

Imperial College London
Dyson School of Design Engineering

Improving Binaural Audio Techniques for Augmented Reality

Juan Isaac Engel Alonso-Martínez

A thesis submitted for the degree of Doctor of Philosophy

May 2021

Statement of Originality

The content of this Thesis is the result of the author's own work unless otherwise specified. Parts of this Thesis have been adapted from journal and conference publications, observing the respective copyright agreements. This is indicated at the beginning of the chapters in which it occurs.

Copyright Declaration

The copyright of this Thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY).

Under this licence, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This on the condition that you credit the author.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.



Abstract

Audio augmented reality (AAR) is defined as the extension of a real auditory environment through virtual sound sources. A successful AAR system should create the illusion that virtual sounds actually come from the user's environment, for which several technical challenges must be overcome. First, room acoustics must be simulated accurately to predict the reverberant sound field produced by the virtual source as sound wavefronts reach the user. Second, said sound field must be translated into a pair of sound pressure signals at the user's ears. Finally, this binaural signal must be delivered to the user through an acoustically transparent system without limiting their ability to hear real sources. This process should be able to adapt in real time to user movements in a computationally efficient way, considering that resources may be limited in practice and most of them will likely be allocated to graphics processing (e.g. in a pair of augmented reality glasses).

This Thesis aims to improve current techniques for binaural audio rendering in AAR by exploring the trade-off between computational complexity and perceived quality. Several perception-focused studies were proposed to explore the different parts of the rendering process. First, a prototype AAR system with hear-through functionality was proposed and a pilot experiment was conducted to investigate how users could adapt to it over time. A second study assessed the effect of non-individualised equalisation on the perceived quality of binaural renderings reproduced with open-ear headphones. A third study evaluated several state-of-the-art methods for the binaural rendering of sound fields of limited resolution in the spherical harmonics (Ambisonics) domain. Finally, a fourth study assessed the perceptual effect of simplifying Ambisonics-based binaural reverberation in various ways.

Even though this Thesis focuses on the AAR scenario, the findings herein may be helpful for any application that would benefit from a computationally efficient implementation of binaural audio rendering methods.

Acknowledgements

I would like to thank my supervisors Lorenzo Picinali and Dan Goodman for giving me the opportunity to work at this exceptional university and for their guidance throughout my PhD.

I am also grateful to my past and present colleagues at Imperial College and Facebook Reality Labs, two places where I had the chance to meet many brilliant minds and extraordinary people. Thank you for the fascinating discussions and invaluable advice, but also for the pizza evenings, whisky tastings, and Beat Saber sessions. In particular, I would like to thank David Alon for being an excellent mentor and friend.

Thanks to my family and friends in Málaga and London for supporting me during throughout this journey. Special thanks to my stellar flatmates Sergio and Marc for the countless ping-pong and FIFA evenings, to my family in Caleta for giving me a second home, and to my parents and sister for their encouragement and unconditional support.

Finally, thanks to my fiancée Alba for her love, her patience, and for being a constant source of inspiration all these years.

Contents

List of Tables	xi
List of Figures	xiii
Acronyms	xix
1 Introduction	1
1.1 Motivation and Objectives	1
1.2 Contributions	5
1.3 Thesis structure	7
2 Technical Background	9
2.1 Binaural rendering	9
2.2 On the headphone transfer function	11
2.3 On the room impulse response	11
2.4 On the different coordinate systems	13
2.4.1 Polar coordinate system	13
2.4.2 Lateral-polar coordinate system	13
2.5 On the Ambisonics framework	14
2.5.1 Spherical Fourier transform	14
2.5.2 Binaural rendering of a sound field	16
2.5.3 Order truncation and aliasing frequency	17
2.5.4 Reducing the SH-HRTF's order	19
2.6 On perceptual hearing models	20
3 Pilot Study: User Adaptation to Acoustically Transparent Headphones	23

3.1	Introduction	24
3.2	AAR system	26
3.2.1	System description	26
3.2.2	Frequency response and equalisation	27
3.2.3	Latency and comb filter effect	29
3.3	Methodology	29
3.3.1	Equipment	30
3.3.2	Localisation test	31
3.3.3	Subjective evaluation	32
3.4	Results and discussion	33
3.4.1	Localisation test	33
3.4.2	Subjective evaluation	34
3.5	Conclusions	35
4	The Effect of Generic Headphone Compensation on Binaural Renderings	37
4.1	Introduction	38
4.2	Methods	40
4.2.1	Hardware setup	40
4.2.2	Headphone EQ approach	41
4.2.3	HpEQ and BRIR measurements	43
4.2.4	Audio material	44
4.2.5	HpEQ and BRIR test conditions	45
4.3	Objective evaluation	46
4.4	Subjective evaluation	48
4.4.1	Method	48
4.4.2	Results	51
4.5	Discussion	53
4.6	Conclusions	55
5	Binaural Rendering Methods: a Model-Based Assessment	57

5.1	Introduction	58
5.1.1	Binaural rendering and Ambisonics	58
5.1.2	Research question and contributions	60
5.2	Rendering methods	62
5.2.1	Truncation (Trunc)	63
5.2.2	Equalisation (EQ)	63
5.2.3	Tapering (Tap)	64
5.2.4	Time-alignment (TA)	65
5.2.5	Magnitude least squares (MagLS, MagLS+CC)	67
5.2.6	Spatial subsampling (SpSub, SpSubMod)	68
5.2.7	Bilateral magnitude least squares (BiMagLS)	69
5.2.8	Overview	70
5.3	Evaluation methods	71
5.3.1	Numerical analysis	71
5.3.2	Auditory models	73
5.4	Results	74
5.4.1	Numerical analysis	74
5.4.2	Auditory models	80
5.5	Discussion	85
5.5.1	Comparing rendering methods	85
5.5.2	Validity of the model-based assessment	86
5.5.3	Future work	88
5.6	Conclusions	89
6	Binaural Reverberation: Comparing Ambisonics-Based Approaches	91
6.1	Introduction	92
6.2	Background and motivations	95
6.2.1	Reverb perception: a summary	95
6.2.2	Spatial order perception and hybrid Ambisonics reverb	97
6.2.3	Multiple-source rendering and RVL	99

6.2.4	Contributions	101
6.3	Methods	102
6.3.1	Measurements and RIR generation	102
6.3.2	Binaural rendering	104
6.3.3	Audio material	105
6.4	Numerical Analyses	106
6.4.1	Descriptive analysis of Ambisonics RIR	106
6.4.2	Objective binaural metrics	107
6.4.3	Spectral analysis	109
6.4.4	Loudness stability	110
6.4.5	Summary of numerical analyses	111
6.5	Perceptual evaluation	112
6.5.1	Experiment 1: Paradigm	113
6.5.2	Experiment 1: Results	113
6.5.3	Experiment 2: Paradigm	116
6.5.4	Experiment 2: Results	118
6.6	Discussion	122
6.6.1	Effect of spatial order	122
6.6.2	Dynamic aspects	124
6.7	Conclusions	125
7	Conclusions	127
7.1	Summary of Thesis Achievements	127
7.2	Outlook and Future Work	130
	Bibliography	133
A	Proof of permission to reproduce published work	151

List of Tables

3.1	Test schedule. ‘Loc.’ stands for ‘Localisation test’, ‘Subj.’ stands for ‘Subjective evaluation’, ‘Open’ stands for ‘Open-ear’, and ‘HT’ stands for ‘Hear-Through’.	30
3.2	Characteristics of the audio fragments.	33
5.1	Evaluated binaural rendering methods.	63
5.2	Mean absolute ITD error between each method and the reference, per order (in μs).	77
5.3	Mean absolute ILD error between each method and the reference, per order (in dB).	78
5.4	PSD between each method and the reference, per order (in sones).	80
5.5	Lateral precision per method and order (in degrees). Reference: 3° .	81
5.6	Polar precision per method and order (in degrees). Reference: 16.39° .	82
5.7	Externalisation per method and order. Reference: 1.	83
5.8	SRM per method and order (in dB). Reference: 8.6 dB.	84
6.1	Acoustic parameters of the two measured rooms, including RT per octave band, EDT per octave band and broadband DRR, calculated according to Zahorik (2002).	103

List of Figures

1.1	Example AAR scenario and the challenges that it involves, assuming that a model of the room acoustics is already available. (Top) Listener wearing AR glasses, facing a real source and a virtual source. (Bottom left) The first challenge is to simulate how the sound produced by the virtual source would reach the listener after interacting with the room. (Bottom centre) The second challenge is to render binaural signals from the simulated sound field, here represented as a frontal plane wave. (Bottom right) The third challenge is to deliver a high-quality binaural signal to the listener’s ears through an acoustically transparent device, e.g. open-ear (left) or hear-through (right).	3
2.1	Illustration of a pair head-related transfer functions (HRTFs)	10
2.2	First 130 ms of a RIR, expressed in decibels relative to the peak value. The RIR was simulated with the image source method (Allen and Berkley, 1979) for an omnidirectional point source placed 10 m away from the receiver in a room with an approximate volume of 2342.7 m ³ . The mixing time, estimated according to Lindau, Kosanke, et al. (2012), is indicated.	12
2.3	(Left) Polar coordinate system. (Right) Lateral-polar coordinate system.	14
2.4	SH spectra of the FABIAN HRTF (Brinkmann, Lindau, Weinzierl, et al., 2017): (a) before preprocessing, (b) after time alignment through phase correction by ear alignment (Ben-Hur, Alon, Mehra, et al., 2019), and (c) after setting its phase to zero. The b90 and b99 parameters are shown, indicating the lowest spatial order that contains 90% and 99%, respectively, of the HRTF’s energy for a particular frequency bin. The SH spectrum is defined as the energy of the SH-HRTF’s coefficients at every order n , according to Equation 2.8.	18

3.1	Smartphone-based AAR system, inspired by the work by Härmä et al. (2004).	25
3.2	AAR system (left) and KEMAR (right).	28
3.3	Comparison of open-ear frequency response against unfiltered and equalised hear-through, where 0 dB represents the lowest amplitude value registered across all measurements.	28
3.4	Distribution of sound sources around a subject sitting in the middle of the room. Sources 1–8 are actual speakers, whereas 9–20 are virtual sources generated by panning pairs of speakers.	31
3.5	Results of the localisation test.	33
3.6	Results of the subjective evaluation. Average results of fragments office1–6 are displayed in the top row, and those of fragments street1–4 are shown in bottom row.	34
4.1	Custom floating headphones (Sennheiser MX475) and binaural microphones (Brüel and Kjær 4101-B) mounted on KEMAR head and torso simulator.	41
4.2	Example of HpEQ filter calculation. (a) Headphone transfer function (H), its direct inverse (H^{-1}) and regularised inverse (H_R^{-1}); (b) Regularisation parameters α and σ^2 for that HpTF.	43
4.3	Spectra of the audio materials. Curves were normalised and smoothed with a third-octave window.	45
4.4	Statistics of (a) calculated HpEQ filters and (b) BRIR magnitude spectra, for the left ear. Median, 10th and 90th percentiles across subjects are indicated with dashed lines; generic is indicated with a solid line. Curves were normalised and smoothed with a third-octave window.	46
4.5	Error curves for all test conditions with respect to the reference condition: (a) absolute monaural error; (b) absolute ILD error. Curves were calculated by convolving the BRIR and HpEQ filter, both dependent on the test condition, then subtracting the reference curve (IndBRIR+IndHpEQ), and finally taking the median value across subjects for each frequency bin (final result smoothed with a third-octave window).	47

- 4.6 Results of the MUSHRA test for each of the four assessed attributes. Interquartile ranges are indicated with boxes, medians with circles, and the most extreme data points with thin lines. Background colour (grey/white) denotes the BRIR type. Audio materials are, from left to right, *speech* (red), *pink noise* (green) and *guitar* (blue). 52
- 5.1 Absolute HRTF magnitude (left ear) and interaural phase delay errors, averaged across 110 directions in an approximate Lebedev grid. HRTFs were interpolated from truncated SH-HRTFs (Trunc method) for five different spatial orders (1, 10, 20, 30, 44). The dotted lines indicate an approximation of the just noticeable differences: 1 dB for magnitude and $20\ \mu\text{s}$ for phase delay. 75
- 5.2 Top: ITD in ms and ILD in dB, plotted as a function of azimuth on the horizontal plane for the same HRTFs evaluated in Figure 5.1. Bottom: violin plots showing the absolute ITDs and ILD errors for each HRTF on the horizontal plane, where the dotted lines represent the approximate just noticeable differences in anechoic conditions, according to Klockgether and van de Par (2016). 76
- 5.3 Absolute HRTF magnitude (left ear) and interaural phase delay errors, averaged across 110 directions in an approximate Lebedev grid. HRTFs were interpolated with the nine rendering methods and $N = 3$. The horizontal dotted lines indicate an approximation of the just noticeable differences: 1 dB for magnitude and $20\ \mu\text{s}$ for phase delay. The vertical dashed line indicates the aliasing frequency. 77
- 5.4 Top: ITD in ms and ILD in dB, plotted as a function of azimuth on the horizontal plane for HRTFs interpolated with the nine rendering methods and $N = 3$. Bottom: violin plots showing the absolute ITD and ILD errors for each HRTF on the horizontal plane, where the dotted lines represent the approximate just noticeable differences in anechoic conditions, according to Klockgether and van de Par (2016). 78

5.5	Top: Estimated loudness of the left-ear HRTF (in sones) for the 11950 available directions. The top-left plot shows the Reference (original HRTF) and the rest show HRTFs interpolated with the nine rendering methods and $N = 3$. Bottom: violin plots showing the loudness difference between each method and the reference, for all directions.	79
5.6	Left-ear PSD between the original HRTF and HRTFs that were interpolated with different methods and spatial orders 1 to 44 (in sones; lower is better). The reference is located at minus infinity.	80
5.7	Predicted lateral localisation precision (in degrees; lower is better) for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference data (black dotted line) was obtained from the original HRTF.	81
5.8	Predicted polar localisation precision (in degrees; lower is better) for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference data (black dotted line) was obtained from the original HRTF.	82
5.9	Predicted externalisation (between zero and one; higher is better) for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference is located at 1.	83
5.10	Predicted SRM (in dB), averaged for 180 masker positions, for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference data (black dotted line) was obtained from the original HRTF.	84
6.1	Direct sound path and first-order early reflections as they reach the left ear of a listener in three scenarios: (left) before any head rotation; (middle) canonical rendering after a head rotation of 30 degrees clockwise; and (right) RVL rendering after the same head rotation. Note how in (c) the direct sound path is accurate, whereas the room is head-locked, affecting the incoming direction of reflections. .	100

