Binaural Resynthesis of Acoustic Environments. Technology and Perceptual Evaluation.

Alexander Lindau
Max Planck Institute for Empirical Aesthetics

Binaural Resynthesis of Acoustical Environments.
Technology and Perceptual Evaluation.

vorgelegt von
Alexander Lindau, M.A.
aus Berlin

von der Fakultät I – Geisteswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Hans-Liudger Dienel
Gutachter:	Prof. Dr. Stefan Weinzierl
Gutachter:	Prof. Dr. Michael Vorländer

Tag der wissenschaftlichen Aussprache: 19. Mai 2014

Berlin 2014
D83

Alexander Lindau

Binaural Resynthesis of Acoustical Environments

Binaural Resynthesis of Acoustical Environments.
Technology and Perceptual Evaluation.

Alexander Lindau

Alexander Lindau, Berlin, June 2014

For every complex problem, there is a simple solution that is wrong.

H. L. Mencken

Acknowledgements

I would like to thank …

… for thoughtful supervision, near unlimited support and sudden inspirations: Prof. Dr. Stefan Weinzierl, Dr. Hans-Joachim Maempel, Prof. Dr. rer. nat. Michael Vorländer, Christoph Moldrzyk, Prof. Dr.-Ing. Sebastian Möller, Prof. Dr.-Ing. Sascha Spors, and Prof. Dr.-Ing. Alexander Raake.

… for making it easier: my colleagues Fabian Brinkmann, Vera Erbes, Torben Hohn, Michael Horn, Dr. Steffen Lepa, and Frank Schultz.

… for self-forgetting duty: student assistants Alexander 'break a neck' Fuß, Athanasios Lykartsis, and Reinhild Roden.

… for committed work: former master students an co-authors Jorgos Estrella Cazuriaga, Martin Klemmer, Linda Kosanke, Sebastian Roos, and Zora Schärer, and Martina Vrhovnik.

… for priceless constructions: Frank Kohlrusch.

… for detailed copy-editing: Matthias Frank.

… all partners from the DFG research unit SEACEN (Simulation and Evaluation of Acoustical Environments) consortium which were not already mentioned.

… for suffering silently: my family.

… for occasional encouragement (not in order of appearance): Jens Ahrens, Benjamin Bernschütz, Clemens Kuhn-Rahloff, Karim Helwani, Stefan Klockgether, Johannes Nowak, and Franz Zotter.

Last but not least, I want to express my gratitude to the technical and administrational staff of the Audio Communication Group and the TU Berlin.

Abstract

This dissertation contains a collection of publications devoted to the technical and perceptual improvement of the binaural resynthesis of acoustical environments. Binaural resynthesis is based on the assumption that the auditory impression of being exposed to some arbitrary sound field may deliberately be re-evoked by a reconstruction of the sound pressure at the ear drums. Today, binaural signal reproduction is typically approached by convolving measured or modelled binaural room impulse responses with anechoic audio content using real-time algorithms for time-variant fast convolution which allow accounting for head movements of listeners in a natural fashion (dynamic binaural synthesis, DBS).

Hence, while being of limited technical complexity, binaural reproduction technology has the inherent potential for delivering true-to-live substitutes of arbitrary sound fields. However, before establishing data-based dynamic binaural synthesis as a research tool applicable for a convenient and trustworthy resynthesis of arbitrary acoustic environments still a number of theoretical, instrumental and methodological issues remain to be solved. Achieving progress with respect to this overall goal was the main purpose of this dissertation.

The works conducted in view of this aim are presented in a three-part-structure: As a starting point, Part I of this dissertation is devoted to the *recording of binaural signals* introducing, e.g., an improved binaural measurement device and studies assessing the spatial resolution required for natural head movements or for the convincing representation of spatially distributed sound sources. Part II addresses improvements of *binaural signal reproduction*, e.g., a reduction of spectral coloration through newly developed transaural binaural headphones and means towards their proper equalization. Further, an approach to *post hoc* individualization of the interaural time delay is shown to considerably improve localization performance, crossfade behavior and the system's response latency, while a simplified representation of late reverberation helps reducing the computational demands for dynamic rendering. Additionally, a perceptually optimal solution for integrating ambient sounds in dynamic binaural simulations is discussed. Finally, Part III introduces novel approaches for both integrative and differentiated *perceptual evaluation of Virtual Acoustic Environments* partly demonstrating their usefulness in concurrent evaluations of the improved binaural simulation.

Zusammenfassung

Diese Dissertation umfasst eine Sammlung von Veröffentlichungen, die der technischen und perzeptiven Optimierung der binauralen Resynthese akustischer Umgebungen gewidmet sind. Binaurale Resynthese beruht auf der Grundannahme, dass es möglich ist, den auditiven Eindruck einem beliebigen Schallfeld ausgesetzt zu sein, künstlich zu evozieren, wenn es gelingt, den entsprechenden Schalldruckverlauf an den Trommelfellen technisch zu rekonstruieren. Zu diesem Zweck werden nachhallfrei aufgenommene Audioinhalte mit gemessenen oder simulierten sog. binauralen Raumimpulsantworten mittels eines echtzeitfähigen schnellen Faltungsalgorithmus‘ gefiltert, wobei diese Impulsantworten auch entsprechend den beobachteten Kopfbewegungen eines Hörers nachgeführt werden können (sog. *dynamische* Binauralsynthese).

Bei begrenzter technischer Komplexität ermöglicht die binaurale Wiedergabetechnik damit eine potentiell realitätsgetreue (Re-)Synthese beliebiger Schallfelder. Bevor es jedoch gelingen kann, die dynamische Binauralsynthese als komfortables und verlässliches Forschungswerkzeug zu etablieren, sind vielfältige theoretische, instrumentelle und methodische Probleme zu lösen. Fortschritte in dieser Richtung zu erzielen, war das Hauptziel dieses Dissertationsvorhabens.

Die Vorstellung der durchgeführten Studien erfolgt in drei thematischen Blöcken: Teil I der Dissertation behandelt Themen der *binauralen Aufnahmetechnik*, wie z.B. die Entwicklung des binauralen Messroboters FABIAN, oder die zur Darstellung natürlicher Kopfbewegungen oder für eine überzeugende Simulation ausgedehnter Schallquellen eben notwendige räumliche Auflösung. Teil II präsentiert verschiedene Ansätze zur Verbesserungen der *binauralen Wiedergabetechnik*, wie z.B. die Reduktion von Klangverfärbungen mittels neuentwickelter transauraler binauraler Kopfhörer und perzeptiv optimierter Verfahren für deren Frequenzgangskompensation. Zudem wurden Lokalisationsleistung, Überblendverhalten und Systemlatenz mittels eines Algorithmus‘ zur *post hoc* Individualisierung der interauralen Laufzeitdifferenz wesentlich verbessert, der Rechenaufwand der dynamischen Synthese durch eine vereinfachte Nachhalldarstellung reduziert und ein Verfahren zur Darstellung dynamischer binauraler Klangatmosphären optimiert. In Teil III werden verschiedene neuentwickelte Ansätze zur einerseits integrativen anderseits differenzierten *perzeptiven Evaluation virtueller akustischer Umgebungen* präsentiert und teilweise bereits auf die verbesserte Simulationstechnik angewendet.

List of included publications

The work at hand is a cumulative dissertation. It contains the following publications written partly or in total by the author:

Part I – Binaural Recording

[1] Lindau, Alexander; Weinzierl, Stefan (2006): “FABIAN - An Instrument for Software-based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom”, in: *Proc. of the 24th Tonmeistertagung.* Leipzig, pp. 621-625

[2] Lindau, Alexander; Klemmer, Martin; Weinzierl, Stefan (2008): “Zur binauralen Simulation verteilter Schallquellen (On the Binaural Simulation of Distributed Sound Sources)”, in: *Proc. of the 34th DAGA (in: Fortschritte der Akustik)*. Dresden, pp. 897-898

[3] Lindau, Alexander; Weinzierl, Stefan (2009): “On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical, and Lateral Direction”, in: *Proc. of the EAA Symposium on Auralization*. Espoo

Part II – Binaural Reproduction

[4] Lindau, Alexander; Hohn, Torben; Weinzierl, Stefan (2007): “Binaural Resynthesis for Comparative Studies of Acoustical Environments”, in: *Proc. of the 122nd AES Convention*. Vienna, preprint no. 7032

[5] Lindau, Alexander; Brinkmann, Fabian (2012): “Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-individual Recordings”, in: *J. Audio Eng. Soc.*, **60**(1/2), pp. 54-62

[6] Erbes, Vera; Schultz, Frank; Lindau, Alexander; Weinzierl, Stefan (2012): „An extraaural headphone system for optimized binaural reproduction“, in: *Proc. of the 38th DAGA (in: Fortschritte der Akustik)*. Darmstadt, pp. 313-314

[7] Lindau, Alexander; Estrella, Jorgos; Weinzierl, Stefan (2010): “Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD”, in: *Proc. of the 128th AES Convention*. London, preprint no. 8088

[8] Lindau, Alexander (2009): “The Perception of System Latency in Dynamic Binaural Synthesis”, in: *Proc. of the 35th NAG/DAGA, International*

Conference on Acoustics (in: Fortschritte der Akustik). Rotterdam, pp. 1063-1066

[9] Lindau, Alexander; Kosanke, Linda; Weinzierl, Stefan (2012): "Perceptual Evaluation of Model- and Signal-based Predictors of the Mixing Time in Binaural Room Impulse Responses", in: *J. Audio Eng. Soc.*, **60**(11), pp. 887-898

[10] Lindau, Alexander; Roos, Sebastian (2010): "Perceptual Evaluation of Discretization and Interpolation for Motion-Tracked Binaural (MTB-) Recordings", in: *Proc. of the 26th Tonmeistertagung*. Leipzig, pp. 680-701

Part III – Perceptual Evaluation of Virtual Acoustic Environments

[11] Lindau, Alexander; Weinzierl, Stefan (2012): "Assessing the Plausibility of Virtual Acoustic Environments", in: *Acta Acustica united with Acustica*, **98**(5), pp. 804-810, DOI: http://dx.doi.org/10.3813/AAA. 918562

[12] Brinkmann, Fabian; Lindau, Alexander; Vrhovnik, Martina; Weinzierl, Stefan (2014): "Assessing the Authenticity of Individual Dynamic Binaural Synthesis", in: *Proc. of EAA Joint Auralization and Ambisonics Symposium*, Berlin, p. 62-68, http://dx.doi.org/10.14279/depositonce-11

[13] Lindau, Alexander; Erbes, Vera; Lepa, Steffen; Maempel, Hans-Joachim; Brinkmann, Fabian; Weinzierl, Stefan (2014): "A Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI)", in: *Proc. of the EAA Joint Auralization and Ambisonics Symposium*, Berlin[1]

[1] Shortly before the submission of this thesis, the author was informed that this contribution was proposed to be included in a special issue of *Acta Acustica united with Acustica*. Hence, when submitting the thesis it was still unclear, whether the version of this article as published here would indeed be available from the the EAA symposium's proceedings or – somewhat later – from the *Acta Acustica*. Readers are kindly requested to consider this fact while reading.

Content

Introduction 1

1 Introduction 3

1.1 Motivation 4

1.2 Basic Mechanisms of Binaural Hearing 5

1.3 Short Genealogy of Binaural Technology 7

1.4 Dynamic Binaural Synthesis – State of the Art 10

1.5 Perceptual Evaluation of VAEs – State of the Art 12

1.6 Main Objectives 14

1.7 Methods 15

1.8 Summary of Original Achievements 24

1.9 Perspectives 26

1.10 References 27

1.11 Used Abbreviations 33

Part I Binaural Recording 35

2 FABIAN - An Instrument for Software-based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom 37

2.1 Abstract 37

2.2 Concept 37

2.3 System Design 38

2.4 Measurement Process 40

2.5 Outlook 41

2.6 References 41

3 On the Binaural Simulation of Distributed Sound Sources 43

3.1 Research Question 43

3.2 State of the Art 43

3.3 Methods 44

3.4 Results 46

3.5 Discussion 46

3.6 References 47

4 On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical, and Lateral Direction 49

4.1 Abstract 49

4.2 Introduction 49

4.3 Method 51

4.4 Results ... 56

4.5 Discussion ... 59

4.6 Conclusion ... 61

4.7 Acknowledgements ... 62

4.8 References ... 62

Part II Binaural Reproduction ... 67

5 Binaural Resynthesis for Comparative Studies of Acoustical Environments ... 69

5.1 Abstract ... 69

5.2 Introduction ... 69

5.3 Methods and Tools ... 70

5.4 Evaluation ... 79

5.5 User Interface ... 83

5.6 Applications and Outlook ... 84

5.7 References ... 85

6 Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-individual Recordings ... 89

6.1 Abstract ... 89

6.2 Introduction ... 89

6.3 Methods ... 92

6.4 Listening Test I ... 98

6.5 Results of Listening Test I ... 99

6.6 Discussion of Results of Listening Test I ... 101

6.7 Improving Regularization ... 102

6.8 Listening Test II ... 103

6.9 Results of Listening Test II ... 103

6.10 Conclusion ... 104

6.11 Acknowledgements ... 105

6.12 References ... 105

7 An Extraaural Headphone System for Optimized Binaural Reproduction ... 109

7.1 Introduction ... 109

7.2 The Extraaural Headphone System ... 110

7.3 HpTF Compensation ... 111

7.4 Discussion and Conclusion ... 113

7.5 References ... 114

8 Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD ... 115

8.1 Abstract 115

8.2 Introduction 115

8.3 ITD Individualization 117

8.4 Listening Test 128

8.5 Results 128

8.6 Conclusions 130

8.7 Acknowledgements 130

8.8 References 131

9 The Perception of System Latency in Dynamic Binaural Synthesis 135

9.1 Motivation 135

9.2 Latency in VAEs 135

9.3 Methods 137

9.4 Results 141

9.5 Discussion 143

9.6 Acknowledgements 144

9.7 References 144

10 Perceptual Evaluation of Model- and Signal-based Predictors of the Mixing Time in Binaural Room Impulse Responses 147

10.1 Abstract 147

10.2 Introduction 148

10.3 Methods 151

10.4 Listening Test Results 164

10.5 Regression Analysis 165

10.6 Conclusions 172

10.7 Acknowledgements 173

10.8 References 173

11 Perceptual Evaluation of Discretization and Interpolation for Motion-Tracked Binaural (MTB-) Recordings 177

11.1 Abstract 177

11.2 Introduction 178

11.3 Methods 185

11.4 Listening Test 188

11.5 Results 192

11.6 Summary and Conclusion 196

11.7 Acknowledgements 197

11.8 References 198

Part III Perceptual Evaluation of Virtual Acoustic Environments 201
12 Assessing the Plausibility of Virtual Acoustic Environments 203
12.1 Abstract 203
12.2 Introduction 204
12.3 Plausibility: An Experimental Approach 205
12.4 Listening Test Setup 212
12.5 Results 215
12.6 Discussion 216
12.7 Acknowledgements 219
12.8 References 219
13 Assessing the Authenticity of Individual Dynamic Binaural Synthesis 223
13.1 Abstract 223
13.2 Introduction 224
13.3 Method 226
13.4 Results 235
13.5 Disscussion and Outlook 236
13.6 Summary 238
13.7 Acknowledgments 238
13.8 References 238
14 A Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI) 241
14.1 Abstract 241
14.2 Introduction 242
14.3 Methods 243
14.4 Results 248
14.5 Discussion 250
14.6 Outlook 251
14.7 Acknowledgements 251
14.8 References 252
14.9 Appendix – SAQI-EN 255

Introduction

1 Introduction

This manuscript contains a collection of the author's most relevant publications devoted to the technical and perceptual improvement of the binaural[2] reproduction of acoustical environments. Studies were conducted during the years 2006–2014 while the author was engaged in several research projects at the TU Berlin's Audio Communication Group. More specifically, these were the projects "Basics of Sound Reproduction and Perception" while holding a PhD-Advisor-Mentor-Program scholarship from Deutsche Telekom AG, "The Concert Hall and its Mediatization. Comparative Empirical Study of Binaurally Synthesized Natural and Electro-acoustically Transmitted Musical Performances" granted by the German Research Foundation (DFG WE 4057/1-1), and „Evaluation of Virtual Acoustic Environments" also granted by the German Research Foundation (DFG) within the DFG research unit "Simulation and Evaluation of Acoustical Environments (SEACEN)" (DFG WE 4057/3-1).

The chapter at hand is intended to give an overview of the works included in this dissertation while interpreting, evaluating and discussing the obtained results in the light of current research. Section 1.1 reviews the general principle of binaural reproduction and motivates the thesis by explaining the strong demand in the acoustic research community for the continued development of the binaural method. Section 1.2 shortly introduces the most relevant mechanisms of binaural hearing. Section 1.3 presents a short genealogy of binaural technology while section 1.4 concludes the historical review by giving a description of the state of the art of binaural synthesis. Section 1.5 gives an overview of quality criteria and existing approaches to the evaluation of Virtual Acoustic Environments (VAEs) and discusses their appropriateness. Section 1.6 identifies relevant problem areas and formulates the main objectives pursued in this dissertation. Section 1.7 presents methods and outcomes of the included studies. Section 1.8 summarizes the original achievements of this dissertation, and section 1.9 finally discusses future research perspectives.

[2] from Latin "with both ears"

1.1 Motivation

Binaural signal reproduction is based on the assumption that the auditory impression of being exposed to an arbitrary sound field may be re-evoked by reconstructing the sound pressure observed at the ear drums (Møller, 1992). This widely cited statement can – without exaggeration – be denominated the 'lemma' of binaural reproduction technique. It may be put into practice straight forward by recording the sound pressure with microphones positioned at the ear drums of a real or artificial human head, and then playing back the recordings to a listener using headphones.

Hence, while being of limited technical complexity – as, e.g., compared to current loudspeaker-array-based approaches to sound field synthesis such as vector-based amplitude panning (VBAP), ambisonics panning, wave field synthesis (WFS) or higher order ambisonics (HOA) (see Spors et al. [2013] for a recent review) – binaural reproduction technology has the inherent potential for delivering true-to-live substitutes of arbitrary sound fields. Numerous applications can be thought of where such an acoustic simulation technique allowing for instantaneous, reproducible and transportable, instrumental or perceptual assessments would be highly desired. For example, it enables the convenient comparative assessments of spatially separated acoustic environments (as, e.g., concert halls, Lehmann and Wilkens, 1980; Maempel and Lindau, 2013) or of different types and setups of acoustic media devices (see, e.g., Lindau and Lepa, 2014) by singular or multiple, co-located or distributed listeners and/or musicians (see, e.g., Schärer Kalkandjiev and Weinzierl, 2013). From the viewpoint of an audio engineer the usage as an 'out-of-the-box laboratory reference of the acoustic reality' might appear most appealing. Further application examples were reviewed by Møller (1992) or Kleiner et al. (1993).

This overview should have helped explaining why binaural technology is regarded such a promising and universal acoustic research tool. However, as the following sections will show, there are numerous smaller and larger pitfalls associated with the realization of a perceptually accurate binaural simulation, making it a challenge, still. However, before presenting the main objectives targeted, the methods applied and the results obtained in this work, the basic physical, physiological and psychological foundations and the historical evolvement of binaural technology will be shortly reviewed.

1.2 Basic Mechanisms of Binaural Hearing

Research on binaural effects in hearing can be dated back as far the end of the 18th century (see Wade and Deutsch, 2008, for a review). Wade and Deutsch portrayed – amongst others – Giovanni Battista Venturi (1746-1822), who suggested localization to be due to inequalities of the sound at the two ears, William Charles Wells (1757-1817), who theoretically devised the impressions of binaurally divergently presented melodic patterns, or Somerville Scott Alison (1813-1877), the inventor of the stethophone, a device which allowed feeding different signal to the two ears, as pioneers in research of binaural hearing.

However, it was Lord Rayleigh who earned the reputation to have given 'birth' to the first consistent theory of binaural perception. In 1907 (Strutt, 1907) he summed up his findings in what is called the Duplex theory today. Accordingly, the ability to localize sounds originates from two morphologically induced characteristics of the ear signals: On the one hand, depending on direction (and frequency content) of a sound source, the listener's head acts as an obstacle to sound propagation inducing interaural intensity (or level) differences (IID, ILD) by acoustic shadowing. On the other hand, the displacement of the two ears with respect to the direction of sound incidence induces so-called interaural time differences (ITD) between both ears' signals (see Figure 1-1).

Stern et al. (2006) shortly reviewed research on binaural localization cues. They explain that ITD and ILD are exploited somehow complementary for human auditory localization: For the diameter of a typical human head and due to the periodic

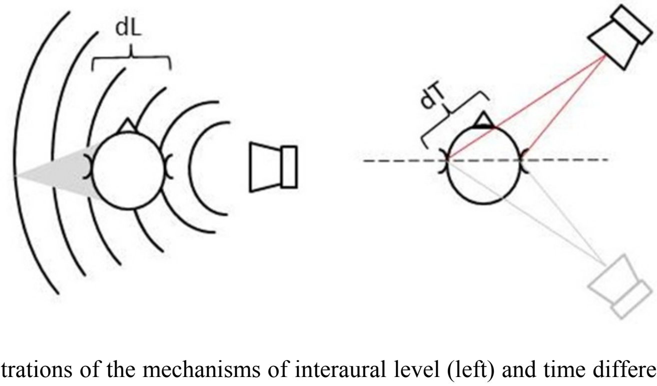


Figure 1-1. Illustrations of the mechanisms of interaural level (left) and time difference (right). Additionally, the figure on the right shows a second source (grey) which was positioned on a 'cone of confusion' by mirroring the first source at the subject's frontal plane. Hence, this second source provokes similar interaural time and level differences as the first one.

nature of pure tones, time differences are becoming increasingly ambiguous for frequencies above approximately 1500 Hz. At the same time, ILDs, which – due to diffraction – are comparatively small at lower frequencies, increase, as head-shadowing becomes more effective for decreasing wave lengths. However, there exits also evidence for a 'transitional mechanism' as it was shown that envelope delays of high frequency carriers that were amplitude-modulated by a lower frequency could be exploited for lateralization (i.e., localization between the ears, Henning, 1974).

Furthermore, in considering the origins of ITD and ILD Lord Rayleigh could explain why front-back discrimination for sources emitting pure tones is sometimes difficult. Hence, ITD and ILD are not exclusively linked to a certain direction of sound incidence; instead, there exist an infinite number of sound source positions resulting in identical ITD and ILD cues. As the entirety of these positions forms conical surfaces – originating at the ears and being rotationally symmetric with respect to the interaural axis – they are called 'cones of confusion'. However, Rayleigh found that the cones of confusion lose their ambiguity when moving one's head. Thus, willingly induced changes of interaural information help clarifying whether a sound source lies in front or behind the listener.

In the special case of sound waves arriving from the median plane, nearly no ITDs and ILDs are induced at all. Instead, specific direction dependent changes of the spectral content of (known) sound sources are interpreted as height information. These timbral changes result from the individual morphology (i.e. the torso, the head and especially the fine structured outer ears or *pinnae*) interacting with the sound field and are typically denoted as (monaural) spectral cues (SC).

Another auditive cue often mentioned in relation with localization or spatial hearing is the interaural correlation (IC), which is often quantified by calculating the correlation coefficient of the sound pressure at the two ears (interaural cross correlation coefficient, IACC). It is related to the perception of the extent of sound sources and to the degree of perceived envelopment in sound fields.

Stern et al. (2006, p. 152) also gave an overview on just noticeable differences (JNDs) for ILD, ITD and IC. Hence, the JND for the ILD (and also for monaural SCs) is in the order of 1 dB, whereas the JND for the ITD (for low frequency pure tones) is about 10 µs. The JNDs for the IACC depend on the starting condition, thus, whereas a decrease from 1 to 0.96 is well discernable (for broadband noise), for a reference condition of zero interaural correlation JNDs of the order of 0.35 have been reported (Kim et al., 2008).

ITD, ILD and IC are good examples of a predominantly bottom-up driven model of perception: Specific physical properties of acoustic stimuli may be related straight forward to certain dimensions of auditory perception. However, as indicated already with the experience-based evaluation of spectral cues (SC), top-down effects or higher cognitive processes such as experience, memory, and reflection may play a relevant role in forming our final percepts, too. The perception of distance is a particularly demonstrative example of top-down processing. Propagation over increasing distances changes a sound field in different ways: the sound pressure level decreases, higher frequencies become increasingly damped, the direct-to-reverberance ratio decreases (in echoic environments). The resulting perception of distance is formed by an interaction of the perceived sound field characteristics, the listener's familiarity with the type of source signal and his/her former experience of distance effects. Occasionally, the result of this interaction may be observed to fail in daily life, when, for example – for a short moment – unexpected noises are perceived to be located in totally different than their actual distances.

1.3 Short Genealogy of Binaural Technology

The advent of binaural reproduction technology may be dated to the first half of the 20th century when researchers as Harvey Fletcher (Bell Laboratories), Alvar Wilska (University of Helsinki), or Kornelis de Boer and Roelof Vermeulen (Philips Research Laboratory) started experimenting with recording devices that imitated the

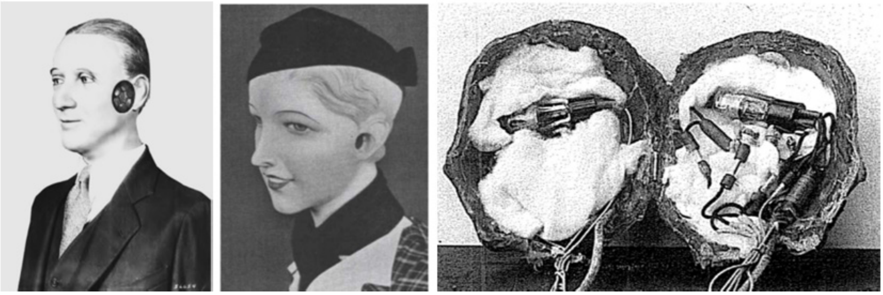


Figure 1-2. Left: "Oscar", artificial head and torso applied for binaural transmissions of musical concerts by Bell Laboratories (Hammer and Snow, 1932). Middle: Display mannequin with microphones placed in rudimentary outer ears used by de Boer and Vermeulen (1939). Right: Inner view of an early artificial head molded from a human corpse's head by Alvar Wilska (1938) in the course of his doctoral thesis.

outer appearance of human heads (and torsi) and which were equipped with microphones at the position of the outer ears (see Paul, 2009, for a review). However, due to the state of development of the recording technology in general, as, e.g., in terms of microphone sensitivity, signal-to-noise ratio, frequency bandwidth, and dynamic behavior, the transmission fidelity was still limited. Furthermore, the way of attaching the microphones to the artificial heads (Figure 1-2, left) or the limited morphological accuracy of outer ears (Figure 1-2, middle) will have caused distorted spectral cues (SC).

The next step in development was taken by American and, particularly, German research groups (the latter being situated in Göttingen, Berlin, and Aachen) in the late 1960s and early 1970s (cf. Figure 1-3). For the first time, ambitious quality criteria for the accuracy of binaural reproduction were formulated: Researchers targeted at a “correct reproduction of acoustical information at any place in a room” while aiming at the elimination of specific shortcoming such as front-back confusion, in-head localization, poor distance perception or an overestimation of reverberation (Kürer, Plenge and Wilkens, 1969). Genuit (1981) explicitly targeted an artificial head recording system to be useable for acoustic measurement purposes and thus overcoming the limitations technical reproduction fidelity of its antecessors. Additionally, the human morphology was reproduced more exactly, as, e.g., by using gypsum or rubber molds from real heads, and by consultation of anthropometric databases (cf. Figure 1-3).

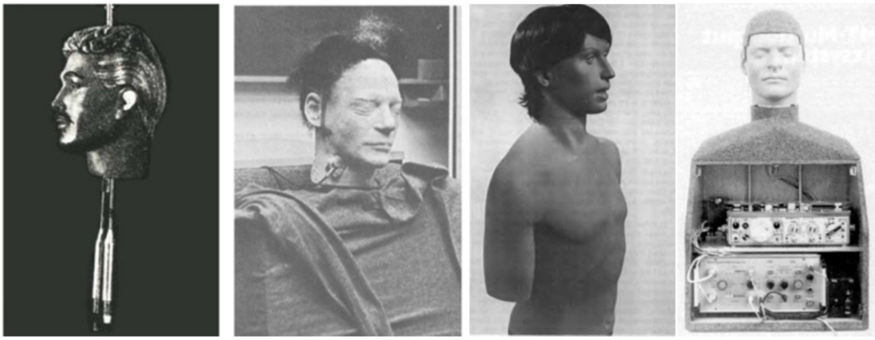


Figure 1-3. From left to right: (1) Mould of a display mannequin’s head with partly mechanically reworked outer ears (Damaske and Wagener, 1969). (2) Artificial head as a combination of gypsum and rubber molds of a real individual (Kürer, Plenge und Wilkens, 1969). (3) KEMAR (Knowles Electronic Manikin for Acoustic Research) with a wig (Burkhardt and Sachs, 1975). (4) Prototype of Genuit’s artificial head recording system intended for acoustic measurement purposes (Genuit, 1981).

As an application example for binaural recordings, German researchers used their improved apparatus for resynthesizing the acoustical characteristics of different concert halls (Lehmann and Wilkens, 1980) and let test subjects assess them in subsequent listening tests. Hence, for the first time, such research questions could be treated independently from limitations of time and place and by using instant auditive comparisons.

More recently, remaining perceptual shortcomings were found to be largely due to the binaural reproduction not responding in a natural way to the head movements of the listeners (Wenzel, 1996; Mackensen, 2004). Hence, the realization of interactive binaural simulations in the 1990s marks the beginning of modern binaural technology. Binaural signals are now created through filtering an anechoic acoustic signal with filters describing the acoustical transfer paths from sound sources in a certain reproduction scenario – such as, e.g., free field listening, a recording studio control room, or a concert hall – to the ears. Further, if head movements of the listener are detected, the simulation updates the acoustical characteristics correspondingly. This "process of rendering audible […] the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the […] space" was called *auralization* by Kleiner et al. (1993). Sandvad (1996) was probably one of the first who (synonymously) introduced the term *dynamic binaural synthesis* (DBS), while the term 'dynamic' referred to a simulation reacting to head movements.

The realization of dynamic auralization or dynamic binaural synthesis was eventually afforded by a multitude of recent technical advancements; some of the most relevant were, e.g.: (a) the availability of powerful digital signal processing units, (b) the development of optimized low-latency algorithms for fast convolution with time variable filters (Gardner, 1995; Müller-Tomfelde, 2001), (c) the availability of convenient means for head tracking (see Welch and Foxlin, 2002 for a review), (d) the advancement of digital (FFT-based) acoustic measurement techniques (Müller and Massarani, 2001), and, finally, (e) the development of modern head-and-torso simulators (HATS) whose heads may be re-oriented above the torso (Moldrzyk, 2002; Mackensen, 2004; Lindau and Weinzierl, 2006; Hess and Weishäupl, 2014; cf. Figure 1-4).

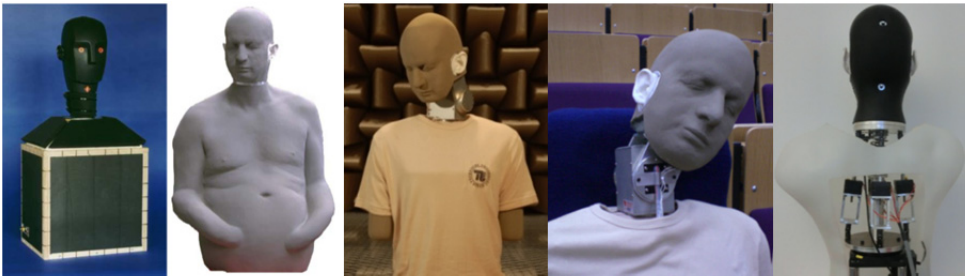


Figure 1-4. Artificial heads or HATS devices that may move their head in one (photo no. 1: rotatable artificial head [Neumann Ku100] used by Mackensen [2004]; photo no. 2: HATS built by Moldrzyk [2002]) or more degrees of freedom (photos no. 3 and 4, HATS FABIAN, courtesy of the author, photo no. 5, rear view of HATS by Hess and Weishäupl, [2014]).

Early implementations of so-called interactive Virtual Auditory Displays (VADs) or Virtual Acoustic Environments (VAEs) which employed dynamic binaural synthesis have been described by Wenzel et al. (1990), Reilly and McGrath (1995) or Karamustafaoglu et al. (1999).

1.4 Dynamic Binaural Synthesis – State of the Art

Note: *In order to avoid redundancy in writing and as a matter of the chosen presentation order in this chapter, the documentation of the state of the art disregards thematically related works of the author which are included in this dissertation. Instead, these will be introduced in a bundled form in the 'Methods' section 1.7.*

The sound transmission scenarios which are of interest in the scope of this dissertation may be described with the help of linear system theory. Accordingly, the acoustic transmission paths from sound sources to receiver positions in free field or in a reverberant environment may be characterized by a finite and real valued impulse response, which can be approximated in the digital domain by an FIR filter (Oppenheim et al., 2004). Such impulse responses may either be measured in real rooms (*data-based* DBS, Karamustafaoglu et al., 1999) or modeled numerically with the help of a CAD model of the acoustic environment (*model-based* DBS, Vorländer, 2007). The main focus of this dissertation is on the resynthesis of acoustical environments by dynamic auralization of binaural data that have been *measured* in real rooms. However, many of the presented findings may be generalized to the auralization of modeled room acoustics, too.

1.4.1 Recording Binaural Data

For the *in situ* measurement of binaural transmission paths, loudspeakers are commonly used as sound sources, while real human subjects or HATS with a changeable orientation of the head above the torso (Figure 1-4) serve as receivers. Furthermore, it is important to distinguish between binaural recordings that were made with a subject's own ears (*individual* binaural synthesis) or that of another subject or artificial head (*non-individual* binaural synthesis), as in case of non-individual recordings, binaural cues will be distorted to an amount that will typically be audible when compared to listening with one's own ears (Møller et al, 1996). In either case, binaural transfer paths are measured discretely for each source-receiver combination and for each relevant head orientation using an inaudibly fine angular resolution. Results may be stored as data sets of so-called Binaural Room Impulse Responses (BRIRs) or Binaural Room Transfer Functions (BRTFs, their frequency domain equivalent), respectively.

When following the lemma of binaural signal reproduction introduced before, intuitively, one would attempt recording binaural signals by using microphones at the position of the ear drums (Burkhardt and Sachs, 1975). However, more recent studies (Middlebrooks et al., 1989; Hammershøi and Møller 1996; Algazi et al., 1999) empirically confirmed the sound propagation to the ear drums to be independent of direction already as far as 6 mm outside the ear canal entrances. Furthermore, Hammershøi and Møller showed the difference between binaural signals recorded at the open and at the blocked ear canal to be independent of direction, too. Hence, for practical reasons – and independently from using real or artificial human heads – nowadays, mostly, signals dedicated for binaural sound field reproduction are recorded at the blocked ear canal.

1.4.2 Dynamic Rendering of Binaural Data

As mentioned earlier, in dynamic binaural synthesis the BRIRs used for the time-variant convolution process are continuously and – at best – inaudibly exchanged according to the current head orientation of the listener. To this end, head movements have to be tracked in one or more degrees of freedom using available sensors. As a result, listeners will be able to rotate their heads in a natural fashion while sound sources are perceived as remaining at their expected location (e.g.: in front), and do not follow the head movements as is typically the case with headphone reproduction of conventional stereophonic audio material. It is basically this 'trick' dynamic auralization owes its potentially astounding degree of realism, all the more as, consequently – at least for reverberant simulated environments – a realistic perception of distance, i.e. a perception of sound sources outside one's head (externalization) arises.

Additionally, and because fast convolution can be implemented as a real-time process, the audio content to be convolved may be changed instantly. Furthermore, as the algorithm also allows switching between different filters, the room information (BRIRs) may be rapidly interchanged, too. Hence, dynamic binaural synthesis allows for a convenient 'switching' between arbitrary audio stimuli played back in arbitrary acoustic environments.

1.4.3 Binaural Signal Reproduction

Binaural signals are typically conveyed to the listener via headphones although approaches for transaural reproduction have been applied, too (Gardner, 1997; Lentz, 2006). In either case the acoustic transfer path has to be equalized. In order to design compensation filters for headphones, typically the so-called Headphone Transfer Function (HpTF) of human listeners or HATS devices is measured and then appropriately inverted. Furthermore, for achieving best acoustic compatibility it was devised that recordings made at the blocked ear canal entries will have to be played back while the acoustic radiation impedance as seen from the ear canal entries of a subject approaches that of free air (Møller, 1992; Møller et al., 1995). Headphones approaching this criterion were denoted 'FEC' headphones (**F**ree air **E**quivalent **C**oupling to the ear) by Møller et al. (1995)

1.5 Perceptual Evaluation of VAEs – State of the Art

Note: *In order to avoid redundancy in writing and as a matter of the chosen presentation order in this chapter, the documentation of the state of the art disregards thematically related works of the author which are included in this dissertation. Instead, these will be introduced in a bundled form in the 'Methods' section 1.7.*

VAEs may be evaluated either *physically* (e.g., by measuring system characteristics such as transfer functions or response latencies, or by examining the proper reproduction of binaural and monaural cues, cf. section 1.2) or *perceptually* (e.g., by assessing fundamental auditory qualities such as sense of direction, of distance or perceived spectral coloration, or more general auditory impressions such as preference, presence, or perceived naturalism, for a short review see also Nicol, 2010, pp. 58).

Physical evaluation is mostly applied in cases where a reference sound field is available which allows both computing difference measures (mostly for spectral deviation, Nicol, 2010, pp. 58) and judging them, e.g. in terms of known just noticeable differences (JNDs). Hence, also in case of a 'purely physical' evaluation meaningful interpretations of the observed deviations will require further knowledge about their psychoacoustic (and maybe even psychological) implications.

Perceptual evaluation has historically been carried out with an emphasis on assessments of fundamental auditive qualities, for instance when assesseing localization performance or the observed number of front-back-confusions (Møller, 1996; Møller 1996b; Lokki and Järveläinen, 2001; Minnaar et al., 2001). However, the selection of these criteria appears to be – at least partly – driven by their methodological accessibility, thereby leaving it unclear in how far they are suitable as perceptual measures of VAEs as a whole. For visual virtual environments more integrative criteria as 'immersion' or 'a sense of presence' have been proposed (for a review, see Lombard and Ditton, 1997). These concepts are generally interpreted as multidimensional constructs related to the overall 'illusion of non-mediation' (ibid.) while including secondary aspects such as the experience of spatial presence ('being there'), a task-related sense of 'involvement' and a judgment of 'realness' (Schubert et al., 2001). However, due to their multidimensional nature, it remains unclear which part of the simulation is actually addressed by the user's evaluation.

Concurrently, 'authenticity' (indistinguishability from an externally provided reference), and 'plausibility' (indistinguishability from an imagined or 'inner' reference) have been proposed as promising integrative quality measures for VAEs (Blauert, 1997; Pellegrini, 2001; Kuhn-Rahloff, 2011). However, empirical test designs rigorously operationalizing the concepts of authenticity or plausibility have not been applied for the evaluation of VAEs, or were not yet available, respectively.

Additionally, recently, more detailed catalogues of auditory qualities have been proposed for the perceptual evaluation of spatial audio reproduction techniques or VAEs. Yet, those vocabularies were either not entirely targeting VAEs (e.g., Lorho, 2005; Berg and Rumsey, 2006), were created without applying a substantiated methodological approach (i.e., they were produced *ad hoc* by the authors, e.g., Pellegrini, 2001) or their focus was limited to specific types of VAEs only (e.g., Silzle, 2007).

1.6 Main Objectives

Based on the described state of the art of binaural reproduction technology, by identification of specific problems in the course of own initial perceptual evaluations of a state of art implementation for data-based DBS (Lindau et al. 2007), and driven by the practical need to further increase the applicability of DBS (cf. section 1.1) several promising research areas were identified and examined in the course of this dissertation. They will be shortly discussed in the following.

As typical for data-based approaches, binaural technology suffers from being tedious and time-consuming. Furthermore, reducing the required amount of data will be helpful when simulating increasingly complex acoustic scenes or when being limited to computationally restricted platforms (e.g., in mobile applications). However, any targeted simplification leading to a *reduction of the measurement effort* will require thorough perceptual evaluation. Similarly, any – potentially related – *reduction of computational effort* would be helpful.

Furthermore and maybe most importantly, it was found that *all remaining audible deviations between a binaural simulation and the acoustic reality demand a qualitative identification,* which have to be followed by respective *refinement of methodology and instrumentation.* Difference qualities can be expected to involve, e.g., spectral coloration, localization instability, or response latency.

Incidentally, virtual acoustic scenes are said to convey an 'artificial' and somewhat 'aseptic' overall impression. One reason might be the fact that the described binaural method relies on measurements of transfer paths between individual pairs of sources and receivers, making the binaural simulation of naturalistic acoustic 'background atmospheres' or 'soundscapes' difficult. Therefore, *methods for increasing ecological validity* of binaural simulations are demanded.

Finally, improvements of binaural simulations require *suitable approaches to perceptual evaluation* such as, e.g., for proving inaudibility of undertaken simplifications, for a qualitative and quantitative characterization of achieved improvements or for benchmarking alternative solutions. Furthermore, these approaches should have both a theoretical foundation and allow for practical operationalization.

Hence, while it was outlined before that establishing data-based dynamic binaural synthesis as a research tool applicable for a convenient and trustworthy resynthesis of arbitrary acoustic environments is a strong demand of the acoustic community, still a number of theoretical, instrumental and methodological issues remain to be solved. Achieving progress with respect to this overall goal was the main motivation of this dissertation. The main research objectives pursued in view of this overall goal can be summarized in the form of work assignments:

(1) Reduce the measurement effort related with data-based binaural simulations,

(2) reduce the computational demands related with data-based binaural simulations,

(3) qualify remaining perceptual differences between binaural simulations and the acoustic reality,

(4) find remedies for identified perceptual shortcomings,

(5) increase ecological validity (i.e. support a more natural auditive impression) of data-based binaural simulations, and

(6) develop theoretically substantiated approaches suitable for both integrative and differentiated perceptual evaluations of the achieved improvements.

1.7 Methods

The majority of the presented studies were concerned with *non-individual binaural synthesis* (i.e. simulations using the BRIRs of the HATS FABIAN). *Individual binaural synthesis* – while still constituting a major procedural effort – was assessed only once in this dissertation (cf. section 13, Brinkmann et al., 2014). Furthermore, it is emphasized that throughout all listening tests conducted in the course of this dissertation *dynamic* auralizations (i.e. binaural simulations accounting for head movements) were used, a fact not explicitly indicated in the following anymore.

Following the presentation order in sections 1.4 and 1.5, this dissertation is organized in three larger parts each one focusing on improvements with specific regard to one of the aspects of *binaural recording, binaural reproduction*, or the *perceptual evaluation of binaural simulations*. The pursued rationales are shortly resumed in the following three subsections.

1.7.1 Part I: Binaural Recording

Part I of this dissertation was mainly concerned with aspects of binaural signal acquisition. It adressed problems related to the 1st and 2nd research objective i.e., the reduction of the measurement effort and of the computational demands for rendering.

Constituting a step forward in binaural measurement technology, in the course of his master's thesis the author developed the binaural measurement robot **FABIAN** ("**F**ast and **A**utomatic **B**inaural **I**mpulse response **A**cquisitio**N**", Figure 1-4, photos no. 3 and 4, Lindau and Weinzierl, 2006) which allows for a convenient measurement of large data sets of BRIRs reflecting different degrees of freedom for head movements. FABIAN differs from its historical antecessors by providing timely audio quality, while its ability to rotate the head in all three degrees of freedom above the torso by means of a servo motorized neck joint distinguishes it from its contemporaries. Its technology is presented here again for reasons of completeness.

Measurement effort and computational demands for rendering will increase linearly with the number of individually virtualized sound sources. Hence, one of the earliest studies in this dissertation (Lindau and Klemmer, 2008) was concerned with the possibility to reduce the amount of individually rendered sound sources. It was assessed in how far several horizontally displaced sound sources in a reverberant environment could be simulated with a reduced number of virtual sources. A listening test was conducted using an adaptive three-alternative-forced-choice (3AFC) paradigm for assessing the just noticeable opening angle between frontally positioned sound sources at two typical listening distances in a concert-hall-like environment and for two types of audio stimuli. Results showed that the just noticeable opening angle increased with the listening distance (i.e. with a decreasing direct-to-reverberant ratio) and for natural signals (as compared to noise pulses). However, it was concluded that, in order to assure an undisturbed reproduction for the majority of the listeners, only sources with an angular distance of less than 5° should be combined into a singular virtual sound source (at least for certain, well discernable stimuli and small DR-ratios). This implies that for typical ensembles of natural sound sources (such as, e.g., a string quartet) a reduction of the number of measured/simulated sources appears not recommendable.

The last study presented in Part I of this dissertation addressed the required measurement effort with respect to head orientations. Therefore, the just noticeable discretization of BRIR grid resolution was assessed in three exemplary acoustic environments (anechoic, a ‘dry’ recording room, a ‘wet’ lecture hall) and separately for all three rotational degrees of freedom (Lindau and Weinzierl, 2009). To date, the HATS FABIAN is the only device that allows for automated measurement of the respective BRIRs (cf. Figure 1-4). Listening tests were again conducted following an adaptive 3AFC procedure. In all cases, the base grid resolution was 1° and could be reduced in steps of 1°. Throughout all conditions BRIRs were auralized using linear cross fading between different head orientations. It was found that – at least for certain stimuli and specific directions of sound incidence – the ability to detect a reduced spatial resolution in discretely measured BRIR data is very similar for all directions of head movements. Furthermore, the thresholds showed only little variation with the size and reverberation time of the measured acoustic environments. It was concluded, that a VAE providing a BRIR grid resolution of 2° for horizontal, 1° for vertical and 1° for lateral head movements will be sufficiently accurate even for very sensitive listeners and when using worst-case audio signals.

1.7.2 Part II: Binaural Reproduction

Part II of this dissertation was mainly concerned with aspects of binaural signal reproduction. It addressed problems related to the 2nd, 3rd, 4th, and 5th research objective i.e., the reduction of computational demands, the qualification and equalization of perceptual issues and a more natural appeal of binaural simulations.

As a starting point, in the first study included in Part II, a state-of-the-art framework for dynamic binaural signal acquisition and rendering was introduced and results of an initial perceptual evaluation were discussed (Lindau et al., 2007). The framework was set up by the author while benefiting from the development of a software package for dynamic fast convolution in the course of a related research project. As a first proof of concept, non-individual dynamic binaural simulations (i.e. using BRIRs measured with the FABIAN HATS) were perceptually evaluated in an *in situ* AB detection task comparing simulation and reality. Subjects were asked to decide which one of two stimuli was the real one, and, afterwards, to describe auditive deviations that guided their decisions. While delivering promising first results (small but significant remaining detection rate of 52,9 %), improvements of both the listening test design and the binaural reproduction were found necessary: Thus, the AB listening test paradigm was found to be theoretically invalid as it does not allow to differentiate between a consistent detectability of an auditive difference and the (actually targeted) acceptability of a simulation as real. Additionally, and as demanded by the 3rd research objective, obvious auditive deviations from reality were revealed, e.g., with respect to spectral coloration, stability of localization, naturalism of reverberation, the perception of latency or cross fade artifacts (enumeration according to frequency of mentioning).

Motivated by this first perceptual evaluation, the next four studies presented in Part II of this dissertation dealt with the 4th research objective (finding remedies for perceptual deviations from acoustic reality).

Spectral coloration was identified as a primary issue when comparing acoustic reality and non-individual binaural simulations. This was not surprising, as the magnitude spectra of head related transfer functions (HRTFs, the free field equivalent to a BRTF) were shown to differ between individuals in clearly audible ranges (Møller et al., 1995b). Besides these non-individual binaural transfer functions, there are further reasons for audible spectral coloration as, e.g., the individual headphone-transfer function (HpTF), or the influence of non-perfect transducers in the electroacoustic transmission chain (e.g., microphones, loudspeakers). In this dissertation, especially the appropriate compensation of the HpTF was addressed. To this end, insert microphones were developed to be used at the blocked ear canal of individuals. Subsequently, different approaches to inverse filtering were assessed in a criterion-free comparative listening test allowing for direct comparison to real sound fields (Lindau and Brinkmann, 2012). As a result, a fast frequency approach to high-pass regularized least-mean-squares inversion using a minimum-phase target function was found to be a perceptually optimal inversion algorithm. Surprisingly, it was found that – for non-individual binaural synthesis – using the HpTF of the recording subject (i.e. of the HATS FABIAN) was to be preferred over using individual headphone compensation (i.e. over a filter based on the listener's own HpTF). We explained this observation with the recording head's filter causing a kind of spectral 'de-individualization'. Hence, while the HpTF of the recording head closely resembles a near-field HRTF, its inverse application resulted in a reduction of spectral features in the non-individual binaural simulation which are typical for the morphology of the used recording head. In turn, other listeners will perceive this simulation as more similar to listening with their own ears. On the contrary, individual headphone filters will result in a 'best possible' presentation of the non-individual i.e. the wrong BRIRs at a subject's ears. Physical evidence for this hypothesis was found by comparing the spectral deviations that were present under both reproduction conditions. However, further empirical validation of this hypothesis is still a topic of future research.

Furthermore, a novel headphone system for binaural signal reproduction, the BKsystem, was developed (Erbes et al., 2012). It features a dedicated DSP-controlled power amplifier (BKamp), and headphones with transaural transducers (BK211) allowing both for unobstructed auditive comparisons with real sound fields and a convenient application of in-ear microphones. The system complies with the FEC-criterion, provides high bandwidth and dynamic range at an inaudible self-noise, and allows for an extension of its low frequency response by a subwoofer. Additionally, the BK211 headphones allow for a convenient integration of typical VAE accessories (as, e.g., head-mounted displays (HMDs), 3D glasses, head tracking sensors) while the acoustic transmission is especially insensitive to small variations in mechanical fit.

Localization instability observed in the initial evaluations was identified to be predominantly due to non-individual recordings conveying distorted ITD cues. Hence, if a subject's head size was smaller than that of the used artificial head, the resynthesized ITD was larger as naturally expected. In this case and for fixed head orientations, sound events will be located at larger horizontal offsets. Additionally, when rotating one's head, sound events which should remain at a stable position will be perceived as moving contrary to one's head movements. Described errors will be reversed in direction if the individual's head size is larger than that of the artificial head. As a solution, an approach to *post hoc* individualization of the ITD was proposed by Lindau et al. (2010). Thereby, the non-individual ITD is extracted from BRIR data sets via onset detection and re-introduced during binaural reproduction via scalable time stretching. Listening tests have been conducted in order to generate an empirical model predicting an individual ITD scaling factor from the measurement of a subject's head diameter. The proposed approach led to a considerable perceptual improvement, all the more as cross-fade artifacts, which were reported to be audible during head movements before, were also strongly reduced when using the resulting quasi-minimum-phase representations of BRIRs for dynamic auralization.

Another issue mentioned by Lindau et al. (2007) was a perception of latency when moving one's head. Hence, by manipulating the end-to-end latency in a discrete and controlled manner, the detection threshold for latency was assessed using an adaptive 3AFC listening test (Lindau, 2009). The lowest observed threshold (52 ms) was smaller than all results reported in previous literature. Results further led to the practical conclusion that the acoustic propagation delay (i.e., the delay related to the distance of the source) should be excluded from measured BRIRs before auralization.

Another study was concerned with the 2nd research objective, i.e. the perceptually motivated reduction of the rendering effort: In 2012, Lindau et al. more closely examined the fact that the reverberant field of ergodic rooms becomes randomly intermixed after some time. With respect to the computational effort in binaural synthesis this behavior can be exploited, e.g., by substituting the dynamic simulation of the late reverberation by a static simulation. To this end, the instant in time had to be identified, from which on the late reverberation did not contain any direction-dependent information. This so-called perceptual mixing time was assessed in nine exemplary acoustic environments which were independently varied in relative reverberance and volume (each in 3 steps). Again, listening tests were conducted following an adaptive 3AFC procedure. It was assessed, at which earliest point in time dynamically updated early parts of a BRIR can be concatenated with a static reverberant tail before being distinguishable from an auralization that dynamically updates the complete BRIR. To this end, the mixing time was altered in steps of 6 ('dry' rooms), or 12 ms ('wet' rooms), respectively. Results showed that the perceptual mixing time does not depend on the damping of an enclosure but on its volume, i.e. on the mean free path length, or on the order of undergone reflections, respectively. Furthermore, listening test results were used to test different model- and data-based predictors for their ability to predict the perceptual mixing time at different levels of sensitivity, revealing satisfactory candidates (74.7% - 83.5% explained variance). Hence, perceptual mixing time can now conveniently be predicted with desired strictness based on geometrical information or measurements, allowing balancing the computational effort and the desired perceptual accuracy.

The last study presented in Part II was related to the 5th objective (increasing ecological validity of data-based binaural synthesis): In 2004, Algazi and colleagues proposed a new method for recording pseudo-binaural signals suitable for dynamic rendering and denoted it Motion Tracked Binaural Sound (MTB). Recording MTB signals typically requires a rigid spherical shell of approximately a human head's size which is equipped with an equatorial array of microphones. Multichannel MTB recordings are then played back by interpolating the signals of pairs of microphones located closest to the positions of the ears of a listener as estimated from head tracking. Although reproduced binaural cues are more or less distorted (ITD tied to sphere diameter, SC corrupted by missing pinnae) MTB is currently the only procedure for recording natural acoustic atmospheres allowing a later rendering with (at least some) dynamic binaural cues. Hence, due to its potential of increasing the perceived naturalism of data-based binaural simulations – i.e. by adding dynamic binaural background atmospheres – a MTB device was built and software allowing for interactive real-time playback was developed. Additionally, a first formal perceptual evaluation of the minimum required number of microphones and the optimum crossfade algorithm was conducted (Lindau and Roos, 2010). Results showed that perceived reproduction accuracy was nearly independent from the number microphones and the type of audio content used, at least, when using a minimum number of eight microphones together with the best interpolation algorithm (a two-band procedure using time-domain linear interpolation below 1500 Hz and spectral interpolation with phase switching above this frequency).

1.7.3 Part III: Perceptual Evaluation of Virtual Acoustic Environments

The three studies presented in Part III of this dissertation were devoted to the 6th research objective (development of suitable integrative and differentiated perceptual assessment methods for VAEs).

Having spent a considerable amount of work into the improvement of different aspects of data-based dynamic binaural synthesis, it was found increasingly relevant to quantify the achieved overall gain. Intuitively, *whether a virtual acoustic scene may be perceived as real or not* by a listener appears to be a fundamental criterion of simulation accuracy suitable for VAEs. Apparently, the degree of 'perceived realness', or the decision whether a certain auditory event is caused by a simulated or a real sound source, relates to an assessment with respect to a subject's conviction or 'inner reference', which in turn is built on former experience and expectations (cf. Kuhn-Rahloff, 2011). In close relation, the notion of *plausibility* has been discussed in the literature (cf. sect. 1.5).

Hence, in the first study presented in Part III of this dissertation (Lindau and Weinzierl, 2012) a plausible simulation was defined to be perceived as being *in accordance with the expectations towards a corresponding real acoustic event.* It was mentioned before that the AB listening test paradigm fails in validly assessing the inner reference, being biased by consistent but maybe only accidentally correct 'detections' of the simulation. Hence, concepts for empirically assessing the 'acceptability as real' had to be reworked. It was found that testing plausibility resembled conducting a discrimination task with respect to an inner reference, in turn requiring presentations of singular stimuli and Yes/No (or 'real'/'simulated') decisions from the subjects. While using acoustically transparent headphones, Lindau and Weinzierl presented a random order of simulated and real sound sources to the test subjects and asked them – individually for each stimulus – whether they perceived the stimuli as real or not. However, the problem with the Yes/No-task is that, *per se*, results are susceptible to being biased by a subject's answering criterion, i.e. his/her disposition to respond – independent of the actual sensation – 'Yes' or 'No' more often. However, if both simulated and real stimuli may be conveyed in an alternating but concealed fashion, objectively true and false decisions may be identified. Furthermore, the answering behavior may be analyzed using signal detection theory (SDT, Green and Swets, 1974) allowing for an independent quantification of the subject's response bias and true sensory discriminability. Additionally, it has to be noticed that from the view point of inferential statistics our original research objective (proving the indistinguishability from reality) aimed at proving a non-effect. To this end, a conservative minimum effect hypothesis stating a near negligible residual detectability of the simulation (55% detection rate) was assessed in a fair and highly powered binomial test. It could be shown, that the improved binaural simulation passed this strict test of plausibility whereas the previous simulator from Lindau et al. (2007) missed the *a priori* criterion shortly.

The concept of plausibility refers to the fact that in their daily lifes listeners seldom get the opportunity to compare artificial acoustic stimuli to the corresponding real references (e.g., musical recordings vs. real concerts). Instead, listeners will commonly rate a reproduction's accuracy with respect to their former experience or expectations, making the proposed test of plausibility particularly relevant for many (practical) applications (as, e.g., acceptability studies or effort reduction). However, for other applications (as, e.g., benchmarking or technical development) it might be more relevant to compare a simulation to some given reference stimulus. For denoting simulations which are perceptually indistinguishable from externally provided acoustic references the term 'authentic' was introduced before. Hence, while resembling in itself an important benchmark, the next study assessed, under which conditions binaural dynamic synthesis might be perceptually indiscernible from a given reference.

In contrast to plausibility, the operationalization of authenticity is straight forward: Any criterion-free (i.e. blind) detection test (as, e.g., the ABX test, Leventhal 1986) may be used. Former research results (Lindau and Brinkmann, 2012) already revealed that non-individual binaural simulations are clearly discernable from individual acoustic reality. Nevertheless, individual binaural simulations were still assumed to potentially provide perceptual authenticity. In order to test this hypothesis, Brinkmann et al. (2014) assessed the sensory discriminability of individual dynamic binaural simulations from the individual acoustic reality. Individual BRIRs of nine subjects were measured in a music recording room for two sound sources and for horizontal head movements in a range of ±34° in 2° steps using the insert microphones from Lindau and Brinkmann (2012). Binaural signals were presented using the BK211 headphones (Erbes et al., 2012) and individualized HpTF compensation. The *a priori* effect size was assumed to be large, thus, the ABX tests were designed to allow a fair and reliable revelation of an individual detection rate of at least 90%. Results showed that while for noise stimuli the individual binaural simulation is still clearly distinguishable from reality, for real world (male speech) stimuli and for a less problematic direction of sound incidence subjects on average performed at threshold level.

Whereas plausibility and authenticity constitute valuable integrative measures of simulation accuracy they do not give detailed insight into specific shortcomings as would be required, for, e.g., a directed technical improvement, or for qualified benchmarking. Hence, the last study included in this dissertation was devoted to the development of a consensual and differentiated descriptive language suitable for the perceptual assessment of VAEs and other applications of spatial audio technologies (Lindau et al., 2014). To this end, 21 German experts for virtual acoustics were gathered for several Focus Group discussions. Within 56 hours of discussions the experts produced a vocabulary comprising 48 terms describing auditive qualities (SAQI – Spatial Audio Quality Inventory) which could be sorted into eight categories (timbre, tonalness, geometry, room, time behavior, dynamics, artifacts, and general impression). Furthermore, for each quality additional short circumscriptions, end labels for rating scales, and – if felt needed – further illustrative audio samples were produced. Moreover, the expert group provided a list of typical assessment entities within VAEs (relating to scene elements, and the used technical apparatus) and a specification of how to deal with perceived modifications of qualities with respect to time variance and interactivity. Results were checked for semantic unambiguity by five additional German experts. Subsequently, the vocabulary was translated into English with the help of eight international experts which had been selected for their abilities in German and English language and their expertise in the field.

1.8 Summary of Original Achievements

Within the course of this doctoral thesis the following original results were obtained:

As an approach towards reducing the measurement effort required for data-based DBS, the just detectable opening angle between sound sources was assessed in an exemplary concert-hall-like scenario. Whereas the just detectable opening angle was shown to increase with distance (or decreasing D/R-ratio, respectively) and for specific stimuli, yet, for typical application scenarios a combination of several originally spatially separated sound sources into a singular new one appeared not recommendable (Lindau et al., 2008).

Furthermore, the minimum required angular resolution for head movements in non-individual data-based dynamic binaural synthesis was assessed for all rotational degrees of freedom (Lindau and Weinzierl, 2009). Results revealed similar sensitivity for all direction of head movement, at least when considering specific directions of sound incidence.

Spectral coloration – often perceived when comparing non-individual data-based dynamic binaural synthesis to the individual acoustic reality – was reduced by developing a perceptually optimal approach to frequency response compensation of the headphone transfer path (Lindau and Brinkmann, 2012). Furthermore, it was shown that for highest transparency of non-individual DBS, the compensation should be based on the HpTF of the subject/HATS that delivered the original binaural recordings.

A novel extraaural headphone system was developed to comply with numerous requirements which are (partly) specific for binaural VAEs: the compliance with the FEC criterion, a high transparency to exterior sound fields, insusceptibility to changes in mechanical fit, a frequency response being non-problematic for equalization, a high compatibility with typical accessories of multimodal VAEs, a wide frequency response, and high sound pressure levels at an inaudible self-noise (Erbes et al., 2012).

Localization instability – often perceived when comparing non-individual data-based dynamic binaural synthesis with the individual acoustic reality – was reduced by introducing an approach for the *post hoc* individualization of the ITD (Lindau et al., 2010).

Additionally, cross fade artifacts often perceived in dynamic binaural synthesis were reduced by using quasi-minimum-phase representations of measured BRIRs for auralization (Lindau et al., 2010).

Furthermore, the response latency just detectable in dynamic binaural synthesis could be shown to be lower as reported in previous literature (Lindau, 2009). Results further implied that the distance delay should be removed from BRIRs intended for dynamic auralization.

Model- and signal-based predictors have been empirically assessed for their ability to predict the perceptual mixing time. Results may conveniently be applied for designing both more efficient data- and model-based VAEs (Lindau et al., 2012).

A perceptual evaluation of the optimal parameterization of Motion Tracked Binaural Sound (MTB) now provides empirically substantiated guidelines for efficient recording and perceptually accurate playback of natural soundscapes with dynamic binaural cues (Lindau and Roos, 2010).

Concepts of plausibility and authenticity were identified as suitable integrative perceptual measures of simulation accuracy. Listening test designs were developed operationalizing (a) the construct of plausibility by means of Yes/No detection tasks and an analysis based on signal detection theory (Lindau and Weinzierl, 2012), and (b) the construct of authenticity by means of an ABX detection tasks and an analysis based on binomial tests (Brinkmann et al., 2014).

Using a Yes/No listening test paradigm and a statistical analysis based on signal detection theory, plausibility was shown to be achievable for an improved version of *non-individual* data-based dynamic binaural synthesis (Lindau and Weinzierl, 2012).

Using an ABX test paradigm test and a statistical analysis based on the binomial distribution, authenticity was shown to be achievable – at least for some individuals and certain audio contents – for an improved version of *individual* data-based dynamic binaural synthesis (Brinkmann et al., 2014).

