z\Delta}\right), \tag{5.6}$$

where Δ is a time delay proportional to the acoustic lag between loudspeakers and listener position. The block diagram form in fig. 5.2 shows the minimization problem that occurs while obtaining optimal CTC filters.

Substituting eq. (5.5) in eq. (5.6) one has

$$\boldsymbol{d} = \left(\boldsymbol{HC} - \boldsymbol{I} \cdot e^{-z\Delta}\right) \boldsymbol{b}, \tag{5.7}$$

which highlights the dependence of the optimal CTC filter on the input binaural signal. Kirkeby and Nelson (1999) suggest substituting $\boldsymbol{b}$ by a delta function, thus obtaining a filter for the "worst case" scenario where the input binaural signal contains energy in the entire frequency spectrum.

For the two loudspeaker setup shown in fig. 5.1, the crosstalk cancellation matrix that minimizes the reproduction error can be obtained from eq. (2.7). Assuming that $\boldsymbol{H}$ is invertible, this is given by

$$\boldsymbol{C} = \boldsymbol{H}^{-1} e^{-z\Delta}. \tag{5.8}$$

The binaural signals do not necessarily have to be reproduced using only two loudspeakers (Bauck and Cooper, 1992). If N loudspeakers are used instead, $\boldsymbol{H}$ expands to

$$\boldsymbol{H} = \begin{bmatrix} H_{1\mathrm{L}}(z) & H_{2\mathrm{L}}(z) & \cdots & H_{NL}(z) \\ H_{1\mathrm{R}}(z) & H_{2\mathrm{R}}(z) & \cdots & H_{NR}(z) \end{bmatrix}.$$

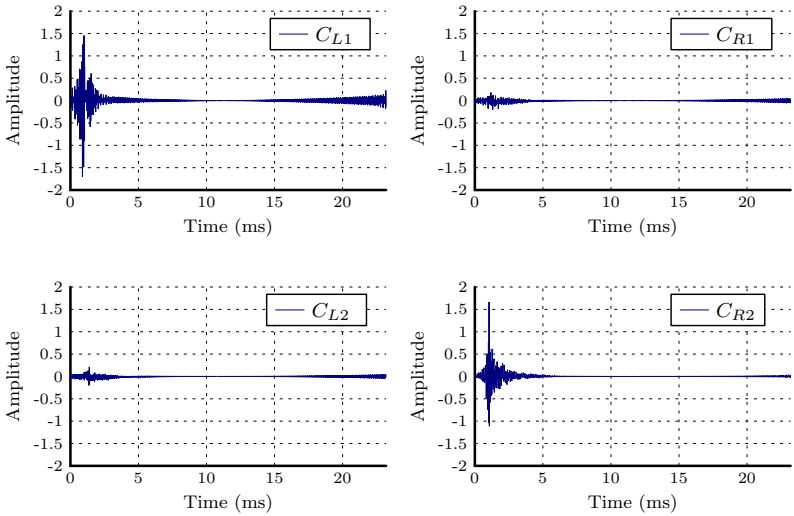


Figure 5.3: Time response of $\boldsymbol{C}$ for two loudspeakers placed at $\phi = \pm 45°$ calculated with the regularized equation eq. (5.24) using $\mu = 0.005$ for all frequencies and $\Delta = 3.4\,\text{ms}$. Noncausal oscillations are clearly visible in all four filters, even though a time delay proportional to the distance between loudspeakers and head was used.

$\boldsymbol{H}$ now represents an underdetermined system and the CTC filters obtained by a least-squares minimization are given by (see appendix B)

$$\boldsymbol{C} = \boldsymbol{H}^* \left(\boldsymbol{H}\boldsymbol{H}^*\right)^{-1} e^{-z\Delta}. \tag{5.9}$$

The CTC formulation could easily be expanded for multiple listeners by concatenating $\boldsymbol{e}$, $\boldsymbol{H}$, and $\boldsymbol{b}$ without actually altering the filter calculation scheme (Kim et al., 2006; Masiero and Qiu, 2009).Bauck and Cooper (1992) studied a number of multiple listeners CTC configurations, for instance a setup with fewer loudspeakers than listener's ears (overdetermined system) and a very interesting setup for cinema application that uses one central loudspeaker and several distributed dipole loudspeaker placed behind the listeners' heads. This chapter, however, will focus only on the one listener setup.

5.2 Channel Separation

The channel separation (CS) has been proposed to describe the quality of CTC systems (Gardner, 1997) and has been defined as the logarithmic difference between the signals at the ipsilateral and the contralateral ear (Bai and Lee, 2006). Thus, assuming

$$\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

an ideal CTC system will generate a signal only at the left ear. Equation (5.5) then reduces to

$$e_{\mathrm{L}} = H_{1\mathrm{L}} C_{\mathrm{L}1} + H_{2\mathrm{L}} C_{\mathrm{L}2}, \tag{5.10a}$$

$$e_{\mathrm{R}} = H_{1\mathrm{R}} C_{\mathrm{L}1} + H_{2\mathrm{R}} C_{\mathrm{L}2}. \tag{5.10b}$$

and the channel separation for the left ear can be calculated as

$$\mathrm{CS}_{\mathrm{L}} = 20 \log_{10} \left(\frac{|H_{1\mathrm{L}} C_{\mathrm{L}1} + H_{2\mathrm{L}} C_{\mathrm{L}2}|}{|H_{1\mathrm{R}} C_{\mathrm{L}1} + H_{2\mathrm{R}} C_{\mathrm{L}2}|} \right). \tag{5.11}$$

The same applies to the right ear. This definition for CS suggests that a larger CS results in a better CTC system, which follows the definition used by Akeroyd et al. (2007) and Qiu et al. (2009), but is contrary to the definition used by Bai and Lee (2006) and Parodi and Rubak (2010). Note that the CS is given separately for each frequency and that the CS is usually averaged over a defined frequency range in order to obtain a single valued quality metric for the CTC system (Bai and Lee, 2006; Akeroyd et al., 2007).

Without the use of CTC filters, i.e., assuming that $\boldsymbol{C} = \boldsymbol{I}$, the channel separation for the left ear would be

$$\widehat{\mathrm{CS}}_{\mathrm{L}} = 20 \log_{10} \left(\frac{|H_{1\mathrm{L}}|}{|H_{1\mathrm{R}}|} \right). \tag{5.12}$$

Again, the same definition applies to the right ear. $\widehat{\mathrm{CS}}$ represents the natural CS caused by head shadowing and it is equivalent to the CS observed using a simple stereophonic reproduction system. The $\widehat{\mathrm{CS}}$ depends directly on the system's loudspeaker position. It is also frequency-dependent and it has its maximum of approximately 30 dB at higher frequencies (Blauert, 1997).

For an ideal CTC system, the obtained CS is expected to be substantially larger than $\widehat{\mathrm{CS}}$. However, in practical applications, the

setup HRTFs used to design the CTC filters in $\boldsymbol{C}$ are not always identical to the *playback* HRTFs contained in $\boldsymbol{H}$. In these cases, a poorer performance of the CTC system can be expected, as shown in section 6.2.2.

5.3 Filter Design

As discussed in the previous section, CTC filters can be calculated either in the time or in the frequency-domain. The frequency-domain solution given in eq. (5.8) can be recast in the time-domain as

$$\widehat{\boldsymbol{C}} = \widehat{\boldsymbol{H}}^{-1}\boldsymbol{I}(\Delta). \tag{5.13}$$

where $\widehat{\boldsymbol{H}}$ is the concatenation of the convolution matrices of each HRIR in $\boldsymbol{H}$, $\widehat{\boldsymbol{C}}$ is the concatenation of the impulse response from the CTC filters, and $\boldsymbol{I}(\Delta)$ is a block diagonal matrix with two delayed delta functions in its diagonal (Kirkeby and Nelson, 1999; Parodi, 2010).

If the available HRIRs are N samples long and the desired CTC filters are M samples long, then $\widehat{\boldsymbol{H}}$ will be a $2(M+N-1)\times 2M$ matrix for the two loudspeaker CTC configuration. This matrix is overdetermined and applying eq. (2.7) would result in the inversion of a single $2M \times 2M$ real matrix. The same problem occurs in the frequency-domain. Assuming that the filter length for both $\boldsymbol{H}$ and $\boldsymbol{C}$ is $M+N-1$, would require $M+N-1$ times the inversion of a 2×2 complex matrix.[2]

As a matrix inversion has a computational complexity $O(n^3)$,[3] an inversion in the frequency-domain has the advantage that its computational requirements are considerably smaller than computation in the time-domain, even considering the required FFTs. This is a major advantage for real-time VR systems since these types of systems require constant filter updating. Already for medium-size filters (around 500 coefficients), it is usually more efficient to repeat the inversion of a small matrix several times, as in the case of the inversion in frequency-domain, than to invert a large matrix only once, which would be done when obtaining the filters directly in time-domain.

[2] Please note that $\boldsymbol{H}$ and $\boldsymbol{C}$ are a three-dimensional tensor, while $\boldsymbol{e}$, $\boldsymbol{v}$, and $\boldsymbol{R}$ are two-dimensional tensors. As the addition and multiplication operations can be conducted independently in the frequency dimension, for each frequency, the three-dimensional tensors can be considered matrices and the two-dimensional tensors can be considered vectors.

[3] There are faster algorithms for matrix inversion with a computation complexity as low as $O(n^{2.3727})$. However, these algorithms usually only produce a considerable speedup for very large matrices.

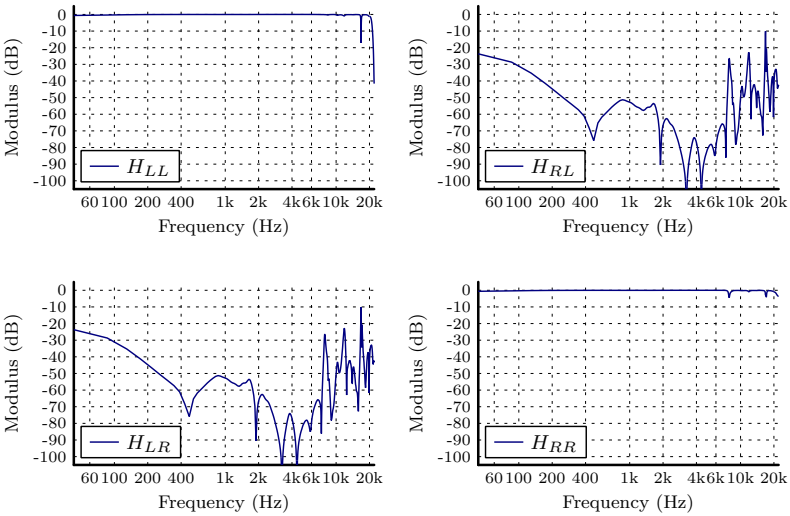


Figure 5.4: Frequency response of the complete transfer-path between the binaural signals and the ear signals for the filters shown in fig. 5.3. The diagonal elements are ideally 0 dB, the off-diagonal elements are ideally $-\infty$ dB. Deviation to the ideal result is caused by the regularization applied at the CTC filter calculation.

When calculating the CTC filter for a two loudspeaker setup in the frequency-domain, there is only one set of filters that can force the reproduction error to be exactly zero. If more than two loudspeakers are available, the transmission matrix $\boldsymbol{H}$ turns into an underdetermined matrix and in this case, there is an infinite number of CTC filters that can deliver an ideal reproduction. The least-square minimization (LSM) used so far chooses (from this infinite group of filters) the CTC filter combination with minimum Euclidean norm (energy) solution by solving the minimization problem

$$\begin{aligned} &\underset{\boldsymbol{c}_j}{\text{minimize}} && \|\boldsymbol{c}_j\|_2 \\ &\text{subject to} && \boldsymbol{H}\boldsymbol{C} = \boldsymbol{I}, \end{aligned} \tag{5.14}$$

where $\boldsymbol{c}_j$ is the j^{th} column of $\boldsymbol{C}$. This minimization can be solved by applying the Lagrangian multipliers, as explained in appendix B.

Instead of minimizing the ℓ_2-norm, one could minimize the ℓ_1-norm of $\boldsymbol{c}_j$, obtaining a set of filters with its coefficients sparsely distributed in the loudspeaker dimension, i.e., at each frequency the least possible number of loudspeakers will be active. This results in CTC filters whose energy is compactly distributed in the spectrum.[4] The result obtained using the ℓ_1-norm minimization resembles the optimal source distribution (OSD) setup, as Takeuchi and Nelson (2007) suggests using only two distinctly positioned loudspeakers for each frequency band as well.

The opposite situation would be to minimize the ℓ_∞-norm of $\boldsymbol{c}_j$. In this case, the obtained set of filters will have its energy equally distributed between all loudspeakers and therefore also along the whole spectrum, differing mainly in the phase response. As many loudspeakers are active and playing almost the same signal, the ear signals are obtained from a very intricate superposition of the many arriving wavefronts and a narrow sweet spot can be expected. Note that this formulation differs from the minimax CTC filter design proposed by Rao et al. (2007).[5]

ℓ_1 and ℓ_∞ minimization problems have, in contrast to the LSM, no analytical solution and must be solved using computationally intensive iterative methods (Boyd and Vandenberghe, 2004).

The frequency-domain LSM calculation is the fastest method to obtain the CTC filters and is therefore used for real-time binaural reproduction setups. The generalized frequency-domain solution given by eq. (5.9) will deliver in the shortest time the best possible channel separation for a given listener-loudspeaker setup. As ideal CTC filters are infinitely long when calculated using the DFT (Kirkeby et al., 1998b), the obtained CTC filters might suffer from cyclic aliasing and noncausality. To minimize such artifacts, Kirkeby et al. (1998a) proposed to extend the frequency resolution of the original transfer matrix $\boldsymbol{H}$, which increases the calculation time, and/or to apply a regularization constraint, which adds unwanted ringing artifacts. A framework is now described for obtaining causal filters from frequency-domain calculations even when regularization is applied.

[4] A filter that is sparse in the frequency-domain will have an IR that is spread in time.

[5] The objective of Rao et al. (2007) is to find, in the time-domain, a set of *short* CTC filters (that per se will not ensure a perfect channel separation) whose minimum channel separation is maximized. In contrast to the frequency-domain problems described in this section, their formulation is an overdetermined problem, thus the ℓ_∞ constraint is applied to the reproduction error and not the coefficients of the CTC filter.

5.3.1 Causality

If the acoustic travel time between the loudspeakers and ears is not compensated for, the resulting CTC filters will be noncausal. This problem can be easily solved by introducing an additional latency Δ in the filters. However, regardless of this time compensation, since HRTFs are not minimum-phase, their inverse will contain a noncausal component.

Calculations of the CTC filters in the time-domain will produce a filter set which is causal, but delivers reduced channel separation in comparison to the ideal filters obtained by frequency-domain calculation eq. (5.9). If the acoustic lag is not compensated for, a noncausal filter would be required and since the product of the time-domain calculation is strictly causal, no filter would be calculated in this case. But besides the compensation of the acoustic lag, it was observed that an extra delay of approximately 1 ms will allow the presence of a certain amount of pre-ringing in the CTC filters which results in improved channel separation. This pre-ringing, originated per se from the filter calculation, will cancel out itself at the ear signal, differently than the pre-ringing originated from regularization which will remain present at the ear signal.

The frequency-domain calculation will deliver an optimal channel separation, but these filters will suffer from time aliasing. Once the filters are shifted and windowed to allow a causal response, the channel separation will also deteriorate.

To combine fast calculation time with causal filter response, a causality constraint can be imposed in the frequency-domain calculation. Using the identity

$$(\cdot)^{-1} = \mathrm{adj}(\cdot)/\det(\cdot), \tag{5.15}$$

where $\mathrm{adj}(\cdot)$ is the adjugate of a matrix[6] and $\det(\cdot)$ its determinant, eq. (5.9) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{L} - \boldsymbol{C}D(f) = 0, \tag{5.16}$$

where $\boldsymbol{L} = \boldsymbol{H}^* \,\mathrm{adj}\,(\boldsymbol{H}\boldsymbol{H}^*)\, e^{-z\Delta}$ and $D(f) = \det(\boldsymbol{H}\boldsymbol{H}^*)$. When written in time-domain, each element of $\boldsymbol{Y}$ is given by

$$y_{ij}(t) = l_{ij}(t) - c_{ij}(t) * d(t) = 0. \tag{5.17}$$

[6] Note that for the special case of a 2×2 matrix the adjugate can be obtained without further calculation.

As discussed by Papoulis (1977, p. 340), a causal constraint can be applied to eq. (5.17) resulting in

$$y_{ij}(t) = l_{ij}(t) - \int_0^\infty d(t-\tau) c'_{ij}(\tau) \mathrm{d}\tau = 0. \tag{5.18}$$

Because of the causal constraint, $y_{ij}(t) = 0$ is valid only for $t > 0$. The integral in eq. (5.18) is clearly a convolution of $d(t)$ with $c'_{ij}(t)$, where $c'_{ij}(t) = 0$ for $t < 0$.

According to Papoulis (1977), "it suffices to find a causal function $c'_{ij}(t)$ and an anti-causal function $y_{ij}(t)$ satisfying eq. (5.18)." He does that by transforming eq. (5.18) in the Laplace-domain and arguing that $Y_{ij}(s)$ must be analytic for $\Re\{s\} < 0$ and $C'_{ij}(s)$ must be analytic for $\Re\{s\} > 0$. The transform of $l_{ij}(t)$ and $d(t)$ are uniquely determined in term of their spectra. Thus, the transform results in

$$Y_{ij}(s) = L_{ij}(-js) - C'_{ij}(s) D(-js). \tag{5.19}$$

Papoulis (1977) finds the solution by first factoring $D(-js)$ so that

$$D(-js) = A^+(s) A^-(s), \tag{5.20}$$

where $A^+(s)$ and its inverse $1/A^+(s)$ are analytic for $\Re\{s\} > 0$ and $A^-(s)$ and its inverse $1/A^-(s)$ are analytic for $\Re\{s\} < 0$.

The next step is to factor the ratio $L_{ij}(-js)/A^-(s)$ as the sum

$$L_{ij}(-js)/A^-(s) = B^+(s) + B^-(s), \tag{5.21}$$

where function $B^+(s)$ is analytic for $\Re\{s\} > 0$ and function $B^-(s)$ is analytic for $\Re\{s\} < 0$.

The desired causal constrained filters are then given by

$$C'_{ij}(s) = B^+(s)/A^+(s). \tag{5.22}$$

Papoulis (1977) concludes the solution proving that $C'_{ij}(s)$ is, as desired, analytic for $\Re\{s\} > 0$ because functions $B^+(s)$ and $1/A^+(s)$ are analytic by construction for $\Re\{s\} > 0$ and proving that $Y(s) = B^-(s)A^-(s)$ is analytic for $\Re\{s\} < 0$ because functions $B^-(s)$ and $A^-(s)$ are also analytic by construction for $\Re\{s\} < 0$.

Equation (5.22) can be rewritten in matrix form as

$$\boldsymbol{C}' = \frac{1}{\det(\boldsymbol{H}\boldsymbol{H}^*)^+} \left[\frac{\boldsymbol{H}^* \operatorname{adj}(\boldsymbol{H}\boldsymbol{H}^*) e^{-z\Delta}}{\det(\boldsymbol{H}\boldsymbol{H}^*)^-} \right]_+, \tag{5.23}$$

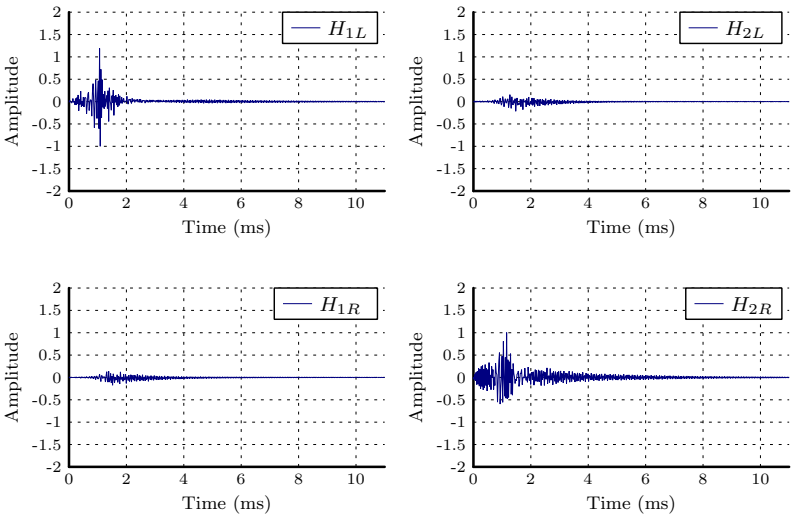


Figure 5.5: Time response of $\boldsymbol{C}$ for two loudspeakers placed at $\phi = \pm 45°$ calculated with the framework presented in section 5.3 using $\mu = 0.005$ for all frequencies and $\Delta = 3.4\,\text{ms}$. The resulting filters are strictly causal.

where $(\cdot)^+$ and $(\cdot)^-$ are, respectively, the minimum causal stable and minimum anti-causal stable parts of the determinant. As $\boldsymbol{H}\boldsymbol{H}^*$ is a Hermitian matrix, $\det(\boldsymbol{H}\boldsymbol{H}^*)$ is real and even. In this case, the Wiener-Hopf decomposition can be efficiently implemented in the cepstral-domain allocating the first half of the cepstrum to the causal stable part and the second half for the anti-causal stable part. Further, $[\cdot]_+$ denotes the estimate of the causal part of each IR which can be obtained by windowing out the second half of the IR. Kim and Wang (2003) make the comment that as eq. (5.18) has a "linear convolution operator, not circular, care should be taken in calculating the convolution in the digital domain."

5.3.2 Regularization

Unfortunately, the transfer matrix $\boldsymbol{H}$ is not always well-conditioned, in which case the CTC filters might produce very high gains causing not only a loss of dynamic range, but also generating the so-called "ringing frequencies" (Nelson and Rose, 2006). Kirkeby et al. (1998a) pro-

posed the use of regularization to limit these high gains, thus limiting the energy of the loudspeaker signal and consequently reducing loudspeaker fatigue and nonlinear behavior as well. As already discussed in section 2.2, regularization is obtained by adding a constraint on the maximum energy of the CTC filters (see appendix B.1).

The optimum filters that satisfy these constraints are given by

$$\boldsymbol{C} = \boldsymbol{H}^* \left(\boldsymbol{H}\boldsymbol{H}^* + \mu \boldsymbol{I}\right)^{-1} e^{-z\Delta}. \tag{5.24}$$

The regularization parameter μ now acts as a trade-off factor between channel separation and dynamic loss. As a by-product, regularization reduces the size of the CTC filters while increasing the noncausal behavior of the filters (Kirkeby et al., 1998a).

Kirkeby and Nelson, 1999 showed how regularization can also be applied to calculations in the time-domain and be made frequency-dependent by filtering the *control effort* with a filter $R(z)$, resulting in

$$\boldsymbol{C} = \boldsymbol{H}^* \left(\boldsymbol{H}\boldsymbol{H}^* + \mu R(z)^* R(z) \boldsymbol{I}\right)^{-1} e^{-z\Delta}, \tag{5.25}$$

where $R(z)$ attenuates all frequencies that should not be regularized. Usually, $R(z)$ has the form of a band-stop filter when it is used to design CTC filters. Note that $R(z)^* R(z)$ is real-valued and acts only as a shape-factor of the regularization that determines which frequencies are to be regularized. Assuming that the same filter will be applied to all channels, $\mu R(z)^* R(z)$ is abbreviated as $\mu(z)$ in the remainder of this thesis.

5.3.3 Minimum-Phase Regularization

Regularization introduces pre-ringing in both the CTC filters and the resulting ear signals (Fielder, 2003; Norcross and Bouchard, 2007). As the regularization parameter is increased, the maximum amplitude of the pre-ringing component increases and the decay rate of the pre-ringing also increases.[7] This pre-ringing can result in audible artifacts if the filters are heavily regularized at certain frequencies. Since the human auditory system has a much longer post-masking behavior than pre-masking (Fastl and Zwicker, 2007), it is desirable to alter the regularization procedure so that (at least part of) the pre-ringing is converted into post-ringing.

[7] It is important to stress that increasing the regularization parameter will nevertheless reduce the total filter length.

As discussed in section 2.3.2, the regularized deconvolution of a single channel can be interpreted as the direct spectral inversion multiplied by $A(z)$, the regularization shape-factor, given by

$$A(z) = \frac{1}{1 + \mu(z) / |H(z)|_2^2}, \tag{5.26}$$

which has a real spectrum (as $\mu(z)$ is real) and, therefore, exhibit a symmetric and noncausal associated IR. Norcross and Bouchard (2007) suggest substituting $A(z)$ with its minimum-phase equivalent $A_{\mathrm{mp}}(z)$ to avoid noncausal artifacts caused by filtering the inverse of $H(z)$ with $A(z)$, and to ensure a frequency regularization without any noncausal artifacts caused by the regularization.