- 6.2 Comparison between the average execution time of the convolution stage in standard Ambisonics and RVL binaural rendering, as a function of the number of rendered sources, for two different RTs. A random input signal with a length of 1024 samples was used as input. Simulations were done in MATLAB using the overlap-add method (Oppenheim et al., 2001), running on a quad-core processor at 2.8 GHz. 101
- 6.3 The two measured rooms. (Left) The library is shown during a BRIR measurement. (Right) The trapezoid is shown during a RIR measurement. 103
- 6.4 Spatial RIR of zeroth to fourth spatial orders (N) for a source placed in front of a listener in the library (top) and trapezoid (bottom), as a function of time and azimuth. The division between direct sound and reverb (τ_{dir}) and the perceptual mixing time (τ_{mix}) are indicated. The direct sound ($t < \tau_{\text{dir}}$) is represented by a discrete spatial delta at ($t = 0\text{ s}, \varphi = 0^\circ$) to indicate that it was excluded from the Ambisonics rendering and generated through convolution with an HRIR. . . 107
- 6.5 Objective binaural metrics of left-ear BRIRs at ($\varphi = 30^\circ, \theta = 0^\circ$) for different rooms and test conditions. (Left) The IACC per octave band, separated per BRIR segment. (Right) The BQI for early reflections ($\tau_{\text{dir}} < t < \tau_{\text{mix}}$) and late reverb ($t > \tau_{\text{mix}}$). 108
- 6.6 (Left) Long-term average spectra of two test signals (speech and drums) convolved with left-ear BRIRs, corresponding to the different test conditions, for a source at ($\varphi = 30^\circ, \theta = 0^\circ$). (Right) The absolute difference between each spectrum and the reference ($N = 4$), averaged across 42 ERBs. 110
- 6.7 Loudness, K-weighted, relative to full scale (in dB) of the reverberant sound field, generated by convolving pulsated pink noise with a BRIR that had the direct sound removed, for different listener head orientations over the horizontal plane. Each line represents a different condition of experiment 2: first-order Hybrid Ambisonics, first-order Hybrid Ambisonics (static) and first-order RVL. 111

6.8	Results from experiment 1 represented by violin plots (Hintze and R. D. Nelson, 1998), which show the probability density of the data, median (circle), interquartile range (box), and mean (horizontal line). (Top) Both test locations pooled together. (Bottom) Separated per test location (from left to right, ‘in situ’ and ‘lab’). From left to right, both rooms pooled together, library and trapezoid data are shown. For this visualisation, data were scaled on a per-trial basis by setting the lowest and highest rating of every trial to 0 and 100. The vertical dotted lines indicate that the groups on the left are significantly different ($p < 0.05$) from the groups on the right (before normalisation).	114
6.9	Scatterplots showing all relative ratings (absolute ratings minus reference ratings) from experiment 1 and the amount of head rotation (azimuth) that listeners performed for each rating. A linear fit of the data is displayed in red, and the Pearson correlation coefficient is indicated at the top. Data for both rooms were pooled together.	116
6.10	Results from experiment 2 represented by violin plots (Hintze and R. D. Nelson, 1998), which show the probability density of the data, median (circle), interquartile range (box) and mean (vertical line). (Top) Both test locations pooled together. (Bottom) Separated per test location (from top to bottom, ‘in situ’ and ‘lab’). From left to right, both rooms pooled together, library and trapezoid data are shown. Ratings range from ‘definitely prefer A’ (−2) to ‘definitely prefer B’ (2). Asterisks indicate whether the mean is significantly different from zero: * for $p < 0.05$ and ** for $p \leq 0.005$	119
6.11	Scatterplots showing all paired comparisons’ ratings (except null pairs) from experiment 2 and the amount of head rotation (azimuth) that listeners performed for each rating. A linear fit of the data is displayed in red, and the Pearson correlation coefficient is indicated at the top. Data were separated per room (library on the left and trapezoid on the right).	121

Acronyms

AAR Audio augmented reality [iii](#), [2](#), [5–7](#), [23–27](#), [29](#), [30](#), [35–37](#), [57](#), [89](#), [91](#), [127–131](#)

AMT Auditory modeling toolbox [61](#), [73](#), [74](#)

AR Augmented reality [1](#), [2](#), [59](#), [128](#), [131](#)

BiMagLS Bilateral magnitude least squares [58](#), [62](#), [63](#), [69](#), [70](#), [76–86](#), [89](#), [90](#), [128](#), [129](#), [131](#)

BQI Binaural quality index [108](#), [109](#), [111](#), [112](#), [122](#)

BRIR Binaural room impulse response [10](#), [12](#), [38–40](#), [43–47](#), [50–56](#), [96–100](#), [102](#), [104–107](#),
[109–112](#), [117](#), [122](#), [124](#), [128](#)

COL Colouration [49–54](#)

DFT Discrete Fourier transform [109](#)

DIR Direction [49–54](#)

DIS Distance [49–54](#)

DRR Direct to reverberant ratio [98](#), [101](#), [103](#), [105](#), [110](#), [111](#), [123](#), [126](#)

EDT Early decay time [103](#)

EQ Equalisation [4–7](#), [27](#), [36–40](#), [44](#), [55](#), [56](#), [63–65](#), [67](#), [69](#), [70](#), [77–84](#), [87](#), [103](#), [117](#), [127](#), [128](#), [130](#)

ERB Equivalent rectangular bandwidth [44](#), [72](#), [109](#)

FDTA Frequency-dependent time-alignment [67](#), [69](#)

FIR Finite impulse response [26](#)

GenBRIR Generic binaural room impulse response [45](#), [48](#), [53](#), [54](#)

GenHpEQ Generic headphone equalisation [45](#), [47–49](#), [52–54](#)

HpEQ Headphone equalisation [11](#), [39–41](#), [43–47](#), [50–55](#), [128](#), [130](#), [131](#)

HpTF Headphone transfer function [11](#), [38](#), [39](#), [41](#), [43](#), [44](#), [55](#), [128](#)

HRF HRTF-related filters [63–65](#), [69](#)

HRIR Head-related impulse response [3](#), [6](#), [9](#), [10](#), [27](#), [58](#), [59](#), [65–67](#), [71–73](#), [93](#), [97](#), [99](#), [102–106](#), [108](#), [113](#), [117](#), [125](#)

HRTF Head-related transfer function [3](#), [4](#), [10](#), [14](#), [16–20](#), [25](#), [30](#), [41](#), [46](#), [48](#), [53](#), [58–60](#), [62–72](#), [74](#), [75](#), [77](#), [85](#), [88](#), [97](#), [103](#), [131](#)

IACC Interaural cross-correlation coefficient [66](#), [107–109](#), [122](#)

ILD Interaural level difference [4](#), [9](#), [20](#), [34](#), [47](#), [48](#), [53](#), [54](#), [64](#), [67](#), [72](#), [75–78](#), [81](#), [85](#)

IMU Inertial measurement unit [105](#)

IndBRIR Individualised binaural room impulse response [45](#), [47](#), [48](#), [50–53](#)

IndHpEQ Individualised headphone equalisation [45](#), [47–53](#)

ISFT Inverse spherical Fourier transform [15](#), [19](#), [69](#), [71](#)

ITD Interaural time difference [4](#), [9](#), [20](#), [34](#), [48](#), [65](#), [67](#), [72](#), [75](#), [76](#), [79–81](#), [86](#), [104](#)

JND Just noticeable difference [75](#), [86](#)

MagLS Magnitude least squares [63](#), [67–70](#), [76–89](#), [129](#)

MagLS+CC Magnitude least squares with covariance constraint [63](#), [67](#), [68](#), [70](#), [77–85](#), [88](#), [131](#)

- MaxIACCe** Maximum IACC calculated on the energy envelopes [72](#)
- MUSHRA** Multiple stimulus test with hidden reference and anchor [40](#), [48](#), [50](#), [113–115](#), [117](#), [118](#), [123](#)
- NoHpEQ** No headphone equalisation [45](#), [47–49](#), [52–54](#)
- OVS** Overall similarity [49–54](#)
- PSD** Perceptual spectral difference [72](#), [79](#), [82](#), [83](#), [85](#), [87](#), [89](#)
- RIR** Room impulse response [11](#), [12](#), [93](#), [94](#), [96–100](#), [102–104](#), [106](#), [107](#), [109](#), [111](#), [117](#), [122](#), [124](#)
- RM-ANOVA** Repeated measures analysis of variance [51](#), [52](#), [114](#), [115](#), [119](#)
- RT** Reverberation time [96](#), [103](#), [104](#), [106](#), [112](#), [131](#)
- RVL** Reverberant virtual loudspeaker method [92](#), [94](#), [95](#), [99–102](#), [104](#), [105](#), [107–112](#), [116–122](#), [124–126](#), [129](#)
- SAQI** Spatial Audio Quality Inventory [49](#)
- SD** Standard deviation [112](#), [113](#), [115](#), [118](#), [120](#)
- SFT** Spherical Fourier transform [14](#), [15](#), [63](#), [69](#)
- SH** Spherical harmonics [3–5](#), [14–19](#), [59](#), [60](#), [62–66](#), [68](#), [92](#), [93](#), [97](#), [98](#)
- SH-HRTF** Head-related transfer function in the spherical harmonics domain [16](#), [18–20](#), [62–71](#), [77](#), [78](#)
- SHF** Spherical head filters [64](#)
- SpSub** Spatial subsampling [63](#), [68–70](#), [77–87](#)
- SpSubMod** Modified spatial subsampling [63](#), [68–70](#), [77–85](#), [87](#)
- SRM** Spatial release from masking [74](#), [83](#), [85](#)

TA Time-alignment [63](#), [66](#), [67](#), [69](#), [70](#), [76–87](#), [89](#), [90](#)

Tap Tapering [63–65](#), [70](#), [72](#), [77–85](#), [87](#)

Trunc Truncation [20](#), [63](#), [70](#), [74](#), [76–84](#), [87](#), [88](#)

VR Virtual reality [1](#), [59](#), [128](#)

Chapter 1

Introduction

1.1 Motivation and Objectives

We define augmented reality (AR) as an interactive experience in which a user's perception of the real world is *extended* through an overlay of computer-generated elements. This should not be confused with virtual reality (VR), where the user's perception of the real world is *completely replaced* by a simulated environment. For instance, a videoconferencing application in AR might let the user see a hologram of their interlocutor that is apparently standing in front of them in the real world, whereas the counterpart in VR might transport both interlocutors to a virtual conference room. Recently, other terms such as 'mixed reality' or 'extended reality' have been coined to define experiences with varying level of exposure to real-world elements. However, for simplicity, this Thesis will just use 'AR' to refer to all such experiences that involve any level of integration of real and virtual elements.

In order to alter the way that a user perceives their environment, an AR system must be able to stimulate one or more of their senses, typically, sight, hearing, or touch. This could involve, for instance, displaying three-dimensional objects that can be localised by their sounds and that respond to interaction through haptic feedback. Having this functionality available in a system with a convenient form factor, like a pair of glasses, could potentially redefine how humans interact with their devices, e.g. by replacing the smartphone's or computer's screen by

an interactive high-resolution virtual display. Although some commercial products, such as the Microsoft Hololens and Magic Leap One headsets, already provide some of these features to a certain degree, they seem to have found limited use among the general public so far, perhaps due to their high cost, large form factor or inconvenient user interface when compared to devices such as smartphones or desktop computers. Whether AR will be mass-adopted by the general public is still uncertain, but it seems clear that further research in various disciplines is still required for its technology to reach the level of maturity needed to make that jump.

This Thesis focuses on the auditory part of AR, referred hereafter as audio augmented reality (AAR), which is here defined as the extension of a user's auditory environment through the addition of virtual sound sources. A successful AAR system should be able to deliver plausible spatial audio, meaning that the user (listener) should perceive the virtual sound sources as if they were at specific locations in their environment (Wightman and Kistler, 1989b). This poses several challenges, of which three of will be addressed in this Thesis and which are illustrated in Figure 1.1.