Finally, using an expert Focus Group approach, detailed and consensual German and English descriptive vocabularies (Spatial Audio Quality Inventory, SAQI) were developed for the qualification and quantification of the auditory impression produced by VAEs when being compared amongst each other or with respect to some – imagined or externally provided – references (Lindau et al., 2014).

1.9 Perspectives

At the end of this dissertation's résumé current and future research question of interest shall be discussed.

To date, it is common for model-based VAEs to use HRTF datasets for auralization which do not account for different head-above-torso orientations (as occurring in real life listening with head movements). However, recently we have shown that a morphological correct consideration of the torso leads to an audible effect (Brinkmann et al., 2014b, accepted). Furthermore, in order to find an *efficient* representation of such HRTFs accounting for the torso effect, different interpolation approaches have been be perceptually evaluated (Brinkmann et al., 2014c, accepted).

Additionally, the effect of *different acoustical environments* on the perceived authenticity of individual data-based DBS is being assessed in ongoing listening tests.

Moreover, in order to characterize in more detail the state of the art in dynamic binaural synthesis which was achieved in this dissertation, both individual and non-individual approaches for data-based DBS are currently being evaluated with respect to deviations from reality, using questionnaires based on the newly developed SAQI vocabulary (Lindau et al. 2014b, accepted). Remaining systematic deviations may motivate continued technical improvements.

Furthermore, after having qualified and quantified remaining perceptual deviations from reality, essential requirements for using data-based DBS as a *reference simulation* in perceptual evaluations of VAEs will be fulfilled. As a first application a planned 'International Round Robin on Auralization' will involve – for a number of numerical algorithms for the simulation of room acoustics – *in simu*[3] assessments of the achieved simulation accuracy.

In the near future, it will also be of interest to evaluate the SAQI itself, e.g., by evaluating relevant statistial item characteristics and by studying its inter-language

[3] *In simu* is used here in contrast to *in situ*: Whereas the latter case refers to the perceptual evaluation of simulations against real sound fields, the former refers to evaluations where all stimuli (references and test stimuli) are simulations.

reliability. Furthermore, its usability may be increased by constructing a database of suitable training and anchor stimuli.

As another future topic it will be interesting to explore in how far simulations may differ from reality before becoming implausible. As in this case the auditory reference is an imagined, or better, a remembered one (i.e. a cognitive representation of what might still be real), the perceptual assessment of just tolerable limits will reveal – so far unknown – 'just memorable differences' (JMDs). It will be interesting to compare them to (known) just noticeable differences (JND).

When being asked to forecast future trends in DBS, one could state that a further increase in the perceptual accuracy of dynamic binaural synthesis can be expected from a full individualization of binaural cues (i.e. through the usage of individual HRTFs). While being still costly in terms of the involved measurement procedures, convenient alternatives might be expectable from a future combination of improved optical scanning techniques for the rapid acquisition of individual morphology and accelerated methods for the numerical calculation of HRTFs (see Jin et al., 2014).

1.10 References

Algazi, V. R.; Avendano, C.; Thompson, D. M. (1999): “Dependence of Subject and Measurement Position in Binaural Signal Acquisition”, in: *J. Audio Eng. Soc.*, **47**(11), pp. 937-947

Algazi, V. R.; Duda, R. O.; Thompson, D. M. (2004): “Motion-Tracked Binaural Sound”, in: *J. Audio Eng. Soc.*, **52**(11), pp. 1142-1156

Berg, J.; Rumsey, F. (2006): “Identification of quality attributes of spatial audio by repertory grid technique”, in: *J. Audio Eng. Soc.*, **54**(5), pp. 365-379

Blauert, J. (1997): *Spatial Hearing. The Psychophysics of Human Sound Localization*. 2.ed. Cambridge, MA.: MIT Press

Brinkmann, F.; Lindau, A.; Vrhovnik, M.; Weinzierl, S. (2014): “Assessing the Authenticity of Individual Dynamic Binaural Synthesis”, in: *Proc. of EAA Joint Auralization and Ambisonics Symposium*, Berlin, pp. 62-68, http://dx.doi.org/10.14279/depositonce-11

Brinkmann, F.; Roden, R.; Lindau, A.; Weinzierl, S. (2014b): “Audibility of different head-above-torso -orientations in head-related transfer functions”, accepted for *Forum Acusticum,* Krakow, Poland

Brinkmann, F.; Roden, R.; Lindau, A.; Weinzierl, S. (2014c): “Interpolation of head-above-torso-orientations in head-related transfer functions”, accepted for *Forum Acusticum,* Krakow, Poland

Burkhard, M. D.; Sachs, R.M. (1975): “Anthropometric manikin for acoustic research”, in: *J. Acoust. Soc. Am.,* **58**(1), pp. 214-222

Damaske, P.; Wagener, B. (1969): "Richtungshörversuche über einen nachgebildeten Kopf", in: Acustica, Vol. 21, pp. 30-35

de Boer, K.; Vermeulen, R. (1939): "Eine Anlage für einen Schwerhörigen", in: Philips’ Gloeilampenfabrieken: *Philips’ Technische Rundschau*, Vol. 4, No. 11, pp. 329-332

Erbes, V.; Schultz, F.; Lindau, A.; Weinzierl, S. (2012): “An extraaural headphone system for optimized binaural re-production”, in: *Proc. of the 38th DAGA (in: Fortschritte der Akustik).* Darmstadt, pp. 313-314

Gardner, W. G. (1995): “Efficient Convolution without Input-Output Delay”, in: *J. Acoust. Soc. Am.,* **43**(3), pp. 127-136

Gardner, W. G. (1997): *3-D Audio Using Loudspeakers*. doct. diss. MIT Media Laboratories. Cambridge

Genuit, K. (1981): “Ein Beitrag zur Optimierung eines Kunstkopfsystems”, in: *Proc. of the 12th Tonmeistertagung*. München, pp. 218-243

Green, D. M.; Swets, J. A. (1974): *Signal Detection Theory and Psychophysics*. Huntington: Krieger

Hammer, K.; Snow, W.: *Binaural Transmission System at Academy of Music in Philadelphia*. Memorandum MM 3950, Bell Laboratories, 1932

Hammershøi, D.; Møller, H. (1996): “Sound transmission to and within the human ear canal”, in: *J. Acoust. Soc. Am.*, **100**(1), pp. 408-427

Henning, G. Bruce (1974): “Detectability of interaural delay in high-frequency complex waveforms”, in: *J. Acoust. Soc. Am.,* **55**(1), pp. 84-90

Hess, W.; Weishäupl, J. (2014): “Replication of Human Head Movements in 3 Dimensions by a Mechanical Joint”, to be published in *Proc. of the International Conference on Spatial Audio,* Erlangen

Jin, Craig T. et al. (2014): "Creating the Sydney York Morphological and Acoustic Recordings of Ears Database", in: *IEEE Transactions on Multimedia*, **16**(1), pp. 37-46

Karamustafaoglu, A.; Horbach, U.; Pellegrini, R. S.; Mackensen, P.; Theile, G. (1999): “Design and applications of a data-based auralization system for surround sound”, in: *Proc. of the 106th AES Convention*. München, preprint no. 4976

Kim, C.; Mason, R.; Brookes, T. (2008): “Initial investigation of signal capture techniques for objective measurement of spatial impression considering head movement”, in: *Proc. of the 124th AES Convention*. Amsterdam, preprint no. 7331

Kleiner, M.; Dalenbäck, B.-I.; Svensson, P. (1993): “Auralization - An Overview”, in: *J. Audio Eng. Soc.,* **41**(11), pp. 861-875

Kürer, R.; Plenge, G.; Wilkens, H. (1969): “Correct spatial sound perception rendered by a special 2-channel recording method”, in: *Proc. of the 37th AES Convention*. New York, preprint no. 666

Kuhn-Rahloff, Clemens (2011): *Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen. Ein Beitrag zum Verständnis der „inneren Referenz“ perzeptiver Messungen.* Doct. diss., Technische Universität Berlin

Lehmann, P.; Wilkens, H. (1980): „Zusammenhang subjektiver Beurteilung von Konzertsälen mit raumakustischen Kriterien“, in: *Acustica*, **45**, pp. 256-268

Lentz, T. (2006): “Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments”, in: *J. Audio Eng. Soc.,* **54**(4), pp. 283-294

Lindau, A.; Weinzierl, S. (2006): “FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom”, in: *Proc. of the 24th Tonmeistertagung.* Leipzig, pp. 621-625

Lindau, A.; Hohn, T.; Weinzierl, S. (2007): “Binaural Resynthesis for Comparative Studies of Acoustical Environments”, in: *Proc. of the 122nd AES Convention.* Vienna, preprint no. 7032

Lindau, A.; Klemmer, M.; Weinzierl, S. (2008): “Zur binauralen Simulation verteilter Schallquellen (On the Binaural Simulation of Distributed Sound Sources)”, in: *Proc. of the 34th DAGA (in: Fortschritte der Akustik)*. Dresden, pp. 897-898

Lindau, A. (2009): “The Perception of System Latency in Dynamic Binaural Synthesis”, in: *Proc. of the 35th NAG/DAGA, International Conference on Acoustics (in: Fortschritte der Akustik).* Rotterdam, pp. 1063-1066

Lindau, A.; Weinzierl, S. (2009): “On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical, and Lateral Direction”, in: *Proc. of the EAA Symposium on Auralization.* Espoo

Lindau, A.; Estrella, J.; Weinzierl, S. (2010): “Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD”, in: *Proc. of the 128th AES Convention*. London, preprint no. 8088

Lindau, A.; Roos, S. (2010): “Perceptual Evaluation of Discretization and Interpolation for Motion-Tracked Binaural (MTB-) Recordings”, in: *Proc. of the 26th Tonmeistertagung*. Leipzig, pp. 680-701

Lindau, A.; Brinkmann, F. (2012): “Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-individual Recordings”, in: *J. Audio Eng. Soc.,* **60**(1/2), pp. 54-62

Lindau, A.; Kosanke, L.; Weinzierl, S. (2012): “Perceptual Evaluation of Model- and Signal-based Predictors of the Mixing Time in Binaural Room Impulse Responses”, in: *J. Audio Eng. Soc.,* **60**(11), pp. 887-898

Lindau, A.; Weinzierl, S. (2012): “Assessing the Plausibility of Virtual Acoustic Environments”, in: *Acta Acustica united with Acustica*, **98**(5), pp. 804-810, DOI: http://dx.doi.org/10.3813/AAA. 918562

Lindau, A.; Lepa, S. (2014): „Dynamische Binauralsynthese – ein Verfahren der virtuellen Akustik als Ansatz zur Untersuchung technologiebezogener Hypothesen im Rahmen medienpsychologischer Rezeptionsexperimente“, to be published in: *Tagungsband der 15. Tagung der Fachgruppe Methoden der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft*, in print

Lindau, A.; Brinkmann, F.; Erbes, V.; Maempel, H.-J.; Lepa, S.; Weinzierl, S. (2014): “A Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI)”, accepted for *EAA Joint Auralization and Ambisonics Symposium*, Berlin

Lindau, A; Brinkmann, F.; Weinzierl, S. (2014b): “Qualitative and Quantitative Deviations of Individual and Non-individual Dynamic Binaural Synthesis from Acoustic Reality”, accepted for *Forum Acusticum,* Krakow, Poland

Lokki, T.; Järveläinen, H. (2001): “Subjective evaluation of auralization of physics-based room acoustics modeling”, in: *Proc. of ICAD 2001 - Seventh Meeting of the International Conference on Auditory Display*. Espoo

Lombard, M.; Ditton, T. (1997): “At the Heart of It All: The Concept of Presence”, in: *J. Computer Mediated-Communication*, **3**(2)

Lorho, G. (2005): “Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction”, in: *Proc. of the 119th AES Convention*. New York, preprint no. 6629

Mackensen, P. (2004): *Auditive Localization. Head Movements, an additional cue in Localization*. Doct. diss. Technische Universität Berlin

Maempel, H.-J.; Lindau, A. (2013): “Opto-acoustic simulation of concert halls – a data-based approach”, in: *Proc of the 27. Tonmeistertagung*. Köln, pp. 293-309

Middlebrooks, J. C.; Makous, J. C.; Green, D. M. (1989): “Directional sensitivity of the sound-pressure level in the human ear canal”, in: *J. Acoust. Soc. Am.*, **86**(1), pp. 89-108

Minnaar, P. et al. (2001): “Localization with Binaural Recordings from Artificial and Human Heads”, in: *J. Audio Eng. Soc.*, **49**(5), pp. 323-336

Møller, H. (1992): “Fundamentals of binaural technology”, in: *Applied Acoustics*, **36**(3-4), pp. 171-218

Møller, H. et al. (1995): “Transfer Characteristics of Headphones Measured on Human Ears”, in: *J. Audio Eng. Soc.*, **43**(4), pp. 203-221

Møller, H.et al. (1995b): “Head-Related Transfer Functions of Human Subjects”, in: *J. Audio Eng. Soc.,* **43**(5), pp. 300-332

Møller, H. et al. (1996): “Binaural Technique: Do We Need Individual Recordings?”, in: *J. Audio Eng. Soc.,* **44**(6), pp. 451-469

Møller, H. et al. (1996b): “Using a Typical Human Subject for Binaural Recording”, in: *Proc. of the 100th AES Convention*. Kopenhagen, preprint no. 4157

Moldrzyk, C. (2002): “Ein neuartiger Kunstkopf zur Verifikation einer akustischen Entwurfsmethodik für Architekten”, in: *Proc. of the 22th Tonmeistertagung.* Hannover

Müller, S.; Massarani, P. (2001): “Transfer-Function Measurement with sweeps”, in: *J. Audio Eng. Soc.,* **49**(6), pp. 443-471

Müller-Tomfelde, C. (2001): “Time varying Filters in non-uniform Block Convolution”, in: *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFX-01).* Limerick

Nicol, R. (2010): *Binaural Technology*. AES Monograph. New York: Audio Engineering Society Inc.

Oppenheim, A. V.; Schafer, R. W. (2004): *Zeitdiskrete Signalverarbeitung.* 2.ed. München et al.: Pearson Education

Paul, S. (2009): "Binaural Recording Technology: A Historical Review and Possible Future Developments", in: A*cta Acustica united with Acustica*, **95**(5), pp. 767-788

Pellegrini, R. S. (2001): *A virtual reference listening room as an application of auditory virtual environments.* Doct. diss., Institut für Kommunikationsakustik der Fakultät für Elektrotechnik und Informationstechnik der Ruhr-Universität Bochum. Berlin: dissertation.de

Reilly, A.; McGrath, D. (1995): “Real-Time Auralization with Head Tracking”, in: *Proc. of the 5th Australian Regional AES Convention. Sydney*, preprint no. 4024

Sandvad, J. (1996): “Dynamic Aspects of Auditory Virtual Environments”, in: *Proc. of the 100th AES Convention*. Kopenhagen, preprint no. 4226

Schärer Kalkandjiev, Z.; Weinzierl, S. (2013): “Room acoustics viewed from the stage: Solo performers' adjustments to the acoustical environment”, in: *Proc. of the International Symposium on Room Acoustics (ISRA)*, Toronto, paper no. 86

Schubert, T.; Friedmann, F.; Regenbrecht, H. (2001): “The Experience of Presence: Factor Analytic Insights”, in: *Presence: Teleoperators and Virtual Environments.* Cambridge, MA.: MIT Press, **10**(3), pp. 266-281

Silzle, A. (2007): *Generation of Quality Taxonomies for Auditory Virtual Environments by Means of Systematic Expert Survey*. Doct. Diss., Fakultät für Elektrotechnik und Informationstechnik an der Ruhr-Universität Bochum, Aachen: Shaker

Spors, S.; Wierstorf, H.; Raake, A.; Melchior, F.; Frank, M.; Zotter, F. (2013): “Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State”, in: *Proc. of the IEEE*, **101**(9), pp. 1920-1938

Stern, Richard M.; Brown, Guy J.; Wang, Deliang (2006): “Binaural Sound Localization”, in: Wang, DeLiang; Brown, J. Guy (eds.): *Computational Auditory Scene Analysis. Principals, Algorithms, and Applications*. Hoboken, NJ: John Wiley & Sons, pp. 147-185.

Strutt (Lord Rayleigh), J. W. (1907): "On Our Perception of Sound Direction", in: *Philosophical Magazine*, Series 6, pp. 214–232

Vorländer, M. (2007): *Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality.* Berlin et al.: Springer

Wade, N. J.; Deutsch, D. (2008): “Binaural Hearing—Before and After the Stethophone”, in: *Acoustics Today*, **4**(3), pp. 16-27

Welch, G.; Foxlin, E. (2002): “Motion Tracking: No Silver Bullet, but a Respectable Arsenal”, in: *IEEE Computer Graphics and Applications*, Nov./Dec., pp. 24-38

Wenzel, E. M. (1996): “What Perception Implies About Implementation of Interactive Virtual Acoustic Environments”, in: *Proc. of the 101st AES Convention*. Los Angeles, preprint no. 4353

Wenzel, E. M. et al. (1990): “A System for Three-Dimensional Acoustic Visualization in a Virtual Environment Workstation”, in: *Proc. of the First IEEE Conference on Visualization*. San Francisco, pp. 329-337

Wilska, Alvar (1938): „Untersuchungen über das Richtungshören“, in: *Acta Societas Medicorum Fennicae „Dudodecim“*, Vol. A, Tom. XXI, Fasc. 1.

1.11 Used Abbreviations

3AFC	three alternative forced choice (task)
ABX	duo-trio detection task with constant reference mode
BK211	transaural headphones belonging to the BKsystem
BKamp	power amplifier belonging to the BKsystem
BKsystem	headphone system for binaural signal reproduction comprising BKamp power amplifier and BK211 headphones
BRIR	binaural room impulse response
BRTF	binaural room transfer function
CAD	computer aided design
(D)DBS	(data-based) dynamic binaural synthesis
FABIAN	**F**ast and **A**utomatic **B**inaural **I**mpulse response **A**cquisitio**N**, automatic head and torso simulator of the TU Berlin
FEC	free air equivalent coupling
FFT	fast Fourier transform
HATS	head and torso simulator
HMD	head mounted display
HpTF	headphone transfer function
HRTF	head-related transfer function

IC	interaural correlation
ILD	interaural level difference
ITD	interaural time delay
JMD	just memorable difference
JND	just noticeable difference
MTB	motion-tracked binaural (sound)
SAQI	Spatial Audio Quality Inventory
SC	spectral cue(s)
SDT	signal detection theory
VAE	virtual acoustic environment

Part I
Binaural Recording

2 FABIAN - An Instrument for Software-based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom

The following chapter is an authorized reprint of the article

> Lindau, Alexander; Weinzierl, Stefan (2006): “FABIAN - An Instrument for Software-based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom”, in: *Proc. of the 24th Tonmeistertagung.* Leipzig, pp. 621-625.

The article has been faithfully reproduced from the author’s post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, changes to order and position of figures, use of US-American spelling, typographic and stylistic corrections.

2.1 Abstract

FABIAN is an instrument for the **F**ast and **A**utomatic **B**inaural **I**mpulse response **A**cquisitio**N**. It uses a new head and torso simulator whose orientation can be controlled in multiple degrees of freedom via a servo-motorized neck joint, while the whole torso can be rotated on a motorized turntable device. A Matlab® application is controlling the measurement process acquiring binaural room impulse responses in high spatial resolution. They shall be used for the simulation of natural sound fields as well as electroacoustic reproduction setups by auralization through binaural technology.

2.2 Concept

Nearly all perceptual information used for orientation in our auditory environment is coded in the sound pressure at our eardrums. Hence, natural or artificial sound fields can be simulated on the basis of binaural room impulse responses (BRIRs), representing the acoustical transmission path from a sound source to the listener. Precision and stability of source localization as well as the depth of immersion during auralization can be increased by means of dynamic head tracking to account for movements of the listener in the simulated environment.

Available head and torso simulators (HATS) for the measurement of BRIRs such as the KEMAR-Manikin [1], the Head Acoustics HMS-II/III series, the Bruel & Kjaer 4100, and the former Cortex (now Metravib) Mk1 as well as systems in the academic area such as the VALDEMAR-HATS [2] or the Aachen-head [3] are static systems with only limited possibilities to emulate movements of head, shoulders, and torso. For the development of the BRS-processor a Neumann KU100 was used to measure BRIRs for horizontal and vertical directions by manual step-by-step reorientation [4]. A HATS-system developed at the TU Berlin [5] was the first to allow for an automated measurement of BRIRs for a horizontal rotation range of ±75°.

The measurement of complete sets of BRIRs in high spatial resolution is a lengthy process already for single-point auralizations enabling different orientations of head and torso relative to each other as well as absolute torso orientations. When the listener shall be allowed to move freely on a two-dimensional grid of BRIR-sets during binaural simulation, speed and automation of the measurement process become even more relevant.

2.3 System Design

To this end, an existing HATS system [6] was extended. Horizontal rotation and tilting of the dummy head in arbitrary angles is now possible by means of a servo-motorized neck joint (cf. Figure 2-1). By reversing the joint's position the third rotational degree of freedom (lateral flexion) becomes accessible. When mounted on a special turntable the torso can be rotated as a whole (cf. Figure 2-2). Thus, any spherical grid of head and torso orientations can be defined for the measurement of binaural impulse responses. The used devices allow exact and fast reorientation with negligible effect on the actual measurement duration.

A new unisex corpus has been designed according to anthropometric data representing the 18–65 year old German population's median values [7]. For measuring in sitting or upright position the corpus was designed to be modular to some extent and can be detached from turntable and supporting stands.

The outer silhouette was formed according to anthropometric models from [8] to get a more human-like appearance (cf. Figure 2-1). Since most HRTFs measured on randomly selected individuals result in better localization performances than those measured on all commercially available artificial heads [9], the head used for FABIAN is a gypsum mold from a human individual [6]. Its perceptual performance has already been evaluated [10]. The complex fine structure of the outer ear is preserved by silicone molds made from individual human ears.

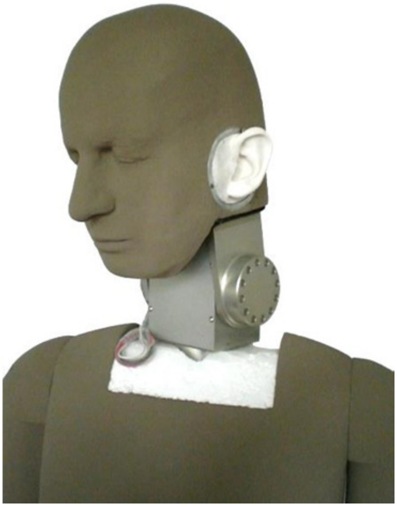


Figure 2-1. Close-up view of the FABIAN dummy head with uncovered neck joint.

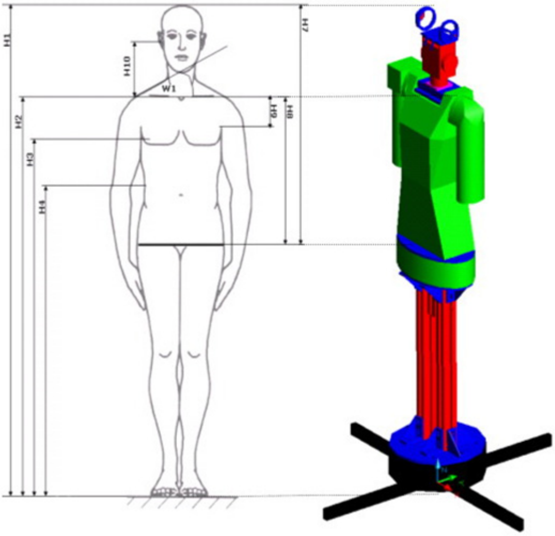


Figure 2-2. Anthropometric measures and the resulting 3D-CAD drafts of the modular torso.

The ears are exchangeable and equipped with low noise miniature condenser microphones DPA 4060 (ø 5.6 mm).

The dummy head's microphones are located at the bottom of the cavum conchae at the beginning of the ear canal entrance. From this point to the eardrums there exists no directional dependence of the transfer functions [11]. The influence of microphones and reproduction setup has to be compensated by post-equalizing the BRIRs.

2.4 Measurement Process

A mobile PC controls head and body movements. It also conducts impulse response measurements using swept sine technique with additional noncyclic IR-deconvolution [12]. A custom multi-channel impulse response measurement application has been implemented. Hard- and software is supporting 44.1 to 96 kHz sampling frequencies and 8 to 32 bit word length for audio data. The two input signals of the dummy head's microphones are acquired while up to 8 simultaneously connected outputs can be used to drive different source-positions (cf. Figure 2-3). By successively stimulating the sources before re-orientating the dummy the measurement process is accelerated considerably.

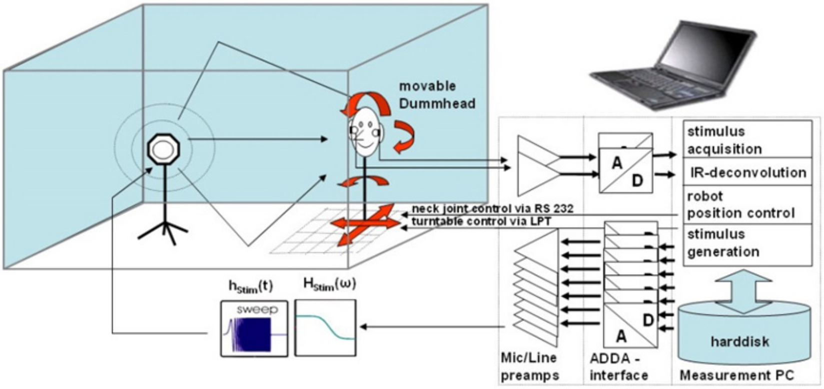


Figure 2-3. Complete setup for the measurement of room impulse responses in multiple degrees of freedom.

Measurements properties as level, duration (FFT-block-size) and spectral coloration of the stimulus and the number of averages can be chosen to adapt the measurement to ambient conditions such as noise level or reverberation time. In this respect the swept-sine measurement has convincing advantages compared with other measurement methods as MLS or TDS [12]. If input clipping is detected, the stimulus level is lowered and the measurement will be repeated automatically. If the signal-to-noise ratio (SNR) temporarily falls below a given threshold while acquiring data a repetition of the last measurement is initiated, too.

During a first trial a 7-channel surround setup has been measured in a recording studio environment. 14.000 BRIRs with 1° horizontal and 5° vertical resolution were collected at the sweet spot during 33 hours of unsupervised measurements. A mean SNR of 107 dB ipsilateral resp. 99 dB contralateral was reached using a bass-emphasized linear sweep of FFT-order 17 with two averages.

2.5 Outlook

Further research will be focused on the comparison of natural auditive perception versus electroacoustic representations by different recording and reproduction techniques, simulation-based *in situ* comparison of room acoustics, and auralization of sound fields where the listener is allowed to move freely in binaurally sampled acoustical environments.

2.6 References

[1] Burkhard, M.D.; Sachs, R.M. (1975): “Anthropometric manikin for acoustic research”, in: *J. Acoust. Soc. Am.,* **58**(1), pp. 214-222

[2] Christensen, F.; Jensen, C. B.; Møller, H. (2000): “The Design of VALDEMAR - An Artificial Head for Binaural Recording Purposes”, in: *Proc. of the 109th AES Convention*, Los Angeles, preprint no. 4404

[3] Schmitz, A. (1995): “Ein neues digitales Kunstkopfmeßsystem”, in: *Acustica*, **81**, p. 416-420

[4] Mackensen, P. (2003): *Auditive Localization. Head Movements, an additional cue in Localization.* Doct. dissertation, Berlin: Technische Universität

[5] Moldrzyk, C. (2002): “Ein neuartiger Kunstkopf zur Verifikation einer akustischen Entwurfsmethodik für Architekten”, in: *Proc. of the 22nd Tonmeistertagung,* Hannover: 2002

[6] Moldrzyk, C.; Ahnert, W.; Feistel, S.; Lentz, T.; Weinzierl, S. (2004): “Head-Tracked Auralization of Acoustical Simulation”, in: *Proc. of 117th AES Convention*, San Francisco, preprint no. 6275

[7] DIN 33402-2 E (2005): *Körpermaße des Menschen - Teil 2: Werte*. Berlin: Deutsches Institut für Normung

[8] IEC TR 60959 (1990): *Provisional head and torso simulator for acoustic measurements on air conduction hearing aids*. Geneva: International Electrotechnical Commission

[9] Minnaar, P.; Olesen, S. K.; Christensen, F.; Møller, H. (2001): "Localization with Binaural Recordings from Artificial and Human Heads", in: *J. Audio Eng. Soc.*, **49**(5), pp. 323-336

[10] Moldrzyk, C.; Lentz, T.; Weinzierl, S. (2005): "Perzeptive Evaluation binauraler Auralisationen", in: *Proc. of the 31th DAGA*, Munich

[11] Møller, H. (1992): "Fundamentals of binaural technology", in: *Applied Acoustics*, **36**(3/4), pp.171-218

[12] Müller, S.; Massarani, P. (2001): "Transfer-Function Measurement with sweeps", in: *J. Audio Eng. Soc.*, **49**(6), pp. 443-471

3 On the Binaural Simulation of Distributed Sound Sources

The following chapter is based on the author's postprint of the abstract-reviewed article:

> Lindau, Alexander; Klemmer, Martin; Weinzierl, Stefan (2008): "Zur binauralen Simulation verteilter Schallquellen", in: *Proc. of the 34th DAGA (in: Fortschritte der Akustik)*. Dresden, pp. 897-898.

The article was originally published in German. For reading convenience, it is presented here in an English translation.

3.1 Research Question

The binaural simulation of spatially distributed sound sources (orchestras, ensembles for chamber music, choirs etc.) would ideally require a large number of binaural data sets, recorded individually for each combination of sound source and listener position. Aiming at a reduction of the effort of both the measurement and the auralization of complex sound scenes, it was examined for which spatial configurations of sound sources it is actually required to use data sets of binaural room impulse responses which have been recorded individually for each sound source. To this end a listening test was conducted, to determine the just detectable opening angle between two or four sound sources, respectively, in different listening distances distances and for different source signals.

3.2 State of the Art

Only few systematic studies could be found to be concerned with the localizability of multiple incoherent sound sources in diffuse sound fields (for an overview see, e.g., [1], chapter 3). However, results showed that localization performance depended on both the stimulus type, and the sound source's distance (and hence on the amount of diffuse reverberation at the listener's position). Further, the perceived source width had been reported to increase with distance [2], [3].

While applying room acoustic modeling techniques and binaural auralization two recent studies showed that a symphonic orchestra simulated with multiple individual sound sources is perceived as more realistic as compared to being simulated using a singular monopole sound source or a distribution of multiple coherent monopoles [4], [5]. Both studies used static auralization, i.e. a binaural simulation not responding to head movements of the listeners. While studies could establish a rank order with respect to plausibility of the different types of representing an orchestra, it was not assessed if there was a listening distance below which an auralization using spatially distributed sound sources could reliably be distinguished from an auralization using only a singular sound source.

3.3 Methods

In order to assess the minimum needed opening angles (as seen from the listener) which are needed for perceiving sound sources as spatially separated, data sets of binaural room impulse responses (BRIRs) were measured in a large hall (auditorium maximum of TU Berlin, V = 8600 m^3, RT = 2.1 s, r_{crit} = 3.6 m) with the help of the binaural measurement robot FABIAN [6]. BRIRs were measured at two listening positions and for different opening angles between several sound sources. The distance between sources and receiver was chosen as (a) once the critical distance (4 m) and (b) its four-time equivalent (16 m). Based on these data sets threshold values for the angle of just noticeable separation of the sound sources were determined in a listening test applying dynamic (with head tracking) binaural synthesis. To this end, the opening angle between the sound sources was altered adaptively, depending on the detection performance of the subjects. As stimuli, two- and four-channel recordings of either noise pulses (each channel filtered with a different bandpass of the width of one octave, pulse duration: 250 ms, pause duration: 750 ms, different temporal offset for each channel) or 'natural' stimuli (recordings of a string quartet recorded at the anechoic chamber at the Technical University Berlin, anechoic recordings from woodwind instruments and recordings of a speaker, duration: 5 s) were presented. BRIRs were measured with an angular resolution of 1° in the horizontal plane allowing horizontal head movements in a range of ± 80°. Before conducting the main experiment, suitable angular ranges and increments of the opening angle between two sound sources were determined in a pre-test using five subjects. From results of this pre-test the angular increment was chosen to be 2°, whereas the maximum opening angle was 20° or 22°, resp., depending on the chosen distance (Figure 3-1, Figure 3-2).

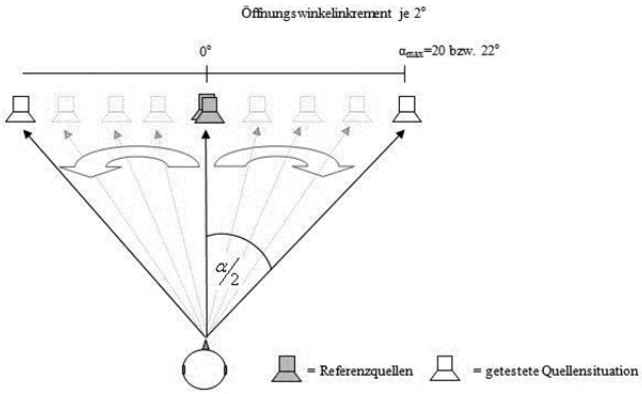

Öffnungswinkelinkrement je 2°
0°
α_{max}=20 bzw. 22°
$\alpha/2$
= Referenzquellen
= getestete Quellensituation

Figure 3-1. Situation 1 simulated in the listening test (two sound sources): During each trial in the adaptive listening test procedure subjects had to listen twice to the reference stimulus and once to the 'hot' sample (i.e. a situation where the sound sources were arranged with a variable opening angle *α*) while having to detect the 'hot' sample. Shaded loudspeakers in the plot indicate the reference situation.

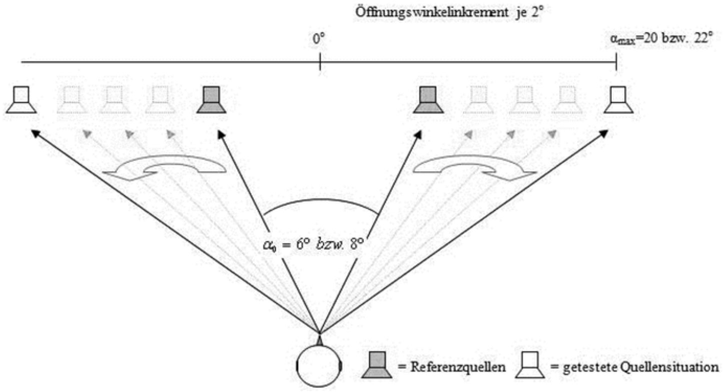

Öffnungswinkelinkrement je 2°
0°
α_{max}=20 bzw. 22°
α_0 = 6° bzw. 8°
= Referenzquellen
= getestete Quellensituation

Figure 3-2. Situation 2 simulated in the listening test (four sound sources): The reference stimulus consisted of the two inner (shaded) sound sources whereas the opening angle between the two outer sound sources was successively altered.

The listening test was conducted with 22 subjects (90% male); most of them had received some sort of musical education. The test was designed to be one-dimensional, three-way (distance, audio stimulus, and number of sources), two-group (close/distant, noise/natural, two/four) and completely varied. For all subjects the just noticeable opening angle was measured under all factor-combinations (fully repeated measures design). Hence, for each subject, the just detectable opening angle was measured as a function of distance (4 m vs. 16 m), audio stimulus (noise pulse complex vs. natural audio example) and the number of sound sources (two vs. four). At the beginning of the experiments, individual training sessions were carried out using easily discriminable example stimuli.

3.4 Results

The inter-subject reliability was satisfactory high (Cronbach's α 0.779). Thresholds values *per subject* were distributed normally (Kolmogoroff-Smirnoff test). The standard deviation was relatively small (σ = 2.84°), the average of the just detectable opening angles was 10.42°. However, the distribution of thresholds values *across all subjects* was not normal. Additionally, homogeneity of variances could not be shown (Levene's test). Thus, main effects were examined using non-parametric statistical tests (Wilcoxon signed-rank tests for dependent samples). The results of the statistical analysis can be summarized as follows:

(1) The just noticeable opening angle between sound sources w*as not significantly affected by the number of individual sound sources* (within the examined variation of two and four sources).

(2) The just noticeable opening angle *increased with the distance to the source* (or, with increasing diffuse-reverberant ratio, $p_{Wilcoxon} = 0.101$, Δ = 1° for the two tested distances of 4 m or 16 m, resp.).

(3) As compared to natural stimuli the just noticeable opening angle *was significantly smaller for noise pulses* ($p_{Wilcoxon} = 0.044$, Δ = 1.4 °).

The first result may be interpreted as an indication that the higher cognitive effort required for tracking more sound events does not come at the cost of a decreasing localization performance. Thus, the just noticeable opening angle between different sound sources might be discussed regardless of the number of sound sources actually being presented. Figure 3-3 shows the cumulated distribution of all individual threshold values (solid curves), interpretable as an estimate of population's psychometric function. Additionally, and according to results two and three stated above, the plots in Figure 3-3 further differentiate the results according to observed effects of distance and audio content (broken and dotted curves in Figure 3-3).

3.5 Discussion

Found results allow formulating criteria for the use of separately simulated sound sources in binaural auralizations. Thus, if two sound sources are separated by an opening angle of $\alpha \leq 10°$ for the majority of listeners (i.e. 50 %) a combined simulation (i.e. by mono-summation at an intermediate location) may not by discernable from a simulation using individual sound sources. To ensure indistinguishability for nearly all of the listeners (i.e. 95 %) only sound sources originally separated by an opening angle of $\alpha \leq 5°$ may be combined into a singular simulated sound source.

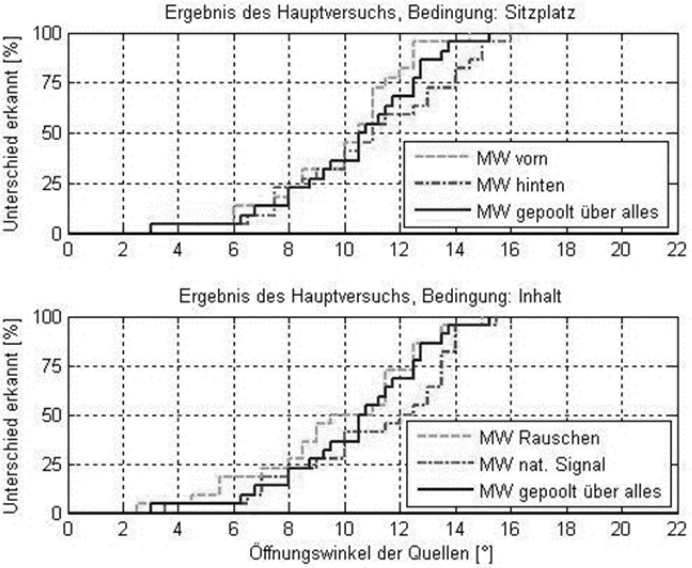


Figure 3-3. Cumulative distribution of thresholds values for the just noticeable opening angle between sound two or four sound sources (solid lines in both plots). Upper plot: Results aggregated for the smaller listening distance (dashed line), and for the larger listening distance (dash-dotted line). Lower plot: Results aggregated for the noise pulse complex (dashed line), and for the 'natural' audio stimuli (dash-dotted line).

Hereby, the above stated relations apply to well localizable source signals (such as pulsed noise) and to an approximately balanced ration of direct and reverberant energy. For music and speech signals or for an increased proportion of reverberant energy the just noticeable opening angle might increase by about one to two degree. Therefore, when simulating a string quartet at a frontal listener seat in a typical concert hall all sound sources should be simulated by individual sound sources. However, when simulating at a comparably large listening distance (i.e. a rear seat) a presentation using two or three individual sound sources might be sufficient.

3.6 References

[1] Blauert, J. (1997): Spatial Hearing. The Psychophysics of Human Sound Localization, 2nd ed., Cambridge, MA.: MIT Press.

[2] Damaske, P. (1967): “Subjektive Untersuchungen von Schallfeldern”, in: Acustica, **19**, pp. 199-213.

[3] Wagener, B. (1971): “Räumliche Verteilung der Hörrichtungen in synthetischen Schallfeldern”, in: Acustica 25, S. 203-219.

[4] Vigeant, M. C.; Wang, L. M.; Rindel, J. H. (2007): “Investigations of Multi-Channel Auralization Technique for Solo Instruments and Orchestra”, in: Proc. of the Int. Congress on Acoustics (ICA 2007), Madrid

[5] Witew, I.B.; Paprotny, J.; Behler, G. (2006): “Auralization of orchestras in concert halls using numerous uncorrelated sources”, in: Proc. of the Institute of Acoustics of the RWTH Aachen, **28**(2), pp. 293-296

[6] Lindau, A.; Weinzierl, S. (2007): “Fabian - Schnelle Erfassung binauraler Raumimpulsantworten in mehreren Freiheitsgraden”, in: Proc. of the 33rd DAGA, Stuttgart, pp. 633-634.

4 On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical, and Lateral Direction

The following chapter is an authorized reprint of the full-paper peer-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander; Weinzierl, Stefan (2009): "On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical, and Lateral Direction", in: *Proc. of the EAA Symposium on Auralization,* Espoo.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, use of US-American spelling, typographic and stylistic corrections.

4.1 Abstract

Dynamic binaural synthesis based on binaural room impulse responses (BRIRs) for a discrete grid of head orientations can provide an auralization naturally responding to head movements in all rotational degrees of freedom. Several experiments have been conducted in order to determine thresholds of just detectable BRIR grid resolution for all three rotational directions of head movements using an adaptive 3AFC procedure. Different audio stimuli as well as BRIR datasets measured in different acoustic environments were used. The results obtained reveal a high sensitivity of listeners towards discretization effects not only in horizontal, but also in vertical and lateral directions. Values indicate a minimum spatial resolution necessary for a plausible binaural simulation of acoustic environments.

4.2 Introduction

The simulation of acoustic environments by means of dynamic binaural synthesis based on measured BRIRs can provide a very high degree of realism [1]. An integral prerequisite for perceptual quality is the realistic interaction between the listener's head movements and the synthesized sound field. It is important whether the virtual acoustic environment (VAE) is able to track all rotational head movements (cf. Figure 4-1) and how fine angular movements can be resolved with respect to the underlying BRIR data set resolution.

At the same time, higher resolutions of BRIR data sets bring about longer measurement times for acquisition, as well as higher computational cost and memory size for auralization. Therefore, measured thresholds of just noticeable BRIR grid granularity are crucial in order to optimize the effort for measuring and auralizing binaural data without introducing perceptual artefacts.

The spatial resolution of the human auditory system has been operationalized with different measures. These include the localization blur, i.e. the mean error made when identifying the spatial position of a sound source, and the minimum audible angle (MAA), i.e. the minimum detectable displacement of a sound source. In anechoic environments, MAAs of 1°–10° have been found, depending on frequency and direction of sound incidence [2]. However, none of these measures can directly be used to derive a necessary resolution of BRIR data sets as natural sound fields contain reflections from all directions of incidence and listeners are free in orientation and velocity of their head movements.

Today, most VAEs track horizontal and sometimes also vertical head movements. The provided spatial resolution is different between implementations (Table 4-1). Moreover, it is common to use HRTF or BRIR datasets with lower resolution interpolated to finer grid sizes [3]–[6].

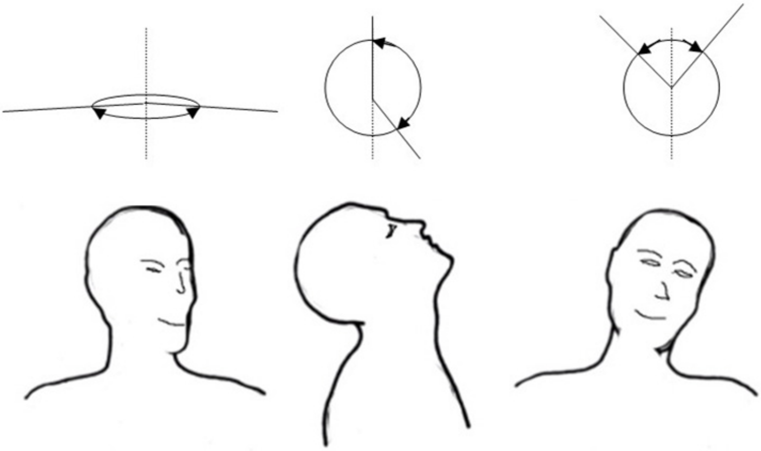


Figure 4-1. Rotational degrees of freedom of head movements (from left to right): Horizontal, vertical and lateral rotation and typical movement ranges.

Here, studies on the audibility of interpolations between HRTF data [7] showed that an original data set with 2° horizontal and vertical resolution could be inaudibly replaced by linear interpolations of a data set reduced to 2°–36° resolution for static sources and 8°–90° resolution for moving sound sources. Again, the values strongly depended on the direction of incidence. The (measured or interpolated) resolution required for a plausible auralization of acoustical environments has, however, never been investigated under realistic conditions, i.e. for natural source signals, for natural (non-anechoic) environments, and for all rotational degrees of freedom of head movements.

Table 4-1. Resolution and ranges of HRTF/BRIR data sets provided with some recent VAEs. (*missing full sphere elevation data generated by repetition, **original data range unknown, ***extrapolated from restricted original dataset)

System	Resolution (hor./ vert.)	Range	Ref.
EASE 4	5-30° / 10°	hor. ±180°, ver: [-45°; 90°]*	[8], [9]
ODEON 9	5° / 5.6°	hor. ±180°, ver: [-45°; 90°]*	[10], [11]
Raven	1° / 5° (interp. to 1°/2°)	full sphere	[12], [13]
IKA-SIM	15° / 10° (interp. to 5°)	full sphere	[3]
DIVA	10° / 15° (interpolated)	full sphere**	[14]
SLAB	10° / 10° (interpolated)	full sphere***	[15], [16]
SSR	5° (interp. to 1°)	hor. ±180°	[4], [17]
BRS	6° (interp. to 1°)	hor. ±43°	[5]

4.3 Method

4.3.1 BRIR Data Sets

At present, only the Berlin HATS (head and torso simulator) FABIAN [1] provides a fast and automated measurement of binaural room impulse responses in all degrees of freedom of head movement. For the present investigation FABIAN has been used to acquire binaural impulse responses in three different acoustical environments, including

- an anechoic chamber,
- a recording studio, and
- two large lecture halls.

Therefore the HATS was seated in a specific distance from a sound source positioned for frontal sound incidence. In the anechoic chamber a distance providing acoustic far field condition was chosen. For the other measurements FABIAN was placed at around twice the critical distance, so that, taking into account the source's directivity, a fairly balanced direct-diffuse field ratio could be expected. In all cases the same two-way active speaker (Meyersound UPL-1) was used as sound source. All datasets were measured with a spatial resolution of 1° for horizontal, vertical and lateral head movements (cf. Table 4-2). Horizontal and vertical movements were measured together, resulting in a two-dimensional BRIR grid (160° x 70°). For mechanical reasons, lateral head movements were measured separately while keeping a constant frontal viewing direction, so that here a 1-dimensional data set (120°) was retrieved.

Table 4-2. Binaural datasets used in the study

Site	**Volume**	**RT**	**r_{crit}**	**Dist.**	**Dataset Ranges hor./ver./lat 1&2**
anechoic	1800 m^3	>60Hz	-	3 m	±80°/±35°/±60°*
studio	235 m^3	0.36 s	1.4 m	2.8 m	±80°/±35°/±60°*
hall 1	8600 m^3	2.1 s	3.6 m	7.5 m	±80°/±35°/0°
hall 2	3000 m^3	0.95 s	3.2 m	10 m	0°/0°/±60°*

* ±30° used for listening test in experiment II

The angular ranges for BRIR acquisition were chosen according to typical values observed for natural hearing [18] and physiologically motivated 'comfort' [19] and 'maximum' values [20].

Whereas a frontal sound source location was shown to be most critical for horizontal and vertical head movements [2], high thresholds could be expected for lateral head movements due to the absence of ILD and ITD differences when moving the head in this direction. Hence, additional data sets were collected for lateral head movements and a sound source directly above the listener. These were measured using a different sound source (Genelec 8030) due to its reduced weight. Moreover, measurements were conducted (a) with, and (b) without the HATS's torso, in order to examine its influence in more detail (not shown here).

4.3.2 Stimuli

All thresholds were determined using two stimuli: (a) pink noise of 5 seconds duration with 20 ms fade in and fade out, and (b) a 5 seconds excerpt from a piece for acoustical guitar (bourrée by J. S. Bach). The latter one had proven to be particularly critical for revealing artefacts in acoustic simulations [1]. Additionally, it was meant to serve as a natural musical stimulus, containing harmonic passages as well as transient components. Pink noise, as also used in [3], [10], [12], [14] was in contrast regarded as being particularly critical for revealing spectral differences induced by a reduced HRTF/BRIR resolution. Due to a bandpass used as compensation target for the headphone's equalization, all stimuli were bandlimited to 50 Hz–20 kHz.

4.3.3 Auralization

A Linux software package for fast partitioned convolution was used to auralize the BRIR sets with a sampling rate of 44.1 kHz. The software uses two partition sizes for the block convolution algorithm: a smaller one for the initial part of the BRIRs and a larger one for the diffuse tail. When head movements of the listener are detected, the initial 2^{14} samples of the BRIR will updated accordingly. Changes in the later diffuse reverberation tail due to different head orientations or source positions were shown to be inaudible [1]. The partition size of the initial BRIR part was set to 256 samples. For the diffuse tail a block size of 8192 samples was chosen. Updating the BRIR is done via parallel spectral domain convolution and time-domain crossfading. In order to avoid switching artefacts a short linear crossfade of 5.8 ms duration (according to the smaller block size) was used. So, the first results of a filter exchange are available one audio block after recognizing a trigger event; whereas the resulting latency of one audio block is already introduced by the underlying JACK audio server architecture (jackaudio.org, last visited at December 3rd, 2013). The time-domain cross fade results are then output blockwise and consecutively. The duration of a full crossfade is therefore as long as the dynamically interchanged early part of the BRIR. Hence, the direct sound is crossfaded with minimum latency, ensuring a minimum response time to head movements, while stretching out the fade process in time [21].

The crossfade time was chosen to avoid audible linear interpolation between adjacent BRIRs while still suppressing switching artefacts. Hence, our study is different from an evaluation of interpolation methods [6], [7] as well as from recent studies on the audibility of abrupt changes in (a) the interaural time delay, and (b) the minimum phase HRTF spectra related to different directions of incidence [23]–[25], respectively. Although these results provide valuable insight into the perception of fundamental localization cues, the (rather artificial) separate variation of ITD and HRTF magnitude was avoided here with respect to the external validity of results for virtual acoustic environments.

The inaudibility of the crossfading between BRIRs was proven by a pre-test on crossfading white noise and sine-tones between identical HRTFs. Since no switching was audible, it was concluded that all artefacts heard later should be due to differences in the BRIRs themselves.

A Polhemus Fastrack head tracker was used, providing an update rate of 120 Hz, i.e. new head positions every 8–9 ms. STAX SR202 Lambda headphones were used for reproduction. They were equalized with a linear phase inverse filter optimized by a least squares criterion [26] based on the magnitude average of ten measurements carried out while repositioning the headphones on the dummy head after each measurement. A recent evaluation of different compensation methods for headphone equalization in binaural synthesis is given in [27].

As the 3AFC listening test design (see below) requires instantaneous switching between data sets with full and reduced resolution, the complete set of BRIRs was held in random access memory (ca. 22 GByte in experiment I).

4.3.4 Subjects

The whole study was split into three successive experiments. In experiment I (horizontal and vertical head movements) 21 subjects (age 24–35, 19 male, 2 female) took part. In experiment II (lateral head movements, frontal sound incidence) 23 subjects (age 23–65, 20 male, 3 female) participated, while experiment III (lateral head movements, vertical sound incidence) was conducted with 20 listeners (age 24–40, 18 male, 2 female). All subjects were experienced in listening tests; most had musical education.

4.3.5 Psychophysical Procedure

An adaptive three alternative forced choice (3AFC) test procedure was chosen [28]. Three stimuli were presented, including the reference situation with 1° resolution twice and a version with reduced grid resolution once in random order for each trial. After an initial training phase (including feedback), each run started with a test stimulus in the middle of the range of provided grid resolutions. The BRIR resolution was then adaptively changed according to the subject's responses using a maximum likelihood adaption rule ("Best-PEST", [29]). The resulting test durations ranged from 35 minutes (exp. II & III) to 1.5 hours (exp. I). Head movements were not restricted, even if no dynamical exchange of BRIRs was provided outside the indicated ranges (Table 4-2). During the training phase, subjects were asked to find individual movement strategies that would maximize their detection rate.

4.3.6 Angular Ranges

BRIR data sets were auralized using the full angular ranges measured (cf. Table 4-2). Only for lateral head movements and frontal sound incidence (exp. II) pretests showed, that for large lateral tilting angles (>35°) noticeable comb filter modulations arise when the ear approaches the shoulder (cf. also Figure 4-4 and section 4.5). These modulations sometimes made even the 1° reference resolution detectable. Since this was not regarded to be a typical listening situation, auralization was limited to a range of ±30°.

Since in experiment I the threshold of just audible grid granularity was to be tested independently for horizontal and vertical grid resolution, the resolution was changed only for one direction. For the other direction the resolution was kept constant at maximum resolution (1°). The datasets used in experiments II and III contained only data for lateral head movements while retaining a frontal head orientation. Due to the 1° resolution of the BRIR sets, the smallest audible grid resolution threshold measurable with an adaptive forced choice procedure was 2°, a value that was reached only three times during the three experiments.

4.3.7 Experimental Design

In experiments I and II thresholds of just audible BRIR grid resolution were collected for horizontal, vertical and lateral head movements for frontal incidence of sound. Additional factors tested were stimulus (2) and rooms (3, including anechoic condition). Both experiments were conducted as full factorial, repeated measures designs, i.e. the thresholds of all subjects were measured under every possible condition. This lead to 3 x 2 x 2 x 21 = 252 threshold values in experiment I and to 2 x 3 x 23 = 138 values in experiment II. Experiment III was conducted with sound incidence from above, since this was assumed to be most critical with respect to lateral head movements. It was also conducted according to a repeated measures design, but while testing only one additional factor (stimulus = 2), thus 2 x 20 = 40 threshold values were obtained.

4.4 Results

Figure 4-2 shows the results of all subjects under all tested conditions in experiments I and II for the noise (above) and the guitar stimulus (below). Since the thresholds were not always normally distributed (conservative Kolmogoroff-Smirnoff test, type I error level = 0.2) medians resp. percentiles were used to indicate central tendency and dispersion. Figure 4-3 shows the lateral threshold values from experiment III (sound incidence from above, only anechoic environment, labelled 'lateral 2') in comparison to the thresholds from experiments I and II under anechoic conditions.

Medians of just audible grid granularity ranged from 4° to 18°, depending on condition. Only three times grid resolutions smaller than 3° could be reliably detected by individual subjects: twice for anechoic environment, vertical head movement and noise stimulus (exp. I), and once for anechoic environment, lateral head movement and noise stimulus (exp. III). For the noise stimulus and all acoustical environments pooled, subjects were almost equally sensitive to a reduced grid resolution in horizontal and vertical direction (medians: 6° and 5°, resp.). For noise, all subjects could reliably detect grid granularities of above 11°, whereas these were much more difficult to detect with the guitar stimulus.

For frontal sound incidence and lateral head movements (exp. II), a reduced grid resolution was much more difficult to detect, and several subjects could not even distinguish the largest granularity from the smallest one, even in direct comparison (30° vs. 1°, cf. Figure 4-2).

A statistical analysis of the trends observable in Figure 4-2 was conducted by means of a 3 x 2 x 2 (exp. I), or a 3 x 2 (exp. II) one-way ANOVA for repeated measures, respectively. Data assumptions (homogeneous error variances, homogeneous correlation under all conditions, [29]) were tested using Mauchly's test of sphericity. Degrees of freedom were corrected if indicated by the test. Reliability of response behavior was very high for experiment I (horizontal and vertical head movements) with Cronbach's α at 0.934. For lateral head movements (exp. II), Cronbach's α was only 0.165, indicating large interindividual differences.

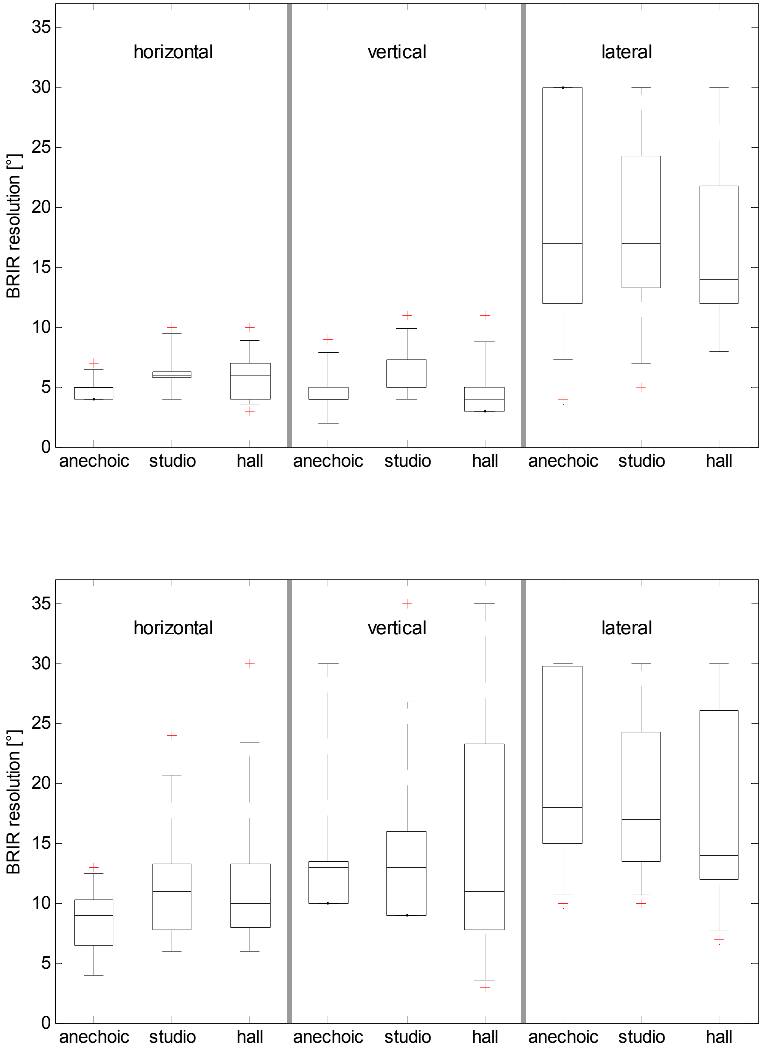


Figure 4-2. The plots displays the just noticeable discretization of BRIR data sets for frontal sound incidence, three rotational degrees of freedom, and three different acoustic environments. Upper plot: noise, lower plot: guitar stimulus. Boxplots indicate medians and interquartile range, whiskers show 90% quantiles; crosses indicate all outliers.

For experiment I a strong interaction between the factors 'direction of movement' and 'stimulus' was observed. When listening to noise, subjects were more sensitive to a reduced grid resolution for vertical head movements (cf. Figure 4-2). The broadband noise signal obviously provided enough spectral information so that two listeners could even reliably detect a grid granularity of 2° in vertical direction. A 2 x 3 ANOVA conducted separately for each stimulus confirmed that vertical thresholds were significantly lower than horizontal thresholds (means: 5° vs. 5.6°) in the noise condition.

When listening to the guitar stimulus, however, listeners were much more sensitive to a reduced grid resolution for horizontal head movements. Here, discretization effects for modulated ITDs and ILDs obviously presented a stronger cue than spectral differences for a narrowband, musical signal.

As expected, a reduced grid resolution for lateral head movements and frontal sound incidence was detected only at very high thresholds for both noise and guitar, due to a lack of binaural cues.

When pooling over all degrees of freedom and all acoustical environments, the effect of 'stimulus' was highly significant (5° vs. 12.3°). In agreement with studies on the MAA the bandwidth of stimuli (here: noise vs. guitar) had a very strong effect in this spatial discrimination task. Only for lateral head movements the stimulus effect is negligible, since the general uncertainty of subjects was reduced only a little by the higher bandwidth of the noise stimulus.

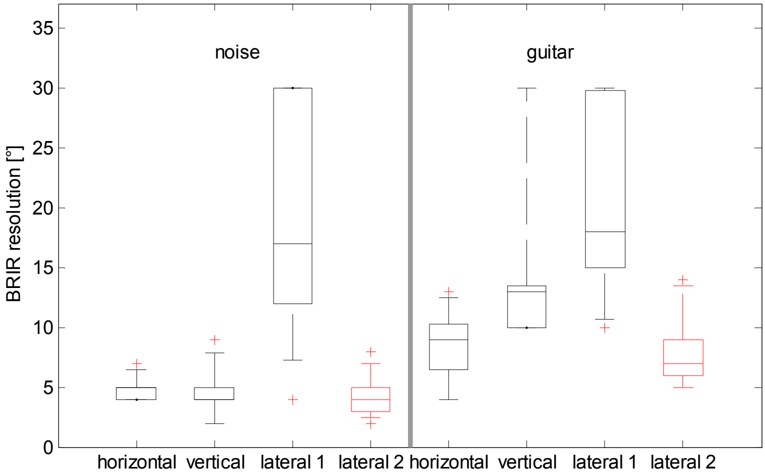


Figure 4-3. The plot displays the just noticeable discretization of BRIR data sets for 3 rotational degrees of freedom using 2 different stimuli in anechoic environment; lateral 1 = frontal sound incidence, lateral 2 = vertical sound incidence.

The factor 'room acoustics' showed no significant influence on the perceptibility of grid resolution; only a trend can be observed. For horizontal and vertical movements a slightly higher sensitivity was found for the anechoic presentations (~1°). On contrast, reduced lateral resolution with frontal sound incidence (exp. II) was easier to detect with increasing room response. Here, room reflections, particularly those from floor and ceiling, obviously provided additional cues compared to the anechoic condition, as was confirmed by the investigation for sound incidence from above (exp. III).

For lateral head movements and vertical sound incidence (exp. III), not surprisingly, thresholds were much lower than those found for frontal sound incidence (cf. Figure 4-3). Again, the noise stimulus lead to significantly lower threshold values than the guitar stimulus (means: 4.25° vs. 7.25°, paired t-test).

The thresholds for horizontal and vertical head movements with frontal source and lateral head movements with sound source above can be regarded as the (presumably) most critical configurations of source, receiver, and direction of head movement. For pink noise as a stimulus, those values are surprisingly similar, with means of 4.8° for horizontal, 4.7° for vertical, and 4.2° for lateral head movements. Although these differences were not significant (independent-samples-ANOVA with conservative post-hoc tests), subjects were, at least by trend, more sensitive to a reduced lateral resolution than to horizontal or vertical resolutions.