For the multi-channel case, the method presented in Norcross and Bouchard, 2007 has the drawback that the minimum-phase correction has to be made for each channel individually. It is possible to approximate a global minimum-phase regularization if eq. (5.15) is expanded to

$$\boldsymbol{C} = \frac{\boldsymbol{H}^* \operatorname{adj}(\boldsymbol{H}\boldsymbol{H}^* + \mu(z)\boldsymbol{I})}{\det(\boldsymbol{H}\boldsymbol{H}^* + \mu(z)\boldsymbol{I})} e^{-z\Delta}. \tag{5.27}$$

As the calculation of the adjugate of a matrix does not involve any division operation, one can assume that $\operatorname{adj}(\boldsymbol{H}\boldsymbol{H}^* + \mu(z)\boldsymbol{I}) \approx \operatorname{adj}(\boldsymbol{H}\boldsymbol{H}^*)$ as long as $\mu(z)$ is small compared to the elements of $\boldsymbol{H}$. Thus, the major influence of regularization occurs at the inversion of the determinant. Similar to eq. (2.11), the effect of regularization can be described by a regularization filter $A(z)$, so that

$$\frac{1}{\det(\boldsymbol{H}\boldsymbol{H}^* + \mu(z)\boldsymbol{I})} \equiv \frac{A(z)}{\det(\boldsymbol{H}\boldsymbol{H}^*)}, \tag{5.28}$$

which equates to

$$A(z) = \frac{\det(\boldsymbol{H}\boldsymbol{H}^*)}{\det(\boldsymbol{H}\boldsymbol{H}^* + \mu(z)\boldsymbol{I})}. \tag{5.29}$$

Again, as the determinant of a Hermitian matrix is real, the numerator and the denominator of eq. (5.29) will be real and thus $A(z)$ will also be real. Substituting the regularization filter $A(z)$ by its minimum-phase equivalent $A_{\mathrm{mp}}(z)$ results in

$$\boldsymbol{C}_{\mathrm{mp}} = \frac{A_{\mathrm{mp}}(z)\boldsymbol{H}^* \operatorname{adj}(\boldsymbol{H}\boldsymbol{H}^* + \mu(z)\boldsymbol{I}) e^{-z\Delta}}{\det(\boldsymbol{H}\boldsymbol{H}^*)}, \tag{5.30}$$

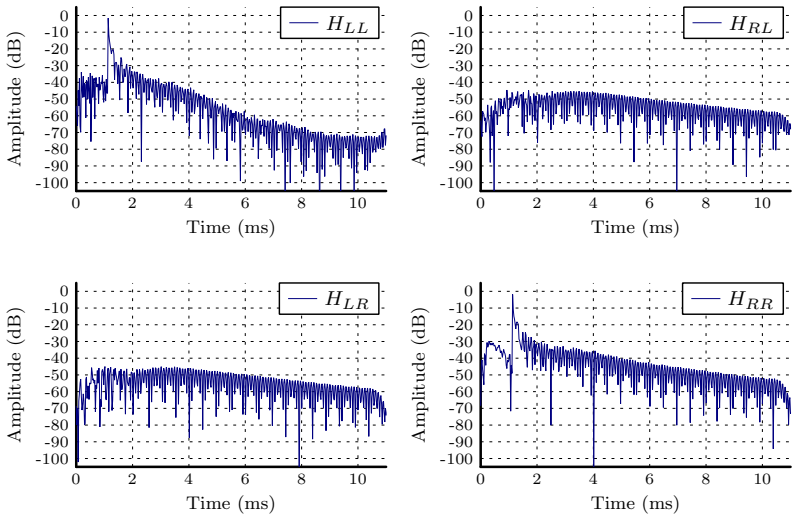


Figure 5.6: Time response of the complete transfer-path between the binaural signals and the ear signals for the filters shown in fig. 5.5. The effect of minimum-phase regularization can be observed in the impulse responses of the diagonal elements, as the impulse responses have a sharp onset (the oscillations prior to the impulse response are caused by noise, as individualized but mismatched HRTFs were used for this calculation).

which has the same amplitude response as eq. (5.25) but with all non-causal artifacts produced by regularization converted in its causal equivalent. It is also possible to combine the zero-phase with the minimum-phase of $A(z)$ in a trade-off between pre- and post-ringing (Norcross and Bouchard, 2007).

By combining eqs. (5.23) and (5.30) a causal CTC filter is obtained

$$\boldsymbol{C}'_{\mathrm{mp}} = \frac{A^{+}_{\mathrm{mp}}(z)}{\det\left(\boldsymbol{H}\boldsymbol{H}^{*}\right)^{+}} \left[\frac{A^{-}_{\mathrm{mp}}(z)\boldsymbol{H}^{*}\operatorname{adj}\left(\boldsymbol{H}\boldsymbol{H}^{*} + \mu(z)\boldsymbol{I}\right) e^{-z\Delta}}{\det\left(\boldsymbol{H}\boldsymbol{H}^{*}\right)^{-}} \right]_{+}, \tag{5.31}$$

which will result in ear signals that are causal and free of pre-ringing.

5.4 Weighting

When designing a CTC reproduction system for immersive VR environments, two loudspeakers will not be sufficient to allow the listener to rotate his/her head freely. If the listener's head points in a direction outside of the arc spanned by both loudspeakers, the CTC system will become unstable (Lentz, 2006). To meet the requirements of an immersive VR environment, Lentz (2006) designed a system with four loudspeakers. However, as he employed the truncated CTC filter calculation algorithm (Köring and Schmitz, 1993), only two loudspeakers could be used to reproduce the binaural signals. Thus, the active pair of loudspeakers had to be exchanged according to the orientation of the listener's head. The switching between each pair of active loudspeakers was made by a soft fading between the filters. This may lead to unwanted artifacts.

To avoid such fading artifacts, all loudspeakers could be used simultaneously. On the other hand, measurements show that "two-channel configurations result in wider controlled area and are more robust to head rotation and frontal displacement than the four-channel configurations" (Parodi and Rubak, 2010). As more sources will interact in a more complex way, smaller displacements will lead to larger errors. Thus it is reasonable to reduce the number of active loudspeakers to two,[8] but with an improved filter fading strategy.

A smoother transition between the active loudspeakers can be obtained by using a weighted matrix inversion where different weights can be applied to each loudspeaker according to the direction in which the listener's head is pointing.

The *weighted* ℓ_2 norm is given by

$$\|\boldsymbol{x}\|_{\boldsymbol{Z}}^2 = \boldsymbol{x}^* \boldsymbol{Z} \boldsymbol{x}, \tag{5.32}$$

where $\boldsymbol{Z}$ is a diagonal matrix containing positive weights for each element of $\boldsymbol{x}$.

The optimum set of filters that minimizes the weighted energy is given by (see appendix B.2)

$$\boldsymbol{C} = \boldsymbol{W}\boldsymbol{H}^* \left(\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^* + \mu(z)\boldsymbol{I}\right)^{-1} e^{-z\Delta}, \tag{5.33}$$

[8] Simulation results suggest that the use of three loudspeakers will increase the robustness of the system (Yang et al., 2003).

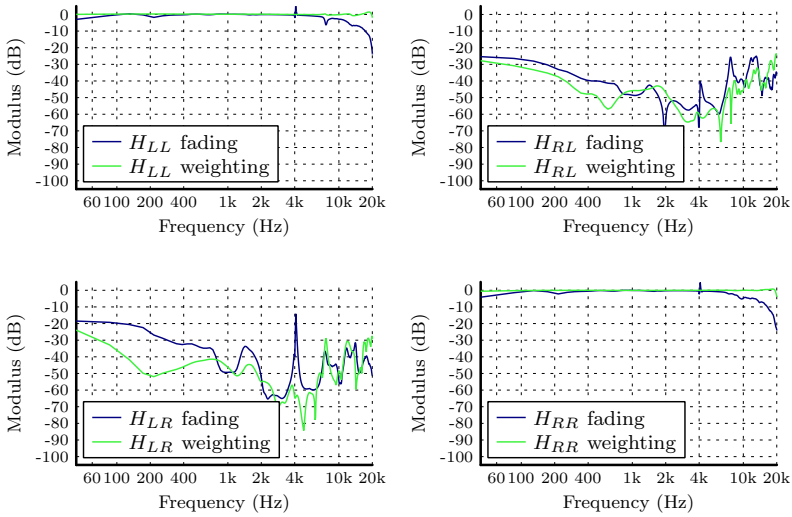


Figure 5.7: Frequency response of the complete transfer-path between the binaural signals and the ear signals using three loudspeaker for reproduction, calculated using both the fading strategy described in the work from Lentz (2007) with the truncated CTC filter calculation algorithm and the weighting strategy presented in this work with a frequency independent regularization parameter $\mu = 0.005$. Under ideal conditions, both results should be identical.

where $\boldsymbol{W} = \boldsymbol{Z}^{-1}$. With this notation, the smaller the weight w_{ii} applied at a loudspeaker, the lower the sound pressure that this loudspeaker is supposed to generate. Thus, switching or specific fading becomes obsolete.

Applying the causality constraint and causal regularization to eq. (5.33) yields

$$\boldsymbol{C}'_{\mathrm{mp}} = \frac{A^{+}_{\mathrm{mp}}(z)}{\det\left(\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{*}\right)^{+}} \left[\frac{A^{-}_{\mathrm{mp}}(z)\boldsymbol{W}\boldsymbol{H}^{*}\mathrm{adj}\left(\boldsymbol{K}\right) e^{-z\Delta}}{\det\left(\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{*}\right)^{-}} \right]_{+}, \tag{5.34}$$

where $\boldsymbol{K} = \left(\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{*} + \mu(z)\boldsymbol{I}\right)$ and $A(z)_{\mathrm{mp}}$ is the minimum-phase version of $A = \det\left(\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{*}\right) / \det\left(\boldsymbol{K}\right)$.

5.5 Discussion

This chapter presents a general framework for the calculation of dynamic crosstalk cancellation (CTC) filters to be applied to binaural reproduction in immersive VR environments using a dynamic CTC setup with multiple loudspeakers. Such setups require high filter update rates. This means that filter calculations are performed in the frequency-domain for higher efficiency.

Since a direct calculation in frequency-domain might yield noncausal artifacts, a causality constraint in the frequency-domain calculation is introduced to avoid undesirable wrap-around effects and echo artifacts. Regularization is commonly applied to the CTC filter calculation in order to limit the output levels at the loudspeakers, which also leads, as a side effect, to noncausal artifacts. These artifacts can be minimized through the proposed minimum-phase regularization. Even though extra calculation steps are added, the calculation time required by this framework is one order of magnitude faster than an equivalent calculation in time-domain for CTC filters with 512 taps and the advantage of frequency calculation tends to increase for larger filters.

Another aspect that is especially critical for dynamic CTC systems is the switch between active loudspeakers in the setup. The use of a weighted filter calculation allows the loudspeakers' contribution to be windowed in space, resulting in a smooth filter transition free of artifacts. Weights can be made frequency-dependent, allowing for a frequency-dependent choice of active loudspeakers (cf. Takeuchi and Nelson, 2007).

All filter calculation described so far assumed a priori knowledge of the transmission matrix to be equalized by the CTC system. As shown in section 6.2, realistic CTC systems will not deliver a channel separation (CS) that is as high as the one obtained using an ideal CTC system. Especially at high frequencies, the obtained CS is often lower than the natural channel separation $\widehat{\mathrm{CS}}$. Gardner (1997, pp. 65,77) already verified this deficiency of nonindividualized CTC systems and suggested that CTC should be used only at low and middle frequencies and that the binaural signal should be played directly via two loudspeakers at high frequencies. He achieved this by bypassing the CTC filters and only equalizing the direct path between loudspeaker and ipsilateral ear. The presented framework could be expanded to include *vector base amplitude panning* (VBAP) for high frequencies, allowing the binaural signal to be smoothly panned between the loudspeakers.

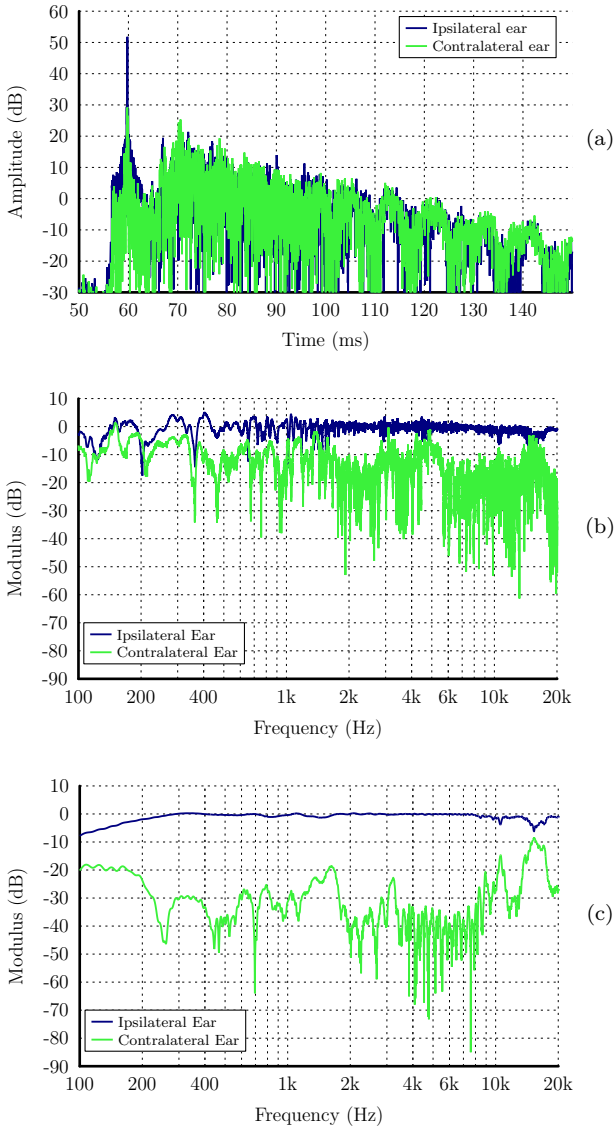


Figure 5.8: Response of a matched CTC system measured in a lightly reverberant room for the left ear of an artificial head. (a) Binaural IR, (b) spectrum of the complete binaural IR and (c) spectrum of the windowed binaural IR containing only the direct sound.

The framework introduced in this chapter does not take the presence of reflections in the reproduction room into account. However, in most practical applications CTC systems are built inside reverberant rooms.

The response of a matched CTC system measured in a lightly reverberant room ($T_{30} \approx 300\,\text{ms}$) can be seen in fig. 5.8. It is possible to observe in the binaural impulse response (fig. 5.8(a)) how the direct sound arriving at the contralateral ear is attenuated by over 20 dB while the room reverberation arrives at both ears with the same levels. The room reflections cause a drop in the observed CS obtained from the spectrum of the entire binaural IR (fig. 5.8(b)) when compared to the CS measured from the spectrum of the windowed binaural IR containing only the direct sound at both ears (fig. 5.8(c)).

Sæbø (2001) studied the influence of room reflections on the localization performance delivered by a nonindividualized CTC system and concluded that room reflections can severely degrade the localization performance. Moreover, he argues that "(it could not be) shown that purely anechoic conditions are necessary (for CTC reproduction). It may well be the case that playback under nearly 'normal' conditions will be acceptable for many applications, but care and thoughtfulness should definitely be exercised."

The room reflections could be canceled by using room binaural responses instead of the HRTF in $\boldsymbol{H}$. This is, however, a more fragile process than crosstalk cancellation alone (Sæbø, 2001). An interesting workaround to this problem is described in a publication by Jungmann et al. (2012) who propose a method to calculate CTC filters that are robust to (small) head displacement and room reflections. They take advantage of the masking effects of the human hearing system to design the filters so that the reflections contained in the binaural response of the system are below the masking threshold caused by the direct sound.

6

Perceptual Evaluation

This chapter presents two listening tests on two distinct aspects of individualized binaural technology. The first experiment, described in section 6.1, studies the plausibility of binaural reproduction via individually equalized headphones (BRvIEH) while the second experiment, described in section 6.2, evaluates the human sound localization performance using individualized and nonindividualized CTC systems.

Experiment I is divided into two parts: a direct and an indirect comparison of the original source, a loudspeaker, and the equivalent binaural auditory display. All 40 subjects participated in both parts of this experiment. The direct comparison was conducted as a three-alternative forced-choice test. Three stimuli were used for this test: noise, speech, and music. Results indicate that at least 50% of all listeners could not distinguish between the auditory event generated from the original source and the auditory event generated from the BRvIEH when presented with a speech or music stimulus. On the other hand, the majority of the subjects could hear a difference in the reproduction method when the presented stimulus was a pulsed pink noise. Further analysis confirmed that the observed difference in error rate between noise and the other two stimuli is significant.

An indirect comparison of the two reproduction methods was carried out in the second part of this experiment where the listeners were asked to say whether the presented stimulus originated from the headphones or one loudspeaker. In such a comparison listeners are less sensitive to differences, as no reference is provided. The pulsed pink noise was the only stimulus used for this test. Results show that no listener was able to distinguish between the original source and the BRvIEH. Furthermore, participants chose the loudspeaker more often than the headphones, which shows the authenticity of the auditory display generated using the BRvIEH.

The second experiment was aimed at testing the sound localization performance using individualized matched, individualized but mismatched, and nonindividualized crosstalk cancellation (CTC) systems.

[§]The results from experiment I were extracted from a broader study on selective auditive attention, which is described in greater depth in (Oberem, 2012).

The individualized matched and individualized mismatched systems were based on two different sets of listener-individual HRTFs. Both sets provided similar binaural localization performance in terms of quadrant errors, polar and lateral errors, suggesting that human sound localization is robust to the HRTF measurement variations—at least to the variation levels observed when using this HRTF measurement setup. The individualized matched CTC system provided performance similar to that from the binaural listening. The localization performance deteriorated when stimuli were presented with the individualized mismatched CTC system and the errors increased even further when the nonindividualized mismatched CTC systems (based on HRTFs of other listeners) were used.

A direction-dependent analysis showed that mismatch and lack of individualization yielded a degraded performance for targets placed outside of the loudspeaker span and behind the listeners. The channel separation (CS) was also analyzed regarding its quality as a predictor for localization performance using CTC systems. The results indicate that CS might be indeed useful when it comes to evaluating mismatched CTC systems with respect to the horizontal plane localization, but a generally weak correlation was observed between the CS and the sagittal plane localization performance.

¶ The virtual reality facility designed for localization tests at the Acoustics Research Institute (ARI) of the Austrian Academy of Science was used for this second experiment. The experiment was designed during a research stay of the author at ARI. To minimize costs, the HRTF measurement setup of ARI was used for this tests instead of the setup described in chapter 3.

‖ The results from experiment II presented in this work have been submitted to publication in Majdak; Masiero, and Fels (2012).

6.1 Experiment I: Authenticity of Binaural Reproduction via Individually Equalized Headphones

The overall quality of binaural reproduction will be influenced by aspects such as similarity of the synthesis HRTFs and the listener's own HRTFs, adequate ambient simulation, compensation of listener's head movements, and adequate sound source equalization.

Localization performance is commonly investigated as an indicator for the quality of binaural auditory displays. Wightman and Kistler (1989) conducted listening tests comparing localization accuracy between real sources and binaural presentation. Even though error rates grew in elevation for binaural reproduction, they stated that the "appropriately synthesized stimuli presented over headphones are judged to have the same spatial positions as stimuli presented in free field." Bronkhorst (1995) also conducted listening tests on this topic. His findings showed that "virtual sound sources can be localized almost as accurately as real sources, provided that head movements can be made and that the sound is left on sufficiently long." He mentions, however, that stimuli containing considerable energy in high frequency produced poorer performance, probably caused by inadequate hardware. Møller et al. (1996) also conducted a similar localization test comparing the localization performance with real sources, binaural reproduction via headphones using individual HRTFs and also non individual HRTFs, concluding "that individual binaural recordings are capable of giving an authentic reproduction for which localization performance is preserved when compared to that of real life." All these studies were conducted using individually equalized headphones and, apart from one test in Bronkhorst's experiment, the stimulus was always presented in a static manner.

Localization is, however, not the only important aspect of a plausible virtual acoustic scene. An auditory event (AE) slightly displaced in relation to the original source can still generate a convincing auditory impression, as long as the AE is well externalized and no strikingly unnatural sound coloration is perceived. Thus, the authenticity of the played scene can be assumed to be a major criterion for a successful binaural reproduction and it is therefore important to examine whether the binaural reproduction can be perceptually distinguished from a real source.

The authenticity can be studied by verifying whether an AE generated by a binaural reproduction via individually equalized headphones (BRvIEH) can be distinguished from an AE generated by the original sound event. The aim of this experiment is thus to analyze if and when the BRvIEH can be distinguished from the original sound source. To achieve the best possible conditions for a plausible binaural reproduction, HRTFs and HpTFs are measured individually and listeners are required not to move their heads during the presentation. As the loudspeaker stimuli are presented to the listeners via open-type headphone, HRTFs are also measured with the listener wearing the same headphones.[1]

6.1.1 Methods

Subjects

A total of 40 listeners participated in this study. All of them stated that they have normal-hearing (no hearing test was conducted) and participated voluntarily in the experiment. All listeners were nonexpert listeners and did not receive any training to improve their localization skills. The study was performed as a blinded experiment, i.e., none of the listeners were the authors and the listeners were not enlightened as to the nature of the experiment.

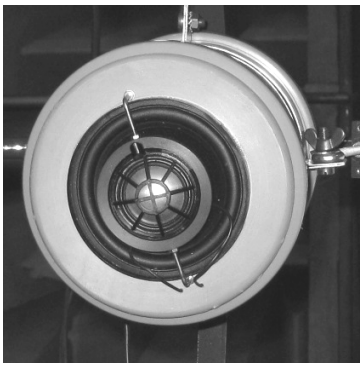
HpTF and HRTF Measurements

Both HpTF and HRTF were measured individually for each listener. The HpTFs were measured eight times including a repositioning of the headphones after every measurement to allow the calculation of a robust headphone equalization filter, as described in section 4.3. An exponential sweep from 100 to 20 kHz lasting 1.73 s was used to measure each HpTF.

The HRTFs were measured at 24 loudspeaker positions (cf. section 6.1.1). The positioning of the listener's head was continuously tracked during the measurement to ensure that the listener remained still during measurement. Listeners wore the open-type headphones during the whole procedure, thus the influence of the headphones on the incoming sound field is already contained in the HRTFs.

[1] Even though the used headphones are of open-type, they do add a considerable coloration to the sound arriving at the listeners ears. A possible workaround would be to use custom-made tube-phones, as the ones described in (Kulkarni and Colburn, 1998).

(a)



(b)

Figure 6.1: Transducers used for HRTF measurement. (a) Miniature microphone fixed with ear plug in ear of participant and (b) one of the 24 custom-made coaxial loudspeakers.

The used excitation signals were exponential sweeps. The same exponential sweep used for the HpTF measurement was used, resulting in a measurement duration of less than 1 min. The influence of the transducers was removed by free-field equalization, cf. section 3.3.1.

The impulse responses of all HpTFs and HRTFs were windowed with an asymmetric Tukey window (fade out of 1 ms) to a 5 ms duration.

Microphones were fixed with an ear plug at the entrance of the listener's ear canals. To ensure a perfect fit of the microphone, the ear plug was shortened in length to be flush with the entrance of the ear canal (fig. 6.1(a)). The microphones' output signals were directly recorded via custom-made pre-amplifiers by the digital audio interface.

Apparatus and Procedure

Even though this experiment did not focus on sound localization, 24 different target positions were used to allow listeners to also use directional cues for their discrimination task. Eight loudspeakers were distributed every 45°, cf. fig. 6.3(a), along three different elevations: −30°, 0° and 30°, cf. fig. 6.3(b). All loudspeakers were placed at a distance of 2 m from the participant fig. 6.4). The target directions were randomly chosen for every new run.

The metal structure holding the loudspeakers was installed inside a fully anechoic chamber. The listener sat on a chair (with a back

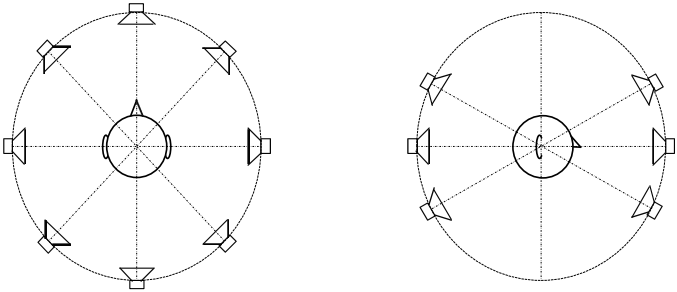
Figure 6.2: Circumaural open-type headphones HD-600 from Sennheiser used for tests presented in this section.

rest, arm rests and an adjustable head rest) placed in the middle of the construction. To minimize the movements the head rest is adjusted for the comfort of every single participant. To take the focus from the visual to the aural sense, lights were turned off during the listening test (cf. Blauert, 1997; Moore, 2012).

The acoustic stimuli were generated using a computer with the ITA-Toolbox[2] at sampling rate of 44.1 kHz and output via an external sound card (Presonus Light pipe) connected to four analog digital converters (ADA 8000, Behringer) with 8 channels each. Twenty-four channels are linked to two custom-made power amplifiers, and further fed to the 24 custom-made coaxial loudspeakers, with a woofer for the frequency range between 100 Hz to 5 kHz and a tweeter for between 3 to 20 kHz, attached in front of the woofer (fig. 6.1(b)). Two channels are linked to the ROBO frontend and further fed to a pair of circumaural open-type headphones (HD-600, Sennheiser). Finally the two miniature microphones (KE-3, Sennheiser), provided with a preamplifier each, were also connected to AD converter.