First, it is necessary to build a *room acoustics simulation* that predicts how the sound waves emitted by virtual sources would interact with the surroundings and produce reverberation before reaching the listener. For this, a room model must be created from a set of parameters (e.g. room geometry, material absorption) that may be acquired through sensors or retrieved from a database. Note that the sensing or estimation of the room acoustics is a challenging research problem in itself, but is outside the scope of this Thesis, in which it is assumed that room acoustics information is available. Once the model of the room is built, its acoustical features can be estimated, which allows to simulate the reverberant sound field that a sound source therein would produce. Said sound field must be realistic enough so that the listener perceives the virtual sources to blend with the real environment, e.g. the reverberation produced by virtual and real sources must match to a certain degree (Olive and Toole, 1989; Catic et al., 2015; Werner et al., 2016). Given the highly interactive nature of AR, the simulation should be frequently updated in real time in order to react to potential movements by the listener or any virtual source, which can be computationally costly if a complex model is employed (Schissler, Stirling, et al., 2017). Considering that an AR device may have limited computational resources

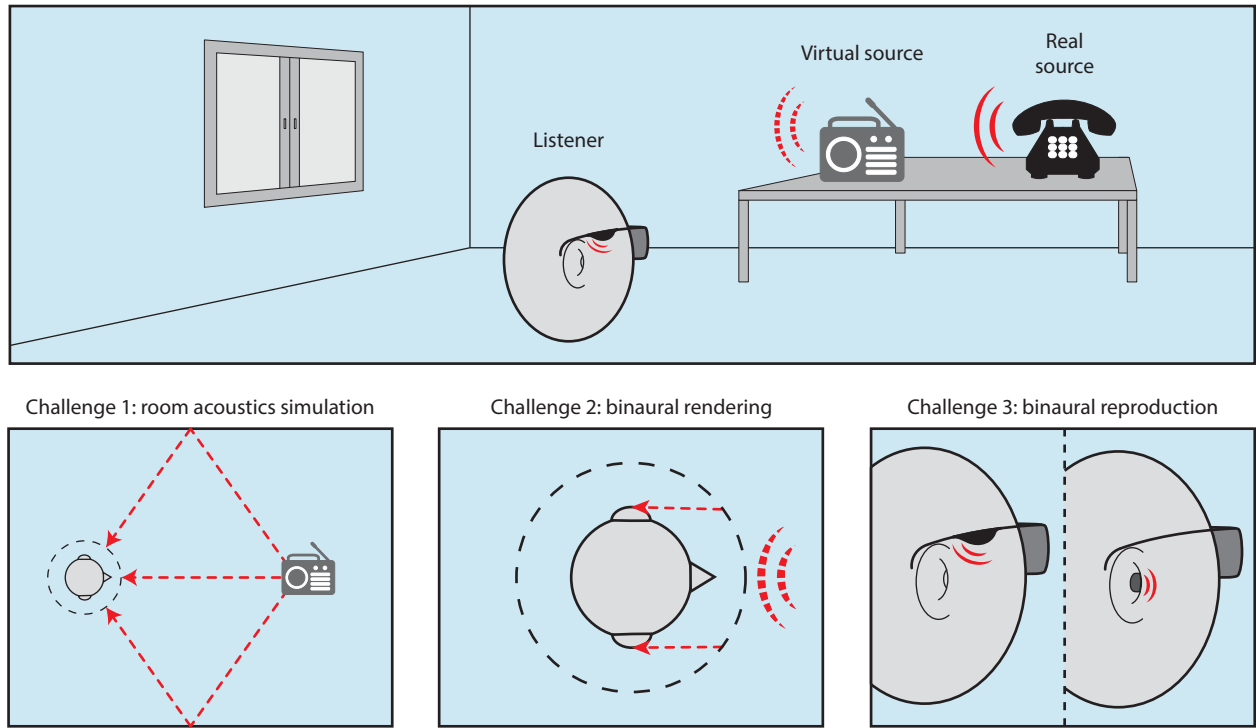


Figure 1.1: Example AAR scenario and the challenges that it involves, assuming that a model of the room acoustics is already available. (Top) Listener wearing AR glasses, facing a real source and a virtual source. (Bottom left) The first challenge is to simulate how the sound produced by the virtual source would reach the listener after interacting with the room. (Bottom centre) The second challenge is to render binaural signals from the simulated sound field, here represented as a frontal plane wave. (Bottom right) The third challenge is to deliver a high-quality binaural signal to the listener’s ears through an acoustically transparent device, e.g. open-ear (left) or hear-through (right).

due to its portability and that a good portion of those resources will likely be allocated to graphics, it is relevant to explore ways to make this process more efficient without degrading the perceived quality of the output. One promising approach is to reduce the spatial resolution of the sound field by means of a mathematical framework known as spherical harmonics (SH), also called Ambisonics in the context of audio production (Gerzon, 1973; Zotter and Frank, 2019). However, the extent to which a reverberant sound field can be simplified in the SH domain while preserving its perceived quality is not yet completely understood.

Second, a *binaural rendering* (also referred to as ‘auralisation’) of the simulated sound field must be performed. This process typically consists in decomposing the sound field into individual audio signals that arrive to the listener from different directions and convolving each one with the corresponding head-related impulse response (HRIR), also referred to as head-related

transfer function (HRTF) if expressed in the frequency domain. The HRTF is a listener-specific, direction-dependent filter that maps a signal emitted by a free-field source to the sound pressure that it produces at the listeners' left and right ears (Møller, 1992). The convolution outputs a binaural audio signal that contains the necessary cues for the human auditory system to localise the virtual sound source, e.g. interaural time differences (ITDs) and interaural level differences (ILDs) (Blauert, 1997). This approach is based on the assumption that the sound field can be expressed as a sum of a finite number of audio signals incoming from different directions, i.e. as provided by ray-tracing or the image source method (Allen and Berkley, 1979), and that multiple-scattering between the listener's head and the room objects is minimal, which is roughly true if the distance between them is large enough.

Alternatively, the binaural rendering may be performed in the SH (Ambisonics) domain. In this case, the sound field is expressed as a continuous function over a sphere surrounding the listener, and is based on the assumption that the HRTF is distance-invariant (not counting overall amplitude changes), which again is roughly true for large enough distances. For this approach, so-called Ambisonics-based binaural rendering methods are required to obtain a SH-based representation of the HRTF, as will be discussed in Chapter 2. It is hypothesised that Ambisonics-based binaural rendering will provide higher accuracy than traditional methods when working with room simulations with reduced spatial resolution, which will be relevant in a low-cost scenario. However, a variety of Ambisonics-based rendering methods have been proposed in the past and it has not yet been established which one performs best.

Third, the binaural signal must be *reproduced* back to the listener through an acoustically transparent hardware system that does not hinder the ability to hear real sources. This may be achieved with an open-ear system that does not occlude the ear canal, or by using a hear-through solution in which the listener's environment is recorded by binaural microphones and immediately played back through an in-ear transducer—in a similar way to how hearing aids work (Härmä et al., 2004). Additionally, the hardware system (or, simply, 'headphones') must be able to deliver a neutral presentation of the binaural signal within the desired frequency bandwidth without introducing spectral variations or other artefacts, which can be achieved through a combination of appropriate hardware and *equalisation* (*EQ*). Previous studies have

shown that headphone EQ filters significantly improved the quality of binaural signals when they were measured on the actual listener (Pralong and Carlile, 1996). However, individual EQ may not always be available and it is not yet well understood whether generic filters would provide similar perceptual benefits.

Overall, the broad goal of this Thesis is to build on the current knowledge of spatial audio perception through perceptual experiments and utilise it to improve the various techniques that are currently used to generate binaural audio in AAR. Nevertheless, it is worth noting that the Thesis' findings may also be relevant for any application concerning binaural audio rendering and reproduction. A particular focus is put on the trade-off between perceived quality and computational efficiency, with the SH framework (Ambisonics) often being the tool of choice for this purpose. Specifically, the main objectives of this Thesis are:

1. Assessing the effectiveness of different types of AAR hardware systems (hear-through and open-ear) in combination with headphone EQ.
2. Review and improve current Ambisonics-based binaural rendering methods.
3. Investigate the perceptual requirements when rendering Ambisonics-based binaural reverberation and propose efficient methods to compute it.

1.2 Contributions

With the aim of achieving the objectives outlined above, several technical developments and experiments were carried out. First, a basic prototype system was built as an initial step to better understand the challenges of AAR, such as the acoustical transparency of the device (Challenge 3). It consisted of a pair of in-ear headphones with hear-through functionality and EQ to compensate for the effect of ear canal occlusion. This prototype was employed to conduct a pilot study to investigate how well users could adapt to the hear-through functionality. The results of this pilot study indicated that users may experience slight adaptation effects,

particularly in terms of self-reported subjective attributes such as externalisation or realism. However, it must be noted that these findings have not yet been verified through a formal study.

Second, the issue of headphone EQ was addressed in more depth by conducting a perceptual study to assess the effect of generic headphone EQ on binaural renderings. To give some context, it is generally advised to employ individual headphone EQ when presenting binaural signals over headphones, to ensure the device's frequency response does not affect the audio content (see Chapter 2). However, individual measurements may not always be available in practice, and it is relevant to investigate if generic EQ could be a valid alternative. The novelty of this study compared to previous ones was that the experimental paradigm considered various perceptual attributes when evaluating the binaural signals, such as colouration and distance perception, rather than a single quality metric. The results showed that, when presenting binaural audio through an open-ear system, generic headphone EQ can still provide an improvement in perceived quality and, therefore, should be employed when individual EQ is not available.

Third, the issue of binaural rendering of order-limited Ambisonics sound fields was investigated (Challenge 2; see Chapter 2 for background on this topic). After conducting a literature review on binaural Ambisonics-based binaural rendering methods, it was found that many of them had been proposed but a thorough comparison had not been carried out yet. Therefore, a study was conducted to do precisely that: compare the most relevant rendering methods in order to establish which one produced the most accurate binaural signals from an Ambisonics sound field with low spatial resolution. Additionally, a novel rendering method was proposed that performed better than the state of the art across most of the evaluated metrics.

Finally, the issue of efficient rendering of binaural reverberation (Challenge 1) was investigated. To this end, a perceptual study was conducted where different simplifications to Ambisonics-based reverberation were implemented and evaluated through numerical analyses and listening tests. One of the main outcomes of this study was finding that second-order Ambisonics might be enough to render binaural reverberation if the direct sound path is computed accurately through convolution with an HRIR. This finding might be of high importance for the implementation of future real-time audio engines in AAR or other applications.

1.3 Thesis structure

The remainder of the Thesis is divided in six chapters, of which four were adapted from scientific articles that were submitted to conferences or journals. The contributions of each chapter are the following:

Chapter 2 provides the technical background that will be useful to provide context and better understand the contents of the other chapters.

Chapter 3 describes the features of the aforementioned prototype AAR system and the pilot study that investigated long-term user adaptation to its hear-through functionality. This study was presented at a conference by the present author and others (Engel and Picinali, [2017](#)).

Chapter 4 describes the perceptual study on generic headphone EQ, which was presented in a peer-reviewed conference contribution by the present author and others (Engel, Alon, Robinson, et al., [2019](#)).

Chapter 5 presents the review and comparison of Ambisonics-based binaural rendering methods using auditory models, a manuscript of which was submitted to the journal *Acta Acustica* in March 2021 and is under review at the time of writing this Thesis (Engel, Goodman, et al., [2021](#)).

Chapter 6 describes the study on Ambisonics-based binaural reverberation, which was published in a peer-reviewed journal article by the present author and others (Engel, Henry, Amengual Garí, Robinson, and Picinali, [2021](#)).

Finally, **Chapter 7** summarises the overall findings of this Thesis and discusses how these may help pushing the state of the art in the field of AAR and other applications. Ideas for future work are also suggested.

Chapter 2

Technical Background

The goal of this chapter is to introduce some technical background and definitions that will be useful to understand the contents of the Thesis.

2.1 Binaural rendering

A key concept employed throughout this Thesis is ‘binaural rendering’, sometimes also referred to as ‘auralisation’. This is defined as the process of recreating the sound pressure variations that would occur at the location of a listener’s left and right ears (in particular, at the entrance of the occluded ear canals, unless otherwise specified) as a consequence of a nearby sound source.

Basically, the binaural rendering is achieved by convolving an anechoic sound signal (i.e. the one originally emitted by the source) with a head-related impulse response (HRIR). The HRIRs are filters that replicate the effects caused by the listener’s anatomy (mainly head, torso and pinnae) on the incoming sounds and they provide the human auditory system with the necessary cues to localise a sound source, such as ITDs, ILDs and monaural cues (Blauert, [1997](#)). HRIRs are generally defined as a pair (one for the left ear and one for the right ear) for a given source position. More formally, the HRIR is calculated by spectral division (i.e. in the frequency domain) between the sound pressure measured at the ears’ positions and the sound pressure

measured at the centre of the interaural axis (origin) without the head present, as illustrated in Figure 2.1. When an HRIR is expressed in the frequency domain, it is referred to as the head-related transfer function (HRTF).

HRTFs can be acoustically measured on a listener or numerically simulated from anatomical data (Brinkmann, Lindau, Weinzierl, et al., 2017). In this Thesis, it is always assumed that HRTFs or HRIRs are measured in anechoic conditions, meaning that they do not include any acoustical reflections from the room where they were measured, or any objects apart from the listeners' anatomy. In cases where the effect of the room is considered (e.g. when discussing reverberation in Chapter 6) it will be explicitly stated that a binaural room impulse response (BRIR), rather than a HRIR, has been employed for the rendering.

A single convolution with an HRIR produces the binaural rendering of a single far-field source. By adding additional convolutions, it is possible to aggregate more sources or reflections produced as a result of the interaction with room, forming a sound field. Note that, unless otherwise specified, this work assumes far-field HRIRs, and therefore near-field effects such as parallax (Cuevas-Rodríguez et al., 2019) are not considered.

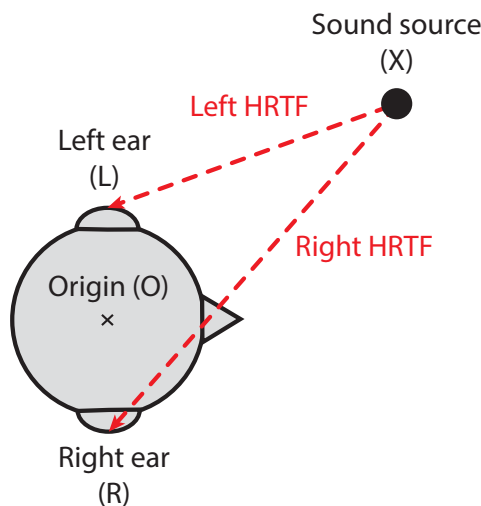


Figure 2.1: Illustration of a pair head-related transfer functions (HRTFs)

2.2 On the headphone transfer function

The headphone transfer function (HpTF) is defined as the acoustical transfer function between the binaural reproduction device (headphones) and the listener’s eardrums. It is typically measured by placing binaural microphones at the entrance of the listener ear canals and playing an excitation signal from the headphones, e.g. a sine sweep (Farina, 2007).

For the reproduction of binaural signals, it is typically preferred that the headphone transfer function (HpTF) is as neutral as possible, or ‘flat’, so that the magnitude and phase of the signal are not modified by effect of the headphones (Pralong and Carlile, 1996). In practice, this is often enforced by compensating the HpTF by means of a headphone equalisation (HpEQ) filter, which is calculated from the HpTF’s inverse (plus a regularisation process, as described later in Chapter 4). Unfortunately, the HpTF is highly dependent on the shape of the listener’s ears (Møller, Hammershøi, et al., 1995) and, therefore, personalised filters must be calculated for each individual listener to achieve an accurate equalisation. Throughout this thesis, we sometimes refer to ‘individual’ (personalised) and ‘non-individualised’ or ‘generic’ (not personalised) HpEQ, in reference to where the headphones were measured before calculating the filters. Even though individual HpEQ is generally preferred, it may not be available in all scenarios (e.g. a typical consumer does not own a pair of binaural microphones to measure their HpTF) so a generic HpEQ must be considered as a practical alternative, as discussed in Chapter 4.