4.5 Discussion

Since we can assume that spectral differences induced by head movements and discrete BRIR data are an important cue for detecting discontinuities in VAEs, these differences have been plotted in Figure 4-4 for every direction of head movement investigated and for anechoic conditions. Magnitude spectra have been normalized towards their smoothed mean spectral magnitude, in order to make only direction-related differences visible.

When we look at modulations induced by vertical head movements, at first sight only minor differences are visible compared to horizontal head movements. However, most spectral variance happens in a small angular region (±10°) close to a neutral (0°) head direction. In the absence of ITD and ILD differences, this variance, caused by different orientations of the pinnae towards the sound source, is obviously high enough to let listeners detect even smaller discretization than for horizontal movements.

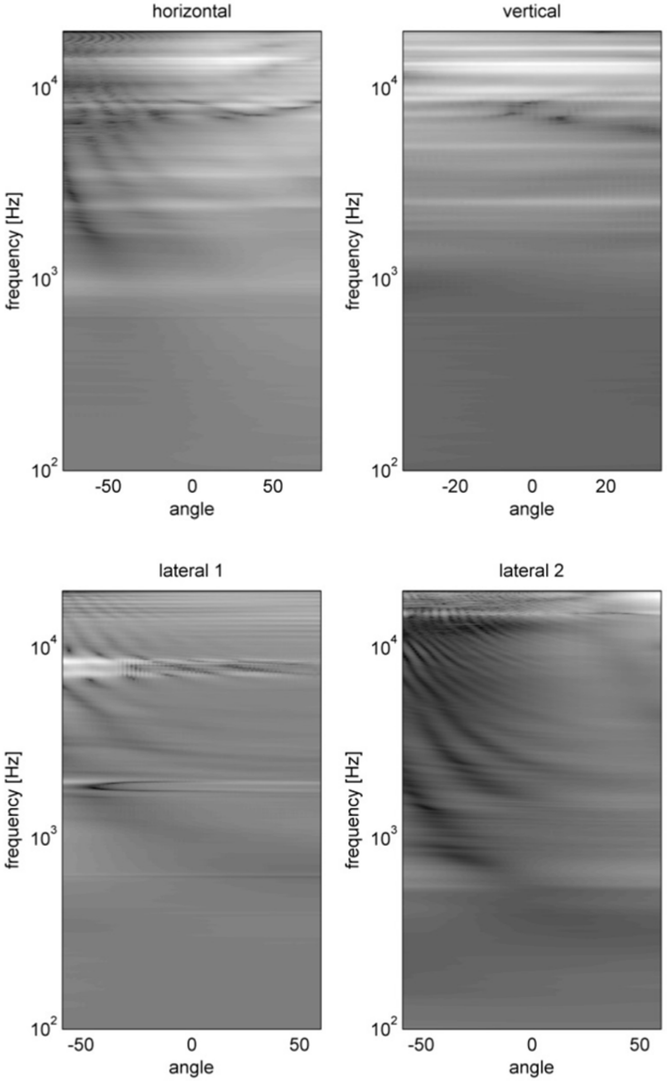


Figure 4-4. Direction dependent magnitude spectra for HRTF data sets acquired in anechoic condition. Each plot has been normalized for its average magnitude spectrum. Above: horizontal and vertical head orientation for a frontal sound source, below: lateral head orientations for a frontal sound source (left) and a sound source above (right). Left ear is shown; negative angles indicate left ward movement, i.e. a downward inclination of the ear.

The magnitude spectra for lateral head movements and frontal sound source ('lateral 1') are largely independent of head orientation. Yet, with decreasing distance to the shoulder (negative angles), a comb filter characteristic is visible, which is most probably due to shoulder and torso reflections, whose path lengths decrease with the head inclined towards the shoulder. This explains why discontinuities due to a reduced BRIR resolution are easier to detect in this area, and indeed many subjects used large lateral tilts in the stipulated discrimination task.

The same comb filter can be observed for horizontal movements and for lateral movements with vertical sound incidence ('lateral 2'). While the modulations of ITDs can be assumed to be very similar for both conditions, the plots indicate a stronger acoustical shadowing for frequencies above 1 kHz (and thus: larger ILDs) and a stronger direction dependent comb filtering in lateral direction, starting already close to the neutral (0°) head direction. This might explain why slightly lower threshold values have been found for lateral than for horizontal head movements.

Since comb filter modulations shown in Figure 4-4 are much less pronounced in measurements without shoulder and torso (not shown here), it can be expected that torso reflections play an important role in spatial discrimination tasks as well as for the authenticity of virtual acoustic environments in general. For a closer examination of this aspect it will be interesting to compare BRIR data measured with and without out FABIAN's torso.

4.6 Conclusion

Thresholds for the minimum audible grid granularity in virtual acoustic environments based on dynamic binaural synthesis have been determined for the three rotational degrees of freedom of head movements. Listening tests revealed thresholds for different acoustic environments and for different audio contents using an adaptive 3AFC psychoacoustic procedure. It could be shown, that, depending on source position, the ability to detect a reduced spatial resolution in discretely measured BRIR data is very similar for all directions of head movements. When listening to a broadband noise stimulus, a reduced grid granularity in vertical and lateral direction was even more critical than for horizontal head movements, a result also supported by [25]. The results of the listening tests are consistent with observations on the magnitude spectra of the HRTFs used, which exhibit higher spectral variance in lateral than in horizontal direction. For musical content with limited and non-stationary bandwidth a reduced spatial resolution of BRIRs was less critical.

The observed thresholds showed only very little variation with the size and reverberation time of the measured acoustic environments. Hence, even natural, partly diffuse spatial environments cannot be synthesized with binaural data of lower resolution than in the anechoic condition.

Since the conditions for a parametric statistical analysis were not always complied with (cf. section 4.4), Table 4-3 summarizes our results as percentiles of the observed distributions of thresholds within our sample of 20–23 subjects for each of the three experiments.

Table 4-3. Grid resolutions just audible for different percentiles of the sample of subjects for horizontal, vertical and lateral (source in front/source above) head movements.

…was audible for	**Noise**	**Guitar**
	hor/ver/lat1/lat2	hor/ver/lat1/lat2
50%	6° x 5° x 16° x 4°	9° x 12° x 16° x 7°
25%	4° x 4° x 12° x 3°	7° x 9° x 12° x 6°
5%	4° x 3° x 8° x 2°	5° x 4° x 8 x 5°
0%	2° x 1° x 3° x 1°	3° x 2° x 3° x 4°

Given the values in Table 4-3, a BRIR grid granularity of 2° for horizontal, 1° for vertical and 1° for lateral head movements should provide a spatial resolution for virtual acoustic environments that is sufficient even for critical listeners, critical audio content, and all possible sound source locations. Further studies on the application of different interpolation algorithms for binaural synthesis can use these granularities as a target for interpolated data sets. For musical content presented in spaces with some reverberation a resolution of 5° for horizontal, 4° for vertical and 5° for lateral head movements will be sufficient to create a plausible simulation for 95% of the listeners.

4.7 Acknowledgements

This work was supported by a grant from the Deutsche Telekom Laboratories and the Deutsche Forschungsgemeinschaft (Grant WE4057/1-1).

4.8 References

[1] Lindau, T. Hohn, S. Weinzierl (2007): "Binaural resynthesis for comparative studies of acoustical environments.", in: *Proc. of the 122nd AES Convention*, Vienna, preprint no. 7032

[2] Mills, A. W. (1958): "On the Minimum Audible Angle", in: *J. Acoust. Soc. Am.*, **30**(4), pp. 237-246,

[3] Silzle, A.; Novo, P.; Strauss, H. (2004): "IKA-SIM: A System to Generate Auditory Virtual Environments", in: *Proc. of the 116th AES Convention,* Berlin, preprint no. 6016

[4] Geier, M.; Ahrens, J., Spors, S. (2008): "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods", in: *Proc. of the 124th AES Convention,* Amsterdam, preprint no. 7330

[5] Mackensen, P. (2004): *Auditive Localization. Head Movements, an additional cue in Localization*, Doct. diss., Technical University of Berlin

[6] Hartung, K.; Braasch, J.; Sterbing, S. J. (1999): "Comparison of different methods for the interpolation of head-related transfer functions", in: *Proc. of the AES 16th International Conference*, Rovaniemi, Finland

[7] Minnaar, P.; Plogsties, J.; Christensen, F. (2005): "Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis", in: J. Audio Eng. Soc., 53(10), pp. 919-929

[8] Gardner, W. G.; Martin, K. (1995): "HRTF measurements of a KEMAR", in: *J. Acoust. Soc. Am.,* **97**(6), pp. 3907-3908

[9] Personal communication with S. Feistel (ADA)

[10] Algazi, V. R. et al. (2001): "The CIPIC HRTF Database", in: *Proc. of the IEEE-WASPAA Workshop*, New Paltz

[11] Personal communication w. Claus L. Christensen (ODEON)

[12] Lentz, T. et al. (2007): "Virtual Reality System with Integrated Sound Field Simulation and Reproduction", in: *EURASIP J. of Advances in Signal Processing* (Article ID 70540)

[13] Personal communication with Frank Wefers (ITA Aachen)

[14] Savioja, L. et al. (1999): "Creating interactive virtual acoustic environments", in: *J. Audio Eng. Soc.*, **47**(9), pp. 675-705

[15] Begault, D. et al. (2006): "Design and Verification of HeadZap, a Semiautomated HRIR Measurement System", in: *Proc. of the 120th AES Convention*, Paris, preprint no. 6655

[16] Wenzel, E. M. et al. (2000): "Sound lab: A real-time, software-based system for the study of spatial hearing.", in: *Proc. of the 108th AES Convention,* Paris: 2000, preprint no. 4150

[17] Personal communication with Jens Ahrens (TLabs Berlin)

[18] Thurlow, W. R.; Mangels, J. W.; Runge, P. S. (1967): "Head movements during sound localization", in: *J. Acoust. Soc. Am.*, 42(2), pp. 489-493

[19] DIN 33408-1 (1987): *Körperumrißschablonen für Sitzplätze*, Berlin: Beuth (German national standard)

[20] Morgan, C. T. et al. (1963): *Human Engineering Guide to Equipment Design*, New York: McGraw-Hill

[21] Müller-Tomfelde, A. (2001): "Time varying Filters in non-uniform Block Convolution", in: *Proc. of the COST G-6 Conference on Digital Audio Effects,* Limerick

[22] Hoffmann, P. F.; Møller, H. (2005): "Audibility of Time Switching in Dynamic Binaural Synthesis", in: *Proc. of the 118th AES Convention*, Barcelona: preprint no. 6326

[23] Hoffmann, P. F.; Møller, H. (2005): "Audibility of Spectral Switching in Head-Related Transfer Functions", in: *Proc. of the 119th AES Convention*, New York, preprint no. 6537

[24] Hoffmann, P. F.; Møller, H. (2006): "Audibility of Spectral Differences in Head-Related Transfer Functions", in: *Proc. of the 120th AES Convention*, Paris, preprint no. 6652

[25] Hoffmann, P. F.; Møller, H. (2008) : "Some Observations on Sensitivity to HRTF Magnitude", in: *J. Audio Eng. Soc.,* **56**(11), pp. 972-982

[26] Kirkeby, O.; Nelson, P. A. (1999): "Digital Filter Design for Inversion Problems in Sound Reproduction", in: *J. Audio Eng. Soc.*, **47**(7/8), pp. 583-595

[27] Schärer, Z.; Lindau, A. (2009): "Evaluation of Equalization Methods for Binaural Signals", in: *Proc. of the 126th AES Convention*, Munich, preprint no. 7721

[28] Leek, M. R. (2001): "Adaptive procedures in psychophysical research", in: *Perception & Psychophysics*, **63**(8), pp. 1279-1292

[29] Pentland, A (1980): “Maximum likelihood estimation: The best PEST”, in: *Perception & Psychophysics*, **28**(4), pp. 377-379

[30] Bortz, J (2005).: *Statistik für Sozial- und Humanwissenschaftler,* 6th ed., Berlin: Springer

Part II
Binaural Reproduction

5 Binaural Resynthesis for Comparative Studies of Acoustical Environments

The following chapter is an authorized reprint of the précis-reviewed article

> Lindau, Alexander; Hohn, Torben; Weinzierl, Stefan (2007): "Binaural Resynthesis for Comparative Studies of Acoustical Environments", in: *Proc. of the 122nd AES Convention*. Vienna, preprint no. 7032.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, use of US-American spelling, typographic and stylistic corrections.

5.1 Abstract

A framework for comparative studies of binaurally resynthesized acoustical environments is presented. It consists of a software-controlled, automated head and torso simulator with multiple degrees of freedom, an integrated measurement device for the acquisition of binaural impulse responses in high spatial resolution, a head-tracked real-time convolution software capable to render multiple acoustic scenes at a time, and a user interface to conduct listening tests according to different test designs. Methods to optimize the measurement process are discussed, as well as different approaches to data reduction. Results of a perceptual evaluation of the system are shown, where acoustical reality and binaural resynthesis of an acoustic scene were confronted in direct A/B comparison. The framework permits, for the first time, to study the perception of a listener instantaneously relocated to different binaurally rendered acoustical scenes.

5.2 Introduction

Auditory perception is often studied on the basis of comparative tests where listeners are asked to assess the difference between different stimuli. Such comparative setups are required in almost every domain of audio communication, from system evaluation, industrial sound design, planning of room acoustics and sound reinforcement systems up to basic research in psychoacoustics or empirical studies of media reception.

Whenever test listeners are supposed to be sensitive to small perceptual differences, instantaneous switching to the new stimulus is required. That is why field studies, e.g., when comparing different room acoustical environments, are largely inappropriate for methodical reasons. Hence, only the synthesis of acoustical environments by means of 3D audio technologies can allow for a direct confrontation of otherwise incompatible acoustical scenes. Since loudspeaker-based approaches such as Wave Field Synthesis (WFS) or Higher Order Ambisonics (HOA) require large-scale, hardly mobile installations, are affected by specific physical artefacts, and are still unable to (re)synthesize 3D sound fields with all their spatial properties, an approach based on binaural technology seems more appropriate, all the more when synthesis is not aimed at a larger audience but for single listeners required in test situations. With the practical realization of a framework as suggested in Figure 5-1, the comparative study of otherwise incompatible situations will become accessible for the first time.

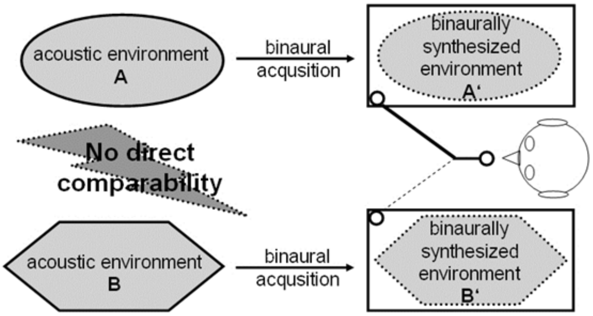


Figure 5-1. Comparison of acoustical environments via binaural resynthesis.

5.3 Methods and Tools

All binaural approaches towards (re)synthesizing a sound field are based on the assumption that nearly all perceptual information used for orientation in our auditory environment is coded in the sound pressure at our eardrums. Hence, natural or artificial sound fields can be simulated on the basis of binaural room impulse responses (BRIRs), representing the acoustical transmission path from a sound source to the listener [1]. When convolving the measured BRIRs with an anechoic source signal the ear signals desired can be (re)created. When all system parameters such as the spatial resolution of the BRIR set, the spectral compensation of all involved transducers [2], [3], and the latency of the dynamic auralization accounting for movements of the listener's head [4]–[6] are carefully controlled, an auditory event can be created that is – even in direct comparison – largely indistinguishable from natural sound fields [7].

In the following we present a tool for comparative studies of acoustical environments based on binaural simulations, including

- a new head and torso simulator (HATS) for software-controlled, automatic measurement of complete sets of binaural impulse responses (BRIRs) in multiple degrees of freedom
- a standardized post-processing of BRIRs
- a fast convolution algorithm allowing instantaneous switching between different BRIR data for multiple source locations
- a user interface for software-controlled listening tests based on different test design options

5.3.1 Acquisition

Several head and torso simulators for the acquisition of binaural impulse responses are available, such as the KEMAR-manikin [8], the Head Acoustics HMS-II/III series, the Bruel & Kjaer 4100 and the former Cortex (now Metravib) Mk1 as well as systems in the academic area such as the VALDEMAR-HATS [9] or the Aachen-head [10]. They are static systems with only limited possibilities to emulate movements of head, shoulders, and torso except by manual step-by-step reorientation [4]. A HATS-system developed at the TU Berlin in 2002 was the first to allow for an automated measurement of BRIRs for a (typical) horizontal rotation range of ±75° [11], [12]. A similar system was presented recently for binaural assessment of car sound, consisting of a modified B&K 4100 head [13].

The encouraging results of a perceptive evaluation of the TU System [12] gave rise to the subsequent development towards a more flexible and universal measurement system for the Fast and Automatic Binaural Impulse response AcquisitioN (FABIAN, cf. Figure 5-2, and [14], [15]).

It now allows for the precise, software-controlled horizontal and vertical orientation of an artificial head while measuring BRIRs using a multichannel impulse response measurement system with swept sine technique and noncyclic IR-deconvolution [16]. The whole device is mounted on a rotatable turntable that allows for the orientations of torso and head to be adjusted independently. Audio quality relating to parameters such as signal level, duration (FFT block size) and spectral coloration of the stimulus as well as the number of averages can be chosen to adapt the measurement to ambient conditions such as noise level or reverberation time. If input clipping is detected or if the signal-to-noise ratio (SNR) temporarily falls below a given threshold, a repetition of the last measurement is initiated. The simultaneous measurement of complete sets of BRIRs for up to 8 sources in high spatial resolution is controlled by a custom Matlab® application and thus considerably accelerated compared to conventional systems.

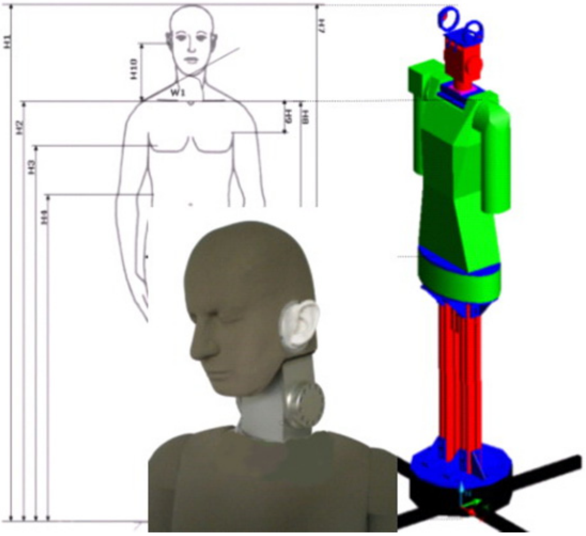


Figure 5-2: Design stages of FABIAN; close up of prototype with automatable neck joint.

5.3.2 Post-processing of BRIRs

The post-processing of BRIRs so far includes the compensation of the headphones' and microphone's transfer function, an amplitude normalization, and a shortening of BRIR lengths.

5.3.2.1 Headphone Compensation

Although a multiplication of the BRIR spectra with the inverted complex transfer function (direct FFT deconvolution) of the mic-headphone-path seems an intuitive approach, two problems arise. First, the system's transfer function is usually not minimum phase, therefore a stable and causal compensation filter may not exist [17]. Second, the highly nonlinear frequency response of headphones shows several narrow peaks up to +12 dB and dips with almost total extinction. Since their center frequencies may shift due to slight changes in the positioning of the headphone, a “perfect” equalization can lead to strong peaks with audible ringing artifacts.

This problem persists when the minimum and maximum phase part of the impulse response are inverted separately (homomorphic technique) [18], although an additional modeling delay applied to the maximum phase part provides an approximation of it’s a-causal inverse, with filters as perfect as those derived from FFT-deconvolution and much smaller filter lengths.

Hence, another strategy was chosen. It is computationally simpler and superior in terms of remaining error energy [18]. The inverse filter is designed in the time domain with a least mean square (LMS) error criterion.

It can be shown [19] that the optimal filter, minimizing the LMS error of a non-perfect compensation and the “effort”, i.e. the energy of the compensation filter, is given by

$$h = (C^T C + \beta B^T B)^{-1} C^T \delta. \tag{5-1}$$

Here, C is the filter to be compensated, β controls the absolute amount of regularization (i.e., non-perfect compensation) used and B is an FIR filter whose frequency response introduces a frequency weighting into the inversion process. Expression (5-1) uses matrix notation where C and B are convolution matrices of the respective filters, h and δ are signal vectors, β is scalar. Regularization will be most effective in the passband of the filter which can be chosen without constraints on phase as only its energy will be effective [20]. To account for non-minimum phase components in the mic-headphone transfer function C a modelling delay is used. Therefore, the response of h is not designed towards reaching an ideal Dirac pulse, but a delayed one, namely $\delta(n-k)$ or δ_k. So the final expression used to design the inverse headphone filter was [eqn.(5-2)]

$$h = (C^T C + \beta B^T B)^{-1} C^T \delta_k. \tag{5-2}$$

The lower cut-off frequency of the compensation is defined by the FIR-filter length, in our case 1024 samples. As the compensation result depends not very strongly on absolute delay position, as long as it is not too close to the filters boundaries [18], the modelling delay was simply set to N/2 = 512 samples.

Since the optimal regularization strongly depends on the impulse response to be compensated, the parameters β and B can only be found by comparative listening [21]. In Figure 5-3 we have set $\beta = 0.8$ and $b(n)$ a highpass filter with -10 dB in the stopband and a long transition range of 2–8 kHz. If regularization on the LMS-error criterion is applied, the compensation of deep notches is affected first (cf. Figure 5-3), whereas third- or sixth-octave band smoothing often used in FFT-deconvolution is less effective in this respect. The quality of the compensation reached was convincing even with a relatively short filter length of 1024.

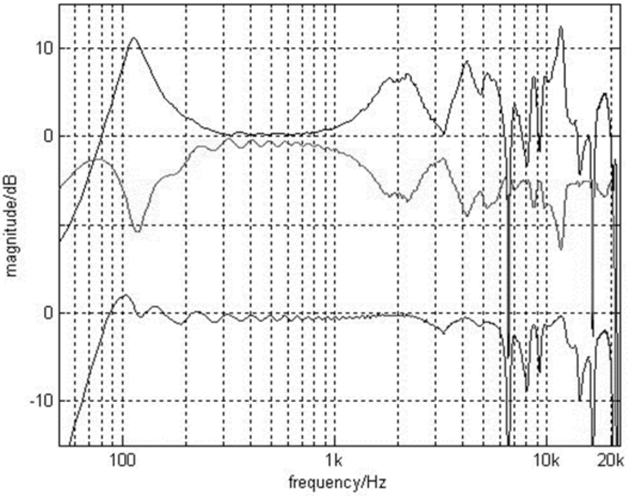


Figure 5-3. Frequency response of STAX SR 202, frequency response of inverse LMS filter (N=1024) with frequency-dependent regularization, calculated frequency response after compensation (from upper to lower).

5.3.2.2 Length of BRIRs

A single set of BRIRs with 1°/5° horizontal/vertical resolution in a ±75°/±45° range and impulse responses of 2 s length sampled at 44.1 kHz in 32 bit floating point format creates a database of 2869 BRIRs allocating ca. 2 GByte of data. Thus, data reduction strategies are crucial, as soon as several sets of BRIRs have to be held in random access memory of the convolution algorithm.

Here, several studies showed that a dynamic, head-tracked auralization is only necessary for the initial part of the BRIR, while the diffuse reverberation tail can be taken from a single direction without perceptual damage. For a small room with short reverberation time (V = 185 m^3, $RT30_{\mathrm{1kHz}}$ = 0.72 s) Meesawat and Hammershøi [22] found an offset of ca. 50 ms for the beginning of a diffuse reverberation tail. Since this offset is expected to increase with the mean free path of the room a similar listening test was repeated for BRIRs measured in a larger hall (V = 10.000 m^3, $RT30_{\mathrm{1kHz}}$ = 2 s). A 3AFC test was done, using a short anechoic speech and drum-set sample as test material. The BRIR sets were measured at a central seat position (third row) for a source placed on stage ca. 8 m away, which is roughly the critical distance of the room including the directivity of the source used. Listeners were asked to detect differences between the original 0°/0°-BRIR and a concatenated BRIR using the diffuse part measured with the head turned left by 75°. An energy-preserving cosine window of 20 ms was used for the crossfade. Whenever listeners correctly identified the concatenated IR, the offset was shifted in steps of 10 ms. After two runs with anechoic speech and a drum-set sample of ca. 5 s duration (Run A, Run B) a third run with the diffuse tail from a source located 20 m away at 130°/30° (worst-case situation in [22]) followed (Run C). Here, only the drum-set sample was used.

Results for 23 subjects are shown in Figure 5-4. Median values for all subjects were 36 and 33 ms for sources with equal distance to the listener (run A/B) and 50 ms for sources with different distance to the listener. These offsets are close to the values determined in [22].

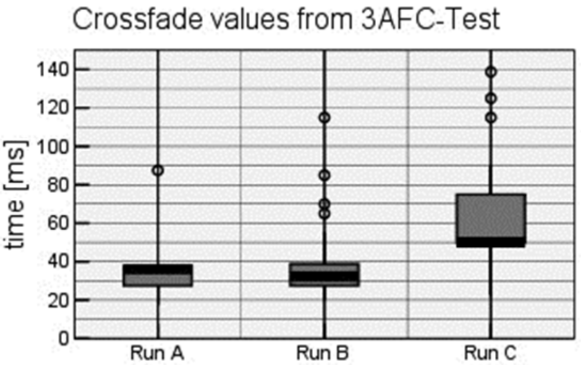


Figure 5-4. Time offset inducing a just noticeable difference when concatenating the early part of a 0°/0°-BRIR with the diffuse tail of (a) the 75°/0°-BRIR of the same source (Run A: speech sample, Run B: drum set sample), and (b) when concatenating the early part of the same 0°/0°-BRIR with the diffuse tail of a sound source located at 130°/30° and located a larger distance (Run C: drum set sample). Boxplots show median, interquartile range, and outliers.

At the same time, there were subjects reliably detecting differences at offsets as late as 140 ms. Obviously, the effect of training (in our test introduced through a slow approach towards the JND) and “finding the right cue” plays an important role, as was reported by listeners reaching significantly different thresholds for different source signals.

Since the system is supposed to work for sensitive and trained listeners also, we use a block size of 16 384 samples (≈ 370 ms) as dynamically refreshed direct part of the auralization. This not only reduces the size of the BRIR database, but is also used to increase the effective SNR of the simulation, because the BRIR used for the reverberation tail can be measured with a higher SNR resulting from several averaged measurements.

5.3.3 Real-time Rendering

The real-time rendering application, running on Linux OS, is implemented as a JACK Audio server client [23]. It is able to render multiple sources at a time depending on computation power and RAM size. It uses fast non-uniform partitioned block convolution and a double static/dynamic caching algorithm to account for latency and memory limitations.

5.3.3.1 Implementation Details

The application is implemented in C++ and optimized for the x86 platform. Optimizing for PPC is quite easy though. Head movements of the listener as reported from the head tracking device are translated into OSC commands [24] by a separate application. For easy OSC integration the *liblo* open source library [25] was used.

The application manages a cache of impulse responses (Figure 5-5). While the maximum number of impulse responses stored in the cache can be configured, a separate cache manager thread watches the current head tracker position and loads impulse responses around the current head position (dynamic cache). As soon as the maximum number of responses is reached, it frees memory associated to the impulse response with maximum distance to the current position. A basic grid of impulse responses (static cache), loaded during program start cannot be unloaded. The cache management allows transitions between different acoustical scenes even if the complete amount of BRIR data is too large to be held in random access memory.

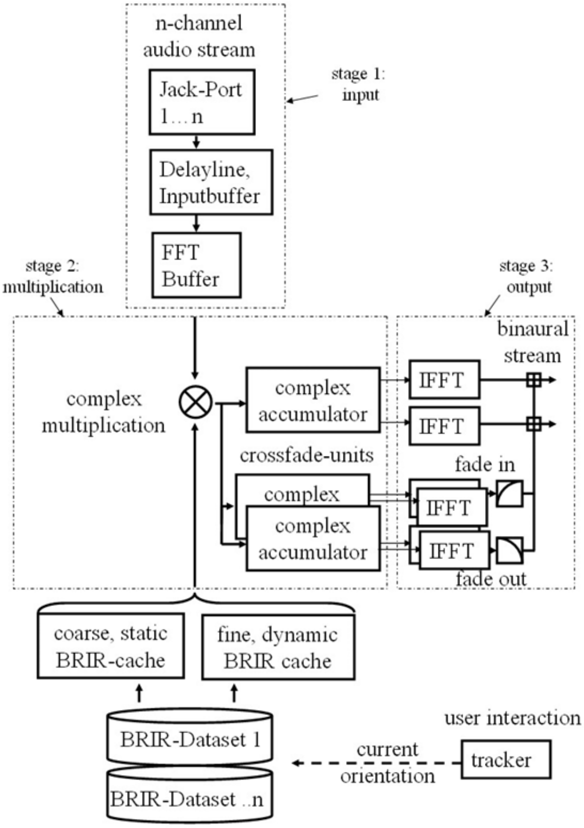


Figure 5-5. Simplified block diagram of rendering application.

To eliminate click artifacts, the exchange of impulse responses due to head movements is crossfaded in the time domain after convolution. The architecture allows doing the mix-down of the different sources as part of the complex multiplication in frequency domain, saving memory as well as memory bandwidth. The software currently uses the fftw3 open source library [26] for the FFT, while the complex multiplication is implemented in a vectorized form with SSE instructions.

5.3.3.2 Complexity

The algorithm can be divided into the input stages (mainly FFT), the complex multiplications, and the output stages (IFFT, crossfades, and mixing). The computational complexity for each of the stages depends on the number of source channels L, the partition size N_{part}, the length of the impulse response N_{BRIR} and the sampling rate f_s. Computational complexity for each stage of the algorithm is given by

$$output = O\left(fs \cdot log\ N_{part}\right) \tag{5-3}$$

$$input = O\left(fs \cdot L \cdot log\ N_{part}\right) \tag{5-4}$$

$$cplxmul = O\left(fs \cdot L \cdot \frac{N_{BRIR}}{N_{part}}\right) \tag{5-5}$$

For common values of $N_{BRIR} = 10^4 \ldots 10^5$ for reverberation times of 0.5 ... 2.5 s and $N_{part} = 10^3 \ldots 10^4$ (see below) the computational costs are dominated by the complex multiplication, increasing proportionally to the sampling rate and the number of source channels while decreasing inversely proportional to the partition size. On an Intel CoreDuo 2GHz Thinkpad with 2 GByte of RAM we could render up to 6 sources in 1°/5° (hor./ver.) resolution with impulse responses of length 2^{17} (3 s) and a block size of 256 samples.

5.3.3.3 Latency

To minimize the latency towards head movements, the minimum partition size of the convolution equals the audio processing block size, as set from inside JACK. The convolution implemented is overlap-add, so the latency of the source signal is one audio block. This latency is already imposed by the JACK audio server system. Head tracking is realized via a Polhemus Fastrack with 120 Hz update rate when using one single sensor.

According to the following tabular overview with worst case latencies at 44.1 kHz sampling rate a reduction of block size below 128 samples would be largely ineffective considering the constant latencies of head tracker and tracker data interface.

Table 5-1. Audio output latencies as calculated for different block sizes.

Block Size (samples)	**128**	**256**	**512**
JACK latency (ms)	2.9	5.8	11.6
Serial port delay (ms)	4.0	4.0	4.0
Head tracker update rate (ms)	8.3	8.3	8.3
Sum (ms)	**15.2**	**18.1**	**23.9**

All sources are rendered continuously. For instantaneous switching between multiple sets of BRIRs, OSC commands are sent from the graphical user interface of the listening test software switching audio outputs on and off (Figure 5-6).

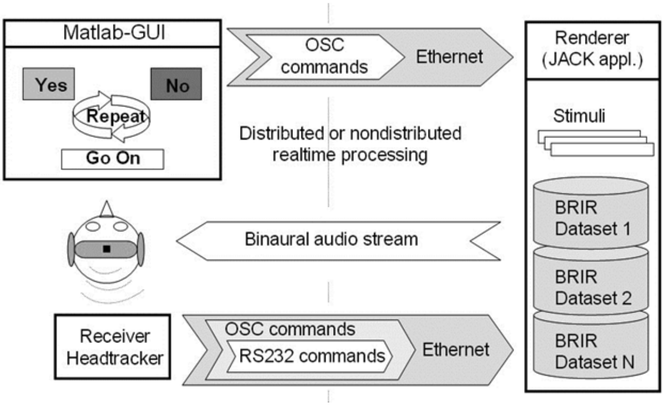


Figure 5-6. Flow diagram of control and audio data in the rendering process.

5.4 Evaluation

To evaluate the plausibility of the binaural synthesis which is a prerequisite for the validity of comparative studies as outlined in Figure 5-1, a direct AB comparison of a natural and a resynthesized acoustic environment was done.

5.4.1 Listening Test Design

As a challenging setup, both in computational and in acoustical respect, a concert-hall-like environment was chosen with the auditorium maximum of the TU Berlin (V = 10.000 m³, $RT30_{1\ \mathrm{kHz}}$ = 2 s). For two source positions and the measurement dummy seated in the third row, ca. 8 m from the sources, a measurement was conducted with 1°/5° horizontal/vertical resolution ranging from ±75°/±30°, resulting in a database of 3926 BRIRs and lasting about 23 hours. Using a bass-emphasized, constant envelope log-sweep of FFT-order 18 with three averages, the resulting

BRIRs were saved with a length of 3.5 s (Figure 5-7). The overall SNR reached was about 95 dB, with a frequency dependence shown in Figure 5-8.

The listening test was designed as a simple forced choice AB comparison test, where each subject listened to 80 randomized pairs of stimuli. Each pair consisted of a natural presentation (loudspeaker) and a binaurally simulated presentation of the same signal.

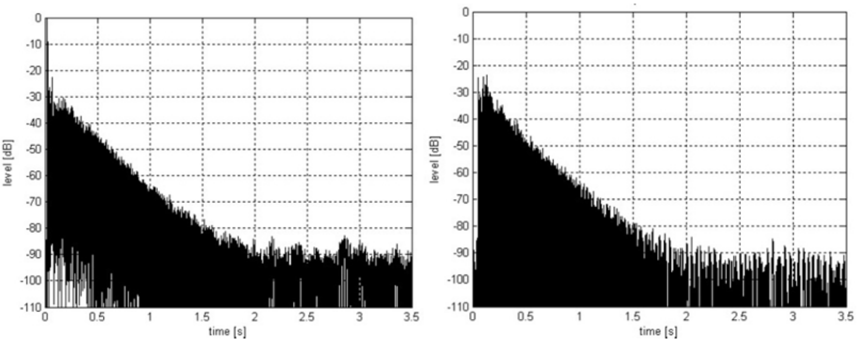


Figure 5-7. Energy-time-curves of two measured left ear BRIRs with the head pointing in 0° direction (left: frontal source, right: rear source). Note the different direct sound levels but similar levels of noise floor.

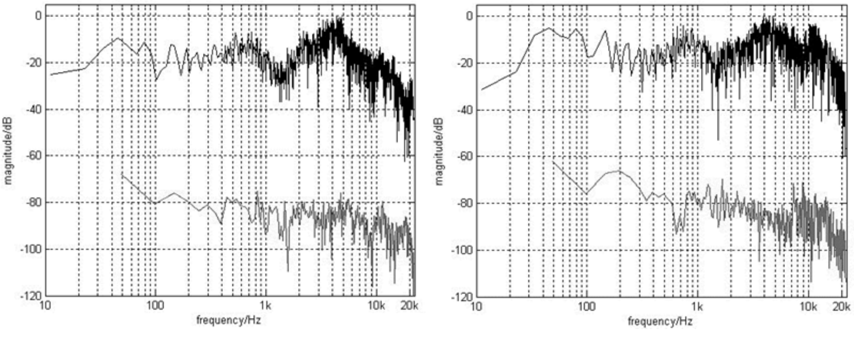


Figure 5-8. Normalized binaural room transfer functions and noise spectra, ipsilateral ear (left) and contralateral ear (right) for a frontal source in 8 m distance, with the head turned 75° left.

After a training period each pair of stimuli was presented once, and the listener was asked to identify which stimulus was the simulation. The duration of each stimulus did not exceed 6 seconds. Acoustically specifically transparent headphones (Stax SR 202) remained attached during the whole test. Accordingly, BRIRs measured with the headphones attached were used to account for the remaining shadowing effect of the external sound field. Some conditions were varied as independent variables for statistical analysis. These included (a) the degrees of freedom in head movement exploited by the rendering engine (horizontal vs. both horizontal and vertical), (b) different anechoic stimuli (male speech, female speech, acoustic guitar, trumpet, and a drum set, all taken from the EBU SQAM CD and the Archimedes collection) and (c) the source position (hor./ver.: 0°/0°, and 130°/30°).

35 subjects took part in this listening test, 12 female and 23 male. Attending a lecture about virtual acoustics and familiar with potential flaws of the technology they could be considered as “expert listeners”.

5.4.2 Results

The overall detection rate including all 2800 decisions (80 pairs from 35 subjects each) was 52.9%. This is a small but – tested on the basis of a Chi²-distribution with 2800 samples – still a statistically significant difference from the guessing rate of 50%. If we look at the detection rate per subject, we find a rather symmetrical distribution ranging from 32% to 72% for the 35 subjects tested (figure 9). Based on a Chi²-distribution for 80 samples the null hypothesis (“subjects are not able to identify the simulation”) has to be rejected for 8 out of 35 subjects (22.8%), i.e. those with detection rates of more than 59%.

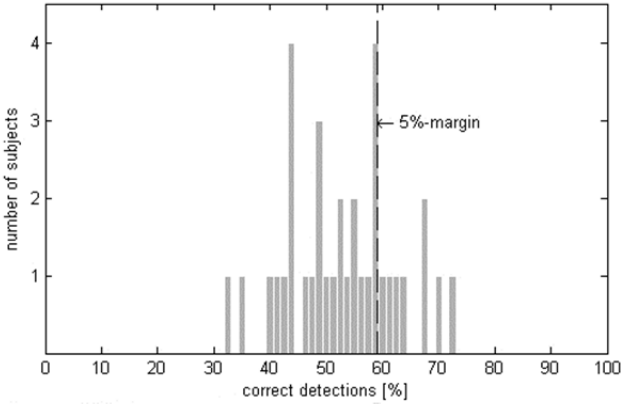


Figure 5-9. Frequency distribution of detection rates in AB-discrimination test, 5% margin for one-sided test is shown.

It is interesting that 5 of those 8 subjects in their questionnaire explicitly indicated that they were mostly guessing during the test. This suggests that they had mainly been better in remembering certain auditory cues and assigning them (rather) consistently to either reality or simulation, while the correct assignment was done merely by chance. This is also supported by the symmetry of the histogram in Figure 5-9.

To find out which attributes were used by the subjects to discriminate between reality and simulation, a questionnaire was filled out right after the test. On the one hand, the answers were analyzed as indications for potential flaws of the system. One the other hand the answers should reveal which features draw most attention when reality and simulation are directly confronted. Looking at the answers of those subjects who could (rather) consistently discriminate between the two conditions (outside of the two-sided 5% Chi² values in Figure 5-9), seven discrepancies were named at least twice. When ordered for frequency of occurrence, these were

1. spectral differences (5x),
2. differences in source localization (4x),
3. differences in reverberant energy (2x),
4. differences in energy on contralateral ear (2x),
5. differences in loudness (2x),
6. latency (2x), and
7. changes in tone during head movements (2x).

It is interesting that, although already considerable effort has been made to compensate the ear canal and headphone transfer function, spectral differences were still most obvious to the subjects.

Since localization differences were reported only with the 130°/30° source and not with the frontal source, they are most likely due to binaural features such as a slight mismatch in interaural time delays (ITD) for certain listeners due to the non-individual HRTFs used.

Concerning the general performance of the binaural simulation it should be kept in mind that the reported cues did not allow any listener to detect the simulation with a probability of more than 72%. With regard to future listening tests the detection rate was also analyzed per audio content, i.e. depending on the anechoic source signal used.

As can be seen in Figure 5-10 the acoustic guitar sample (bourrée by J. S. Bach) and the two speech samples seemed to be most suited to uncover potential artefacts of the simulation, while subjects were less sensitive when the drum sample and the trumpet sample was used. The guitar sound with a combination of transient and tonal features was obviously best suited to make slight differences in timbre as well as in localization audible.

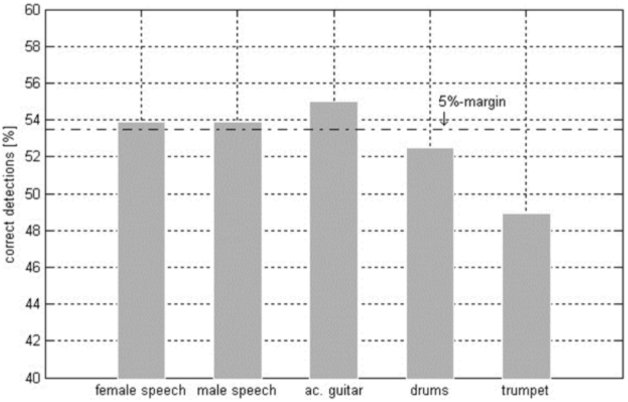


Figure 5-10. Detection rate depending on type of stimulus used.

The detection rate was slightly higher for the 0°/0° source than for the 130°/30° source (53.9% vs. 51.8%) and also when horizontal and vertical head tracking was used (53% vs. 52.7%). While it is known that listeners are most sensitive to directional shifts close to the 0°/0° direction [27], the relevance of BRIRs in two degrees of freedom will be subject to further studies.

5.5 User Interface

For comparative studies of different acoustical environments the binaural rendering engine can be controlled by a specifically designed user interface realized in Matlab® and allowing for different listening test designs. The interface sends OSC commands to the rendering engine via TCP/IP (Figure 5-6).

Test designs already implemented include:

- AB comparisons of different binaurally synthesized or combined real vs. binaural reproductions such as the test presented above. They generate a randomized sequence of stimulus pairs and produce an output report for statistical analysis.
- A qualitative test design according to the repertory grid method (RGT) conducting an automated elicitation of individual attributes by random-

ized triads of binaural stimuli. Creating tables of personal constructs as a basis for subsequent quantitative ratings, the method has been successfully applied to study the perception of different multichannel recording and reproduction techniques [28].

- Rating scales based on individually composed attributes, recommendations such as ITU-R BS 1284 [29], IEC 60268-13 [30], AES 20 [31] or semantic differentials resulting from a qualitative pretest as mentioned above.

5.6 Applications and Outlook

The binaural framework presented will be an efficient tool whenever

- complete sets of binaural room impulse responses (BRIRs) have to be acquired in multiple degrees of freedom, in high directional resolution, and at high speed, and
- perceptual evaluations shall be based on instantaneous switching between different acoustical environments.

We see potential applications in the evaluation of the acoustical impression in different rooms and different listening positions, different configurations of sound reinforcement systems, or in a very fundamental confrontation of a natural acoustic environment, such as a concert hall, and its electroacoustic representation by different recording and reproduction systems. If the focus is extended to the binaural synthesis of computer-modelled environments other scenarios become accessible such as the design of sound installations or even the evaluation of historical developments such as the evolution of concert hall acoustics and its impact on perceptual attitudes [32].

Technical applications could include automotive audio assessment and the enhancement of speech intelligibility in teleconferencing and VoIP applications through binaural synthesis.

Whenever complete sets of BRIRs are acquired for multiple listening positions or even for a given listening area where listeners shall be allowed to move freely over a narrow grid of binaurally sampled listening positions, speed becomes an important issue. This applies to the speed of measurement already provided by the acquisition tool presented, but also to the speed of access to binaural data within the rendering application. Here, efficient strategies of data reduction have to be implemented. It is therefore essential to examine to what extent methods such as interpolation or principal component analysis, successfully applied for “lossless” compression of HRTF data [27], [33], are equally efficient for binaural room impulse responses.

5.7 References

[1] Møller, H. (1992): “Fundamentals of binaural technology”, in: *Applied Acoustics*, **36**(3/4), pp.171-218

[2] Hammershøi, D., Møller, H. (2002): “Methods for Binaural Recording and Reproduction”, in: *Acustica* **88**(3), pp. 303-311

[3] Larcher, V.; Jot, J. M.; Vandernoot, G. (1998): “Equalization methods in binaural technology”, in: *Proc. of the 105th AES Convention*, San Francisco, preprint no. 4858

[4] Mackensen, P. (2003): *Auditive Localization. Head Movements, an additional cue in Localization.* Doct. dissertation. Berlin: Technische Universität

[5] Sandvad, J. (1996): “Dynamic Aspects of Auditory Virtual Environments”, in: *Proc. of the 100th AES Convention*, Copenhagen, preprint no. 4226

[6] Brungart, D. S. et al. (2004): “The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources”, in: *Proc. of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display*. Sydney

[7] Moldrzyk, C.; Lentz, T.; Weinzierl, S. (2005): “Perzeptive Evaluation binauraler Auralisationen”, in: *Proc. of the 31th DAGA*, Munich

[8] Burkhard, M.D.; Sachs, R.M. (1975): “Anthropometric manikin for acoustic research”, in: *J. Acoust. Soc. Am.,* **58**(1), pp. 214-222

[9] Christensen, F.; Jensen, C. B.; Møller, H. (2000): “The Design of VALDEMAR - An Artificial Head for Binaural Recording Purposes”, in: *Proc. of the 109th AES Convention*, Los Angeles, preprint no. 4404

[10] Schmitz, A. (1995): “Ein neues digitales Kunstkopfmeßsystem”, in: *Acustica*, **81**, pp. 416-420

[11] Moldrzyk, C. (2002): “Ein neuartiger Kunstkopf zur Verifikation einer akustischen Entwurfsmethodik für Architekten”, in: *Proc. of the 22nd Tonmeistertagung,* Hannover: 2002

[12] Moldrzyk, C.; Ahnert, W.; Feistel, S.; Lentz, T.; Weinzierl, S. (2004): "Head-Tracked Auralization of Acoustical Simulation", in: *Proc. of 117th AES Convention*, San Francisco, preprint no. 6275

[13] Christensen, F. et al. (2005): "A Listening Test System for Automotive Audio - Part 1: System Description", in: *Proc. of the 118th AES Convention*, Barcelona, preprint no. 6358

[14] Lindau, A.; Weinzierl, S. (2006): "FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom", in: *Proc. of the 24. Tonmeistertagung*, Leipzig, pp. 621-625

[15] Lindau, A.; Weinzierl, S. (2007): "FABIAN - Schnelle Erfassung binauraler Raumimpulsantworten in mehreren Freiheitsgraden", in: *Proc. of the 33th DAGA*. Stuttgart, pp. 633-634

[16] Müller, S.; Massarani, P. (2001): "Transfer-Function Measurement with sweeps", in: *J. Audio Eng. Soc.*, **49**(6), pp. 443-471

[17] Mourjopoulos, J. (1994): "Digital Equalization of Room Acoustics", in: *J. Audio Eng. Soc.,* **42**(11), pp. 884-900

[18] Mourjopoulos, J.; Clarkson, P. M.; Hammond, J.K. (1982): "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals", in: *Proc. of the IEEE ICASSP 1982 (International Conference on Acoustics, Speech, and Signal Processing)*. Paris

[19] Kirkeby, O.; Nelson, P. A. (1999): "Digital Filter Design for Inversion Problems in Sound Reproduction", in: *J. Audio Eng. Soc.*, **47**(7/8), pp. 583-595

[20] Norcross, S. G.; Soulodre, G. A.; Lavoie, M. C. (2002): "Evaluation of Inverse Filtering Techniques for Room/Speaker Equalization", in: *Proc. of the 113th AES Convention,* Los Angeles, preprint no. 5662

[21] Norcross, S. G.; Soulodre, G. A.; Lavoie, M. C. (2003): "Subjective Effects of Regularization on Inverse Filtering", in: *Proc. of the 114th AES Convention*, Amsterdam, preprint no. 5848

[22] Meesawat, K.; Hammershøi, D. (2003): "The time when the reverberant tail in binaural room impulse response begins", in: *Proc. of the 115th AES Convention,* New York, preprint no. 5859

[23] http://jackaudio.org, last visited on Nov. 30th, 2013

[24] Wright, M.; Freed, A.; Momeni, A. (2003): “Open Sound Control: State of the art 2003”, in: *Proc. of the 2003 Conference on New Interfaces for Musical Expression (NIME-03),* Montreal

[25] http://liblo.sourceforge.net, last visited on Nov. 30th, 2013

[26] http://www.fftw.org, last visited on Nov. 30th, 2013

[27] Minnaar, P.; Plogsties, J.; Christensen, F. (2005): “Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis”, in: *J. Audio Eng. Soc.*, **53**(10), pp. 919-929

[28] Berg, J.; Rumsey, F. (2006): “Identification of quality attributes of spatial audio by repertory grid technique”, in: *J. Audio Eng. Soc.*, **54**(5), pp. 365-379

[29] ITU-R BS 1284-1 (2003): *General methods for the subjective assessment of sound quality,* Geneva: International Telecommunication Union

[30] IEC 60260-13 (1985): *Sound system equipment – Part 13: Listening tests on loudspeakers*, Geneva: International Electrotechnical Commission

[31] AES-20 (1996): *AES Recommended practice for professional audio – The subjective evaluation of loudspeakers,* New York: Audio Engineering Society

[32] Weinzierl, S. (2002): *Beethovens Konzerträume: Raumakustik und symphonische Aufführungspraxis an der Schwelle zum modernen Konzertwesen,* Doct. dissertation, Technische Universität Berlin, Frankfurt a. M.: Erwin Bochinsky

[33] Kistler, D. J.; Wightman, F. (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

6 Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-individual Recordings

The following chapter is an authorized reprint of the full-paper peer-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander; Brinkmann, Fabian (2012): "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-individual Recordings", in: *J. Audio Eng. Soc.*, **60**(1/2), pp. 54-62.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, typographic and stylistic corrections.

6.1 Abstract

The headphone transfer function (HpTF) is a major source of spectral coloration observable in binaural synthesis. Filters for frequency response compensation can be derived from measured HpTFs. Therefore, we developed a method for measuring HpTFs reliably at the blocked ear canal. Subsequently, we compared non-individual dynamic binaural simulations based on recordings from a head and torso simulator (HATS) directly to reality, assessing the effect of non-individual, generic and individual headphone compensation in listening tests. Additionally, we tested improvements of the regularization scheme of a LMS inversion algorithm, the effect of minimum phase inverse filters, and the reproduction of low frequencies by a subwoofer. Results suggest that while using non-individual binaural recordings the HpTF of the individual used for the recordings – typically a HATS – should be used for headphone compensation.

6.2 Introduction

6.2.1 Motivation

Binaural reproduction can achieve a high degree of realism. However, when directly comparing dynamic binaural synthesis to the corresponding real sound field we identified spectral coloration as a major shortcoming [1]. In this respect, the common practice to use head and torso simulators (HATS) for creating non-individual binaural recordings is especially problematic. Due to morphological differences, the head related transfer functions (HRTFs) differ from those of the actual listener and result in various distortions of auditory perception as described for instance by Møller et al. ([2]–[4]). Additionally, transducers involved in the binaural recording and reproduction signal chain introduce unwanted spectral coloration. These transducers include loudspeakers and microphones used for binaural measurements, and the headphones used for reproduction. The influence of the headphone transfer function (HpTF) can potentially be compensated by inverse filtering. In an earlier study [5], comparing several inversion approaches for HpTFs, we found highpass-regularized least-mean-square (LMS) inversion [6] approximating a pre-defined bandpass as target function to be a perceptually well-suited inversion algorithm. However, coloration was still audible in these listening tests presumably originating both from using non-individual binaural recordings obtained with our HATS FABIAN [1] and from using non-individual HpTFs for headphone compensation. As an approach to further optimize the headphone compensation in the case of non-individual binaural synthesis, in the present study we examined the effect of using non-individual, generic or individual HpTFs for headphone compensation.

6.2.2 State of the Art

Møller [7] has stated that all spatial information of the sound field is encoded in the sound pressure at the entrance of the blocked ear canal. In turn, the eardrum signal should be perfectly reproducible from the sound pressure measured at the blocked ear canal as long as headphones used for reproduction exhibit a linear frequency response at the blocked ear canal, and an acoustic impedance close to that of free air (free air equivalent coupling, FEC [7]).

To make things difficult, different frequency response target functions deviating considerably from linearity have been defined for headphones [9]–[11]. Kulkarni and Colburn [13] showed that differences can be of the same order as found within HRTFs of different directions of incidence. Moreover, frequency response targets are approached most differently across manufacturers, models and even within batches [5], [8], [12].

For circum- or extraaural headphones the situation is even more complicated: For the same headphone model, the individually differing morphology of the outer ear can cause deviations up to 20 dB between individual HpTFs (inter-individual variability [8], [9]).

Møller et al. [8] found the inter-individual HpTF variability to be reduced when measuring at the blocked ear canal as compared to measuring at the open ear canal.

Transfer functions also vary as headphones are taken on and off repeatedly (intra-individual variability [5], [8], [12]). Therefore, Kulkarni and Colburn [13] recommended compensating the HpTF based on an average of multiple measurements taken while re-seating the headphones in-between. The authors further assumed leakage (e.g., due to incidental gaps in the cushion-head- or cushion-pinna-interface observed with circumaural or supraaural headphones, resp.) to be the dominating cause for intra-individual low-frequency variability observed with re-seating.

Assessing four different headphones in a criterion-free listening test Paquier and Koehl [14] found that positional variability lead to audible deviations.

Wenzel et al. [15] assessed the localization performance achievable with non-individual binaural recordings. Headphones were compensated using the HpTF of the individual whose HRTFs had been used for auralization (termed "reference subject"). Authors stated that recordings would be reproduced the less "faithful" the less the test subjects' HpTFs resembled that of the reference subject. While this might hold true, it must be kept in mind that these were still the wrong (i.e. the non-individual) HRTFs which were reproduced more faithful[4].

The benefit of individual over non-individual headphone compensation while rendering individual HRTFs, was illustrated by Pralong and Carlile [16]. Comparing two subjects' HpTFs they found deviations of up to 10 dB in the region of 3–7 kHz, which would in turn be distorting the HRTFs if applied as non-individual headphone compensation. On the other hand, the authors showed that using individual HpTFs for compensation leads to an assumed perceptually transparent reproduction (mean difference within ±1 dB) of both individual and non-individual binaural recordings.

[4] Additional note not included in the published version of this article: Wenzel's observation that simulations were perceived the more faithful the more the HpTFs of test subject and "reference subject" resembled each other may be explained by the fact that the morphology of the two subjects' outer ears showed a high resemblance. This in turn should also result in a high(er) resemblance of HRTFs, giving a good explanation why such subjects described the non-individual simulation as more 'faithful' than other subjects less well resembling the reference subject in terms of morphology.

Martens [17] assessed the benefit of generic headphone compensation (i.e. a compensation filter based on the average HpTF obtained from several individuals). From a model-based test design the author concluded generic headphone compensation to be sufficient for faithful binaural synthesis.

By means of auditory filter analysis we assessed the differences of non-individual, generic or individual HpTFs for compensation [18]. Results (cf. section 6.3.2) suggested that the fidelity of the equalization increases with the amount of individualiindividualization used in headphone compensation. However, whether this trend was beneficial also for auralization with non-individual recordings remained to be studied.

6.2.3 Scope of the Study

Directly comparing a non-individual binaural simulation to the respective real sound source, we aimed at assessing the effect of non-individual, generic, and individual headphone compensation on the perceived difference. Additionally in [5], subjects occasionally mentioned pre-ringing and high frequency flaws. Therefore, we also assessed the fidelity of binaural reproduction while using minimum phase instead of unconstrained phase filters for compensation, and for several improvements of the highpass regularization inversion scheme to better adapt to the high frequency characteristics of HpTFs. Further on, as a possible future means to extend binaural reproduction beyond the lower cut-off frequency of headphones, we assessed the binaural reproduction's realism when it was combined with a subwoofer reproducing the low frequency components (50–166 Hz).

6.3 Methods

6.3.1 Measuring Individual HpTFs

In [18] we presented custom-built silicone earplugs flush-cast with miniature electret condenser microphones (Knowles FG 23329, ø 2.5 mm,) for measuring individual HpTFs at the blocked ear canal. Inserts were fabricated in three sizes based on anthropometrical data supplied by the manufacturer PHONAK (cf. Figure 6-1).

For validating these earplugs, transfer functions were measured on a silicon-made artificial ear with ear canal, while reinserting the measurement microphone after each measurement. For comparison, these measurements were also conducted using two different types of foam earplugs: the EAR II plug, and a less obtrusive and softer UVEX “com4-fit” foam plug, both commonly reported in binaural literature.

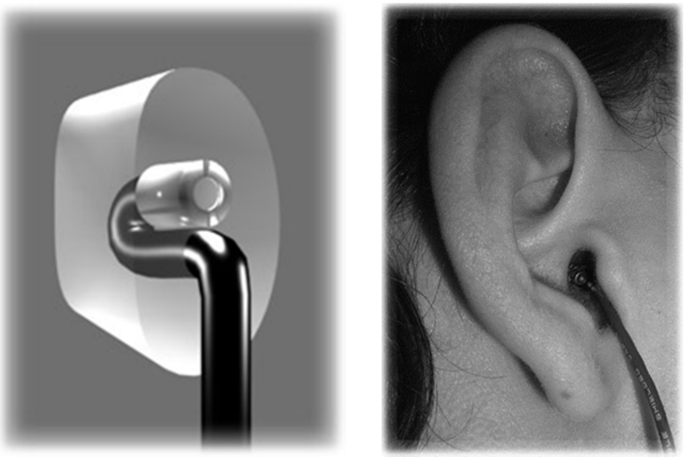


Figure 6-1. Custom-built silicone earplug. Left: CAD model, right: inserted into a subject’s ear (from [18]).

Due to replacement, the foam plug measurements showed deviations up to ± 10 dB and more, whereas deviations obtained with our silicon earplugs were negligible below 8 kHz; above this range deviations reached ± 2 dB (cf. Figure 6-2).

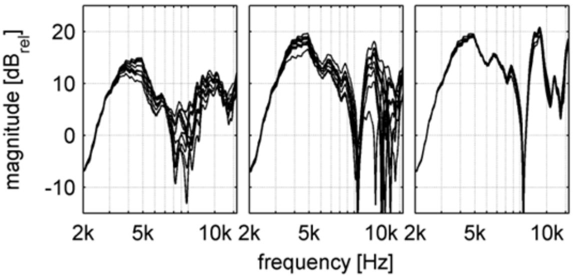


Figure 6-2. Ten transfer functions of an artificial ear measured with Knowles FG 23329 and EAR II foam plug (left), UVEX foam plug (middle), and novel silicone earplug (right), when plugs were reinserted between each measurement (from [18]).

Further on we measured HpTFs of 2 female and 23 male subjects using STAX SR 202 headphones and our silicone earplugs [18]. Subjects had to reposition the headphones before each of ten measurements. The spectral variability of all 250 measurements is depicted in Figure 6-3. Four characteristic frequency ranges could be identified. Below 200 Hz (region I), differences of ±3 dB can primarily be assigned to leakage effects. Up to 2 kHz (region II), differences were smaller than 1 dB. Above 2 kHz and up to 5 kHz (region III), deviations quickly increased to ±3 dB. Above 5 kHz (region IV), the region of narrow pinna notches began. Deviations in that range were distributed asymmetrically between approx. +7 and -11 dB, respectively.

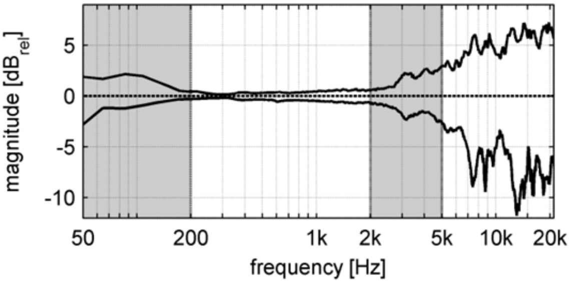


Figure 6-3. Spectral variability of individual HpTFs (left ear only), upper and lower curve enclose the 75 % percentile of the magnitude transfer functions. Shaded areas differentiate between characteristic regions of HpTF curves.

6.3.2 Auditory Modeling of Inversion Results

We used the measured HpTFs to analyze non-individual, generic, and individual headphone compensation in a quantitative manner. Non-individual headphone compensation was realized by filtering each subject's HpTF with the inverse HpTF of a singular subject (here: HATS FABIAN). The inverse of the average HpTF across all 25 subjects served for the generic compensation, whereas for individual compensation the inverse of each subject's own average HpTF was applied. An auditory filter bank of 40 equivalent rectangular bandwidth (ERB) filters was used to model the perceptual deviation between compensated HpTFs and the target function, the latter comprising a bandpass (-6 dB points: 50 Hz, and 21 kHz, 60 dB stopband rejection, cf. [5]).

When comparing compensation results to the overall HpTF variability (i.e. Figure 6-4 to Figure 6-3) it becomes clear, that the non-individual filter provides negligible improvement. As expected, the generic filter is found to symmetrically redistribute the spectral deviations around the 0 dB line, while not reducing the overall amount of spectral variation. Individual compensation promises best results as regions I to III are nearly perfectly equalized.

Only the narrow pinna notches – typically occurring in HpTFs above 5 kHz – remain after compensation. The preservation of notches is perceptually advantageous and directly intended when using the LMS inversion method with highpass-regularization.

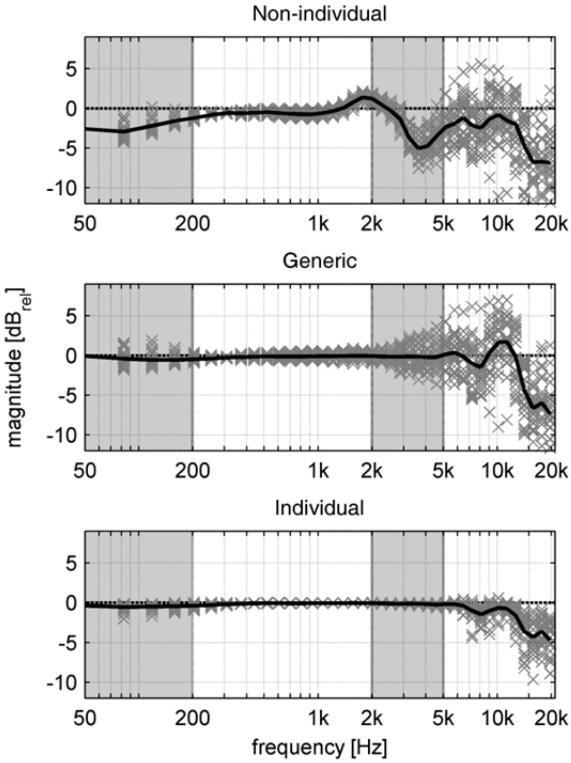


Figure 6-4. Deviations of compensated HpTFs (both ears) of 25 subjects from target function for each band of an auditory filter bank and for three different inversion approaches. Grey crosses: average deviations of singular subjects. Black solid curve: average deviation across all subjects. LMS inversion with highpass-regularization was used throughout (from [18]).