The level of the presented stimuli was 63 dB_{SPL} for both headphone and loudspeaker presentation. The stimulus' level for each presentation

[2]The ITA-Toolbox is a full-fledged MATLAB toolbox. The author participated in its design and development. It is available as an open source project at http://ita-toolbox.org/



(a) (b)

Figure 6.3: Schematic distribution of loudspeakers used in experiment I distributed in (a) the horizontal plane and (b) the median plane.

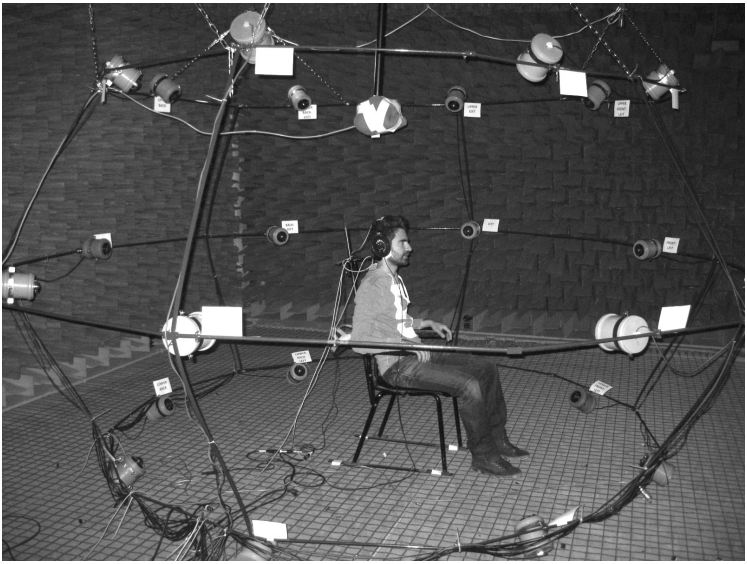


Figure 6.4: Participant sitting in a chair with headrest placed in the center of the structure used to hold the loudspeakers during the listening test.

Translation	Rotation	Application
±10 mm	±2°	during measurement and playback of stimulus
±20 mm	±4°	between measurement and listening test, between input and playback

Table 6.1: Limits of tolerated movements during measurement and listening tests of experiment I.

was randomly roved within the range of ±5 dB.

A tracking sensor (Patriot, Polhemus) was mounted on the headphones' headband, which captured the position and orientation of the head in real time. The tracking data was used to monitor whether listeners remained still during measurement and signal presentation and whether they kept their head at the same position throughout all stages of the experiment. While the participant is not allowed to make greater movements than ±10 mm in translation and ±2° in rotation during the presentation of the stimulus or the measurement, the limits of movement are twice as big for the participant to find the correct position after making a decision or between measurement and listening test (cf. table 6.1).

During the test participants gave their response using a touchscreen tablet (ThinkPad Tab, Lenovo). This tablet was wirelessly connected to the main computer and acted as an extended monitor displaying a graphical user interfaces (GUI) controlled MATLAB. The participants were asked to keep the tablet on their knees during stimulus presentation and to hold it up only while entering their results. During the stimulus presentation the screen was turned off.

Test Description

Experiment I consisted of two different examinations, both with the same aim of investigating naturalness, authenticity and plausible of binaural reproduction via headphones. In the first part of this experiment a real and a synthesized stimulus were directly compared while in the second part an indirect comparison was carried out.

The whole test procedure was conducted in one section and took approximately 40 min including a break between measurements and the listening test.

Part I The first part of this experiment was a three-alternative forced choice (3-AFC) direct comparison test. The condition tested was whether a difference between a stimulus played by loudspeaker and BRvIEH could be heard. Listeners wore headphones throughout the whole test. As the loudspeaker condition was heard through the open-type headphones, the HRTFs used for synthesize the binaural signals also contained the effect of headphone attenuation (cf. section 6.1.1). Even though the attenuation of approximately 10 dB observed for frequencies above 2 kHz could influence the results, a comparison between the systems would otherwise not be possible. Before the experiment, listeners were not instructed about the kind of differences they should be paying attention to and also did not have a training phase

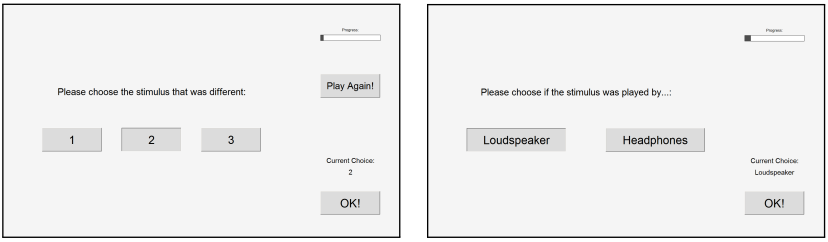
Three stimuli were used for this test. The first stimulus was an anechoic recording of the spoken German word "Wunschdenken" with a duration of 0.8 s. This stimulus was band limited between 200 Hz and 8 kHz. The second stimulus was a music sample, with a duration of 1.8 s. This stimulus was also band limited between 200 Hz and 10 kHz. The last stimulus was a pulsed pink noise covering the frequency range from 200 Hz to 20 kHz and with a duration of 0.8 s (each pulse had a duration of 0.3 s with a fade in and fade out of 50 ms).

Each participant listened to 20 sets of stimuli. For each set the stimulus type, reproduction combination and target direction were all chosen randomly. Furthermore, the level was roved, as explained in the previous section. After hearing a set of stimuli (up to three times), the listener had to decide which of the three presented stimuli was different than the two others and give his answer on a GUI displayed on the tablet, as shown in fig. 6.5(a).

Again, participant's head movements were observed throughout the whole test and in case they exceed the defined limits, the presented set of stimuli was considered invalid and repeated at the end of the test.

Part II In the second part of this experiment an indirect comparison test was carried out. The stimulus is reproduced either using a loudspeaker or a BRvIEH and the listeners' task is to decide whether the sound event was generated by the loudspeaker or the headphones. Part II did not include a training phase as well.

Only one stimulus type was presented in this part, namely the pulsed pink noise (described above). Every listener was presented with five stimuli via headphones and five stimuli via loudspeakers, all played in random order. Head tracking, level roving and randomized target



(a) (b)

Figure 6.5: GUI design: selection menu used in experiment I for (a) the direct comparison in part I, with the instruction "please choose the stimulus that was different", and (b) the indirect comparison in part II, with the instruction " please choose if the stimulus was played by... (Loudspeaker/Headphone)".

direction are handled as in part I. After hearing each stimulus (played only once), the listener had to decide whether the stimulus had been played from a loudspeaker or from the headphones and give his answer on a GUI displayed on the tablet, as shown in fig. 6.5(b).

6.1.2 Results

Direct Comparison: Sound Quality

Prior to the analysis of the results, the collected data was analyzed for consistency. Specifically, all measured HRTF and HpTF were examined regarding abnormal behavior, e.g. dips and peaks in the lower frequency range. From this analysis one participant had to be completely excluded from the study due to an inaccurate HpTF and some single sets were excluded from other participants due to inaccurate HRTF.

The 3-AFC test consists of presenting three stimuli in random order from which two stimuli are the same and one is different. The participants are supposed to answer which of the three presented stimuli is the different one (Macmillan and Creelman, 2005). If participants were only guessing, they would have a chance out of three to choose the correct answer at each presentation, i.e., a 33.33% hit rate. In other words, a percentage of 66.67% wrong answers is expected when the subjects are guessing and therefore did not hear any difference. In case participants answered for one out of three times incorrectly (33.33%), they could

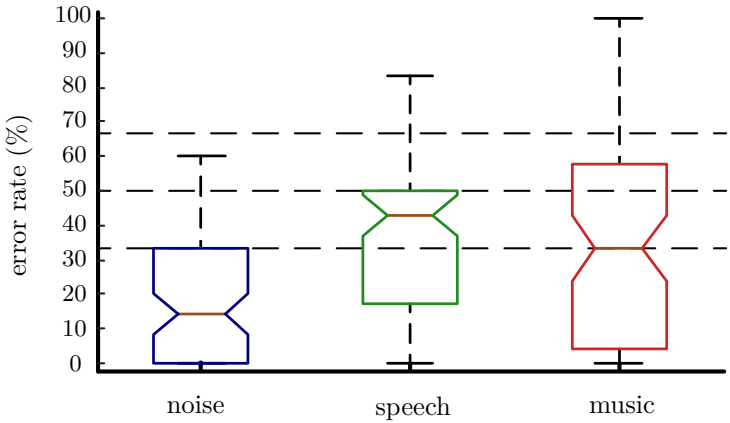


Figure 6.6: Box plot showing distribution of error rate (normalized by the total error rate of each condition) among participants in the 3-AFC discrimination test between loudspeaker reproduction and BRvIEH for three stimulus condition: noise, speech, and music.

not hear a difference for 50% of the presented stimuli. In this case, in relation to all subjects, it can be said that 50% of all listeners did not hear any difference.

Figure 6.6 shows the boxplot results for all participants and all presented stimuli. These results indicate that most of the subjects could hear a difference in the reproduction method when the stimulus was a pink noise. In numbers, 16.74% (38 out of 227) of all sets of pink noise stimuli were not answered correctly. The music stimulus presented an error rate of 35.10% (73 out of 208), slightly higher than the 33.33% limit. Therefore, at least 50% of all listeners could not distinguish between the reproduction methods. For the speech stimulus even more subjects were not able to hear any difference with an error rate of 38.17% (92 out of 241). An ANOVA with the factor condition at the three stimuli was performed. The results for music and speech were significantly different than the ones for noise ($F = 10.77$, $p < 0.001$).

The obtained data was also analyzed according to playing level and target direction. As each subject was presented with only 20 stimuli, there was in average less than two stimuli for each playing level and less than one stimulus for each target direction, making it impossible to conduct an analysis of variance. Nevertheless, simple observation of the bar chart (fig. 6.7) displaying the percentage of wrong answers (normalized by the frequency each stimulus was presented at each condition) indicates no

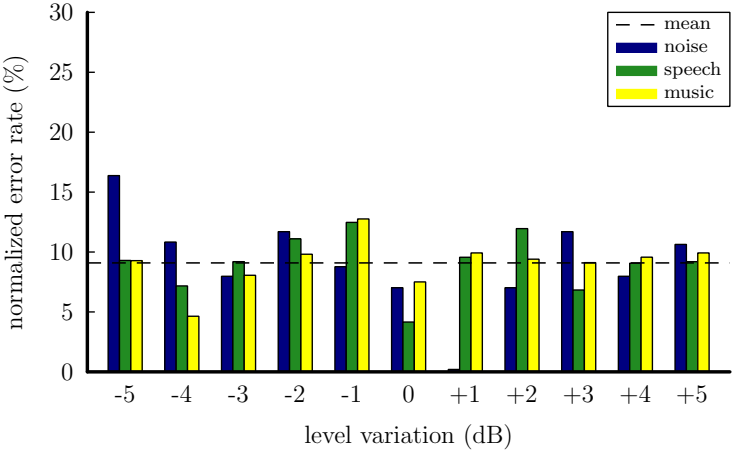


Figure 6.7: Histogram showing the distribution of errors in the 3-AFC discrimination task over different stimulus presentation level, further subdivided into the three stimulus condition: noise, speech, and music. The error rate was normalized by the frequency each stimulus was presented at each condition.

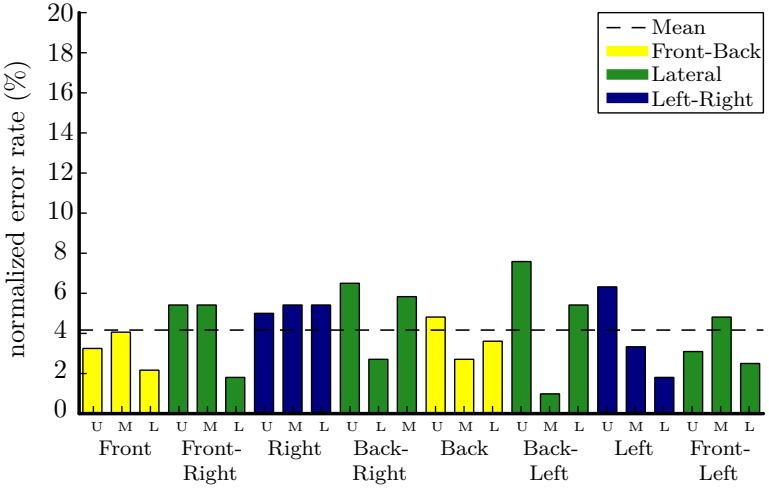


Figure 6.8: Histogram showing the distribution of errors in the 3-AFC discrimination task over different target directions for *speech* as stimulus condition. The error rate was normalized by the frequency each stimulus was presented at each condition. U stands for the *upper*, M the *middle* and L the *lower* loudspeakers.

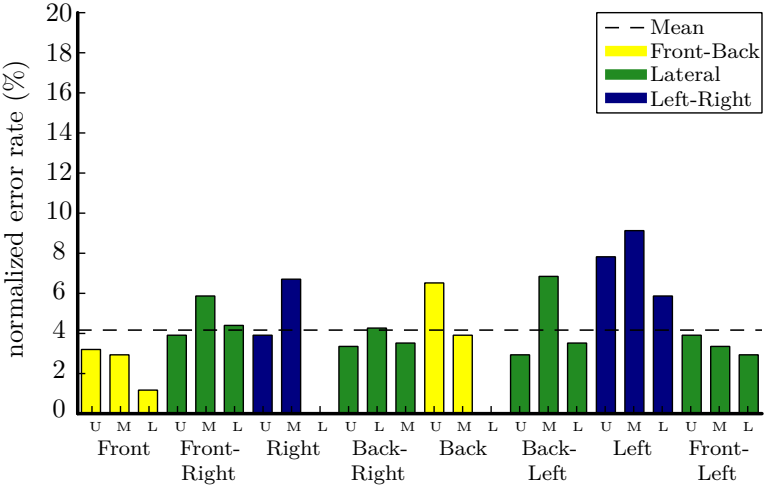


Figure 6.9: Histogram showing the distribution of errors in the 3-AFC discrimination task over different target directions for *music* as stimulus condition. The error rate was normalized by the frequency each stimulus was presented at each condition. U stands for the *upper*, M the *middle* and L the *lower* loudspeakers.

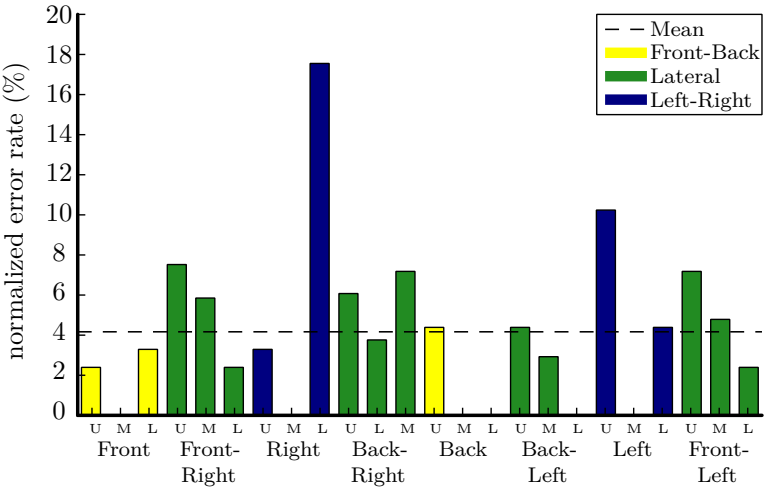


Figure 6.10: Histogram showing the distribution of errors in the 3-AFC discrimination task over different target directions for *noise* as stimulus condition. The error rate was normalized by the frequency each stimulus was presented at each condition. U stands for the *upper*, M the *middle* and L the *lower* loudspeakers.

significant difference between playing levels for all three stimuli. The same can be said about the error distribution according to the target direction, exhibited by the bar charts in figs. 6.8 to 6.10, where no significant difference between directions can be observed.

Indirect Comparison: Loudspeakers VS Headphones

Besides the direct comparison conducted in part I, an indirect comparison between loudspeaker reproduction and binaural synthesis reproduced via headphones was also made.

In a two forced-choice test as is this second part, a total error rate of 50% indicates that subjects are guessing at all times (Macmillan and Creelman, 2005). As the observed total error rate is 51.03% (199 out of 390) with a median of 50% (cf. fig. 6.11(a)), it can be assumed that no subject was able to distinguish between the reproduction methods.

Figure 6.11(b) shows that participants chose the loudspeaker (63.25%) as the reproducing method more often than the headphones (36.75%). While 32% of all stimuli presented by real sources are answered correctly, a rate of only 18% is observed for the BRvIEH.

An analysis according to the playing level was conducted. As for part I, since only 10 stimuli were presented to each participant, there is not sufficient statistical data for an analysis of variance. At a first glance, the bar chart of the total error rates indicates no significant difference in answers at each playing level.

However, an interesting result can be obtained whenever the error rate is further subdivided according to the condition reproduction method. Figure 6.12 shows the percentage of wrong choices normalized by the frequency each level was presented for each reproduction method. The chart indicates that listeners chose more often the loudspeakers as the reproduction method when the stimulus was presented with lower levels while they more often chose the headphones when the stimulus was presented with higher levels. Many listeners reported that, as no reference stimulus was presented, they did not know how to categorize the auditory event and, since they could not observe any other differences, ended up relying on the presentation level to make their decisions.

6.1.3 Discussion

In the first part of this experiment a direct comparison between loudspeaker reproduction and binaural synthesis reproduced via headphones

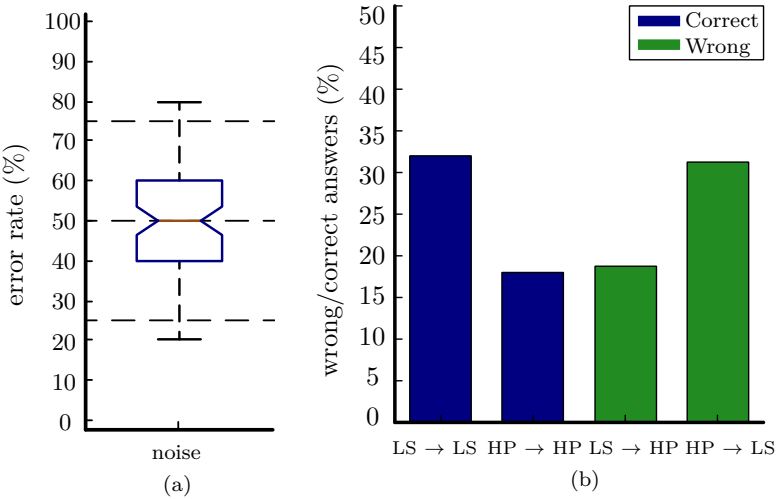


Figure 6.11: Results of the indirect discrimination task. (a) Box plot showing the distribution of the error rate among participants. (b) Histogram showing the distribution of wrong and correct answers for the four combinations of actual and perceived reproduction methods.

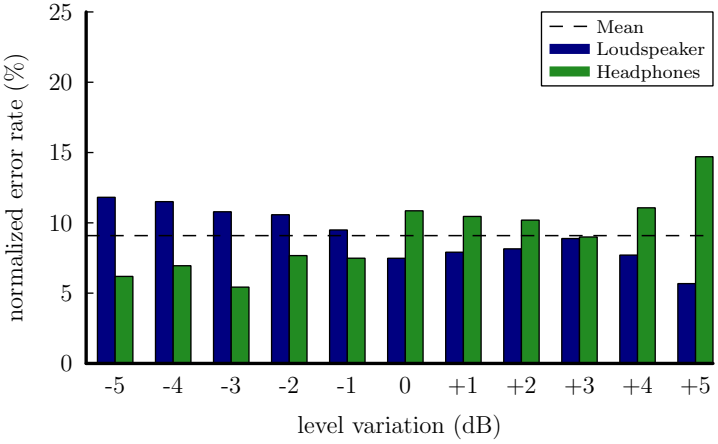


Figure 6.12: Histogram showing the distribution of errors in the indirect discrimination task over different stimulus presentation level, further subdivided into the two presentation method condition: loudspeaker and BRvIEH. The error rate was normalized by the frequency each level was presented for each reproduction method.

was also made using three different stimuli: pulsed pink noise, a speech sample, and a music sample. Results show that different numbers of subjects were able to distinguish between BRvIEH and the original sources according to the type of stimulus. This suggests that the power spectral density (PSD) of each particular stimulus plays a role in the differentiation task. While pink noise is a broad-band signal, speech is quite narrow-banded and music exhibits a PSD dominant in the range between 400 Hz and 2 kHz. It could be shown that listeners committed significantly more mistakes differentiating the reproduction methods when the played stimulus was either speech or music. This suggests that stimuli with dominant high frequency components tend to allow an easier distinction between reproduction methods.

After the listening test subjects were questioned regarding the differences they heard during the test. Most of them reported that for pink noise a different coloration was audible in higher frequencies, facilitating the discrimination task. Since headphone equalization is especially critical for frequencies higher than 4 kHz, these observations made by the listeners are reasonable.

A few subjects (2 out of 39) reported that they could hear a difference in source distance—it was not possible to verify if the sources perceived as closer were reproduced with the headphones. Reasons for this difference between BRvIEH and the original sources could be an inaccurate equalization. Further test would be required to specifically analyze the influence of headphone equalization in the perception of distance.

The most common type of difference reported by the listeners was a variation in perceived source direction of some degrees. Again, these observations are appropriate as the head displacement between measurement and reproduction could be greater than the localization blur (Blauert, 1997) and thus the head movement limitation implemented with the head tracking device was not strict enough. To eliminate the influence of the head displacement either the listeners' head should be fixed in a tighter manner, causing discomfort to the listener, or a dynamic binaural synthesis should be employed, a system with increased complexity and prone to new error influence.

In the second part of this experiment an indirect comparison between loudspeaker reproduction and binaural synthesis reproduced via headphones was made. This test was conducted with only one stimulus: the pulsed pink noise, which is expected to be the stimulus that will make listeners most sensitive to the differences between the two reproduction

methods. However, results showed that not a single listener was able to consistently identify whether the auditory event was generated by a loudspeaker or by the headphones.

Since subjects were not able to find differences for this stimulus, it can be assumed that subjects will also not be able to distinguish between real sources and BRvIEH for stimuli like music and speech.

The fact that the listeners could not tell these two reproduction methods apart indicates that the BRvIEH sounded natural and authentic, i.e., even though differences between the two methods can be heard when stimuli possessing significant content in high frequencies are played, these differences are not big enough to allow listener to perceive that the auditory event binaurally reproduced via individually equalized headphones is actually coming from the headphones and not from the external source.

6.2 Experiment II: Localization Performance with Individualized and Nonindividualized CTC Systems

As discussed in chapter 5, CTC filters are calculated based on the transfer paths between loudspeakers and listener's ears, i.e., the HRTFs. So far, it has been assumed that exactly the same HRTFs are used for the filter calculation and the listening situation, a so-called *matched* CTC system, which provides optimal crosstalk cancellation. In a *mismatched* CTC system, the HRTFs do not exactly match the CTC filters and the performance is assumed to degrade. The actual localization performance of a CTC system has already been investigated in the horizontal plane (Gardner, 1997; Takeuchi et al., 2001; Bai and Lee, 2006; Lentz, 2006), however, little is known about the localization performance in *both* horizontal and sagittal planes provided by a CTC system.

Akeroyd et al. (2007) used HRTFs from other listeners to create mismatched CTC systems and compared their numeric performance, based on the obtained channel separation (CS) with the matched CTC systems. In a simulation of binaural processing, they showed disrupted ITDs and ILDs for the mismatched CTC systems. Through their simulation results they concluded that the mismatched system will probably yield a degraded lateral localization performance, particularly for directions with a high value for the lateral angle α.

Even though Akeroyd et al. (2007) used listener-individual HRTFs to create a matched CTC system, the listener-individual HRTFs do not always yield a matched CTC system. For example, if the HRTF measurement is repeated for the same listener, the HRTF set will still be considered as listener-individual, but acoustic properties of the HRTFs would slightly change, causing a mismatch to the CTC filters. This is actually a common situation, even in individualized CTC systems, where the propagation paths change between the HRTF measurements and the actual use of the CTC system.

Thus, the aim of this listening test is to investigate two-dimensional human localization performance in CTC systems with a special focus on individualized matched, individualized but mismatched, and nonindividualized CTC systems. The individualized but mismatched CTC systems used a second HRTF measurement of the same listeners. The nonindividualized CTC systems used HRTFs from a mannequin and other listeners. Also, the baseline performance was acquired for binaural sound presentation without any CTC filtering.