2.3 On the room impulse response

In order to discuss the role of reverberation in the process of binaural rendering (see Chapter 6), it is important to define the room impulse response (RIR), which is a common concept in room acoustics literature. A RIR is the acoustic transfer function between a sound source and a receiver in a given acoustic space (room), assuming that the system formed by these is linear and time-invariant. It defines the sound pressure measured at the receiver’s position after the source has emitted an impulsive sound with infinite energy and infinitesimal duration (Dirac’s delta), which allows to predict how any sound emitted by the source will be transformed by

the room as it reaches the receiver. To capture spatial information (e.g. direction of arrival of a given echo), one may measure multichannel RIRs whose various channels correspond to different receivers placed around the listener’s position—e.g. see B-format microphones and spherical microphone arrays. In the special case where two receivers are placed at the ear canals of a human listener or a dummy head, the two resulting RIRs form a binaural room impulse response (BRIR).

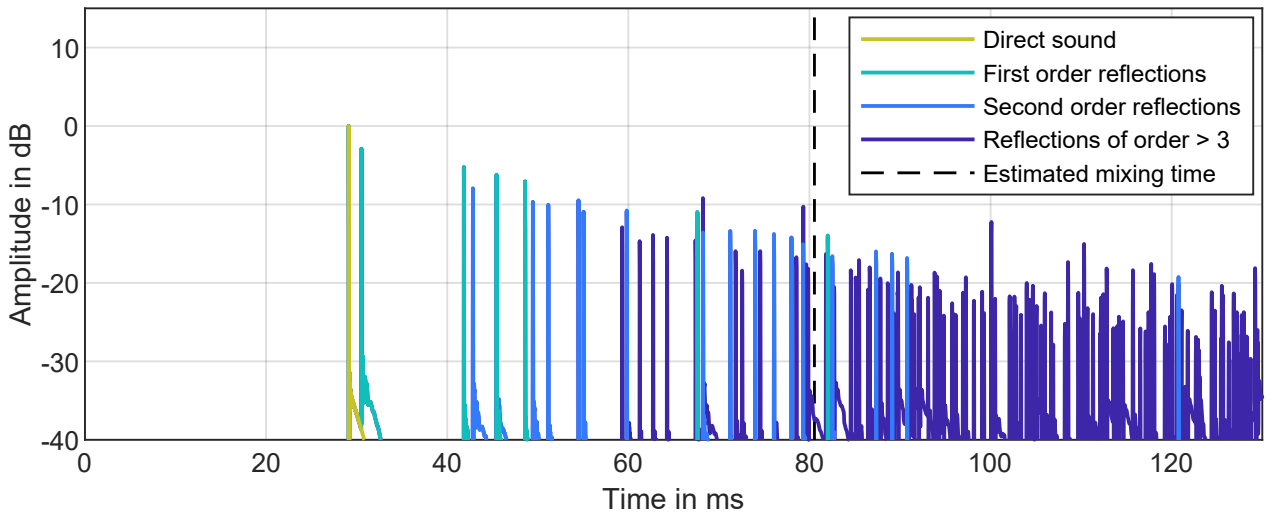


Figure 2.2: First 130 ms of a RIR, expressed in decibels relative to the peak value. The RIR was simulated with the image source method (Allen and Berkley, 1979) for an omnidirectional point source placed 10 m away from the receiver in a room with an approximate volume of 2342.7 m³. The mixing time, estimated according to Lindau, Kosanke, et al. (2012), is indicated.

In a broad sense, an RIR can be split into three temporal segments (see Figure 2.2): the *direct sound*, which is equivalent to an anechoic or free-field response; *early reflections*, which are the first echoes to arrive after the direct sound, generally sparse in their time and direction of arrival; and *late reverberation*, which is the ‘tail’ of the RIR, where echoes have higher temporal density and converge towards a diffuse sound field. The separation between early reflections and late reverberation, also called ‘mixing time’, is not exact and depends on human perception; Lindau, Kosanke, et al. (2012) defined it as the instant after which the RIR does not perceptibly change across different listener’s positions or orientations within the room. Throughout this work, all the parts of the RIR that are not the direct sound (i.e. early reflections and late reverberation) are sometimes simply referred to as ‘reverberation’ or ‘reverb’.

2.4 On the different coordinate systems

Throughout this Thesis, different coordinate systems are employed to describe the position of points relative to the listener. In general, they are egocentric spherical systems where the origin is located at the centre of the listener's interaural axis, which is the line that goes through the left and right ears. A common frame of reference is used regardless of the coordinate system, which includes the median plane and the equator. The median plane is defined as the vertical plane that contains the origin and that separates the left and right sides of the listener's anatomy. The equator is defined as the set of points that are at the same height as the interaural axis. The coordinate systems are listed below and illustrated in Figure 2.3.

2.4.1 Polar coordinate system

The main coordinate system that is used throughout the Thesis is the polar coordinate system, where points are expressed in terms of azimuth angle (ϕ), elevation angle (θ), and radius (r). Given a point X and the vector between the origin and said point $\bar{x}$, the azimuth of X is defined as the angle between $\bar{x}$ and the median plane, measured counter-clockwise from the latter (frontal direction). The elevation of X is defined as the angle between $\bar{x}$ and the equator, measured upwards from the equator unless otherwise stated. Note that azimuth and elevation are essentially defined in the same way as longitude and elevation as used most geographic coordinate systems. Angles are expressed in degrees unless otherwise stated, with azimuth being always positive ($0^\circ < \phi < 360^\circ$) and elevation being positive above the equator and negative below it ($-90^\circ < \theta < 90^\circ$). For instance, the point ($\phi = 90^\circ, \theta = 25^\circ$) is located 90 degrees at the left of the listener and 25 degrees above the horizontal plane.

2.4.2 Lateral-polar coordinate system

A second coordinate system is sometimes used in this Thesis, known as the lateral-polar system. In this case, points are expressed in terms of lateral angle (ψ) and polar angle (ζ). To calculate

them, a vertical plane (i.e. parallel to the median plane) containing the point X is first drawn, which defines the so-called cone of confusion, which is a region where a sound source location cannot be unambiguously determined by looking at interaural differences (Blauert, 1969). The lateral angle is defined as the arc between the median plane said vertical plane, whereas the polar angle is defined as the arc between the equator and X within the cone of confusion.

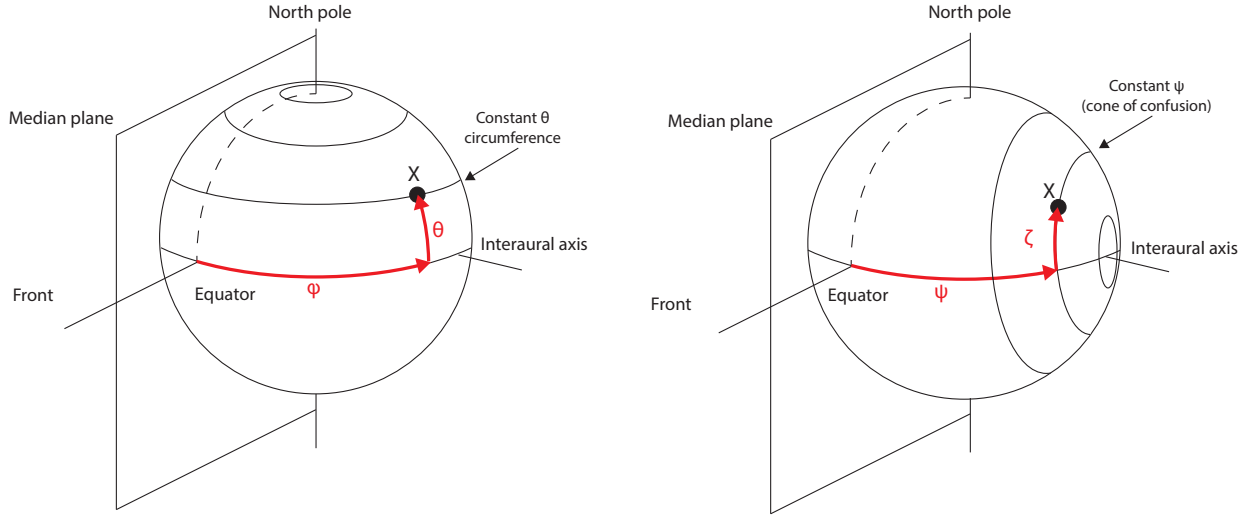


Figure 2.3: (Left) Polar coordinate system. (Right) Lateral-polar coordinate system.

2.5 On the Ambisonics framework

The purpose of this section is to provide some mathematical foundations about the Ambisonics framework which is largely used in Chapter 5 and Chapter 6. Most of the notation is borrowed from Rafaely and Avni (2010), Zotter and Frank (2019) and Bernschütz (2016).

2.5.1 Spherical Fourier transform

The main concept behind the Ambisonics technique is to express spatial audio signals, e.g. a three-dimensional sound field or an HRTF, as spherical functions described by SH coefficients, which enables various useful post-processing and playback options. The process of obtaining the SH coefficients from a spatial audio signal is known as the spherical Fourier transform

(SFT). Similarly to how the Fourier transform is used to express a time-domain signal as a series of frequency coefficients, the SFT can express a signal sampled at discrete directions (i.e. at specific elevation/azimuth angles) over a sphere as a series of SH coefficients (see Rafaely, 2015, Section 1.4). Given a function $x(\theta, \phi)$ sampled at a set of points, where θ is the elevation measured downwards from the north pole and ϕ is the azimuth measured counter-clockwise from the front, and the radius is fixed, its SH coefficients are calculated with the SFT, as (see Rafaely and Avni, 2010, Equation 1):

$$x_{nm} = \mathcal{SFT}\{x(\theta, \phi)\} \equiv \int_0^{2\pi} \int_0^\pi x(\theta, \phi) Y_n^m(\theta, \phi) \sin \theta d\theta d\phi, \quad (2.1)$$

where $Y_n^m(\theta, \phi)$ are the normalised, real-valued SH of order n and degree m (see Zotter and Frank, 2019, Equation A.35):

$$Y_n^m = (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos \theta) y_m, \quad (2.2)$$

with

$$y_m = \begin{cases} \sqrt{2} \sin(|m|\phi) & m < 0 \\ 1 & m = 0 \\ \sqrt{2} \cos(m\phi) & m > 0 \end{cases} \quad (2.3)$$

and where $P_n^m(x)$ is the associated Legendre function, calculated as in (Williams, 1999, Equation 6.29). Applying the SFT to a signal is sometimes called ‘Ambisonics encoding’. Analogously, the inverse spherical Fourier transform (ISFT) or ‘Ambisonics decoding’ is defined as:

$$x(\theta, \phi) = \mathcal{ISFT}\{x_{nm}\} \equiv \sum_{n=0}^{\infty} \sum_{m=-n}^n x_{nm} Y_n^m(\theta, \phi). \quad (2.4)$$

Note that SH conventions vary depending on the scientific field and the author’s style. In

this work, we chose the real-valued formulation employed by Zotter and Frank (2019), which is commonly used in Ambisonics and is more convenient than complex-valued ones, e.g. as in (Rafaely and Avni, 2010), because it does not involve the complex conjugation of the Y_n^m term in Equation 2.1, and is therefore simpler to implement while providing the same results. In contrast, the complex-valued formulation eliminates the need of the y_m term at the cost of working with complex values rather than real ones. The reader is referred to the work by Poletti (2009) and Andersson (2016) for further discussion on SH conventions and their use in Ambisonics.

2.5.2 Binaural rendering of a sound field

We define a sound field as a sum of an infinite number of plane waves (PW) and we describe it with a PW density function, $a(f, \theta, \phi)$, which varies over frequency and direction. For its binaural rendering, the sound pressure at the left ear can be calculated in the frequency domain by multiplying each PW with the corresponding left-ear HRTF $h^l(f, \theta, \phi)$ across all directions (see Rafaely and Avni, 2010, Equation 7):

$$p^l(f) = \int_0^{2\pi} \int_0^\pi a(f, \theta, \phi) h^l(f, \theta, \phi) \sin \theta d\theta d\phi. \quad (2.5)$$

By substituting $a(f, \theta, \phi)$ and $h^l(f, \theta, \phi)$ with their SH representation (Equation 2.4) and applying the SH orthogonality property (see Rafaely, 2015, Equation 1.23), we obtain:

$$p^l(f) = \sum_{n=0}^{\infty} \sum_{m=-n}^n a_{nm}(f) h_{nm}^l(f), \quad (2.6)$$

where $a_{nm}(f)$ and $h_{nm}^l(f)$ are the SH coefficients of $a(f, \theta, \phi)$ and $h^l(f, \theta, \phi)$, respectively (see Rafaely and Avni, 2010, Equations. 8–10). We may also refer to $a_{nm}(f)$ as the Ambisonics signal and to $h_{nm}^l(f)$ as the SH-HRTF. Note that this same process can be performed for the right-ear HRTF $h^r(f)$ to obtain the pressure at the right ear, $p^r(f)$, to produce the complete binaural

signal. Hereafter, the left and right superscripts are omitted for brevity, and it is assumed that both ears are processed separately.