For clarity, regularization means limiting the inversion effort in specific frequency regions. For regularization we used a shelve filter with 15 dB gain and a half-gain frequency of 4 kHz, resulting in a less precise compensation in the amplification region of the filter. As an adverse side effect of this type of regularization the lower plot in Figure 6-4 reveals that high frequency damping in HpTFs is practically left uncorrected, potentially causing a slightly dull or muffled reproduction. In connection with similar statements from subjects in [5], this was our motivation for assessing different improvements of the highpass-regularization scheme in the present study.

6.3.3 Inverse HpTF Filter Design

Throughout this study raw HpTF measurements were shortened to 2048 samples; all inverse filters were designed to have the same length. Before inversion, the measurement microphones' frequency responses were removed from the HpTFs via deconvolution. As target function for the compensated headphones the above mentioned bandpass was defined. The LMS method with highpass-regularization allows designing HpTF inverse filters in the time [6] or in the frequency domain [19]. We used the latter method as it is faster, especially with larger filter lengths. With the conventional LMS methods one typically defines the impulse response (the spectrum, resp.) of a linear phase bandpass as target function. As a result the inverse filters may exhibit considerable pre-ringing (cf. Figure 6-5). Lately, Norcross et al. [20] presented an approach to obtain inverse filters with minimum phase. We also tested this method in the listening test.

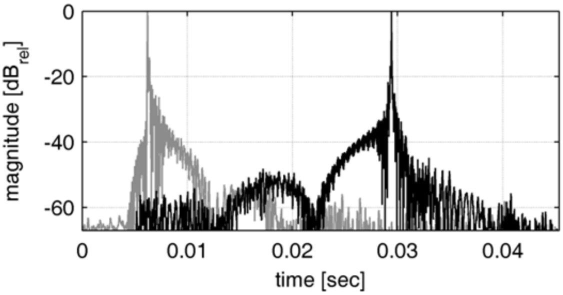


Figure 6-5. Impulse responses of compensated HpTFs. Left/grey: using an LMS filter designed to exhibit minimum phase (acc. to [20]), right/black: using an LMS filter without constraints to filter phase designed to best approximate the (linear phase) target function after compensation.

6.3.4 Subwoofer Integration

The STAX headphones could be equalized to reproduce at moderate levels a frequency range of 50–21 kHz (cf. Figure 6-6, upper curve). Besides, as it might be of future interest to extend the reproduction to the full audio range, we tested integrating an active subwoofer into the binaural playback. Hereby it is assumed that binaural cues conveyed by headphone signals being highpass filtered at some low frequency will still permit a proper perception of spatial sound. This is reasonable as important ILD, ITD and spectral cues are nearly invariant over a larger frequency range (up to ca. 1 kHz). If we could show the "2-way"-reproduction's realism to be comparable to that of the headphones-only mode it might render possible a low-frequency-extended reproduction of binaural content.

The ADAM SUB8 is a small (single 8'' driver) bass reflex design with adjustable gain and low pass cross over frequency. It can be fitted well beneath a listener's chair. For frequency response calibration near field measurements were conducted.

Using two parametric equalizers from a digital loudspeaker controller (Behringer DCX2496) we could establish a nearly ideal 4th order bandpass behavior within 26–150 Hz.

Figure 6-6 shows measurements from the final calibration procedure in the listening test room collected at the ear canal entrance of a test subject while wearing the STAX headphones. Room modes disturbing the fidelity of the subwoofer reproduction were treated by applying two more parametric equalizers of the DCX2496. A smooth summation of subwoofer and headphones output was achieved by level adjustment using the DCX2496 and phase delay adjustment using both a pre-delay in the target bandpass – applied on the subwoofer to shape its lower slope – and the phase switch of the SUB8. After comparing calibration results from the test subject and the FABIAN HATS we assumed the alignment to be nearly invariant across subjects. In the listening test, the subwoofer was driven with a mono sum of the binaural headphone signals which were attenuated by 6 dB to preserve the level calibration.

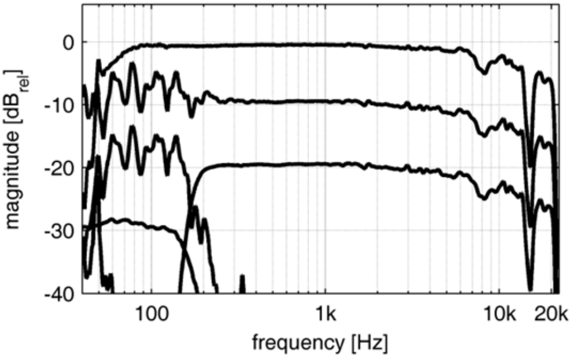


Figure 6-6. Magnitude spectra of compensated HpTF (frequency domain LMS inversion with highpass-regularization) measured at a test subject's right ear. Curves top down: 1) full range headphone reproduction, 2) sum response of 2–way reproduction, 3) 2–way reproduction, subwoofer and headphones shown separately, 4) near field response of subwoofer (all curves 1/24th octave smoothed, 10 dB offsets for clarity).

This way, we were able to present the binaural signals in two different reproduction modes. In (a), the full range mode, only the headphones – equalized to approximate the target bandpass response – were used, whereas in (b), the 2-way reproduction mode, headphone filters were designed to yield a crossover behavior at 166 Hz (-6 dB) reproducing the target bandpass response in summation with the subwoofer.

6.4 Listening Test I

Two listening tests were conducted. In the first we aimed at a perceptual evaluation of the three compensation approaches (non-individual, generic, individual). Conducting the second listening test was found necessary after inspecting the somewhat unexpected results of listening test I as will be explained in section 6.5.

In an acoustically dry recording studio ($RT_{1\ \mathrm{kHz}}$ = 0.47 s, V = 235 m^3), binaural room impulse responses (BRIRs) were measured using the FABIAN HATS. A measurement loudspeaker (Genelec 1030a) was placed frontally in a distance of 2 m, and BRIRs were measured for horizontal head movements within an angular range of ± 80° and a step size of 1°. Our investigation was thus restricted to frontal sound incidence. In our opinion, for detecting spectral deficiencies of headphone compensation though, this familiar source setup made it most easy for subjects to focus on the task, thus resembling an – intended – worst case condition.

During measurements the HATS already wore the STAX headphones. These headphones are virtually transparent to exterior sound fields, in turn allowing simulation and reality to be directly compared without taking them off. Thus, by applying dynamic auralization accounting for horizontal head movements [1] the virtual loudspeaker – presented via differently compensated headphones – could be directly compared to the real loudspeaker.

Besides (a) the three described approaches to headphone compensation (factor 'filter'), we additionally assessed (b) type of content (pink noise and acoustic guitar, factor 'content'), (c) the use of minimum phase versus unconstrained inverse filters (factor 'phase'), and (d) the effect of a 2-way binaural reproduction with the low frequency content being reproduced by a calibrated subwoofer (factor 'reproduction mode') resulting in 3 x 2 x 2 x 2 = 24 test conditions which were assessed in a fully repeated measures design (each subject assessed each condition). As we expected no interactions between tested factors, and while assuming an inter-subject correlation of 0.4, 20 subjects were calculated to be needed for testing a small main effect (E = 0.1) at a type-1 error level of 0.05 and a power of 80 % [21], [22].

Subjects were seated in the former position of FABIAN in front of the real loudspeaker while their absolute position was controlled by aligning their ear canal entries using two perpendiculars. At the beginning of each listening test, the individual HpTFs were measured using our insert microphones and filters were calculated with prepared Matlab® routines. Training was conducted to familiarize subjects with stimuli and the rating process. In a multiple-stimulus ABC/HR listening test paradigm [23] 27 subjects (24 male, 3 female, avg. age 31.7 years) had to (a) detect the simulation and (b) rate its similarity with respect to the real loudspeaker reproduction. On the graphical user interface, subjects found two sliders and three play buttons ("A", "B", "Ref/C") for each stimulus condition. The two buttons adjoining the sliders were randomly playing the test stimulus (HpTF-compensated simulation) or the reference (the real loudspeaker), the third button, "Ref/C", always reproduced the reference. Slider ends were labeled "identical" and "very different" (in German), and ratings were measured as continuous numerical values between five and one. Only one of the two sliders could be moved from its initial position ("identical"), which would also indicate this sample as being identified as the test stimulus. While taking their time at will, subjects compared sub sets of six randomized stimuli using one panel of paired sliders. Within each sub set the audio content was kept constant. The length of the stimuli was about five seconds. For unbiased comparability the frequency response of the real loudspeaker was also limited by applying the target bandpass. Additionally, real time invidualization of the interaural time delay according to [25] was used throughout the listening test. Including HpTF measurement, filter calculation, and training, the test took about 45–60 minutes per subject, of which on average 20 minutes were needed for rating.

6.5 Results of Listening Test I

Results from two subjects were discarded after post-screening: one rated all simulations equally with "very different", another one experienced technical problems while testing. Following [23], results were calculated as difference grades, subtracting the test stimulus' rating from the reference's rating. If the test stimulus was correctly identified all the time, only negative difference ratings would be observed (ranging from 0 = "identical" to -4 = "very different"). For all 24 test conditions average difference ratings and confidence intervals of the remaining 25 subjects are shown in Figure 6-7.

Obviously, the simulation was always clearly detectable (negative average difference grades). This is not surprising as the ABC/HR design provides an open reference (i.e. the real loudspeaker is always played back when hitting the "Ref/C" button). Thus, slightest spectral deviation will enable subjects to rather easily detect the test stimulus, which in turn is likely the case as the binaural recordings were explained to be non-individual (cf. section 6.2.1).

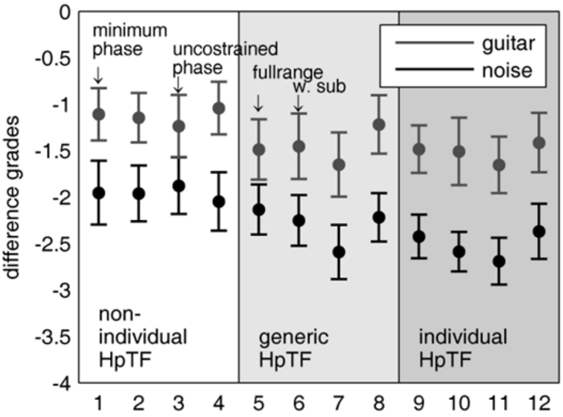


Figure 6-7. Results from listening test I: Difference grades and 95 % confidence intervals for all conditions averaged over all subjects. Shaded columns differentiate between filter types. Ratings for conditions phase and reproduction mode alternate throughout columns as indicated by arrows.

The effect of content is also clearly obvious; moreover, for type of filter a noticeable variation can be seen. Effects of the conditions phase and reproduction mode are less obvious. As no intermediate anchor stimuli were defined, ratings were z-normalized across subjects before being subjected to inferential analysis (repeated measures ANOVA) [23]. In terms of average difference ratings we had formulated the following *a priori* hypotheses for the four main effects a) $\mu_{\text{individual}} > \mu_{\text{generic}} > \mu_{\text{non-indivdual}}$, b) $\mu_{\text{guitar}} > \mu_{\text{noise}}$, c) $\mu_{\text{minimum-phase}} > \mu_{\text{uncostrained-phase}}$, d) $\mu_{\text{1-way}} = \mu_{\text{2-way}}$. The inter-rater reliability was pleasingly high (Cronbach's α 0.944), indicating a sufficient duration of the training phase. We found effects for content and filter to be highly significant. In agreement with [5] and our *a priori* hypothesis overall difference grades were significantly worse for the noise content. This is not surprising as the problematic frequency ranges of the compensated HpTF ranges (cf. Figure 6-6) will be excited much stronger by wide band noise than with the rather limited frequency range of the guitar stimulus. However, the filter effect surprised us, as the simulation compensated with the non-individual HpTF (that of the FABIAN HATS) was rated best. Multiple comparisons with Bonferroni adjustment furthermore showed that generic and individual compensation differed only insignificantly from each other, at least a trend was observed for the individual compensation to be rated worse. No significant effect of phase could be found, although there was a trend for unconstrained phase filters to be rated slightly worse. Additionally, and in accordance with our *a priori* hypothesis, no effect of reproduction mode, i. e. no difference in the amount of perceived similarity with the real sound source could be found between headphones-only and 2-way reproduction mode. While – from post-hoc power calculation – having been able to reject a small effect size of E = 0.0924 with 80 % power, the latter null hypothesis can assumed to be well supported.

6.6 Discussion of Results of Listening Test I

Although the 2-way reproduction showed moderate low frequency distortion (± 4 dB, Figure 6-6), the amount of perceived similarity with the real sound source was of the same order as for full range headphone reproduction. Thus, results support the conclusion that with the application of moderate room equalization, proper level adjustment, and crossover design subwoofers might well be integrated into binaural reproduction for low frequency reproduction. Moreover, a future extension of the reproduction to the full audio range (i.e. down to 20 Hz) should be considered.

Regarding the effect of filter, from verbal responses of subjects we were already informed that when compared to reality, generic and individual compensation were perceived more damped in the high frequencies as compared to the non-individually compensated simulation. In order to understand what happened, we tried to reconstruct the signal differences subjects have perceived when comparing simulations and natural listening. Therefore, in the same setup as in listening test I, we measured five subjects' HpTFs and their BRIRs for frontal head orientation. Two different kinds of headphone compensation: (1) non-individual (HpTF from FABIAN), and (2) individual (HpTFs from each of the five subjects), were applied to the subjects' HpTFs. Afterwards, HpTFs were convolved with FABIAN's frontal BRIR to obtain the signal our simulation would have produced at the five listeners' ears for a neutral head orientation. From comparison of the spectrum of the subjects' own BRIRs and those of the differently compensated simulations (cf. Figure 6-8 for spectral difference plots) we got an impression of the coloration people would have actually perceived in each of these situations.

While admitting that due to the small sample size this examination has an informal character, results confirmed that spectral differences (which were pronounced only above 5 kHz) were on average less in the case of non-individual headphone compensation. An explanation might be that the HpTF of FABIAN, measured with circumaural headphones, closely resembles a near-field HRTF preserving prominent spectral features from the pinna characterizing also FABIAN's BRIRs used in the listening test. Using FABIAN's HpTF to compensate the headphone reproduction of FABIAN's binaural recordings may have resulted in a kind of de-individualization of the binaural simulation, especially compensating FABIAN's dominating high frequency (i.e. pinna-related) spectral characteristics. In contrast, when using the subjects own HpTF (individual compensation), the characteristics of the 'foreign' BRIRs are reproduced nearly unaltered, meaning that inter-individual deviations will become most obvious.

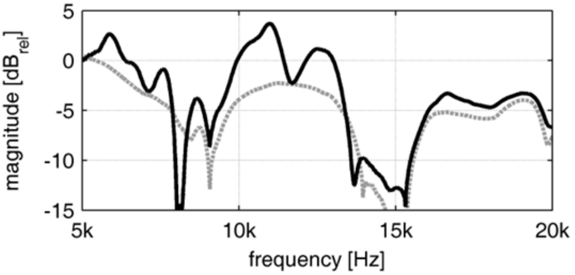


Figure 6-8. Octave smoothed difference spectra of individual BRIRs and BRIRs from two binaural simulations using different headphone compensations (averaged across 5 subjects and both ears). Solid black curve: difference to non-individual BRIR compensated with non-individual HpTF, dashed grey curve: difference to non-individual BRIR compensated with individual HpTF.

We thus concluded that using HpTF of the subject which served also for non-individual binaural recordings was a special case not covered by our prior three-stage classification scheme of the filter types. To test our initial hypothesis again, we set up a new listening test, this time using a "true" non-individual HpTF, selected at random from the sample of listening test I (cf. section 6.8). Summing up, findings indicate that headphone compensation for binaural reproduction cannot be discussed without regarding the actual binaural recordings to be reproduced.

6.7 Improving Regularization

As a new listening test was scheduled, we used the opportunity to test some more hypotheses. At first, we were concerned with improving the highpass-regularization scheme. Two new methods were considered. The first is based on the assumption that a HpTF has to be compensated equally well within the complete passband range (no general limitation in the high frequency range), while still taking care of 1–3 problematic notches typically occurring in HpTFs. A routine was programmed in Matlab®, which allowed us to define a regularization function which is flat on overall except for 1–3 parametric, peaking notch filters at the position of notches in the subject's HpTF. This in turn would limit the inversion effort only at the notches while flattening out all other deviations from linearity (termed "PEQ regularization" in the following). For the second approach, we assumed that regularization should somehow adapt to the HpTF, primarily flattening boosts while being of less effect with occurring notches. This behavior can be achieved by using the inverse average HpTF itself as a regularization function [24]. We already tested this approach in [5] while using an octave smoothed version of the inverse HpTF. We considered inferior perceptual results in [5] to be due to the spectral resolution being too coarse. Therefore, this time we tested a sixth octave smoothed inverse HpTF (cf. [24]) as a regularization function (we termed this approach the "HpTF inverse regularization").

6.8 Listening Test II

In the second listening test, we assessed effects of four factors: (a) the use of "individual" vs. "true non-individual" headphone compensation, with the latter being a HpTF related to neither the current test subject nor the binaural dataset in use (factor 'filter'), (b) the two new regularization schemes (PEQ regularization, HpTF inverse regularization) and the highpass-regularization (factor 'regularization'), (c) again, the susceptibility to filter phase, this time using an assumed more critical stimulus, a drum set excerpt (factor 'phase'), and (d) the type of content (pink noise, drum set, factor 'content'). The listening test design was exactly the same as for test I. Again, the number of tested condition was 2 x 3 x 2 x 2 = 24. Maintaining above mentioned specifications for test sensitivity and power, 27 new subjects (20 male, 7 female, avg. age 27.6 years) participated in the test.

6.9 Results of Listening Test II

No subject had to be discarded in post-screening. The interrater reliability was again high (Cronbach's α 0.919). Average difference ratings and confidence intervals of the 27 subjects are shown in Figure 6-9.

Overall detectability and the effect of content were comparable to test I. The effect of filter was now as expected: The "true" non-individual compensation was rated much worse than the individual condition. From comparison of Figure 6-7 and Figure 6-9, true non-individual compensation can assumed to be the worst choice in any of the tested cases. It though remains untested whether using no headphone compensation at all (cf. [5]) might be even worse.

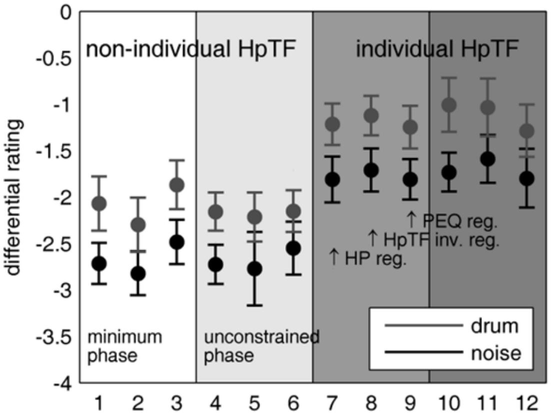


Figure 6-9. Results from listening test II: Difference grades and 95 % confidence intervals for all conditions averaged over all subjects. Lighter/darker shaded columns differentiate between filter types. Ratings for conditions phase and regularization alternate throughout columns as indicated by arrows.

Effects of phase and regularization seem to be negligible. Standardized difference ratings were again subjected to repeated measures ANOVA. Effects for content and filter were found to be highly significant. Again, no susceptibility to filter phase ($p = 0.98$) could be found. Also, types of regularization showed no audible effect ($p = 0.44$), though there was a significant interaction (filter*regularization) indicating using the inverse smoothed HpTF for regularization to be best suited for individual HpTF compensation.

6.10 Conclusion

In two listening tests, we addressed the effect of different aspects of headphone compensation on the perceived difference of non-individual dynamic binaural synthesis when compared to reality. We assessed susceptibility to filter individualization, to filter phase, to audio content, the effect of a hybrid reproduction incorporating a subwoofer and improvements of the highpass-regularized LMS inversion scheme (the latter only for individual and “true” non-individual HpTF compensation). The effect of headphone compensation was found to be not straight forward. Surprisingly, non-individual binaural recordings which were headphone-compensated using the HpTF of the subject used for these recordings were perceived as most realistic. Due to the scope of this study, this conclusion remains limited to the case of non-individual recordings. With individual binaural recordings though, to us there appears to be no reason why the individual HpTF should not be the best choice. A pronounced susceptibility to filter phase could not be found as well as an overall effect of two novel regularization schemes. A significant interaction though indicated the sixth octave smoothed inverse HpTF regularization to be more adequate in case of individual HpTF compensation. Using a cross over network, level, phase, and room correction calibrated at a reference subject's ear canal entrance, a subwoofer was shown suitable for low-frequency reproduction of binaural recordings.

6.11 Acknowledgements

Alexander Lindau and Fabian Brinkmann were supported by grants from the Deutsche Forschungsgemeinschaft (DFG WE 4057/1-1, and DFG WE 4057/3-1, respectively).

6.12 References

[1] Lindau, A.; Hohn, T., Weinzierl, S. (2007): "Binaural resynthesis for comparative studies of acoustical environments", in: *Proc. of the 122nd AES Conv.*, Vienna, preprint no. 7032

[2] Møller, H. et al. (1997): "Evaluation of Artificial Heads in Listening Tests", in: *Proc. of the 102nd AES Conv.,* Munich, preprint no. 4404

[3] Møller, H. et al (1995): "Head-Related Transfer Functions of Human Subjects", in: *J. Audio Eng. Soc.*, **43**(5), pp. 300-332

[4] Møller, H. et al. (1996): "Binaural Technique: Do We Need Individual Recordings?", in: *J. Audio Eng. Soc.*, **44**(6), pp. 451-469

[5] Schärer, Z., Lindau, A. (2009): "Evaluation of Equalisation Methods for Binaural Signals", in: *Proc. of the 126th AES Conv.*, preprint 7721

[6] Kirkeby, O.; Nelson, P. A. (1999): "Digital Filter Design for Inversion Problems in Sound Reproduction", in: *J. Audio Eng. Soc.*, **47**(7/8), pp. 583-595

[7] Møller, H. (1992): "Fundamentals of Binaural Technology", in: *Applied Acoustics*, **36**(3/4), pp. 171-218

[8] Møller, H. et al. (1995): "Transfer Characteristics of Headphones Measured on Human Ears", in: *J. Audio Eng. Soc.*, **43**(4), pp. 203-217

[9] Sank, J. R. (1980): "Improved Real-Ear Tests for Stereophones", in: *J. Audio Eng. Soc.*, **28**(4), pp. 206-218

[10] Theile, G. (1986): "On the Standardization of the Frequency Response of High-Quality Studio Headphones", in: *J. Audio Eng. Soc.*, **34**(12), pp. 956-969

[11] Møller, H. et al. (1995): "Design Criteria for Headphones", in: *J. Audio Eng. Soc.*, **43**(4), pp. 218-232

[12] Toole, F. E. (1984): "The acoustics and psychoacoustics of headphones", in: *Proc. of the 2nd Int. AES Conference*. Anaheim, CA

[13] Kulkarni, A., Colburn, H. S. (2000): "Variability in the characterization of the headphone transfer-function", in: *J. Acoust. Soc. Am.*, **107**(2), pp. 1071-1074

[14] Paquier, M., Koehl, V. (2010): "Audibility of headphone positioning variability", in: *Proc. of the 128th AES Conv.*, London, preprint no. 8147

[15] Wenzel, E. M. et al. (1993): "Localization using nonindividualized head-related transfer functions", in: *J. Acoust. Soc. Am.*, **94**(1), pp. 111-123

[16] Pralong, D., Carlile, S. (1996): "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space", in: *J. Acoust. Soc. Am.*, **100**(6), pp. 3785-3793

[17] Martens, W. L. (2003): "Individualized and generalized earphone correction filters for spatial sound reproduction", in: *Proc. of ICAD 2003*. Boston

[18] Brinkmann, F., Lindau, A. (2010): "On the effect of individual headphone compensation in binaural synthesis", in: *Proc. of the 36th DAGA,* Berlin, pp. 1055-1056

[19] Kirkeby, O. et al. (1998): "Fast Deconvolution of Multichannel Systems Using Regularization", in: *IEEE Transactions on Speech and Audio Processing*, **6**(2), pp. 189-195

[20] Norcross, S.G. et al. (2006): "Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization", in: *Proc. of the 121nd AES Conv.*, preprint no. 6929

[21] Bortz, J., Döring, N. (2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 4. ed., Heidelberg: Springer, pp. 618, pp. 630

[22] Faul, F. et al. (2007): “G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences”, in: *Behavior Research Methods,* **39**(2), pp. 175-191

[23] ITU-R Rec. BS.1116-1 (1997): *Methods for the subjective Assessment of small Impairments in Audio Systems including Multichannel Sound Systems*, Geneva

[24] Norcross, S.G.; Soulodre, G.A., Lavoie, M.C. (2002): “Evaluation of Inverse Filtering Techniques for Room/Speaker Equalization”, in: *Proc. of the 113th AES Conv.*, Los Angeles, preprint no. 5662

[25] Lindau, A.; Estrella, J., Weinzierl, S. (2010): “Individualization of dynamic binaural synthesis by real time manipulation of the ITD”, in: *Proc. of the 128th AES Conv.*, London, preprint no. 8088

7 An Extraaural Headphone System for Optimized Binaural Reproduction

The following chapter is an authorized reprint of the abstract-reviewed article (reproduced from the author's post-print):

> Erbes, Vera; Schultz, Frank; Lindau, Alexander; Weinzierl, Stefan (2012): „An extraaural headphone system for optimized binaural re-production“, in: *Proc. of the 38th DAGA (in: Fortschritte der Akustik).* Darmstadt, pp. 313-314.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, typographic and stylistic corrections.

Author's Note

In the case of the publication at hand the author of this dissertation was not the primary author. However, this publication marks an end point of a longer line of research initiated and supervised by the author. Hence, for the sake of completeness of the whole presentation it was decided to be included.

7.1 Introduction

The transparent binaural reproduction of virtual acousticenvironments requires a headphone system with high spectral bandwidth, with a transfer function as linear as possible, a high signal-to-noise ratio (SNR), sufficient crosstalk attenuation and a frequency response which is robust with respect to repositioning and interindividually differing morphology [1], [2].

Additionally, for *in situ* comparisons with real sound fields and for low-frequency extension with subwoofers, such a system should be sufficiently transparent to exterior sound fields and comply to the 'free air equivalent coupling'-criterion (FEC), i.e. approach the acoustic impedance of free air as seen from the ear canal entrances [1]. Moreover, it should be easy to perform an individual headphone transfer function compensation, e.g., with miniature in-ear microphones [3]. Finally, the system should be combinable with other technical components of virtual environments such as 3D shutter glasses, head mounted displays and head tracking sensors.

For this purpose, an extraaural headphone system was developed (BKsystem comprising BK211 headphones and BKamp power amplifier, cf. Figure 7-1), featuring an extraaural headset, an IIR-/FIR-filter DSP-system and a low noise power amplifier. Measurements show that the system fulfills the requirements specified above better than other commercially available headphones.

7.2 The Extraaural Headphone System

The body of the BK211 headset was fabricated based on a designed 3D-CAD model using the selective laser sintering (SLS) rapid prototyping technique (cf. Figure 7-1, left). For the left and right channel it incorporates an acoustically separated closed-box loudspeaker design with an effective volume of about 300 ml driven by a 2 inch full range electrodynamic transducer (Peerless 830970). Five centimeters were chosen as the average 'transducer to ear canal entrance'-distance while a turning knob allows adjusting the headset to an individual head diameter within a typical range of variation (±15 mm, cf. [4]). To decouple structure-born sound and to increase wearing comfort the headset features resiliently mounted cushions which are tuned to a frequency two octaves below the typical lower cut-off frequency (f_c = 55 Hz) of the headphone system. The weight of a completely assembled headset is approximately one kilogram.

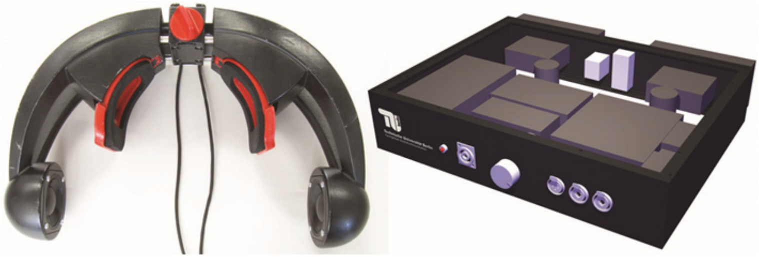


Figure 7-1. BKsystem headphone system: extraaural headset BK211 (left) and DSP-driven amplifier unit BKamp (right).

Already for a first prototype of the extraaural headphone crosstalk attenuation was shown to be 23 dB in average up to 2 kHz and increasing to 60 dB at higher frequencies [4], which was considered to be sufficiently small even without additional crosstalk cancellation. Using boundary element method (BEM) simulations, the FEC criterion was shown to be perfectly fulfilled up to 3 kHz [4] and still FECcompliant according to [1] above this frequency. The headset is to be used with a dedicated driving unit (BKamp) which integrates a two-channel DSP section and low noise power amplifiers in a standard 2U 19” rack housing (cf. Figure 7-1, right). The maximum input voltage of the unit is 4.9 V_{peak} which matches the typical output level of professional sound cards. The whole signal chain is designed to realize an inaudible self noise while providing sufficient gain for spectral compensation to reach the target transfer function – a 4th-order Butterworth high-pass at $f_c = 55$ Hz. Using IIR-filtering this target function is realized on axis at a distance of five centimeters in free field within ±2 dB. A full-scale pink noise with a crest factor of 13 dB – assumed typical for music material – which will be filtered according to the target function, yields 1.1 V_{RMS} at the 4Ω-rated amplifier output. In this case the noise floor is 27.5 μV_{RMS} resulting in 92 dB SNR. Assuming moderate sound pressures of about 85 − 90 dB_{SPL} at the ear canal entrance the noise floor falls below the threshold of hearing. Total harmonic distortion (THD) then reaches < −40 dB above 200 Hz and does not exceed −15 dB for the lowest frequencies of the target function.

7.3 HpTF Compensation

As the BK211 has been linearized for the free field situation, the headphone transfer function (HpTF) measured at the blocked ear canal reveals different sources of frequency response distortion, including the lateral near field head related transfer function (HRTF) and a standing wave pattern originating from the distance between loudspeaker membrane and the head (cf. Figure 7-2, label [b]). When reproducing individual binaural recordings the HpTF of the BK211 should be linearized based on individual HpTF measurements [3]. For HpTF-linearization a bandpass characteristic (Figure 7-2, label [a]) comprising a 55 Hz Butterworth 4th-order high-pass and a 21 kHz Kaiser-Bessel windowed FIR low-pass with 60 dB stop band rejection was used as a target response.

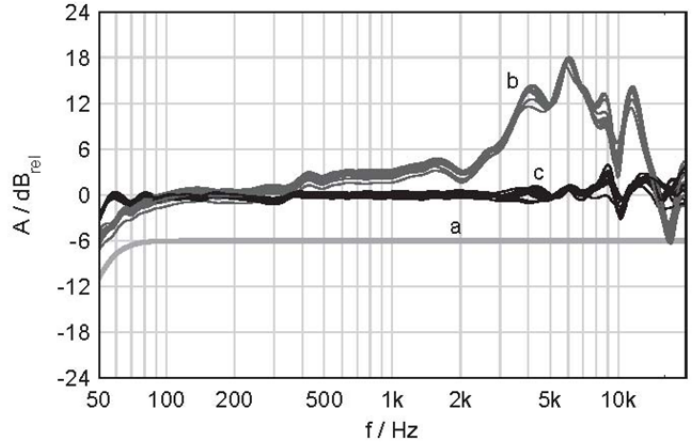


Figure 7-2. Individual linearization of BK211 with repositioning between measurements: (a) desired target response (shifted by -6 dB), (b) 11 measured HpTFs displayed with 1/6 octave smoothing, (c) linearized HpTFs displayed with 1/6 octave smoothing. The compensation filter was derived from the complex average of case (b) using a high-shelve filter regularized LMS-inversion.

Based on the complex average of 11 HpTFs measured with repositioning between measurements an inversion filter (FIR order 2^{12}, 44.1 kHz sampling rate) was generated using a high-shelve (15 dB gain, half-pad gain at 4 kHz) regularized least mean square approach according to [5]. Linearized HpTFs are shown in Figure 7-2 (label [c]). Up to 3 kHz a nearly perfect linearization (variation of ±1 dB) can be observed. At higher frequencies the linearization is affected by repositioning variability and limited compensation due to the high-shelve regularization. Compared to alternative supra- and circumaural headphones, however, the overall irregularity of the linearized HpTFs is considerably reduced [1], [2].

Due to the extraaural design, the BK211 headset does not have to be taken on and off to perform in-ear microphone measurements for compensation. Thus, an even more precise linearization of the HpTF can be achieved with only a single measurement (cf. Figure 7-3).

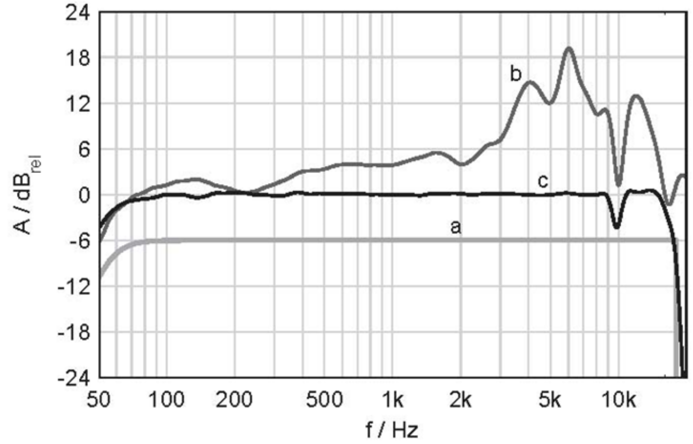


Figure 7-3. Individual linearization of BK211 without repositioning: (a) desired target response (shifted by -6 dB), (b) HpTF displayed with 1/6 octave smoothing, (c) linearized HpTF displayed with 1/6 octave smoothing. The compensation filter was derived from case (b) using a parametrical filter regularized LMS-inversion.

Using a compensation filter designed to linearize the complete spectrum except for the notches at 10 kHz and 17 kHz to avoid potential ringing artifacts due to excessive boosting ('PEQ method' in [3]), 1/6 octave smoothed deviations from the target bandpass (here 55 Hz Butterworth 4th-order highpass, 18 kHz Kaiser-Bessel windowed FIR low-pass with 60 dB stop band rejection) can be show to be within ±0.5 dB outside of the area of the notches. In that case, the -3 dB cut-off frequencies can be stated to be 55 Hz and 16 kHz, respectively, while avoiding highest frequency partial oscillations of the membrane.

7.4 Discussion and Conclusion

We presented the extraaural headphone system BKsystem optimized for binaural sound reproduction in the context of virtual reality applications. It comes as an integrated DSP/amplifier/headset solution featuring a crosstalk attenuation of at least 23 dB and a noise floor below the threshold of hearing while providing a maximum SPL of 101 dBSPL (sine) for a linearized HpTF. The extraaural design was shown to reduce intra-individual HpTF variability and to provide a simple approach to generate accurate individual compensation filters by using miniature in-ear microphones. The internal DSP can be additionally used to provide (a) a variable crossover frequency when extending the low frequency reproduction with a subwoofer, (b) a diffuse-field filter for using the BK211 in a typical recording studio scenario, (c) a storage for individual HpTF compensation filters for optimal binaural reproduction. The BK211 headphone system can easily be equipped with head tracking sensors (pre-built to fit for Polhemus Fastrak sensors) and can be combined with 3D glasses or small head mounted displays for application in multimodal virtual reality environments.

7.5 References

[1] Møller, H. et al. (1995): "Transfer characteristics of headphones measured on human ears", in: J. *Audio Eng. Soc.* **43**(4) (1995), pp. 203-217

[2] Schärer, Z., Lindau A. (2009): "Evaluation of equalization methods for binaural signals", in: *Proc. of the 126th Audio Eng. Soc. Conv.*, Munich, preprint no. 7721

[3] Lindau A., Brinkmann F. (2012): "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings", in: *J. Audio Eng. Soc.* **60**(1/2), pp. 54-62

[4] Schultz, F.; Lindau, A.; Makarski, M.; Weinzierl, S. (2010): „An extraaural headphone for optimized binaural reproduction", in: *Proc. of the 26th Tonmeistertagung.* Leipzig, pp. 702-714

[5] Norcross, S. G.; Bouchard, M.; Soulodre, G. A. (2006): "Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization", in: *Proc. of the 121st AES Convention.* San Francisco, preprint no. 6929

8 Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD

The following chapter is an authorized reprint of the précis-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander; Estrella, Jorgos; Weinzierl, Stefan (2010): "Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD", in: *Proc. of the 128th AES Convention*. London, preprint no. 8088.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, typographic and stylistic corrections.

8.1 Abstract

The dynamic binaural synthesis of acoustic environments is usually constrained to the use non-individual impulse response datasets, measured with dummy heads or head and torso simulators. Thus, fundamental cues for localization such as interaural level differences (ILD) and interaural time differences (ITD) are necessarily corrupted to a certain degree. For ILDs, this is a minor problem as listeners may swiftly adapt to spectral coloration at least as long as an external reference is not provided. In contrast, ITD errors can be expected to lead to a constant degradation of localization. Hence, a method for the individual customization of dynamic binaural reproduction by means of real time manipulation of the ITD is proposed. As a prerequisite, perceptually artifact-free techniques for the decomposition of binaural impulse responses into ILD and ITD cues are discussed. Finally, based on listening test results, an anthropometry-based prediction model for individual ITD correction factors is presented. The proposed approach entails further improvements of auditory quality of real time binaural synthesis.

8.2 Introduction

Virtual acoustic environments (VAEs) are commonly realized via dynamic binaural synthesis. Anechoic audio is convolved in real time with head related impulse responses (HRIRs) or binaural room impulse responses (BRIRs). By definition, HRIRs are measured in anechoic environment for a discrete set of angular positions on a sphere around a real person or a dummy head [1]. In contrast, BRIRs can be measured in (echoic) environments for a discrete number of head orientations by using dummy heads that are motion controlled above their torso [2]. During real time convolution, binaural impulse response filters are inaudibly exchanged according to the listener's head movements as observed by a head tracking device. Thus, realistic auditory experiences such as static or moving sound sources or the impression of the acoustics of a real room can be provided.

8.2.1 Basic Localization Cues

From Lord Rayleigh's [3] duplex theory of hearing it is known, that interaural level (ILD) and time (ITD) differences are primarily exploited for auditory localization. From that point of view, a pair of HRIRs respectively HRTFs (head related transfer functions) may be regarded as a self-contained and frequency dependent description of all auditory localization cues corresponding to a sound event of a distinct direction of incidence. A BRIR additionally contains the complete multitude of reflections excited by the source, including cues for distance and directivity of the source.

According to [4] the ITD is frequency dependent in a way that the ITD for higher frequencies (i.e. for ka_{head} >> 1) is about two-thirds as large as the low frequency ITD [4]. For localization, the ITD below 1500 Hz is evaluated. Above that frequency range, interaural envelope shifts and ILDs increasingly determine the impression of a sound events direction [5]. If a signal contains ambiguous temporal and spectral localization cues, the ITD tends to dominate the perceived sound source direction [6].

8.2.2 Non-individualized Binaural Data

The geometry of an individual's pinna, head and torso influences the ILDs and ITDs of binaural impulse response data sets. The diameter of the head primarily affects the low frequency ITD. Therefore, ITDs differ individually due to varying head geometries. As it is seldom feasible to collect individual data, binaural impulse responses are usually collected with dummy heads or head and torso simulators. If these non-individualized binaural data sets are used for auralization, degradation of localization accuracy and of the perceived stability of sound sources may be observed: If the head used for measurement is smaller than the listener's head, a movement of the sound sources in the same direction as the head's movement is perceivable. If, on the other hand, the head is larger, an apparent movement in the opposite direction becomes audible [7]. This effect can be very annoying, especially as adaptation does not seem to occur. For this reason, a method for the customization of binaural synthesis by means of individual real time manipulation of the ITD as contained in binaural impulse response data sets is proposed.

8.3 ITD Individualization

8.3.1 A 'by-product'

For HRIR-based VAEs, some approaches towards individualization of the ITD have been proposed before [7], [8]. Historically, they have evolved somehow as a by-product of the solution for a another issue of VAEs: HRIRs and BRIRs are usually measured for a discrete and finite number of directions of sound incidence respectively head orientations, therefore ITD differences are also discretized. As a result, impulse response peaks of adjacent HRIRs or BRIRs are not time-aligned but 'jump' back and forth between the different measurement grid positions. This has audible consequences for instance when interpolating HRIRs in the time domain in order to achieve a finer angular resolution than that of the original measurement [9] or when cross fading impulse responses due to different head orientations during convolution [10]. Time domain interpolation of delayed impulse responses will necessarily produce comb filtering effects. As a solution, magnitude and phase information of the HRTFs are often separated. Thus, cross fading is applied to audio signals convolved with time aligned HRIRs while the interaural delay is handled separately. To date, with virtual auditory displays based on HRTFs, ITDs were sometimes reintroduced with sample precision as frequency independent delays extracted from the original dataset or as being calculated from functions based on spherical head models, such as the well-known Woodworth-Schlosberg Formula [11], (8-1) and further extensions of it [12], [13].

$$ITD(\theta) = \frac{r_{head}}{c_0}(\sin(\theta) + \theta) \tag{8-1}$$

With equation (8-1) individualization of the ITD becomes easily applicable. Using the head radius of a certain subject, its individual ITD can be calculated. As (8-1) is a simplification based on a spherical head model, in [7], an empirical optimization of individualized ITDs derived from the Woodworth-Schlosberg formula was presented. There, 'optimized individual head radii' r_{opt} – to be applied with the Woodworth-Schlosberg Formula – were derived from individual sets of HRIRs by minimizing an error criterion between empirical and modeled ITDs while varying the model's r_{head}. Finally, a multiple linear model for the prediction of r_{opt} based on anthropometrical measures was developed.

8.3.2 Scaling of Empirical ITDs

In contrast to this approach, individual customization of the ITD could also be accomplished by simple scaling of the ITD by a constant, frequency independent factor. It might also be possible to derive a person's ITD from that of any other person through scaling. While this certainly is a simplified approach, nevertheless, when comparing it to the spherical head models, it exhibits some advantages over the latter. At first, the Woodworth-Schlosberg model formulates an approximation of the ITD that is valid only for high frequencies ($ka_{head} >> 1$, cf. [4]). Secondly, it does not account for asymmetries. However, in reality, the ear canals entrances are not positioned symmetrically on the ends of a sphere's diameter, as is usually assumed in spherical head models. Instead, they are slightly offset to the back and downwards. Ellipsoidal models of the head try to account for the resulting ITD asymmetries [14]. Further, by evaluating different ITD extraction methods, authors in [15] concluded that modeled ITDs should also account for asymmetries introduced by *individually* differing morphology. Thirdly, when using BRIRs measured for various head orientations with a HATS (head and torso simulator) and for arbitrary source positions in a real room, one usually does not know the exact angular direction of the direct sound incidence. Therefore, deterministic ITD models cannot be applied at all. All these limitations are inherently avoided by the ITD rescaling approach.

8.3.3 Separating ITD and ILD of BRIRs

In order to be able to manipulate the interaural time differences contained in binaural impulse response data sets, we considered a variety of methods suitable for both extraction and removal of time delays from the binaural datasets. Most of these methods have been developed and discussed in due consideration of HRIRs/HRTFs. It will be shown, that not all of these methods can be equally well applied when dealing with BRIRs.

8.3.3.1 Perception of HRFT Phase Spectra

Since the perception of HRTF phase spectra has been investigated quite well, only a short review will be given here. HRTFs can conveniently be discussed using LTI system theory, where the complex frequency response can be described by its magnitude and two phase terms describing the minimum and the excess phase fraction of the transfer function:

$$H(j\omega) = |H(\omega)| e^{j\varphi_{min}(\omega)} e^{j\varphi_{excess(\omega)}}. \qquad (8\text{-}2)$$

The minimum phase fraction can directly be derived from the magnitude response as its natural logarithm is related to the phase via the Hilbert transform [23]. The excess phase term $\varphi_{excess}(\omega)$ further consists of a frequency dependent allpass section $\varphi_{all}(\omega)$ and a pure delay term $\varphi_{lin}(\omega)$:

$$H(j\omega) = |H(\omega)| e^{j\varphi_{min}(\omega)} e^{j\varphi_{all(\omega)}} e^{j\varphi_{lin(\omega)}}. \qquad (8\text{-}3)$$

In [16] it has been shown that the auditory sensitivity to absolute HRTF phase spectrum is low. The decomposition into a minimum phase transfer function and a pure delay derived from $\varphi_{lin}(\omega)$ – thereby neglecting the allpass term $\varphi_{all}(\omega)$ – seemed feasible without compromising spatial perception too much and has since then been widely applied [cf. (8-4)]:

$$H_{ml}(j\omega) = |H(\omega)| e^{j\varphi_{min}(\omega)} e^{j\varphi_{lin(\omega)}}. \qquad (8\text{-}4)$$

Hence, the ITD can be calculated as the difference of the left and right ear's linear phase terms resp. group delays and reintroduced as frequency independent pure delay to the left or right audio channel. Research further revealed that maximum deviation between empirical HRTFs and their minimum phase reconstructions occurs at contralateral angles and for sound incidents from low elevations [16]. A listening test proved audibility only for extreme lateral directions of sound incidence [16], a fact that was confirmed in [17] and [18]. In [17] it was shown that allpass components of up to 30 µs were inaudible, but for some directions of sound incidence, even higher phase lags were observed [18]. Therefore, the audibility of all pass components of HRTFs was formally examined in [18]. Empirically found allpass sections in HRTFs were nearly constant up to 3 kHz, therefore it was assumed, that they could be replaced with the corresponding group delay. In a listening test, minimum phase HRTFs whose allpass fractions had been replaced by the corresponding constant group delay rounded to integer samples[5] and added to

[5] At a sampling rate of 48 kHz allpass fractions did never exceed three samples.

the overall linear phase fraction (i.e. the pure delay term) were compared to the original empirical HRTFs. In this case, none of the subjects could reliably discriminate between manipulated and original HRTFs even at critical directions of sound incidence. It was thus concluded that HRTFs can – for all directions of sound incidence – be sufficiently well represented by a minimum phase transfer function and an ITD corresponding to the group delay difference of the excess phase evaluated at 0 Hz (IGD_0) rounded to integer samples.

8.3.3.2 The Special Case of BRIRs

In order to widen the scope of this discussion towards BRIRs it has to be questioned whether the sample-precise IGD_0 is an appropriate measure for detecting and extracting the ITD from BRIRs for the purpose of manipulation. A BRIR can be regarded as a superposition of numerous acoustic reflections each weighted by a distinct head related transfer function and delayed by its arrival time. Therefore, on the one hand, the IGD_0 would characterize the direct sound's direction only, while being meaningless for all following reflections each producing a distinct ITD/IGD_0. On the other hand, the direct sound's ITD will dominate the perception of direction, thus, when aiming at correcting the individual perception of direction it could be sufficient to manipulate the direct sounds arrival times. Therefore, existing methods for the determination of the delay in impulse responses and the calculation of the ITD respectively the IGD_0 will shortly be discussed, with a special focus on their suitability to separate ITD and ILD, i.e. the pure delay term from the magnitude spectrum, in a way that the BRIR can be reconstructed without artifacts later on.

8.3.3.3 Applicability of ITD Extraction Methods

Most approaches for detecting the ITD in binaural impulse responses have been developed for the purpose of ITD estimation and to a lesser extent for a decomposition of ITD and ILD. In [19] about seven methods for the estimation of the ITD were reviewed and evaluated. There, special attention was given to the fact whether the examined methods were able to account for the audible all pass components in HRTFs. Further, approaches using Hilbert transform based decomposition ([18], [20]) and, more recently, improved versions of the cross correlation method were proposed and evaluated ([15], [21]). As these methods are being described thoroughly in the cited references, they will be given here in tabular form only (Table 8-1).

Of these methods, only two directly lead to a separation of minimum phase spectrum (ILD) and pure delay (ITD). The first one is described in [18], where "the first nonzero value [of an HRIR] is identified by visual inspection". These leading zeros are then treated as the pure delay, whereas the remaining impulse response can be further decomposed into minimum and all pass phase spectrum. For handling of larger datasets, the visual approach can easily be replaced by an automatic onset (or "leading edge") detection algorithm [22]. The second approach uses the Hilbert transform for the extraction of the complete excess phase and the separation of a minimum phase spectrum. As the onset detection – without further processing – cannot extract all pass terms, Hilbert transform based decomposition is the only method that directly separates the complete excess phase (IGD_0) from the minimum phase spectrum. All other methods listed in Table 8-1 can only indirectly be used to separate delay and magnitude spectrum, for instance by simply 'cutting' the determined delay values – after being rounded to integer samples – from the begin of the impulse responses. Moreover, especially for those ITD estimators truly extracting IGD_0 this will lead to another problem: If significant all-pass sections are found, a direct extraction of the pure delay could result in cutting into the direct sounds rising edge by up to three samples (see above).

With regard to empirical impulse response data sets as measured with our HATS FABIAN [2], two more problems related to common ITD detection algorithms became obvious. At first, BRIRs can be quite long. Considering the reverberation times of concert halls, filter lengths of 2^{16} to 2^{18} samples are quite usual. Therefore, the first two methods in Table 8-1 are not applicable as the exhibit strong demands on calculation and are prone to rounding errors [18]. Secondly, binaural room impulse responses are usually highpass filtered to some degree, e.g., due to the limited frequency response of the measurement loudspeakers or due to the limited frequency range of the measurement signals (i.e. from 50 Hz to $fs/2$). Furthermore, DC blocking of audio converter inputs can make the determination of IGD_0 impossible [21]. For those reasons, several of the proposed methods are inapplicable, especially those being designed to derive the ITD from direct determination of IDG_0 (methods 1–4 in Table 8-1).

A commonly encountered problem arises with freely available HRTF databases (e.g., CIPIC [24] or IRCAM [25]): When examining the group delay of the excess phase of some selected HRTFs, it is sometimes impossible to determine valid IDG_0 values. According to [18] this could be due to HRTFs not being created by spectral division with the reference spectrum measured at 'the middle of the head'. This for instance applies to our own BRIR data and HRTFs from the CIPIC data base, but not to the raw IRCAM data[6].

Table 8-1. Overview of ITD extraction methods

no.	short description	ref.
1.	Accumulation of the group delay of first and second order all pass sections	[19]
2.	Group delay of the complete all pass component.	[19]
3.	Centroid of impulse response	[19], [26]
4.	Group delay from gradient of the excess phase	[19]
5.	Linear curve fitting of excess phase	[19], [20]
6.	Maximum of interaural cross correlation function	[19], [27]
7	Maximum of interaural cross correlation function of rectified impulse responses	[15]
8.1	Cross correlation maximization	[21]
8.2	Least squares minimization	[21]
8.3	Weighted excess group delay	[21]
9.	Onset or leading edge detection	[19], [22]
10.	Hilbert transform based decomposition.	[18], [20]

The solution proposed in [18], to read off any group delay value above roll-off until 1500 Hz did not solve the problem, as deviations in this frequency range where often much higher than the desired ±30 µs accuracy. We have not yet systematically examined method 5 and methods 7 to 8.3 in Table 8-1 as they do not lead to a separation of ITD and ILD. With the classical cross correlation approach (method 6), we encountered similar problems as reported in [19]. For lateral sound source positions, due to shadowing at the ipsilateral ear, there is little correlation of impulse responses, giving rise to large errors. It is assumed that similar problems will affect method 7. Besides, the new correlation methods (methods 8.1–8.3 , [21]) were shown to overcome this problem by calculating the ITD as the difference of the positions of the maxima from cross correlation of the left and right ears minimum phase impulse responses with the accompanying original impulse responses. However, regarding binaural room impulse responses we found minimum

[6] However, the measurement loudspeaker used at IRCAM exhibited a steep roll off below ca. 60 Hz

and original phase impulse responses to be much less similar than those obtained with HRIRs.

8.3.3.4 Perceptual Evaluation of ITD Extraction Methods

Only Hilbert transform based decomposition is able to decompose a BRIR into its excess and minimum phase spectrum. Due to the special case of the BRIRs being a superposition of numerous delayed HRIRs (see section 8.3.3.2), the extraction of the complete excess phase could be meaningless. Moreover, it was assumed that the impulse compaction characteristic of Hilbert transform could lead to changes in the BRIR's time structure, audibly impairing the individual reflection structure of the room.

Therefore, at first, a listening test was conducted to assess the audibility of differences in the time structure of the reflections due to Hilbert transformation of BRIRs. Using an ABX test design, ten subjects had to discriminate the minimum phase BRIRs from the original ones. As a deviation of time structure was tested, care was taken to make sure that no localization cues, which could be used to discriminate between minimum phase versions and original BRIRs, remained. As presumably most challenging stimuli three singular BRIRs for frontal sound incidence were selected from three rooms (small, medium, large volume) with different reverberation times (1.2, 1.8, 2 s). Although the selected BRIRs already exhibited nearly no interaural time difference, it was made sure by visual inspection that all localization cues due to pure delays were removed. To avoid folding back of artifacts from Hilbert transform into the BRIR's reverberant tail, zero padding towards double length was applied. Minimum phase BRIRs were constructed from the delay-free original BRIRs using *rceps* function of Matlab®. Afterwards, the zero-padded part containing artifacts was discarded. A short drum sample was used as stimulus. Each participant had to listen 14 times to stimuli of each of the three rooms resulting in 42 decisions per subject. Thus, the H_0 hypothesis stating that a difference was inaudible (corresponding to a detection rate of 50%), could be tested for an effect size of 24% (i.e. a detection rate of 74%) per subject on a 5% significance level with a test power of 95% ([28], [29]). Results are shown in Figure 8-1.

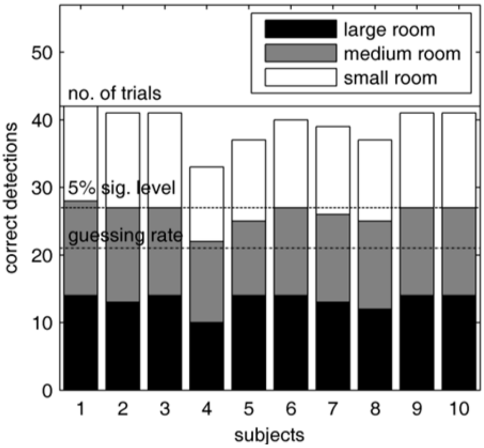


Figure 8-1. Results from ABX listening test for audibility of artifacts from Hilbert transform based decomposition.

The H_0 could be rejected if at least 27 of the 42 decisions were correct. It can be seen that for all participants the difference was obvious. For half of the subjects the detection rate was between 97.5% and 100% and it never fell below 78.5% (subject 4). Subjects described the artifacts as a kind of 'contrast reinforcement' where impulsive sound appeared to be more accentuated or 'raised' over background reverberation. Additionally, the impression of sound source distance was altered in a content specific manner, e.g., the hit of a snare drum appeared to be closer with the minimum phase BRIRs.

Hilbert transform based decomposition was therefore abandoned and onset detection was further examined. In [22] the start of the impulse response is detected by finding the sample where the impulse response, for the first time, exceeds 5% of the maximum value. Subtraction of left and right ears value then gives the ITD. We slightly adapted this procedure, defining the threshold on a log-energy scale. Therefore, the time signal x is computed to be (see also Figure 8-2, upper plot):

$$x(n) = 20log_{10}(|x(n)|). \tag{8-5}$$

The onset threshold is set to a suitable value relative to the peak amplitudes. In Figure 8-2, two HRIRs from a critical lateral position are shown. The threshold value has to be set below the worst-case SNR. For Figure 8-2 it was 20 dB relative to the maximum, and originating from the contralateral HRIR. Onset detection was conducted in the ten times upsampled domain, resulting in a precision of 2.2 µs at a sampling rate of 44.1 kHz, which is meant to guarantee a subliminal quantization of the ITD. An initially visually verified threshold value is then used throughout for automatic onset detection within the complete dataset. In order to avoid cutting into the rising edge of impulse responses exhibiting larger SNRs (i.e. the ipsilateral impulse responses) an overall margin is subtracted from the detected onset position, which was five samples in the case of Figure 8-2. The quasi-minimum phase impulse responses and ITD values were stored in floating point format for further use.

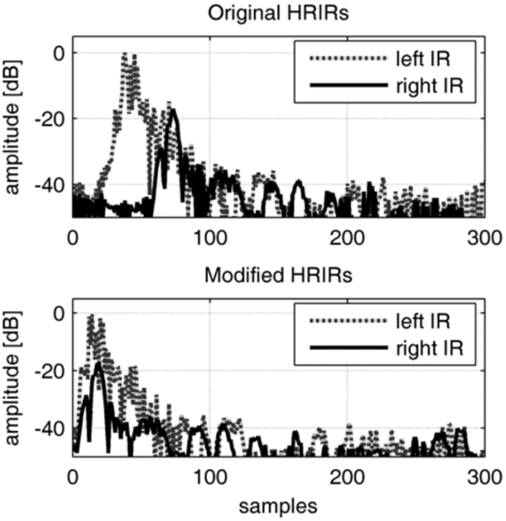


Figure 8-2. Example for the extraction of times of flight from HRIRs via onset detection. Upper plot: HRIRs before onset-based extraction of times of flight. Lower plot: HRIRs after extracting the time of flight individually per ear using a 20 dB threshold criterion and while adding an additional *post hoc* pre-delay of 5 samples (see text).

The described procedure is quite robust even at lateral angles, where other approaches are often affected negatively due to low SNR and high-frequency attenuation of the contralateral impulse response (see also [19]). However, detection results are to some extent depending on the choice of the specific value used as onset threshold. When extracting ITDs from several BRIR sets using method 8.1 of Table 8-1 (incl. upsampling) for informal comparisons, we retrieved very similar ITD functions over angle while observing a slightly higher susceptibility to the lateral angle problematic (more pronounced outliers).

The onset procedure was perceptually evaluated in another listening test. This time, the question was whether the onset-based de- and subsequent re-composition of BRIRs was prone to audible artifacts. As most critical stimuli, several (anechoic) HRIRs were selected from extreme lateral directions of sound incidence from the IRCAM HRTF database. The positions [azimuth, elevation] were: [90°, 0°], [90°, 45°], [90°, -45°], [-90°, 0°], [-90°, 45°], [-90°, -45°], [45°, 45°], [-45°, 45°]. Pure delays were extracted via the described onset technique and introduced again using 10-fold upsampling throughout. Thus, this listening test tested the audibility of (a) applying an up- and down sampling algorithm (Matlabs® *interp* function) and (b) the results of interpolation occurring twice after subsample delay extraction and re-insertion (see small differences between upper and lower plot signals in Figure 8-2 for an example). Again, ten subjects conducted an ABX listening test. This time all eight manipulated HRTFs were tested three times with two different audio stimuli (white noise bursts, male speech) resulting in 48 decisions per subject. While keeping type I and II error levels and the effect size as before, the H_0 could now be rejected if at least 31 of the 48 decisions were correct. Results are shown in Figure 8-3. As expected, no subject could reliably discriminate the reconstructed HRIRs from the original ones.

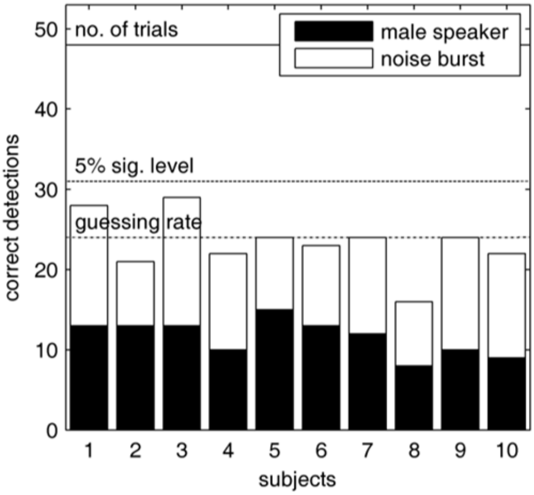


Figure 8-3. Results from ABX listening test for audibility of artifacts stemming from onset detection based de- and re-composition of HRIRs.

Due to its applicability on empirical binaural room impulse response data sets, its ability to deliver a quasi-minimum phase BRIR and a subsample-precise ITD, its robustness, and its perceptually artifact-free results, the onset detection method was selected for further ITD manipulation.

8.3.4 Adjusting the ITD in real time

The individualized binaural resynthesis process was implemented as an extension of the time-variant fast convolution algorithm conventionally used in dynamic binaural synthesis. Figure 8-4 shows a schematic depiction of the implementation. The conventional convolution procedure is split up allowing an independent processing of magnitude and phase information. Thus, the dynamic convolution algorithm is fed the quasi-minimum phase binaural datasets generated by using onset detection as described. Because the crossfade between BRIRs of different head orientations is now conducted on time-aligned signals, characteristic comb filtering artifacts that were quite audible with head movements are largely removed. This results in a clearly audible improvement of the binaural rendering. The output from the fast convolution (signals L’ and R’ in Figure 8-4) is then subjected to a variable delay line (VDL).

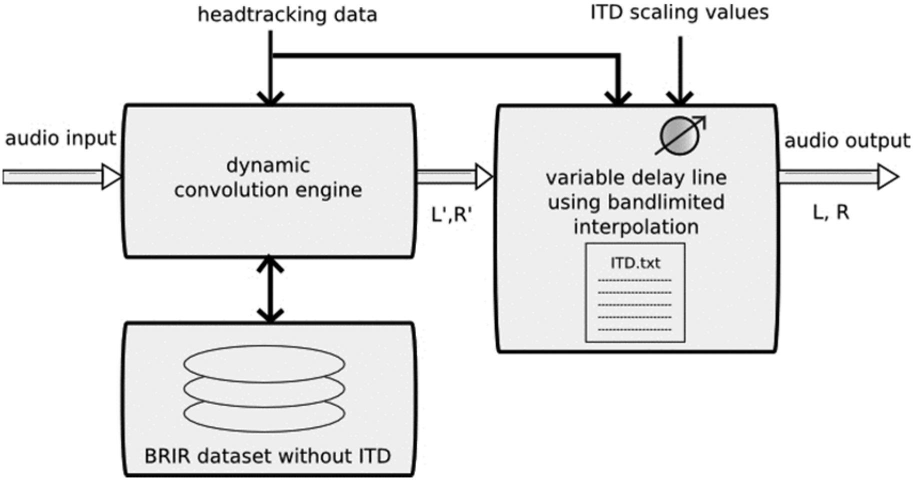


Figure 8-4. Schematic depiction of the signal flow in dynamic binaural synthesis with real time individualization of the ITD.