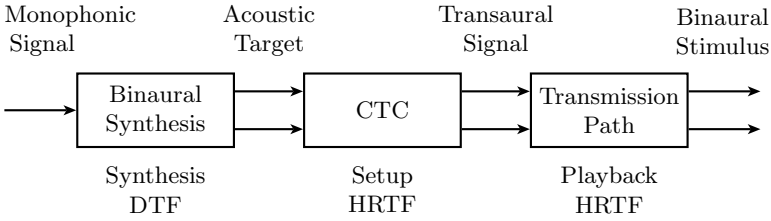


Figure 6.13: Block diagram for the signal processing conducted for the preparation of experiment II. A monophonic signal is first filtered by the individual DTF at the binaural synthesis stage. The resulting acoustic target is further filtered by the CTC filter to generate the transaural signals, which are further filtered by the individual HRTFs corresponding to two loudspeakers, resulting in the binaural signals presented to the listener.

Channel separation (see section 5.2) is commonly used to describe the quality of a CTC system (Gardner, 1997; Bai and Lee, 2006). Akeroyd et al. (2007) showed that much smaller CSs are obtained in mismatched CTC systems compared with matched CTC systems. Recently, Parodi and Rubak (2011) investigated the minimum audible channel separation in an artificial CTC system. However, it is still not clear how the channel separation is related to the localization performance. Thus, a comparison between the channel separation values and the sound-localization performance in CTC systems is conducted to investigate its use as a predictor for the localization performance.

The channel separation, being a frequency-dependent measure, is usually averaged over a frequency range in order to describe a CTC system by a single value. Keeping in mind that different frequency regions contribute differently to the sound localization in the horizontal and sagittal planes, as discussed in section 2.4.2, it was investigated whether channel separation calculated in specific frequency regions better describes different aspects of the sound localization.

Current CTC systems usually suffer from various technical limitations (cf. chapter 5). The cancellation quality depends to a large extent on the listener's alignment between the loudspeakers (Takeuchi et al., 2001) and loudspeaker combinations have been proposed to increase the area of the sweet spot (Takeuchi and Nelson, 2007). Loudspeakers, usually simulated as point sources, have non-ideal transfer function and directionality, which also have a strong effect on the quality of the CTC systems (Qiu et al., 2009). Other potential artifacts are head movements and room reflections.

In order to better control the issues listed above, the localization performance was tested using a *virtual CTC system*. Thus, a binaural simulation of the CTC system (see fig. 6.13) was used, i.e., the stimulus was presented via headphones and individualized HRTFs were used to simulate the propagation paths between loudspeakers and the listener. The virtual CTC system consisted of three different filter stages: 1) listener-individual DTFs, used to create an acoustic target; 2) CTC filters, used to create the transaural signals for the virtual loudspeakers, and; 3) listener-individual HRTFs, used to simulate the virtual loudspeakers. In such a setup, the loudspeaker effects were reduced to that of the HRTF measurement and the listeners' head was always virtually fixed within the sweet spot.

6.2.1 Methods

Subjects

Eight listeners participated in this study,[3] all of them having absolute hearing thresholds within the 20 dB range of the average normal-hearing population in the frequency range between 0.125 and 12.5 kHz. All listeners who participated in this test had participated in previous test and had therefore previous experience with localization tests. They all showed front-back confusion rates below 20% in pre-experiments with their own broadband DTFs. As experiment I, this study was also performed as a blinded experiment. All participants received a financial compensation for taking part in this listening test.

HRTF Measurements

HRTFs were measured individually for each listener. The measurement setup at ARI is also composed of a supporting arc, containing 22 loudspeakers (custom-made boxes with VIFA 10 BGS as drivers) at fixed elevations from -30° to 80°, with a 10° spacing between 70° and 80° and 5° spacing elsewhere. The listener was seated in the center point of the circular arc on a computer-controlled rotating chair. The distance between the center point and each speaker is 1.2 m. Miniature

[3] Compared to exp. I, the number of subjects in exp. II might seem insufficient for an adequate statistical analysis. However, in exp. I each listener was tested for only one condition and did relatively few repetitions. Therefore, many listeners had to be pooled together. Exp. II compares the performance of each individual listener, conducting many repetitions for each condition. Thus, in this case, eight subjects can be considered a satisfactory sample size.

microphones (Sennheiser KE-4-211-2) were inserted into the listener's ear canals and their output signals were directly recorded via amplifiers (FP-MP1, RDL) by the digital audio interface.

The used excitation signal was a multiple exponential sweep, cf. section 3.2.1. A 1.73 s exponential frequency sweep from 0.05 to 20 kHz was used to measure each HRTF. At an elevation of $\theta = 0°$, the HRTFs were measured with a horizontal spacing of $\Delta\phi = 2.5°$ within the range of $\phi = \pm\ 45°$ and with the horizontal spacing of $\Delta\Phi = 5°$ otherwise. According to this rule, the measurement positions for other elevations were distributed with a constant spatial angle, i.e., the azimuthal spacing increased towards the poles. In total, HRTFs for 1550 positions within the full 360° horizontal span were measured for each listener. The measurement procedure lasted approximately 20 minutes. As described in section 3.3.1, first the influence of the transducers was removed by equalizing the HRTFs. Then the directional transfer functions (DTFs) were calculated. Finally, the impulse responses of all HRTFs and DTFs were windowed with an asymmetric Tukey window (fade in of 0.5 ms and fade out of 1 ms) to a 5.33 ms duration.

Two sets of HRTFs were measured for each listener.[4] The first measurements were performed for a previous study—all current participants took part at this previous study—and the second measurements were performed for the present study. The interval between the two measurements was approximately five years.

Acoustic Targets

Lateral and polar angles from the horizontal-polar coordinate system (see fig. 2.6) were used to describe the acoustic target's position (Morimoto and Aokata, 1984). The tested lateral angle ranged from $\alpha = -90°$ (right) to $\alpha = 90°$ (left). The polar angle of the targets ranged from $\beta = -30°$ (front, below eye-level) to $\beta = 210°$ (rear, below eye-level). The targets were pseudo-uniformly distributed on the surface of the sphere by using a uniform distribution for the polar angle and an arcsine-scaled uniform distribution for the lateral angle.

The acoustic targets were Gaussian white noises with a duration of 500 ms and 10 ms fade-in and fade-out, filtered with the listener-specific DTFs. Prior to filtering, the position of the acoustic target was discretized to the grid of the available DTFs.

[4] The measured HRTFs and DTFs are available at http://www.kfs.oeaw.ac.at/hrtf. The listeners are referred throughout the work by the same anonymous identification number used on the online database. NH stands for *normal hearing*.

The level of the presented stimuli was 50 dB above the individual absolute hearing threshold in each condition. The threshold was estimated in a manual up-down procedure individually for each condition using an acoustic target positioned at lateral and polar angle of 0°. As in experiment I, the stimulus level for each presentation was randomly varied within the range of ±5 dB to reduce the possibility of localizing spatial positions based on overall level.

Binaural CTC Simulation

In the tested CTC conditions (see section 6.2.1), the acoustic targets were processed with a binaural CTC simulation. The simulation was used to ensure that subjects were always in the sweet-spot and to fully control the correspondence between the acoustic paths and CTC filters.

The CTC filters were calculated for a pair of virtual loudspeakers with one loudspeaker placed at $\phi = 45°$ left and second loudspeaker placed at $\phi = -45°$ right to the listener, both at $\theta = 0°$. Thus, the loudspeaker span angle was $\Delta\phi = 90°$.

The propagation paths from the loudspeakers to the listener's ears are described by the so-called "setup HRTFs". The corresponding impulse responses were zero padded to 85.33 ms.[5] The CTC filters were calculated in the frequency-domain according to eq. (5.24) with $\beta = 0.005$ for all frequencies.[6] The CTC filters were converted back to the time-domain and circularly shifted by 3.125 ms to avoid noncausality. Finally, the impulse responses where windowed with a one-sided Tukey window with a fade out of 18.6 ms at their end.

The transaural signals were calculated by processing the acoustic target with the CTC network according to eq. (5.4). Then, the transmission of the transaural signals from the loudspeakers to the listener's ears was simulated by filtering the transaural signals using the listener-individual HRTFs, the so-called "playback HRTFs". Note that listener-individual HRTFs were used for the playback HRTFs in all conditions—only the setup HRTFs were varied in this study.

[5] Note that 85.33 ms correspond to 4096 samples. All the signal processing calculations in this experiment were done at the sampling rate of 48 kHz.

[6] To allow comparison of this experiment's results with the results from Akeroyd et al. (2007), the CTC filters here were calculated in the same way as described by them. Therefore, causality constraint (section 5.3.1) and minimum-phase regularization (section 5.3.3) were not used.

Apparatus and Procedure

The virtual acoustic stimuli were presented via headphones (HD 580, Sennheiser) in a double-wall sound-proof room. The headphones were diffuse-field-compensated circumaural headphones and no additional headphone correction was applied,[7] as DTFs were used for the binaural synthesis (Larcher et al., 1998). The listener stood on a platform enclosed by a circular railing. Stimuli were generated using a computer and output via a digital audio interface (ADI-8, RME) with a 48 kHz sampling rate. A virtual visual environment was presented via a head-mounted display (3-Scope, Trivisio). It provided two screens with a field of view of 32° × 24° (horizontal × vertical dimensions). The virtual visual environment was presented binocularly with the same picture for both eyes. A tracking sensor (Flock of Birds, Ascension) was mounted on the top of the listeners' head, which captured the position and orientation of the head in real time. A second tracking sensor was mounted on a manual pointer. The tracking data were used for the 3-D graphic rendering and response acquisition.

The listeners were immersed in a spherical virtual visual environment (Majdak et al., 2010). They held a pointer in their right hand. The projection of the pointer direction on the sphere's surface, calculated based on the position and orientation of the tracker sensors, was visualized and recorded as the perceived target position. The pointer was visualized whenever it was in the listeners' field of view.

Prior to the tests, listeners performed a visual and an acoustic training. The aim of the visual training was to train subjects to perform accurately in the virtual environment. The visual training was a simplified game in the first-person perspective where listeners had to find a visual target, point at it, and click a button within a limited time period. This training was continued until 95% of the targets were found with a root-mean-square (RMS) angular error in the range of 2°. This performance was achieved within a few hundred trials. Then the acoustic training was performed with listener-individual DTF (Majdak et al., 2010). The goal of the acoustic training was to ensure a stable localization performance of the subjects. The acoustic training consisted of 6 blocks, 50 acoustic targets each, lasting approximately 2 hours.

[7] According to Schonstein et al. (2008), the impact of the headphone equalization on the binaural localization performance is still arguable. Unfortunately, their test was conducted with only one subject—who also designed the test—therefore lacking in statistical significance.

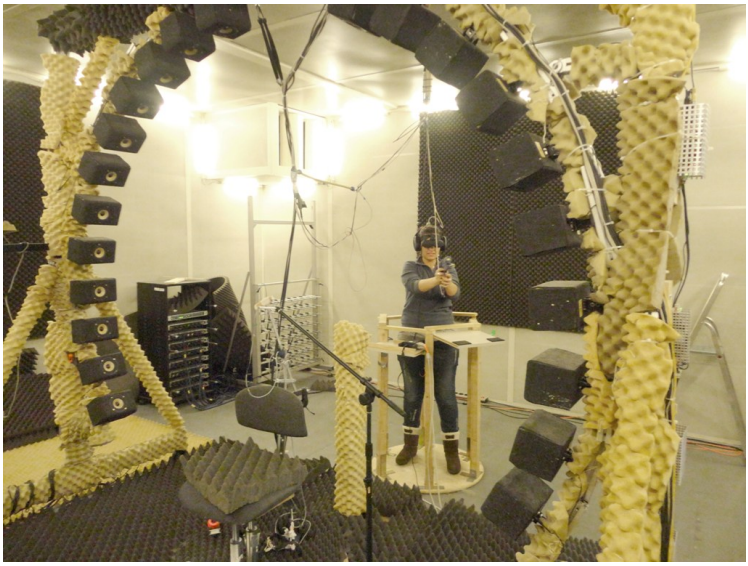


Figure 6.14: Overview of the anechoic chamber at ARI. In the front-plane the HRTF Measurement arc with its 22 loudspeakers and computer-controlled rotating chair. In the background the platform where the localization tests were conducted. A listener wearing the head-mounted displays holds the pointing device. The position of the listener's head and pointing device are tracked by a tracking device whose sender is placed in front of the test platform.

In the actual acoustic tests, at the beginning of each trial, the listeners were asked to align themselves with the reference position and click a button. Only after that the stimulus was presented. During the presentation, the listeners were instructed not to move, cf. section 2.4.2. The listeners were asked to point to the perceived stimulus location and click the button again. This response was recorded for the data analysis. The tests were performed in blocks; each block consisted of 100 acoustic targets and took approximately 15 minutes. Within a block, the targets are first randomly selected (cf. section 6.2.1) and then sampled to the nearest neighbor from the 1550 possible spatial positions. After each block, subjects had a break of approximately 15 minutes. The procedure was controlled by LocaCTC from the ExpSuite.[8]

[8] Available at http://sf.net/projects/expsuite.

Conditions

Eight conditions were tested in three blocks each. The order of the blocks was randomized in such a way that within eight blocks all conditions were in a randomized order.

The first two conditions consisted of pure acoustic targets, i.e., binaural signals without the CTC simulation. The former, *binOwn*, used the same DTFs as those used for the acoustic training while the latter, *binOwnB*, used the more recently measured DTFs.

In the individual matched CTC condition, *ctcOwn*, the acoustic targets were presented via the simulated CTC system using the same setup and playback HRTFs, namely the listener-individual HRTFs from the condition binOwn. The condition ctcOwn corresponds to the matched case from Akeroyd et al. (2007) and represents an *ideal individualized* CTC system where the CTC filters match exactly the acoustic paths between the loudspeakers and the listener.

In the individual but mismatched CTC condition, *ctcOwnB*, the playback HRTFs were the same as in the matched CTC condition, while the setup HRTFs were those from the more recent measurement, corresponding to the DTFs used for condition binOwnB. The condition ctcOwnB represents a *realistic individualized* CTC system where for the calculation of the CTC filters, the listener-individual HRTFs have been measured, but during the signal presentation, the acoustic propagation paths do not exactly match these measured HRTFs. Note that from an acoustic point of view, this condition is a mismatched condition.

The last CTC conditions were *nonindividual* mismatched conditions, i.e., the setup HRTFs were those from other sources, while the playback HRTFs did not change. In the condition *ctcKemar*, the setup HRTFs were those from measurements on a mannequin (Gardner and Martin, 1995). Note that in contrast to all other HRTFs used in study, the mannequin's HRTFs were measured using microphones included in an ear simulator, yielding an HRTF set containing the direction independent ear-canal transfer function. In the remaining nonindividual conditions, the setup HRTFs were those from other listeners, namely, NH57, NH64, and NH68. These particular listeners were also tested with setup HRTFs from NH12 in order to obtain the same number of tested conditions for each listener. Those conditions are referred to as *ctcNH57*, *ctcNH64*, *ctcNH68*, and *ctcNH12*. For the sake of simplicity, all nonindividual conditions are referred to as *ctcOther*.

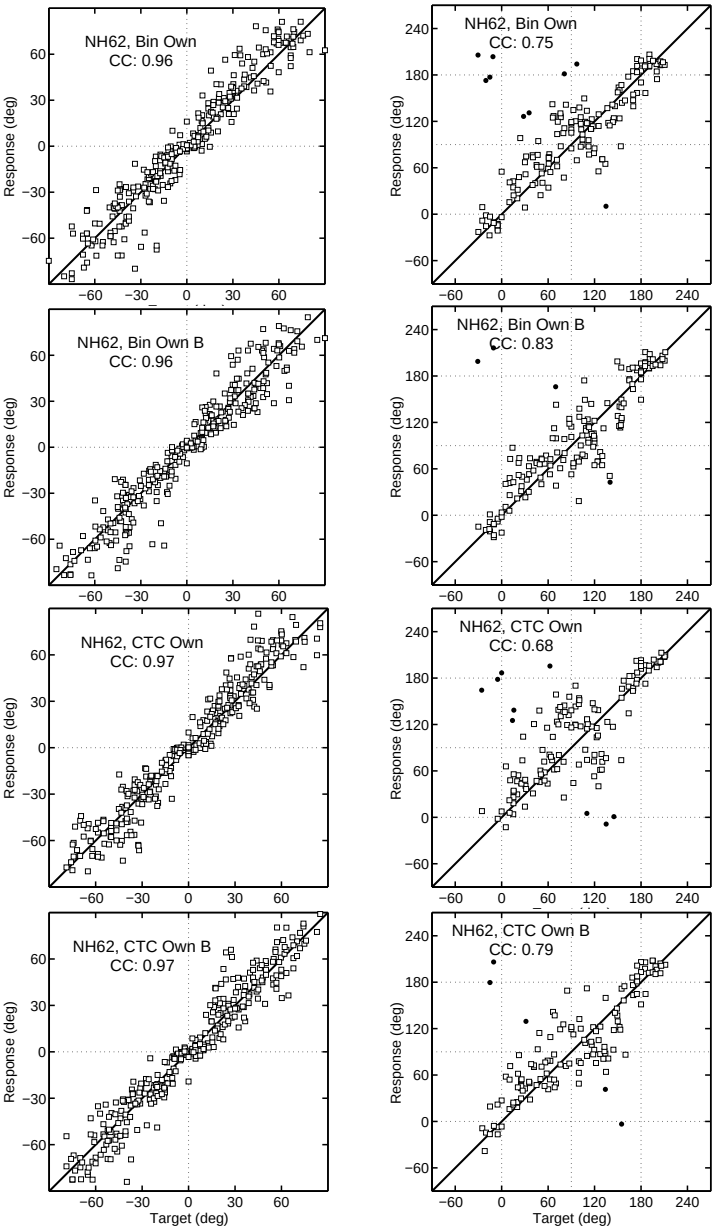


Figure 6.15: Localization results of the exemplary listener NH62 for all tested conditions. Lateral results are plotted in the left panels, the polar results in the right panels. Polar results outside

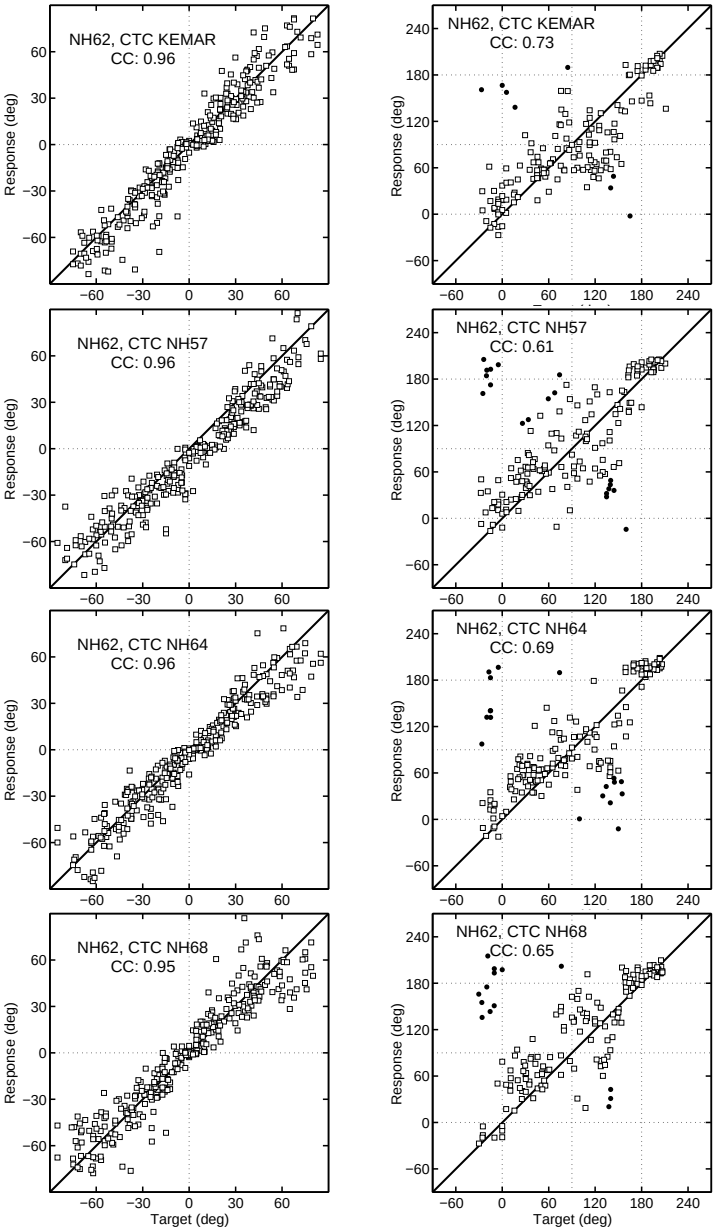


Figure 6.15: (cont.) the lateral range of ± 30° are not shown. Filled circles: Responses with errors outside the ± 90° range. CC: Correlation coefficient between responses and targets.

6.2.2 Results

Localization Performance: Binaural Reproduction

Figure 6.15 shows results of the localization experiment for an exemplary listener (NH64). The target and response angles are shown on the horizontal and vertical axes, respectively, of each panel. For the polar dimension, the results are shown for targets with lateral angles within ± 30° only. Responses that resulted in absolute polar errors larger than 90°, i.e. can be considered as quadrant errors, are plotted as filled circles. All other responses are plotted as open squares. The performance seems to be similar for both binaural conditions and the differences to the ctcOwn condition seem to be negligible. A generally degraded performance can be observed for ctcOwnB and also all other mismatched conditions.

As defined in section 2.4.2, localization errors were calculated by subtracting the target angles from the response angles. The lateral error (LE) is used to measure localization performance in the horizontal plane. In the polar dimension, data is analyzed in regard to confusions between the hemifields, measured with the quadrant error (QE), and the local performance within the correct hemifield, measured with the polar error (PE). Only responses within the lateral range of ± 30° were considered in the polar dimension analysis (Middlebrooks, 1999b).

The results described by the error metrics QE, PE, and LE are shown in tables 6.2, 6.3 and 6.4, respectively. In both binaural conditions, the group average performance was within the range of previously reported performance for localization of virtual broadband noises under comparable conditions. For the sagittal planes, the average QE of 8.6% (for binOwn, 7.1% for binOwnB) was similar to QE of 7.7% from Middlebrooks (1999b) and to QE of 9.4% from Goupell et al. (2010). Also, the average PE of 31.0° (31.6° for binOwnB) was similar to PE of 28.7° from Middlebrooks (1999b) and to PE of 33.8° from Goupell et al. (2010). For the horizontal planes, the average LE of 10.7° (10.4° for binOwnB) was similar to LE of 14.5° from Middlebrooks (1999b) and to LE of 12.4° from Majdak et al. (2011).

Repeated measures (RM) analysis of variance (ANOVA) was used for the statistical analysis of the results. Each of the three tested blocks was treated as within-subject repetition. For the binaural conditions, RM ANOVAs were calculated on the LE, PE, and QE with the factor condition at two levels (binOwn, binOwnB). The analysis showed neither significant effect for the QE ($p = 0.44$), nor for PE ($p = 0.68$), nor for

LE ($p = 0.38$). This indicates that despite of five years of break between the two HRTFs measurements, both HRTF sets provided localization performance at a similar level.

Localization Performance: Individualized CTC systems

In order to investigate the performance in individualized CTC systems, RM ANOVAs with the factor condition at four levels (binOwn, binOwnB, ctcOwn, and ctcOwnB) were performed. The results were significant for the QE ($p = 0.005$) and for the LE ($p < 0.001$), but not for the PE ($p = 0.20$). Tukey-Kramer *post-hoc* tests were used to test the statistical significance of particular levels. The significance was considered at $p < 0.05$. Post-hoc tests showed that only the ctcOwnB condition yielded significantly larger QE and LE compared to all other conditions. Note that even though for PE the differences were not significant, the PE was larger for ctcOwnB (34.2°) than for ctcOwn (31.8°) or the binaural conditions (31.0° and 31.6°).

The lack of significance in the differences between the ideal CTC and binaural systems indicates that the ideal CTC system used in this test provided localization performance at the level of the binaural reproduction systems. In a realistic application of the individualized CTC, this situation is, however, unachievable because the propagation paths would (slightly) change as soon as a listener leaves the HRTF measurement setup and enters the CTC system. This situation was represented by the condition ctcOwnB, where individual HRTFs from the latter measurement were used to calculate CTC filters. Note that this is not a *worst-case scenario* as head movements may induce stronger changes to the actual playback HRTF. The performance in such a realistic CTC system was worse than that in an ideal CTC system in terms of significantly larger QEs and LEs. This demonstrates that a mismatch between the playback and setup HRTFs may result in a degraded localization performance in a CTC system, even when both HRTFs provide a similar performance in a binaural system.

Compared to ctcOwnB, the ctcOwn condition yielded a better performance in the horizontal plane. This result confirms the results for modeling interaural differences in matched and mismatched CTC systems (Akeroyd et al., 2007) where for mismatched CTC systems, the model predicted large ITD and ILD errors.