2.5.3 Order truncation and aliasing frequency

In practice, the infinite summation in Equation 2.6 must be truncated at some finite order N , which yields an approximation of the true binaural signal:

$$\hat{p}(f) = \sum_{n=0}^N \sum_{m=-n}^n a_{nm}^N(f) h_{nm}^N(f), \quad (2.7)$$

where the superscripts indicate that the SH coefficients have been truncated at order N . This order truncation causes a loss of information that can lead to audible artefacts in the binaural signal $\hat{p}(f)$, such as over-emphasised low frequencies or poor localisation of sound sources (Avni et al., 2013). The cause of these artefacts can be intuitively explained by looking at the HRTF's SH spectrum, defined by Ben-Hur, Alon, Mehra, et al. (2019) as the energy of its SH coefficients for each order (n):

$$E_n(f) = \sum_{m=-n}^n |h_{nm}(f)|^2. \quad (2.8)$$

Looking at the SH spectrum in Figure 2.4a, we observe that an HRTF's high-frequency content is mainly stored at high orders, meaning that order truncation will cause a loss of mostly high-frequency content, which explains the spectral colouration described in (Avni et al., 2013). The dotted lines, which roughly indicate the upper boundary of the SH spectrum, increase almost linearly with frequency. In fact, previous work has shown that the minimum truncation order (N_a) required to contain an HRTF's SH spectrum up to a given frequency f_a , follows the linear relation:

$$N_a \simeq \frac{2\pi f_a r}{c}, \quad (2.9)$$

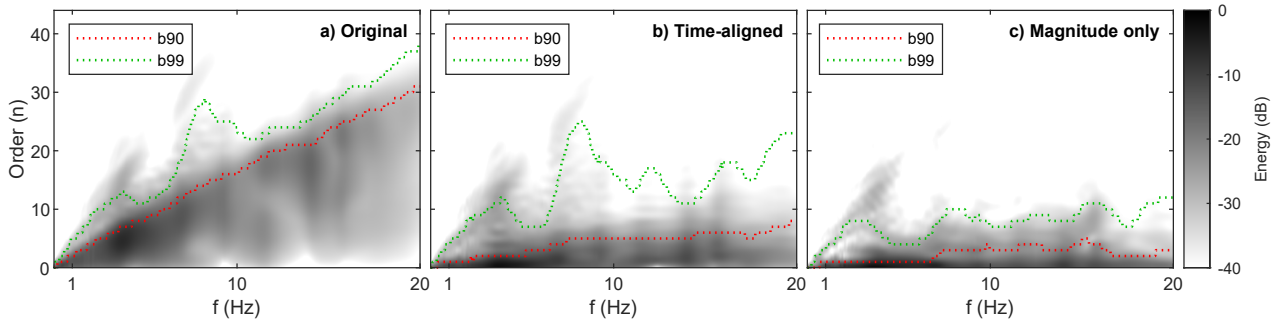


Figure 2.4: SH spectra of the FABIAN HRTF (Brinkmann, Lindau, Weinzierl, et al., 2017): (a) before preprocessing, (b) after time alignment through phase correction by ear alignment (Ben-Hur, Alon, Mehra, et al., 2019), and (c) after setting its phase to zero. The b90 and b99 parameters are shown, indicating the lowest spatial order that contains 90% and 99%, respectively, of the HRTF’s energy for a particular frequency bin. The SH spectrum is defined as the energy of the SH-HRTF’s coefficients at every order n , according to Equation 2.8.

where c is the speed of sound and r is the radius of the smallest sphere surrounding the listener’s head (Ben-Hur, Alon, Mehra, et al., 2019). Conversely, it can also be said that, for a given truncation order N_a , the SH spectrum can only be accurately represented up to the HRTF’s spatial aliasing frequency f_a (Rafaely, 2005; Bernschütz, 2016, Section 3.8):

$$f_a \simeq \frac{N_a c}{2\pi r}. \quad (2.10)$$

Therefore, assuming a speed of sound of $c \simeq 343$ m/s and a nominal head radius of $r \simeq 0.0875$ m (Ben-Hur, Alon, Mehra, et al., 2019), a truncation order of approximately 32 would be needed to accurately represent an HRTF in the SH domain within the audible spectrum (up to 20 kHz), which is in agreement with Figure 2.4a. In other words, truncating the order of an SH-HRTF below 32 could potentially introduce audible artefacts in any binaural signal rendered with it.

However, the truncation order is sometimes imposed in practice as a constraint of the binaural rendering application, usually because the Ambisonics signal is given with a limited order, as discussed in Section 5.1. Since the order of a binaural rendering is dictated by the lowest order between $a_{nm}(f)$ and $h_{nm}(f)$, as established by Ben-Hur, Sheaffer, et al. (2018), the Ambisonics signal will impose its lower order even if the SH-HRTF has a higher one—the opposite could also happen, but it is less common. Therefore, the SH-HRTF’s order must be reduced.

2.5.4 Reducing the SH-HRTF's order

The most straightforward way to reduce the an SH-HRTF's spatial order to N , so it matches the Ambisonics signal, is to simply truncate it by removing all SH coefficients from $N + 1$ onwards. In practice, this is typically done by applying the discrete version of the spatial Fourier transform as in (Ben-Hur, Alon, Mehra, et al., 2019):

$$\mathbf{h}_{\text{nm}}^{\text{N}} = \mathbf{h} \mathbf{Y}^{\text{N}\dagger}, \quad (2.11)$$

$$\mathbf{h}_{\text{nm}}^{\text{N}} = \begin{bmatrix} h_{00}(f_1) & h_{0(-1)}(f_1) & \dots & h_{NN}(f_1) \\ \vdots & \ddots & & \\ h_{00}(f_F) & & & h_{NN}(f_F) \end{bmatrix}$$

$$\mathbf{h} = \begin{bmatrix} h(f_1, \theta_1, \phi_1) & \dots & h(f_1, \theta_D, \phi_D) \\ \vdots & \ddots & \\ h(f_F, \theta_1, \phi_1) & & h(f_F, \theta_D, \phi_D) \end{bmatrix}$$

$$\mathbf{Y}^{\text{N}} = \begin{bmatrix} y_{00}(\theta_1, \phi_1) & \dots & y_{00}(\theta_D, \phi_D) \\ y_{0(-1)}(\theta_1, \phi_1) & \ddots & \\ \vdots & & \\ y_{NN}(\theta_1, \phi_1) & & y_{NN}(\theta_D, \phi_D) \end{bmatrix}$$

where $\mathbf{h}_{\text{nm}}^{\text{N}}$ is a matrix representation of the truncated SH-HRTF with as many rows as frequency coefficients (F) and as many columns as SH coefficients $(N + 1)^2$; $\mathbf{h}$ is a matrix representation of the HRTF with F rows and as many columns as measured directions (D); $\mathbf{Y}^{\text{N}}$ is a matrix containing the SH up to order N sampled at the HRTF's directions; and $\dagger$ denotes the pseudoinverse. Note that there is also a discrete version of the ISFT, which is typically

employed to interpolate an HRTF ($\hat{\mathbf{h}}$) for a desired set of directions (Ben-Hur, Alon, Mehra, et al., 2019):

$$\hat{\mathbf{h}} = \mathbf{h}_{nm}^N \mathbf{Y}^N. \quad (2.12)$$

Truncating and interpolating an HRTF without preprocessing (see Trunc method in Section 5.2) leads to the audible artefacts discussed earlier. Several approaches have been proposed to reduce the order of an SH-HRTF in such a way that such artefacts are alleviated and binaural renderings are more accurate—these are the other methods reviewed in Section 5.2.

Note that the coarse sampling of the sound field or the HRTF can also lead to spatial aliasing errors, especially when dealing with low-order microphone array recordings (Bernschütz, 2016). However, this work assumes that both $a_{nm}(f)$ and $h_{nm}(f)$ are alias-free, e.g. as in a plane-wave-based audio engine that has access to a high-order HRTF (Schissler, Stirling, et al., 2017), and focuses on truncation-related errors. For a discussion on aliasing mitigation methods, the reader is referred to the work by Lübeck (2019).

2.6 On perceptual hearing models

Perceptual hearing models (or auditory models) allow us to simulate how a human listener hears, processes or even responds to auditory stimuli. Typically, models have a series of stages that reflect the human auditory system, such as cochlear filtering, which is typically simulated through a filter bank (Glasberg and Moore, 1990).

We distinguish between monaural models, which are those that consider a single-channel audio signal input, and binaural models, which consider two inputs (left and right ears). Binaural models allow us to exploit interaural differences (ITDs,ILDs) and other important cues for spatial audio perception, such as interaural coherence. In this Thesis, we employ binaural models as a quick and efficient way to simulate the results of listening tests, such as localisation tasks with the model by Reijniers et al. (2014), externalisation assessment with the model by

Baumgartner and Majdak (2020) and speech intelligibility in noise with the model by Jelfs et al. (2011). The usage of auditory models as an evaluation method for binaural signals is further discussed in Chapter 5.

Chapter 3

Pilot Study: User Adaptation to Acoustically Transparent Headphones

At the initial stages of the research project that led to this Thesis, the focus was put on building a pair of acoustically transparent headphones that served as an initial hardware prototype for AAR. To that end, a system with binaural recording and reproduction capabilities was built with off-the-shelf components and custom software. This prototype system allowed to present binaural audio to the listener while letting them hear the outside world thanks to low-latency hear-through functionality. This chapter describes an initial exploration that was performed to investigate how much listeners could adapt to such hear-through functionality over time.

This chapter was adapted from an article by the present author and others (Engel and Picinali, 2017) that was presented at the 24th International Congress on Sound and Vibration in London, United Kingdom, in 2017. This was done with permission of the rights holder, the Institute of Acoustics (United Kingdom); the permission letter can be found in the appendix.

Abstract

Audio augmented reality (AAR) consists of extending a real auditory environment with virtual sound sources. This can be achieved using binaural earphones/microphones. The microphones, placed in the outer part of each earphone, record sounds from the user's environment, which are

then mixed with virtual binaural audio, and the resulting signal is finally played back through the earphones. However, previous studies show that, with a system of this type, audio coming from the microphones (or hear-through audio) does not sound natural to the user. The goal of this study is to explore the capabilities of long-term user adaptation to an AAR system built with off-the-shelf components (a pair of binaural microphones/earphones and a smartphone), aiming at achieve perceived realism for the hear-through audio. To compensate the acoustical effects of ear canal occlusion, the recorded signal is equalised in the smartphone. In-out latency was minimised to avoid distortion caused by comb filtering effect. To evaluate the adaptation process of the users to the headset, two case studies were performed. The subjects wore an AAR headset for several days while performing daily tests to check the progress of the adaptation. Both quantitative and qualitative pilot evaluations, i.e. localising real and virtual sound sources and analysing the perception of pre-recorded auditory scenes, were carried out, finding slight signs of adaptation, especially in the subjective tests.

3.1 Introduction

AAR consists in extending a real auditory environment with virtual sound sources (Härmä et al., 2004; Karjalainen et al., 2008; Tikander, 2009). One of the biggest challenges in making an AAR system is that it must be acoustically transparent, meaning that it can deliver sound to the ears without modifying the signal or altering the user's natural hearing capabilities, e.g. localisation accuracy (Härmä et al., 2004; Tikander, 2009; Brungart et al., 2003). Different potential approaches exist for AAR, and the present authors plan to undertake long-term research on the most promising ones, being this study the first step on that path. Future stages of the project will explore technologies such as open-fitting headphones, hearing-aid-based systems, or receiver-in-the-canal devices (as discussed in later sections).

This study explores an acoustically transparent hearing device based on insert-type earphones with embedded binaural microphones (see Figure 3.1). The user's acoustic environment is captured by the microphones and equalised, and the resulting signal (from now on, hear-through audio) is mixed with auralised virtual audio and presented through the earphones (Härmä et al., 2004; Karjalainen et al., 2008). As sound is recorded near the entrance of the ear canal, loss of

spatial information is minimised (Tikander, 2009; Brungart et al., 2003). However, the presence of earplugs alters the natural resonance in the ear canals (Rona et al., 2015) and sound may leak from outside to inside the earphones (particularly at low frequencies), all of which causes that hear-through audio shows ‘colouration’ with respect to natural hearing (Karjalainen et al., 2008). Therefore, to accurately replicate the acoustics of an open ear, the recorded signals must be equalised before mixing them with virtual audio (Karjalainen et al., 2008; Tikander, 2009; Hoffmann et al., 2013; Hiipakka et al., 2012; Albrecht et al., 2011).

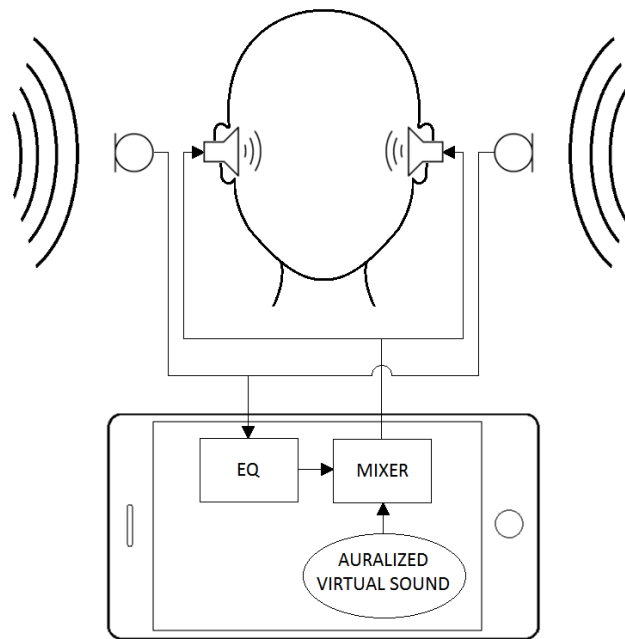


Figure 3.1: Smartphone-based AAR system, inspired by the work by Härmä et al. (2004).

Tikander (2009) obtained positive results after testing an equalised hear-through function on a portable AAR system in real life situations, and observed some adaptation after the subjects wore the device for at least 1.5 hours. Previous studies already suggested that humans can adapt to changes in their HRTF if they are exposed to them for an extended period of time (Blauert, 1997; Wanrooij and Opstal, 2005; Mendonça et al., 2012). The goal of this study is to further explore the effect of long-term user adaptation to an AAR system built with off-the-shelf components (rather than custom-made ones), aiming at achieve perceived realism in an equalised hear-through function. Two individual case studies were performed where subjects used a AAR headset for several days while performing periodic control tests (localisation accuracy and subjective perception) to check the progress of the adaptation.

The following sections will outline the proposed AAR device, the adaptation test methods, and the results of the experiment. The chapter ends with concluding thoughts and ideas for future work.

3.2 AAR system

3.2.1 System description

The proposed AAR system uses insert-type earphones and binaural microphones to implement a hear-through function, a concept that was first introduced by Härmä et al. (2004) and further developed by Karjalainen et al. (2008) and Tikander (2009). Contrary to those approaches, which implemented an AAR mixer/equaliser with custom analog circuits, this device is smartphone-based and only uses commercial off-the-shelf components. Therefore, it is more accessible to the public and can be easily integrated with other applications, such as music listening, hands free calling or games, which is one of the main areas of improvement suggested by test subjects in previous studies (Tikander, 2009; Albrecht et al., 2011).