The previously extracted time delay can be re-inserted with subsample accuracy according to the current head position, thereby re-establishing the ITD between left and right ear signals. Subsample accuracy is guaranteed by means of a band limited interpolation (i.e. fractional resampling) method [30]. For implementation, an open-source sample rate converter library [31] was used. It allows for glitch-free time stretching while maintaining a bandwidth of 97% and a signal to noise ratio of

97 dB. Moreover, as changes of the ITD are realized by fractional sample rate conversion, the Doppler Effect is correctly imitated for the direct sound.

8.4 Listening Test

The described customizable rendering method was used in a listening test, in order to determine individually adequate scaling factors of a foreign dataset's ITD for different subjects. A set of BRIRs was measured in an acoustically damped room (V = 155 m^3, RT = 0.47 s) using the HATS FABIAN. ITDs were extracted using the onset procedure described in section 8.3.3.4, quasi-minimum phase BRIRs were stored. A Genelec 1030A loudspeaker was positioned frontally at a distance of 2 m. BRIRs were measured for horizontal head rotations within ±90° in angular steps of 1°. To allow for an undisturbed direct comparison between the real sound source and its individually adjustable auralization, the HATS wore acoustically transparent headphones during the measurement.

In the listening test, subjects were seated at the position of the dummy head. Using the method of adjustment, the subject's task was – while instantly switching between simulation and reality – to adjust the foreign ITD by interactively changing a multiplicative scaling factor until localization and source stability was perceived to be similar for reality and simulation. During training subjects were instructed to rotate the head widely, as audibility of artifacts due to misaligned interaural delay is maximized at head positions with large ITDs. Eleven subjects (one female, ten male) took part in the test. Their average age was 28 years. The ITD could be modified in a range of 0-200% at a resolution of 1% using up-down pushbuttons as interface. To minimize the impact of concurrent ILD, low-pass filtered white noise bursts were used as stimulus (f_{stop} = 1.5 kHz). Each subject conducted ten runs starting from randomized ITD scaling-factors each time. For the adjustment, participants could take their time at will.

8.5 Results

Despite training, the task turned out to be difficult for some of the subjects. This is also reflected by the rather large confidence intervals in Figure 8-5. By means of residual analysis and outlier tests, two of the eleven subjects had to be excluded from the final analysis.

When applying our approach later on, individually correct scaling factors could be established in lengthy manual adjustment procedures. A generic prediction would therefore be more convenient. Following the approach in [7], for establishing a functional relation between the head's proportions and the ITD scaling factor, four anthropometric measures were taken from each subject: width, height and depth of head and the intertragus distance, which is the distance between both ears' incisura anterior marking the tragus' upper end. In addition to [7] this new measure was chosen due to the tragus' proximity to the ear channel and its simple and reliable determination.

Individual scaling values were predicted from these anatomical measures by means of multiple regression analysis. Explained variance (adj. R^2) indicated that a single predictor – the intertragus distance – was best suited for predicting the individual ITD scaling factors. In this case, the explained variance was about 70.3 %. From Figure 8-5 it can be seen, that the individual ITD scaling factors show a clear and linearly increasing relation to the head diameter described by the intertragus distance. Figure 8-5 depicts all nine mean individual scaling values together with their 95% confidence intervals (CI). The linear regression model is also shown with 95% CIs. The regression formula derived for the individual scaling factor S was:

$$S = 0.00304 d_i + 0.5792 \tag{8-6}$$

with the intertragus distance d_i, specified in millimeters. It has to be emphasized though, that this model is valid only for binaural datasets acquired with our HATS FABIAN. The model could possibly be generalized to arbitrary binaural data by scaling the predicted values by a certain ratio of the intertragus distances of the foreign and our artificial head. However, this approach has not been evaluated so far. A formal evaluation of the achieved localization improvement is subject to future work, too.

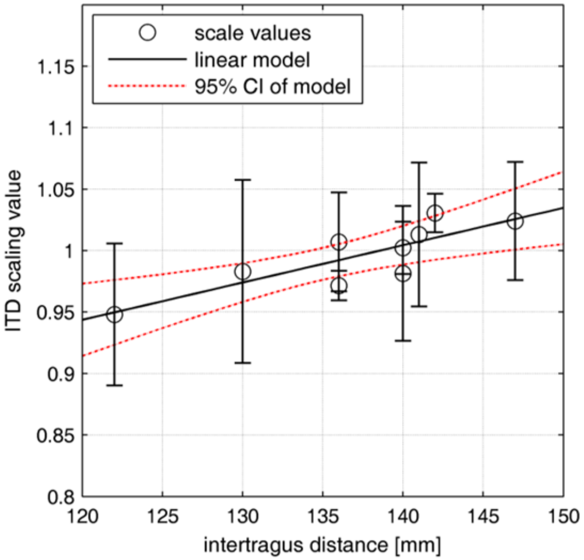


Figure 8-5. Modeling of listening test results: Mean individual ITD scale values plotted over intertragus distance with 95% CIs. Linear regression model is shown with hyperbolic 95% CIs.

8.6 Conclusions

A method for the individual customization of dynamic binaural reproduction by means of real time manipulation of the ITD was proposed. The suitability of different approaches towards the de- and re-composition of BRIRs into ITD and ILD cues was discussed in detail. Promising approaches such as Hilbert transform based decomposition or onset detection were perceptually evaluated. Furthermore, an anthropometry-based prediction model for an individual ITD correction factor was suggested. The presented approach exhibits further advantages for real time binaural resynthesis: Besides stabilization of localization, the elimination of cross fade comb filtering by using quasi-minimum phase audio signals is most obvious. Moreover, the separation of magnitude spectrum and phase creates the possibility of using different spatial resolution and interpolation methods for both temporal and spectral cues. At last, a correct simulation of the Doppler Effect for the direct sound of virtual sound sources by means of sample rate conversion is obtained. In sum, overall perceptual quality of dynamic binaural synthesis could be noticeably improved.

8.7 Acknowledgements

Alexander Lindau was supported by a grant from the Deutsche Telekom Laboratories and by the Deutsche Forschungsgemeinschaft (DFG, grant WE 4057/1-1).

8.8 References

[1] Møller, H. (1992): "Fundamentals of binaural technology", in: *Applied Acoustics*, **36**(3-4), pp. 171-218

[2] Lindau, A.; Hohn, T.; Weinzierl, S. (2007): "Binaural resynthesis for comparative studies of acoustical environments", in: *Proc. of the 122nd AES Conv.*, Vienna, preprint no. 7032

[3] Strutt, J. W. (Lord Rayleigh) (1907): "On Our Perception of Sound Direction", in: *Philosophy Magazine*, **13**, pp. 214–232

[4] Kuhn, G.F. (1977): "Model for the interaural time differences in the azimuthal plane", in: *J. Acoust. Soc. Am.*, **62**(1), pp. 157-167

[5] Blauert, J. (1997): *Spatial Hearing. The Psychophysics of Human Sound Localization,* 2nd ed., Cambridge, MA.: MIT Press

[6] Wightman, F.L.; 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

[7] Algazi, V. R. et al (2001): "Estimation of a spherical-head model from anthropometry", in: *J. Aud. Eng. Soc.*, **49**(6), pp. 472-479

[8] Jot, J.-M.; Walsh, M.; Philip, A. (2006): "Binaural Simulation of Complex Acoustic Scenes for Interactive Audio", in: *Proc. of the 121st AES Conv.*,. San Francisco, preprint no. 6950

[9] Wenzel, E. M.; Foster, S.H. (1993): "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis", in: *Proc. of the IEEE WASPAA 1993 (Workshop on Applications of Signal Processing to Audio and Acoustics)*, pp. 102-105

[10] Müller-Tomfelde, C. (2001): "Time varying Filters in non-uniform Block Convolution", in: *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*. Limerick

[11] Woodworth, R. S.; Schlosberg, H. (1962): *Experimental Psychology*, New York: Holt, Rinehard and Winston

[12] Larcher, V.; Jot, J.-M. (1997): "Techniques d'interpolation de filtres audionumériques Application à la reproduction spatiale des sons sur écouteurs", in: *Proc. of the Congrès Français d'Acoustique*. Marseille

[13] Savioja, L. et al. (1999): “Creating interactive virtual acoustic environments”, in: *J. Audio Eng. Soc.*, **47**(9), pp. 675-705

[14] Duda, R. O.; Avendano C.; Algazi, V. R. (1999): “An Adaptable Ellipsoidal Head Model for the Interaural Time Difference”, in: *Proc. of the IEEE ICASSP 1999 (International Conference on Acoustics, Speech, and Signal Processing)*, Phoenix, AZ, pp. II 965-968

[15] Busson, S.; Rozenn, N.; Katz, B.F.G. (2005): “Subjective investigations of the interaural time difference in the horizontal plane”, in: *Proc. of the 118th AES Conv.*, Barcelona, preprint no. 6324

[16] Kulkarni, A.; Isabelle, S. K.; Colburn, H. S. (1999): “Sensitivity of human subjects to head-related transfer-function phase spectra”, in: *J. Acoust. Soc. Am.*, **105**(5), pp. 2821-2840

[17] Minnaar, P.et al. (1999): “Audibility of All-Pass Components in Binaural Synthesis”, in: *Proc. of the 106th AES Conv.*, Munich, preprint no. 4911

[18] Plogsties, J. et al. (2000): “Audibility of All-Pass Components in Head-Related Transfer Functions”, in: *Proc. of the 108th AES Conv.*, Paris, preprint no. 5132

[19] Minaar, P. et al. (2000): "The Interaural time difference in binaural synthesis." In: Proc. of the. 108th AES Conv., Paris, preprint no. 5133

[20] Jot, J.-M.; Larcher, V.; Warusfel, O. (1995): “Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony”, in: *Proc. of the 98th AES Conv.,* Paris, preprint no. 3980

[21] Nam, J.; Abel, J.S.; Smith III, J.O. (2008): “A Method for Estimating Interaural Time Difference for Binaural Synthesis”, in: *Proc. of the 125th AES Conv.*, San Francisco, preprint no. 7612

[22] Sandvad, J.; Hammershøi, D. (1994): “Binaural Auralization. Comparison of FIR and IIR Filter Representation of HIRs”, in: *Proc. of the 96th AES Conv.*, Amsterdam, preprint no. 3862

[23] Preis, D. (1982): “Phase Distortion and Phase Equalization in Audio Signal Processing”, in: *J. Audio Eng. Soc.*, **30**(11), pp. 774-794

[24] Algazi, V. R. et al. (2001): “The CIPIC HRTF Database”, in: *Proc. of the IEEE WASPAA 2001 (Workshop on Applications of Signal Processing to Audio and Acoustics)*, New York, pp. 99-102

[25] IRCAM, Room Acoustics Team: *Listen HRTF Database*, http://recherche.ircam.fr/equipes/salles/listen/system_protocol.html, 2010, last visited Dec. 5th 2013

[26] Georgopoulos, V.C.; Preis, D. (1998): "A Comparison of Computational Methods for Instantaneous Frequency and Group Delay of Discrete-Time Signals", in: *J. Audio Eng. Soc.*, Vol. **46**(3), pp. 152-163

[27] Kistler, D.J.; 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

[28] Leventhal, L. (1986): "Type I and Type 2 Errors in the Statistical Analysis of Listening Tests", in: *J. Audio Eng. Soc.*, **34**(6), pp. 437-453

[29] Burstein, H. (1988): "Approximation Formulas for Error Risk and Sample Size in ABX Testing", in: *J. Audio Eng. Soc.*, **36**(11), pp. 879-883

[30] Smith, J. O.: "Digital Audio Resampling Home Page", Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, [2010-02-03], http://www-ccrma.stanford.edu/~jos/resample, last visited Dec. 5th 2013

[31] de Castro Lopo, E.: "Secret Rabbit Code", http://www.mega-nerd.com/SRC/index.html, last visited Dec. 5th 2013

9 The Perception of System Latency in Dynamic Binaural Synthesis

The following chapter is an authorized reprint of the abstract-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander (2009): "The Perception of System Latency in Dynamic Binaural Synthesis", in: *Proc. of the 35th NAG/DAGA*, *International Conference on Acoustics (in: Fortschritte der Akustik)*. Rotterdam, pp. 1063-1066.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, typographic and stylistic corrections.

9.1 Motivation

In an interactive virtual acoustic environment the total system latency (TSL) plays an important role for the authenticity of simulations, be it a purely acoustic [1] or an audiovisual simulation [2]. Knowledge about thresholds of just detectable latency will allow for adequate adjustment of the rendering effort. Moreover, headroom available for additional audio processing will be determined. Most former studies examined latency by means of auralization based on anechoic head related transfer functions (HRTFs). Thus, as no reliable thresholds exist for the binaural simulation of reverberant acoustic environments this study was conducted for different acoustic environments, and while using different stimuli in a criterion free adaptive psychoacoustic procedure.

9.2 Latency in VAEs

Early studies on latency in virtual acoustic environments (VAEs) did not directly evaluate the detectability of latency but measured localization accuracy or user response times as a function of altered TSL [3]–[5]. However, as localization accuracy was shown to be barely degraded by latencies as high as 96 [3], 150 [4], or even 250 ms [5], it can hardly be regarded as a good predictor for the detectability of system latency. In localization tasks, latency mainly increases the response times of the subjects [4], [6]. So, some of the differences in values from cited studies ([5], [6]) were suspected to be related to limited stimulus duration. Only recently, the minimum detectable latency in VAEs was directly investigated by different authors [1], [7], [8]. An overview of these results is given in Table 9-1.

Table 9-1. Results from recent studies on just audible latency in VAEs (*minimum observed threshold, **method of determination unmentioned, mTSL = minimum total system latency realized in the respective test).

Ref.	Test paradigm	Stimulus	Subjects	mTSL	Threshold*
[1]	Yes/No	noise	9	12 ms	60 ms
[7]	2AFC	castanets	17	50 ms**	75 ms
[8]	paired comp.	multitone	9	9.9 ms**	70 ms

Alternatively, in [9] it was suggested to deduce minimum detectable latency from psychoacoustic results on the minimum audible movement angle (MAMA, [10]). This quantity describes the angle a moving source has to cover if it is to be detected as different from a stationary source. MAMAs have been found to increase (slowly) with source velocity [10] and decrease with audio stimuli bandwidth [11]. Inferring that MAMAs hold true also for a moving receiver (i.e. a rotating head) and a stationary (virtual) source, minimum latency can be calculated from MAMA. Thus, a VAE with

$$TSL < \frac{MAMA}{v_{head}} \text{ (in [s])} \qquad (9\text{-}1)$$

should be able to render stable virtual sources, i.e. yield an inaudibly low latency. MAMAs as low as 5.9°, 8.2°, and 9.6° as reported in [10] for source velocities of 90°/s, 180°/s, and 360°/s would thus demand system latencies of 65.6 ms, 45.6 ms, and 26.7 ms resp. To examine the predicted interrelation between head movements and latencies all head tracking data was recorded in this study.

Following [5], latency and update rate are distinct, but practically related parameters of VAEs. Total system latency is thus defined as the temporal delay between an event such as a distinct head movement and the corresponding reaction of the VAE, i.e. convolving anechoic audio with updated HRTFs/BRIRs. Update rates are introduced when temporal sampling of the auditory scene happens. This can happen either inside the renderer, which for instance, calculates scene updates for fixed instances of time, or at the input sensing devices, which is most often a head tracker, typically exhibiting an update rate of between 120–180 Hz. Since several elements contribute to the total system latency, VAE response times are typically distributed around a mean and have to be determined empirically [12].

In contrast to applications in augmented reality where visual or auditory real-life cues without latency are concurrently presented within the simulation, a pure audio VAE represents a worst case latency task for subjects [1]. In this case a minimum TSL (mTSL) of at least 30–40 ms should be proven to be obtainable (see Table 9-1).

9.3 Methods

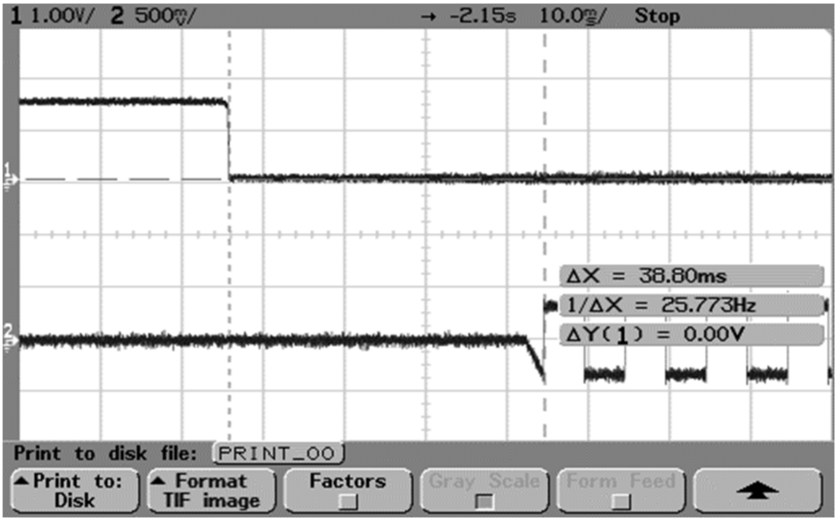
Binaural impulse response datasets for auralization were measured in a large lecture hall (V = 8600 m^3, RT = 2.1 s) and in an anechoic chamber using the automatic FABIAN HATS (head and torso simulator) [13], which is able to move its head freely above the torso. Datasets were measured and auralized for frontal sound incidence (sound source: Meyersound UPL-1) and for a grid of horizontal and vertical head movements (horizontal ±80°, vertical ±35°) with a virtually inaudibly fine angular resolution of 2° x 2° [14]. As acoustic travel times included in field-measured BRIRs, would directly increase latency, these have been discarded from datasets. An onset detection algorithm was used to find the earliest direct sound in each dataset. This delay, reduced by 50 samples for safely preserving the onsets, was then the removed from the datasets.

The used auralization system [13] is a Linux package for fast partitioned convolution. It uses two block sizes, a smaller one for the early part of the BRIRs, and a larger one for the diffuse reverberation tail. Updating the BRIR is done via parallel spectral domain convolution and time-domain crossfading. In order to avoid switching artefacts a short linear cross fade, corresponding to the smaller block size, is used. Thus, the first results of a filter exchange are available one audio block after recognizing a trigger event.

Before adequately operationalizing latency for the purpose of listening tests, the actual minimum TSL has to be determined. As shown in [9], and [12] multiple elements contribute to TSL such as: head tracker update rate, serial port latency, tracker library latency, network latency, scheduling of crossfading in the convolution engine, block size used in convolution engine, and delays introduced by time of flight in BRIR datasets and headphone compensation filters.

In order to realize a minimum system latency, the block size of the fast convolution algorithm was set to the minimum possible value (128 samples) while still preventing dropouts. Frequency response compensation of the STAX SRS 2050II headphones was realized with a frequency domain least squares approach with high pass regularization, whose good perceptual properties were shown in [15]. To eliminate the filter's modeling delay introduced by this method, a minimum phase approach from [16] was used, again adding 20 samples of delay for onset protection.

For head tracking a Polhemus Fastrack device was used. The specified update rate of 120 Hz could be confirmed by measurement with the serial port set to maximum speed of 115 kBaud. A mean message delay of 8.34 ms (1000 measurements, σ = 1.7 ms, min: 3.3 ms, max: 13.5 ms) was observed.



1 1.00V/ 2 500mV/ → -2.15s 10.0ms/ Stop

ΔX = 38.80ms

1/ΔX = 25.773Hz

ΔY(1) = 0.00V

Print to disk file: PRINT_00

Figure 9-1. Screenshot from minimum TSL measurement, upper trace: falling edged indicates a starting tracker movement, lower trace: output of the convolution. After a fade-in of the duration of one audio block size an alternating square wave is rendered.

The minimum TSL was measured using a setup similar to that in [9]. Therefore, the head tracking sensor was attached to a mechanical swing-arm apparatus. When moving the swing-arm from its initial position (comparable to starting a head movement), an electrical circuit was broken, which caused an immediate voltage drop at a simple voltage divider circuit. The convolution engine was set up with identical parameters as in the listening test, yet a specially designed IR-dataset was used, which lead to rendering silence when tracker indicated initial position (0°/0°). Deviations of more than 1° from this position immediately caused a square wave output to be rendered (see Figure 9-1). Due to using this artificial dataset, the 20 +

50 = 70 samples from onset protection included in the listening test's original BRIR datasets had to be added to measurement results to obtain the actual mTSL.

As mentioned, due to the interaction of different contributors' update rates or processing latencies TSL becomes a distribution of values. Figure 9-2 shows 60 measured values from our system. In the graph the missing 70 samples from onset protection within the BRIR data have already been added. The mean mTSL was thus determined to be 43 ms (σ = 3.8 ms, range: 16 ms) which would be just sufficient, as was also confirmed by the listening test.

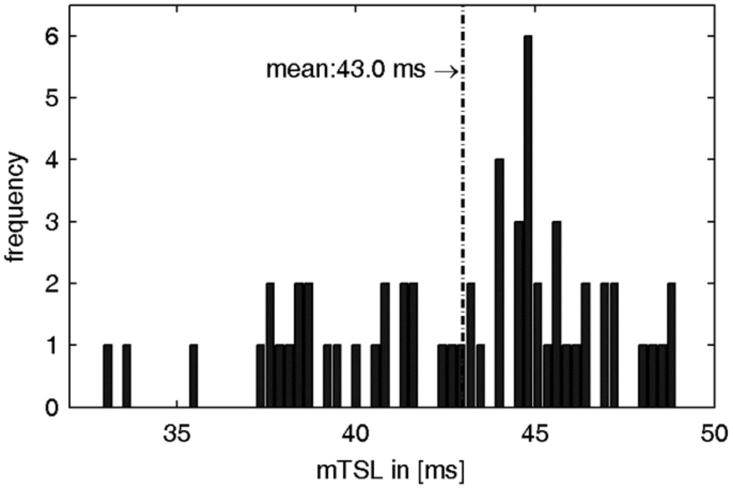


Figure 9-2. Distribution of measured minimum TSL values, the broken line indicates the mean value (resulting true end-to-end latency as realized in the listening test is shown including additional 70 samples of delay from BRIR datasets and headphone compensation).

For operationalizing latency in a listening test, FIFO buffering of the quasi-continuous stream of control data (i.e. head tracking data) should be employed. In contrast, the prolongation of the system response time by dropping tracker update cycles – which is technically easy to implement and was used for instance in [4] and [7] – mixes up delayed response time with reduced spatial resolution and should be avoided. In our case tracker position events, encapsulated in open sound control (OSC) messages, were FIFO cued using the 'pipe' object in the *puredata* software. Measurements proved this method to be able to reliably delay the OSC stream of head tracking data in arbitrary increments of milliseconds.

Stimulus selection was evaluated in pretests (pink noise, narrow band noise, pink noise pulses, acoustic guitar, speech, and castanets). In contrast to [7], castanets did not show to be particularly suited to elicit low latency thresholds. Instead, excerpts from pink noise (≈ 6 s) and male speech (≈ 4.5 s) were chosen as they induced lowest latency thresholds.

For the listening test an adaptive three alternative forced choice (3AFC) test procedure was used. Three stimuli were presented concurrently and without repetition including the reference situation with minimum TSL twice and a stimulus with increased latency once while being randomized a-new for each trial. After an initial training phase including feedback, each run started with a test stimulus taken from the middle of the range of provided latency values. Latency was then adaptively changed according to the subject's responses using a maximum likelihood adaption rule ("Best-PEST", [17]).

Pretests were conducted in order to suitably set up the test parameters. Thus, latency values were presented within a range of [mTSL; mTSL+225ms] and adapted with a step size of 3 ms. A maximum trial number of 20 was set as stopping criterion for the "Best-PEST" adaptation process. Head movements were possible within the BRIR data ranges (controlled for by aural operator guidance). During the training phase, subjects were encouraged to find individual movement strategies that would maximize their detection rate.

Finally, 22 subjects took part in this study (21 male, 1 female, avg. age: 29.8 yrs.); 90% of them had former experience in listening tests and received some musical education.

To test for stimulus and reverberation effects the test was designed as a full factorial 2 x 2 (2 stimuli, 2 acoustic environments) repeated measures test. Hence, all subjects had to evaluate all four possible situations in randomized order, resulting in 4 x 22 = 88 threshold values. The individual test duration was about 45 minutes.

The listening test was implemented in Matlab®, providing user interfaces for training and listening test. The auralization engine, the insertion of latency, and the audio playback were remotely controlled via OSC messages. Head tracking data, received as OSC messages were also recorded by Matlab® with a sampling rate of 50 Hz. The head tracker was reset before each run.

9.4 Results

The distribution of all obtained threshold values is shown in Figure 9-4. Results are depicted for a range starting from mTSL (43 ms). The lowest threshold of 52 ms, which is only 3 increments above mTSL, was observed once for the reverberant/speech condition. For anechoic/speech and reverberant/noise lowest thresholds found were equally 64 ms; for anechoic/noise it was 73 ms. These findings are in good agreement with the figures from in Table 9-1. The largest threshold of 187 ms was also found for the anechoic/noise condition.

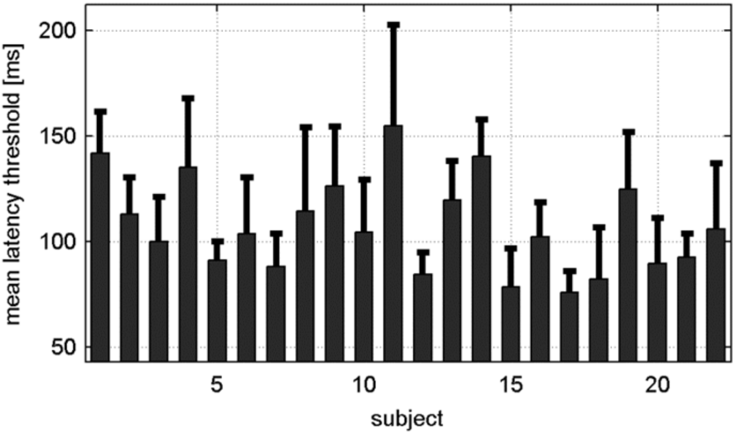


Figure 9-3. Average just detectable latencies per subject with standard deviations, and pooled over all 4 conditions.

Threshold values could be assumed to be normally distributed (Kolmogoroff-Smirnov Test, $p > 0.2$) for all conditions. Means and standard variations were very similar within the four conditions (overall mean: 107.6 ms, σ: 30.4 ms). However, individual thresholds were very different (see also [1], [8], and Figure 9-3). A low reliability value (Cronbach's α: 0.7) supported the latter finding.

Threshold data were analyzed by means of a 2 x 2 univariate ANOVA for repeated measures. The ANOVA showed no effects of stimulus or acoustic environment. Hence, the four runs of each subject were regarded as repetitions and all data were pooled for further analysis.

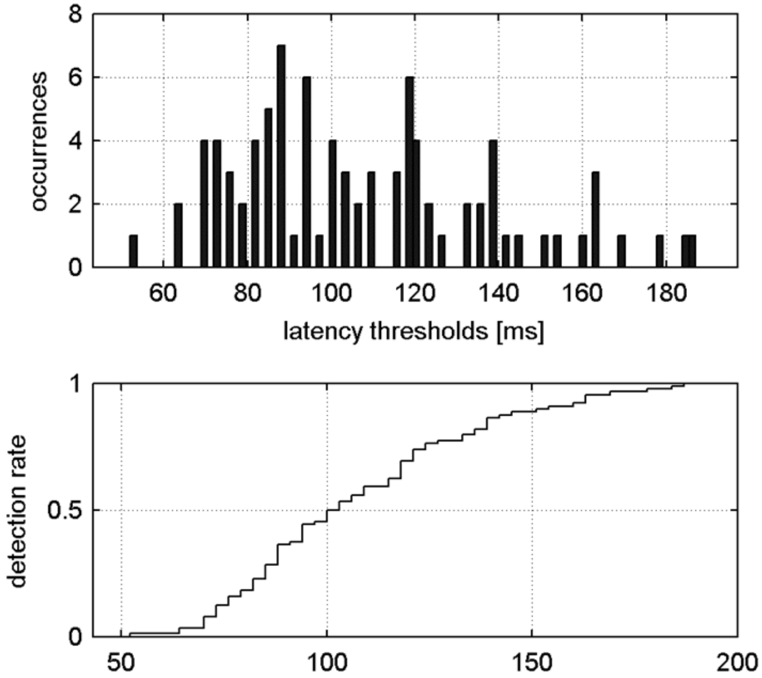


Figure 9-4. Above: Histogram of all latency thresholds. Below: Cumulated detection rate of latency thresholds serving as an estimate of the underlying psychometric function (pooled for all conditions as no related effects could be observed).

To further clarify the missing effects some subjects' comments on their own results shall be citied: Whereas some individuals reported the anechoic/noise situation as being best suited for a critical and undisturbed assessment of latency, others explained that speech in reverberant environment was the most natural stimulus, which made it easy to detect any unnatural behavior within the simulation. From theory (see section 9.2) it was expected to obtain lower thresholds with stimuli of higher spectral bandwidth (i.e. from noise).

As mentioned, head tracking data were recorded for all subjects throughout the listening test whenever a stimulus was present. Horizontal and vertical angular head position was recorded with a sampling frequency of 50 Hz. From head position data acceleration and velocity traces were retrieved, too. Mean horizontal movement ranges were ±26°, mean velocity was 190°/s, and mean acceleration

92°/s² (excluding times were head stood still). Mean vertical movement ranges were ±11°, mean vertical velocity was 20°/s, and mean acceleration 10°/s². Further, histograms revealed different individual strategies in exploiting the movement range.

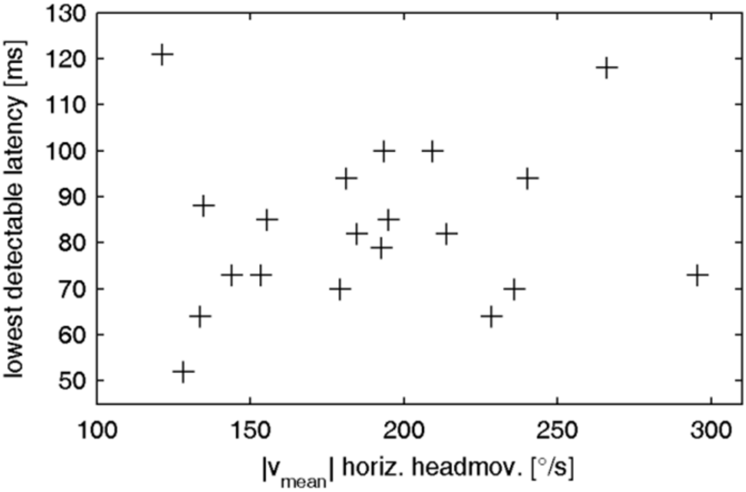


Figure 9-5. Scatter plot of lowest latency threshold reached by individuals vs. mean magnitude of their horizontal head movement velocity in the related run (1 subject omitted due to missing data).

From visual inspection of recorded movement traces and as values for vertical head movements were by far smaller than those for horizontal movements it is assumed, that vertical head position was used more or less as a constant offset and not altered consciously during the test. The maximally observed horizontal head movement velocity was 942°/s, maximum acceleration was 612.5°/s², whereas 428°/s respectively 206°/s² were maximally observed for vertical head movements.

Somewhat surprisingly, the individual latency thresholds showed no linear relation to the observed mean horizontal head movement velocities (see Figure 9-5, the best subject was a 'slow mover'). Indeed, a linear regression showed a best fit to a constant. When analyzing the maximum horizontal head movement velocities the same inconclusive behavior was observed.

9.5 Discussion

Thresholds for the detection of latency in a VAE were determined in a criterion free listening test design, providing a distinct minimum TSL and operationalizing latency with a fine temporal resolution. A minimum threshold of 53 ms was observed once. Mean and standard deviation of pooled threshold values were 107.63 ms resp. 30.39 ms. As normal distribution could be assumed, from these values interval estimates for the test population can be calculated. Thus, the 95% confidence interval of the mean is [101.3 ms; 114 ms]. Only three times – which is 3.4% of all measured thresholds – latencies ≤ 64 ms were found to be detectable. No effect could be observed either for anechoic vs. reverberant environments or for noise vs. speech stimulus. From reported findings on the minimum audible movement angle (MAMA) it was expected that higher bandwidths and faster head movements would lead to lower latency threshold values. The missing stimulus effect is thus somehow unexpected, although it is admitted that a group of two stimuli constitutes no systematic variation of bandwidth. In [1] it was argued, that three subjects with lowest thresholds also showed maximal rotation speeds. Contradictory, from the published data can also be seen, that this behavior was reversed for all 6 remaining subjects. Likewise from our data a relation between mean respectively maximum head movement velocity and latency thresholds could not be found; maybe a different predictor will be better suited. Until then, a VAE is assumed a complex system, and the underlying processes that lead to a perceptibility of latency do not seem to be reducible to stimulus bandwidth and mean or maximum velocity of head movements.

9.6 Acknowledgements

Alexander Lindau was supported by a grant from the Deutsche Telekom Laboratories.

9.7 References

[1] Brungart, D. S.; Simpson, B. D.; Kordik, A. J. (2005): “The detectability of headtracker latency in virtual audio displays”, in: *Proc. of ICAD 2005 - 11th Meeting of the International Conference on Auditory Display*, Limerick

[2] Meehan, M. et al. (2003): “Effect of Latency on Presence in Stressful Virtual Environments”, in: *Proc. of the IEEE Virtual Reality Conference 2003*. Los Angeles, pp. 141

[3] Bronkhorst, A. W. (1995): “Localization of real and virtual sound sources”, in: *J. Acoust. Soc. Am.*, vol. 98, No. 5, pp. 2542-2553

[4] Sandvad, J. (1996): “Dynamic Aspects of Auditory Virtual Environments”, in: *Proc. of the 100th AES Convention*, Kopenhagen, preprint no. 4226

[5] Wenzel, E. M. (2001): “Effects of increasing system latency on localization of virtual sounds”, in: *Proc. of ICAD 2001 - Seventh Meeting of the International Conference on Auditory Display*. Boston

[6] Brungart, D. S. et al. (2004): “The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources”, in: *Proc. of ICAD 2004 - 10th Meeting of the International Conference on Auditory Display*. Sydney

[7] Mackensen, P.: *Auditive Localization. Head Movements, an additional cue in Localization*. Doct. dissertation, Technische Universität Berlin, 2004

[8] Yairi, S.; Iwaya, Y.; Suzuki, Y. (2006): “Investigations of system latency detection threshold of virtual auditory display”, in: *Proc. of ICAD 2006 - 12th Meeting of the International Conference on Auditory Display*. London, pp. 217-222

[9] Wenzel, E. M. (1997): “Analysis of the Role of Update Rate and System Latency in Interactive Virtual Acoustic Environments”, in: *Proc. of the 103rd AES Convention*, New York, preprint no. 4633

[10] Perrott, D. R.; Musicant, A. D. (1977): “Minimum audible movement angle: Binaural localization of moving sound sources”, in: *J. Acoust. Soc. Am.*, **62**(6), pp. 1463-1466

[11] Chandler, D. W.; Grantham, D. W. (1992): “Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity”, in: *J. Acoust. Soc. Am.*, **91**(3), pp. 1624-1636

[12] Miller, J. D. et al. (2003): “Latency measurement of a real-time virtual acoustic environment rendering system”, in: *Proc. of ICAD 2003 - 9th Meeting of the International Conference on Auditory Display*. Boston

[13] Lindau, A., Hohn, T., Weinzierl, S. (2007): “Binaural resynthesis for comparative studies of acoustical environments”, in: *Proc. of the 122nd AES Convention*, Vienna, preprint 7032

[14] Lindau, A.; Maempel, H.-J.; Weinzierl, S. (2008): “Minimum BRIR grid resolution for dynamic binaural synthesis”, in: *Proc. of the Acoustics '08,* Paris, pp. 3851-3856

[15] Schärer, Z.; Lindau, A. (2009): "Evaluation of Equalization Methods for Binaural Signals", in: *Proc. of the 126th AES Convention*, Munich, preprint no. 7721

[16] Norcross, S. G.; Bouchard, M.; Soulodre, G. A. (2006): "Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization", in: *Proc. of the 121st AES Convention,* San Francisco, preprint no. 6929

[17] Pentland, A. (1980): "Maximum likelihood estimation: The best PEST", in: *Perception & Psychophysics*, **28**(4), pp. 377-379

10 Perceptual Evaluation of Model- and Signal-based Predictors of the Mixing Time in Binaural Room Impulse Responses

The following chapter is an authorized reprint of the full-paper peer-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander; Kosanke, Linda; Weinzierl, Stefan (2012): "Perceptual Evaluation of Model- and Signal-based Predictors of the Mixing Time in Binaural Room Impulse Responses", in: *J. Audio Eng. Soc.*, **60**(11), pp. 887-898.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style, typographic and stylistic corrections.

Authors' Note

The first presentation of this study at the 128th AES Convention was awarded a Student Technical Paper Award.

10.1 Abstract

The mixing time of room impulse responses denotes the moment when the diffuse reverberation tail begins. A diffuse ("mixed") sound field can physically be defined by (1) equidistribution of acoustical energy and (2) a uniform acoustical energy flux over the complete solid angle. Accordingly, the perceptual mixing time could be regarded as the moment when the diffuse tail cannot be distinguished from that of any other position or listener's orientation in the room. This, for instance, provides an opportunity for reducing the part of binaural room impulse responses that has to be updated dynamically in Virtual Acoustic Environments. Several authors proposed model- and signal-based estimators for the mixing time in rooms. Our study aims at an evaluation of all measures as predictors of a perceptual mixing time. Therefore, we collected binaural impulse response data sets with an adjustable head and torso simulator for a representative sample of rectangular shaped rooms. Altering the transition time into a homogeneous diffuse tail in real time in

an adaptive, forced-choice listening test, we determined just audible perceptual mixing times. We evaluated the performance of all potential predictors by linear regression and finally obtained formulae to estimate the perceptual mixing time from measured impulse responses or physical properties of the room.

10.2 Introduction

Room impulse responses are usually considered to comprise three successive parts: direct sound, early reflections, and a tail of stochastic reverberation. The transition point between early reflections and the stochastic reverberation tail is often called mixing time (t_m) [1]. Due to increasing reflection density and diffuseness of the decaying sound field the perceptual sensitivity for the temporal and spectral structure of room impulse responses decreases during the decay process, and individual reflections become less and less distinguishable [2]–[4]. Moreover, auditory suppression effects, as, e.g., level-, direction-, and time-dependent simultaneous and nonsimultaneous masking or the precedence-effect further affect the audibility of reverberation fine structure. Further, Olive and Toole [5] pointed out the role of the audio content in discriminating room reflections.

Computational demands for Virtual Acoustic Environments (VAEs) will be reduced with the amount of early reflections to be rendered. A common method to achieve this is to replace the individual reverberation tail of binaural room impulse responses (BRIRs) – after an instant when perceptual discrimination is no longer possible – with an arbitrary and constant reverberation tail. Also for efficient loudspeaker array-based sound field synthesis it is relevant to know how many individual early reflections have to be rendered and when a stochastic tail is sufficient. In the following, this instant will be referred to as the perceptual mixing time t_{mp}.

Already in an early publication on dynamic auralization it was proposed to split the convolution process into a time variant convolution of the early impulse response (IR) parts and a static convolution with an arbitrary reverberation tail [2]. For precalculated BRIRs of a concert hall, in [2], this split point was set after 4000 samples corresponding to a transition point at 83 ms.

In [4], for a lecture hall (V = 420 m^3, RT = 1.0 s) and for less critical audio material, a split point of 80 ms was found sufficient for crossfading into an artificially designed reverberant tail.

For a small room (V = 185 m³, RT = 0.72 s), Meesawat and Hammershøi examined t_{mp} for different combinations of early reflections and diffuse tails [3]. Manipulating the crossfade instances authors determined perceptual mixing times for (a) interchanged tails of the two ears, (b) tails from different receiver positions while keeping the same relative source position, (c) tails from the same receiver position but different horizontal source angles, and finally (d) tails from different receiver and source positions. Stimuli were convolved with accordingly manipulated binaural impulse responses, resulting in static auralization for presentation. For a listening test, the method of constant stimuli was used. Results lead to the conclusion that – for this room – t_{mp} was about 40 ms and independent from all position changes tested.

Since we expected higher mixing times for larger rooms, in a past study [6], we assessed t_{mp} for a large auditorium (V = 8500 m³, RT = 2 s), also using static auralization but an adaptive listening test design. The perceptual mixing time was indeed higher (up to 140 ms). In addition to findings in [3], we found no effect of taking a tail from the same receiver position but for different head orientations. However, the effect of taking the tail from a different source position and a different head orientation (case not tested by Meesawat and Hammershøi, [3]) led to considerably increased perceptual mixing times; potential reasons will be discussed in Section 10.2.2. Additionally, it turned out that listeners were most sensitive when a specific drum set sample with strong transients was used.

The aim of our present study was to find the perceptual mixing times for approximately shoebox shaped rooms of differing volume and average absorption while utilizing state of the art dynamic auralization and an adaptive, forced-choice listening test design. Subsequently, we examined several predictors of the physical mixing time for their ability to predict the perceptual mixing time.

10.2.1 The Concept of Physical Mixing Time

Diffusion and mixing are often used synonymously for the characteristic of sound fields in real enclosures to become more and more stochastic over time.

The transition from early reflections into a stochastic reverberation tail is a gradual one. Every time a sound wave hits a wall, it is reflected. Depending on the surface properties, this reflection can be specular, partly, or fully diffuse. In an ideally diffuse reflecting room, the sound energy continuously spreads over the whole volume in time. Finally, the ideal diffuse sound field is characterized by a uniform angular distribution of sound energy flux and a constant acoustical energy density over the whole space [8]. The process of mixing, on the other hand, is usually illustrated in terms of particle trajectories, when, over time, position and direction of two initially adjacent rays become statistically independent. A requirement for a room to become mixed is ergodicity, which means that – at some point in time – the statistical behavior at all points in the space equals that at one point over time (time average equals ensemble average, [1], [9]), i.e. the sound field has completely "lost any memory" of its initial state.

The duration of the diffusion process, i.e. the physical mixing time, increases with room size as – due to larger free path lengths – the time intervals between individual reflections are increased. This effect is further pronounced if the room is lacking any diffusing obstacles [7].

10.2.2 Real-World Limitations of Complete Mixing

Ergodicity was shown to be dependent on the shape of the enclosure and the surface reflection properties [10]. Examples for non-ergodic rooms are perfectly rectangular non-diffusing rooms (particle directions remain deterministic) or non-diffusing spherical rooms (due to focusing not all positions will be reached by a particle). The process of mixing within the decaying room impulse response may further be disturbed in rooms with non-uniform distribution of large areas with highly varying absorption coefficients (for instance, when large windowpanes are combined with highly absorbing audience seats). As shown by Pollack [1], absorbing rooms can never be perfectly diffuse, because there always remains a net energy flow in the direction of the losses (i.e., toward the absorbing walls). Also coupled rooms, highly damped rooms, and very small rooms may lack mixing in their decay.

Inherently, the whole concept of mixing is further confined to a frequency range where the theory of geometrical and statistical acoustics applies. In real rooms, these assumptions are violated by modal behavior in the low-frequency range. Another problem might arise from proximity to room boundaries (sidewalls, floors). In this case reflections may form distinct comb filters whose spectra depend on the exact receiving position violating the assumption of positional independence of the diffuse sound field. In summary, it can be stated that perfect mixing (or total diffusion) is an idealization never fully encountered in real rooms.

10.2.3 Physical and Perceptual Mixing Time

With reference to the physical definition we propose to say that a room is perceptually mixed, if the stochastic decay process at one position in the enclosure cannot be distinguished from that of any other position and/or listener orientation. Due to auditory and cognitive suppression effects as well as properties of the audio content mentioned above, the perceptual mixing time can expected to be equal or smaller than the physical mixing time, no matter how the latter is determined. An operationalization of mixing time still has to be defined and should take into account the intended application. Below, we will introduce an experimental concept aiming at applications in binaural technology.

10.3 Methods

Section 10.3.1 recapitulates common model-based estimators of the mixing time. Section 10.3.2 gives an overview of four recently proposed signal-based parameters. Section 10.3.3 explains the motivation for selecting the rooms for the listening tests, whereas Section 10.3.4 describes the measurement of binaural room impulse responses for the listening test. Section 10.3.5 gives details about the actual calculation of these parameters and the treatment of practical issues we encountered. Section 10.3.6 explains the listening test, and finally, Section 10.3.7 explains the listener selection procedure.

10.3.1 Model-Based Estimators of Mixing Time

Several estimators of the perceptual mixing time t_{mp} have been suggested in literature. *Ad hoc* values as for instance 50 ms [11], or 80 ms ([2], [12]) have been proposed regardless of further room properties. Other authors suggest time ranges of 100–150 ms [13], 150–200 ms [14], or 50–200 ms [15], to take into account different room properties.

Some theoretically motivated estimators of t_{mp} explicitly refer to properties of the auditory system such as time resolution or being “free from flutter” and assume reflection densities from 250 s^{-1} [17], 400 s^{-1} ([14], [18]), 1000 s^{-1} [19], 2000 s^{-1} [20], 4000 s^{-1} [21] up to 10 000 s^{-1} [22] to be sufficient to render stochastic reverberation tails. Setting the reflection density dN/dt as derived from the mirror source model of the rectangular room

$$\frac{dN}{dt} = \frac{4\pi \cdot c_0^3 \cdot t^2}{V} \qquad (10\text{-}1)$$

(c_0: sound velocity in m/s, V: room volume in m³) to 400 s^{-1}, and solving for t, the following popular estimation of the mixing time was proposed in [16]:

$$t_{mp} \approx \sqrt{V}, \text{ with } t_{mp} \text{ in ms.} \tag{10-2}$$

Rubak and Johansen [21] introduced a different view, relating the instant of perceptual mixing to the concept of the mean free path length l_m:

$$l_m = 4\frac{V}{S}, \tag{10-3}$$

where S is the total surface area of the enclosure in m². The rationale of this approach is that the sound field is assumed to be virtually diffuse, if every sound particle has on average undergone at least some (e.g., [21]: four) reflections. Thus their estimation of t_{mp} (in ms) reads:

$$t_{mp} \approx 4l_m\frac{10^3}{c_0} = 4\cdot\left(\frac{4V}{S}\right)\cdot\frac{10^3}{c_0} \approx 47\cdot\frac{V}{S}. \tag{10-4}$$

Recently, Hidaka et al. [23] proposed a linear regression formula that fits results from a larger study on physical mixing times determined empirically from impulse responses, including 59 concert halls of different shape and size. The formula predicts the mixing time t_{m500Hz} for the 500 Hz octave band (in ms) from the room's reverberation time

$$t_{m500Hz} = 80\cdot RT_{500Hz}, \tag{10-5}$$

Thus, all suggested estimators depend on only three room specific quantities: volume, surface area, and reverberation time. They can therefore further be generalized to

$$t_{mp1} = k_{refl}\cdot\sqrt{V}, \tag{10-6}$$

for the reflection density relation (10-2). The general mean free path length relation similarly reads

$$t_{mp2} = k_{path} \cdot \frac{V}{S}, \tag{10-7}$$

and the general estimation from the reverberation time can be rewritten as

$$t_{mp3} = k_{reverb} \cdot RT_i. \tag{10-8}$$

Thus, three basic model-based relations of the mixing time remain to be subjected to a perceptual evaluation.

10.3.2 Signal-Based Predictors of Physical Mixing Time

Recently, several algorithms were proposed for calculating the physical mixing time from empirical room impulse responses. We included four of these approaches into our evaluation.

Abel and Huang (2006)

Abel and Huang [25] proposed an approach based on the assumption that the sound pressure amplitudes in a reverberant sound field assume a Gaussian distribution. For determining the mixing time, a so-called "echo density profile" is calculated. With a short sliding rectangular window of 500–2000 samples, the empirical standard deviation of the sound pressure amplitudes is calculated for each sample index. In order to determine how well the empirical amplitude distribution approximates a Gaussian behavior, the proportion of samples outside the empirical standard deviation is determined and compared to the proportion expected for a Gaussian distribution. With increasing time and diffusion, this echo density profile should increase until it finally – at the instant of complete diffusion – reaches unity. With larger window sizes, the overall shape of the echo density profile stays similar whereas smoothing of the fine structure can be observed. We chose a rectangular window of 2^{10} samples (23 ms), as suggested by the authors while referring to the auditory temporal resolution. The mixing time can be defined as the instant where the echo density profile becomes unity for the first time (criterion I). In order to account for minor fluctuations Abel and Huang modified this criterion to account for the instant when the reflection density is within $1 - \sigma_{late}$ (σ_{late} being the standard deviation of the late echo density, criterion II). We evaluated both stopping criteria, while calculating σ_{late} from the last 20% of the impulse responses before reaching the noise floor.

Stewart and Sandler (2007)

Following an idea proposed in [25], Stewart and Sandler [26] suggested measuring the kurtosis of the sound pressure amplitudes and comparing this value to that expected for a Gaussian distribution. As a second order cumulant, the kurtosis γ_4 is a measure of the "non-Gaussianess" contained in a signal. In the normalized form, γ_{4n} is given as:

$$\gamma_{4n} = \frac{E\{x-\mu\}^4}{\sigma^4} - 3. \qquad (10\text{-}9)$$

where E is the expectation operator, μ is the mean, and σ is the standard deviation of the process. For increasingly Gaussian-like processes, the normalized kurtosis must approach zero. We calculated this instant with identical settings as for the echo density profile. Although not clearly stated in [26], we concluded from the authors' discussion, that the instant when the normalized kurtosis γ_{4n} reached zero for the first time should assumed to be the mixing time.

Hidaka et al. (2007)

Hidaka et al. [23] proposed a frequency-domain approach for the estimation of the instant when a room impulse response has become diffuse. Therefore, the time-frequency energy distribution of the impulse response $p(t)$ is calculated according to

$$E(t,\omega) = \left|\int_t^\infty p(\tau)e^{j\omega\tau}\, d\tau\right|^2. \qquad (10\text{-}10)$$

When averaging over a frequency range $\Delta\omega$, (10-10) can be shown to be identical to the Schroeder integration [24]. The energy distribution $E(t,\omega)$ is calculated for impulse responses beginning with the direct sound; initial delays are removed in advance. With increasing time t, $E(t,\omega)$ will progressively contain fewer early reflections and more stochastic reverberation. In a second step, Pearson's product-moment correlation $r(t)$ is calculated as a continuous function of time for

$E(0{:}\infty, \Delta\omega)$ and $E(t{:}\infty, \Delta\omega)$ in arbitrary frequency bands. This will describe the similarity between (a) the energy decay process including the initial state and (b) the energy decay process with beginning from any time t afterward in one particular frequency band. Hidaka et al. define the "transition time" into stochastic reverberation as the instant when $r(t) = e^{-1} = 0.368$. Thus, we calculated $E(t, \omega)$ and $r(t)$ for octave bands from 125 Hz to 16 kHz, and detected the mixing time at the moment when $r(t) \leq 0.368$ for the first time. For ease of computation we limited the temporal resolution to 100 samples ($\Delta t = 2.3\ ms$).

Defrance et al. (2009)

Recently, Defrance et al. [27] suggested a new procedure for estimating the physical mixing time from room impulse responses. Their method is based on the assumption that, over time, the reflection density at an observing point in an enclosure becomes so large, that singular reflections begin to overlap and cannot be distinguished anymore. The authors propose a technique ("Matching Pursuit") somewhat similar to wavelet decomposition to decompose room impulse responses into singular reflections (called "arrivals"). As a result they obtain a function of the cumulative number of arrivals, which – as can be derived from time integration of (10-1) – should show a cubic increase. Following the authors' argumentation the decomposition process should more and more fail to distinguish superimposed reflections resulting in a slope changing from cubic to linear. The instant of the changing slopes would then equal the physical mixing time.

Considering all reflections to be more or less copies of the direct sound impulse only the direct sound itself is used as wavelet in the "Matching Pursuit." Decomposition is conducted by correlating the impulse response with the direct sound while shifting the latter along the impulse response to all possible instances in time. At the instance of maximum correlation, the direct sound is subtracted from the impulse response weighted by the corresponding correlation value. The decomposition process is repeated until the energy ratio SRR (signal residual ratio) of the reconstructed signal (reconstructed from the direct sound and the time vector of correlation coefficients) and remaining impulse response signal (the residuum) rises above a certain value. To avoid a decomposition wrongly favoring the early parts of the impulse response, its energy decay has to be compensated before running the decomposition. Finally, the mixing time is determined by applying a reflection distance criterion on the reconstructed impulse response. Therefore, Defrance et al. argued that the mixing time would be equivalent to the moment were the first two reflections are spaced equal or less than the so-called “equivalent duration” of the direct sound.

Using the software provided by the authors, we were able to calculate the Matching Pursuit decomposition using their original Matlab® code. Additionally, we implemented the energy decay compensation and the calculation of the equivalent duration of the direct sound. According to recommendations in [27] we used a SRR of 5 dB as stopping criterion for all decompositions.

10.3.3 Room Selection

The main purpose of this study was to find reliable predictors for the perceptual mixing times for a broad range of acoustical environments. The physical mixing times derived in [23] for a large selection of concert halls were at maximum for shoebox shaped rooms. From the theory of mixing (cf. Section 10.2.2), their regular shape and their long unobstructed path lengths suggests them to be most critical in terms of late mixing times. Therefore, we confined this study to largely rectangular rooms. Coupled enclosures were avoided. Wall surface materials varied from only little diffusing concrete, glass or gypsum to considerably structured wood panels. Floors were made from linoleum, parquet or granite. The smaller (lecture) rooms were equipped with chairs and tables, whereas all the larger rooms included extended audience seating areas. For the sake of simplicity, we calculated surface area from the three main dimensions of the considered ideal shoebox room, neglecting additional surfaces of galleries or furniture.

We selected nine rooms, aiming at a systematic variation of both volume and average absorption coefficient (α_{avg}), each in three steps (cf. Table 10-1). This so-called complete variation would permit an independent statistical assessment of both influences by means of two-way ANOVA.

Due to physical interrelation, it is difficult to vary room volume independently from absolute amount of reverberation, i.e. from the reverberation time. However, while varying the average absorption coefficient, we could at least assess the influence of the relative amount of reverberation independently of volume. Step sizes of reverberation time were additionally chosen to exceed at least a just noticeable difference of 10%. The most important parameters of the selected rooms are listed in Table 10-1. Additionally, Figure 10-1 shows true to scale floor plans of the rooms.

The three small rooms were, in order of increasing reverberation time, the Electronic Studio of the TU Berlin (room 1), and two small lecture rooms, EN-111 and EN-190 (room 2 and 3). The medium size rooms were the TUB's lecture halls H-104 (room 4), and HE-101 (room 5), and the large recording room of the Teldex Studio Berlin (room 6). The three large venues comprised the concert hall of the University of Arts in Berlin (room 7), the auditorium maximum of the TUB (room 8), and the Jesus-Christus Church in Berlin-Dahlem (room 9). Rooms 4, 6, 7, 8, and 9 are regularly used as musical performance spaces.

Table 10-1. Volume, average absorption coefficient, and reverberation time of the nine selected rooms.

	large α **(RT)**	**med. α** **(RT)**	**small α** **(RT)**	**avg. Vol.**
small V	**room 1** 216 m³ α: 0.36 (0.39 s)	**room 2** 224 m³ α: 0.26 (0.62 s)	**room 3** 182 m³ α: 0.17 (0.79 s)	**207 m³**
medium V	**room 4** 3300 m³ α: 0.28 (1.15 s)	**room 5** 5179 m³ α: 0.23 (1.67 s)	**room 6** 3647 m³ α: 0.2 (1.83 s)	**4042 m³**
large V	**room 7** 8298 m³ α: 0.33 (1.52 s)	**room 8** 8500 m³ α: 0.23 (2.08 s)	**room 9** 7417 m³ α: 0.23 (2.36 s)	**8072 m³**
avg. α **(RT)**	**0.32** (1 s)	**0.24** (1.45 s)	**0.2** (1.66 s)	

10.3.4 Binaural Measurements

In order to provide high quality dynamic binaural simulations of all environments, we measured binaural room impulse responses in all nine rooms using the automatic head and torso simulator FABIAN [6].

As sound source, a 3-way dodecahedron loudspeaker, providing high signal to noise ratio and optimal omnidirectional directivity, was placed in the middle of the stage, which was typically located at one of the narrow ends of the rooms. To obtain a wide frequency range for the BRIRs, the loudspeaker was equalized to a linear frequency response within ± 3 dB from 40 Hz to 17 kHz.

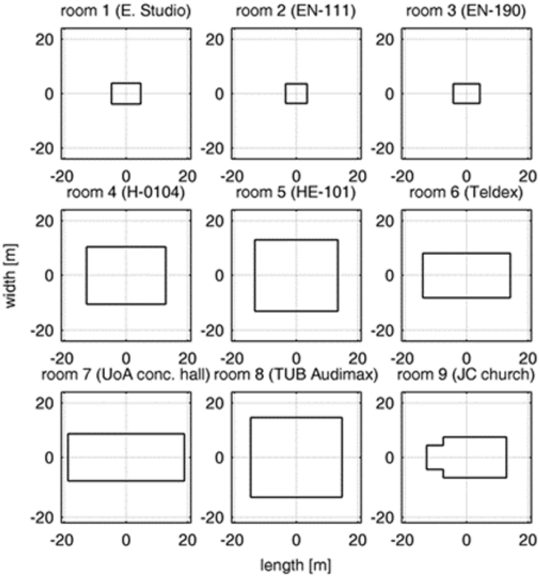


Figure 10-1. Floor plans of the nine rooms (true scale).

In a first step, monaural room impulse responses were measured at three different positions in the diffuse field using an omnidirectional microphone. Second, the three major room dimensions length, width, height were measured for calculating the volume. Then, the reverberation times displayed in Table 10-1 were calculated *in situ* as an average over the octave bands from 125 Hz to 4 kHz, and all three measurement positions. Now, the critical distance could be derived and FABIAN was seated on a place on the room's longitudinal symmetry axis directly facing the loudspeaker at twice the critical distance, where, according to Kuttruff [28], a random sound field can be expected. BRIRs were collected for horizontal head orientations within ±80° in angular steps of 1°.

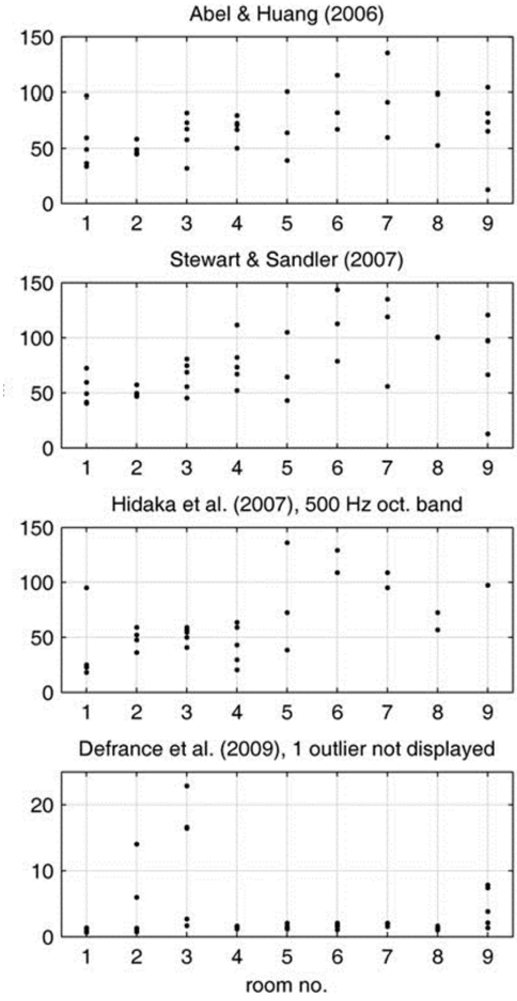


Figure 10-2. Positional variability obtained with automatic t_m detection using signal-based parameters (five values per room: left and right ears BRIR, three impulse responses from omnidirectional microphone measurements).

10.3.5 Practical Considerations when Calculating Signal-Based Mixing Time Parameters

We calculated the signal-based mixing time parameters from the dummy head's left and right ears' impulse response of the neutral head orientation (i.e. when facing the sound source) and from the three measurements collected with the omnidirectional microphone resulting in five signal-based mixing time estimates for each room. We considered all four approaches introduced in Section 10.3.2.

The mixing time according to Abel and Huang [25] was calculated using both mentioned stopping criteria I and II. The estimator proposed by Hidaka et al. [23] was calculated individually for the eight octave bands between 125 Hz–16 kHz.

A major issue observed with signal-based parameters was the variability of t_m values with measurement position. All mixing times derived from the four approaches are depicted in Figure 10-2 (Abel and Huang only for criterion I, Hidaka et al. only for 500 Hz octave band). As can be seen, values vary by factors up to two or three and even more as, e.g., in the case of Defrance et al.

As discussed already in Section 10.2.2 such positional variances might indicate imperfect mixing caused by close room boundaries, changing low frequency modal patterns, or residual coupled volumes. Positional variability of measures has partly been subject of discussion in the original publications and is not uncommon for certain room acoustical parameters (see, e.g., [29]). Since for BRIRs there are always two impulse responses available, we also tried to reduce this variability by subjecting the mean of the mixing time values as estimated from both channels of a BRIR to later statistical analysis.

Moreover, in case of the results from the Matching Pursuit decomposition [27], most values determined automatically were implausibly low (often around 2–5 ms, cf. Figure 10-2, bottom). This behavior was described already by Defrance et al. [27], when discussing the dependency of the estimates and their spread on the chosen signal residual ratio (SRR) in the Matching Pursuit decomposition (see also Section 10.3.2).

Apart from purely physical explanations for the observed positional variability, from examination of the plots of echo density profiles, normalized kurtosis or cumulated arrival function (cf. Figure 10-3) we suspected the different criteria for determination of the mixing time to be another reason for the positional instability of t_m measures. As can be seen from Figure 10-3 (upper and middle plot), echo density and normalized kurtosis did not always approach their target values (1 or 0, resp.) in a continuous manner, but jumped occasionally. Thus, measuring the time until the target value is reached for the first time might not always be a reasonable criterion.

Besides, from parameters' profile plots of Abel and Huang's, Stewart and Sandler's, and Defrance et al.'s method the mixing time could also be determined visually. In reading off mixing times from the location of the last noticeable jump toward the target value in the profile plots of echo density or normalized kurtosis, respectively, we hoped to get more stable results. Regarding Defrance et al.'s method, reading off the point where the slope of the cumulated arrival function changes from cubic to linear was often difficult (cf. Figure 10-3, bottom).