Condition	NH12	NH14	NH15	NH57	NH62	NH64	NH68	NH72	Mean
binOwn	3.3	5.3	7.4	25.4	5.9	2.7	6.3	12.5	8.6
binOwnB	1.7	5.2	1.5	17.5	2.8	6.1	13.3	9.1	7.1
ctcOwn	4.2	7.4	6.8	15.2	6.7	1.4	13.5	8.9	8.0
ctcOwnB	0.6	6.3	5.0	33.0	4.0	11.0	32.3	16.7	13.6
ctcOther	6.3	9.7	16.3	25.8	10.0	13.2	33.2	20.7	16.9
ctcKemar	2.0	7.8	10.1	23.1	5.4	12.2	36.6	5.1	12.8
ctcNH57	10.6	11.5	14.2	-	11.8	4.4	36.6	23.9	16.1
ctcNH64	0.0	4.4	18.5	28.3	10.8	-	29.8	23.2	16.4
ctcNH68	21.9	12.1	27.1	23.3	9.1	23.1	-	18.3	19.3
ctcNH12	-	-	-	28.3	-	14.2	25.3	-	22.6

Table 6.2: Quadrant errors (QE) in % for all listeners and conditions tested. The condition ctcOther represents the median of the nonindividual conditions.

Condition	NH12	NH14	NH15	NH57	NH62	NH64	NH68	NH72	Mean
binOwn	28.3	26.5	30.5	37.0	27.0	28.5	31.1	38.9	31.0
binOwnB	25.0	30.1	31.2	35.6	26.6	34.8	34.2	35.5	31.6
ctcOwn	26.7	25.8	36.2	33.9	35.0	32.0	26.7	38.4	31.8
ctcOwnB	26.8	35.6	32.2	36.7	29.0	32.0	41.2	39.8	34.2
ctcOther	35.4	31.9	40.1	39.4	34.1	33.7	37.4	41.5	36.7
ctcKemar	35.2	26.0	39.9	33.7	35.2	31.3	40.6	36.8	34.8
ctcNH57	35.5	31.4	40.4	-	34.4	33.5	42.9	42.7	37.2
ctcNH64	26.1	32.3	36.0	42.2	31.3	-	32.3	43.6	34.8
ctcNH68	36.8	32.8	41.1	38.2	33.7	33.9	-	40.4	36.7
ctcNH12	-	-	-	40.6	-	38.0	34.2	-	37.6

Table 6.3: Local polar error (PE) in degrees for all listeners and conditions tested. The condition ctcOther represents the median of the nonindividual conditions.

Condition	NH12	NH14	NH15	NH57	NH62	NH64	NH68	NH72	Mean
binOwn	8.1	10.1	10.9	16.7	9.7	9.4	10.0	10.7	10.7
binOwnB	8.2	9.6	12.7	13.6	9.9	9.2	9.4	10.6	10.4
ctcOwn	8.0	9.1	11.7	14.0	9.6	8.7	10.4	11.1	10.3
ctcOwnB	10.8	13.4	14.3	15.5	9.5	10.2	11.8	18.2	13.0
ctcOther	11.0	11.3	14.4	15.5	10.9	10.8	13.0	14.0	12.6
ctcKemar	9.8	9.8	14.1	14.8	9.4	11.3	13.7	9.3	11.5
ctcNH57	11.9	8.9	14.8	-	12.8	10.2	13.6	14.4	12.4
ctcNH64	10.0	12.8	13.7	16.2	10.4	-	12.3	13.5	12.7
ctcNH68	13.0	16.2	17.2	17.6	11.4	15.8	-	14.6	15.1
ctcNH12	-	-	-	14.7	-	10.0	11.4	-	12.0

Table 6.4: Lateral error (LE) in degrees for all listeners and conditions tested. The condition ctcOther represents the median of the nonindividual conditions.

Looking more closely at the simulation results presented in (Akeroyd et al., 2007, fig. 10; reproduced in fig. 6.16), it seems that the errors are large for large negative interaural differences only. For smaller interaural differences, the errors seem to be negligible, which would suggest a correct reproduction of central targets.

In order to investigate this issue, the targets were grouped to those within (central) and those outside (lateral) the loudspeaker span and the LEs were calculated as a function of the target lateral angle (fig. 6.17(a)). While in the ctcOwnB condition the performance seems to slightly degrade for the central targets, the performance appears to be much worse for the lateral targets.

An RM ANOVA was performed on the LEs for the factors target direction (central, lateral) and condition (ctcOwn, ctcOwnB). Both main effects ($p < 0.001$) and their interaction ($p = 0.048$) were significant. The significant interaction suggests a different impact of the condition for the two target directions. The post-hoc test showed that the only significant difference was that for the lateral targets tested with ctcOwnB (17.2°) when compared to the lateral targets tested with ctcOwn (12.5°) or when compared to the central targets (11.5° for ctcOwnB and 9.7° for ctcOwn).

This indicates that while for targets placed outside the loudspeaker span the mismatch significantly affects the lateral localization performance, for targets placed inside the span the mismatched CTC system

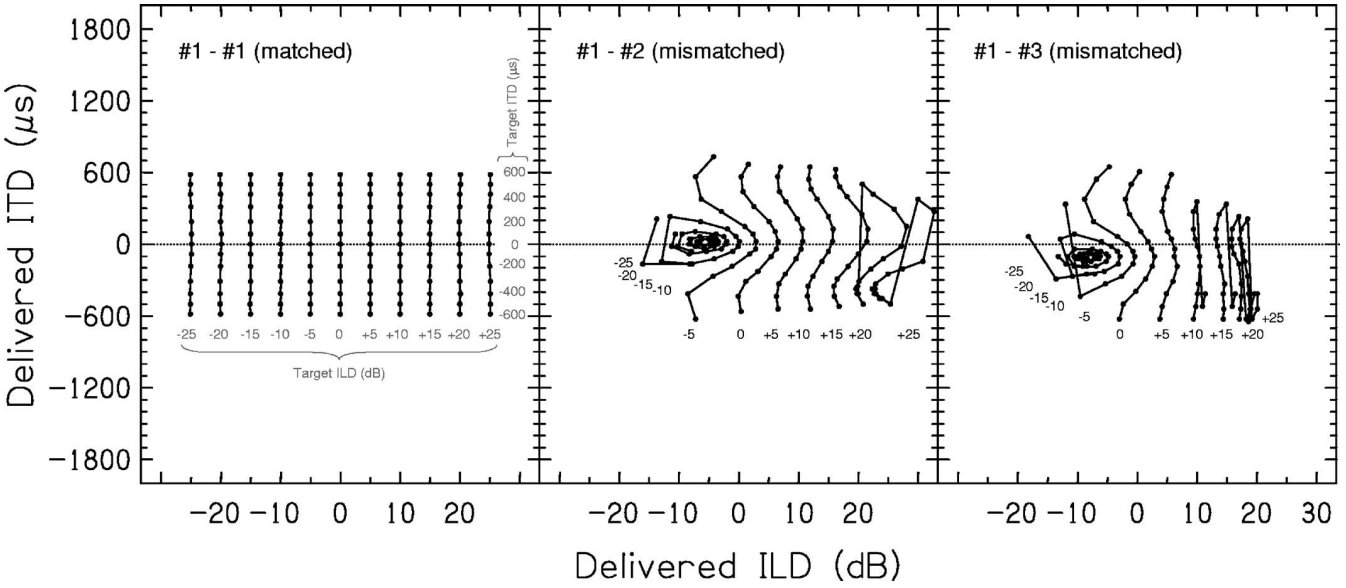


Figure 6.16: The binaural performance of a virtual CTC system, calculated for a matched system (left panel) and two mismatched systems (middle and right panels). Each panel shows the ongoing ITD (ordinate) and ILD (abscissa) delivered by the simulation for a large set of combinations of target ITD and ILDs (parameters); the lines join points with the same target ILD. The analysis was run at an auditory-filter frequency of 1000 Hz.

Reprinted with permission from M. A. Akeroyd et al. (2007). "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics". In: *J. Acoust. Soc. Am.* 121.2, pp. 1056–1069. Copyright 2007, Acoustical Society of America.

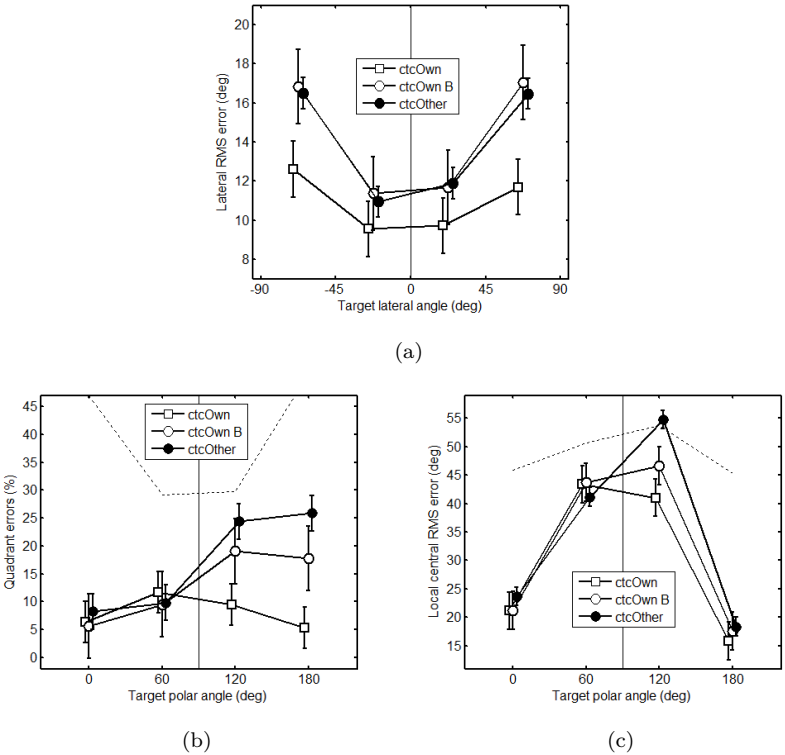


Figure 6.17: Localization error as a function of target position. (a) Lateral error as a function of the lateral angle. (b) Quadrant errors and (c) polar errors as functions of the polar target angle. The dashed lines show the errors which would result from random responses.

may yield a similar performance as the matched CTC system. This seems to confirm the above-mentioned observations concerning the details of binaural modeling described in Akeroyd et al. (2007). It could be speculated that there exist a correspondence between central targets in a mismatched CTC system and phantom sources in a stereophonic reproduction system.

Targets placed at elevations near the loudspeakers may also correspond to a phantom-source stereophonic reproduction. If such targets were well-localized in sagittal planes even in a mismatched CTC system then the difference between ctcOwn and ctcOwnB would depend on the

target polar angle, with a larger difference in the performance for targets placed behind the listener.

Figure 6.17(b) shows the QEs as a function of the polar angle, with targets grouped to four groups with a polar angle span of 60° (starting at $\beta = -30°$). The QE seems to increase with the polar distance between the targets and the loudspeaker elevation.

An RM ANOVA was performed with factors target hemifield (frontal targets: $-30° \leq \beta \leq 30°$, rear targets: $150° \leq \beta \leq 210°$) and condition (ctcOwn, ctcOwnB). The main factors ($p < 0.018$ for both) and their interaction ($p = 0.023$) were significant. The significant interaction suggests a different impact of the conditions in the two target hemifields. The post-hoc test showed that while for the frontal targets, the difference between the conditions was not significant (5.5% for ctcOwn, 5.9% for ctcOwnB), for the rear targets, highly-significantly ($p < 0.005$) more QEs occurred in the ctcOwnB (16.8%) than in the ctcOwn (5.7%) condition. This indicates a strong impact of the mismatch on the localization performance for the targets placed behind the listener. This is in agreement with the result presented by Nelson et al. (1997), who conducted a localization test using a mismatched CTC system using real loudspeakers and verified that when the system worked poorly, the front-back inversion rate (thus, the quadrant error) increased substantially.

All in all this means that, for the targets placed in the same hemisphere as the loudspeakers, the individual but mismatched condition ctcOwnB yielded a sagittal plane performance similar to that for the ideal matched condition ctcOwn. For the targets placed in the opposite hemisphere as the loudspeakers, the QE were substantially larger. This indicates that mismatched but individualized CTC systems might be able to provide a good performance only for the frontal targets. The exact match of the CTC filters to the propagation paths seems to be highly relevant for targets virtually placed at the other hemisphere than the loudspeakers. Furthermore, results indicate that only an ideal, individualized, matched CTC system provides a correct reproduction of the spectral cues required for accurate sagittal plane sound localization in both hemifields.

Localization Performance: nonindividualized CTC systems

The localization performance in the nonindividualized and thus mismatched CTC systems is usually assumed to be worse than that in individualized CTC systems (Akeroyd et al., 2007). However, an individualized CTC system can also be matched or mismatched, depending

on whether an ideal or a realistic CTC system is being studied. Thus, in the following the nonindividualized CTC systems (ctcOther) are compared with both the ideal CTC system (ctcOwn) and the realistic CTC system (ctcOwnB).[9]

The LE increased from 10.3° (ctcOwn) to 13.0° (ctcOwnB), but then it decreased to 12.6° (ctcOther), indicating a weak impact of the individualization on the horizontal plane localization performance (see table 6.4). However, there might have been some differences at particular target directions. Thus, the targets were grouped to those within (central) and those outside (lateral) the loudspeaker span and the LE were calculated as a function of the target lateral angle (fig. 6.17(a)). LE was still similar for all the mismatched (individual and nonindividual) conditions. Thus, for the horizontal plane localization, there seems to be no difference between the two types of mismatched CTC systems and an individualized CTC system seems to be of no advantage.

The QE increased from 8.0% (ctcOwn) to 13.6% (ctcOwnB) and then further to 16.9% (ctcOther, see table 6.2). Also, the PE increased from 31.8° (ctcOwn) to 34.2° (ctcOwnB) and then further to 36.7° (ctcOther, see table 6.3). The RM ANOVAs were performed with the factor condition at three levels (ctcOwn, ctcOwnB, and ctcOther) on the QEs and PEs. The factor condition significantly affected the QEs ($p < 0.001$) and the PEs ($p = 0.019$). The post-hoc tests showed that while ctcOther yielded significantly larger errors compared with ctcOwn, the errors were not significantly different when compared with ctcOwnB. At first glance, this might indicate that for the sagittal plane localization the performance is independent of the individualization of the filters. However, differences at particular positions might have been expected as was the case for the individualized mismatched condition. Thus, the targets were grouped into four groups and the QE and PE were calculated as a function of the polar angle (figs. 6.17(b) and 6.17(c)). For ctcOther, the QE increased with the increasing distance between the targets and the loudspeakers more than it did for ctcOwnB.

An RM ANOVA with factors target hemifield (frontal targets: -30° to 30°, rear targets: 150° to 210°) and condition (ctcOwnB and ctcOther) was performed on the QEs. The main factors condition ($p = 0.048$) and target hemifield ($p < 0.001$) were significant, but their interaction was not ($p = 0.23$). For the front targets, the QE was 5.8% and 8.1% for ctcOwnB and ctcOther, respectively. For the rear targets, the QE was

[9] The performance also varied across particular nonindividual conditions (see ctcKEMAR versus ctcNH68). However, in order to increase the statistical power, all nonindividual CTC conditions were pooled together in the presented analysis.

16.8% and 26.1% for ctcOwnB and ctcOther, respectively. While the not significant interaction shows that none of the hemifield-condition combinations was significantly different from the others, the significance in the factor condition shows that ctcOwnB indeed yielded a significantly better performance than ctcOther—when separately analyzed for the two hemifields. The 50% larger QE in ctcOther for the rear targets further supports this evidence.

A similar situation was revealed by the RM ANOVA performed on the PEs with factors target hemifield (frontal targets: -30° to 90°, rear targets: 90° to 210°), condition (ctcOwnB and ctcOther), and their interaction. The interaction ($p = 0.012$) was significant and the post-hoc test showed that while for the frontal targets, the difference between the conditions was not significant (32.4° for ctcOther and 33.6° for ctcOwnB), for the rear targets, significantly ($p < 0.01$) larger PE occurred in the ctcOther (40.8°) than in the ctcOwnB (33.7°) condition. Thus, the individualization of the CTC systems was able to substantially reduce the PE for rear targets.

All in all, the presented analysis demonstrates that in mismatched CTC systems, the sagittal plane localization performance improves when individualized CTC filters are considered, especially for targets placed behind the listener.

Channel Separation

The channel separation (CS) was calculated according to eq. (5.11) for all conditions and listeners. The CS, averaged in the frequency range 0.3 to 8 kHz (Akeroyd et al., 2007; Parodi and Rubak, 2011) is shown in table 6.5 for each listener and condition. The last two rows of table 6.5 show the corresponding $\widehat{\mathrm{CS}}$—calculated according to eq. (5.12) for the left ear and by analogy for the right ear—averaged over the same frequency range and over both ears.

For ctcOwn, the CS was large, on average 68.4 dB, and in the range of those reported previously for matched CTC systems (Bai and Lee, 2006; Akeroyd et al., 2007). For ctcOwnB, the CS was substantially lower, on average 14.7 dB, and in the range of $\widehat{\mathrm{CS}}$ for both measured HRTF sets (14.9 and 14.5 dB). This indicates that individualized but mismatched CTC systems and reproduction systems without any CTC yield similar averaged CS.

For ctcOther, the CS was on average 15.0 dB and in the range of those reported previously for nonindividualized CTC systems (Akeroyd

Frequency Range	0.3 to 8 kHz									0.3 to 2 kHz	4 to 16 kHz
CS	NH12	NH14	NH15	NH57	NH62	NH64	NH68	NH72	Average ± Std. Dev.	Average ± Std. Dev.	Average ± Std. Dev.
ctcOwn	70.9	67.6	67.9	67.9	68.3	67.0	68.8	68.6	68.4 ± 1.2	50.4 ± 2.2	58.5 ± 2.6
ctcOwnB	16.2	13.5	11.5	15.6	19.2	18.2	11.4	12.3	14.7 ± 3.0	16.5 ± 3.4	14.2 ± 1.2
ctcKemar	17.2	15.6	16.8	15.9	18.1	18.0	12.6	18.9	16.6 ± 2.0	16.5 ± 1.9	12.6 ± 0.6
ctcNH57	13.6	17.4	17.9	-	13.9	16.3	11.9	13.7	14.9 ± 2.3	15.2 ± 2.8	13.2 ± 0.6
ctcNH64	17.2	15.1	16.1	16.2	17.7	-	11.0	17.1	15.8 ± 2.3	16.6 ± 2.3	13.6 ± 1.5
ctcNH68	14.5	11.5	12.0	11.9	12.2	11.0	-	12.9	12.3 ± 1.2	12.4 ± 2.1	11.9 ± 0.7
ctcNH12	-	13.6	13.9	13.6	16.3	17.2	14.5	18.9	15.4 ± 2.1	14.1 ± 1.2	14.5 ± 1.1
binOwn	16.4	15.3	15.1	14.4	14.3	14.5	13.8	14.6	14.8 ± 0.8	8.1 ± 0.3	17.1 ± 1.1
binOwnB	15.3	14.6	14.1	15.3	14.6	13.9	13.8	14.8	14.5 ± 0.6	8.2 ± 0.3	17.4 ± 1.4

Table 6.5: Channel separation CS in dB averaged over three frequency ranges. The last two rows show the natural channel separation $\widehat{CS}$ averaged over both ears. Conditions not tested in the localization experiments are shown italic.

et al., 2007, average of 17.1 dB). It was also similar to that for ctcOwnB (14.7 dB), which might lead to the conclusion, that an individualization of the CTC systems is not necessary at all. Note that such a conclusion would not be consistent with the results from the previous section.

One reason for the similar CS in the individualized and nonindividualized CTC systems might be the choice of the frequency range used for averaging the frequency-dependent CS. In order to investigate the use of other frequency ranges, the CS was averaged over frequencies from 0.3 to 2 kHz (low-frequency CS) and from 4 to 16 kHz (high-frequency CS). Both low- and high-frequency CSs averaged over the listeners are shown in the right-most column of table 6.5. As averages over listeners, low- and high-frequency CSs showed a similar trend to the mid-frequency CS (0.3 to 8 kHz) when compared across the conditions. The correlation coefficient between all low-frequency and high-frequency CSs was 0.98.

Such a high correlation might arise, however, because of large and small CSs in the matched and mismatched conditions, respectively, and such large differences may dominate the correlation. Thus, the correlation was calculated separately for the matched and mismatched conditions. For the matched conditions only, the correlation coefficient between the low-frequency and the high-frequency CSs was 0.51. For the mismatched conditions only, the correlation coefficient was 0.24. A further comparison of the CS and $\widehat{\text{CS}}$ revealed that for the low-frequency range, the average CS was 15.0 dB and thus, larger than the corresponding average $\widehat{\text{CS}}$ of 8.15 dB. This indicates that the CTC indeed increased the CS in the frequencies below 2 kHz. For the high-frequency range, the average CS was 13.2 dB and thus *smaller* than the corresponding average $\widehat{\text{CS}}$ of 17.25 dB. This means that the CTC actually *decreased* the CS in the frequencies above 4 kHz for the tested CTC setup. Thus, if no CTC is applied in the frequency range above 4 kHz, the mismatched CTC systems tested in this experiment would show a larger CS.

This finding is in agreement with Gardner (1997), who limited his CTC system to the low frequencies only. This observation can also explain the results from Bai et al. (2007), who band-limited their CTC to 6 kHz and obtained a lateral localization performance similar to that of the full-bandwidth CTC. While their choice for the band limitation was based on computational issues, the lack of the mismatched CTC at higher frequencies, and thus, no decrease in the CS at these frequencies might also have contributed to their findings. Generally, it seems that a frequency-dependent amount of CTC might be useful in order to avoid a decrease of the CS in mismatched CTC systems.

Localization Performance: Channel Separation

On the one hand, one quality aspect of a CTC system is the localization performance the system is able to provide. On the other hand, the CS is usually employed to describe the general quality of a CTC system. However, not much is known about the relation between the CS and localization performance.

The CS is calculated between the two ears and is, in principle, an interaural metric. Therefore, it might indeed have the potential to describe the horizontal plane localization performance, which also depends on interaural cues. The ITDs, being the most salient cues for sound localization in the horizontal planes (Wightman and Kistler, 1992) are assumed to contribute in frequencies up to approximately 2 kHz (Macpherson and Middlebrooks, 2002). The contribution of ITDs to the localization performance was evaluated by comparing the horizontal plane performance with the low-frequency CS (0.3 to 2 kHz). The ILDs, also salient cues for the horizontal plane localization, are large in frequencies above approximately 3 kHz. The contribution of both ILDs and ITDs was evaluated by means of the mid-frequency CS (0.3 to 8 kHz). Even though CS is an interaural metric, the use of the CS to describe the sagittal plane localization performance was also investigated. The spectral cues, being the most salient cues for the sagittal plane localization (Langendijk and Bronkhorst, 2002; Macpherson and Middlebrooks, 2002) are assumed to contribute the most in the frequency range from 4 to 16 kHz (Carlile and Pralong, 1994; Perrett and Noble, 1997; Middlebrooks, 1999b). Thus, the sagittal plane localization performance was compared with the high-frequency CS (4 to 16 kHz).

At first sight, the relation between the CS and the performance seems to be weak. For example, for ctcOwn, ctcOwnB, and ctcOther, the average QE was 8.0%, 13.6%, and 16.9%, respectively, correspondent to a mid-frequency CS of 68.4, 14.7, and 15.0 dB, respectively. While from ctcOwn to ctcOwnB, the increase in QE is well represented by the decrease in CS, the further increase in QE from ctcOwnB to ctcOther is not. Generally, the CS in the range of 50 dB corresponds to a good localization performance. However, smaller CS (in the range between 13 to 18 dB) did not provide any statement on the localization performance. One example is NH72, who for quite different CSs (12.3 and 17.1 dB) showed nearly the same QE (23.2% and 23.9%). Other example is NH15, who for similar CSs (11.5 and 12.0 dB) showed completely different QEs (5% and 27.1%). Note that NH15 also showed a QE of 6.8% for the

matched condition with a CS of 67.9 dB. This demonstrates the rather complex relation between CS and the localization performance in terms of QEs.

In order to estimate the statistical relation between the CS and the localization performance, the correlation between the CS and the localization errors is analyzed. The correlation coefficients calculated for the mid-, low, and high-frequency CS, are shown in table 6.6. For all tested conditions, the correlation coefficient between mid-frequency CS and QE, PE, and LE was -0.35, -0.32, -0.43, respectively (all significant, $p < 0.025$). Similar correlations resulted for the low- and high-frequency CS. Such weak correlations suggest that the CS is generally not a good predictor for the localization performance.

The low correlations between the CS and localization performance could, however, be put down to the listener-individual performance in the localization task. In order to compensate for the individual performance, the correlation coefficients were calculated between the CS and the performance relative to that obtained from the binOwn condition. All correlation coefficients increased (see table 6.6), with the largest coefficient being at -0.49. Hence, CS seems to be a poor predictor for the localization performance even when compensated for the listener-individual localization performance.