The components (see Figure 3.1) are:

1. A pair of Roland CS-10EM binaural earphones/microphones (Hamamatsu, Japan).
2. A TASCAM iXJ2 audio interface (Los Angeles, CA, United States), to power the microphones and provide stereo audio input.
3. An Apple iPhone 5 smartphone (Cupertino, CA, United States) running iOS 10.

Audio processing and routing is performed by a custom app based on the open source library AudioKit (<https://audiokit.io>, last viewed April 21, 2021). The app includes a sound equaliser, which consists of four second-order parametric filters (one low-shelf filter and three peak/notch filters) that the user can adjust as needed. This approach was chosen over more complex solutions, e.g. a high-order finite impulse response (FIR) filter, to minimise in-out latency as discussed in Subsection 3.2.3.

3.2.2 Frequency response and equalisation

If an unfiltered hear-through function was implemented in the AAR system, the user would perceive a ‘coloured’ version of the acoustic environment which would sound unnatural. This is due to the ear canal’s resonance being modified by the earphones and to sound leakage incoming from outside the headphones, especially at the lower frequency spectrum (Karjalainen et al., 2008; Tikander, 2009; Rona et al., 2015; Hoffmann et al., 2013; Hiipakka et al., 2012; Albrecht et al., 2011). To accurately replicate the experience of natural hearing, the hear-through signal must be equalised. To find the right EQ curve, frequency response was calculated by measuring HRIRs on a KEMAR head and torso simulator (GRAS, Holte, Denmark) in two scenarios:

1. without the AAR system (from now on, open-ear) and
2. with the AAR system running an unequalised hear-through function.

The result of subtracting the first curve from the second one (in dB, i.e. through division of the respective spectra) gives the response of the audio filter that must be implemented in the iOS app. A third scenario was measured later, using the equalised hear-through function. HRIRs (0° azimuth, 0° elevation) were measured using the sine sweep technique (Farina, 2007). An Equator Audio D5 speaker (Los Angeles, CA, USA) was used to play sweeps from 20 Hz to 20 kHz and a length of 4 s. It was placed approximately 2 m away from the KEMAR (see Figure 3.2), at the same height and facing one another. The speaker input and the KEMAR output were connected to a computer through a MOTU UltraLite Mk3 audio interface (Cambridge, MA, United States).

The filter to be implemented was not direction-dependent so, even though off-axis positions were considered only frontal HRIRs were ultimately measured. For the sake of consistency, two different pairs of Roland CS-10EM were used, performing 10 HRIR measurements with each one in each scenario, for a total of 20 results per scenario, which were then averaged. Considering that the HRIR could vary depending on the earphone fit in the ear, earplugs were removed and reinserted again before each measurement in the hear-through scenario.

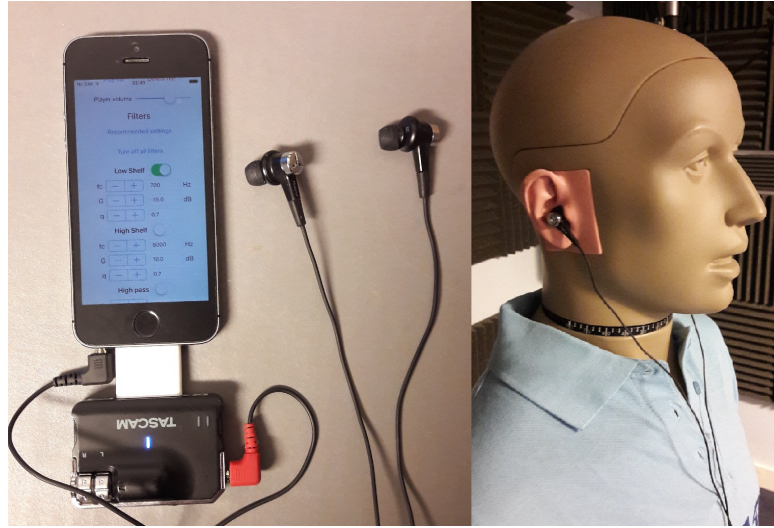


Figure 3.2: AAR system (left) and KEMAR (right).

The equaliser in the iOS app was adjusted according to the results (the filter parameters were adjusted by the authors after visual inspection of the frequency response) and this configuration was used in the adaptation experiments. The same filter was used for left and right channel. Frequency response curves for unfiltered and equalised hear-through functions are shown in Figure 3.3.

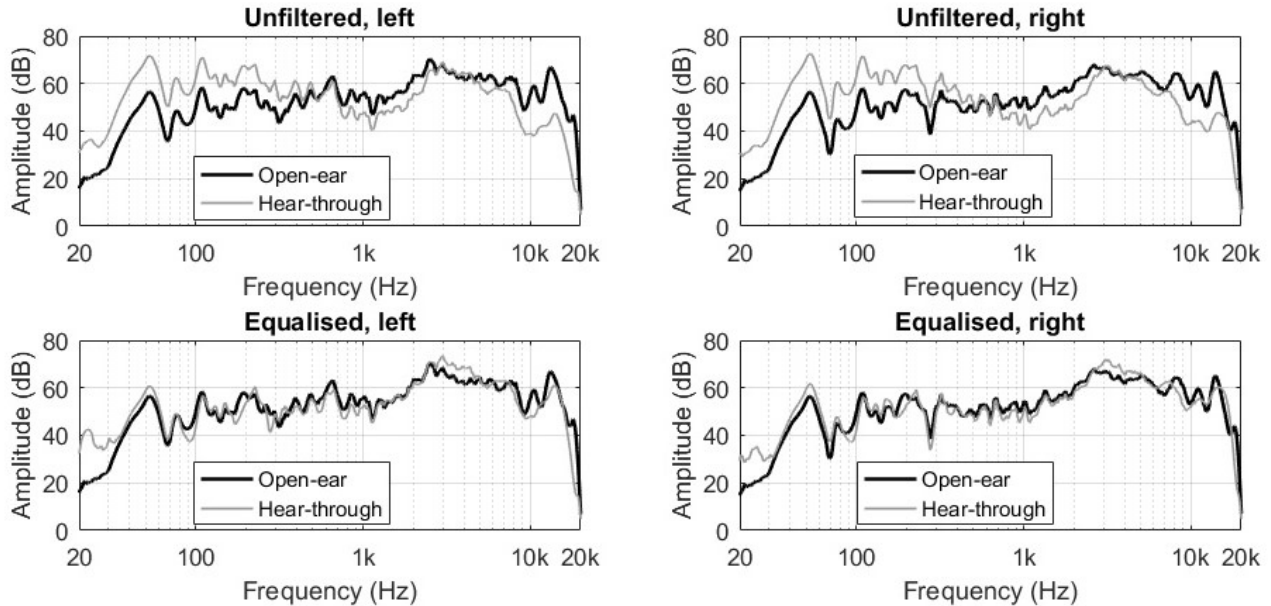


Figure 3.3: Comparison of open-ear frequency response against unfiltered and equalised hear-through, where 0 dB represents the lowest amplitude value registered across all measurements.

3.2.3 Latency and comb filter effect

Total in-out latency of the system is defined as the time lapse between the arrival of sound leaked through the earphone and its hear-through counterpart. This time lapse was measured to be approximately 10 ms. A high latency of that size could cause a phenomenon known as comb filter effect, which would add a ‘metallic echo’ to the perceived sound, potentially deteriorating the hearing experience (Härmä et al., 2004; Karjalainen et al., 2008; Rämö and Välimäki, 2012). Such issue can be difficult to eliminate through additional filtering, but it can be mitigated by reducing latency and/or increasing earphone attenuation (Rämö and Välimäki, 2012). For the Roland CS-10EM, attenuation was measured to be in the range of 5–20 dB for frequencies below 1 kHz, and 20–35 dB for the rest of the spectrum, meaning that low frequencies were the most affected. Latency was minimised in the iOS app, albeit it being constrained by the phone’s analog/digital converters. In practice, it was found that comb filter effect was hard to eliminate completely but it could become barely noticeable when the earphones were well fitted inside the ear, achieving a higher attenuation and therefore less leakage. Ultimately, the subjects did not find it to be a problem during the field test. Potential phase effects introduced by the filters were not found to be noticeable in informal listening tests, which confirms the findings of Lindau and Weinzierl (2012), who found that listeners were not able to distinguish between minimum-phase headphone equalisation filters and unconstrained-phase ones in a perceptual evaluation.

3.3 Methodology

Two case studies were performed to test the hypothesis that it is possible for a person to adapt to hearing comfortably through an AAR system. Two male subjects with no reported hearing impairments, both with previous experience on binaural audio and one of them being the author of this study wore the device for three days, between seven and eight hours per day, performing four daily control sessions: two in the morning, before and after putting the earphones in, and two in the evening, before and after removing them (see Table 3.1). The following control tests

were used (both will be further discussed in later subsections):

1. A localisation test, to measure the subject’s accuracy to localise sounds in the surroundings.
2. A subjective evaluation, to estimate changes in the subject’s perception of pre-recorded auditory scenes.

Table 3.1: Test schedule. ‘Loc.’ stands for ‘Localisation test’, ‘Subj.’ stands for ‘Subjective evaluation’, ‘Open’ stands for ‘Open-ear’, and ‘HT’ stands for ‘Hear-Through’.

Day	Time		Test ID
1	Morning	Before inserting the earphones (baseline)	Open1 (1 Loc.+1 Subj.)
		After inserting the earphones (null adaptation)	HT1 (1 Loc.+1 Subj.)
	Evening	Before removing the earphones	HT2 (1 Loc.+1 Subj.)
		After removing the earphones	Open2 (1 Loc.+1 Subj.)
2	Morning	Before inserting the earphones	Open3 (1 Loc.+1 Subj.)
		After inserting the earphones	HT3 (1 Loc.+1 Subj.)
	Evening	Before removing the earphones	HT4 (1 Loc.+1 Subj.)
		After removing the earphones	Open4 (1 Loc.+1 Subj.)
3	Morning	Before inserting the earphones	Open5 (1 Loc.+1 Subj.)
		After inserting the earphones	HT5 (1 Loc.+1 Subj.)
	Evening	Before removing the earphones (max. adaptation)	HT6 (1 Loc.+1 Subj.)
		After removing the earphones	Open6 (1 Loc.+1 Subj.)
Total		Open-ear	6 Loc. + 6 Subj.
		Hear-through	6 Loc. + 6 Subj.

Considering the hypothesis, the subjects were expected to adapt to their modified HRTF (Blauert, 1997; Wanrooij and Opstal, 2005; Mendonça et al., 2012), so a progressive evolution should be observed along the hear-through sessions, being the first one (HT1) the farthest away from open-ear results, and the last one (HT6) being the closest. On the other hand, all open-ear sessions were expected to obtain similar results, although it was possible that some degradation occurred due to adaptation to the AAR device.

3.3.1 Equipment

A surround sound setup of eight Equator Audio D5 speakers was used, positioned as shown in Figure 3.4 (positions 1–8) in an acoustically dampened room of approximate dimensions of 4x4x4m and operated by a computer through a Focusrite RedNet 2 audio interface (High

Wycombe, United Kingdom). The subjects interacted with the system through a user interface made with Max 7 (Cycling '74, Walnut, CA, United States).

3.3.2 Localisation test

This test evaluated how accurately could the subject pinpoint the location of surrounding sound sources. A total of twenty sources were defined around the subject (see Figure 3.4). Sources 1–8 were the actual speakers, whereas 9–20 were virtual sources, generated by panning pairs of speakers, e.g. a sound played with the same loudness from speakers 1 and 2 appears to come from source 9.

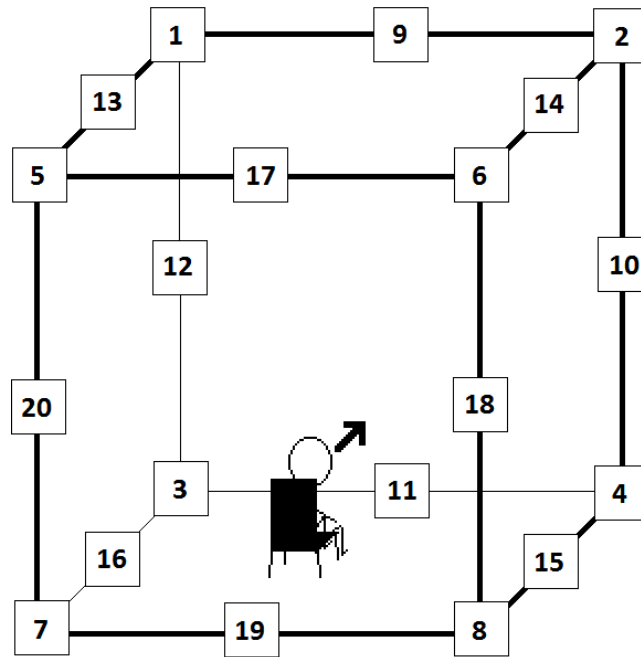


Figure 3.4: Distribution of sound sources around a subject sitting in the middle of the room. Sources 1–8 are actual speakers, whereas 9–20 are virtual sources generated by panning pairs of speakers.

The subject sat with his head at the same distance of the eight speakers and listened to a sequence of 40 sound samples. The possible sound source positions were numbered from 1 to 20 and were visible to the subject. The goal was to guess which source was used each time by selecting the corresponding number from a visual user interface. Sound samples were a combination of white noise and speech, and one second long. The test was divided in two parts:

1. Part one: head movement was not allowed. All 20 sources were used once in a random order.
2. Part two: head movement was permitted to improve accuracy. Again, all 20 sources were used once, but in a different random order than before.

Results were analysed in terms of ratio of wrong answers, front-back confusions and up-down confusions. Both open-ear and hear-through scenarios were compared.