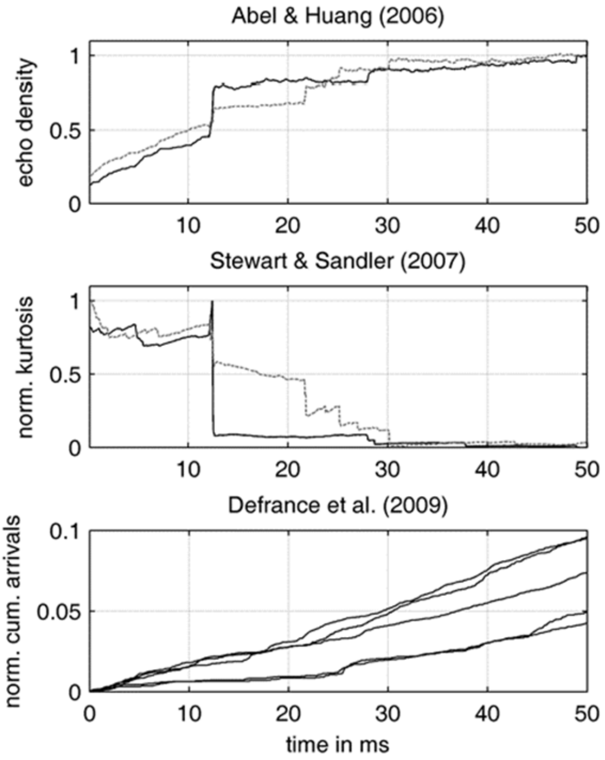


Figure 10-3. Positional variability of signal-based parameters' profiles (plots 1 and 2: from both BRIR channels in room 2; plot 3: from all five impulse responses in room 5).

Despite these problems, for the three approaches we additionally determined mixing time values by visual inspection of the corresponding curves and subjected them to our later statistical analysis.

10.3.6 Listening Test

Perceptual mixing times were determined using an adaptive 3AFC (three-alternative forced-choice) listening test procedure. Subjects were confronted with three stimuli in random order. Two of them were the reference simulation, where the complete BRIRs were updated in real time according to head movements. One of them was the manipulated simulation, where only the early part of the BRIRs was dynamically updated, whereas the late reverberation tail – taken from the BRIR corresponding to frontal head orientation – was concatenated with a linear cross-fade within a window size corresponding to the early block size of the fast convolution engine. This block size also corresponded to the step width the mixing time could be altered with (cf. Figure 10-4).

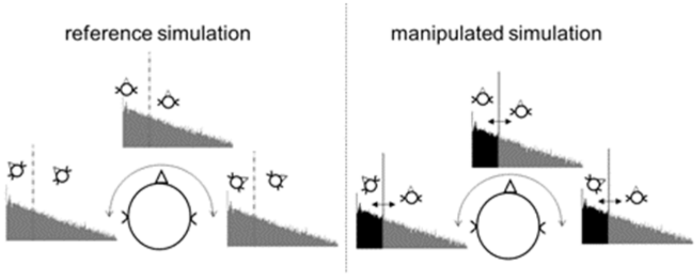


Figure 10-4. The two stimulus conditions presented in the listening test. Reference sound field/left: The complete BRIR is continuously updated according to the current head orientation. Manipulated sound field/right: Only the early BRIR part is continuously updated; the late reverberation always corresponds to frontal head orientation. The concatenation point between early and late part of the BRIRs was adaptively altered in the listening test.

Thus, the concatenation point of the updated early and the static late BRIR could be changed in increments of 5.8 ms (small rooms, nos. 1 to 3), or 11.6 ms (medium and large rooms, nos. 4 to 9). Whenever the subjects could correctly identify the manipulated simulation, the duration of the early dynamic part of the BRIR was increased; otherwise it was reduced, forcing the transition time to converge to the just noticeable point.

Following the definition proposed in Section 10.2.3, if the rooms were totally mixed everywhere and in the complete frequency range, BRIRs from every position in the rooms could have delivered the static reverberant tail. When assessing reverberant tails from different measurement positions in pretests, however, low frequency differences between original and manipulated tails were clearly audible, leading to high perceptual mixing times. As discussed in Section 10.2.2, different effects may have disturbed mixing of the sound field, such as comb filters caused by near room boundaries or a position-dependent low frequency response of the rooms due to modal patterns below the Schroeder frequency. Hence, the BRIRs of the neutral head orientation (i.e., from the same location as the dynamic BRIR datasets used for auralization) were used in order to avoid positional dependencies still observable in the late stochastic tail.

Although the noise floor was below -80 dB relative to the direct sound for all measurements, we limited the length of the BRIRs to about three-quarters of the duration of the decay to noise floor, because a slightly different background noise level and spectral coloration for different BRIRs was audible when comparing reference and manipulated stimuli and would thus have biased the detection task. Hence, BRIRs had a length between 14 000 (i.e., 0.375 s for room 1) and 100 000 (i.e., 2.26 s for room 9) samples, maintaining approximately 60 dB decay for binaural simulations of all rooms.

Loudness differences between simulated rooms were minimized through normalization of BRIR datasets. Electrostatic headphones (STAX SR-2050II) were frequency compensated using fast frequency deconvolution with high pass regularization from measurements on our dummy head FABIAN [30], [31]. Subjects were allowed to adjust sound pressure during training to a convenient level. This level was then kept constant throughout the listening test.

The listening test was conducted using the WhisPER toolbox [32]. As adaptation method for the threshold value (here: the just audible transition time), a Bayesian approach that closely matches the ZEST procedure [33] was chosen due to its unbiased and reliable results. The *a priori* probability density function was a Gaussian distribution with its mean in the middle of the stimulus range; the standard deviation was adapted in pretests.

As stimulus, the critical drum set sample from [6] was used again (length: 2.5 s without reverberation tail). The three stimuli where played back successively in each trial, without the possibility to repeat a trial. Subjects had to assess all nine rooms in an individually randomized order. A run was stopped after 20 trials, resulting in a test duration of about 60 minutes per person.

10.3.7 Subjects

During pretests, the listening test turned out to be a difficult task for some subjects. Consequently, we introduced a criterion to select “expert listeners” as those who were able to detect the right perceptual cue in order to perform the difference detection task successfully. Therefore, we regarded those subjects as experts, who were able to achieve, in all of the nine tested rooms, thresholds that were larger than the earliest possible concatenation instant (i.e. when concatenating only the dynamically updated direct sound with a static diffuse tail).

Finally, results of ten expert listeners (two female, eight male) with an average age of 27.8 years were taken into account for further statistical analysis. Most of the subjects had a musical education background and all had participated in listening tests before. During training subjects were instructed to rotate their head widely for maximizing the difference between original and manipulated reverberation tails. To increase statistical power we used a repeated measures design (every subject assessed every stimulus condition).

10.4 Listening Test Results

For each room the just detectable perceptual mixing times t_{mp} were calculated as the moment corresponding to the middle of the crossfade window between early and late BRIR at the cross fade instant the adaptive algorithm had converged to after 20 trials. Figure 10-5 shows the average perceptual mixing times t_{mp50} and confidence intervals ordered according to the two tested conditions volume and average absorption coefficient. As expected, t_{mp50}-values were found to increase with room volume. As indicated by the growing confidence intervals of rooms 7–9 the variation among subjects increased, too.

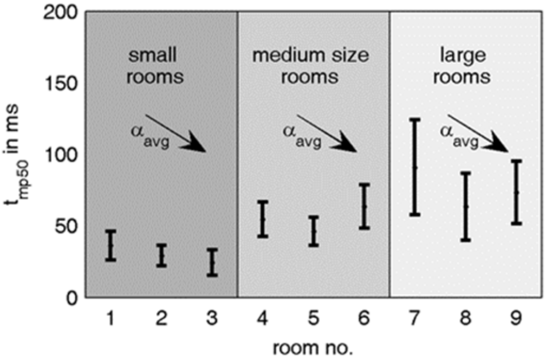


Figure 10-5. Average perceptual mixing times t_{mp50} per room with 95% CIs.

The ANOVA for repeated measures proved the volume effect to be significant at p = 0.001. Trend analysis confirmed a significant positive linear relation. An effect of the average absorption coefficient (i.e. the relative reverberance independent of volume) could not be found. This is in accordance with theory, as the amount of reverberation is in principle not related to diffusion or mixing. However, our sample size allowed only testing of intermediate effects ($f_{posthoc}$ = 0.307).

10.5 Regression Analysis

In Section 10.5.1, results of a regression analysis conducted to test the power of the three most important model-based relations (6)–(8) to predict t_{mp}, are presented. In Section 10.5.2, regression results for the signal-based parameters are discussed. In both cases, regression analysis was conducted for the average perceptual mixing time (t_{mp50}). Additionally, regression analysis was conducted while regressing on the 95%-point of the assumed normal distribution of the listening test results (t_{mp95}). While the t_{mp95}-regression formulae are intended to guarantee a perceptively perfect, close-to-authentic simulation, the t_{mp50}-predictions will guarantee a transparent simulation for at least half of the expert listeners.

Linear regressions were calculated as the least squares fit of the empirical t_{mp} values, derived in the listening test, to the t_m values as predicted by the above introduced model- and signal-based predictors. Thus, models of the form

$$t = b_1 t_m + b_2 \qquad \text{(10-11)}$$

were derived. Although the intercept term b_2 cannot be easily interpreted in physical terms (a zero physical mixing time should predict a zero perceptual mixing time), a higher explained variance was obtained by allowing an intercept term. All regression results were evaluated by means of the explained variance R^2, and the significance of regression, i.e., the possibility of rejecting the null hypothesis of a zero slope value b_1 at $p \leq 0.05$.

The number of nine rooms was rather low for linear regression analyses, a fact that is reflected by the confidence intervals displayed with the models. This shortcoming is, however, counterbalanced by the systematic and wide variation applied in selecting the rooms and by the selection of expert listeners, yielding a reliable measurement of perceptual mixing times with low variance among subjects.

10.5.1 Regression Results for Model-Based Predictors of Physical Mixing Time

As model-based predictors of the perceptual mixing time (a) the square root of volume, (b) the mean free path length, and (c) the reverberation time were subjected to regression analysis. Additionally, we tested the volume V, the surface area S (calculated from the three major room dimensions) and the average absorption coefficient α_{mean}.

Stepwise multiple and univariate linear regression analyses were conducted. Depending on the selection of variables, models containing one or three predictors resulted. The latter could be rejected, as the additional linear coefficients were insignificant, exhibited collinearity problems (high intercorrelation), and confidence intervals spanning to zero.

Thus, t_{mp50} was best predicted by the ratio V/S, the kernel of the mean free path length formula (10-7(10-7). In this case, the explained variance R^2 reached 81.5%. Regression on $\sqrt{V}$ (i.e. the reflection density relation) reached 78.6%, whereas volume alone achieved an R^2 of 77.4%. The reverberation time turned out to be unsuitable as predictor of the perceptual mixing time, since the explained variance of 53.4% can be completely attributed to confounded volume variation, while the average absorption coefficient α_{mean} shows nearly no linear relation to t_{mp50} (R^2 = 0.8%). All regressions were significant, except the one derived for α_{mean}. Figure 10-6 shows t_{mp50} values and linear regression models including 95% confidence intervals of both data and models.

The regression formula for the best predictor of t_{mp50} (in ms) was:

$$t_{mp50} = 20 \cdot V/S + 12 \qquad (10\text{-}12)$$

Thus, when comparing (10-4) and (10-12) and neglecting the constant term of the regression model, one derives that after approximately two reflections sound fields were experienced as being diffuse. Additionally – and while also neglecting the constant model term – from the second best predictor found, a just audible reflection density can be estimated by substituting t in (10-1) with the first addend of the regression formula

$$t_{mp50} = 0.58 \cdot \sqrt{V} + 21.2. \qquad (10\text{-}13)$$

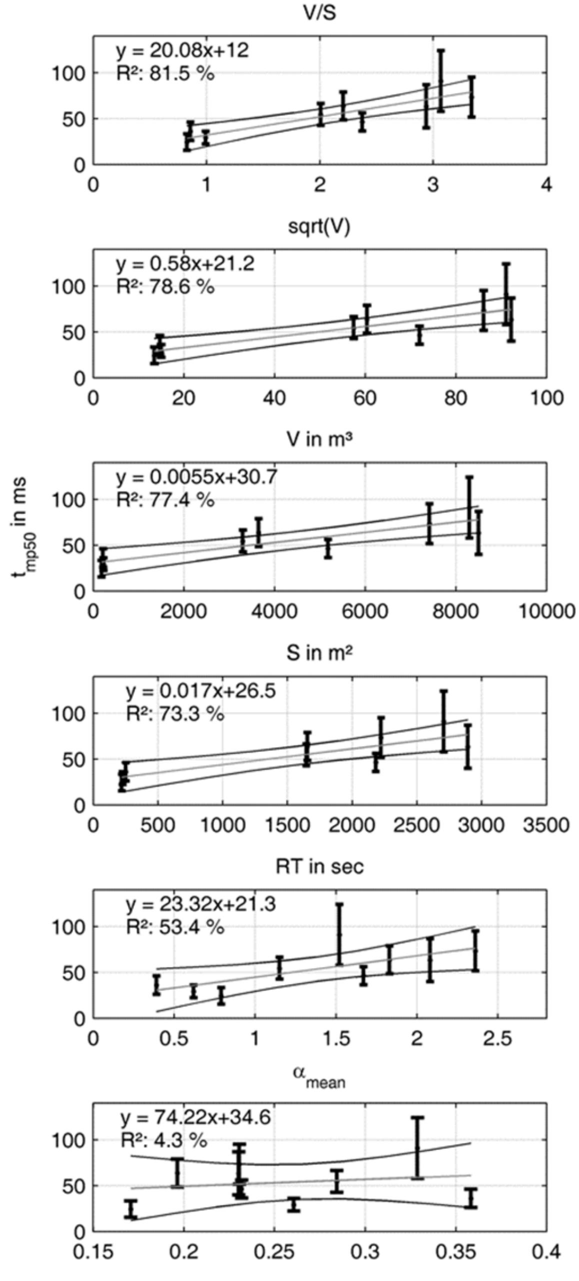


Figure 10-6. Average perceptual mixing times t_{mp50} in ms (incl. 95% CIs) plotted over different model-based predictors, and resulting linear regression models (incl. hyperbolic 95% CI curves).

Thus, with $c_0 = 343\,m/s$, the just audible reflection density can be estimated as $dN/dt = 171\,s^{-1}$. These values are considerably lower than those traditionally suggested in the literature (cf. Section 10.3.1).

However, it must be emphasized, that these are not measured physical quantities but are inferred from the model-based relations (10-6) and (10-7). Moreover, the inferred just audible quantities might be true only in the case of large rooms where the neglected constant term of the linear models becomes more and more irrelevant.

When regressing on the stricter t_{mp95} values all models were significant, too, except for the one derived from the average absorption coefficient α_{mean}.

The perceptual mixing time t_{mp95} (in ms) was best predicted by volume ($R^2 = 78.7\%$):

$$t_{mp95} = 0.0117 \cdot V + 50.1. \qquad (10\text{-}14)$$

Results for further parameters are displayed in

Table 10-2.

10.5.2 Regression Results for Signal-Based Predictors of Physical Mixing Time

Both, the mixing time values calculated from the left and right ears' BRIR and their average, determined either (a) visually, or (b) using the described deterministic detection criteria (cf. Section 1.5) were subjected to linear regression analyses. Again, regression analysis was conducted for an average (t_{mp50}), and a strict (t_{mp95}) perceptual mixing time criterion.

For the algorithm of Defrance et al. [27], most of the estimated mixing time values were implausibly low (cf. Figure 10-2), especially when assuming the signal-based approaches to be directly estimating the physical mixing time. A comparative visual inspection of the cumulative arrival functions suggests that the equivalent pulse duration criterion does not always lead to a correct detection of the inflection point of the cumulative arrival function. Therefore, and although in principal we consider this method as an attractive approach we did not subject its results to regression analysis.

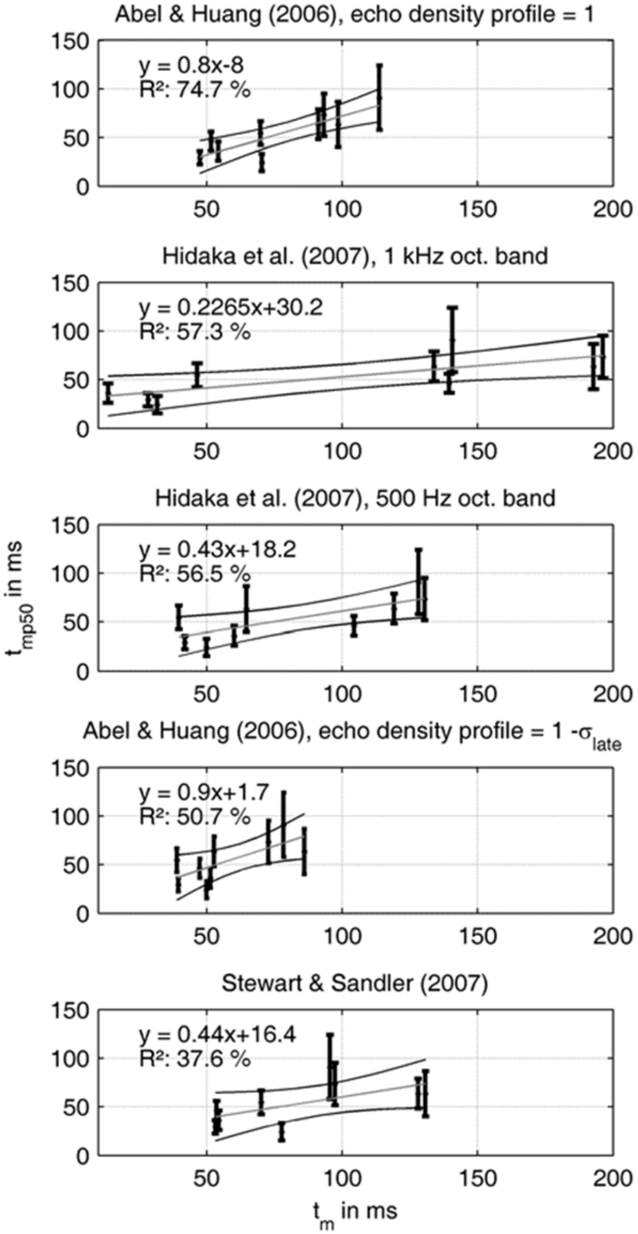


Figure 10-7. Average perceptual mixing times t_{mp50} in ms (incl. 95% CIs) plotted over signal-based tm-predictors (mean values from both channels of a BRIR), and resulting linear regression models (incl. hyperbolic 95% CI curves).

Although some of the regression models derived from values of a single ear's BRIR reached higher values of explained variance, this happened randomly for the left or the right ear. A systematic relation could not be found, thus all further results are solely based on the average mixing time calculated from both channels of the BRIRs.

All regression models were significant, except the one derived from Stewart's and Sandler's [26] kurtosis-based method. The echo density approach of Abel and Huang [25] (criterion I) achieved a R^2 of 74.7% (cf. Figure 10-7, plot 1). The obtained regression formula was

$$t_{mp50} = 0.8 \cdot t_{mix-Abel-I} - 8. \tag{10-15}$$

Therefore, we can recommend this estimator for signal-based determination of t_{mp50}. Regression models of the other approaches are depicted in Figure 10-7, where results are presented in descending order of performance. The correlation approach from Hidaka et al. [23] reaches minor prediction performance but at least R^2 values of 56.5% to 57.3% for the mid frequency octave bands (500 Hz and 1 kHz).

Visually reading off mixing time values from the profile plots of (a) reflection density, (b) normalized kurtosis, or (c) cumulative arrivals resulted in regression models with considerably less explained variance. Moreover, as this procedure is very time consuming it cannot be recommended.

Prediction results for t_{mp95} are also shown in

Table 10-2 ordered for performance. Again, all models were significant, despite that one derived from the kurtosis approach [26].

The approach of Abel and Huang again showed a superior performance, with an explained variance of 83.7%. The regression formula reads

$$t_{mp95} = 1.8 \cdot t_{mix-Abel-I} - 38. \tag{10-16}$$

Table 10-2. Ranking of model- and signal-based mixing time estimators in predicting perceptual mixing times t_{mp50} and t_{mp95}.

	Model-based predictors		Signal-based predictors	
#	t_{mp50}	t_{mp95}	t_{mp50}	t_{mp95}
1	V/S R^2 81.5%	V R^2 78.7%	Abel I R^2 74.7%	Abel I R^2 83.7%
2	$\sqrt{V}$ R^2 78.6%	V/S R^2 75.7%	Hidaka 1k R^2 57.3%	Abel II R^2 66.7%
3	V R^2 77.4%	$\sqrt{V}$ R^2 73.4%	Hidaka 500 R^2 56.5%	Hidaka 1k R^2 55%
4	S R^2 73.3%	S R^2 69.5%	Abel II R^2 50.7%	Hidaka 500 R^2 49.2%
5	RT R^2 53.4%	RT R^2 46.5%	Stewart R^2 37.6%	Stewart R^2 40.3%
6	α_{mean} R^2 4.3%	α_{mean} R^2 4.8%		

Further measures performed less well, though the echo density with criterion II [25] in this case worked better than the correlation measures of Hidaka et al. (cf.

Table 10-2). For assessing how well the t_{mp50}-values are directly predicted by the signal-based parameters t_m (and not as part of a regression model), Figure 10-8 displays all assessed parameters in the $t_m t_{mp50}$-plane. If predicted values of t_{mp50} were identical to the estimated mixing times t_{mp}, points should scatter along the angle bisector of the $t_m t_{mp50}$-plane. As can be seen, this is again best achieved by Abel and Huang's approach (criterion I, [25]).

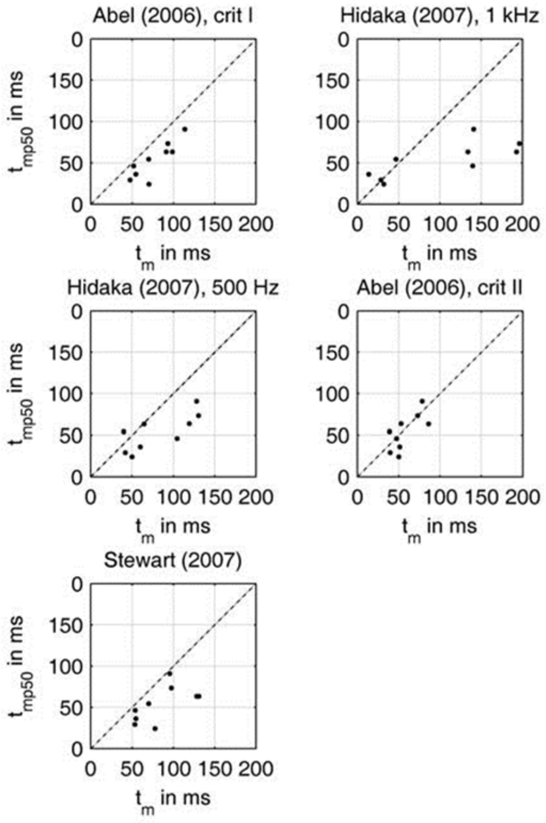


Figure 10-8. Average perceptual mixing times t_{mp50} (mean values from both channels of a BRIR) plotted over signal-based tm-predictors.

10.6 Conclusions

The perceptual mixing time was assessed for the first time by means of a high quality dynamic binaural simulation. BRIR data sets have been acquired for nine acoustical environments, systematically varied in volume and average absorption. Both model- and signal-based estimators of the mixing time were evaluated for their power to predict the listening test results of a group of expert listeners. As a result, linear regression models predicting either (a) the average, or (b) the more critical 95%-point of the perceptual mixing times were presented, yielding predictors for situations, where (1) only the dimensions of the room are known (for instance in the case of model-based auralization), or (2) when an impulse response is available for instance in the case of data-based auralization).

Results show that for shoebox shaped rooms the average perceptual mixing time can be well predicted by the enclosure's ratio of volume over surface area V/S [equation (10-12)] and by $\sqrt{V}$ [equation (10-13)] being indicators of the mean free path length, and the reflection density, respectively. The linear factors in our regression models suggest that a time interval corresponding to about two mean free path lengths, i.e., on average two orders of reflection, and a reflection density of less than 200 s^{-1} is perceived as diffuse even by trained listeners. Any dependence on reverberation time turned out to be due to its implicit co-variation with room volume.

If an impulse response is available, average perceptual mixing times can be optimally predicted by regression formula (10-15) using values calculated from the echo density approach of Abel and Huang [25] applying the stopping criterion I (the echo density profile becoming equal to unity). For increased reliability of the prediction, the input value should be an average over several measurement positions.

The presented regression formulae for a perceptual mixing time can be applied to reduce the rendering effort of both loudspeaker- or headphone-based high quality VAEs and plausible auralization on limited platforms such as mobile audio devices.

For convenient application of the presented predictors we made appropriate Matlab® source code publicly available[7].

10.7 Acknowledgements

Alexander Lindau was supported by a grant from the Deutsche Telekom Laboratories and by the Deutsche Forschungsgemeinschaft (DFG), grant WE 4057/1-1.

10.8 References

[1] Polack, J.-D. (1992): "Modifying Chambers to play Billiards the Foundations of Reverberation Theory", in: *Acustica*, **76**, pp. 257-272

[2] Reilly, A.; McGrath, D. (1995): "Real-Time Auralization with Head Tracking", in: *Proc. of the 5th Australian Regional AES Conv.*, Sydney, preprint no. 4024

[7] www.ak.tu-berlin.de/menue/digitale_ressourcen/research_tools/mixing_time_prediction

[3] Meesawat, K; Hammershøi, D. (2003): “The time when the reverberant tail in binaural room impulse response begins”, in: *Proc. of the 115th AES Conv.*, New York, preprint no. 5859

[4] Menzer, F.; Faller, C. (2010): “Investigations on an Early-Reflection-Free Model for BRIRs”, in: *J. Audio Eng. Soc.*, **58**(9), pp. 709-723

[5] Olive, S. E.; Toole, F. E. (1989): “The Detection of Reflections in Typical Rooms”, in: *J. Audio Eng. Soc.*, **37**(7/8), pp. 539-553

[6] Lindau, A.; Hohn, T.; Weinzierl, S. (2007): “Binaural resynthesis for comparative studies of acoustical environments”, in: *Proc. of the 122nd AES Conv.*, Vienna, preprint no. 7032

[7] Kuttruff, H. (2000): *Room Acoustics. 4th ed.*, New York: Routledge Chapman & Hall

[8] Schroeder, M. R. (1959): “Measurement of Sound Diffusion in Reverberation Chambers”, in: *J. Acoust. Soc. Am.*, **31**(11), pp. 1407-1414

[9] Blesser, B. (2001): “An Interdisciplinary Synthesis of Reverberation Viewpoints”, in: *J. Audio Eng. Soc.*, **49**(10), pp. 867- 903

[10] Joyce, W. B. (1975): “Sabine's reverberation time and ergodic auditoriums”, in: *J. Acoust. Soc. Am.*, **58**(3), pp. 643-655

[11] Begault, D. (1992): “Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems”, in: *J. Audio Eng. Soc.*, **40**(11), pp. 895-904

[12] Bradley, J. S.; Soulodre, G. A. (1995): “The influence of late arriving energy on spatial impression”, in: *J. Acoust. Soc. Am.*, **97**(4), pp. 2263-2271

[13] Kuttruff, H. (1993): “Auralization of Impulse Responses Modeled on the Basis of Ray-Tracing Results”, in: *J. Audio Eng. Soc.*, **41**(11), pp. 876-880

[14] Reichardt, W.; Lehmann, U. (1978): “Raumeindruck als Oberbegriff von Räumlichkeit und Halligkeit, Erläuterungen des Raumeindrucksmaßes R”, in: *Acustica*, **40**(5), pp. 277-290

[15] Hidaka, T.; Okano, T.; Beranek, L. L. (1995): “Interaural cross-correlation, lateral fraction, and low- and -high-frequency sound levels as measures of acoustical quality in concert halls”, in: *J. Acoust. Soc. Am.*, **98**(2), pp. 988-1007

[16] Cremer, L.; Müller, H. A. (1978): *Die wissenschaftlichen Grundlagen der Raumakustik. Bd. 1: Geometrische Raumakustik. Statistische Raumakustik. Psychologische Raumakustik.* 2nd ed., Stuttgart: Hirzel

[17] Schmidt, W.; Ahnert, W. (1973): "Einfluss der Richtungs- und Zeitdiffusität von Anfangsreflexionen auf den Raumeindruck", in: *Wiss. Zeit. d. TU Dresden*, **22**, pp. 313

[18] Polack, J.-D. (1988): *La transmission de l'énergie sonore dans les salles*, Thèse de Doctrorat d'Etat. Le Mans: Université du Maine

[19] Schroeder, M. R. (1962): "Natural sounding artificial reverberation", in: *J. Audio Eng. Soc.*, **10**(3), pp. 219-223

[20] Schreiber, L. (1960): "Was empfinden wir als gleichförmiges Rauschen?", in: *Frequenz*, **14**(12), pp. 399

[21] Rubak, P.; Johansen, L. G. (1999): "Artificial Reverberation based on a Pseudo-random Impulse Response II", in: *Proc. of the 106th AES Conv.*, Munich, preprint no. 4900

[22] Griesinger, D. (1989): "Practical Processors and Programs for Digital Reverberation", in: *Proc. of the 7th International AES Conference: Audio in Digital Times,* Toronto

[23] Hidaka, T.; Yamada, Y.; Nakagawa, T. (2007): "A new definition of boundary point between early reflections and late reverberation in room impulse responses", in: *J. Acoust. Soc. Am.*, **122**(1), pp. 326-332

[24] Schroeder, M. R. (1965): "New method of measuring reverberation time", in: *J. Acoust. Soc. Am.*, **37**, pp. 409-412

[25] Abel, J. S.; Huang, P. (2006): "A Simple, Robust Measure of Reverberation Echo Density", in: *Proc. of the 121st AES Conv.*, San Francisco, preprint no. 6985

[26] Stewart, R.; Sandler, M. (2007): "Statistical Measures of Early Reflections of Room Impulse Responses", in: *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, Bordeaux

[27] Defrance, G.; Daudet, L.; Polack, J.-D. (2009): "Using Matched Pursuit for Estimating Mixing Time Within Room Impulse Responses", in: *Acta Acustica united with Acustica*, **95**(6), pp. 1071-1081

[28] Kuttruff, H. (1991): “On the audibility of phase distortion in rooms and its significance for sound reproduction and digital simulation in room acoustics”, in: *Acustica*, **74**, pp. 3-7

[29] de Vries, D; Hulsebos, E. M.; Baan, J. (2001): “Spatial Fluctuations in measures for spaciousness”, in: *J. Acoust. Soc. Am.*, **110**(2), pp. 947-954

[30] Schärer, Z.; Lindau, A. (2009): “Evaluation of Equalization Methods for Binaural Signals”, in: *Proc. of the 126th AES Conv.*, Munich, preprint no. 7721

[31] Lindau, A.; Brinkmann, F. (2012): “Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings”, in: *J. Audio Eng. Soc.*, **60**(1/2), pp. 54-62

[32] Ciba, S.; Wlodarski, A.; Maempel, H.-J. (2009): “WhisPER – A new tool for performing listening tests”, in: *Proc. of the 126th AES Conv.*, Munich, preprint 7749

[33] King-Smith, P. E. et al. (1994): “Efficient and Unbiased Modifications of the QUEST Threshold Method: Theory, Simulations, Experimental Simulations, Experimental Evaluation and Practical Implementation”, in: *Vision Research*, **34**(7), pp. 885-912

11 Perceptual Evaluation of Discretization and Interpolation for Motion-Tracked Binaural (MTB-) Recordings

The following chapter is based on the article:

> Lindau, Alexander; Roos, Sebastian (2010): "Perceptual Evaluation of Discretization and Interpolation for Motion-Tracked Binaural (MTB-) Recordings", in: *Proc. of the 26th Tonmeistertagung*. Leipzig, pp. 680-701.

Author's Note

The original article was presented at the German Tonmeistertagung, a platform known for presenting of both scientific and applied topics. As a result, a substantial amount of the original article was devoted to the practical application of MTB recording and playback. However, in the context of the thesis at hand – focusing on fundamental research questions in the field of audio perception – most of these remarks were not considered relevant. Hence, the presentation here is a shortened version, aiming at presenting only those parts of the article which are relevant for the subject of perceptual evaluation of discretization and interpolation of MTB recordings.

Further, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g. reworking the citation style, typographic and stylistic corrections.

11.1 Abstract

In 2004, Algazi and colleagues introduced motion-tracked binaural sound (MTB, [1]) as a new method for capturing, recording, and reproducing spatial sound. The typical MTB recording device consists of a rigid sphere with the diameter of an average human head which and is equipped with a circular array of equidistant microphones at its circumference. Recordings are played back via headphones while being interpolated in real time according to the listener's current horizontal head orientation. Whereas the technical development of MTB has already arrived at a quite elaborate stage, a formal perceptual evaluation of the sound quality with regard to relevant system parameters is still missing. Therefore, we conducted a listening test with 26 subjects rating the degree of perceived naturalism of synthetically created and interactively rendered MTB recordings. Using a MUSHRA-like

[12] comparative test design, we tested the effect of the number of microphones (8, 16, 24, 32), interpolation scheme (5 methods), and audio content (pink noise, speech, music). Although we could find a clearly superior configuration, the naturalism of MTB reproduction was found to be highly interdependent on all three parameters.

11.2 Introduction

In 2004, Algazi et al. [1] proposed a new binaural recording technique: motion tracked binaural sound (MTB). The typical MTB recording device consists of a rigid sphere with the diameter of an average human head. Several microphone capsules are evenly distributed around the horizontal circumference of the sphere (cf. Fig. 1).

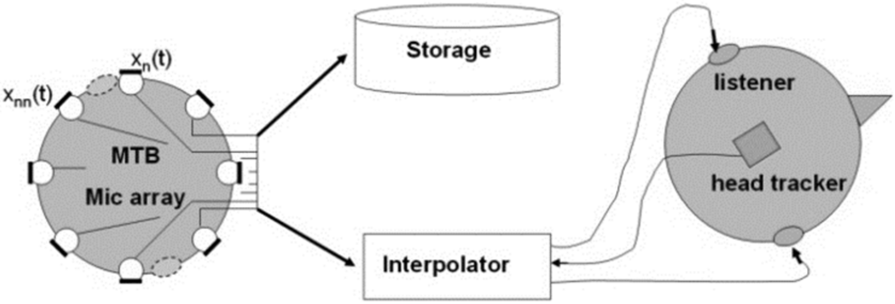


Figure 11-1. Sketch of MTB reproduction principle: A head tracker determines the position of the listener's ears. The signals of the two microphones which are closest to the position of an ear, $x_n(t)$ and $x_{nn}(t)$, are selected and ear signal are constructed by interpolating between these microphone signals while continuously following the head orientation. Ear signals can be constructed instantaneously or off-line from stored MTB recordings (drawing after [5]).

Two audio signals are generated from the MTB's multichannel signal continuously approximating the sound pressure function at the location of the listener's ears by means of interpolation. Therefore, the rigid sphere of the MTB array acts as an obstacle for sound propagation, introducing pseudo-binaural cues to the recorded signal, as, e.g., frequency dependent interaural level (ILD) and time (ITD) differences similar to those occurring in human binaural hearing. Depending on the complexity of the interpolation algorithm, the interactive ear signals can be constructed in real-time. In contrast to impulse response based dynamic binaural synthesis [2]–[4] the MTB technique allows direct binaural 'field' recording while preserving the possibility of later dynamization. This has advantages especially when aiming at capturing ambient acoustic scenes with binaural cues, as, e.g., in the context of soundscape research. Of course, this advantage has to be traded against a limited flexibility, as – in contrast to conventional dynamic binaural synthesis – audio content and acoustic scene setup cannot be manipulated separately after recording.

11.2.1 Foundations of Spatial Hearing Relevant for MTB Reproduction Quality

In order to better understand MTB's perceptual limitations, the most relevant interaural signal features exploited for spatial hearing are recapitulated here shortly. According to Lord Rayleigh's duplex theory [6], to determine a sound source's location the ITD is evaluated for frequency ranges below approximately 1.5 kHz. Above that frequency, ILD cues are interpreted accordingly. However, there exist infinite non-identical positions in space for which ITD and ILD are the same (due to geometrical shape referred to as 'cones of confusion'). Naturally, these ambiguities are resolvable via deviations induced by head movements. Moreover, for wide band stimuli containing conflicting ITD and ILD cues, Wightman and Kistler [7] showed the dominance of the ITD in determining source localization. For source locations on the median plane no ILD and ITD cues exist. Elevation of sound sources is therefore detected by evaluating monaural spectral differences introduced by torso, head and pinna [7], [8] (i.e. spectral cues, SC). If the pinna geometry differs from that of the listener, or, as in the case of MTB pinnae are completely absent, errors in determining source elevation can be expected.

11.2.2 Interpolation Algorithms for MTB Signals

Relevant parameters of the MTB can be chosen deliberately. Some of these parameters have been discussed before: (1) the number of microphones, (2) the interpolation algorithm, (3) the array sampling grid, and (4) the angular position aimed at when interpolating microphone signals. The potential perceptual impacts of these parameters have been discussed by Algazi et al. [1] using plots of physical error measures. Effects were discussed for five different approaches to interpolation while using 8, 16, or 32 microphones respectively. However, the perceptual effects were discussed on the base of informal listening tests. Therefore, in the present study we independently assessed the impact of the number of microphones and the type of interpolation algorithm in a formal listening test. In order to better understand the tested conditions, the five interpolation algorithms from [1] will be shortly explained here in an order of increasing complexity.

11.2.2.1 Full Range Nearest Microphone Selection (FR-NM)

The probably simplest way to reconstruct the ear signals is to select the signals of the microphones which are located next to the listener's ear positions (instant switching). As a result, the acoustic space will be fragmented into N (N being the number of available microphones) angular sectors each of a size of $\Delta\varphi$ = 360°/N in each of which the signal is not adapted to the listeners head orientation. Thus, the acoustic image will 'jump' at the sectors' boundaries. Additionally, switching artifacts can be expected to become audible each time when crossing such a boundary.

11.2.2.2 Full Range Linear Interpolation (FR-LI)

In a next step, the discontinuous behavior of the FR-NM method can be avoided by linearly cross fading between two adjacent microphones' signals. Hence, the signal $x(t)$ at the ear's position can be interpolated from $x_n(t)$ being the output of the nearest microphone and $x_{nn}(t)$ being the output of the next nearest microphone (cf. Figure 11-1) by a linear inverse-distance cross-fade function (from [1]):

$$x(t) = (1 - w)x_n(t) + wx_n(t). \tag{11-1}$$

To this end, the interpolation weight w is determined as the ratio of the angle between the ear and the (currently) nearest microphone β and the average angular microphone distance $\Delta\varphi$:

$$w = \beta/\Delta\varphi. \tag{11-2}$$

Linearly combining signals from closely aligned microphones is prone to comb filtering artifacts. Moreover, the resulting audible spectral coloration will vary with head movements and direction of sound incidence. From an analysis of the comb filter system resulting for grazing sound incidence Algazi et al. [1] derived a criterion for the minimum number of microphones needed to keep MTB's magnitude response within ±3dB deviation below a certain frequency f_{up}:

$$N_{min} = 8\pi r_{MTB}\, f_{up}/c_0. \tag{11-3}$$

Assuming an MTB radius of r_{MTB}= 87.5 mm, c_0 as the sound velocity, and setting f_{up} to 20 kHz one would need a number of at least 128 microphones. As this is a rather large number, the next three interpolation methods exploit aspects of human hearing to achieve a satisfactory sound quality in a more efficient manner. According to (11-3), comb filtering artifacts can be pushed beyond spectral regions of approximately 1.2 kHz when using a number of at least eight microphones. Hence, by applying a low pass filter, perceptually appropriate low frequency ear signals – preserving the relevant low frequency ITD cues – can be reconstructed from such a MTB signal. However, high frequency spectral energy and ILD cues need to be reproduced, too. Therefore, the next three interpolation approaches from [1] employ two-band treatments while differing in the way the high frequency components are restored.

11.2.2.3 Two Band Fixed Microphone Interpolation (TB-FM)

First and most simply, higher frequencies could be reproduced from the high pass filtered signal of a fixed omnidirectional microphone (‘complementary microphone’). Hence, high frequency spectral energy will be restored, but, as this signal is monaural, all spatial information (ITD, ILD) will be lost. Whereas - due to weak sensitivity to high frequency phase differences - the loss of high frequency ITDs might be less disturbing, missing high frequency ILDs will strongly disturb auditory spatial perception. Algazi et al. [1] informally reported the perception of ‘split’ acoustic images, with low frequency content reproduced correctly, while high frequency content is located inside the head. Additionally, the question arises, where the complementary microphone should be located best? Depending on the acoustic scene and the amount of reverberation its exact location will have a strong effect on the overall sound color.

11.2.2.4 Two Band Nearest Microphone Selection (TB-NM)

Alternatively, the high frequency information can also be restored by applying the nearest microphone selection procedure (i.e. instant switching, cf. 11.2.2.1) to the upper audio range (Figure 11-2). This approach can be expected to restore dynamic high frequency ILD cues helping to avoid the perception of ‘split images’. However, as proven also by results of our listening test, depending on the audio content and the number of microphones high frequency sectorial switching will lead to audible artifacts, too.

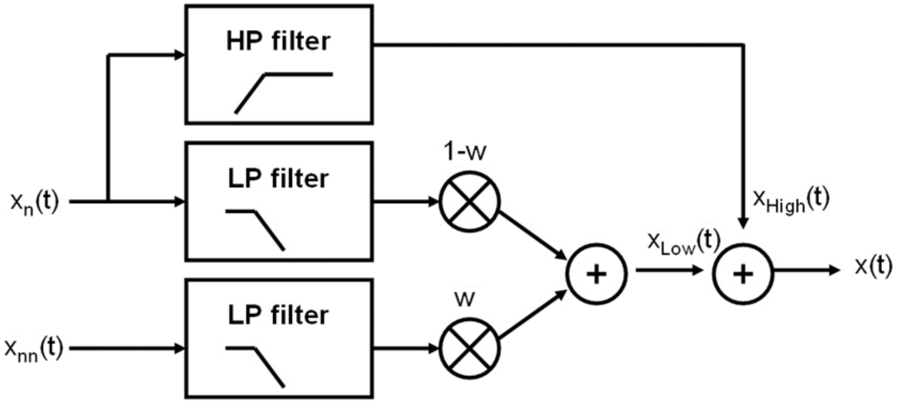


Figure 11-2. Block diagram of TB-NM algorithm: Low frequency information is restored from continuously crossfading between the two microphones' signals $x_n(t)$ and $x_{nn}(t)$ closest to the respective ear. High frequency information is restored by switching between the respective nearest microphone signals, e.g., here $x_n(t)$ (after [5]).

11.2.2.5 Two Band Spectral Interpolation Restoration (TB-SI)

The last approach uses the short-time (fast) Fourier transform to conduct real-time linear interpolation in the spectral domain. With $M_n(\omega)$ and $M_{nn}(\omega)$ being the magnitudes of the short-time Fourier transform of the high frequency content of the microphone signals $x_n(t)$, and $x_{nn}(t)$ respectively (cf. Figure 11-3), (11-1) becomes

$$M_c(\omega) = (1 - w)M_n(\omega) + wM_n(\omega). \qquad (11\text{-}4)$$

Three different procedures for the restoration of high frequency content by spectral interpolation have been proposed and discussed thoroughly in [5]. A general flow diagram is given in Figure 11-3. The procedures differ with respect to (a) the compatibility to continuous block wise processing necessary for real-time operation, and (b) the reconstruction of the temporal waveform (block 'Waveform synthesis' in Figure 11-3).

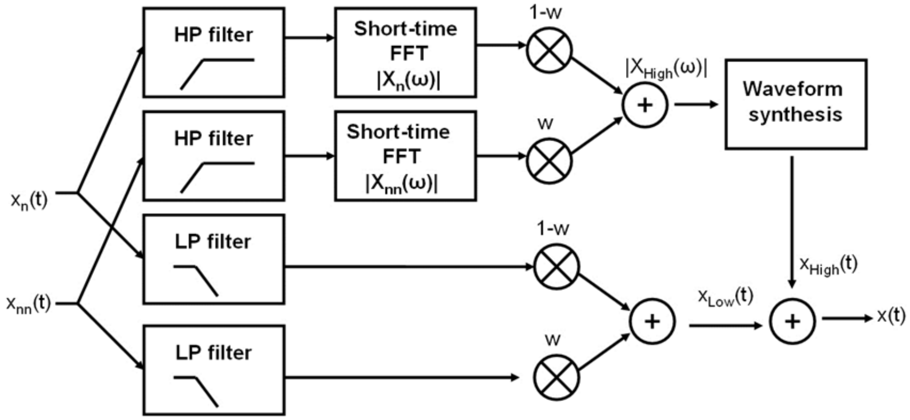


Figure 11-3. Block diagram of TB-SI algorithm: Low frequency information is derived from continuously crossfading between the two microphones' signals $x_n(t)$ and $x_{nn}(t)$ closest to the respective ear. High frequency information is derived through interpolation of short-time magnitude spectra and phase reconstruction during waveform synthesis (after [5]).

The two real-time methods (weighted-overlap-and-add (WOLA), and least-squares-error estimation of the modified short-time Fourier transform (LSEE-MSTFT), cf. [5] for further details), only deliver an interpolated short-time magnitude spectrum, so a suitable phase spectrum has still to be derived. The approach pursued in [5] is to select the phase of the nearest microphone ('phase switching'). For our listening test we chose WOLA with phase switching as approach to Two Band Spectral Interpolation Restoration (TB-SI, cf. also sect. 11.3.1).

11.2.3 MTB's Perceptual Defects and Potential Remedies

MTB's perceptual defects and probable causes have been discussed thoroughly by Melick et al. ([10], cf. Table 2 there). Most relevant issues were found to be due to (1) the missing pinnae, (2) mismatch in head and array diameter, (3) shortcomings of the interpolation algorithms, and (4) mismatch of microphone and ear location.

First, the missing pinnae result in both erroneous monaural high frequency spectral detail and high frequency ILD cues. Thus, in comparison to a head with pinnae, MTB signals will produce spectral coloration, erroneous elevation cues and horizontal localization mismatches. The missing pinnae though offer the advantage that an MTB array has no preferred spatial orientation within the horizontal plane. Hence, from a single MTB recording, ear signals for a plurality of listeners with individual head orientations can be rendered synchronously.

Second, the mismatch between the MTB's and a listener's head diameter will produce erroneous ITD cues. As a result, horizontal source location will not be perceived as stable. If the listener's head diameter is smaller than that of the array, ITD cues are larger than natural, resulting in a perceived motion of the source in retrograde direction of the listeners head movements. If the head size is larger, the inverse effect occurs and is perceived as the source 'lagging behind' the head movements [1]. Whereas for impulse response based dynamic binaural synthesis a promising approach to post-hoc individualization of head size has been presented in [9], for MTB recordings a generic ('one-fits-all') sphere diameter is commonly used. Algazi et al. [1] proposed to use a diameter of 175 mm, from which 98% of the U.S. American population deviate by ±15%.

Third, the possibility to interpolate the MTB signal at arbitrary in-between microphone positions allows manipulating both ITD and timbre. Hence, both ears' virtual positions may be shifted forward and backward on the circumference. Delocating the virtual ear position from the frontal plane (either for- or backwards) will decrease the low frequency ITD in the MTB signal as microphones become closer to each other [10]. Simultaneoulsy, – for frontal sound incidence – moving the virtual ear position backwards will lead to high frequency shadowing, while moving it forward will increase pressure stasis leading to high frequency boosting. For the two-band interpolation algorithms Melick et al. [10] proposed to (a) move the virtual ear position backwards on the circumference for the low pass filtered signal path, this way achieving a more natural ear position, while (b) shifting the ear position slightly forward for the high pass filtered signal path partly compensating the missing pinnae with the resulting pressure stasis effect.

11.2.4 *A priori* Perceptual Ranking of MTB Interpolation Algorithms

From theoretical considerations and informal listening tests Algazi et al. [1] derived a perceptual ranking of the five interpolation approaches. Therefore, they considered how faithful each approach preserves psychoacoustic cues (ILD, ITD, SC), or in how far it was prone to auditive artifacts (switching, comb filtering, split imagery). They concluded that 'two band spectral interpolation' should result in perceptively best reproduction, followed by 'two band nearest-microphone selection', whereas 'full band interpolation' is seen on the third rank. 'Full band nearest microphone' and 'two-band fixed microphone' are both considered worse but efficient in terms of bandwidth and computational effort. They stated that the number of microphones would not have to be large in order to achieve a 'strong sense of realism and presence' [1]. They further assumed that the overall acceptability of the interpolation approaches would depend on the audio content and would, in turn, have to be traded against the available bandwidth. While assuming that spatial discontinuities would make reproduction unacceptable especially for musical content, authors also argued that, if the required sound quality was low and the available bandwidth was limited, a number of eight microphones might suffice for general purposes.

11.2.5 Aims of this Study

So far we have introduced MTB as a method for recording and interactive rendering of spatial acoustic scenes. We explained its basic technology, reviewed its perceptual shortcomings and potential remedies for them. We also showed that results from perceptual evaluation of even the most relevant system parameters (number of microphones, interpolation strategies, and interaction with different audio content) are sparse and mostly informal. Therefore, we assessed the perceived degree of naturalism of the MTB method as a function of the number of microphones, the type of interpolation algorithm and the audio content.

11.3 Methods

It was decided, that a most demanding comparison of MTB approaches would need an interactive real-time capable listening test environment. Therefore, two dedicated real-time applications the ‘MTB player’ and the ‘MTB renderer’ were implemented (cf. 11.3.1). Further, MTB recordings were created for several acoustic scenes using different numbers of microphones. For achieving most unbiased comparability all recordings were synthesized from room impulse responses measured with a virtual MTB microphone array and an arrangement of multiple loudspeakers (cf. 11.3.2).

11.3.1 Implementation and Parameterization of MTB Rendering Software

Our MTB applications make use of the Jack audio connection kit (http://jackaudio.org) as backbone for real-time audio streaming. The MTB renderer application allows interactive rendering of live or recorded MTB signals while instantaneously choosing between different interpolation methods. So far, the renderer offers the five interpolation algorithms described in section 11.2.2 (FR-NM, FR-LI, TB-FM, TB-NM, and TB-SI) while accepting MTB signals with 8, 16, 24, or 32 microphone channels. For the two band methods the crossover frequency can be chosen at will. For the listening test we chose a 1.5 kHz crossover frequency realized by 4th order Linkwitz-Riley filters. As there was no respective recommendation in [1], we deliberately chose the microphone pointing into the array’s frontal orientation for high frequency reconstruction for the two band fixed microphone (TB-MF) algorithm. For the spectral interpolation (TB-SI) method we chose the weighted overlap and add (WOLA) real-time method with phase switching described in [5]. Being a block based implementation of the continuous spectral interpolation different analysis/synthesis windows widths and sizes can be chosen from (rectangular, triangular, Hanning, Blackman, Hamming). For our listening test we used WOLA with a 128 taps Hanning window with 75% overlap. The MTB player application allows simultaneous playback of MTB recordings made with different numbers of microphones. If the player is interfaced to the renderer one may instantaneously switch between available MTB recordings made with different numbers of channels. Further, most parameters of MTB player and renderer may be instantly controlled using either a graphical user interface, or the OSC remote control protocol [11]. For the listening test the MTB renderer was further configured to render the ear signals for diametrically opposed positions at the maximum sphere diameter. Head tracking data were transmitted and received via OSC messaging.

11.3.2 Synthetic MTB Recordings

MTB recordings were created in a lecture hall of the TU Berlin (room HE 101, RT = 1.67 s, V = 5200 m^3). HE-101 – for which Lothar Cremer was the acoustic consultant – is a slightly fan-shaped shaped room with a sloped seating area, rising softly towards the back wall. The side walls are covered with perforated, wooden lattice structure acting both diffusing and absorbing, whereas the ceiling has a stepped, stairs-like surface. For a flexible design of acoustic test scenes, we placed eight small active wideband loudspeakers (Fostex 6301, 10 cm diaphragm) at more or less random locations in the hall (cf. Figure 11-4, right). Due to the dimensions of the loudspeakers their directivity of can assumed to roughly mimic that of a human talker. The virtual MTB microphone was placed centrally in the hall at approximately two meters above floor level with the frontal microphone position pointing at the stage end of the hall. As MTB microphone we used a singular full range and omnidirectional electret microphone capsule (Panasonic WM-61A) attached at the circumference of a rigid plastic sphere of 180 mm in diameter (cf. Figure 11-4, left). From substitution measurements in an anechoic chamber it was found, that the WM-61A had a sufficiently flat and wide frequency response (0 Hz–19.5 kHz, ±1.5 dB, with a 2 dB roll-off above 3 kHz), which was not corrected for. This single-microphone array was mounted on a base that could be rotated horizontally with high angular precision using a servo motor device as descried in [3]. With this setup we measured the impulse responses for all eight loudspeakers to the MTB array using the swept-sine measurement software from [3] at a sampling rate of 44.1 kHz. In order to obtain a reasonable signal to noise ratio while preventing damage from the rather small loudspeakers our measurement signal was high pass filtered at 200 Hz. Every time the eight measurements were completed, the microphone was rotated another angular step with a step size according to the intended MTB array solution (i.e. by 360/8 degrees for the virtual 8-microphone-array) and the measurements were repeated. This way we collected impulse responses of all eight loudspeakers at all possible microphone positions of the four virtual MTB arrays. The measurement duration could be reduced as the equiangular 8 and 16 channel microphone arrays form a symmetric subset of the 32 channel MTB array.

Using Matlab®, the four sets of multichannel MTB impulse responses were – independently for each loudspeaker – convolved with anechoic audio stimuli. Thus we obtained ‘virtual’ multichannel MTB recordings which could instantly be rendered using the MTB player and MTB renderer applications. By summing the corresponding MTB channels, we could further generate MTB recordings of several loudspeakers playing different audio material at the same time. From informal auditive assessments, the following three scenarios were chosen for the listening test:

(1) A series of pink noise bursts of 4.5 s length emitted from loudspeaker 1,

(2) a number of sentences spoken by a German male speaker emitted from loudspeaker 1, and

(3) a string quartet playing an excerpt from a tango (1st violin from loudspeaker 8, 2nd violin from loudspeaker 2, viola from loudspeaker 5, violoncello from loudspeaker 4).

Admittedly, the last stimulus does not resemble a very typical spatial setting for a string quartet, but as we also wanted to assess a spatially distributed scene of several synchronously playing sources, it was decided to use the anechoic string quartet recordings which were at hand.

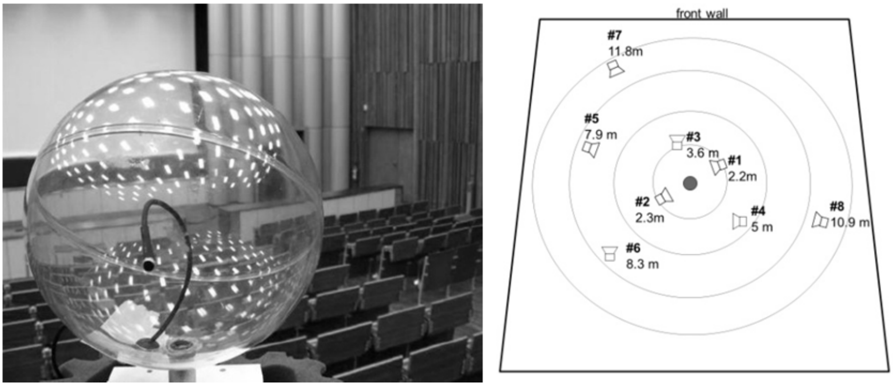


Figure 11-4. Left: Virtual MTB mounted on a remotely rotatable basement in the middle of a lecture hall. Notice the singular microphone mounted in the plastic sphere. Right: Sketch of loudspeaker arrangement used for impulse response measurements with the virtual MTB array (distances are given relative to the virtual MTB's location; circles indicate multiples of the critical distance which was 3.17 m).

11.4 Listening Test

11.4.1 Approaching MTB Sound Quality

The aim of our assessment was to achieve well-differentiated overall quality ratings of our stimuli regarding the involved number of microphones, interpolation algorithm and content. To this end, it was decided to assess MTB sound quality in a MUSHRA-like [12] comparative test design, with MUSHRA (**MU**ltiple **S**timulus with **H**idden **R**efernce and **A**nchors) being a recommended test practice if the deviations within the assessed stimulus pool can assumed to be mostly well perceivable. For each stimulus, subjects are asked to directly rate the intensity of a specific impression with the help of sliders. Typically, multiple stimuli are assessed simultaneously, thus subjects can compare several stimuli while rating. MUSHRA is often used to assess the severity of degradations or deteriorations due to some treatment. Therefore, usually an untreated version of the signal is available as reference. This reference is typically presented to subjects both open and hidden. Additionally, low quality anchors (strongly degraded stimuli) are added to the stimulus pool on purpose. This two-side anchoring strategy supports full usage of the degradation scale and increases inter- and intraindividual consistency. Further, it helps identifying insensitive and unreliable subjects and allows an interpretation of the observed amount of degradation in terms of an absolute difference regarding the untreated signal. However, in case of MTB recordings providing a proper reference stimulus was perceived as problematic, as naturally this would have been the acoustic reality itself, i.e. the impression of listening with one's own ears (cf. last paragraph of sect. 11.6). Hence, as we were mainly interested in differences between the alternatives for MTB sound reproduction, we assumed an explicit external reference not to be necessarily required. Instead, we instructed our listeners to relate their impression of the stimuli to internal references of their own, i.e. to rate the stimuli's perceived deviations from (imagined) equivalent real acoustic events. Two-side anchoring was though – at least partly – realized by using hidden low and high quality stimuli (cf. sect. 11.4.2). This way we consider our test procedure to be an assessment the perceived degree of naturalism, while admitting that the validity of absolute values – e.g., a rating of 60% realism for a specific method – should not be overestimated. However, as results will show, ratings derived from this procedure were pleasingly consistent and allowed a differentiated perceptual analysis of MTB reproduction methods.

11.4.2 Listening Test Design

Three independent variables were assessed within our listening test (tested variations shown in brackets):

(1) DISCRETIZATION: number of microphones (8, 16, 24, 32),

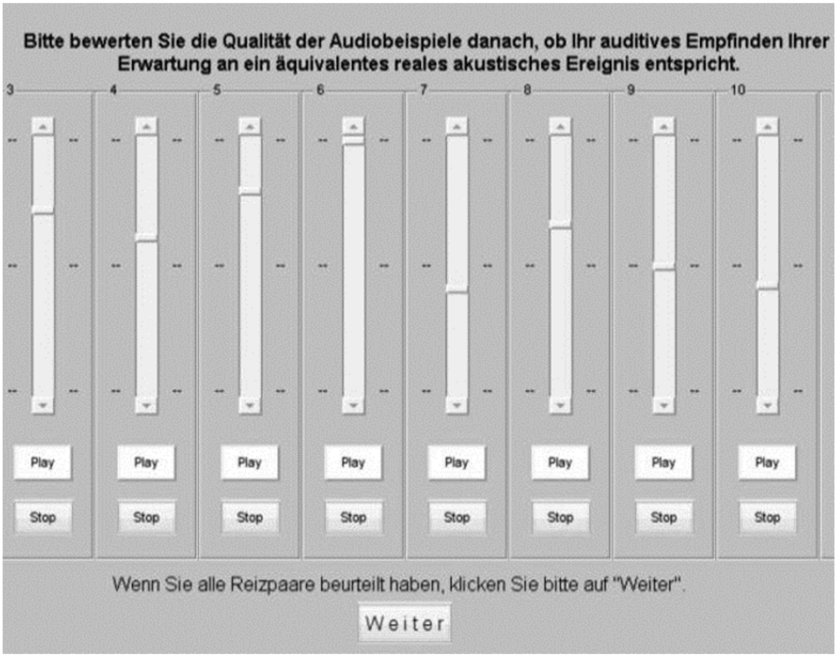
(2) INTERPOLATION: type of interpolation algorithm (TB-FM, FR-NM, FR-LI, TB-NM, TB-SI), and

(3) CONTENT: combining both content and spatial arrangement (pink noise, male speech, string quartet).

The spatial arrangements related to the specific audio contents were described already in sect. 11.3.2. The settings of the interpolation algorithms used in the listening test can be found in sect 11.3.1. In a repeated measures design all 4 (DISCRETIZATION) x 5 (INTERPOLATION.) x 3 (CONTENT) = 60 stimulus combinations were assessed by each of our subjects. As explained before, supporting a wide usage of rating scales, two stimuli were always presented within one set of stimuli rated at a time serving as either a presumably very low (8 channel TB-FM) or high (32 channel TB-SI) quality anchor, respectively.

As the listening test involved only headphone presentations, it was conducted in a small lecture room lacking any special acoustic treatment. The used headphone model was a Sennheiser HD 800; a non-individual headphone compensation created from measurements on our FABIAN artificial head [3] was applied. Audio signals were played back at 44.1 kHz sampling frequency using an M-AUDIO

Delta Audiophile 192 sound card. Headphone playback was calibrated to 65 $\mathrm{dB_{SPL}}$ for the pink noise stimuli; the remaining stimuli were perceptually adjusted toward a comparable loudness. An Intersense InertiaCube head tracker was fixed to the top of the headphones and connected with the rendering computer. Subjects were seated in front of a laptop computer displaying the listening test GUI (cf. Figure 11-5) realized in Matlab®. All audio processing was done on a dedicated rendering workstation (8-threaded IntelCore i7, Linux OS, 12 GB RAM). OSC messages controlling the listening test progress and all rendering parameters were sent via Ethernet from the laptop to the rendering computer.



Bitte bewerten Sie die Qualität der Audiobeispiele danach, ob Ihr auditives Empfinden Ihrer Erwartung an ein äquivalentes reales akustisches Ereignis entspricht.

Wenn Sie alle Reizpaare beurteilt haben, klicken Sie bitte auf "Weiter".

Figure 11-5. Graphical user interface for of MUSHRA test. The briefing sentence (in German, see text for translation) was always displayed at the top of the panel of sliders.

The GUI consisted of a row of continuous sliders and ‘play’ and ‘stop’ buttons for each stimulus. Subjects could switch instantly between different stimuli, taking their time for rating at will. Stimuli were presented in 6 successive panels each displaying 12 sliders (in sum 72 presentations). Within two successively presented panels of sliders the audio content was kept constant. Panel order and stimulus-slider assignment was though randomized for each subject. The number of 72 presentations occurs as the 'high' and 'low' quality anchor stimuli were included additionally in each panel. At the beginning of the listening test, a training panel was presented to make each subject known to the variety of stimuli within the listening test. As internal references cannot be monitored and will depend on individual experiences and expectations. As an attempt to balance this influence, all subjects received a similarly instruction. Using a written German text, naturalism was introduced as the amount a stimulus corresponded to the expectation of an equivalent real acoustic event. Throughout the whole test procedure a short briefing sentence, repeating the instruction in German, was displayed above the sliders of the rating panel (cf. Figure 11-5, translation: "Please rate the quality of the audio examples in terms of how far your auditory sensation equals your expectation of a corresponding real event.") Listeners were instructed to rate stimuli in a holistic manner using the sliders labeled from 'fully' to 'not at all' (in German) at the corresponding ends. Experimenters further emphasized that, as far as possible, before switching to the next panel, slider positions should reflect a rank order of the perceived degree of naturalism for the stimuli assessed within a panel.