Since the localization performances of the matched and mismatched CTC systems differ tremendously, the CS might better correlate with the performance when compared separately for the matched and mismatched CTC conditions. For the matched condition (ctcOwn in table 6.6), the most correlation coefficients were low and not significantly different from zero, i.e. were uncorrelated. The only significant ($p = 0.018$) correlation coefficient (-0.79) was found between the high-frequency CS and the LE relative to that obtained for binOwn. This might suggest that in matched CTC systems, the high-frequency CS is able to predict the horizontal plane localization performance relative to the listener-individual performance.[10]

For the mismatched CTC systems, the largest significant correlation coefficients were -0.50 (QE and mid-frequency CS), -0.33 (PE and low-frequency CS), and -0.60 (LE and mid-frequency CS). These correlations did not improve when the relative localization performance was considered and show the extent to which CS might act as a predictor for the localization performance. Especially for the horizontal plane localization

[10] Note that despite the statistical support, this correlation is based on a small sample size (n=8) and that such a conclusion is to be treated with caution.

Frequency Range	0.3 to 8 kHz (mid)			0.3 to 2 kHz (low)			4 to 16 kHz (high)		
Condition	QE	PE	LE	QE	PE	LE	QE	PE	LE
all tested	**−0.35**	**−0.32**	**−0.43**	**−0.35**	**−0.33**	**−0.43**	**−0.32**	**−0.31**	**−0.38**
all tested. relative to binOwn	**−0.37**	**−0.36**	**−0.45**	**−0.39**	**−0.38**	**−0.49**	**−0.35**	**−0.35**	**−0.42**
matched	0.00	−0.34	−0.31	0.21	0.03	0.51	0.30	−0.40	0.26
matched. relative to binOwn	−0.25	−0.35	−0.27	−0.57	−0.36	−0.43	−0.25	−0.33	**−0.79**
mismatched	**−0.5**	−0.28	**−0.6**	**−0.35**	**−0.33**	**−0.42**	**−0.33**	−0.16	−0.27
mismatched. relative to binOwn	**−0.48**	−0.20	**−0.55**	**−0.38**	−0.24	**−0.53**	−0.33	−0.02	**−0.25**

Table 6.6: Correlation coefficients for the correlation between the localization errors and the channel separation. Coefficients significantly ($p < 0.05$) different from zero are shown bold. The matched condition was ctcOwn. The mismatched conditions were ctcOwnB and ctcOthers.

performance, the correlation of -0.6 might be useful in further evaluations of CTC systems, depending on the application criteria. For the sagittal plane localization performance, the CS seems to be a poor predictor. This is not surprising, considering that monaural, not interaural, cues are the most salient cues for the sagittal plane localization.

6.2.3 Discussion

In experiment II the sound-localization performance in CTC reproduction systems was studied and its CTC filters were calculated under various conditions. The performance was compared to the baseline binaural condition. Channel separation, an objective measure for the quality of a CTC system, was calculated for the tested conditions and compared with the localization performance.

Under binaural conditions, the localization performance in terms of quadrant errors, polar errors, and lateral errors was within the range of previously reported performance. This was also the case when they were tested using HRTFs obtained from a measurement that was repeated approximately five years later, even though training was conducted only with the first HRTF set. This suggests that the human auditory localization system is robust to HRTF measurement variability—at least for the measurement setup used in this experiment.

With the matched CTC systems, the performance was similar to that from the binaural conditions. With the individualized but mismatched CTC systems, where CTC filters were based on the repeated HRTF measurements, the listeners showed a degraded localization performance in terms of larger lateral, polar, and quadrant errors. This shows that the propagation paths from the loudspeakers to the ears must *exactly* match the filters in a CTC system in order to provide localization performance at a similar level as the binaural reproduction. The direction-dependent analysis of the localization performance showed that in the mismatched CTC systems, the performance deteriorated especially for targets placed outside the loudspeaker span and/or behind the listener. With the nonindividualized CTC systems, the quadrant errors further increased for the rear targets and the performance for the frontal targets was in the range of that for the individualized but mismatched CTC system.

These findings show that for targets placed within the loudspeaker span and in the same hemisphere as the loudspeakers, the quality of the CTC system is not critical regarding localization and the amount of CTC can be reduced in order to provide a better timbre reproduction.

This might lead to a deteriorated apparent source width, though. Much attention, however, should be attached to the CTC systems for targets placed at directions outside the loudspeaker span. In particular for the rear targets, a work around to the currently unachievable matched CTC system, could be a second CTC system with loudspeakers placed behind the listener (Parodi and Rubak, 2011). This appears indeed to be a promising approach. For the more lateral targets, additional loudspeakers at lateral positions might help. They could, combined with the loudspeakers in the rear, form a ring of loudspeakers around the listener. Such a system would have to consider all available loudspeaker combinations[11] to choose the most adequate CTC filter for each source to be reproduced and might thus be seen as an extension of the vector base amplitude panning (VBAP, Pulkki, 1997) to binaural reproduction. To the best of the author's knowledge, such a combination of systems has not been scientifically investigated yet.

A common quality metric for CTC systems is the channel separation. The results from this experiment show a substantial difference in channel separation between the matched and the mismatched CTC systems. However, channel separation was similar in both individualized and nonindividualized mismatched CTC systems, even though the sagittal plane localization performance was not. For the mismatched CTC systems, the channel separation was in the range of the natural channel separation provided by a stereophonic reproduction. The mismatched CTC systems improved the channel separation in frequency range below approximately 2 kHz but degraded the channel separation in the frequency range above approximately 4 kHz, suggesting that mismatched CTC should be avoided in the higher frequency regions. The matching had only little impact on the low-frequency channel separation. Hence, future efforts with regard to the matching should focus on the mid- and high-frequency regions, at least for the tested loudspeaker span.

A generally weak correlation (up to -0.35) was observed between the channel separation and the sagittal plane localization performance. This was also the case (up to -0.39) when compensating for the listener-individual localization performance in the binaural condition. The correlation increased to -0.5 when only mismatched CTC system were considered. This confirms the evidence that channel separation, being an interaural metric, is not an appropriate predictor for the sagittal plane localization performance.

For the horizontal plane, a better correlation between channel separation and localization performance could be expected. It was -0.49 in

[11] A method for the dynamic implementation of such filters is described in section 5.3.

general and increased to -0.79 when only matched CTC systems were considered. This correlation was between the high-frequency channel separation and the lateral errors relative to the baseline performance, it was significantly correlated, however, it was based on only eight samples. For mismatched CTC systems, the correlation of -0.6 was found for a more convincing sample size of 40 samples. Such a correlation indicates that the channel separation might be indeed useful when evaluating mismatched CTC systems with respect to the horizontal plane localization.

7

Conclusion

7.1 Summary

The main objective of this work was to improve the quality of binaural-based virtual acoustics systems. One of the key aspects to achieve this goal is individualization. Therefore, two components of individualized binaural technology were addressed in this thesis: the efficient acquisition of individual head-related transfer functions (HRTFs) and the adequate individual equalization of binaural reproduction systems.

This work started with the design of a measurement setup for the fast acquisition of individual HRTFs. The proposed solution was to construct a setup composed of an arc that can hold up to 40 loudspeakers and a turntable to rotate the subject inside that arc. This combination is assumed to be the best trade-off between hardware costs and measurement duration.

Binaural-based virtual reality systems commonly neglect the effect of near-field HRTFs. This is, nevertheless, a very important aspect when it comes to improving the realism of acoustic scenes. The range extrapolation technique, based on acoustic spherical holography, makes it possible to describe the HRTF's distance dependence based on measurements at a single distance. This setup is one of the first of its kind designed to be compatible with the range extrapolation technique. To avoid reflections coming from the supporting arc, it was built as a truss structure, expected to be acoustically transparent at the audible frequency range. The loudspeakers were designed in a drop-like shape to avoid edge diffraction. Finally, care was taken to choose a loudspeaker driver with broadband response that radiated as similar as possible to an omnidirectional source.

A reduced measurement time is not only more comfortable for the subject being measured, it can also help to reduce the measurement variability. The use of multiple loudspeakers and excitation signals in parallel will already produce a shorter total measurement time. Instead of orthogonal pseudo-random sequences, which are very sensitive to nonlinearity, the more robust multiple exponential sweep method (MESM) was

used. This method was then further optimized, taking into consideration the time structure of the measured impulse response and allowing a more flexible distribution of the unwanted harmonic impulse responses along the raw impulse response. Measuring an HRTF dataset with 40 positions in elevation and 100 positions in azimuth in a sequential manner with an exponential sweep of length 1.34 s (assuming the turn table takes 0.3 s to reach its next position), would take 90 min. The same HRTF dataset measured with the original MESM would take just over 12 min to complete and applying the optimizations described in this work would reduce the measurement time even further, to less than 6 min. The use of interleaved sweeps has no influence on the obtained signal-to-noise ratio.

Measurements made with the MESM will result in a raw interleaved impulse response that needs further post-processing. After extracting each direction's transfer function, the signals must be further equalized to eliminate the influence of the transducers. At this point, the range extrapolation can be applied to account for the radial dependency of the HRTF. To verify the quality of the newly developed setup, an artificial head, whose HRTF dataset had already been measured with a previous generation setup composed of only one loudspeaker attached to a moving arm, was also measured with the newly constructed setup. The comparison of time, frequency, and spatial data showed a good agreement between the two measurement setups. The measurement with the new setup took indeed less than 6 min to complete, making this setup the fastest existing individual HRTF measurement system the author is aware of.

Individual HRTFs provide the basis for the binaural synthesis. After the binaural signals have been synthesized using these HRTFs, they must now be adequately played back to the listener. Binaural signals can then be reproduced either via headphones or via loudspeakers.

Reproduction via headphones is the more straightforward of the two methods, as headphones are able to feed each binaural signal directly to the respective ear. However, to deliver an authentic auditory impression without additional spectral coloration, the reproduction via headphones must be adequately equalized. Repeated measurements of the headphone transfer function (HpTF) confirmed that when the listeners are allowed to place the headphones themselves, at what they consider to be the most comfortable position, then HpTF variance drops considerably. On the other hand, measurements with several listeners also confirmed that HpTF varies considerably among subjects. A framework was developed for the design of individual headphone equalization filters. It includes

several measurements (about ten) of the listener's HpTF, where the listeners are asked to take off the headphones in between each measurement. The magnitude spectra of these HpTFs are averaged and deep spectral dips are smoothed. The resulting magnitude spectrum is then inverted and finally the equivalent minimum-phase spectrum is obtained through the Hilbert transform.

The quality of the proposed equalization filters was verified with the aid of perceptual tests. The first part of this experiment consisted of a direct comparison (in a three-alternative forced-choice setup) of a stimulus reproduced through a loudspeaker—played in anechoic conditions—and its binaural synthesis reproduced via individually equalized open-type circumaural headphones. Three stimuli were used for this test: a pulsed pink noise, a speech sample and a music sample. Results showed that at least 50% of all listeners could not distinguish between the reproduction methods when hearing to speech (error rate: 38.17%) and music (error rate: 35.10%). Meanwhile, when listening to the pulsed pink noise—which was the only stimulus that contained spectral components above 10 kHz—listeners made significantly less mistakes (error rate: 16.74%), which suggests that stimuli with dominant high frequency components tend to allow an easier distinction between reproduction methods. This result is reasonable as HpTFs show higher variance exactly at this frequency range.

The second part of this experiment was an indirect comparison, where listeners were asked to say whether the presented stimulus—this time only the pulsed pink noise—came from the loudspeaker or the headphones, excited as in the first part of this experiment. Results showed that listeners were guessing at almost all times (error rate: 51.03%), indicating that the binaural signals reproduced via individually equalized headphones sounded natural and authentic. Thus, even though differences between the original anechoic sound source and its binaural auditory display could be heard when the stimulus contained high frequency components, these differences were not big enough to allow the listeners to state clearly whether the auditory event that was binaurally reproduced via individually equalized headphones was actually coming from the headphones and not from an external source.

The other way to play back binaural signals to a listener is via (at least two) loudspeakers. This method is commonly preferred for applications where the use of headphones might hinder the sense of immersion, e.g. in virtual reality environments. Binaural reproduction via loudspeakers suffers from crosstalk, which mixes the spatial cues contained in the binaural signal. This effect can be compensated for by using crosstalk

cancellation (CTC) filters, which generate the desired channel separation between the listener's ears. In virtual reality applications, users should have complete freedom of movement. For such systems, a dynamic CTC with multiple loudspeakers is required and the CTC filters must be constantly updated according to the tracked head position. As frequency-domain calculations are usually more efficient (and faster) than their time-domain counterparts, dynamic CTC systems tend to be implemented in frequency-domain. This has the drawback that such filters display noncausal artifacts, which is not the case in time-domain calculations.

A framework was proposed for the calculation of causal CTC filters in the frequency-domain. An approximation of the causal solution is obtained by using a time-domain calculation which is in turn based on the Wiener-Hopf decomposition. It is known that CTC filters show a large gain boost at certain frequencies and regularization is used as a gain limiter. However, the regularization has the drawback that it adds noncausal artifacts both in the CTC filters and in the resulting impulse response in the ears. These noncausal artifacts can be eliminated—or rather be transformed in causal artifacts—through the minimum-phase regularization. Instead of the channel-dependent solution proposed with the original minimum-phase regularization method, the proposed framework applies a new global minimum-phase regularization strategy.

Dynamic CTC systems need multiple loudspeakers to be able to switch between active loudspeakers, thus avoiding instability of the CTC filters. It was shown that simple panning between two configurations can affect the system's resulting frequency response. The proposed framework incorporates spatial fading in the filter calculation stage through a weighted matrix inversion, which provides a smooth transition between active loudspeakers.

CTC filters are calculated from the transfer functions between loudspeakers and listener's ears, i.e., the HRTFs. Therefore, the CTC filters are subject to individualization as well. Nevertheless, many CTC systems use generic transfer functions for its filter calculation. A localization test was conducted to evaluate the influence of individualization on the localization performance of CTC systems. So far, only one similar evaluation has been conducted, but it used an auditory model simulation instead of a perceptual test. The localization performance was tested with regard to quadrant errors (QE), polar errors (PE), and lateral errors (LE). Listeners had their HRTF measured twice (within an interval of five ears). Acoustic targets were Gaussian white noises, filtered with the listener-specific directional transfer functions. Baseline tests showed

that the localization performance was within the range of previously reported performance for both HRTF sets (set 1: LE=10.7°, PE=31.0°, QE=8.6%; set 2: LE=10.4°, PE=31.6°, QE=7.1%), even though training was provided using only the first of these sets, suggesting that the human auditory localization system is robust to small HRTF measurement variability.

In this experiment, CTC systems were virtually rendered and presented via headphones. Results showed that individualized matched CTC systems (the same HRTFs are used for the filter calculation and the loudspeaker rendering) provided performance similar to that from the binaural listening (LE=10.3°, PE=31.8°, QE=8.0%). For individualized mismatched systems (two different HRTF datasets from the same listener are used for the filter calculation and the loudspeaker rendering) the localization performance deteriorated (LE=13.0°, PE=34.2°, QE=13.6%). And for nonindividualized mismatched systems (the CTC filters are calculated with the HRTFs from other listeners) the sagittal localization errors increased further (PE=36.7°, QE=16.9%). The direction-dependent analysis showed that mismatch and lack of individualization yielded a degraded performance for targets placed outside of the loudspeaker span (LE=17.2°) and behind the listeners (PE=40.8°, QE=26.1%), indicating the relevance of individualized CTC systems for such targets.

It is commonly assumed that binaural reproduction through loudspeakers will only work if the CTC filters can provide a sufficient channel separation. Thus, channel separation is very often used as a predictor for the quality of a CTC system. The channel separation was calculated for all conditions evaluated in this localization test for different frequency ranges. The results showed a substantial difference in channel separation between the matched and the mismatched CTC systems, but similar values in both individualized and nonindividualized mismatched CTC systems, which does not match the observed variations in localization performance. It was observed that the mismatched CTC systems improved the channel separation in frequency range below approximately 2 kHz, but degraded the channel separation in the frequency range above 4 kHz, an observation that is in agreement with the practical knowledge that mismatched CTC should be avoided at the higher frequency regions. Results showed that channel separation might be indeed useful for the evaluation of mismatched CTC systems with respect to the horizontal plane localization, but it is not an appropriate predictor for the sagittal plane localization performance.

All in all, this thesis extended the current knowledge on the efficient measurement of individual head-related transfer functions and highlighted the importance of individual equalization filters in binaural reproduction, both via loudspeakers and headphones. Moreover, an integrated framework for the calculation of such equalization filters was presented.

7.2 Outlook

Some aspects in the field of individual binaural technology could be improved in the course of this thesis, but a number of other questions were left unanswered and many new questions were raised. This section presents some ideas that could further improve the quality of the acquisition and post-processing of individual HRTFs as well as the quality of binaural reproduction methods.

The proposed individual HRTF measurement setup used a head-rest to stabilize the listener's head. This apparatus did reduce the listener's movement during measurement, but could not completely eliminate it. A possible solution to compensate for the influence of small head movements during measurement would be to track the position of the listener's head and compensate for the observed movements in a post-processing step, using a measurement grid for the spherical harmonic interpolation made with the tracked head position at the time of each measurement.

The chosen design of the arc proved not to be ideal. From a mechanical point of view, the used truss structure was too delicate and the solder joints constantly gave away. From an acoustical point of view, the arc itself vibrated during measurement, acting as an unwanted acoustic source. Even though the designed drop-like loudspeakers did have an adequate radiation pattern, they reflected the sound coming from the neighboring speakers. Thus, another design of the arc should be investigated, eventually attaching the loudspeaker to a more rigid arc with a continuous acoustically absorbing material. Furthermore, the loudspeaker drivers displayed a time-variant behavior, unacceptable for this kind of measurement. A new loudspeaker driver should be chosen that does not display this kind of behavior.

The truncation of the spherical harmonic order during interpolation or range extrapolation results in a spatial smoothing of the data. It is known that HRTFs smoothed in frequency-domain are still able to provide correct spatial impression (Kulkarni and Colburn, 1998;

Xie and Zhang, 2010), so it is also possible that a spatially smoothed HRTF dataset will also provide a correct spatial impression. Thus, a psychoacoustic evaluation of the required spherical harmonic order to reconstruct a perceptually correct HRTF dataset is necessary, as currently available studies on the required spherical harmonic order that adequately describe an HRTF dataset, take only physical aspects into account. It would also be important to investigate whether this limit varies with distance, as this would directly influence the truncation limits of the range extrapolation stage.

Even though the constructed HRTF measurement setup is already quite fast, it could be made even faster if the listeners were constantly rotated during the measurement as is the case in the continuous HRTF measurement methods proposed by Enzner (2009) and the plenacoustic interpolation method proposed by Ajdler et al. (2007). Theoretical calculations suggest that a dynamic measurement scheme using the MESM would reduce the measurement duration of 4000 HRTFs from the current 6 min to a mere 2 min, keeping the robustness to nonlinearity. However, the HRTFs measured that way will be spatially blurred. It is still to be evaluated if this blur is of perceptual relevance. Nevertheless, first steps towards compensating the influence of rotation in the dynamic measurement have been reported by Krechel (2012).

The fact that loudspeakers adjacent in elevation are sitting on opposite halves of the arc can lead to strong phase ripples in the HRTF's spatial representation, which are harmful to the spherical harmonic processing of the data. This will happen if the subject being measured is not placed with its longitudinal axis exactly centered with the turntable's rotation axis. As such a precise positioning of the listener is practically impossible, other methods should be evaluated on how to correct this effect in a post-processing stage. One possibility would be to conduct a dynamic reference measurement, as proposed by Krechel (2012). Another alternative, possibly more precise, would be to search for the acoustic center of the points from each half of the arc independently—e.g. with a search algorithm as the one proposed by Ziegelwanger (2012)—and then shift both halves together.

The limited number of measurement points can lead to spatial aliasing when post-processing the HRTF dataset. A possible way to deal with this spatial aliasing would be to use the compressive sensing framework. A preliminary investigation showed that spherical harmonics are not an adequate basis to describe the HRTFs in a sparse manner (Masiero and Pollow, 2010). Work has still to be done in searching for an adequate

basis for the sparse representation of HRTFs. A possible basis could be spherical wavelets or the Slepian functions.

Besides individual HRTFs, the binaural reproduction could also be improved. It is clear that binaural reproduction via headphones is a well-established technique and this thesis even confirmed that auditory displays presented via individually equalized headphones sound realistic and authentic. However, it would be interesting to evaluate how the lack of equalization will actually influence the auditory impression as, e.g., the influence of headphone equalization in localization is still debatable. This research could lead to the design of a new pair of headphones, which could provide an authentic auditory impression without individual equalization, thus facilitating the introduction of binaural technology in the consumer market.

On the other hand, the binaural reproduction via loudspeakers still has some hurdles to clear before it is ready for the consumer market. The first, and probably easiest of them, is an efficient low cost head tracking device. This could be easily implemented with a camera and face tracking software. First steps in this direction have been taken by Fundalewicz (2012).

A second very important aspect—deliberately ignored throughout this thesis—is the influence of the reproduction room, as unwanted reflections play a major role in the quality of the reproduced binaural signal. An interesting approach proposed to control this effect is to generate CTC filters that also compensate for room reflections, but only on a time range where reflections are above the temporal masking threshold of the human auditory system (Jungmann et al., 2012). The use of psychoacoustical knowledge to lessen the restrictions imposed on CTC systems seems to be a promising field of studies.

In this context, virtual reality reproduction systems could profit from a hybrid scheme, ideally combining the advantages of different spatial audio reproduction techniques. Such a hybrid method was first proposed by Pelzer; Masiero, and Vorländer (2011), who suggested the use of a binaural CTC system to reproduce only the direct sound and early reflections and the use of a system like low-order ambisonics to reproduce the reverberant tail of a simulated room impulse response.

A further possible improvement of CTC systems can be derived from the results presented in this thesis, namely that localization of sources within the loudspeakers' span was not significantly deteriorated by a mismatched CTC system. First, the speculation that this situation occurs because central targets in a mismatched CTC system are localized

with the same psychoacoustical process used to localize phantom sources in a stereophonic reproduction system should be investigated. Furter, in a setup with multiple loudspeakers, each reflection of a room impulse response could be played back by the group of loudspeakers chosen in accordance with the direction this reflection is arriving from, much in the same way as vector base amplitude panning does its active triangle selection.

Finally, as it was shown that channel separation does not seem to be an adequate predictor for the localization performance provided by a CTC system, a new type of CTC quality predictor should be developed. Auditory models could come in handy to help define such a predictor.

A

Regularization as a Gain Limiter

Farina (2007) suggests the use of a time-packing filter to control the size of an inverse filter's impulse response. He achieved this by conducting a regularized inversion in frequency-domain, as follows

$$C(z) = H(z)^*/(H(z)^*H(z) + \mu). \tag{A.1}$$

Regularization acts, however, not only as a time-packing tool. It also works as a gain limiter in the frequency-domain. This can be verified by taking the derivative of the amplitude of $C(z)$ in relation to the amplitude of $H(z)$

$$\frac{\partial|C(z)|}{\partial|H(z)|} = \frac{-|H(z)|^2 + \mu}{(|H(z)|^2 + \mu)^2}. \tag{A.2}$$

From eq. (A.2), when $|H(z)|^2 = \mu$, the maximum value of $|C(z)|$ is

$$|C(z)| = \frac{1}{(2\sqrt{\mu})}. \tag{A.3}$$

Thus, if $|C(z)|$ should be no greater than x dB, μ should be chosen to be

$$\mu = \frac{1}{(2 \cdot 10^{x/20})^2}. \tag{A.4}$$

Also in the multi-channel regularized inversion μ can be understood as a gain limiter. We know that the Euclidean norm of a matrix is given by

$$\|\boldsymbol{H}\|_2 = \sigma_{\max}(\boldsymbol{H}), \tag{A.5}$$

where $\sigma_{\max}$ is the largest singular value of $\boldsymbol{H}$. Assuming that the singular value decomposition from $\boldsymbol{H}$ is

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \tag{A.6}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices and $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values of $\boldsymbol{H}$, then the regularized matrix inversion

$$\boldsymbol{C} = (\boldsymbol{H}^*\boldsymbol{H} + \mu\boldsymbol{I})^{-1}\boldsymbol{H}^*. \tag{A.7}$$

can be rewritten as

$$\boldsymbol{C} = \boldsymbol{V}(\boldsymbol{\Sigma}'\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-1}\boldsymbol{\Sigma}'\boldsymbol{U}^*, \tag{A.8}$$

where $\boldsymbol{\Sigma}'$ is the transpose of $\boldsymbol{\Sigma}$ and the singular values from $\boldsymbol{C}$ are thus

$$\sigma_i(\boldsymbol{C}) = \frac{\sigma_i(\boldsymbol{H})}{(\sigma_i(\boldsymbol{H})^2 + \mu)}. \tag{A.9}$$

Golub and Van Loan (1996) argue that the amplitude of the largest element of a matrix is smaller or equal to the Euclidean norm of this matrix, thus

$$\max_{i,j} |c_{i,j}(z)| \leq \|\boldsymbol{C}\|_2 = \sigma_{\max}(\boldsymbol{C}) = \frac{\sigma_{\mathrm{j}}(\boldsymbol{H})}{(\sigma_{\mathrm{j}}(\boldsymbol{H})^2 + \mu)}, \tag{A.10}$$

where $\sigma_{\mathrm{j}}(\boldsymbol{H})$ is the singular value of $\boldsymbol{H}$ that results in the largest singular value of $\boldsymbol{C}$.