3.3.3 Subjective evaluation

The aim of this test was to evaluate the evolution of the subject's spatial hearing perception when presented different auditory scenes during the adaptation process. The scenes were recorded with a first-order Ambisonics microphone (Oktava MK-4012, Tula, Russia) and decoded to the eight-speaker surround system. Two scenarios were used: a quiet office and a busy street in London. A total of 10 audio fragments (see Table 3.2), all with a length of 30s, were extracted from the recordings. In each test, all fragments were presented in a random order and the subject was asked to rate using a scale between 1 and 5 for each of the following attributes (inspired by the list proposed by Simon et al. (2016), used to evaluate the quality of binaural recordings):

1. *brightness*: 'how bright the sounds feel, overall' (1 for very dark and 5 for very bright),
2. *externalisation*: 'perception of sounds located outside your head' (1 for inside the head and 5 for outside the head),
3. *immersion*: 'feeling of yourself being located in the middle of the audio scene' (1 for not immersive at all and 5 for very immersive),
4. *realism*: 'feeling of sounds coming from real sources located around yourself' (1 for very unrealistic and 5 for very realistic), and

5. *relief*: ‘feeling of distance between the closest sound objects and the farthest’, (1 for very compact and 5 for very spread out).

Table 3.2: Characteristics of the audio fragments.

Audio fragments	Characteristics and location of sounds with respect to the microphone
office1–2	Two people talking (left), plastic bag noise (right), pen writing on paper (right).
office3–6	Two people talking (left), laptop playing classical music (back-right), plastic bag noise (right), mobile phone ringing (right), door slamming (front-right).
street1–4	Three people talking (back) vehicles passing by (all directions).

3.4 Results and discussion

3.4.1 Localisation test

Figure 3.5 shows the evolution of the localisation test results along the 6 sessions, in terms of ratio of wrong answers, ratio of front-back confusions, and ratio of up-down confusions, for both open-ear and hear-through situations.

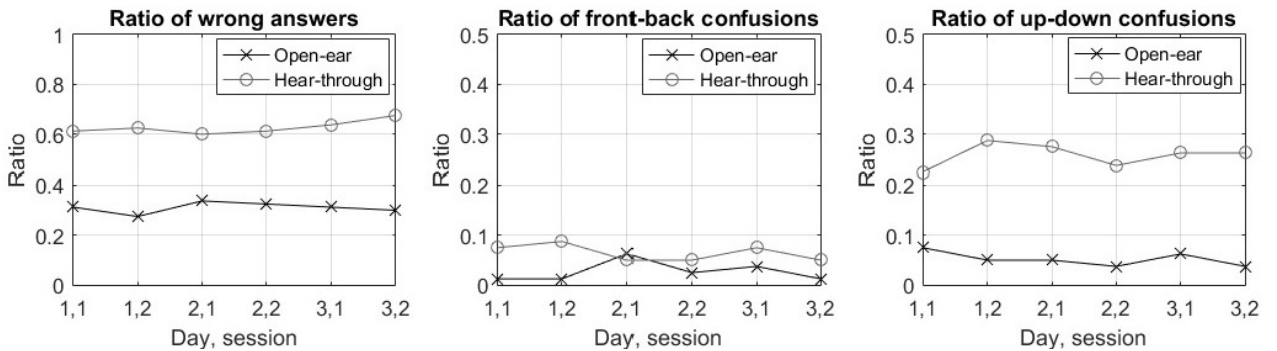


Figure 3.5: Results of the localisation test.

The hear-through condition showed a 30% increase in localisation error with respect to open-ear, meaning an overall poorer localisation performance. This result is in line with the findings of Marentakis and Liepins (2014). Up-down confusions were observed to be a major cause of errors when wearing the device, showing a 20% higher ratio than in the open-ear condition. In the case of front-back confusions, they were also found more frequent in hear-through than in open-ear, but the difference was less prominent (less than 8%). In general, the increase of

elevation errors and number of front-back confusions in the hear-through condition may have been due to the position of the microphones being slightly off the entrance of the ear canal, altering the subject's natural localisation cues. On the other hand, azimuth errors were found minimal, arguably because the device did not cause a modification of the ITDs and ILDs, which are dominant in lateral localisation.

Neither degradation or improvement were observed for open-ear and hear-through cases over the six sessions. The curves for both hearing conditions do not show a clear tendency to converge, therefore no evident signs of adaptation were found. However, further testing with more subjects and a longer process of adaptation could lead to more conclusive results. Once more data is collected in future studies, these results should be confirmed by means of appropriate inferential analysis, which here was not conducted due to the lack of enough available data.

3.4.2 Subjective evaluation

Figure 3.6 shows a summary of the results of the subjective evaluation, displayed as the evolution of the average score for each attribute in both the office scene (calculated as the mean of the scores for fragments office1–6), and the street scene (mean of the scores for fragments street1–4).

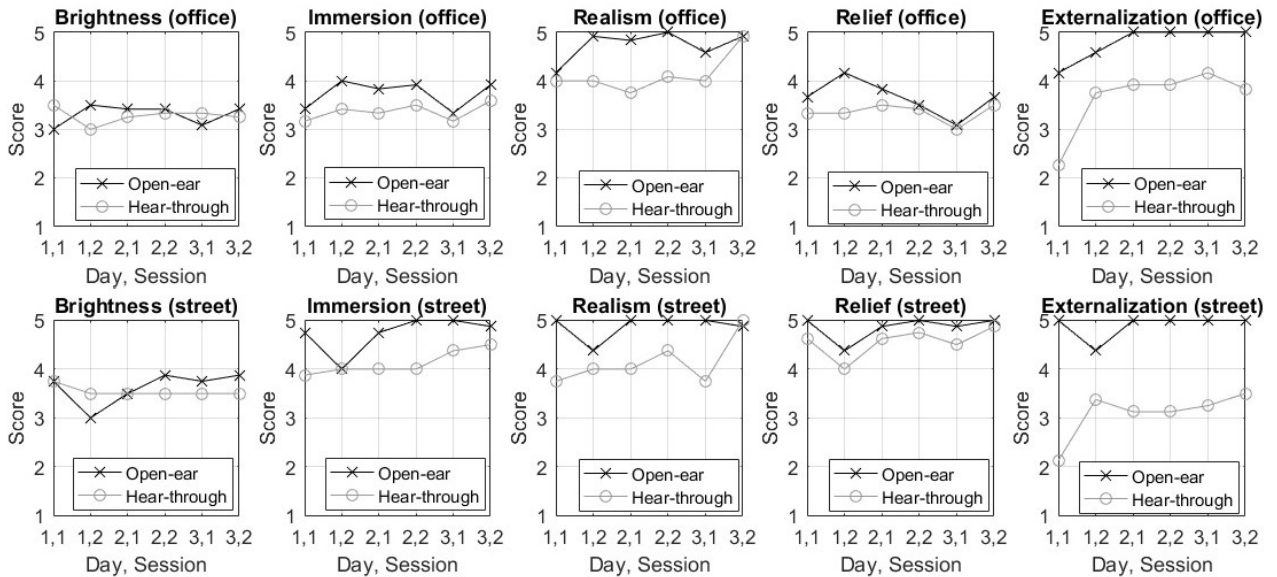


Figure 3.6: Results of the subjective evaluation. Average results of fragments office1–6 are displayed in the top row, and those of fragments street1–4 are shown in bottom row.

It was observed that wearing the AAR device had an immediate effect in the perception of all attributes except for brightness, which showed little variation between open-ear and hear-through. The most notable difference between the two hearing conditions was found in externalisation, with significantly lower scores in the hear-through case. In fact, subjects seemed to identify this attribute as the defining factor for the overall quality of the ‘hearing experience’, meaning that better externalisation often translated to more naturalness.

Adaptation was observed to happen to some extent in all attributes (again except brightness), showing signs of convergence in the hear-through and open-ear curves, particularly in the case of realism, which obtained equal scores for both hearing conditions in the last session. The most notable evolution happened in externalisation between the first and second sessions, for both ‘street’ and ‘office’ scenes. No evident signs of degradation over time were observed in the open-ear results, meaning that the subjects did not ‘reject’ their natural hearing condition during the adaptation process. This should be confirmed by more extensive tests with a higher number of subjects and proper inferential analysis.

3.5 Conclusions

The outcomes of this pilot study suggested that slight effects of adaptation, especially in subjective attributes (e.g. externalisation), might take place after a prolonged usage of the hear-through functionality of the proposed acoustically transparent headphones. Regardless, it is important to emphasise that this was indeed a preliminary exploratory study with a very limited number of subjects and its results should not be strictly interpreted as scientific findings, given that their reproducibility and reliability might be limited. In fact, further experiments would be needed to obtain more conclusive results. For future work, increasing the number of subjects and the duration of the field tests, in order to maximise the effects of long-term adaptation, is suggested, as well as conducting appropriate inferential analysis of the results tests. Furthermore, it would be relevant to compare the proposed system in terms of usability for AAR applications to other hardware solutions such as open-fitting headphones, bone conduction,

and receiver-in-the-canal devices.

In addition to these findings, the pilot study provided important insights into the challenges of AAR which are discussed in this Thesis. For instance, informal listening tests performed by the authors suggested that the EQ process could have an essential role in the perceived quality not only of the hear-through audio, but also in any simulated binaural content, which was not covered in the present chapter. After the pilot study, it was decided to shift the research direction away from hardware usability to focus on digital signal processing techniques for spatial audio rendering in AAR. Thus, the next study explored in more depth the topic of headphone EQ and its effect on binaural audio, but this time on a different kind of AAR hardware: open-ear headphones.

Chapter 4

The Effect of Generic Headphone Compensation on Binaural Renderings

After the initial exploration of the challenges of implementing an AAR system, discussed in Chapter 3, it was observed that headphone EQ could play a crucial role in the perceived quality of the binaural audio presented to the listener, not only in a hear-through scenario but also for auralisations in general. This chapter covers a study that looked at the perceptual effect of headphone compensation techniques when applied to binaural renderings.

This chapter was adapted from the article by the present author and others (Engel, Alon, Robinson, et al., [2019](#)), which was peer-reviewed as a complete manuscript and presented at the Audio Engineering Society Conference on Immersive and Interactive Audio in York, United Kingdom, in 2019. This was done with permission of the rights holder, the Audio Engineering Society (United States); the permission letter can be found in the appendix. The work described in this chapter was done as part of an internship at Facebook Reality Labs in Redmond, WA, United States.

Abstract

Binaural rendering allows us to reproduce auditory scenes through headphones while preserving spatial cues. The best results are achieved if the headphone effect is compensated with

an individualised filter, which depends on the headphone transfer function, ear morphology and fitting. However, due to the inconvenience of remeasuring a new filter every time the user repositions the headphone, generic compensation may be of interest. In this study, the effects of generic headphone equalisation in binaural rendering are evaluated objectively and subjectively, with respect to unequalised and individually-equalised cases. Results show that generic headphone equalisation yields perceptual benefits similar to individual equalisation for non-individual binaural renderings, and it increases overall quality, reduces colouration, and improves distance perception compared to unequalised renderings for the particular case of open-ear headphones.

4.1 Introduction

Binaural rendering allows us to reproduce auditory scenes through headphones while preserving all their spatial cues, by using measured or simulated binaural impulse responses (Wightman and Kistler, 1989a; Møller, 1992; Begault and Trejo, 2000), and is therefore widely used for synthesising 3D sound in virtual and augmented reality applications. For instance, for a static listener and a single sound source in a room, the transfer function between the source and the listener's ears can be measured as a pair of filters (left and right), which are referred to as the BRIR as defined by Møller (1992).

If a dry audio signal is convolved with a BRIR and presented through headphones, the listener should get the sensation that a real source is producing the sound at the corresponding location and with all the room acoustics preserved. However, for the synthesised signal to be indistinguishable from the real sound field, it should not be altered by the headphones, which have a non-flat frequency response. Therefore, it is necessary to use an EQ filter to compensate the effect of the HpTF. Essentially, the transfer function of the filter should be the inverse of the HpTF, so that when playing the equalised signal through the headphones, both transfer functions cancel out and the listener receives an unaltered version of the rendered binaural audio (Pralong and Carlile, 1996). In this process, magnitude response has been found to be more relevant than phase response, e.g. Lindau and Brinkmann (2012) showed that listeners could

distinguish between headphone filters with unconstrained phase and filters with minimum phase in a perceptual test. Therefore, in this work the EQ filters aim to compensate the magnitude of the HpTF first and foremost, rather than the phase.

It is important to note that the HpTF compensation approach differs from other EQ methods recommended in the literature. A well known example is the Harman target curve, which was designed to simulate the response of a stereo loudspeaker system in a reverberant room, based on the assumption that music recordings are often optimised for such a setup (Olive, Welti, et al., 2013). In the case of binaural rendering, however, the room response is already included in the BRIR and therefore should not be taken into account when designing the headphone EQ curve.

The BRIR and headphone transfer function (HpTF) are highly dependent on the morphology of the ear (Møller, Hammershøi, et al., 1995); therefore, the highest degree of authenticity is only achieved when an individualised pair of BRIR and HpEQ filter are used (Pralong and Carlile, 1996). Previous research has shown that when individualised filters are used static listeners could not distinguish between a real and a rendered audio source in a discrimination task (Langendijk and Bronkhorst, 1999). Other studies have claimed that discrimination rates are higher (a) for broadband noise stimuli than for speech or music, (b) if listeners are given unlimited listening time, or (c) if head movements are allowed (Oberem et al., 2016; Brinkmann, Lindau, and Weinzierl, 2017).

However, if a non-individualised BRIR is used, it is no longer obvious which kind of HpEQ optimises the quality of the binaural simulation. Lindau and Brinkmann (2012) claim that the best practice is to use an HpEQ filter measured on the same head as the BRIR; the second-best choice would be an individualised HpEQ, and the least preferable option would be to use a non-individualised HpEQ measured on a different subject than the BRIR was. Nevertheless, in the perceptual evaluation, they used a single universal attribute to measure similarity between simulated and real audio, and it is therefore unclear whether listeners were paying more attention to the timbral characteristics of the audio content or to its spatial features. Also, in the ABC/HR type of comparison which was used, tested conditions were compared to the real loudspeaker

but not to each other, so small perceptual differences between HpEQ types could have been lost.

In this study, the effects of individual and generic headphone compensation on the quality of a binaural rendering were further investigated, both objectively and perceptually. The contributions can be summarised as follows:

1. Instead of a single global rating, several features (overall similarity, colouration, distance, direction) were assessed separately, to better understand how all the binaural audio content characteristics are affected by the HpEQ type.
2. Evaluations were performed in both ideal and non-ideal binaural rendering scenarios (individualised and non-individualised BRIR, respectively).
3. A multiple stimulus test with hidden reference and anchor (MUSHRA), which is robust for measuring small and intermediate differences (ITU-R, [2015a](#)), was used. Furthermore, this test allowed for direct comparison between the different test conditions (individual EQ, generic EQ and unequalised).