11.4.3 Statistical Hypotheses and Calculation of Sample Size

As theoretical knowledge about perceptual foundations of MTB is rather complete, directed statistical hypotheses could be formulated *a priori*. Additionally, for the determination of sample size, practically relevant effect sizes were defined. Basically, we expected the following main effects: Firstly, perceived naturalism should increase with decreasing stimulus bandwidth (i.e. from noise stimulus over the string quartet to the speech stimulus). Secondly, naturalism should increase with the number of microphones, and thirdly (in accordance with [1], cf. sect. 11.2.4) it should increase for the following order of interpolation algorithms: TB-FM, FB-NM, FB-LI, TB-NM, TB-SI (in the following also designated algorithms 1 to 5). We also expected some interaction effects, as, e.g., that for certain combinations of stimulus and interpolation algorithm an increasing number of microphones would not increase perceived naturalism any further (saturation, 2nd order interaction). In the scope of our study beign fundamental research we considered small effects as practically relevant. Using the Gpower software [13] we calculated sample sizes for our repeated measures design allowing testing small effects at a 5% type-1 error level with a power of 80% ([14], p. 606, pp. 618, pp. 630). Hence, if conservatively assuming an average correlation of ρ_{mean} = 0.25 for all pairwise dependent series of ratings, testing a small effect for least gradated main effect (CONTENT) would require 13 subjects. To test a small effect for the highest (i.e. 2nd) order interaction

(DISCRETIZATION x INTERPOLATION x CONTENT) 29 subjects would be needed.

11.4.4 Participants and Test Duration

Finally, 26 subjects (84% male) of an average age of 29.5 years took part in our test. Most of them had prior experience with listening tests, and a musical education of on average more than ten years. Normal hearing was assessed via self-reports. Including instruction and training the average test duration was approximately 40 minutes.

11.5 Results

Following recommendations in [12] we post-screened our raw data. After checking normality of ratings with Matlab's Lilliefors test, we conducted two-sided Grubb's outlier tests. Both tests were done at a 5% type-1 error level. Subsequently we checked distributions of individual ratings to identify subjects lacking variability in their ratings or showing clusters of extreme ratings. The Grubb's test found an increased number of outlier ratings for three subjects (eight, eight and five times). However, as eight times was considered a rather small fraction of the 60 ratings per individual and as there were no other problems identified, we concluded that there was no reason to exclude any subject from further analysis.

With the help of the SPSS® software package we calculated the intraclass correlation $ICC(2,k)$ [16] for the raw ratings as a measure of our subjects' agreement. The observed value of 0.919 was fairly high. In conjunction with the well-differentiated test results, it indicates, that test design and instructions enabled subjects to rate stimuli in a highly consistent manner.

As our stimuli contained no intermediate quality anchor we followed recommendations from [17] to standardize individual ratings. All further results from inferential statistics were derived from these standardized ratings. Eventually, we found that statistical results from both raw and standardized data were nearly identical. Figure 11-6 shows the results as average relative raw ratings with 95% confidence intervals.

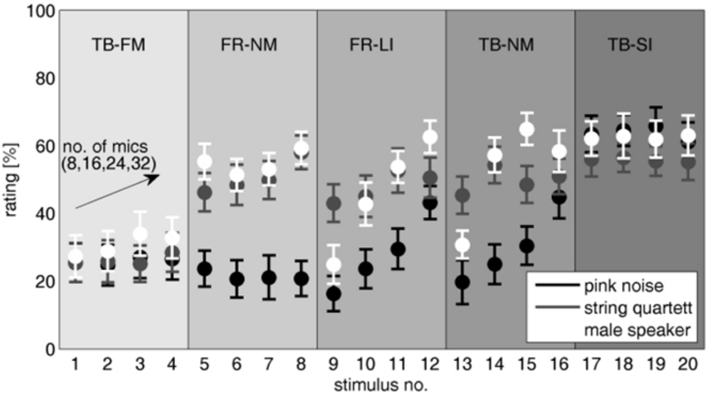


Figure 11-6. Listening test results: Raw ratings of relative perceived naturalism of MTB stimuli. Results are displayed as means with 95% confidence intervals. Results are ordered for the three independent variables (1) number of microphones, (2) interpolation algorithm and (3) audio content. Results are ordered from interpolation algorithm 1 to 5 in the grey subsections. Within each subsection the number of microphones increases from 8 to 32.

For both INTERPOLATION and DISCRETIZATION stimuli have been ordered from left to right in order of increasing overall ratings as expected from our *a priori* hypotheses (cf. sect. 11.4.3). Further, from non-aggregated data we found that the scales' ranges were fairly well exploited, although we observed a tendency toward clustering at the bottom end. Hence, average ratings seldom exceeded 60%. This fact should not be overestimated though, because, as explained already in section 11.4.1, an external reference was missing. Moreover, subjects were well aware of listening to recordings only; they did not expect to hear a stimulus being fully identical to their expectation which might have led to avoiding the upper scale range. However, in written comments collected after the listening test, several subjects mentioned a permanent perception of elevation and of unstable source localization. From section 11.2.3 we know, that these artifacts are due to missing pinnae and a mismatch in sphere diameter resulting in erroneous ITD cues. These artifacts may also have caused a limitation of MTB's naturalism ratings. When averaging over all three contents, our anchor stimuli (no. 1 and no. 20 in Figure 11-6) were indeed rated similar as the worst or best stimuli, respectively. They thus can be expected to have served as intended.

All effects were tested for significance by means of a repeated-measures analysis of variances (ANOVA). Data requirements for ANOVA were verified using Mauchly's test of sphericity. Accordingly, significance ratings were corrected for degrees of freedom where needed. It was found that all main effects and all first order interactions were highly significant. Directive contrasts for the main effects further approved all *a priori* hypotheses at a highly significant level. Hence, on average, perceived naturalism of MTB increased with the number of microphones, with decreasing stimulus bandwidth, and following the perceptually motivated order of interpolation algorithms as proposed by Algazi et al. [1]. However, *post hoc* pairwise comparisons with Bonferroni adjustment for multiple comparisons put these results into perspective, as (1) the CONTENT effect was governed mainly by the noise stimuli's ratings differing from that of the other two stimuli, (2) the INTERPOLATION effect was formed by threefold grouping of the five interpolation algorithms (algorithm 1 was rated worse than all others, followed by algorithms 2 to 4 evaluated to be of similar performance (with FR-LI being on average rated even slightly worse than FR-NM) and algorithm 5 rated best), and (3) the 8-microphone condition (DISCRETIZATION) was rated worse than all other combinations, while both 24 and 32 microphones were on average rated similarly well. However, as all 1st order interactions (DISCRETIZATION x INTERPOLATION, DISCRETIZATION x CONTENT, and INTERPOLATION x CONTENT) were significant too, they were analyzed for contradictions before accepting these main effects.

We will start the discussion with the 1st order interaction DISCRETIZATION x INTERPOLATION as it is both instructive and most interesting as seen from the scope of the study. This interaction is expressed by the group-wise similar ratings of interpolation algorithms with regard to the number of microphones (Figure 11-6). Hence, despite an increasing number of microphones ratings remained nearly constant for algorithms 1, 2, and 5 (TB-FM, FR-NM, TB-SI), whereas for algorithms 3 and 4 (FR-LI, TB-NM) average ratings increase with the number of microphones as expected. Ratings for algorithms 1 and 5 (TB-FM, TB-SI) were – independent from content –either constantly worse or rather good. For algorithm 2 (FR-NM) ratings were also independent from the number of microphones, yet they were depending on content, which will be discussed further below.

Already from these results we can immediately identify a clear 'winner' algorithm. Two Band Spectral Interpolation Restoration (TB-SI) performed better or at least as good as all other combinations of microphones and interpolation algorithms. What is even more, TB-SI performed equally well even for most critical signals and a minimum number of microphones. Hence, upper band magnitude crossfading and phase switching prove to be perceptually superior approaches even for coarse angular resolution. The robustness of magnitude crossfading to spatial discretization is most probably due to missing pinnae of the MTB, leading to a reduction of direction-dependent spectral detail. Additionally, for lower number of microphones phase switching obviously leads to less audible artifacts than full-range switching, upper band switching or linear crossfading.

In contrast, Two Band Fixed Microphone Interpolation (TB-FM) was rated inferior under all tested conditions. This is most probably due to the monaural high frequency components leading to constant in-head-localization and a 'split' perception of low and high frequency content. The overall annoyance of FR-NM's (algorithm 2) full range switching artifacts was also not improved by increasing the number of microphones. This might have been due to the fact that, although the number of microphones influences the frequency of switching events during head movements, the velocity of head movement and in turn the switching frequency is seldom constant.

As mentioned already, algorithms 3 and 4 (FR-LI, TB-NM) formed a second group of the interaction DISCRETIZATION x INTERPOLATION. Both were similarly perceived as, on overall, increasingly natural with increasing number of microphones. For FR-LI this is explained straight forward as cross fading artifacts are being pushed into ever-higher and less audible frequency regions. However, for TB-NM (high frequency switching) this overall effect can be explained better with regard to the audio content (see below).

As typical symptom of the second significant 1st order interaction INTERPOLATION x CONTENT algorithms 1 and 5 were rated nearly independent from content, whereas for algorithms 2 to 4 average ratings for noise were always worse. While exhibiting longer steady state sections, with noise, artifacts of algorithm 2 (full range switching) were always clearly audible. In contrast, with speech and music switching appeared less audible, probably as these signals contain more amplitude modulation. Indeed subjects mentioned that, depending on content, switching artifacts were sometimes hard to hear at all, an effect, which was also observed for algorithm 4 (upper band switching).

The third and last significant 1st order interaction DISCRETIZATION x CONTENT was expressed in ratings of the speech stimulus being – when averaged over all interpolation algorithms – sensitive above average to the number of available microphones. Speech ratings showed the widest spread, in the case of eight microphones it was on average rated even worse than the musical stimulus. This pronounced sensitivity to naturalism of speech could be explained with being an exceptionally familiar type of signal.

Additionally, we found a trend (not significant) toward the 2nd order interaction INTERPOLATION x DISCRETIZATION x CONTENT. However, while considering practical relevance, we refained from further discussions.

11.6 Summary and Conclusion

In 2004, Algazi et al. [1] introduced motion-tracked binaural (MTB) sound as a new method for capturing, recording, and reproducing spatial sound. Their study was accompanied with thorough quantitative analyses of potential approaches to interpolation and discretization. Melick et al. [10] discussed perceptual shortcomings of MTB and presented possible remedies. Nevertheless, both studies lacked a formal perceptual assessment of MTB sound quality. Starting from here, we introduced the perceived degree of naturalism as a suitable criterion for assessing interpolation and discretization of MTB sound. We described a method to synthesize MTB stimuli systematically varying in (1) audio content, (2) interpolation algorithm, and (3) the number of used microphone channels and presented results from a listening test with 26 subjects. In a thorough analysis of results we showed the naturalism of motion-tracked binaural sound to be highly interdependent on all three parameters.

For the five tested interpolation algorithms the degree of naturalism indeed increased as expected by Algazi et al. [1]. However, the highly content-dependent ratings of algorithm 2 (Full Range Nearest Microphone Selection, FR-NM) appeared to be a special case. For modulated natural signals only it was even found to be suited second best. The benefit of the number of microphones has – although naturalism increased on average with the number of microphones – to be judged with regard the specific interpolation algorithm. Differences in audio content ratings seemed to a lesser degree be due to differences in bandwidths than due to amount of modulation in the stimuli, with amplitude modulated signals being far less demanding for reproduction. Moreover, the speech stimulus was found a special case as its ratings were most sensitive for the number of microphones used.

The probably most surprising finding was the clearness of the superior behavior of Two Band Spectral Interpolation Restoration (TB-SI). If audio quality is of prime importance and there are no constraints in available processing power, the TB-SI algorithm should always be preferred, especially as its perceptual performance is nearly independent from the number of microphones (from 8 up to 32) and type of audio content (steady and modulated).

If processing power is limited, but quality of critical signals is still important, algorithms 3 and 4 (Full Range Linear Interpolation FR-LI, Two Band Nearest Microphone Selection, TB-NM) perform nearly equally well, but signals should be recorded with the highest possible number of microphones. Algorithm 1 and 2 (Two Band Fixed Microphone Interpolation TB-FM, Full Range Nearest Microphone Selection, FR-NM) should never be used if high quality transmission of critical audio content is aimed at. However, if the application is limited to transmission of speech, algorithms 2, 3 and 4 (FR-NM, FR-LI, TB-NM) may be used whereas algorithms 3 and 4 should not be applied using less than 16 microphones. If bandwidth and processing power are severely limited, for speech transmission at least, algorithm 2 (FR-NM) applied with 8 microphones might be recommendable.

With the identification of a superior configuration for MTB sound reproduction, two questions were found promising for future examination: First, it would be interesting to assess MTB's perceptual transparency in absolute terms, i.e. in comparison to listening with one's own ears. Such a comparison can be realized by means of individual dynamic binaural synthesis. Besides, the operational requirements for adequate simulations are extensive: Using insert microphones one would have to measure sets of individual binaural room impulse responses for controlled head orientations. However, the effort could be limited as potentially only few subjects would be needed for such an assessment. Secondly, – and at best within the same listening test – it would be interesting to assess, formally and in detail, qualities and intensities of perceptual shortcomings occurring with MTB rendering and to discover in how far they constitute limitations of MTB's naturalism. A qualitative expert vocabulary which could be used for that purpose is currently being developed by our group.

11.7 Acknowledgements

Alexander Lindau was supported by a grant from the Deutsche Telekom Laboratories and by the Deutsche Forschungsgemeinschaft (DFG), grant WE 4057/1-1.

11.8 References

[1] Algazi, V. R.; Duda, R. O.; Thompson, D. M. (2004): “Motion-Tracked Binaural Sound”, in: *J. Audio Eng. Soc.,* **52**(11), pp. 1142-1156

[2] Karamustafaoglu, A.; Horbach, U.; Pellegrini, R. S.; Mackensen, P.; Theile, G. (1999): “Design and applications of a data-based auralization system for surround sound”, in: *Proc. of the 106th AES Convention*, Munich, preprint no. 4976

[3] Lindau, A.; Hohn, T.; Weinzierl, S. (2007): “Binaural resynthesis for comparative studies of acoustical environments”, in: *Proc. of the 122nd AES Convention*, Vienna, preprint no. 7032

[4] Smyth, S. M. (2006): *Personalized headphone Virtualization* (US Patent Application Publication). US 2006/0045294 A1

[5] Hom, R. C.-M.; Algazi, V. .; Duda, R. O. (2006): “High-Frequency Interpolation for Motion-Tracked Binaural Sound”, in: *Proc. of the 121st AES Convention*. San Francisco, preprint no. 6963

[6] Strutt, J. W. (1907): “On Our Perception of Sound Direction”, in: *Philosophical Magazine,* **13**, pp. 214–232

[7] Wightman, F.; 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

[8] Takemoto, H.; Mokhtari, P.; Kato, H.; Nishimura, R. (2012): “Mechanism for generating peaks and notches of head-related transfer functions in the median plane”, in: *J. Acoust. Soc. Am.*, **132**(6), pp. 3832-3841

[9] Lindau, A.; Estrella, J.; Weinzierl, S. (2010): “Individualization of dynamic binaural synthesis by real time manipulation of the ITD”, in: *Proc. of the 128th AES Convention*. London, preprint no. 8088

[10] Melick, Joshua B. et al. (2004): “Customization for Personalized Rendering of Motion-Tracked Binaural Sound”, in: *Proc. of the 117th AES Convention*. San Francisco, preprint no. 6225

[11] Wright, M.; Freed, A.; Momeni, A. (2003): “Open Sound Control: State of the art 2003”, in: *Proc. of the 2003 Conference on New Interfaces for Musical Expression (NIME-03),* Montreal

[12] ITU-R Rec. BS.1534-1 (2003): *Method for the subjective assessment of intermediate quality level of coding systems*, Geneva: International Telecommunication Union

[13] Faul, Franz et al. (2007): "G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences", in: *Behavior Research Methods*, **39**(2), pp. 175-191

[14] Bortz, Jürgen; Döring, Nicola (2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*, 4th ed., Heidelberg: Springer

[15] Bortz, Jürgen (2005): *Statistik für Sozial- und Humanwissenschaftler*. 6th ed., Heidelberg: Springer

[16] Shrout, Patrick E.; Fleiss, Joseph L. (1979): "Intraclass Correlations: Uses in Assessing Rater Reliability", in: *Psychological Bulletin*, **86**(2), pp. 420-428

[17] Bech, S.; Zacharov, N. (2006): *Perceptual Audio Evaluation: Theory, Method and Application.* Chichester: Wiley

Part III
Perceptual Evaluation of Virtual Acoustic Environments

12 Assessing the Plausibility of Virtual Acoustic Environments

The following chapter is an authorized reprint of the full-paper peer-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander; Weinzierl, Stefan (2012): "Assessing the Plausibility of Virtual Acoustic Environments", in: *Acta Acustica united with Acustica*, **98**(5), pp. 804-810, DOI: http://dx.doi.org/10.3813/AAA. 918562.
>
>

The definitive publisher-authenticated version is available online at http://www.ingentaconnect.com/content/dav/aaua. Please note that readers must contact the S. Hirzel Verlag for reprint or permission to use the material in any form.

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style and minimal typographic corrections.

12.1 Abstract

Aiming at the perceptual evaluation of virtual acoustic environments (VAEs), 'plausibility' is introduced as a quality criterion that can be of value for many applications in virtual acoustics. We suggest a definition as well as an experimental operationalization for plausibility, referring to the perceived agreement with the listener's expectation towards a corresponding real acoustic event. The measurement model includes the criterion-free assessment of the deviation from this non-explicit, inner reference by rating corresponding real and simulated stimuli in a Yes/No test paradigm and analyzing the results according to signal detection theory. The specification of a minimum effect hypothesis allows testing of plausibility with any desired strictness. The approach is demonstrated with the perceptual evaluation of a system for dynamic binaural synthesis in two different development stages.

12.2 Introduction

Dynamic binaural synthesis has reached a high degree of realism in the simulation of acoustic environments. The actual quality of binaural simulations is, however, typically assessed by observing only singular aspects of spatial hearing, e.g., by comparing the localization accuracy in real and simulated sound fields [1]–[6] or by comparing to references which are simulations themselves [7]. The outcome of such experiments seems, therefore, hardly suited as a measure in how far the binaural synthesis as a whole is able to provide substitutes for real sound fields.

More holistic measures are provided by analyzing the ‘immersion’ or a ‘sense of presence’ of subjects in virtual environments. Widely used in the evaluation of visual virtual environments (and up to now only incidentally assessed in acoustics), these features are generally interpreted as multidimensional constructs including aspects such as an experience of spatial presence (‘being there’), a sense of ‘involvement’ and a judgment of ‘realness’ [8], [9]. Some of these underlying facets are, however, strongly related to the quality of the presented content, the targeted task, the provided modes of interaction, the usability of the interface applied, and the personality of the users addressed – hence it might remain unclear which part of the simulation is actually addressed by the user’s evaluation. For the purpose of system development and evaluation, these constructs thus seem less appropriate.

As a system-oriented criterion, ‘authenticity’ was suggested by Blauert ([10], p. 373) referring to the perceptual identity between simulation and reality. Among all potential perceptual quality criteria for VAEs, authenticity would thus make maximum demands on the performance of virtual environments. The necessary immediate comparison between simulation and reality can, however, not always be realized experimentally, and it is, at the same time, not required for applications where users do not have access to any external reference. An appropriate criterion for most applications could be the ‘plausibility’ of a virtual environment, which we would thus define as

a simulation in agreement with the listener’s expectation
towards a corresponding real event.

Referring to an inner reference as the result of each listener’s personal experience and expectation [11] rather than to the exact physical or perceptual identity of reality and simulation, plausibility corresponds well to the situation in which most users are exposed to virtual environments. Moreover, it does not require the rating of potentially ambiguous qualities such as 'immersion' or 'spatial presence' but only a certain familiarity with the corresponding real acoustic environments. Thus, a high consistency and external validity of the test can be expected.

The concept of plausibility for the evaluation of virtual environments has been suggested quite frequently. According to Pellegrini, a plausible simulation of a given environment would include "a suitable reproduction of all required quality features for a given specific application" rather than a copy of "an existing environment in all its physical aspects" [12]. In the context of game audio, Reiter considered a plausible simulation to be given, "as long as there is no obvious contradiction between the visual and the acoustic representation of a virtual scene", allowing "the human senses [to] merge auditory and visual impressions" [13]. Both authors thus emphasize the role of the application and the task for the assessment of plausibility, which is not contradictory to our definition, if one takes into account, that the listener's expectation towards an acoustic environment will, of course, be influenced by other perceptual modalities and by the task he is supposed to perform. The challenge, however, lies in finding an appropriate measurement model to quantify the degree of plausibility achieved, including a definition, an operationalization and a statistical analysis of perceptual data.

12.3 Plausibility: An Experimental Approach

The plausibility of virtual environments could theoretically be rated directly on a linear scale with values between '0' and '1'. However, due to personal theories about the credibility of virtual realities and the performance of media systems in general, a strong and inter-individually different response bias can be expected, in particular when no explicit reference is available. Therefore, a criterion-free assessment of plausibility is essential. Following the definition given above, requiring the evaluation of a simulation with regard to an inner reference, any forced choice paradigm is precluded, because it would require a direct comparison with an external, given reference. Decisions with regard to an inner reference can, however, be collected by using a Yes/No paradigm for the evaluation of both simulated and real stimuli and by removing the response bias *ex post* with an analysis according to signal detection theory.

12.3.1 Plausibility as a Signal Detection Problem

Signal detection theory (SDT), originally used as a perceptual model for the detection of weak signals in the presence of internal noise [14], has later been generalized to model perceptual and decisional processing in stimulus identification, categorization or similarity judgment [15]. Hence, it seems tempting to adapt the SDT concept to the discrimination task involved in evaluating the plausibility of a virtual environment. In our case, the reality takes the role of the ‘no signal’ condition, while the simulation represents the ‘signal’ condition, assuming that the latter contains small artifacts (the ‘signal’) compared to the original. This is, however, an arbitrary assignment which has no effect on the statistical analysis.

As in standard SDT approaches, we use a simple observer model which assumes Gaussian probability density distributions of equal variance representing the ‘reality’ (noise alone) and the ‘simulation’ (signal plus noise) condition on a horizontal axis measuring the internal response to the presented stimulus (Figure 12-1). The sensory difference of the two stimuli is expressed by the distance between the two distributions (sensitivity d' in Figure 12-1). The individual response bias, i.e. the tendency to regard the simulation as reality or vice versa, is reflected by an individually differing position of the response criterion λ_i. If the sensation level is perceived to be above this criterion, the observer will give a positive response.

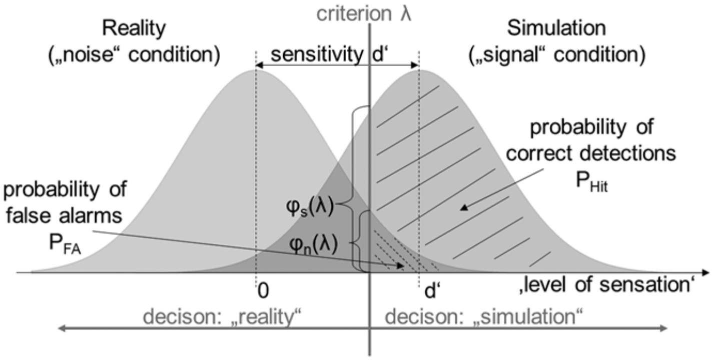


Figure 12-1. Parameters of the equal-variance Gaussian signal detection model, adapted to the evaluation of ‘plausibility’ for virtual environments.

Hence, observers with $\lambda_i > d_i'/2$ show a conservative answering behavior, i.e. a tendency to believe in the realness of the stimulus, whereas subjects with $\lambda_i < d_i'/2$ will respond more progressively, i.e. consider even the real stimulus as simulated. A criterion of $\lambda_i = d_i'/2$ would indicate a perfectly balanced (‘fair’) observer. Applying the inverse cumulative normal distribution $Z(p)$, the individual criterion λ_i and the sensitivity d_i' can be estimated (with ^ indicating estimated variables) based on the rate of false alarms p_{FA} and correct detections p_{Hit} as

$$\hat{\lambda}_i = Z(1 - p_{FA_i}), \tag{12-1}$$

and

$$\hat{d}_i' = Z(p_{Hit_i}) - Z(p_{FA_i}). \tag{12-2}$$

An alternative measure of bias is the ratio of the values of normalized noise and signal probability density at the position of the criterion,

$$\hat{\beta}_i' = \varphi_s(\hat{\lambda}_i)/\varphi_n(\hat{\lambda}_i) = \varphi(\hat{\lambda}_i - \hat{d}_i')/\varphi(\hat{\lambda}_i). \tag{12-3}$$

In contrast to λ_i it allows a more direct interpretation of the bias value (independent of d_i'): Subjects exhibiting $\beta_i < 1$ have the tendency to report "Yes" (simulation), whereas $\beta_i > 1$ indicates a "No" (reality) tendency.

12.3.2 Minimum Effect Hypothesis and Optimal Sample Size

In terms of the SDT observer model, showing 'perfect' plausibility would require proving a sensitivity of $d' = 0$. From the view of inferential statistics, however, a direct proof of the null hypothesis H_0 is impossible. One can only reject a specific alternative hypothesis H_1 by rejecting an effect that is small enough to be regarded as perceptually irrelevant (a minimum-effect hypothesis, [16]).

Values of d' cannot easily be interpreted for the formulation of a meaningful minimum effect hypothesis. They can, however, be directly related to detection rates of nAFC test paradigms, as both Yes/No and nAFC paradigms can be described using the same probabilistic representation [17]. For an equal variance Gaussian signal detection model the probability of correct responses P_c in the 2AFC paradigm and the sensitivity d' are related by

$$P_c = \Phi(d'/\sqrt{2}), \tag{12-4}$$

and

$$d' = \sqrt{2} \cdot Z(P_c), \tag{12-5}$$

with $\Phi(z)$ the cumulative standard normal distribution. These relations allow formulating hypotheses more intuitively in terms of 2AFC detection rates P_c (see Table 12-1 for corresponding values).

Table 12-1. 2AFC detection rate P_c and corresponding sensitivity parameter d' for the equal variance Gaussian signal detection model

$\boldsymbol{P_c}$	0.50	0.55	0.60	0.65	0.70	0.75
$\boldsymbol{d'}$	0	0.1777	0.3583	0.5449	0.7416	0.9539

For the exemplary listening test described below we assumed the simulation to be plausible if the probability of correct responses in an equivalent 2AFC test design were less than P_c = 0.55, i.e. exceeding the pure guessing rate by less than 5%. It should be noted that this yields a far stricter criterion than determining the inflection point of the psychometrical function (at P_c = 0.75) commonly targeted as a population's difference or detection threshold.

For determining the sample size necessary to test a specific effect size d' with a given type I and type II error level, a model for the variance of d' is needed. The usual approach is to start with an estimate of the variance of the hit rate p_{Hit} and of the false alarm rate p_{FA} which follow a multinomial distribution, i.e.

$$\hat{\sigma}^2(p_{Hit}) = \frac{p_{Hit}(1-p_{Hit})}{N_s} \tag{12-6}$$

and

$$\hat{\sigma}^2(p_{FA}) = \frac{p_{FA}(1-p_{FA})}{N_n}, \tag{12-7}$$

with N_n and N_s denoting the number of noise and signal presentations, respectively ([17], p. 202). Assuming that d' is the sum of two statistically independent terms in (12-2), the variance of d' is the sum of the variance components. Approximating the nonlinear relation $y = Z(x)$ with a Taylor series truncated after the linear term, then yields

$$\hat{\sigma}^2(\hat{d}') = \frac{\hat{\sigma}^2(p_{Hit})}{\varphi^2(\hat{d}'-\hat{\lambda})} + \frac{\hat{\sigma}^2(p_{FA})}{\varphi^2(\hat{\lambda})}, \qquad (12\text{-}8)$$

with $\varphi(z)$ denoting the standard normal distribution

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} . \qquad (12\text{-}9)$$

Since the derivation of (12-8) is based on a Gaussian model for the detection task, it can be regarded as a large-sample approximation which will be inaccurate when the number of trials is small and when p_{Hit} or p_{FA} approach one or zero. The variance of d' thus depends on the value of d', on the position of the criterion λ, and on the hit and false alarm rates resulting from the presented ratio of noise and signal conditions. For an equal number of noise and signal conditions ($N_n = N_s = 0.5N_{tot}$) and an unbiased observer ($\lambda = d'/2$ and $p_{Hit} = 1 - p_{FA}$), (12-8) reduces to

$$\hat{\sigma}^2(\hat{d}') = \frac{4\Phi(\hat{d}'/2)\Phi(-\hat{d}'/2)}{N_{tot}\varphi^2(\hat{d}'/2)}, \qquad (12\text{-}10)$$

and for small sensory differences ($\hat{d}' \to 0$), approaches

$$\hat{\sigma}^2(\hat{d}') = \frac{2\pi}{N_{tot}}. \qquad (12\text{-}11)$$

Since any hypothesis testing of specific d' values is based on the estimated variance, the required sample size depends on the type I error (rejecting H_0 although it is true) and the type II error (rejecting H_1 although it is true) that seems acceptable for the current investigation. The optimum sample size can be calculated from letting the one-sided confidence intervals of the null hypothesis ($d_0' = 0$) and minimum effect hypotheses ($d' = d_{min}'$) for the given type I and II error levels touch each other, i.e. when

$$d'_0 + se(d'_0)z_\alpha = d'_{min} - se(d'_{min})z_\beta \ , \qquad \text{(12-12)}$$

with the standard error $se(d') = \sqrt{\sigma^2_{d'}}$ and z_α, z_β denoting the z values for the given type I (α-) and type II (β-) error. With $d'_0 = 0$ and using (12-11) for $\hat{\sigma}^2(\hat{d}')$ one can solve (12-12) for N_{opt}, obtaining

$$N_{opt} = (z_\alpha + z_\beta)^2 \frac{2\pi}{\hat{d}'^2_{min}} \ , \qquad \text{(12-13)}$$

Since the approximation for small d' values is valid for $d' \leq 1$ ([17], p. 205), (12-13) should not be applied for large effect sizes ($d' > 1$). Instead, (12-10) could be used to derive a more complex expression for N_{opt}. Because (12-13) is based on the assumption of a perfectly unbiased observer with stationary response behavior and an equal number of noise and signal conditions presented, any – in practice inevitable – deviation from this assumption will result in an increased variance of hit and false alarm rates and thus in larger sample sizes required to confirm the targeted effect. The estimated sample sizes should thus be regarded as lower boundaries. For inferential statistical analysis of the collected data this is, however, not relevant, since the t-test to be conducted with the collected d' values (see below) takes into account all sources of variance.

As an illustration, Figure 12-2 shows the resulting sample sizes N_{opt} as a function of d'_{min} (the minimum effect hypothesis) for different combinations of type I and type II error levels.

If neither of the two hypotheses shall be favored, both, type I and type II error levels (i.e. α and β) could be set to 5%. If the test aims at proving the null hypothesis (the system is plausible), it can, however, be acceptable to reduce the demands on the type I error level to 10%, 20% or even 25%, if the required sample size otherwise gets unreasonably large. For the test described below, we tested an assumed 2AFC detection rate of $P_c = 0.55$ corresponding to $d' = 0.1777$ with a type I/type II error level of 0.25/0.05, i.e. with 95% test power, resulting in a number of 1071 singular decisions required according to (12-13).

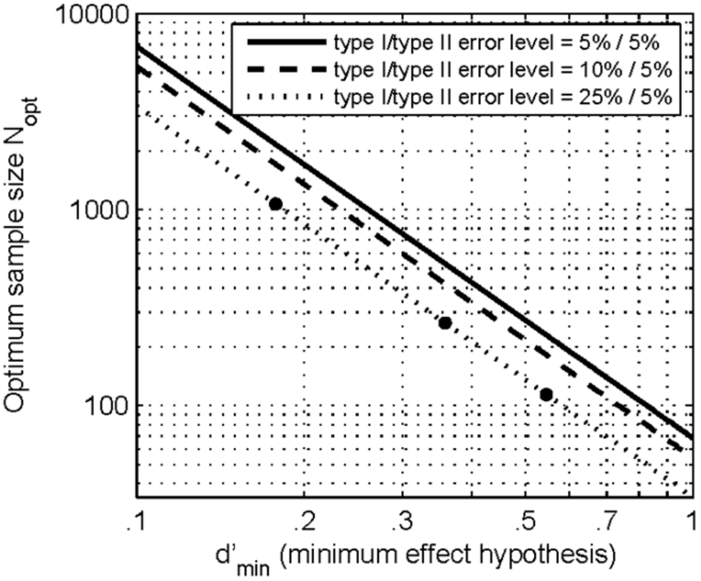


Figure 12-2. Optimum sample size N_{opt} allowing a clear decision in favor of the null hypothesis (d'= 0, 'plausible') or in favor of the alternative hypothesis ($d' = d'_{min}$, 'implausible') for a given type I/type II error level and d'_{min} = [0.1, 1]. We have assumed an equal number of noise and signal conditions (N_n = N_s = 0.5N_{tot}) and an unbiased observer ($\lambda = d'/2$ and $p_{Hit} = 1 - p_{FA}$). Markers on the 25% / 5%-line are highlighting N_{opt} and d' values corresponding to the 2AFC detection rates P_c = [.55 .60 .65] (from left to right).

12.3.3 Aggregation of SDT Indices

SDT tests are typically conducted with only few, well trained subjects. If the results analyzed individually for each subject lead to the same conclusions, they are considered to be generalizable. In the context of our study, it appeared more adequate to evaluate the plausibility of VAEs using a larger and therefore more representative sample of subjects. As the calculation of d' involves nonlinear transformations the calculation of an average sensitivity d'_{avg} from the pooled Yes/No decisions of all observers does not give the same result as when averaging over d'_i values of different observers. To obtain a measure of the average sensitivity of the group, individual values for d'_i have to be calculated first and then averaged to get d'_{avg}. The within-group variance of the different d'_i values will then measure both the individual and the intersubject variability [17]. If not only the overall performance but also the individual results are of interest, e.g., because the simulation might be plausible for some subjects and not for others, the individual d'_i values can be analyzed separately. In this case, a number of at least 100 decisions is considered necessary for obtaining stable individual SDT parameters [18]; this is a result of the hit and false alarm otherwise depending on very small figures. In the current test we therefore used a sample of 11 subjects with 100 decisions each, resulting in a total of 1100 decisions.

12.4 Listening Test Setup

12.4.1 Realizing the Binaural Simulation

The presented approach towards assessing the plausibility of virtual environments requires a test setup where real and simulated stimuli can be presented in the same listening environment. To guarantee constant test conditions, both real and simulated stimuli were generated from pre-recorded audio material and electro acoustical sound sources placed in a large lecture hall (TU Berlin, auditorium maximum, V = 8500 m^3, RT = 2.0 s, r_{crit} = 3.6 m). Five mid-size active loudspeakers (Meyersound UPL-1) were placed at different positions on the stage, floor, and balcony areas (Table 12-2).

The head and torso simulator (HATS) FABIAN [20] with freely moveable artificial head above torso was placed at a central seat in the frontal half of the audience area.

Table 12-2. Positions of loudspeakers used in the listening test given in coordinates relative to the listener position (right handed spherical coordinate system, azimuth/elevation = 0°/0° equals frontal viewing direction).

Loudspeaker	Distance	Azimuth	Elevation
1	9.5 m	0°	6°
2	12 m	-62°	-1°
3	16.5 m	-113°	20°
4	13.3 m	-175°	34°
5	11.5 m	131°	7°

Datasets of binaural room impulse responses (BRIRs) were measured individually for each of the five loudspeakers and for horizontal head movements in a range of ±80° with an angular resolution of 1°. Distances between loudspeakers and the central listener seat varied between approximately 3–5 times r_{crit}.

For the listening test, subjects were placed at the same seat as the HATS. Dynamic auralization was realized using a fast convolution algorithm with head-tracking

[20]. To hide the presentation mode, subjects kept their headphones on throughout the test. This was enabled by letting the dummy head wear acoustically relatively transparent headphones (STAX SR-202 Basic) throughout the BRIR measurements.

12.4.2 Testing Two Development Stages of a VAE

A system for dynamic binaural synthesis was perceptually evaluated in 2007 [20] with encouraging but not fully satisfying results (termed: basic simulation). Following qualitative reports of perceived deficiencies such as spectral coloration, latency, instability of localization, and cross fade artifacts, several technical improvements were implemented (termed: improved simulation). These include a perceptually optimized headphone compensation [21], a reduced system latency below the just audible threshold [22], a reduction of cross fade artifacts and localization instability by individualizing the interaural time delay by means of ITD extraction and manipulation and by replacing BRIRs with minimum phase representations [23], and perceptually validated thresholds for the transition of dynamic and static parts of the room impulse response [24]. For an exemplary and comparative realization of the test design introduced above, plausibility was tested for both simulator stages in two independent listening tests. Average system latency was reduced from 112 ms to 65 ms, measured according to the procedure described in [22]. Individualization of the ITD was realized using the procedure described in [23], i.e. by measuring individual head diameters (intertragus distances) of subjects prior to the listening test.

12.4.3 Listening Test Procedure

According to the sample size calculation above, each of the two tests was conducted with eleven subjects, which were not the same across tests. One hundred real and simulated stimuli were presented to each subject in individually randomized order. The actual sequence of the presentation mode (real vs. simulated) was – again individually for each subject – drawn from a uniform dichotomous random distribution. As a result, the proportions of simulated and real stimuli varied among subjects between 0.42:0.58 and 0.56:0.44. Slightly unequal proportions were tolerated to minimize interdependence of succeeding answers and to prevent subjects from making assumptions about the absolute amount of correct “Yes” and “No” answers in the test. After each presentation subjects had to decide whether, in their opinion, the presentation was simulated or not. Before beginning the test, subjects were allowed to take headphones off once while playing either stimulus. This was necessary as it was observed in pre-tests that stimuli were so similar that people actually thought they never heard a simulation.

As subjects wore the STAX headphones throughout the test, the perception of absolute timbre was filtered by the headphones' transfer function for exterior sound fields. Since this coloration was a moderate and constant source of error under all tested conditions, and as subjects were not familiar with the presented stimuli, the authors have no reason to believe that the subjects' ability to distinguish between simulated and real loudspeakers was significantly disturbed.

Subjects were encouraged to make use of horizontal head rotations without exceeding an angular range of ±80° given by the BRIR measurement. Due to this instruction, which was controlled by monitoring the head tracking data during the test, not all loudspeaker locations (i.e. #4, #5; #3 at best only partially) were visible to the subjects during the test. In order to suppress memory effects of minor auditive differences, which could potentially bias individual results in either direction, stimuli were randomly varied in content (20) and source location (5), so that a particular combination of content and source position was presented only once in each test, either as real or simulated stimulus.

Contents varied from artificial signals such as steady state or burst noises, over male and female speech in foreign and native language, to recordings of single instruments and monophonic down mixes of pop songs. Loudness differences between stimuli were compensated beforehand. The playback level of the frontal loudspeaker (#1) was adjusted to 65 $\mathrm{dB_{SPL}}$ at the listener position for the pink noise stimulus. Since BRIR data sets had been measured with all loudspeakers set to identical gain sound pressure level differences between the loudspeakers were not equalized. A moderate playback level was chosen, because nonlinear distortions of the speakers would not have been reproduced by the auralization engine and could have influenced the detection task. Equal loudness between real and simulated presentation mode was established through pre-adjustment by two expert listeners. The duration of the stimuli was set to approximately six seconds (including reverberation), which sufficed to move one's head in a reasonable range. Each stimulus was presented only once. To help maintaining a constant level of concentration, subjects could decide when to proceed to the next stimulus. As stimuli were rather short and stimulus replay was not allowed, none of the subjects needed more than 15 minutes to complete the 100 trials.

12.4.4 Subjects

Across both listening tests subjects were of an average age of 29 years (86.5% male). Subjects had an average of six years of musical education and more than half of them had already taken part in listening tests with dynamic binaural technology. Hearing acuity was assessed based on self-reports. Subjects could thus be regarded as an experienced sample of a typical, untrained population.

12.5 Results

Estimates of the individual sensitivities d'_i were calculated from individual hit and false alarm rates according to (12-2). Their mean value, i.e. the average ty d'_{avg}, shows values above zero for both development stages of the simulator (see Table 12-3).

Table 12-3. Average sensitivity values, biases and respective standard deviations for the two groups of subjects assessing either simulator stage.

	basic simulation	**improved simulation**
d'_{avg}	0.2956	0.0512
$\hat{\sigma}_{d'}$	0.4504	0.1456
β_{avg}	1.0777	1.0186
$\hat{\sigma}_{\beta}$	0.0845	0.0540

Hence, the test points to a slight sensory difference between simulation and reality – in the direction that the simulation was identified as such more often than the real sound field. Average values for the response bias β, derived from estimates of the individual bias values β_i according to (12-3), show a moderate shift into the "No" direction (β_{avg} > 1), i.e. listeners tended to believe in the 'realness' of the stimuli, independent of the actual sensory difference.

T-tests, conducted to verify the statistical significance of the observed differences, showed that the d' values for the *improved simulation* (d'_{avg} = 0.0512) were significantly smaller than d'_{min} = 0.1777 considered to represent a meaningful effect (one-sided test, t = -2.882, p = 0.0082). Before, data were tested for normality using a Kolmogorov–Smirnov test (p = 0.964). For the basic simulation a t-test was unnecessary because the observed sensitivity (d'_{avg} = 0.2956) was already larger than the stated minimum effect. When comparing the results of both simulator stages, statistical significance was missed just barely (p = 0.056, one-sided t-test for independent samples, see also slightly overlapping 90%-CIs in Figure 12-3).

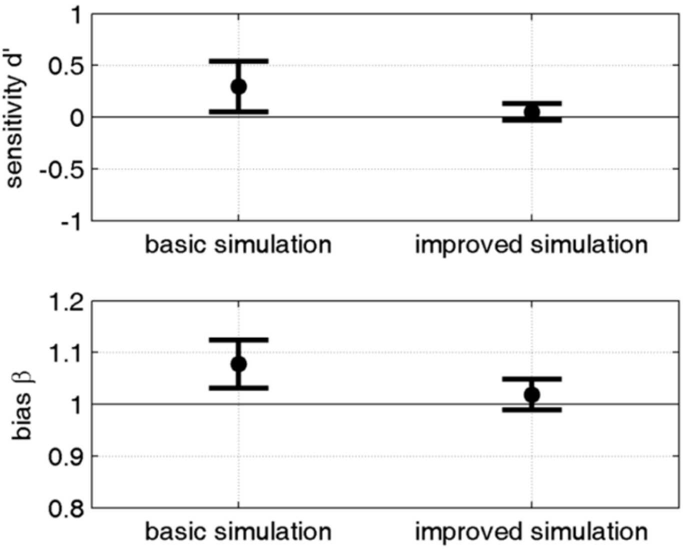


Figure 12-3. Average values and 90% confidence intervals of individual sensitivities d'_i (above) and biases β_i (below) as found for the two groups of subjects assessing either simulator stage.

Figure 12-4 illustrates the response behavior of the two observer groups assessing the two development stages of the simulator as modeled by the SDT equal variance observer model.

12.6 Discussion

We suggested a definition for the ‘plausibility’ of virtual realities as the agreement of a simulation with an inner reference of the corresponding real environment, as well as an operationalization that can be experimentally applied to evaluate simulations. It uses a Yes/No discrimination task, followed by an analysis based on signal detection theory, thus separating the sensory difference from the response bias of the test subjects.

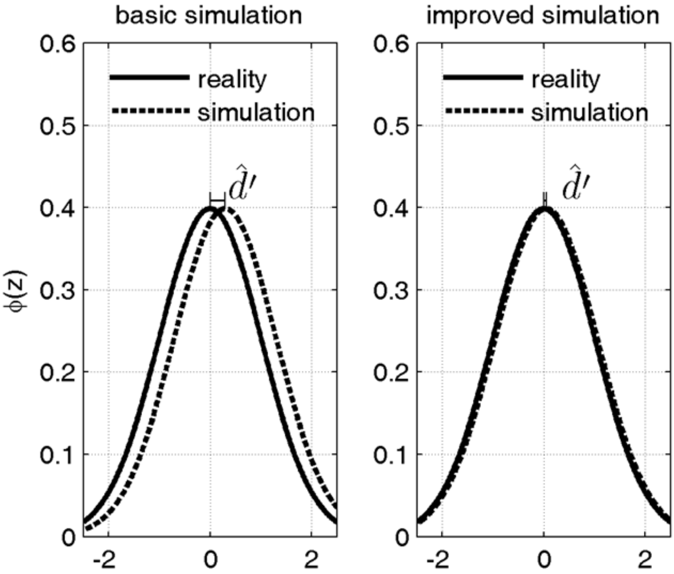


Figure 12-4. Probability density distributions modeling the groups' performances in detecting the 'simulation' condition in both simulator stages.

Although the test could in principle be used to evaluate any kind of virtual environment, including visual or audio-visual displays, the pre-condition, that reality and simulation can be presented in the same technical and spatial environment, will in practice preclude the application for most optical systems based on screen projections or head mounted displays as well as the evaluation of loudspeaker arrays for sound field synthesis, which can hardly be made transparent for an exterior environment. It works, however, well for binaural simulations, where listeners may stay 'wearing' the simulator (i.e. their headphones) no matter whether an audio signal is actually played back from the headphones or whether it is coming from outside. Defining a minimum effect hypothesis in terms of a detection rate considered to be relevant in the context of a certain application, plausibility can be tested with any desired strictness. Moreover, the bias of listeners to believe in the 'realness' or the 'simulatedness', which may depend on a variety of psychological, technical, cultural and content-related variables, can be analyzed independently of the sensory difference.

We applied the suggested measurement model to evaluate the overall performance of a system for data-based dynamic binaural synthesis and to demonstrate the efficiency of technical improvements of the system. The results show, that data based dynamic binaural simulations, particularly if certain improvements in signal processing are implemented, are able to provide a very high degree of plausibility. When asked to discriminate between simulation and reality, subjects were almost perfectly guessing. Whereas there is evidence [2] that spectral differences resulting from the non-individual morphology of the HATS and an imperfect spectral compensation of the signal chain interfere with the perceived ‘authenticity’, i.e. the perceptual identity of reality and simulation, our results imply that these deficiencies do not impair the perception of ‘plausibility’. Plausibility seems, instead, to be sensitive to excessive latency, cross fade artifacts and instable localization, as suggested by comparing the performance of the two implementations under test. The observed tolerance towards spectral coloration can probably be attributed to the weak memory of timbre ([27]–[29]), making spectral differences not immediately obvious to the listener without direct comparison with an external reference.

Results also imply that, with a state of the art implementation frequently claimed ‘systematic artifacts’ of binaural simulations such as a perception of elevation ([30], [31]) or a lacking externalization ([32], [32]) do not occur to a degree where the simulation could be recognized as such.

It is tempting to think about an application of the test procedure to the criterion-free measurement of other perceptual constructs which are defined by or shall be assessed in relation to some inner reference. The application of signal detection theory, however, requires a four-field-matrix, where Yes/No answers can be objectively classified as false or true. This excludes, according to our notion, attributes with an evaluative, personal component, such as ‘immersion’ or ‘sense of presence’. For the measurement of perceptual attributes which are related more closely to physical properties, such as ‘instability of localization’ or ‘perceived latency’, the test will possess only limited validity as the correctness of respective judgments can only be concluded indirectly from comparative assessments of real and simulated sound fields; the mentioned percepts might subjectively be truly perceivable under both conditions.

For the central question, in how far acoustical simulations as a whole are convincing substitutes for real acoustical environments, the suggested test might provide an attractive approach.

12.7 Acknowledgements

The authors would like to thank Dr. Hans-Joachim Maempel for inspiring discussions during the design of this study. The investigation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG WE 4057/3-1).

12.8 References

[1] Bronkhorst, A. W. (1995): "Localization of real and virtual sound sources", in: *J. Acoust. Soc. Am.*, **98**(5), pp. 2542-2553

[2] Møller, H. et al. (1996): "Binaural Technique: Do We Need Individual Recordings?", in: *J. Audio Eng. Soc.*, **44**(6), pp. 451-469

[3] Møller, H. et al. (1997): "Evaluation of Artificial Heads in Listening Tests", in: *Proc. of the 102nd AES Conv.*, Munich, preprint no. 4404

[4] Djelani, T. et al. (2000): "An Interactive Virtual-Environment Generator for Psychoacoustic Research II: Collection of Head-Related Impulse Responses and Evaluation of Auditory Localization", in: *Acta Acustica united with Acustica*, **86**, pp. 1046-1053

[5] Minnaar, P. et al. (2001): "Localization with Binaural Recordings from Artificial and Human Heads", in: *J. Audio Eng. Soc.*, **49**(5), pp. 323-336

[6] Liebetrau, J. et al. (2007): "Localization in Spatial Audio - from Wave Field Synthesis to 22.2", in: *Proc. of the 123rd AES Conv.*, New York, preprint no. 7164

[7] Pulkki, V.; Merimaa, J.: "Spatial Impulse Response Rendering: Listening tests and applications to continuous sound", in: *Proc. of the 118th AES Conv.*, Barcelona, peprint no. 6371

[8] Schubert, T.; Friedmann, F.; Regenbrecht, H. (2001): "The Experience of Presence: Factor Analytic Insights", in: *Presence: Teleoperators and Virtual Environments*, **10**(3), pp. 266-281

[9] Lessiter, J. et al. (2001): "A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory", in: *Presence: Teleoperators and Virtual Environments*, **10**(3), pp. 282-297

[10] Blauert, J. (1997): *Spatial Hearing. The Psychophysics of Human Sound Localization*, 2nd ed., Cambridge, MA.: MIT Press

[11] Kuhn-Rahloff, C. (2012): *Realitätstreue, Natürlichkeit, Plausibilität: Perzeptive Beurteilungen in der Elektroakustik*, Berlin: Springer Verlag

[12] Pellegrini, R. S. (2001): *A virtual reference listening room as an application of auditory virtual environments*, Doct. dissertation, Ruhr-Universität Bochum, Berlin: dissertation.de

[13] Reiter, U. (2011): "Perceived Quality in Game Audio", in: Grimshaw, M. (ed.): *Game Sound Technology and Player Interaction: Concepts and Developments,* Hershey, New York: IGI Global

[14] Green, D. M.; Swets, J. A. (1974): *Signal Detection Theory and Psychophysics*, Huntington: Krieger

[15] Ashby, Gregory F. (2000): "A Stochastic Version of General Recognition Theory", in: *J. Math. Psych.*, **44**, pp. 310-329

[16] Murphy, K. R.; Myors, B. (1999): "Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model", in: *J. Appl. Psychol.*, **84**(2), pp. 234-248

[17] Wickens, T. D. (2002): *Elementary Signal Detection Theory*, New York: Oxford University Press

[18] Kadlec, H. (1999): "Statistical Properties of d' and ß Estimates of Signal Detection Theory", in: *Psychological Methods*, **4**(1), pp. 22-43

[19] Lipshitz, S. P.; Vanderkooy, J. (1981): "The Great Debate: Subjective Evaluation", in: *J. Audio Eng. Soc.*, **29**(7/8), pp. 482-491

[20] Lindau, A.; Hohn, T., Weinzierl, S. (2007): "Binaural resynthesis for comparative studies of acoustical environments", in: *Proc. of the 122nd AES Conv.*, preprint no. 7032

[21] Lindau, A.; Brinkmann, F. (2010): "Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings", in: *Proc. of the 3rd Int. Workshop on Perceptual Quality of Systems*, Dresden, pp. 137-142

[22] Lindau, A. (2009): "The Perception of System Latency in Dynamic Binaural Synthesis", in: *Proc. of 35th DAGA*, Rotterdam, pp. 1063-1066

[23] Lindau, A.; Estrella, J.; Weinzierl, S. (2010): “Individualization of dynamic binaural synthesis by real time manipulation of the ITD”, in: *Proc. of the 128th AES Conv.*, London, preprint no. 8088

[24] Lindau, A.; Kosanke, L.; Weinzierl, S. (2010): “Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses”, in: *Proc. of the 128th AES Conv.*, London, preprint no. 8089

[25] Faul, Franz et al. (2007): “G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences”, in: *Behavior Research Methods*, **39**(2), pp. 175-191

[26] Cohen, Jacob (1988): *Statistical power analysis for the behavioral sciences,* 2nd. Ed., NJ et al.: Hillsdale

[27] Cowan, N. (1984): “On short and long auditory stores”, in: *Psychological Bulletin*, **96**(2), pp. 341-370

[28] Starr, G. A.; Pitt, M. A. (1997): “Interference effects in short-term memory for timbre”, in: *J. Acoust. Soc. Am.*, **102**(1), pp. 486-494

[29] Winkler, I.; Cowan, N. (2005): “From Sensory to Long-Term Memory. Evidence from Auditory Memory Reactivation Studies”, in: *Exp. Psych.*, **52**(1), pp. 3-20

[30] Begault, D. R. (1992): “Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems”, in: *J. Audio Eng. Soc.*, **40**(11), pp. 895-904

[31] Härmä, A. et al. (2004): “Augmented Reality Audio for Mobile and Wearable Appliances”, in: *J. Audio Eng. Soc.*, **52**(6), pp. 618-639

[32] Kim, S.-M.; Choi, W. (2005): “On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach”, in: *J. Acoust. Soc. Am.*, **117**(6), pp. 3657-3665

[33] Brookes, T.; Treble, C. (2005): “The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binaural recordings”, in: *Proc. of the 118th AES Convention*, Barcelona, preprint no. 6439

13 Assessing the Authenticity of Individual Dynamic Binaural Synthesis

The following chapter is an authorized reprint of the full-paper peer-reviewed article (reproduced from the author's post-print):

> Brinkmann, Fabian; Lindau, Alexander; Vrhovnik, Martina, Weinzierl, Stefan (2014): "Assessing the Authenticity of Individual Dynamic Binaural Synthesis", in: *Proc. of EAA Joint Auralization and Ambisonics Symposium*, Berlin, p. 62-68, http://dx.doi.org/10.14279/depositonce-11

The article has been faithfully reproduced from the author's post-print. However, in order to achieve a consistent typographic style throughout the whole dissertation minor modifications have been necessary, as, e.g., reworking the citation style and minimal typographic corrections.

Author's Note

In the case of the publication at hand the author of this dissertation was not the primary author. However, this publication marks an end point of a longer line of research initiated and supervised by the author. Hence, for the sake of completeness of the whole presentation it was decided to be included.

13.1 Abstract

Binaural technology allows capturing sound fields by recording the sound pressure arriving at the listener's ear canal entrances. If these signals are reconstructed for the same listener the simulation should be indistinguishable from the corresponding real sound field. A simulation fulfilling this premise could be termed as perceptually *authentic*.

Authenticity has been assessed previously for static binaural resynthesis of sound sources in anechoic environments, i.e. for HRTF-based simulations not accounting for head movements of the listeners. Results indicated that simulations were still discernable from real sound fields, at least, if critical audio material was used.

However, for *dynamic* binaural synthesis to our knowledge – and probably because this technology is even more demanding – no such study has been conducted so far. Thus, having developed a state-of-the-art system for individual dynamic auralization of anechoic and reverberant acoustical environments, we assessed its perceptual authenticity by letting subjects directly compare binaural simulations and real sound fields. To this end, individual binaural room impulses were acquired for two different source positions in a medium-sized recording studio, as well as individual headphone transfer functions. Listening tests were conducted for two different audio contents applying a most sensitive ABX test paradigm. Results showed that for speech signals many of the subjects failed to reliably detect the simulation. For pink noise pulses, however, all subjects could distinguish the simulation from reality. Results further provided evidence for future improvements.

13.2 Introduction

As overall criteria for the quality of virtual acoustic environments, the perceived *plausibility* and *authenticity* has been proposed [1], [2]. Whereas the plausibility of a simulation refers to the degree of agreement with the listener's expectation towards a corresponding real event (agreement with an inner reference), authenticity refers to the perceptual identity with an explicitly presented real event (agreement with an external reference). While a non-individual data-based dynamic binaural synthesis has already been shown to provide plausible simulations [3], a dynamic synthesis based on individual binaural recordings appears to be a particularly promising candidate for a perceptually authentic acoustical simulation. Further, a formal assessment of the authenticity of state-of-the-art binaural technology would be of great practical relevance: Since nearly all currently known approaches to sound field synthesis (such as wave field synthesis, or higher order ambisonics) can be transcoded into binaural signals, a perceptually authentic binaural reproduction would provide a convenient reference simulation required for the strict, reliable and comprehensive evaluation of a wide variety of simulation approaches and systems [4].

Three empirical studies were found to be concerned with the authenticity of binaural simulations. However, all three studies assessed static auralization, i.e., simulations not accounting for natural head movements of the listeners. In order to allow for a convenient comparability, statistical significance of the observed results was assessed based on exact Bernoulli test statistics, if not initially given.

Langendijk and Bronkhorst [5] assessed the authenticity of individual binaural reproduction for six sound sources distributed evenly around the listener. Binaural signals were reproduced utilizing small earphones placed 1 cm in front of the concha with only little influence on the sound field of external sources. Band limited white noise bursts (500 Hz–16 kHz) were presented in a four interval 2AFC (alternative forced choice) paradigm where each sequence of four noise bursts contained three identical and one 'oddball'-stimulus in either second or third position, that had to be detected by the subjects. Detection rates across subjects were slightly but significantly above chance ($p_{\text{correct}} = 0.53$, 6 subjects, $N_{\text{total}} = 1800$ trials).

Moore et al. [6] conducted a similar listening test. Subjects participated twice in the experiment, and were considered untrained in the first run and trained in the second. A frontal sound source was auralized using cross-talk canceled (transaural) reproduction of individual binaural recordings. When presenting click or noise stimuli to trained subjects detection rates were again slightly but significantly above chance ($p_{\text{corr. click}} = p_{\text{corr. noise}} = 0.594$, 8 subjects, $N_{\text{total}} = 192$). *Untrained subjects*, however, were not able to detect the binaural simulation reliably ($p_{\text{corr. click}} = 0.5$, $p_{\text{corr. noise}} = 0.54$, $p_{\text{corr.testable}} = 0.675$ @ $\alpha = 0.05$ with 95% power, Dunn-Sidák corrected for multiple testing). Moreover, when using a *synthetic vowel sound*, the simulation was indistinguishable for both trained and untrained subjects ($p_{\text{corr.observed}} = 0.48$, $p_{\text{corr.testable}}$ as mentioned above).

Masiero [7] tested authenticity in a 3AFC test paradigm utilizing 24 sound sources distributed evenly around the listeners. Individual binaural signals were presented to 40 subjects through circumaural open headphones using noise, speech and music stimuli. Average detection rates were $p_{\text{corr. noise}} = 0.87$, $p_{\text{corr. speech}} = 0.74$, and $p_{\text{corr.music}} = 0.71$ (transformed to 2AFC detection rates for better comparability). While not being given originally by the authors, a *post hoc* inferential statistics analysis of the raw data revealed that for all three stimulus conditions detections rates were significantly above chance. Further, an ANOVA conducted by Masiero showed the stimulus effect to be significant.

All three studies used some kind of head rest to control the subjects' head position. In addition, Moore et al. and Masiero monitored the subjects' head position with optic or magnetic tracking systems. Throughout his study, Masiero allowed for head movements between ±1°–2° rotation, and ±1–2 cm translation, respectively. Additionally, Masiero allowed his subjects to listen three times to the sequence of test stimuli whereas in the other two studies each condition was presented only once.

While – technically – being a far more demanding reproduction mode than static auralization, perceptual authenticity of *dynamic* binaural synthesis has not been assessed before. Moreover, a success of such an assessment has become more likely as number of technical improvements has been introduced recently: For example, new extraaural binaural headphones were presented (*BKsystem*, [8]) along with a perceptually optimized approach to the compensation of the headphone transfer function [9]. Further, an in-ear measurement systems for the reliable acquisition of individual binaural transfer functions (*PRECISE*, [9]) has been developed, and crossfade artifacts of dynamic binaural rendering have been minimized [10].

Further, as shown above, former studies achieved high statistical test power by cumulating test results over individuals and repeated trials while omitting a priori discussions of practical effect size and required test power. However, in order to limit the required methodological effort, and as individual performance was expected to be potentially quite different, we aimed at designing our test to produce practically meaningful results already on the level of individual subjects (cf. section 13.3.5).

13.3 Method

13.3.1 Setup

The listening tests were conducted in the recording studio of the *State Institute for Music Research*[8], Berlin ($V = 122\ \mathrm{m}^3$, $RT_{1\mathrm{kHz}} = 0.65\ \mathrm{s}$). Subjects were seated on a customized chair with an adjustable neck rest and a small table providing an armrest and space for placing the tactile interface used throughout the test (*Korg nanoKONTROL* Midi-Interface). An LCD screen was used as visual interface and placed 2 m in front of the subjects at eye level.

Two active near-field monitors (*Genelec 8030a*) were placed in front and to the right of the subjects at a distance of 3 m and a height of 1.56 m, corresponding to source positions of approximately 0° azimuth, 8° elevation (source 1) and 90° azimuth, 8° elevation (source 2). With a critical distance of 0.8 m and a loudspeaker directivity index of ca. 5 dB at 1 kHz, the source-receiver distance results in a slightly emphasized diffuse field component of the sound field. The height was adjusted so that the direct sound path from source 1 to the listening position was not blocked by the LCD screen. The source positions were chosen to represent conditions with minimal and maximal interaural time and level difference at a neutral head orientation (see test setup, Figure 13-1).

[8] Staatliches Institut für Musikforschung, http://www.sim.spk-berlin.de/

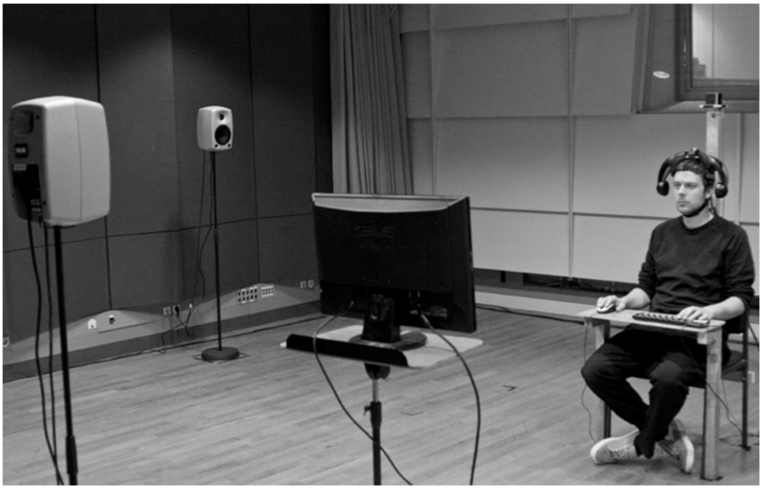


Figure 13-1. Listening test environment and used setup.