Taking the maximum of $\sigma_{\max}(\boldsymbol{C})$ with regards to $\sigma_{\mathrm{j}}(\boldsymbol{H})$, as was done for the one channel case, results in

$$\max_{i,j} |c_{i,j}(z)| \leq \frac{1}{(2\sqrt{\mu})}. \tag{A.11}$$

Thus, just as for the single channel inversion, also for the multi-channel inversion the regularization parameter μ acts as an upper bound for every resulting inverse filters.

B

Least-Square Minimization

If the transfer matrix $\boldsymbol{H}$ is underdetermined, i.e. it has more columns than rows, there will be an infinite number of CTC filter combinations that can drive the error energy $\boldsymbol{d}$ to zero. In this case, besides the minimization of the error energy, the *control effort*, i.e. the energy of the loudspeaker signals, is also minimized. This extra constraint added to the cost function leads to a single optimal solution to this minimization problem.

Such minimization requirements can be cast as a constrained optimization problem using Lagrange multipliers (Nelson and Elliott, 1995). The cost function to be minimized is now

$$J(z) = -\boldsymbol{v}^*\boldsymbol{v} - \boldsymbol{d}^*\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\boldsymbol{d}, \tag{B.1}$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers. This equation can be expanded to reveal its dependency on $\boldsymbol{C}$.

$$\begin{aligned} J(z) \quad = \quad & -\boldsymbol{b}^*\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{b} - \boldsymbol{b}^*\left(\boldsymbol{H}\boldsymbol{C} - \boldsymbol{I}\cdot e^{-z\Delta}\right)^*\boldsymbol{\lambda} - \\ & \boldsymbol{\lambda}^*\left(\boldsymbol{H}\boldsymbol{C} - \boldsymbol{I}\cdot e^{-z\Delta}\right)\boldsymbol{b}. \end{aligned} \tag{B.2}$$

The filters for each ear are optimized independently. As the optimization depends on the input signal, Kirkeby and Nelson (1999) suggest to set $b_j(n) = \delta(n)$, the Dirac delta function, as this gives the worst-case scenario for the optimization. The new cost function to be minimized is now

$$J(z) = -\boldsymbol{c}_i^*\boldsymbol{c}_i - \left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right)^*\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right), \tag{B.3}$$

where $\boldsymbol{c}_i$ is the i^{th} column of $\boldsymbol{C}$ and $\boldsymbol{y}_i$ is the i^{th} column of $\boldsymbol{I}\cdot e^{-z\Delta}$.

The derivative with respect to a vector is defined as

$$\frac{\partial z}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial z}{\partial x_1} \\ \vdots \\ \frac{\partial z}{\partial x_n} \end{bmatrix}. \tag{B.4}$$

According to Nelson and Elliott (1995), deriving J with respect to both $\boldsymbol{c}$ and $\boldsymbol{\lambda}$ results in

$$\partial J/\partial \boldsymbol{c} = -2\boldsymbol{c}_i - 2\boldsymbol{H}^*\boldsymbol{\lambda}, \tag{B.5}$$

$$\partial J/\partial \boldsymbol{\lambda} = -2\left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right). \tag{B.6}$$

The minimum value of $\boldsymbol{c}_i$ is given by setting both derivatives to zero and substituting them into each other to isolate $\boldsymbol{c}_i$. Assuming that $\boldsymbol{H}\boldsymbol{H}^*$ is not singular, then

$$\boldsymbol{c}_i = \boldsymbol{H}^*\left(\boldsymbol{H}\boldsymbol{H}^*\right)^{-1}\boldsymbol{y}_j, \tag{B.7}$$

that is equivalent in matrix form to

$$\boldsymbol{C} = \boldsymbol{H}^*\left(\boldsymbol{H}\boldsymbol{H}^*\right)^{-1} e^{-z\Delta}. \tag{B.8}$$

B.1 Regularized Least-Square Minimization

The Lagrange multipliers $\boldsymbol{\lambda}$ can be related to the effort made by the filters to satisfy the constraints imposed in eq. (B.3). This minimization can be regularize by restricting the effort made by the CTC filters, resulting in the new cost function

$$J(z) = -\boldsymbol{c}_i^*\boldsymbol{c}_i - \left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right)^*\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right) + \mu\boldsymbol{\lambda}^*\boldsymbol{\lambda}. \tag{B.9}$$

Equation (B.5) remains unaltered while eq. (B.6) changes to

$$\partial J/\partial \boldsymbol{\lambda} = -2\left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right) + 2\mu\boldsymbol{\lambda}, \tag{B.10}$$

which yields

$$\boldsymbol{C} = \boldsymbol{H}^*\left(\boldsymbol{H}\boldsymbol{H}^* + \mu\boldsymbol{I}\right)^{-1} e^{-z\Delta}. \tag{B.11}$$

B.2 Weighted Regularized Least-Square Minimization

Substituting the ℓ_2 norm of $\boldsymbol{c}_i$ by the weighted norm eq. (5.32) in eq. (B.9) yields

$$J(z) = -\boldsymbol{c}_i^*\boldsymbol{Z}\boldsymbol{c}_i - \left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right)^*\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\left(\boldsymbol{H}\boldsymbol{c}_i - \boldsymbol{y}_i\right) + \mu\boldsymbol{\lambda}^*\boldsymbol{\lambda}. \tag{B.12}$$

The larger the weight z_i given for a given loudspeaker, the higher the effort made by the algorithm to minimize this loudspeaker's energy and thus the smallest the energy of the filters related to this loudspeaker.

Equation (B.10) remains unaltered while eq. (B.5) changes to

$$\partial J/\partial \boldsymbol{c} = -2\boldsymbol{Z}\boldsymbol{c}_i - 2\boldsymbol{H}^*\boldsymbol{\lambda}, \tag{B.13}$$

which yields

$$\boldsymbol{C} = \boldsymbol{Z}^{-1}\boldsymbol{H}^* \left(\boldsymbol{H}\boldsymbol{Z}^{-1}\boldsymbol{H}^* + \mu\boldsymbol{I}\right)^{-1} e^{-z\Delta} \tag{B.14}$$

as long as $\boldsymbol{Z}^{-1}$ exists, which is the case if $\boldsymbol{Z}$ is a diagonal matrix and $\forall z_i > 0$.

If, however, $\boldsymbol{Z}$ is not directly invertible, another solution, based in the method described in (Ruffini et al., 2002) can be used. First $\partial J/\partial \boldsymbol{\lambda}$ is multiplied by $\boldsymbol{H}^*$ and added to $\partial J/\partial \boldsymbol{c}$, giving

$$(\boldsymbol{Z} + \boldsymbol{H}^*\boldsymbol{H})\boldsymbol{c}_i = \boldsymbol{H}(\boldsymbol{\lambda} + \boldsymbol{y} + \mu\boldsymbol{\lambda}). \tag{B.15}$$

The dependency on $\boldsymbol{\lambda}$ is eliminated by multiplying $\partial J/\partial \boldsymbol{c}$ by $\boldsymbol{H}$, isolating $\boldsymbol{\lambda}$ and substituting it into eq. (B.15), which, after some algebraic manipulations, results in

$$\boldsymbol{C} = \left[\boldsymbol{H}^*\boldsymbol{H} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Z} + \mu\boldsymbol{P}\boldsymbol{Z}\right]^{-1} e^{-z\Delta}, \tag{B.16}$$

where $\boldsymbol{P} = \boldsymbol{H}^*(\boldsymbol{H}\boldsymbol{H}^*)^{-1}\boldsymbol{H}$.

List of References

Ajdler, T.; Sbaiz, L., and Vetterli, M. (Sept. 2007). "Dynamic measurement of room impulse responses using a moving microphone." In: *J. Acoust. Soc. Am.* 122.3, p. 1636 (cit. on pp. 65, 149).

Akeroyd, M. A.; Chambers, J.; Bullock, D.; Palmer, A. R.; Summerfield, A. Q.; Nelson, P. A., and Gatehouse, S. (2007). "The binaural performance of a cross-talk cancellation system with matched or mismatched setup and playback acoustics". In: *J. Acoust. Soc. Am.* 121.2, pp. 1056–1069 (cit. on pp. 3, 83, 116, 117, 120, 123, 127, 129–132, 134).

Algazi, V. R.; Duda, R.; Thompson, D., and Avendano, C. (2001). "The CIPIC HRTF database". In: *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 99–102 (cit. on pp. 2, 21, 23, 28).

Aretz, M. (2012). "Combined Wave And Ray Based Room Acoustic Simulations In Small Rooms". PhD. RWTH Aachen University (cit. on p. 52).

Atal, B. S.; Hill, M., and Schroeder, M. R. (1966). *Apparent Sound Source Translator* (cit. on p. 3).

Bai, M. R. and Lee, C.-C. (2006). "Objective and subjective analysis of effects of listening angle on crosstalk cancellation in spatial sound reproduction". In: *J. Acoust. Soc. Am.* 120.4, p. 1976 (cit. on pp. 83, 116, 117, 134).

Bai, M. R.; Shih, G.-Y., and Lee, C.-C. (2007). "Comparative study of audio spatializers for dual-loudspeaker mobile phones". In: *J. Acoust. Soc. Am.* 121.1, p. 298 (cit. on p. 136).

Bauck, J. L. and Cooper, D. H. (1992). "Generalized Transaural Stereo". In: *93rd AES Convention*. San Francisco, USA (cit. on pp. 81, 82).

Bauck, J. L. and Cooper, D. H. (1993). "On Transaural Stereo for Auralization". In: *95th AES Convention*. Vol. 3728. New York (cit. on p. 3).

Bauck, J. L. and Cooper, D. H. (1996). "Generalized transaural stereo and applications". In: *J. Audio Eng. Soc.* 44.9, pp. 683–705 (cit. on p. 77).

Bauer, B. B. (1961). "Stereophonic Earphones and Binaural Loudspeakers". In: *J. Audio Eng. Soc.* 9.2, pp. 148–151 (cit. on p. 3).

Begault, D. R.; Wenzel, E. M.; Lee, A. S., and Anderson, M. R. (Oct. 2000). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". In: *108th AES Convention*. Vol. 49. 10 (cit. on p. 14).

Blauert, J. (1969). "Sound localization in the median plane". In: *Acustica* 22, pp. 205–213 (cit. on p. 18).

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT Press (cit. on pp. 1, 15, 17, 18, 45, 47, 83, 104, 114).

Borish, J. and Angell, J. B. (1983). "An Efficient Algorithm for Measuring the Impulse Response using Pseudorandom Noise". In: *J. Audio Eng. Soc.* 31 (cit. on p. 11).

Bouchard, M.; Norcross, S. G., and Soulodre, G. (2006). "Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization". In: *121st AES Convention*. San Francisco, USA (cit. on pp. 9, 13).

Boyd, S. and Vandenberghe, L. (June 2004). *Convex Optimization Theory*. New York: Cambridge University Press, p. 730 (cit. on p. 86).

Breebaart, J. and Kohlrausch, A. (2001). "The perceptual (ir) relevance of HRTF magnitude and phase spectra". In: *110th AES Convention*. Amsterdam, The Netherlands: Audio Engineering Society; 1999 (cit. on p. 47).

Bronkhorst, A. W. (1995). "Localization of real and virtual sound sources". In: *J. Acoust. Soc. Am.* 98.5, p. 2542 (cit. on pp. 2, 21, 23, 101).

Brungart, D. S. and Rabinowitz, W. M. (Oct. 1999). "Auditory localization of nearby sources. Head-related transfer functions." In: *J. Acoust. Soc. Am.* 106.3 Pt 1, pp. 1465–79 (cit. on pp. 2, 24, 52).

Bucklein, R. (1981). "The audibility of frequency response irregularities". In: *J. Audio Eng. Soc.* 29.3, pp. 126–131 (cit. on p. 72).

Carlile, S. and Pralong, D (1994). "The location-dependent nature of perceptually salient features of the human head-related transfer functions". In: *J. Acoust. Soc. Am.* 95, pp. 3445–3459 (cit. on p. 137).

Cruzado, C. G. M. (2002). "Influence of the Acoustic Impedance of the Headphone on Psychoacoustic Effects". M.Sc. RWTH Aachen University (cit. on p. 69).

Daniel, J. (2003). "Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format". In: *23rd International Conference: Signal Processing in Audio Recording and Reproduction*. Copenhagen, Denmark (cit. on p. 1).

Dellepiane, M; Pietroni, N; Tsingos, N; Asselot, M, and Scopigno, R (2008). "Reconstructing head models from photographs for individualized 3D-audio processing". In: *27th Computer Graphics Forum*. Vol. 27. 7. Trier, Germany, pp. 1719–1727 (cit. on p. 2).

Dietrich, P.; Masiero, B., and Vorländer, M. (2012a). "On the Optimization of the Multiple Exponential Sweep Method". In: *J. Audio Eng. Soc.* Submitted (cit. on pp. 21, 33, 36, 40).

Dietrich, P.; Masiero, B.; Pollow, M.; Krechel, B., and Vorländer, M. (2012b). "Time Efficient Measurement Method for Individual HRTFs". In: *Fortschritte der Akustik – DAGA*. Darmstadt, Germany (cit. on pp. 38, 40).

Driscoll, J. and Healy, D. (1994). "Computing Fourier transforms and convolutions on the 2-sphere". In: *Advances in Applied Mathematics* (cit. on p. 47).

Duraiswami, R.; Zotkin, D. N., and Gumerov, N. (2004). "Interpolation and range extrapolation of HRTFs". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4, pp. iv–45–48 (cit. on pp. 2, 18, 47, 48, 50, 51).

Enzner, G. (2009). "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering". In: *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pp. 325–328 (cit. on pp. 65, 149).

Evans, M. J.; Angus, J. A. S., and Tew, A. I. (1998). "Analyzing head-related transfer function measurements using surface spherical harmonics". In: *J. Acoust. Soc. Am.* 104.4, pp. 2400–2411 (cit. on pp. 47, 48).

Farina, A. (2007). "Advancements in impulse response measurements by sine sweeps". In: *122nd AES Convention*. Vol. 122. Vienna, Austria (cit. on pp. 12, 153).

Fastl, H. and Zwicker, E. (2007). *Psychoacoustics: Facts and Models*. Springer, p. 462 (cit. on p. 90).

Fazi, F. M. (2010). "Sound Field Reproduction". PhD thesis. Uuiversity of Southampton, p. 297 (cit. on p. 9).

Fels, J. (2008). "From children to adults: How binaural cues and ear canal impedances grow". Ph.D. RWTH Aachen University (cit. on pp. 2, 13).

Fels, J. and Masiero, B. (June 2011). "Binaural reproduction technologies for studies on dichotic and selective binaural hearing : Headphone reproduction". In: *Proceeding of Forum Acusticum*. Aalborg, Denmark (cit. on p. 67).

Fielder, L. D. (2003). "Analysis of traditional and reverberation-reducing methods of room equalization". In: *J. Audio Eng. Soc.* 51.1/2, pp. 3–26 (cit. on p. 90).

Freeden, W. and Windheuser, U. (Jan. 1997). "Combined Spherical Harmonic and Wavelet Expansion—A Future Concept in Earth's Gravitational Determination". In: *Applied and Computational Harmonic Analysis* 4.1, pp. 1–37 (cit. on p. 66).

Freeland, F. P.; Biscainho, L. W., and Diniz, P. S. R. (2007). "HRTF interpolation through direct angular parameterization". In: *Proc. 2007 IEEE Intern. Symposium on Circuits and Systems* May, pp. 1823–1826 (cit. on p. 47).

Fukudome, K; Suetsugu, T; Ueshin, T; Idegami, R, and Takeya, K (Aug. 2007). "The fast measurement of head related impulse responses for all azimuthal directions using the continuous measurement method with a servo-swiveled chair". In: *Applied Acoustics* 68.8, pp. 864–884 (cit. on p. 65).

Fundalewicz, J. (2012). "Mobiles transaurales Wiedergabesystem mit videobasiertem Head-Tracking". Diploma. RWTH Aachen (cit. on p. 150).

Gardner, W. G. (1997). "3-D audio using loudspeakers". PhD. Massachusetts Institute of Technology (cit. on pp. 3, 78, 83, 95, 116, 117, 136).

Gardner, W. G. and Martin, K. D. (1995). "HRTF measurements of a KEMAR". In: *J. Acoust. Soc. Am.* 97.6, pp. 3907–3908 (cit. on p. 123).

Gerzon, M. (1973). "PERIPHONY: WITH-HEIGHT SOUND REPRODUCTION". In: *J. Audio Eng. Soc.* 21.1, pp. 2–10 (cit. on p. 1).

Golay, M. (Apr. 1961). "Complementary series". In: *IEEE Transactions on Information Theory* 7.2, pp. 82–87 (cit. on p. 11).

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press (cit. on p. 154).

Goupell, M.; Majdak, P., and Laback, B. (Feb. 2010). "Median-plane sound localization as a function of the number of spectral channels using a channel vocoder". In: *J. Acoust. Soc. Am.* 127.2, pp. 990–1001 (cit. on p. 126).

Guillon, P; Zolfaghari, R; Epain, N; van Schaik, A; Jin, C. T.; Hetherington, C; Thorpe, J, and Tew, A (2012). "Creating the Sydney York Morphological and Acoustic Recordings of Ears Database". In: *2012 IEEE International Conference on Multimedia and Expo*, pp. 461–466 (cit. on p. 2).

Hammershøi, D. and Møller, H. (2005). "Binaural Technique—Basic Methods for Recording, Synthesis and Reproduction". In: *Communication Acoustics*. Ed. by J. Blauert. Springer-Verlag, p. 379 (cit. on pp. 41, 46, 64, 71).

Hartmann, W. M. (Nov. 1999). "How We Localize Sound". en. In: *Physics Today* 52.11, p. 24 (cit. on pp. 15, 19).

Hofman, P. M.; Van Riswick, J. G., and Van Opstal, A. J. (Sept. 1998). "Relearning sound localization with new ears." In: *Nature neuroscience* 1.5, pp. 417–21 (cit. on p. 14).

Jungmann, J. O.; Mazur, R.; Kallinger, M.; Mei, T., and Mertins, A. (Aug. 2012). "Combined Acoustic MIMO Channel Crosstalk Cancellation and Room Impulse Response Reshaping". In: *IEEE Transactions on Audio, Speech and Language Processing* 20.6, pp. 1829–1842 (cit. on pp. 97, 150).

Katz, B. F. G. (Nov. 2001). "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation". en. In: *J. Acoust. Soc. Am.* 110.5, p. 2440 (cit. on p. 2).

Kim, S.-M. and Wang, S. (2003). "A Wiener filter approach to the binaural reproduction of stereo sound". In: *J. Acoust. Soc. Am.* 114.6, p. 3179 (cit. on p. 89).

Kim, Y.; Deille, O., and Nelson, P. A. (Oct. 2006). "Crosstalk cancellation in virtual acoustic imaging systems for multiple listeners". In: *Journal of Sound and Vibration* 297.1-2, pp. 251–266 (cit. on p. 82).

Kirkeby, O. and Nelson, P. A. (1999). "Digital Filter Design for Inversion Problems in Sound Reproduction". In: *J. Audio Eng. Soc.* 47.7/8, pp. 583–595 (cit. on pp. 3, 81, 84, 90, 155).

Kirkeby, O.; Nelson, P. A.; Hamada, H., and Orduna-bustamante, F. (Mar. 1998a). "Fast deconvolution of multichannel systems using regularization". In: *IEEE Transactions on Speech and Audio Processing* 6.2, pp. 189–194 (cit. on pp. 86, 89, 90).

Kirkeby, O.; Nelson, P. A., and Hamada, H. (1998b). "The 'stereo dipole' a virtual source imaging system using two closely spaced loudspeakers". In: *J. Audio Eng. Soc.* 46.5, pp. 387–395 (cit. on pp. 78, 86).

Kistler, D. J. and Wightman, F. L. (1992). "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction". In: *J. Acoust. Soc. Am.* 91.3, pp. 1637–1647 (cit. on pp. 47, 66).

Kleber, J and Vorländer, M (2001). "Messung von Gehöreingangsimpedanzen des freien Ohres und des abgeschlossenen Ohres mit Otoplastiken, Im-Ohr-Hörgeräten oder Kopfhörern". In: *Fortschritte der Akustik – DAGA* (cit. on p. 69).

Köring, J and Schmitz, A. (1993). "Simplifying Cancellation of Cross-Talk for Playback of Head-Related Recordings in a Two-Speaker System". In: *Acustica* 79.December 1992, pp. 221–232 (cit. on pp. 3, 93).

Krechel, B. (2012). "Schnelle Messung von individuellen HRTFs mit kontinuierlichen MIMO-Verfahren". M.Sc. RWTH Aachen (cit. on pp. 30, 31, 64, 65, 149).

Kulkarni, A. and Colburn, H. (1998). "Role of spectral detail in sound-source localization". In: *Nature* 396.December, pp. 747–749 (cit. on pp. 47, 102, 148).

Kuttruff, H. (2000). *Room Acoustics*. CRC Press (cit. on p. 11).

Langendijk, E. H. A. and Bronkhorst, A. W. (2000). "Fidelity of three-dimensional-sound reproduction using a virtual". In: *J. Acoust. Soc. Am.* 107.1, pp. 528–537 (cit. on p. 47).

Langendijk, E. H. A. and Bronkhorst, A. W. (2002). "Contribution of spectral cues to human sound localization". In: *J. Acoust. Soc. Am.* 112.4, p. 1583 (cit. on p. 137).

Larcher, V.; Jot, J.-M., and Vandernoot, G. (1998). "Equalization methods in binaural technology". In: *105th AES Conference*. Vol. 4858. San Francisco, USA (cit. on pp. 75, 121).

Lentz, T. (2006). "Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments". In: *J. Audio Eng. Soc.* 54.4, pp. 283–294 (cit. on pp. 3, 78, 93, 116).

Lentz, T. (2007). "Binaural technology for virtual reality". PhD thesis. Institut für Technische Akustik, RWTH-Aachen (cit. on pp. 1, 3, 18, 21, 23, 47, 50, 51, 54–57, 59, 78, 94).

Li, Z. and Duraiswami, R. (2006). "Headphone-based reproduction of 3D auditory scenes captured by spherical/hemispherical microphone arrays". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Toulouse, France, pp. 337–340 (cit. on p. 2).

Macmillan, N. A. and Creelman, C. D. (2005). *Detection Theory*. Mahwah, New Jersey: LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS (cit. on pp. 108, 112).

Macpherson, E. A. and Middlebrooks, J. C. (2002). "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited". In: *J. Acoust. Soc. Am.* 111.5, p. 2219 (cit. on pp. 18, 137).

Majdak, P.; Balazs, P., and Laback, B. (2007). "Multiple exponential sweep method for fast measurement of head-related transfer functions". In: *J. Audio Eng. Soc.* 55.7/8, p. 623 (cit. on pp. 2, 21, 23, 33–37, 41, 63).

Majdak, P.; Goupell, M., and Laback, B. (Feb. 2010). "3-D localization of virtual sound sources: effects of visual environment, pointing method, and training". In: *Atten. Percept. Psychophys.* 72.2, pp. 454–469 (cit. on pp. 14, 121).

Majdak, P.; Goupell, M., and Laback, B. (2011). "Two-dimensional localization of virtual sound sources in cochlear-implant listeners". In: *Ear & Hearing* 32, pp. 198–208 (cit. on p. 126).

Majdak, P.; Masiero, B., and Fels, J. (2012). "Human sound localization performance in individualized and non-individualized crosstalk cancellation systems". In: *J. Acoust. Soc. Am.* Submitted (cit. on p. 100).

Mammone, R. (Nov. 1999). "Inverse Problems and Signal Reconstruction". In: *Digital Signal Processing Handbook*. Ed. by V. K. Madisetti and D. B. Williams. Electrical Engineering Handbook. Boca Raton: CRC Press, pp. 1–4 (cit. on p. 7).

Masiero, B. and Fels, J. (Mar. 2011a). "Equalization for Binaural Synthesis with Headphone". In: *Fortschritte der Akustik – DAGA*. Düsseldorf, Germany, pp. 675–676 (cit. on p. 67).

Masiero, B. and Fels, J. (May 2011b). "Perceptually Robust Headphone Equalization for Binaural Reproduction". In: *130th AES Convention*. London, England, pp. 1–7 (cit. on pp. 67, 71).

Masiero, B. and Pollow, M. (May 2010). "A review of the compressive sampling framework in the lights of spherical harmonics: applications to distributed spherical arrays". In: *Second International Symposium on Ambisonics and Spherical Acoustics*. Paris, France (cit. on pp. 66, 149).

Masiero, B. and Qiu, X. (Mar. 2009). "Two Listeners Crosstalk Cancellation System Modelled by Four Point Sources and Two Rigid Spheres". In: *Acta Acustica united with Acustica* 95.2, pp. 379–385 (cit. on p. 82).

Masiero, B. and Vorländer, M. (2012). "A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters". In: *IEEE Transactions on Audio, Speech and Language Processing*. Submitted (cit. on p. 77).

Masiero, B.; Pollow, M., and Fels, J. (June 2011a). "Design of a Fast Broadband Individual Head-Related Transfer Function Measurement System". In: *Proceeding of Forum Acusticum*. Aalborg, Denmark (cit. on pp. 21, 28).