For binaural reproduction, low-noise DSP-driven amplifiers and extraaural headphones were used, which were designed to exhibit only minimal influence on sound fields arriving from external sources while providing full audio bandwidth (*BKsystem*, [8]). Headphones were worn during the entire listening test, i.e. also during the binaural measurements, this way allowing for instantaneous switching between binaural simulation and corresponding real sound field. The subjects' head position was controlled using head tracking with 6 degrees of freedom (x, y, z, azimuth [head-above-torso orientation], elevation, lateral flexion) with a precision of 0.001 cm and 0.003°, respectively (*Polhemus Patriot*). A long term test of eight hours showed no noticeable drift of the tracking system.

Individual binaural transfer functions were measured at the blocked ear canal using *Knowles FG-23329* miniature electret condenser microphones flush cast into conical silicone earmolds. The molds were available in three different sizes, providing a good fit and reliable positioning for a wide range of individuals [9]. Phase differences between left and right ear microphones did not exceed ±2° avoiding audible interaural phase distortion [11].

The experiment was monitored by the investigator from a separate room with talkback connection to the test environment.

13.3.2 Reproduction of Binaural Signals

The presence of headphones influences the sound field at the listeners' ears. Having considered an additional filter for compensating this effect [12], Moore et al. [6] concluded that head-phones should not be used for direct comparisons of simulation and reality and consequently used transaural sound reproduction for their listening tests on authenticity. In contrast, we argue that a test on authenticity is not compromised as long as wearing the headphones (a) would affect real sound field and simulation in an identical manner and (b) would not mask possible cues for dis-criminating between the two. Condition (a) will be fulfilled by wearing the headphones both during measurement and simulation. For assessing condition (b), binaural room impulse responses (BRIRs) were measured with and without three types of head-phones (*BKsystem, STAX SRS 2050 II, AKG K-601*) using a source at 2 m distance, 45° azimuth and 0° elevation for head-above-torso orientations in the range of ±80° azimuth. For this purpose, the head and torso simulator FABIAN equipped with a computer controlled neck joint for high precision and automated control of the head-above-torso orientation was used [13]. The headphone's influence was analyzed based on differences in the magnitude responses, and with respect to deviations of interaural time and level differences (ITD, ILD). For the *BKsystem*, magni-tude response differences (Figure 13-2, top left) show an irregular pattern with differences between approx. ±7.5 dB.

Whereas differences in magnitudes might influence localization in the median plane [14] the perceivable bandwidth of the signal remains largely unaffected making it unlikely that potential cues for a direct comparison would be eliminated. ITD and ILD differences are displayed in Figure 13-2 (middle and bottom) and are believed to be inaudible for most head orientations. Assuming just audible differences of approximately 10-20 µs and 1 dB, respectively [1], only at 45°, where the ipsilateral ear is fully shadowed by the headphone, ILD differences slightly exceed the assumed threshold of audibility.

The observed differences are comparable to those found by Langendijk and Bronkhorst [5] who used small earphones near to the concha. Additionally, it is worth noting that differences were more than twice as high if conventional headphones were used (see Figure 13-2).

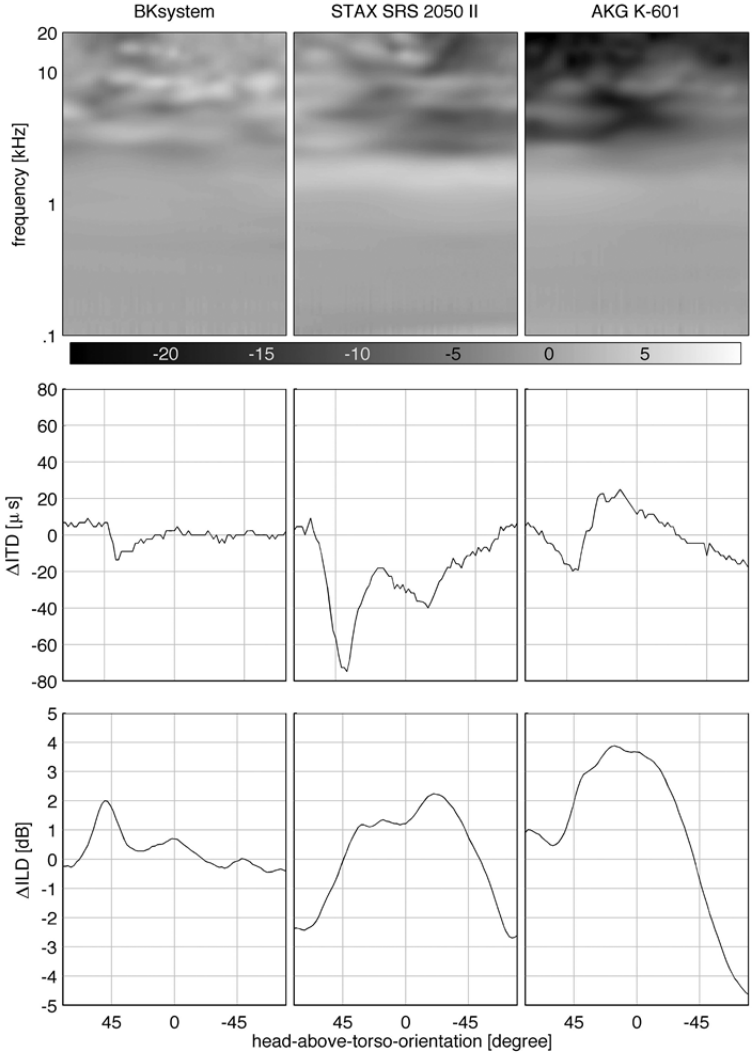


Figure 13-2. Differences observed in BRIRs when measured with and without headphones for head-above-torso-orientations of between ±80° and for a source at -45° azimuth and 0° elevation. Top: Magnitude spectra (3rd octave smoothed, right ear, gray scale indicates difference in dB); Middle: ITDs; Bottom: ILDs.

13.3.3 Measurement of Individual Binaural Transfer Functions

Binaural room impulse responses and headphone transfer func-tions (HpTFs) were measured and processed for every subject prior to the listening test. *Matlab®* was used for audio playback, recording and processing the input signals. The head position of the subject was monitored using *Pure Data*. Communication between the programs was done by UDP messages. All audio processing was conducted at a sampling rate of 44.1 kHz.

Before starting the measurements, subjects put on the head-phones and were familiarized with the procedure. Their current head position, given by azimuth and x/y/z coordinates was dis-played on the LCD screen along with the target position given only by azimuth. Additionally, an acoustic guidance signal was played back through the headphones helping subjects finding the target azimuth for the subsequent measurement. The head tracker was calibrated with the test subject looking at a frontal reference position marked on the LCD screen. Subjects were instructed to keep their eye level aligned to the reference position during measurement and listening test, this way establishing also indirect control over their head elevation and roll. For training proper head-positioning, subjects were instructed to move their head to a specific azimuth and hold the position for 10 seconds. All subjects were quickly able to maintain a position with a precision of ±0.2° azimuth.

Then, subjects inserted the measurement microphones into their ear canals until they were flush with the bottom of the concha. Correct fit was inspected by the investigator. The measurement level was adjusted to be comfortable for the subjects while also avoiding limiting of both the DSP-driven loudspeakers and headphones.

BRIRs were measured for head-above-torso orientations be-tween ±34° in azimuth and with a resolution of 2° providing smooth adaption to head movements [15]. The range was restricted to allow for a comfortable range of movements and convenient viewing of the LCD screen. Sine sweeps of an FFT order 18 were used for measuring transfer functions achieving a peak-to-tail signal-to-noise ratio (SNR) of approx. 80 dB for the BRIR at neutral head orientation without averaging [16].

The subjects started a measurement by pressing a button on the MIDI-interface after moving their head to the target position and reached it within ±0.1°. For the frontal head orientation, the target orientation had to be met also within 0.1 cm for the x/y/z-coordinates. For all other head orientations the translational positions naturally deviate from zero; in these cases subjects were instructed to meet the targeted azimuth only. During the measurement, head movements of more than 0.5° or 1 cm would have led to a repetition of the measurement, which rarely happened. These tolerance levels were set in order to avoid audible artifacts introduced by imperfect positioning [1] (p. 39), [17].

Thereafter, ten individual HpTFs were measured per subject. To *a priori* account for potential positional variance in the transfer functions, subjects were instructed to move their head to the left and right in between individual headphone measurements. After all measurements, which took about 30 minutes, the investigator removed the microphones without changing the position of the headphones.

13.3.4 Post-Processing

In a first step, times-of-flight were removed from the BRIRs by means of onset detection and ITDs were calculated and stored separately. ITDs were reinserted in real time during the listening test, avoiding comb-filter effects occurring in dynamic auralization with non-time-aligned BRIRs and reducing the overall system latency [10]. Secondly, BRIRs were normalized with respect to their mean magnitude response between 200 Hz and 400 Hz. Due to diffraction effects BRIRs exhibit an almost constant magnitude response in this frequency range making normalization especially robust against measurement errors and low-frequency noise. In a last step, BRIRs were truncated to 44100 samples with a squared sine fade out.

Individual HpTF compensation filters were designed using a weighted regularized least mean squares approach [18]. Filters of an FFT order 12 were calculated based on the average of ten HpTF per subject. Regularization was used to limit filter gains if perceptually required, the used approach is shortly explained here: HpTFs typically show distinct notches at high frequencies which are most likely caused by anti-resonances of the pinna cavities [19]. The exact frequency and depth of these notches strongly depends on the current fit of the headphones. Already a slight change in position might considerably detune a notch, potentially leading to ringing artifacts of the applied headphone filters [9]. Therefore, individual regularization functions were composed after manually fitting one or two parametric equalizers (PEQs) per ear to the most disturbing notches. The compensated headphones approached *a target band-pass* consisting of a 4th order Butterworth high-pass with a cut-off frequency of 59 Hz and a 2nd order Butterworth low-pass with a cut-off frequency of 16.4 kHz.

Finally, presentations of the real loudspeaker and the binaural simulation had to be matched to evoke equal loudness impressions. If assuming that signals obtained via individual binaural synthesis closely resemble those obtained from loudspeaker reproduction (cf. Figure 13-3), loudness matching can be achieved by simply matching the RMS-level of simulation and real sound field. Hence, matching was pursued by adjusting the RMS-level of five second pink noise samples recorded from loudspeakers and headphones while the subject's head was in the frontal reference position. To account for the actual acoustic reproduction paths in the listening test, prior to loudness-matching, the headphone recordings were convolved with the frontal incidence BRIRs and the headphone compensation filter whereas the loudspeaker recordings were convolved with the target band-pass.

13.3.5 Test Design

The ABX test paradigm as part of the N-AFC test family provides an objective, criterion-free and particularly sensitive test for the detection of small differences [20], and thus seems appropriate also for a test on the authenticity of virtual environments. ABX-testing involves presenting a test stimulus (A), a hidden reference stimulus (B) and an open reference stimulus (X). Subjects may either succeed (correct answer) or fail (incorrect answer) to identify the test stimulus. Being a Bernoulli experiment with a (2AFC) guessing rate of 50%, the binomial distribution allows the calculation of exact probabilities for observed detection rates enabling tests on statistical significance.

If ABX tests are used to prove the authenticity of simulations, one should be aware that this corresponds to proving the null hypothesis H_0 (i.e., proving equality of test conditions). Strictly speaking, this proof cannot be given by inferential statistics. Instead, the approach commonly pursued is to establish empirical evidence that *strongly supports* the H_0, e.g. by rejecting an alternative hypothesis H_1 stating an effect of irrelevant size, e.g. a minimal increase of the empirical detection rate above the guessing rate (i.e., negating a minimum-effect hypothesis [21]).

When testing a difference hypothesis H_1, two kinds of errors can be made in the final decision: The type 1 (alpha) error refers to the probability of wrongly concluding that there was an audible difference although there was none. The type 2 (beta) error is made, if wrongly concluding that there was no audible difference although indeed there was one. The test procedure (i.e. the number of AFC decisions requested) is usually designed to achieve small type 1 error levels (e.g. 0.05), making it difficult (especially for smaller differences) to produce significant test results. If we aim, however, at proving the H_0 such a design may unfairly favor our implicit interest ('progressive testing'). In order to design a fair test we first decided about a practically meaningful effect size to be rejected and then aimed at balancing both error levels in order to statistically substantiate both the rejection and the acceptance of the null hypothesis, i.e. the conclusion of authenticity.

For the current listening test, a number of 24 trials was chosen per subject and for each test condition (i.e., one combination of source direction and stimulus type), ensuring that for 18 or more correct answers, the H_0 ($p_{\mathrm{corr.}} = 0.5$) can be rejected, while for less than 18 correct answers, a specific H_1 of $p_{\mathrm{corr.}} = 0.9$ can be rejected for one test condition, both at equal (i.e., fair) type 1 and type 2 error levels. The chosen statistical design also accounted for the fact that each subject had to conduct 4 repeated tests (i.e. error levels of 5% for individual tests were established by suitable Bonferroni correction). The rather high detection rate of $p_{\mathrm{corr.}} = 0.9$ chosen to be rejected corresponds to our expectation that even small differences would lead to high detection rates, considering the very sensitive test design and the trained subjects available.

13.3.6 Test Procedure

Nine subjects with an average age of 30 years (6 male, 3 female) participated in the listening test, 3 of them were fairly and 6 of them highly experienced with dynamic binaural synthesis. No hearing anomalies were reported and all subjects had musical background (average 13 years of education). They could thus be regarded as expert listeners.

During the listening test three buttons (A/B/X) were displayed on the screen. Audio playback started, if the one of the buttons on the MIDI interface was pressed. To give the answer “A equals X”, the corresponding button had to be pressed and held for a short time. Subjects could take their time at will and repeatedly listen to A, B and X before answering, controlling all interaction with the tactile MIDI interface.

Two audio contents were used: a pulsed pink noise (0.75 s noise, 1 s silence, 20 ms ramps) and an anechoic male speech recording (5 s). The latter was chosen as a familiar ‘real-life’ stimulus, while noise pulses were believed to best reveal potential flaws in the simulation. Further, the bandwidth of the stimuli was restricted using a 100 Hz high-pass to eliminate the influence of low frequency background noise on the binaural transfer functions. As mentioned already, four ABX tests were conducted per subject (2 sources x 2 contents) each consisting of 24 trials. The presentation order of content and source was randomized and balanced across subjects. On average, the test took about 45 minutes. To avoid a drift in head position, subjects were instructed to move their head back to the reference position once between each trial and to keep the head’s orientation at approx. 0° elevation throughout the test.

Dynamic auralization was realized using the fast convolution engine *fWonder* [13] in conjunction with an algorithm for real-time reinsertion of the ITD [10]. *fWonder* was also used for applying (a) the HpTF compensation filter and (b) the loudspeaker target band-pass. The playback level for the listening test was set to 60 dB(A). BRIRs used in the convolution process were dynamically exchanged according to the subjects' current head-above-torso orientation, and playback was automatically muted if the subject's head orientation exceeded 35° azimuth.

13.3.7 Physical Verification

Prior to the listening test, acoustic differences between test con-ditions were estimated based on measurements with the FABIAN dummy head. Therefore, FABIAN was placed on the chair and BRIRs and HPTFs were measured and post-processed as de-scribed above. In a second step, BRIRs were measured as being reproduced by the headphones and the simulation engine described above. Differences between simulation and real sound field for the left ear and source 1 are depicted in Figure 13-3.

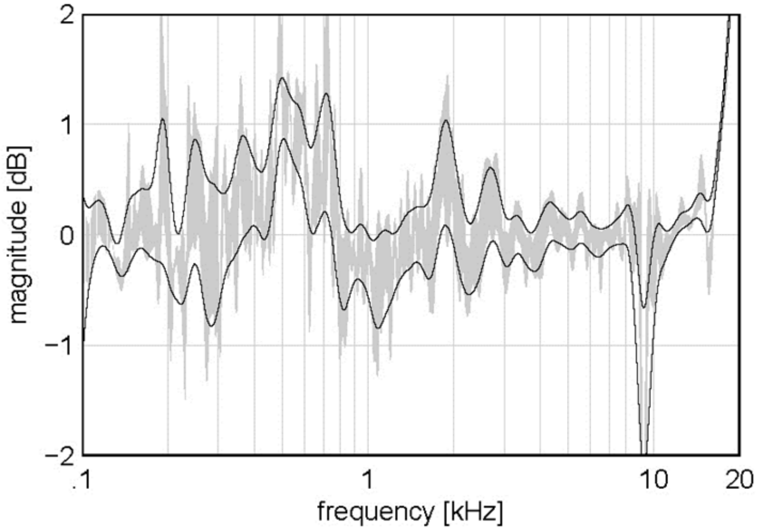


Figure 13-3. Differences between binaural simulation and real sound field for source 1 and left ear. The grey area encloses the range of differences observed for all head-above-torso orientations between ±34°. For ease of in-terpretation, the range of differences is shown again after applying 6th octave smoothing (black lines).

At a notch frequency in the HpTF at 10 kHz, differences reached up to 6 dB. However, this was assumed to be perceptually irrele-vant since the bandwidth of the notch was less than a 10th octave. Above 3 kHz differences were in a range of ±0.5 dB. Somewhat larger and presumably audible deviations of up to ±2 dB were observed between 100 Hz and 3 kHz which were potentially caused by time variance of electro-acoustic transducers. Alto-gether, Figure 13-3 shows comparable error patterns as Fig. 7b in Moore et al. [6].

13.4 Results

Results of the ABX listening test are summarized in Figure 13-4 for all subjects. A clear difference in detection performance was found between contents: While for the pulsed noise subjects were able to discriminate simulation and real sound field (all individual tests were statistically significant, see sect. 13.3.5 for the description of the statistical test), for the speech stimulus about half of them were not (55% significant tests). This increased uncertainty is also reflected in larger variance across subjects. Moreover, a tendency for higher detection rates ($p_{\text{corr.}}$) was found for source 2 ('s2') compared to source 1 ('s1'). Although statistical analysis of detectability was conducted on the level of individual subjects, observed average detection rates are given for better comparability to earlier studies: $p_{\text{corr. noise s1}} = 0.978$, $p_{\text{corr. noise s2}} = 0.991$, $p_{\text{corr. speech s1}} = 0.755$, and $p_{\text{corr. speech s2}} = 0.829$.

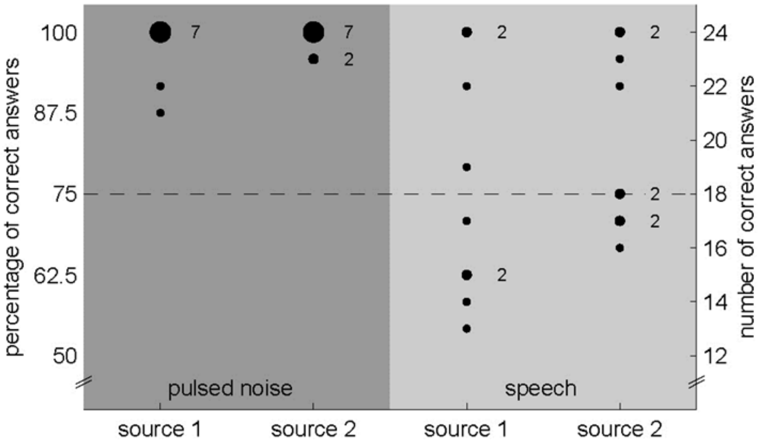


Figure 13-4. Listening test results of nine subjects and for each test condition. Dots indicate percentage/number of correct answers for each tested condition; singular num-bers indicate subjects with

identical detection results. Dots on or above the dashed line indicate statistically significant differences.

Differences between stimuli could also be found when com-paring the average duration needed for making decisions (significantly higher for speech: 38 s vs. 15 s, $p < 0.01$, Wilcoxon signed rank test for dependent samples). Furthermore, increased head movements were found for speech (interquartile range 20° vs. 8° azim., $p < 0.01$, Wilcoxon signed rank test for dependent samples), indicating an extended search behavior adopted by subjects.

During auralization, BRIRs were selected solely based on the subjects' head-above-torso orientation. Hence, unobserved dif-ferences in the remaining degrees of freedom (x, y, z, elevation, lateral flexion) might have caused audible artifacts. Therefore, head tracker data were recorded and used for a post hoc analysis of deviations between head position during binaural measurements and ABX tests: For x, y, z coordinates, deviations were found to have been smaller than 1 cm for 95% of the time and never exceed 2 cm which is well within limits given by Hiekkanen et al. [17]. Differences in head elevation (tilt) and in lateral flection (roll) rarely exceeded 10° and were below 5° for 90% of the time. This may have caused audible artifacts occasionally [1], (p. 44), but a systematic influence on the results is unlikely.

When asked for the qualities of perceived differences between simulation and reality after the listening test, subjects named coloration (7x), slight differences in loudness (2x), and spaciousness (1x). Furthermore, two subjects reported a hissing or resonating sound in the decay of the noise pulses.

13.5 Disscussion and Outlook

In the present study we assessed whether a state-of-the-art indi-vidual dynamic binaural simulation of an echoic environment can still be discriminated from the corresponding real sound field (test of 'perceptual authenticity'). To this end, measurement and post-processing of individual binaural transfer functions was demonstrated to be feasible within a reasonable amount of time, while obtaining a sufficient SNR and avoiding excessive test subject fatigue. Further, listening tests were conducted immediately after the measurements (i.e., – due to the minimization of deviations caused by time variability – resembling a best case scenario when aiming at proving authenticity) using a sensitive ABX test paradigm.

In accordance with earlier studies, we found that for a pulsed pink noise sample all subjects could reliably detect a difference between reality and simulation (individual detection rates between 87.5% and 100%). In case of the speech sample, however, only about half of the subjects still perceived a difference (individual detection rates between 54% and 100%). The higher detectability for the noise stimulus can be explained by its broad-band and steady nature, supporting the detection of coloration, which, according to the subjects, was perceived as the major difference. Further, in considering this, also the mentioned loudness differences might be related to remaining spectral deviations.

Furthermore, higher detection rates were observed for source 2 as compared to source 1. These could be explained by occasionally observed slight discontinuities in the extracted ITD, most probably due to lower SNR at the contralateral ear. Additionally, low SNR might have led to larger measurement errors potentially perceivable as coloration.

Further, a tendency for interaction between source and type of stimulus was observed, as across all subjects, detection rate was by far lowest for source 1 and the speech stimulus ($p_{corr.s1.noise}$ = 75.5%). The observed value indicates that for this condition the group's detection performance was at threshold level (discrimination between simulation and reality in 50% of the cases, equalling 75% in a 2AFC paradigm).

On overall, the observed detection rates were higher than those reported in previous studies, although the precision of the binaural reproduction was comparable ([6], there: Fig. 7b). Hereby, our test design allowing subjects to switch at will between stimuli before making final decisions, may be assumed to be much more sensitive to small flaws of the simulation than sequence-based presentations applied in previous studies. This is also indicated by the fact that six subjects reported to have felt to be merely guessing although four of them produced significant detection results for one source of the speech stimulus. In addition, results indicate that it is still more demanding to realize an authentic interactive real time simulation as compared to static auralization. This was somehow expectable as extended abilities of a simulation naturally go together with extended potential for perceptual issues (e.g., with respect to crossfading, latency, or spatial discretization).

Moreover, and in contrast to former studies, our test included simulating a reverberant environment. Future tests which are planned to be conducted in an anechoic chamber and a concert hall will reveal whether the simulation of reverberant environments resembles a specific challenge.

The 'hissing' sound perceived by two subjects might be an artefact related to slightly mistuned headphone filters, indicating the potential for future improvements of our simulation as e.g. with respect to perceptually more robust headphone filter design. Further, an optimization of individual ITD modelling appears advisable and will be pursued in the future.

13.6 Summary

A test of authenticity was conducted for the first time for a dynamic individual binaural simulation. Results showed that when by applying a sensitive test design the simulation was always clearly distinguishable from the real sound field, at least for critical sound source positions and if presenting noise bursts. However, for male speech, resembling a typical 'real-life' audio content and for a non-critical source position, half the subjects failed to reliably discriminate between simulation and reality, and averaged across subjects performed at threshold level.

13.7 Acknowledgments

This work was funded by the German Research Foundation (DFG WE 4057/3-1).

13.8 References

[1] Blauert, J. (1997): *Spatial Hearing. The Psychophysics of Human Sound Localization*, 2nd ed., Cambridge, MA: MIT Pres

[2] Pellegrini, R. S. (2001): *A virtual reference listening room as an application of auditory virtual environments*, doct. diss., University of Bochum

[3] Lindau, A.; Weinzierl, S. (2012): "Assessing the plausibility of virtual acoustic environments", in: *Acta Acustica united with Acustica*, **98**(5), pp. 804-810

[4] Wierstorf, H.; Raake, A.; Geier, M.; Spors, S. (2013): "Perception of focused sources in wave field synthesis", in: *J. Audio Eng. Soc.*, **61**(1/2), pp. 5-16

[5] Langendijk, E. H. A.; Bronkhorst, A. W. (2000): "Fidelity of three-dimensional-sound reproduction using a virtual auditory display", in: *J. Acoust. Soc. Am.*, **107**(1), pp. 528-537

[6] Moore, A. H.; Tew, A. I., Nicol, R. (2010): “An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments”, in: *J. Audio Eng. Soc.* (Engineering Report), **58**(1/2), pp. 36-45

[7] Masiero, B. (2012): *Individualized Binaural Technology. Measurement, Equalization and Perceptual Evaluation,* doct. dissertation, RWTH Aachen, 2012.

[8] Erbes, V.; Schulz, F.; Lindau, A.; Weinzierl, S. (2012): „An extraaural headphone system for optimized binaural repro-duction“, in: *Fortschritte d. Akustik: proc. of the 38th DAGA (German annual acoustic conference),* pp. 313-314, Darmstadt

[9] Lindau, A.; Brinkmann, F. (2012): “Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings”, in: *J. Audio Eng. Soc.*, **60**(1/2), pp. 54-62

[10] Lindau, A.; Estrella, J., Weinzierl, S. (2010): “Individualization of dynamic binaural synthesis by real time manipulation of the ITD”, in: *Proc. 128th of the AES Convention*, preprint no. 8088, London

[11] Mills, A. W. (1958): “On the minimum audible angle”, *J. Acoust. Soc. Am.*, **30**(4), pp. 237-246

[12] Moore, A. H.; Tew, A. I., Nicol, R. (2007): “Headphone transparification: A novel method for investigating the externalisation of binaural sounds”, in: *Proc. of the 123rd AES Convention*, preprint no. 7166, New York

[13] Lindau, A.; Hohn, T.; Weinzierl, S. (2007): “Binaural resynthesis for comparative studies of acoustical environments”, in: *Proc. of the 122th AES Convention,* preprint no. 7032, Vienna

[14] Hartmann, W. M.; Wittenberg, A. (1996): “On the externalization of sound images”, in: *J. Acoust. Soc. Am.*, **99**(6), pp. 3678-3688

[15] Lindau, A.; Weinzierl, S. (2009): “On the spatial resolution of virtual acoustic environments for head movements on horizontal, vertical and lateral direction”, in: *Proc. EAA Symposium on Auralization*, Espoo, Finland

[16] Müller, A.; Massarani, P. (2001): “Transfer function measurement with Sweeps. Directors's cut including previously unreleased material and some corrections”, originally published in: *J. Audio Eng. Soc.*, **49**(6), pp. 443-471

[17] Hiekkanen, T.; Mäkivirta, Ä.; Karjalainen, M. (2009): "Virtualized listening tests for loudspeakers", in: *J. Audio Eng. Soc.*, **57**(4), pp. 237-251

[18] Norcross, S. G.; Bouchard, M; Soulodre, G. A. (2006): "Inverse Filtering design using a minimal phase target function from regularization", in: *Proc. of the 121th AES Convention*, preprint no. 6929, San Francisco

[19] Takemoto, H.; Mokhtari, P.; Kato, H.; Nishimura, R., Iida, K. (2012): "Mechanism for generating peaks and notches of head-related transfer functions in the median plane", in: *J. Acoust. Soc. Am.*, **132**(6), pp. 3832-3841

[20] Leventhal, L. (1986): "Type I and type 2 errors in the statistical analysis of listening tests", in: *J. Audio Eng. Soc.*, **34**(6), pp. 437-453

[21] Murphy, K. R.; Myors, B. (1999): "Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model", in: *J. Appl. Psychol.,* **84**(2), pp. 234-248

14 A Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI)

The following chapter is an authorized reprint of the full-paper peer-reviewed article (reproduced from the author's post-print):

> Lindau, Alexander; Erbes, Vera, Lepa, Steffen; Maempel, Hans-Joachim; Brinkmann, Fabian; Weinzierl, Stefan (2014): "A Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI)", in: *Proc. of the EAA Joint Auralization and Ambisonics Symposium*, Berlin.

The article has been faithfully reproduced from the author's post-print.

Author's Note

Shortly before the submission of this thesis, the author was informed that this contribution was proposed to be included in a special issue of *Acta Acustica united with Acustica*. Hence, when submitting the thesis it was still unclear, whether the version of this article as published here would indeed be available from the the EAA symposium's proceedings or – somewhat later – from the *Acta Acustica*. Readers are kindly requested to consider this fact.

14.1 Abstract

The perceptual evaluation of virtual acoustic environments may be based on overall criteria such as *plausibility* and *authenticity* or by using catalogues of more detailed auditory qualities as, e.g., loudness, timbre, localization, etc. However, only the latter will be suitable to reveal specific shortcomings of a simulation under test and allow for a directed technical improvement. To this end a common vocabulary of relevant perceptual attributes appears desirable. Existing vocabularies for the evaluation of sound field synthesis, spatialization technologies and virtual environments were often generated *ad hoc* by the authors or have focused only on specific perceptual aspects. To overcome these limitations, we have developed a Spatial Audio Quality Inventory (SAQI) for the evaluation of virtual acoustic environments. It is a consensus vocabulary comprising 48 verbal descriptors of perceptual qualities assumed to be of practical relevance when comparing virtual environments to real or imagined references or amongst each other. The vocabulary was generated by a Focus Group of 20 German speaking experts for virtual acous-

tics. Five additional experts helped verifying the unambiguity of all descriptors and the related explanations. Moreover, an English translation was generated and verified by eight bilingual experts. The paper describes the methodology and the outcome, presenting the vocabulary in the English version.

14.2 Introduction

14.2.1 Objective

The perceptual assessment of virtual acoustic environments (VAEs, [1]) can be based on overall quality criteria such as plausibility [2] or authenticity ([3], p. 373), with operational definitions and experimental designs suggested in [4] and [5]. These measures differentiate between assessments with respect to an inner reference resulting from former experience (*plausibility*) or to an external, explicitly given reference (*authenticity*). They give, however, no insight into specific perceptual deviations and related technical shortcomings which would be required for a further improvement of the systems under test. Hence, we aimed at developing a descriptive vocabulary allowing for detailed perceptual assessments of VAEs. Hereby, we understand VAEs in a wide sense as all possible combinations of algorithms and instrumentation for the simulation, measurement, coding, processing and reproduction of spatial sound fields. Notions of Virtual Auditory Display (VAD, [4]), and Auditory Virtual Environments (AVE, [7]) we understand to be synonyms for VAE, whereas we would refrain from using a psychological notion ('auditory') for describing a technical apparatus.

14.2.2 State of the Art

Descriptive sensory vocabularies have been developed for various fields of interest in audio, as e.g. room acoustics, loudspeakers, multichannel recording and reproduction systems ([8], [9]), as well as for room correction systems, audio codecs, algorithms for headphone spatialization [10] and also for VAEs as a whole ([1], [11], [12]). Previous studies often applied initial vocabularies generated *ad hoc* from experience/knowledge of the authors and were later reduced by factor analysis of the listeners' ratings. Only recently, methodically advanced procedures for the development of consensual attributes have been applied, such as Quantitative Descriptive Analysis – QDA [9], Free Choice Profiling – FCP [10], or the Repertory Grid Technique – RGT [8].

Only a single study [12] was directly concerned with the development of a descriptive vocabulary ('quality features') for VAEs. It did not use a stimulus- but an expert-based approach (Delphi method, [13]). However, its scope was limited to the quality of dynamic auralizations from geometrical sound field simulations. Although the term ‘quality’ was understood quite broadly as the degree of agreement with a desired characteristic, it was narrowed by referring to three specific use cases (localization test, chat room, edutainment scenario).

The items collected in the five thematically most closely related studies ([1], [9]–[12]) include attributes for spectral coloration, spaciousness, localizability, steadiness of movements, source width, loudness, loudness balance, source distance, internalization vs. externalization, impulse-like artifacts, and dynamic responsiveness.

However, for a comprehensive perceptual evaluation of virtual environments, the above cited vocabularies do not seem sufficiently complete, nor do they cover all aspects of particular importance. E.g., while there is usually a descriptor for the perceived width of individual sources, there is none for the dimensions of complex ensembles of sources or of the spatial environment, such as the height or depth of rooms. Elementary problems of the spatial rendering are not covered, such as offsets in perceived location. A reason for these gaps might be that none of the authors explicitly targeted *comparative* assessments, either with regard to reality or between different VAEs.

For a differential diagnosis of technical shortcomings of virtual acoustic environments of any type, it therefore seemed mandatory to develop a consensus vocabulary (CV) utilizing an empirically substantiated approach for its generation, based on an agreement within a larger group of experts, and covering all aspects of particular relevance for the perceived quality of the different technologies involved.

14.3 Methods

14.3.1 General Considerations

Techniques for the spoken elicitation of descriptive consensus vocabularies as summarized in [14] (pp. 43), may be divided into individual (RGT, FCP, Flash Profiling) and group-based ap-proaches (QDA, Flavour Profile, Texture Profile, Sensory Spectrum). Whereas individual elicitation methods are always con-fronted with the problem of merging individual vocabularies into a valid group vocabulary, group methods directly aim at deriving such a consensual language. Mostly, panels of naïve subjects are instructed to develop a descriptive language by discussing sensory impressions of selected stimuli under guidance of a moderator. Such procedures are time-consuming [9] and the chosen sample of stimuli is critical with respect to the representativeness of results. Therefore, in [12] two established approaches for non-stimulus based CV generation were distinguished: the Delphi method [13] and the Focus Group method [15]. Both procedures rely on the assumption that the superior practical and theoretical experience of an expert compensates for the lack of adequate stimuli, of which a representative sample might not easily be obtainable. The Delphi method is a survey-based procedure for finding an agreement in a course of repeated interrogation sessions (interviews, questionnaires), i.e. without direct contact between experts. In contrast, the Focus Group may be applied, if experts can be accessed for face-to-face moderated roundtable discussions. While the first is deemed to reduce group biases, the latter is expected to lead to a more vivid and potentially more effective discussion.

14.3.2 The Focus Group Approach

As partners in a larger research consortium for virtual acoustics (SEACEN, Simulation and Evaluation of Acoustical Environments[9]) the authors had comparably easy access to experts in the field in order to arrange face-to-face meetings, making a Focus Group approach feasible. Methodologically, a Focus Group may be considered a combination of guided interviews and group discussions. This combination is particularly well-suited for the elicitation of expert-knowledge, as experts are routinely used to discourse-based revelation of consensual knowledge [16]. The incompleteness of results and irrelevancy during discussions can be reduced by using the so-called *dual-moderator* setting, where one moderator guides the discussion, while a co-moderator keeps track of the pre-defined agenda and discussion guidelines.

The moderator is supposed to control for unwanted group effects, e.g. by restraining ‘leading’ and motivating ‘hiding’ discussants, and being sensitive to non-verbal communication. The co-moderator is supposed to monitor the moderator’s behavior and the general compliance with the discussion guidelines.

Experimenter bias and group effects may be further addressed by extending the scheme to a so-called *two-way Focus Group*. There, during the first part of each discussion round the panel is split up in two groups, one group discussing (in dual-moderator scheme) and the other group observing the discussion from a remote room without interfering directly (for instance via one-way AV-monitoring). The

[9] http://www.seacen.tu-berlin.de

observer group acts as a further control mechanism: Less exposed to group effects, the observers are supposed to follow the discussion more objectively and rationally. In the second part, observers and discussants are brought together, discussing the observers' comments on the preceding discussion. If the targeted objective cannot be reached within a single discussion a *serial* Focus Group scheme allows for repeated discussion rounds.

14.3.3 Panel

As discussants we were able to invite several German speaking experts who were mostly members of the SEACEN consortium (PhD candidates, post-docs and professors) representing a wide professional experience regarding recording, simulation, reproduction and evaluation of spatial sound fields. While there were some changes over the different meetings regarding group size and composition, our panel size of 10–15 participants (20 experts in total, aged 25 to 67 yrs., 1–2 females per meeting) may be considered as optimal [17]. According to [17] the panel may further be regarded as a 'homogenous real group', i.e. with discussants coming from similar educational background and being known to each other before. In contrast to groups of differing backgrounds, homogenous groups are expected to lead more effective discussions. Moreover, real groups may (a) be more influenced by given hierarchies and role models, and (b) be more prone to 'private' conversations than randomly assigned groups, an effect that has to be controlled for by the moderator.

Discussions were held at four meetings in Berlin and Aachen over a period of six months. During those meetings repeated discussion rounds were scheduled for between one and four days. Discussions were conducted in the two-way dual-moderator scheme. After separating the experts into panel and observer group, the results collected so far were continuously updated on a projection screen, with the discussion audiovisually transmitted to the observer group along with the current state of the discussion. The AV-transmission was recorded for documentation purposes.

14.3.4 Main Discussion Objectives

As the main objective of the group, the vocabulary was defined as aiming at a consensual *psychological* measurement instrument, i.e. a questionnaire to be used for listening tests. As primary objects of assessment, VAEs of all kinds and in all stages of development were considered. The sensory scope was defined to comprise all perceivable differences of VAEs with respect to reality (be it imagined or explicitly given) or between different VAEs themselves. At last, some typical intended applications of the future vocabulary were given such as technical improvement, effort reduction, and benchmarking. These main objectives were aggregated into a mission statement formulated as the

creation of a consensus vocabulary for evaluating apparatus-related perceptual differences between technically generated acoustic environments or with respect to a presented or imagined acoustic reality.

The experts were instructed to aim at completeness of the overall vocabulary, while considering the assumed practical relevance at the same time. Descriptors should preferably be formulated semantically unidimensional and mutually exclusive. Since the terms should be self-explanatory to other experts in the field, the use of established names was to be preferred over inventing new ones. If the experts found it difficult to give a self- explanatory term, a short supplementary statement could be added to the descriptor. In particularly difficult cases it could also be agreed upon creating some illustrative audio examples later on. The group was instructed to propose scale label only if they were assumed to be necessary for clarity, leaving operationalization details mainly to the authors of the study.

14.3.5 Discussion Rules and Agenda

All major decisions were to be agreed upon by simple majority, with the moderator excluded from voting. However, after thorough and sometimes long discussions, a full consensus could be reached in most cases. After each discussion round, feedback rounds were conducted protocolling group comments on the moderator's performance, the progress of the discussion, setting and organization. Comments of the observer group were kept for the record until discussed to the satisfaction of the group.

After defining the discussion rules, an initial agenda was created by means of 20-minute-brainstorming sessions conducted at the beginning of the first two discussion rounds and resulting in a list of about 62 initial descriptors serving as a basis for the subsequent discussion rounds.

These lasted about 3.5 hours, with 2 rounds conducted per day. In total, 16 discussion rounds (56 hours) were completed. Each discussion round began with a random separation into discussants and observers. Then, the observed discussion began and lasted for about 90 minutes. Afterwards, a 20 minutes consolidation break was given to the observers. Finally, discussants and observers joined for a roundtable discussion of 90 minutes. If the discussion round was the last of a meeting, it was concluded by the feedback round.

Following recommendations in [16], the group was involved in the development of objectives, agenda and rules, thus both increasing involvement of participants and maintaining a thematically focused discussion. Thus, for instance, the group repeatedly specified the mission statement more precisely, added a fair rotation rule for panel and observer group, motivated the moderator to – in case of lively discussions – apply a speaker list or demanded from the observers to present their statements as an organized and consolidated list of pleas.

14.3.6 Finalization

Discussions resulted in a preliminary vocabulary which was finalized during the following four-stage procedure:

First, a post-check of semantic consensus was carried out. To this end, short written circumscriptions were created for all descriptors. Those experts that took part in the majority of discussion rounds (12 participants) were invited again to comment on the proposed circumscriptions in a written on-line discussion moderated by the author. In parallel, requested illustrative audio stimuli for three of the descriptors were created and discussed. Finally, consensual circumscriptions were agreed on for all descriptors.

Second, the vocabulary was subjected to an external evaluation of understandability. For this purpose, five additional VAE experts were asked to individually explain what descriptors meant to them while being given only the descriptor names, the (optional) short supplementary statements and audio examples, as well as objective and method of the vocabulary development. They were also asked to state, whether audio examples were adequate.

Third, after receiving all written statements (ca. 250) those were analyzed and checked for semantic equivalence with the group's circumscriptions. For about a third of the attributes minor semantic problems were identified.

In a fourth step, corrections derived from the analysis of additional expert's comments were agreed upon during final face-to-face discussion of a core group of five experts. Thereby, one attribute was confirmed to be obsolete. Moreover, label for rating scales for each item were agreed upon, considering, where available, earlier proposals of the group. Further, it was agreed upon including the final circumscriptions into the vocabulary as in most cases this was supposed to resolve remaining uncertainties identified by external experts. Additionally, a first informal English translation of the vocabulary was agreed upon. The German SAQI was published at the 40th Annual German Congress on Acoustics (DAGA 2014, [18]).

14.3.7 Translation to English

As descriptive vocabularies have been shown to be sensitive to language, care has to be taken in translation to conserve the original meaning [14] (p. 46). Translators should hence be sensitive to both the obviously meant (denotation), and the, potentially inter-culturally differing, *ascribed* meaning (connotation). According to guidelines related to test translation [19], it is recommended to invite as translators at least two experts in the field which are fluent in both languages. To ensure validity of the translation it is proposed to back-translate the questionnaire and to consider some more experts for a final review.

Regarding the target language, and as a result of the international publication process, we assumed a 'technical community language' to exist in the field of acoustics, which is neither a real US, UK nor any other native English. Thus, we would consider any scientist in the field a 'native' speaker of this 'community language'. Accordingly, as translators, we invited one native US, one native UK, one Greek and two Dutch acousticians. Half of them were researchers in virtual acoustics and all had good knowledge of German. They were provided with the German descriptors, the pre-translated circumscriptions, and the audio examples and were asked to produce adequate English terms. Translations were finally discussed in an audiovisual teleconfer-ence attended by translators and the authors. Additionally, three more German experts for virtual acoustics living for a longer time in English-talking countries produced back-translations which were in turn finally semantically analyzed by the authors. Besides being a test of the semantic compatibility of English and German version, the back-translating also resembled a test of the assumed 'bilingualism' of experts in the 'community English'. Thus, finding the back-translated versions to only minimally differ in meaning from the original German SAQI was considered as empirical evidence in support of our above hypothesis.

14.4 Results

The final vocabulary is termed Spatial Audio Quality Inventory for Virtual Acoustic Environments (SAQI, cf. Table 14-1). It consists of 48 qualitative descriptors which are sorted into 8 categories (timbre, tonalness, geometry, room, time behavior, dynamics, artifacts, and general impressions) and are to be considered as describing 'perceived differences with respect to [descriptor name]'.

Some attributes reflect a 'bottom-up' perspective of perception, being closely related to temporal or spectral properties of the audio signal. Some attributes are related to the specific VAE technology, appearing either as percepts of their own or resembling modifications of already existing perceptual qualities. Finally, some descriptors reflect a 'top-down' perspective of the comparative assessment task, representing supra-modal, affective, aesthetic or attitudinal aspects.

Each descriptor is complemented by a short written clarifying circumscription and suitable dichotomous, uni- or bipolar scale label, respectively. For three of the descriptors, illustrative audio examples were created ('roughness', 'comb-filter-likeness', 'compressor effects'). For handling possibly overlooked or newly emerging aspects of VAEs, an open category ('Other') is included in the vocabulary, to be named by subjects of the listening tests. Readers are cordially invited to share their experience with this category with the corresponding author.

It seemed further reasonable to make a perceptual assessment addressable to specific reference objects of a VAE: Five basic assessment entities were defined, providing an ideal-type ontology of the presented scene such as: *foreground sources, background sources,* the simulated *room acoustical environment,* the *reproduction system* (e.g. loudspeaker artifacts, amplifier noise) and the *laboratory environment* (HVAC noise, environmental sounds). In combination, these five entities are thought to incorporate all possible objects of interest (Table 14-2).

Finally, perceptual qualities of the observed VAEs may be further differentiated with respect to their time-variance. Thus, perceived differences might be either *constant* or *time-varying*. The time-variance might be *periodically or otherwise rule-based* or *non-regular* and it might be *continuous or discontinuous*. Moreover, perceived differences may either *depend on user interaction, depend on scene events* or on none of them (*independent*, Table 14-3).

Both, assessment objects and temporal modifications of qualities might be additionally queried in listening tests. Practical guidelines are referred to in the outlook section.

14.5 Discussion

Throughout the current investigation, the expert Focus Group approach proved to be an effective method for deriving a consensual vocabulary for the assessment of virtual acoustic environments. It yielded not only a comprehensive semantic differential but also some valuable extensions such as a systematics for reference entities, for the temporal behavior of auditory attributes, audio examples, and a glossary of terms.

Objectivity of the approach was addressed by different mechanisms for self-control. These turned out to be important, e.g. in cases, where the moderator tended to influence the discussion or where the ‘power of persuasion’ was not fully balanced within the group. In these cases, both the comments of the observer group and the external evaluations were perceived as valuable.

Construct validity of the discussions benefited from recent debates on quality measures for VAEs ([4], [5], [20]).Thus, – e.g. from discussing the mission statement – disputants were sensitized regarding a clear separation between perceptual impressions and physical measures and were aware of the relevance of assessments regarding both inner and external references (i.e. issues of plausibility and authenticity). Further, content validity was believed high as our panel of discussants covered a substantial and diverse range of expertise regarding the topic under discussion.

To obtain an impression of the completeness of the derived vocabulary, it can be confronted with results from previous studies as summed up in the first paragraph of section 14.2.2. From there, it can be verified that all previously identified aspects are also covered by the SAQI.

Although the Focus Group discussions took more time than expected, the overall duration of the vocabulary development was comparable to those reported for other approaches (i.e. for QDA [9], RGT [14], or the Delphi method [12]).

The notion of a ‘community language’ for the presented English translation might reduce the effort also for further translations: As most experts in the field can be considered ‘bilingual’ in their native and the community language, they should be able to produce valid translations by themselves. Hence the initial English translation might serve as a ‘bridge’ to the complete community.

14.6 Outlook

The SAQI questionnaire is intended to be helpful, e.g., when aiming at a directed technical improvement of VAE systems, when seeking opportunities for a perceptually motivated reduction of simulation effort, or when benchmarking competing VAE systems. It is currently applied in several experiments of the SEACEN consortium providing first impressions considering its applicability and usability. By providing a sample of identical stimuli to be included in different listening tests, the retest reliability of SAQI items shall be assessed, also across both languages. Based on these results, we will also be able to analyze the interdependency of items. For a future increase of overall test quality, a database of instructive training stimuli could be developed.

The German and English versions of the SAQI have been incorporated in the Matlab® listening test environment whisPER v1.8.0 [21] which is freely available[10]. It allows for an easy modification of the questionnaire, e.g., by selecting language, test paradigm (paired comparison or direct assessment), descriptors, and assessment entities or aspects of time variance an interactivity to be asked for (cf. whisPER User Manual). Detailed practical guidelines (e.g., for test subject training, test customization) are given in the extensive *SAQI Test Manual* [21]. Using both standardized software package, and the Test Manual for administration, training and instruction of subjects is supposed to increase objectivity and reliability of tests and is thus strongly recommended.

14.7 Acknowledgements

This investigation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG WE 4057/3-1). The authors – all of them members of the Focus Group themselves – would like to cordially thank the following individuals for participation (in alphabetical order): Benjamin Bernschütz, Clemens Büttner, Diemer de Vries, Alexander Fuß, Matthias Geier, Martin Guski, Michael Horn, Stefan Klockgether, Johannes Nowak, Rob Opdam, Zora Schärer, Frank Schultz, Michael Vorländer, and Hagen Wierstorf. We also would like to thank the five external VAE experts for providing their semantic circumscriptions of the pre-final qualitative descriptors: Matthias Frank, Markus Noisternig, Sönke Pelzer, Andreas Silze, and Franz Zotter. Moreover we would like to thank the five translators: Jude Brereton, Kees de Visser, Brian Gygi, Charalampos Saitis, and Steven van de Par while

[10] http://www.ak.tu-berlin.de/whisper, http://dx.doi.org/10.14279/depositonce-31

our final regards go to the three back-translators Frank Melchior, Nils Peters and Ulrich Reiter.

14.8 References

[1] Wenzel, E. M.; Foster, S. H. (1990): "Real-time Digital Synthesis of Virtual Acoustic Environments", in: *Proc. of the ACM Symposium on Interactive 3D Computer Graphics*, pp. 139-140

[2] Pellegrini, R. S. (2001): "Quality Assessment of Auditory Virtual Environments", in: *Proc. of ICAD 2001 - Seventh Meeting of the International Conference on Auditory Display*, Espoo

[3] Blauert J. (1997): *Spatial Hearing. The psychophysics of human sound localization*, MIT Press, 2nd Edition, Massachusetts, USA, 1997

[4] Lindau, A; Weinzierl, S (2012): "Assessing the Plausibility of Virtual Acoustic Environments", in: *Acta Acustica united with Acustica*, **98**(55), 804-810

[5] Brinkmann, F.; Lindau, A.; Vrhovnik, M., Weinzierl, S. (2014): "Assessing the Authenticity of Individual Dynamic Binaural Synthesis", in: *Proc. of EAA Joint Auralization and Ambisonics Symposium*, Berlin, p. 62-68, http://dx.doi.org/10.14279/depositonce-11

[6] Wenzel, E. M.; Wightman, F. L.; Foster, S. H. (1988): "A Virtual Display System for Conveying Three-Dimensional Acoustic Information", in: *Proc. of the Human Factors and Ergonomics Society Annual Meeting,* pp. 86-90

[7] Cohen, E. A.: "Technologies for three dimensional sound presentation and issues in subjective evaluation of the spatial image", in: *Proc. of the 89th AES Convention.* Los Angeles, preprint no. 2943

[8] Berg, J.; Rumsey, F. (1999): "Spatial Attribute Identification and Scaling by Repertory Grid Technique and other methods", in: *Proc. of the 16th International AES Conference: On Spatial Sound Reproduction*. Rovaniemi

[9] Koivuniemi, K; Zacharov, N (2001): "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training", in: *Proc. of the 111th AES Convention*. New York, preprint no. 5424

[10] Lorho, G (2005): “Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction”, in: *Proc. of the 119th AES Convention*. New York, preprint no. 6629

[11] Lokki, T; Järveläinen, H (2001): “Subjective evaluation of auralization of physics-based room acoustics modeling”, in: *Proc. of ICAD*. Espoo

[12] Silzle, Andreas (2007): “Quality Taxonomies for Auditory Virtual Environments”, in: *Proc. of the 122nd AES Convention*. Vienna, preprint no. 6993

[13] Dichanz, H. (2005): “Delphi-Befragung“, in: Mikos, L.; Wegener, C. (eds.): *Qualitative Medienforschung*. Konstanz: UVK, p. 297-303

[14] Bech, S.; Zacharov, N. (2006): *Perceptual Audio Evaluation: Theory, Method and Application*, Chichester: Wiley

[15] Stewart, D. W.; Shamdasani, P. N.; Rook, D. W. (1990): *Focus groups: Theory and practice*. Newbury Park, CA: Sage

[16] Bogner, A.; Leuthold M. (2005): “ ‘Was ich dazu noch sagen wollte...‘. Die Moderation von Experten-Fokusgruppen“, in: Bogner, A.; Littig, B; Menz, W. (eds.): *Das Experteninterview*. Wiesbaden: VS Verlag für Sozalwissenschaften

[17] Dürrenberger, G.; Behringer, J. (1999): *Die Fokusgruppe in Theorie und Anwendung*, Stuttgart: Rudolph-Sophien-Stift gGmbH

[18] Lindau, A. et al. (2014): „Ein Fokusgruppenverfahren für die Entwicklung eines Vokabulars zur sensorischen Beurteilung virtueller akustischer Umgebungen“, in: *Fortschritte der Akustik: Proc. of the 40th DAGA*. Oldenburg

[19] Hambleton, R. K. (2001): “The Next Generation of the ITC Test Translation and Adaptation Guidelines”, in: *Europ. J. Psych. Ass.*, **17**(3), pp. 164-172

[20] Maempel, H.-J.; Weinzierl, S. (2012): “Demands on measurement models for the perceptual qualities of virtual acoustic environments”, in: *Proc. of the 59th Open Seminar on Acoustics (OSA)*. Boszkowo (PL)

[21] Ciba, S.; Wlodarski, A.; Maempel, H.-J. (2009): “WhisPER – A new tool for performing listening tests”, in: *Proc. of the 126th AES Convention*. Munich, preprint 7749, http://dx.doi.org/10.14279/depositonce-31

[22] Lindau, A. (2014): *SAQI. Test Manual*, http://dx.doi.org/10.14279/depositonce-1

14.9 Appendix – SAQI-EN

Table 14-1. Spatial Audio Quality Inventory (SAQI) - English version

	perceptual quality	**circumscription**	**scale end label**
	Difference	Existence of a noticeable difference.	none – very large
Timbre	Tone color bright-dark	Timbral impression which is determined by the ratio of high to low frequency components.	darker – brighter
	High-frequency tone color	Timbral change in a limited frequency range.	attenuated – emphasized
	Mid-frequency tone color	Timbral change in a limited frequency range.	attenuated – emphasized
	Low-frequency tone color	Timbral change in a limited frequency range.	attenuated – emphasized
	Sharpness	Timbral impression which e.g., is indicative for the force with which a sound source is excited. Example: Hard/soft beating of percussion instruments, hard/soft plucking of string instruments (class. guitar, harp). Emphasized high frequencies may promote a 'sharp' sound impression.	less sharp – sharper
	Roughness*	Timbral impression of fierce or aggressive modulation/vibration, whereas individual oscillations are hardly distinguishable. Often rated as unpleasant.	less rough – more rough
	Comb filter coloration*	Often perceived as tonal coloration. 'Hollow' sound. Example: speaking through a tube.	less pronounced – more pronounced
	Metallic tone color	Coloration with pronounced narrow-band resonances, often as a result of low density of natural frequencies. Often when exciting metallic objects such as Gongs, bells, rattling tin cans audible. Applicable to room simulations, plate reverb, spring reverb, too.	less pronounced – more pronounced
Tonalness	Tonalness	Perceptibility of a pitch in a sound. Example for tonal sounds: voiced speech, beeps.	more unpitched – more pitched
	Pitch	The perception of pitch allows arranging tonal signals along a scale "higher - lower".	lower – higher
	Doppler effect	Continuous change of pitch (see above). Often perceived as a 'continuous detuning'. Example: 'Detuned' sound of the siren of a fast-moving ambulance.	less pronounced – more pronounced
Geometry	Horizontal direction	Direction of a sound source in the horizontal plane.	shifted anticlockwise – shifted clockwise (up to 180°)
	Vertical direction	Direction of a sound source in the vertical plane.	shifted up – shifted down (up to 180°)
	Front-back position	Refers to the position of a sound source before or behind the listener only. Impression of a position difference of a sound source caused by 'reflecting' its position on the frontal plane going through the listener.	dichotomous scale: not confused / confused
	Distance	Perceived distance of a sound source.	closer – more distant
	Depth	Perceived extent of a sound source in radial direction.	less deep – deeper
	Width	Perceived extent of a sound source in horizontal direction.	less wide – wider

	perceptual quality	**circumscription**	**scale end label**
	Height	Perceived extent of a sound source in vertical direction.	less high – higher
	Externalization	Describes the distinctness with which a sound source is perceived within or outside the head regardless of their distance. Terminologically often enclosed between the phenomena of in-head localization and out-of-head localization. Examples: Poorly/not externalized = perceived position of sound sources at diotic sound presentation via headphones, good/strongly externalized = perceived position of a natural source in reverberant environment and when allowing for movements of the listener.	more internalized – more externalized
	Localizability	If localizability is low, spatial extent and location of a sound source are difficult to estimate, or appear diffuse, resp. If localizability is high, a sound source is clearly delimited. Low/high localizability is often associated with high/low perceived extent of a sound source. Examples: sound sources in highly diffuse sound field are poorly localizable.	more difficult – easier
	Spatial disintegration	Sound sources, which - by experience - should have a united spatial shape, appear spatially separated. Possible cause: Parts of the sound source have been synthesized/simulated using separated algorithms/simulation methods and between those exists an unwanted offset in spatial parameters. Examples: fingering noise and playing tones of an instrument appear at different positions; spirant and voiced phonemes of speech are synthesized separately and then reproduced with an unwanted spatial separation.	more coherent – more disjointed
Room	Level of Reverberation	Perception of a strong reverberant sound field, caused by a high ratio of reflected to direct sound energy. Leads to the impression of high diffusivity in case of stationary excitation (in the sense of a low D/R-ratio). Example: The perceived intensity of reverberation differs significantly between rather small and very large spaces, such as living rooms and churches.	less – more
	Duration of Reverberation	Duration of the reverberant decay. Well audible at the end of signals.	shorter – longer
	Envelopment (by reverberation)	Sensation of being spatially surrounded by the reverberation. With more pronounced envelopment of reverberation, it is increasingly difficult to assign a specific position, a limited extension or a preferred direction to the reverberation. Impressions of either low or high reverberation envelopment arise with either diotic or dichotic (i.e., uncorrelated) presentation of reverberant audio material.	less pronounced – more pronounced

	perceptual quality	circumscription	scale end label
Time behavior	Pre-echoes	Copies of a sound with mostly lower loudness prior to the actually intended the starting point of a sound.	less intense – more intense
	Post-echoes	Copies of a sound with mostly decreasing loudness after the actually intended the starting point of a sound. Example: repetition of one's own voice through reflection on mountain walls.	less intense – more intense
	Temporal disintegration	Sound sources, which - by experience - should have a united temporal shape, appear temporally separated. Causes similar to "Spatial disintegration", however, here: due to timing-offsets in synthesis. Example: fingering noise and playing tones of an instrument appear at different points in time.	more coherent – more disjointed
	Crispness	Characteristic which is affected by the impulse fidelity of systems. Perception of the reproduction of transients. Transients can either be more soft/more smoothed/less precise, or - as opposed - be quicker/more precise/ more exact. Example for ‘smoothed’ transients: A transmission system that exhibits strong group delay distortions. Counter-example: Result of an equalization aiming at phase linearization.	less pronounced – more pronounced
	Speed	A scene is identical in content and sound, but evolves faster or slower. Does not have to be accompanied by a change in pitch. Examples of technical reasons: rotation speed, sample rate conversion, time stretching, changed duration of pauses between signal starting points; movements proceed at a different speed.	reduced – increased
	Sequence of events	Order or occurrence of scene components. Example: A dog suddenly barks at the end, instead - and as opposed to the reference - at the beginning.	unchanged – changed
	Responsiveness	Characteristic that is affected by latencies in the reproduction system. Distinguishes between more or less delayed reactions of a reproduction system with respect to user interactions.	lower – higher
Dynamics	Loudness	Perceived loudness of a sound source. Disappearance of a sound source can be stated by a loudness equaling zero. Example of a loudness contrast: Whispering vs. Screaming.	quieter – louder
	Dynamic range	Amount of loudness differences between loud and soft passages. In signals with a smaller dynamic range loud and soft passages differ less from the average loudness. Signals with a larger dynamic range contain both very loud and very soft passages.	smaller – larger
	Dynamic compression effects*	Sound changes beyond the long-term loudness. Collective category for a variety of percepts caused by dynamic compression. Examples: More compact sound of sum-compressed music tracks in comparison to the unedited original. ‘Compressor pumping’: Energy peaks in audio signals (bass drums, speech plosives) lead to a sudden drop in signal loudness which needs a susceptible period of time to recover.	less pronounced – more pronounced

	perceptual quality	circumscription	scale end label
Artifacts	Pitched artifact	Perception of a clearly unintended sound event. For example, a disturbing tone which is clearly not associated with the presented scene, such as an unexpected beep.	less intense – more intense
	Impulsive artifact	Perception of a clearly unintended sound event. For example, a short disturbing sound which is clearly not associated with the presented scene, such as an unexpected click.	less intense – more intense
	Noise-like artifact	Perception of a clearly unintended sound event. For example, a noise which is clearly not associated with the presented scene, such as a background noise from of a fan.	less intense – more intense
	Alien source	Perception of a clearly unintended sound event. Examples: an interfering radio signal, a wrongly unmuted mixing desk channel.	less intense – more intense
	Ghost source	Spatially separated, nearly simultaneous and not necessarily identical image of a sound source. A kind of a spatial copy of a signal: a sound source appears at one or more additional positions in the scene. Examples: two sound sources which are erroneously playing back the same audio content; double images when down-mixing main and spot microphone recordings; spatial aliasing in wave field synthesis (WFS): sound sources are perceived as ambivalent in direction.	less intense – more intense
	Distortion	Percept as a result of non-linear distortions as caused e.g. by clipping. Scratchy or ‘broken’ sound. Often dependent on signal amplitude. Perceptual quality can vary widely depending on the type of distortion. Example: clipping of digital input stages.	less intense – more intense
	Tactile vibration	Perception at the border between auditory and tactile modality. Vibration caused by a sound source can be felt through mechanical coupling to supporting surfaces. Examples: Live Concert: bass can be ‘felt in the stomach’; headphone cushions vibrate noticeably on the ear/head.	less intense – more intense
General	Clarity	Clarity/clearness with respect to any characteristic of elements of a sound scene. Impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected. The term is thus to be understood much broader than the in realm of room acoustics, where Clarity is used to predict the impression of declining transparency with increasing reverberation.	less pronounced – more pronounced
	Speech intelligibility	Impression of how well the words of a speaker can be understood. Typical of low speech intelligibility: station announcements. Typical for high speech intelligibility: Newscaster.	lower – higher
	Naturalness	Impression that a signal is in accordance with the expectation/former experience of an equivalent signal.	lower – higher
	Presence	Perception of ‘being-in-the-scene’, or 'spatial presence'. Impression of being inside a presented scene or to be spatially integrated into the scene.	lower – higher

	perceptual quality	circumscription	scale end label
	Degree-of-Liking	Difference with respect to pleasantness/unpleasantness. Evaluation of the perceived overall difference with respect to the degree of enjoyment or displeasure. Note that 'preference' might not be used synonymously, as, e.g., there may be situations where something is preferred that is - at the same time - not liked most.	lower – higher
	Other	Another, previously unrecognized difference.	less pronounced – more pronounced

*sound examples may be downloaded from http://dx.doi.org/10.14279/depositonce-1

Table 14-2. Hierarchical description system for assessments entities

All audible events				
Intended audible events (elements of the presented virtual scene)			Unintended audible events	
Foreground sources	Background sources	Room acoustic environment	Reproduction system	Laboratory environment

Table 14-3. Hierarchical description system for modifications of perceptual qualities

The perceived difference is …		
… constant	… varying periodically or otherwise rule-based with time	… varying non-regularly with time
	… in a continuous / discontinuous **manner**	
… **and** depending on scene events / user interaction / independent.		