Masiero, B.; Fels, J., and Vorländer, M. (2011b). "Review of the crosstalk cancellation filter technique". In: *Proceedings of the International Conference on Spatial Audio*. Ed. by M. Kob. Detmold, Germany (cit. on p. 77).

Masiero, B.; Pollow, M.; Dietrich, P., and Fels, J. (2012). "Design of a Fast Measurement System for the Range Extrapolation of HRTFs". In: *J. Audio Eng. Soc.* Accepted (cit. on p. 21).

Middlebrooks, J. C. (Sept. 1999a). "Individual differences in external-ear transfer functions reduced by scaling in frequency." In: *J. Acoust. Soc. Am.* 106.3 Pt 1, pp. 1480–92 (cit. on p. 47).

Middlebrooks, J. C. (Sept. 1999b). "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency." In: *J. Acoust. Soc. Am.* 106.3 Pt 1, pp. 1493–510 (cit. on pp. 14, 17, 19, 20, 126, 137).

Mills, A. W. (1972). *Auditory localization*. Ed. by T. J. Vol. 2. New York Academic Press (cit. on p. 19).

Minnaar, P.; Flemming, C.; Møller, H.; Olesen, S. K., and Plogsties, J. (1999). "Audibility of All-Pass Components in Binaural Synthesis". In: *106th AES Convention* (cit. on p. 73).

Mokhtari, P.; Takemoto, H.; Nishimura, R., and Kato, H. (2007). "Comparison of Simulated and Measured HRTFs: FDTD Simulation Using MRI Head Data". In: *123rd AES Convention* (cit. on p. 2).

Møller, H. (1992). "Fundamentals of binaural technology". In: *Applied Acoustics* 36.3-4, pp. 171–218 (cit. on pp. 1, 3, 67–69).

Møller, H.; Sørensen, M. F.; Hammershøi, D., and Jensen, C. B. (1995a). "Head-Related Transfer Functions of Human Subjects". In: *J. Audio Eng. Soc.* 43.5, pp. 300–310 (cit. on pp. 2, 13, 21, 23, 25).

Møller, H.; Hammershøi, D.; Jensen, C. B., and Sørensen, M. F. (1995b). "Transfer Characteristics of Headphones Measured on Human Ears". In: *J. Audio Eng. Soc.* 43.4, pp. 203–217 (cit. on pp. 3, 22, 68).

Møller, H.; Sørensen, M. F.; Jensen, C. B., and Hammershøi, D. (1996). "Binaural technique: Do we need individual recordings?" In: *J. Audio Eng. Soc.* 44.6, pp. 451–469 (cit. on pp. 13, 101).

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing*. Emerald Group (cit. on p. 104).

Morimoto, M. and Aokata, H. (1984). "Localization cues in the upper hemisphere". In: *J Acoust Soc Jpn (E)* 5, pp. 165–173 (cit. on pp. 18, 119).

Müller, S. (1999). "Digitale Signalverarbeitung für Lautsprecher". Ph.D. RWTH Aachen University, p. 263 (cit. on p. 72).

Müller, S. (2008). "Measurement of Transfer Functions and Impulse Responses". en. In: *Handbook of Signal Processing in Acoustics*, ed. by D. Havelock; S. Kuwano, and M. Vorländer. Vol. -1. New York, NY: Springer New York, pp. 65–85 (cit. on p. 10).

Müller, S. and Massarani, P. (2001). "Transfer-function measurement with sweeps". In: *J. Audio Eng. Soc.* 49.6, pp. 443–471 (cit. on pp. 10, 11, 33).

Nelson, P.; Kirkeby, O.; Takeuchi, T., and Hamada, H. (July 1997). "Sound Fields for the Production of Virtual Acoustic Images". In: *Journal of Sound and Vibration* 204.2, pp. 386–396 (cit. on p. 132).

Nelson, P. A. and Elliott, S. J. (1995). *Active Control of Sound*. 3rd. San Diego, CA: Academic Press, p. 436 (cit. on pp. 8, 155, 156).

Nelson, P. A. and Rose, J. F. W. (Sept. 2006). "The time domain response of some systems for sound reproduction". In: *Journal of Sound and Vibration* 296.3, pp. 461–493 (cit. on pp. 77, 89).

Norcross, S. G. and Bouchard, M. (2007). "Multichannel Inverse Filtering with Minimal-Phase Regularization". In: *123rd AES Convention*, pp. 1–8 (cit. on pp. 9, 90–92).

Oberem, J. (2012). "Analysis of different equalization methods for binaural reproduction". M.Sc. RWTH Aachen University (cit. on pp. 69, 99).

Oppenheim, A. and Schafer, R. (1989). *Discrete-Time Signal Processing*. 1st. Vol. Signal Pro. Prentice Hall (cit. on pp. 5–7).

Papoulis, A. (1977). *Signal Analysis*. McGraw-Hill, p. 431 (cit. on p. 88).

Paquier, M. and Koehl, V. (2010). "Audibility of headphone positioning variability". In: *128th AES Convention* (cit. on p. 69).

Parodi, Y. L. (2008). "Analysis of design parameters for crosstalk cancellation filters applied to different loudspeaker configurations". In: *125th AES Convention*. San Francisco, USA (cit. on p. 78).

Parodi, Y. L. (2010). "A Systematic Study of Binaural Reproduction Systems Through Loudspeakers: A Multiple Stereo-Dipole Approach". PhD. Aalborg University (cit. on p. 84).

Parodi, Y. L. and Rubak, P. (2011). "A Subjective Evaluation of the Minimum Channel Separation for Reproducing Binaural Signals over Loudspeakers". In: *J. Audio Eng. Soc.* 59.7-8, pp. 487–497 (cit. on pp. 117, 134, 141).

Parodi, Y. L. and Rubak, P. (Sept. 2010). "Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers." In: *J. Acoust. Soc. Am.* 128.3, pp. 1045–55 (cit. on pp. 83, 93).

Paul, S. (Sept. 2009). "Binaural Recording Technology: A Historical Review and Possible Future Developments". In: *Acta Acustica united with Acustica* 95.5, pp. 767–788 (cit. on pp. 2, 10).

Peltonen, T. (2000). "A Multichannel Measurement System for Room Acoustics Analysis". M.Sc. Helsinki University of Technology (cit. on p. 11).

Pelzer, S.; Masiero, B., and Vorländer, M. (2011). "3D Reproduction of Room Acoustics using a Hybrid System of Combined Crosstalk Cancellation and Ambisonics Playback". In: *Proceedings of the International Conference on Spatial Audio*. Ed. by M. Kob. Detmold, Germany, pp. 297–301 (cit. on p. 150).

Perrett, S and Noble, W (Oct. 1997). "The effect of head rotations on vertical plane sound localization." In: *J. Acoust. Soc. Am.* 102.4, pp. 2325–32 (cit. on p. 137).

Pollow, M.; Nguyen, K.-V.; Warusfel, O.; Carpentier, T.; Müller-Trapet, M.; Vorländer, M., and Noisternig, M. (Jan. 2012a). "Calculation of Head-Related Transfer Functions for Arbitrary Field Points Using Spherical Harmonics Decomposition". In: *Acta Acustica united with Acustica* 98.1, pp. 72–82 (cit. on pp. 24, 47, 50, 51, 63).

Pollow, M.; Masiero, B.; Dietrich, P.; Fels, J., and Vorländer, M. (2012b). "Fast Measurement System for Spatially Continuous Individual HRTFs". In: *Spatial Audio in Today's 3D World - AES 25th UK Conference*, pp. 1–8 (cit. on p. 21).

Pollow, M.; Dietrich, P.; Masiero, B.; Fels, J., and Vorländer, M. (2012c). "Modal sound field representation of HRTFs". In: *Fortschritte der Akustik – DAGA*. Darmstadt, Germany (cit. on p. 65).

Pralong, D and Carlile, S. (Dec. 1996). "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space." In: *J. Acoust. Soc. Am.* 100.6, pp. 3785–93 (cit. on p. 3).

Pulkki, V. (1997). "Virtual sound source positioning using vector base amplitude panning". In: *J. Audio Eng. Soc.* 45.6, pp. 456–466 (cit. on pp. 1, 141).

Qiu, X.; Masiero, B., and Vorländer, M. (2009). "Experimental Study on Channel Separation of Crosstalk Cancellation System with Mismatched Sound Sources". In: *The 10th Western Pacific Acoustics Conference*. Beijing, China (cit. on pp. 83, 117).

Qu, T.; Xiao, Z.; Gong, M.; Huang, Y.; Li, X., and Wu, X. (Aug. 2009). "Distance-Dependent Head-Related Transfer Functions Measured With High Spatial Resolution Using a Spark Gap". In: *IEEE Transactions on Audio, Speech and Language Processing* 17.6, pp. 1124–1132 (cit. on p. 24).

Rao, H. I. K.; Mathews, V. J., and Park, Y.-C. (Nov. 2007). "A Minimax Approach for the Joint Design of Acoustic Crosstalk Cancellation Filters". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.8, pp. 2287–2298 (cit. on p. 86).

Rayleigh, L. (Feb. 1907). "XII. On our perception of sound direction". In: *Philosophical Magazine Series 6* 13.74, pp. 214–232 (cit. on p. 18).

Ruffini, G.; Marco, J., and Grau, C. (2002). *Spherical harmonics interpolation, computation of Laplacians and Gauge Theory*. arXiv: 0206007v1 [arXiv:physics] (cit. on pp. 50, 157).

Rui, Y.; Yu, G., and Xie, B. (Sept. 2012). "Approximately calculate individual near-field head-related transfer function using an ellipsoidal head and pinnae model." en. In: *J. Acoust. Soc. Am.* Vol. 132. 3, p. 1997 (cit. on p. 2).

Rumsey, F. (2012). *Spatial Audio*. CRC Press, p. 256 (cit. on p. 1).

Sæbø, A. (2001). "Influence of Reflections on Crosstalk Cancelled Playback of Binaural Sound". Ph.D. Norwegian University of Science and Technology (cit. on p. 97).

Sartor, M. (2010). "Entwurf von Lautsprecher und Messbogen zur HRTF-Messung". Studienarbeit. RWTH Aachen (cit. on p. 26).

Schmidt, S. (2009). "Finite Element Simulation of External Ear Sound Fields for the Optimization of Eardrum-Related Measurements". Ph.D. Ruhr-Universität Bochum, p. 145 (cit. on pp. 67, 71).

Schonstein, D.; Ferr, L., and Katz, B. F. G. (2008). "Comparison of headphones and equalization for virtual auditory source localization". In: *Acoustics '08, Paris* (cit. on p. 121).

Schröder, D.; Wefers, F.; Pelzer, S.; Rausch, D. S.; Vorländer, M., and Kuhlen, T. (2010). "Virtual Reality System at RWTH Aachen University". In: *Proceedings ICA 2010, 20th International Congress on Acoustics :* ed. by M. Burgess. Sydney, Australia: Australian Acoustical Society, NSW Division (cit. on p. 1).

Schroeder, M. R. (Aug. 1979). "Integrated-impulse method measuring sound decay without using impulses". en. In: *J. Acoust. Soc. Am.* 66.2, p. 497 (cit. on p. 11).

Seeber, B. (2002). "Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode". Ph.D. TU München (cit. on p. 18).

Slepian, D (1964). "Prolate spheroidal wave functions, Fourier analysis and uncertainty–IV: Extension to Many Dimensions; Generalized Prolate Spherical Functions". In: *Bell Syst. Tech. J* 43.6, pp. 3009–3057 (cit. on p. 66).

Takeuchi, T. and Nelson, P. A. (2007). "Subjective and objective evaluation of the optimal source distribution for virtual acoustic imaging". In: *J. Audio Eng. Soc.* 55.11, p. 981 (cit. on pp. 3, 77, 86, 95, 117).

Takeuchi, T.; Nelson, P. a., and Hamada, H. (2001). "Robustness to head misalignment of virtual sound imaging systems". In: *J. Acoust. Soc. Am.* 109.3, p. 958 (cit. on pp. 116, 117).

Tohyama, M. and Koike, T. (1998). *Fundamentals of Acoustic Signal Processing.* Academic Press (cit. on pp. 5, 13).

Toole, F. E. (1984). "The Acoustics and Psychoacoustics of Headphones". In: *2nd AES International Conference.* Anaheim, USA (cit. on p. 69).

Vanderkooy, J. (Nov. 2010). "Rapid In-Place Measurements of Multichannel Venues". In: *129th AES Convention* (cit. on p. 33).

Völk, F. (2011a). "Inter- and Intra-Individual Variability in blocked auditory canal transfer functions of three circum-aural headphones". In: *131st AES Convention* 8465, p. 10 (cit. on p. 69).

Völk, F. (2011b). "System Theory of Binaural Synthesis". In: *131st AES Convention* 8568, p. 17 (cit. on p. 69).

de Vries, D. (1988). *Wave Field Synthesis*. Audio Engineering Society (cit. on p. 1).

Wang, L.; Yin, F., and Chen, Z. (2009). "Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights". In: *Acoustical Science and Technology* 30.6, pp. 395–403 (cit. on p. 47).

Weinzierl, S.; Giese, A., and Lindau, A. (2009). "Generalized Multiple Sweep Measurement". In: *126th AES Convention* (cit. on p. 36).

Wenzel, E. M.; Arruda, M.; Kistler, D. J., and Wightman, F. L. (July 1993). "Localization using nonindividualized head-related transfer functions". In: *J. Acoust. Soc. Am.* 94.1, pp. 111–123 (cit. on pp. 2, 13, 14).

Wightman, F. L. and Kistler, D. J. (Feb. 1989). "Headphone simulation of free-field listening. II: Psychophysical validation". In: *J. Acoust. Soc. Am.* 85.2, pp. 868–878 (cit. on pp. 2, 13, 101).

Wightman, F. L. and Kistler, D. J. (1992). "The dominant role of low-frequency interaural time differences in sound localization." In: *J. Acoust. Soc. Am.* 91.3, pp. 1648–1661 (cit. on p. 137).

Williams, E. G. (1999). *Fourier Acoustics: sound radiation and nearfield acoustical holography*. Academic Press (cit. on pp. 49, 51).

Xiang, N and Schroeder, M. R. (2003). "Reciprocal maximum-length sequence pairs for acoustical dual source measurements". In: *J. Acoust. Soc. Am.* 113, p. 2754 (cit. on p. 33).

Xie, B. and Zhang, T (2010). "The Audibility of Spectral Detail of Head-Related Transfer Functions at High Frequency". In: *Acta Acustica united with Acustica* (cit. on pp. 47, 149).

Yang, J.; Gan, W., and Tan, S.-E. (2003). "Improved sound separation using three loudspeakers". In: *Acoustic Research Letters* 4.April, pp. 47–52 (cit. on p. 93).

Zhang, W.; Zhang, M., and Kennedy, R. (2012). "On High-Resolution Head-Related Transfer Function Measurements: An Efficient Sampling Scheme". In: *IEEE Transactions on Audio, Speech and Language Processing* 20.2, pp. 575–584 (cit. on pp. 30, 31, 43, 49, 65).

Ziegelwanger, H. (2012). "Efficient modeling of the time-of-arrival in binaural reproduction of virtual sound sources". Diplom. Universit für Musik und darstellende Kunst Graz, p. 98 (cit. on pp. 29, 45, 149).

Zotkin, D. N.; Duraiswami, R.; Grassi, E., and Gumerov, N. A. (2006). "Fast head-related transfer function measurement via reciprocity". In: *J. Acoust. Soc. Am.* 120.4, p. 2202 (cit. on pp. 2, 21, 22).

Zotter, F. (Jan. 2009). "Analysis and Synthesis of Sound-Radiation with Spherical Arrays". Ph.D. University of Music and Performing Arts, Graz, Austria (cit. on pp. 24, 31, 49, 65).

Zotter, F. (2010). "Sampling Strategies for Acoustic Holography/Holophony on the Sphere". In: *Fortschritte der Akustik – DAGA* (cit. on pp. 31, 50).

Acknowledgments

It is now time to thank and acknowledge the many helping hands that led to the conclusion of this thesis.

First, I'm obliged to thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the financial support and for making it possible for me to conduct my research in a highly renowned institution in the field of acoustics.

I would like to thank Prof. Dr. rer. nat. Michael Vorländer, head of the Institute of Technical Acoustics (ITA) of the RWTH Aachen University, for welcoming me to his institute and accepting to take over the role of my scientific adviser. Your ability to disseminate your wisdom and your determination to expand the field of acoustics have impressed me very much.

I also would like to thanks Prof. Philip Nelson, FREng. for accepting the invitation to be the second referee for this thesis and providing insightful comments on the thesis.

Thanks to Dr. Gottfried Behler for his willingness to discuss any problem that might occur and almost always coming up with a practical solution. And thanks to Prof. Dr. Ing. Janina Fels for giving me the opportunity to stay one year longer at ITA participating in one of her projects and for the many suggestions and advice in the final stages of this dissertation.

During the first year of this thesis I was lucky to work with Dr. Xiaojun Qiu. Thank you for the very enriching collaboration. And during the last year of my thesis I was again lucky to work with Dr. DI Pjotr Majdak at the Acoustics Research Institute of the Austrian Academy of Sciences. I would like to express my greatest gratitude for you having me there and for the great collaboration. Also many thanks to Michael Mihocic for running the localization tests and helping out in all possible ways.

Special thanks go to the staff of the mechanical and electronic workshops at ITA, namely Rolf Kaldenbach, Hans-Jürgen Dilly, Uwe Schlömer, Thomas Schäfer, and all apprentices, for the outstanding work of bringing to life ideas that otherwise would have stayed only on paper; and to the institute's secretariat, namely Ulrike Görgens (*in memoriam*),

Wilma Vonhoegen, and Karin Charlier, for making our life easier and less bureaucratic.

It is said that "a chain is only as strong as its weakest link". I'm glad I was in an environment with colleagues that were very "strong" in the very many different field of acoustics. I am greatly indebted to some colleagues who made a substantial contribution to this thesis: to Martin Pollow for endless discussions about all aspects of dealing with "spherical acoustics" and for the words of advice about doing a "round the world" trip; to Pascal Dietrich who helped in all matters related to acoustic measurement and initiated all self-help groups we had at ITA; and to the VR-Crew Frank Wefers and Sönke Pelzer who always supported me in the many challenges of spatial audio reproduction. Special thanks to Johannes Klein for helping with the daunting dark pictures. For all other colleagues—Dr. Andreas Frank, Dr. Dirk Schröder, Elena Shabalina, Dr. Elzbieta Nowicka, Ingo Witew, Jan Köhler, Dr. Marc Aretz, Markus Müller-Trappet, Martin Guski, Matthias Lievens, Oliver Strauch, Ramona Bomhardt, Renzo Vitale, Rob Opdam, Roman Scharrer, Dr. Sebastian Fingerhuth, and Xun Wang—thanks for the friendly and productive atmosphere.

I also would like to thank Dr. Marcio de Avelar Gomes for opening up the first opportunity for me to come to Aachen and Dr. João Henrique Diniz Guimarães for all the help in my first days in Aachen.

During the five year I spent in ITA I had the chance to work with many very skilled students who contributed in one way of another to the development of this thesis. Björn Kutzner, Souptik Barua, Srikanth Korse, Thomas Bierbaums, Gengliang Han, Malte Sartor, Jörg Seidler, Mario Otten, Jochen Giese, Johnny Nahas, Johannes Fundalewicz, Benedikt Koppers Benedikt Krechel. I would like to thank every single one of you for your trust and hard work. Special thanks to Josefa Oberem for so orderly conducting the perceptual tests.

On a more personal level, I would like to thank all players of the Rugby Club Aachen for providing me with such a good atmosphere where I could (almost always) forget about the pressures of my daily work.

Being away from home is not always easy. The Pfaff family was my step-family throughout these years. They were always there when I needed them, be it when I broke a tooth or had a shoulder surgery or just when I did not have any plans for Easter.

Last, but above all, I would like to thank my family. Even though you were not always physically present, you were always there for me. Special thanks goes to my father for always asking when I was going to

be done with this dissertation, for proof-reading and greatly improving the only parts of this dissertation that will probably be read by anyone else, and for introducing me to the world of buying wine. I would also like to thank my little sister for talking with me from the other side of the world when I was just feeling down and for making sure that I did not work all through the night just before the deadline. And to my mother for being the loving person that holds our family that is spread all over the world together. This thesis is dedicated to the three of you!

Curriculum Vitæ

Personal Data

	Bruno Sanches Masiero
20.12.1981	born in São Paulo, Brazil as son of Vera Lúcia und Paulo Masiero

Education

02/1989–12/1996	Primary School in São Carlos (SP), Brazil
02/1997–12/1999	Secondary School in São Carlos (SP), Brazil

Higher Education

02/2000–12/2005	Bachelor's degree in Electrical Engineering at Universidade de São Paulo, Brazil Major: Telecommunication
10/2004–06/2005	Free mover at the RWTH Aachen University
01/2006–07/2007	Master's degree in Electrical Engineering at Universidade de São Paulo, Brazil Specialization: Electronic Systems

Employments

06/2003–05/2004	Student worker at the Department of Music, Universidade de São Paulo, Brazil
06/2004–09/2004	Internship at Siemens AG, Munich
02/2005–05/2005	Student worker at the Institute of Technical Acoustics, RWTH Aachen University
10/2007–09/2011	Stipendiary from CNPq, Brazil, at the Institute of Technical Acoustics, RWTH Aachen University
10/2011–10/2012	Research Assistant at the Institute of Technical Acoustics, RWTH Aachen University

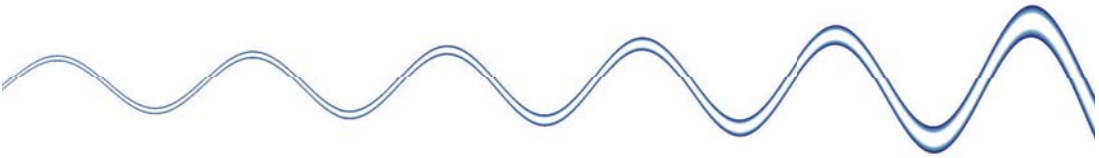
Bisher erschienene Bände der Reihe

Aachener Beiträge zur Technischen Akustik

ISSN 1866-3052

1	Malte Kob	Physical Modeling of the Singing Voice ISBN 978-3-89722-997-6 40.50 EUR
2	Martin Klemenz	Die Geräuschqualität bei der Anfahrt elektrischer Schienenfahrzeuge ISBN 978-3-8325-0852-4 40.50 EUR
3	Rainer Thaden	Auralisation in Building Acoustics ISBN 978-3-8325-0895-1 40.50 EUR
4	Michael Makarski	Tools for the Professional Development of Horn Loudspeakers ISBN 978-3-8325-1280-4 40.50 EUR
5	Janina Fels	From Children to Adults: How Binaural Cues and Ear Canal Impedances Grow ISBN 978-3-8325-1855-4 40.50 EUR
6	Tobias Lentz	Binaural Technology for Virtual Reality ISBN 978-3-8325-1935-3 40.50 EUR
7	Christoph Kling	Investigations into damping in building acoustics by use of downscaled models ISBN 978-3-8325-1985-8 37.00 EUR
8	Joao Henrique Diniz Guimaraes	Modelling the dynamic interactions of rolling bearings ISBN 978-3-8325-2010-6 36.50 EUR
9	Andreas Franck	Finite-Elemente-Methoden, Lösungsalgorithmen und Werkzeuge für die akustische Simulationstechnik ISBN 978-3-8325-2313-8 35.50 EUR

10	Sebastian Fingerhuth	Tonalness and consonance of technical sounds ISBN 978-3-8325-2536-1 42.00 EUR
11	Dirk Schröder	Physically Based Real-Time Auralization of Interactive Virtual Environments ISBN 978-3-8325-2458-6 35.00 EUR
12	Marc Aretz	Combined Wave And Ray Based Room Acoustic Simulations Of Small Rooms ISBN 978-3-8325-3242-0 37.00 EUR
13	Bruno Sanches Masiero	Individualized Binaural Technology. Measurement, Equalization and Subjective Evaluation ISBN 978-3-8325-3274-1 36.50 EUR



In this work the importance of individualization in binaural technique is investigated. The results extend the present knowledge on the efficient measurement of individual head-related transfer functions (HRTFs) and highlight the importance of individual equalization filters in binaural reproduction, using both loudspeakers and headphones.

An innovative measurement setup is developed to allow the fast acquisition of individual HRTFs. The hardware is designed to be compatible with the range extrapolation technique. An individual HRTF dataset with 4000 directions can be measured in less than 6 minutes with this new setup.

Further, a framework is presented that integrates causality constraints to the regularized frequency domain calculation of crosstalk cancellation (CTC) filters. This framework also addresses the switching of active loudspeakers applying a weighted filter calculation method. A sound localization test showed that individualized CTC systems provide performance similar to that of binaural listening while nonindividualized CTC systems provide a significantly lower localization performance.

Finally, a robust individual headphone equalization method is proposed. Perceptual tests showed that, in all but one of the tested situations, no audible differences between the original sound source and its binaural auditory display could be perceived.

λογος

ISSN 1866-3052

ISBN 978-3-8325-3274-1

Logos Verlag Berlin