4.2 Methods

A total of 12 subjects (ages 26–55, 2 female) took part in this study. All reported normal hearing and had previous experience with listening tests. Individual HpEQ filters and BRIRs were measured for all subjects with a pair of open headphones, as described in the following Subsections. Several test conditions were defined combining different types (generic/individual) of HpEQ and BRIR, which were analysed objectively and subjectively in a perceptual experiment.

4.2.1 Hardware setup

Custom so-called floating headphones were built, similar to the ones used by Langendijk and Bronkhorst ([1999](#)) and by Romigh et al. ([2015](#)), consisting of a pair of Sennheiser MX475 earbuds (Wedemark, Germany) attached to a headband (see Figure 4.1). The purpose of this design

was to leave the listener’s ear canal unoccluded (therefore, avoiding changing its impedance) while minimising the effect of the hardware on the HRTF (Langendijk and Bronkhorst, 1999). Another reason for choosing these headphones was that they have similar frequency response characteristics to air-conducted built-in speakers, which are prevalent in current virtual and augmented reality headsets such as the Oculus Go™, Microsoft HoloLens™, and Bose AR™. This consideration seemed relevant given that virtual and augmented reality are a common application area for binaural rendering.



Figure 4.1: Custom floating headphones (Sennheiser MX475) and binaural microphones (Brüel and Kjær 4101-B) mounted on KEMAR head and torso simulator.

Subjects were seated in the centre of a reverberant room ($RT30_{[400\text{ Hz}-1250\text{ Hz}]} = 244\text{ ms}$), wearing a pair of Brüel and Kjær 4101-B in-ear microphones (Nærum, Denmark). A Genelec 8020 loudspeaker (Iisalmi, Finland) was employed as the sound source and placed at 45° azimuth, 0° elevation, and a distance of 2 m from the subject’s head. An adjustable chair and a laser alignment system were used to make sure that the head was at the right position and orientation during the measurements.

4.2.2 Headphone EQ approach

In order to compensate for the HpTF, HpEQ filters were calculated using frequency-dependent regularisation (Kirkeby and P. A. Nelson, 1999), which has been shown to perform better

than other methods in perceptual tests (Schärer and Lindau, 2009). In general, the goal of regularisation is to avoid the excessive boost of certain frequencies that happens if the direct inverse of the transfer function is used as a compensation filter, thus preventing distortion and sensitivity to measurement errors (Kirkeby and P. A. Nelson, 1999). In the particular case of headphone compensation, regularisation prevents the inversion of narrow notches at high frequencies, which could lead to ringing artefacts if the headphones are repositioned after the measurement (Schärer and Lindau, 2009). For this reason, a frequency-dependent regularisation parameter must be set to a higher value at frequencies where those notches are present. While this parameter has traditionally been adjusted by expert listeners (Schärer and Lindau, 2009), Bolaños et al. (2016) have proposed a procedure by which to calculate it, demonstrating positive objective and perceptual results. In this study, the latter approach is used, calculating the regularised inverse $H_R^{-1}(\omega)$ of a headphone response $H(\omega)$ as

$$H_R^{-1}(\omega) = \frac{H^*(\omega)}{|H(\omega)|^2 + [\alpha(\omega) + \sigma^2(\omega)]} D(\omega) \quad (4.1)$$

where $D(\omega)$ is a modelling delay to ensure that the filter is causal, $\alpha(\omega)$ is the parameter which defines the bandwidth and maximum amplification of the filter, and $\sigma^2(\omega)$ is an estimator of the amount of regularisation needed within the inversion bandwidth. We define

$$\alpha(\omega) = \alpha_0 + \frac{1}{|W(\omega)|^2} - 1 \quad (4.2)$$

$$\sigma(\omega) = \begin{cases} |\hat{H}(\omega)| - |H(\omega)| & \text{if } |\hat{H}(\omega)| \geq |H(\omega)| \\ 0 & \text{if } |\hat{H}(\omega)| < |H(\omega)| \end{cases} \quad (4.3)$$

$\alpha(\omega)$ is calculated from a unity-gain passband filter $W(\omega)$, which delimits the bandwidth within which the headphones are equalised. α_0 is a scalar that limits the amount of amplification allowed by the filter (zero means no limit). Note that $\sigma(\omega)$ was defined as the negative deviation of the headphone response $H(\omega)$ from a smoothed version $\hat{H}(\omega)$, which will be larger in zones with

narrow notches and therefore will provide an estimation of the required amount of regularisation required, as stated above (Bolaños et al., 2016). In this study, the same method proposed by Bolaños et al. was implemented, using the parameters outlined in the next Subsection.

As seen in Figure 4.2, the HpEQ filter (regularised inverse) is similar to the direct inverse of the HpTF, except that amplification is reduced outside the defined headphone bandwidth (200–20000 Hz in this case) and in zones with narrow notches; this is particularly noticeable around 11 and 17 kHz.

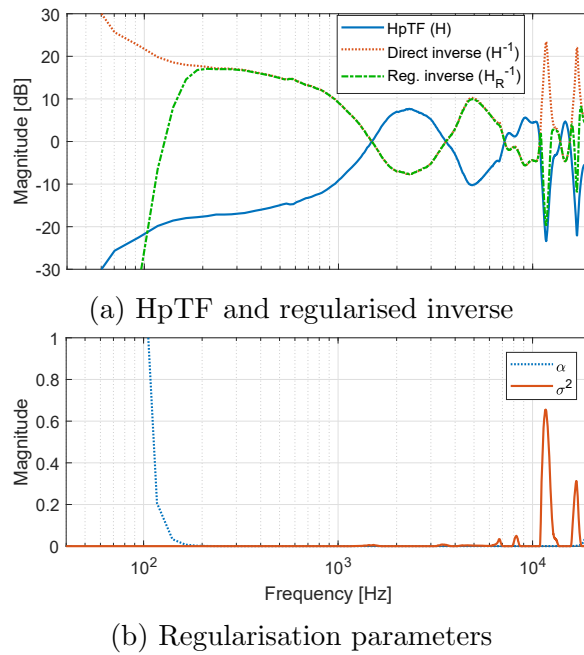


Figure 4.2: Example of HpEQ filter calculation. (a) Headphone transfer function (H), its direct inverse (H^{-1}) and regularised inverse (H_R^{-1}); (b) Regularisation parameters α and σ^2 for that HpTF.

4.2.3 HpEQ and BRIR measurements

Measurements were performed with the sine sweep technique (Farina, 2007), using a sweep length of 2 s for the HpTF and 8 s for the BRIR, with a frequency range from 10 to 24000 Hz. HpEQ filters (H_R^{-1}) were calculated from the HpTF (H) through Equation 4.1, Equation 4.2, and Equation 4.3 while using the following parameters:

1. $W(\omega)$: fifth-order Butterworth bandpass filter (200–20000 Hz), according to the headphone

bandwidth. The phase of this filter is not expected to have a effect on the resulting EQ, given that only its magnitude response is employed in Equation 4.2.

2. α_0 : $2.5 \cdot 10^{-4}$, which limits the amplification to 30 dB within the inversion range.
3. $\hat{H}(\omega)$: smoothing window of one equivalent rectangular bandwidth (ERB), following (Kohlrausch and Breebaart, 2001), which gave good results in preliminary tests.

The HpTF and BRIR were measured for each subject and for a KEMAR head and torso simulator, which was used for the ‘generic’ conditions. Measurements on subjects were done immediately before performing the test, and headphones were not repositioned or removed until the experiment was finished. An overview of the results is detailed in Section 4.3 and shown in Figure 4.4.

4.2.4 Audio material

In order to evaluate the effect of HpEQ for various scenarios, three different dry audio materials were used in the listening test:

1. *Speech*: anechoic male speech recording from the Archimedes project (Hansen and Munch, 1991).
2. *Pink noise*: a sequence of four broadband noise bursts of length 750 ms with 500 ms of silence between them. Each noise pulse was faded in and out with a 50 ms raised cosine window.
3. *Guitar*: anechoic guitar recording, also from the Archimedes project (Hansen and Munch, 1991).

All signals were of length 5 s and were faded in and out with 50 ms raised cosine windows. To ensure that all spectral content could be played at a reasonable level (60 dB A-weighted) without distortion, a band-pass filter between 500 and 16000 Hz was applied in order to match the reproduction bandwidth of the floating headphones (see Figure 4.3).

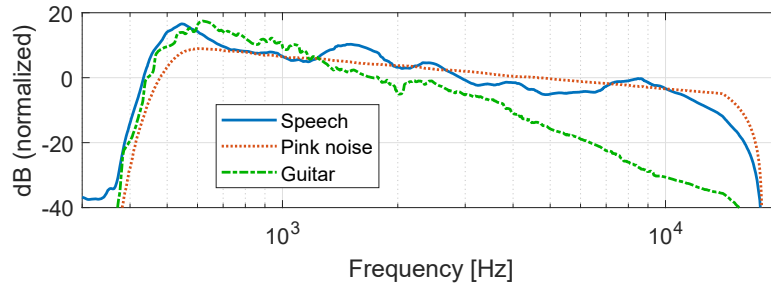


Figure 4.3: Spectra of the audio materials. Curves were normalised and smoothed with a third-octave window.

4.2.5 HpEQ and BRIR test conditions

In this study, we wanted to evaluate the effect of HpEQ for the ideal case where the individualised BRIR is available, as well as for the case where a generic BRIR is used. Therefore, two independent variables were tested:

1. *BRIR type*: individualised binaural room impulse response (IndBRIR) / generic binaural room impulse response (GenBRIR).
2. *HpEQ type*: individualised headphone equalisation (IndHpEQ) / generic headphone equalisation (GenHpEQ) / no headphone equalisation (NoHpEQ).

This makes a total of six test conditions, from now on referred to as xxxBRIR+yyyHpEQ, e.g. IndBRIR+GenHpEQ.

Although the study initially included the real loudspeaker as a reference condition, it was finally removed for two reasons: (1) for a fair comparison of the source direction between the loudspeaker and rendering, the listener's head needed to be completely static, i.e. with the help of a chin rest, during the whole duration of the experiment, which was found to be sufficiently uncomfortable so as to introduce a bias in the responses due to listener fatigue, and (2) IndBRIR+IndHpEQ was considered a suitable reference condition, as it was found to be almost indistinguishable from the real loudspeaker in informal listening tests performed by the authors, which is in agreement with the findings from Langendijk and Bronkhorst (1999).

4.3 Objective evaluation

Spectra of the three dry audio materials are shown in Figure 4.3. It can be seen that *pink noise* and *speech* offer the widest content in terms of frequency range, and therefore may be more effective in revealing colouration changes across different test conditions. The *guitar* stimulus, on the other hand, might give the listeners less spectral information, which could make the perceptual evaluation more challenging.

Figure 4.4 shows the statistics (10th and 90th percentiles and median) of HpEQ filters (Figure 4.4a) and the BRIRs in the frequency domain (Figure 4.4b), based on measurements from the 12 subjects. In addition, the generic HpEQ filter and BRIRs measured on KEMAR are also shown. In Figure 4.4a, it can be observed how curves are very similar across subjects for frequencies below 2 kHz, and start to diverge above that point, which indicates that individual features tend to appear at higher frequencies, as found by Pralong and Carlile (1996). The overall shape of the HpEQ curves hints at the limitations of the custom floating headphones when unequalised, given the > 20 dB difference in gain between the low (300 Hz) and mid (2 kHz) frequencies.

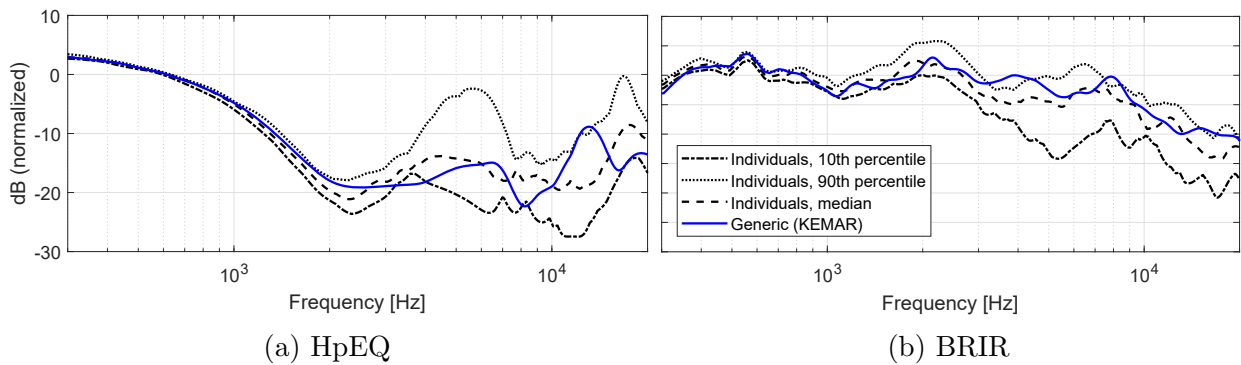


Figure 4.4: Statistics of (a) calculated HpEQ filters and (b) BRIR magnitude spectra, for the left ear. Median, 10th and 90th percentiles across subjects are indicated with dashed lines; generic is indicated with a solid line. Curves were normalised and smoothed with a third-octave window.

In Figure 4.4b, similarly to the HpEQ plot, curves seem to converge in the lower part of the spectrum, and diverge for higher frequencies. This is due to the nature of the HRTFs, which have small variability at low frequencies and higher variability at high frequencies. It can be seen that the generic BRIR differs from the individual measurements by up to 5 dB for frequencies

as low as 1.5 kHz, and by up to 15 dB for higher frequencies.

To get a better understanding of the impact of the HpEQ and BRIR on the in-ear pressure level, and perhaps some insights on the perceived differences between the tested conditions, it may be interesting to analyse their effect on monaural and binaural cues. To that end, the complete binaural response, i.e. taking into account both the BRIR and the HpEQ filter, was computed for each test condition, across all subjects. Taking the IndBRIR+IndHpEQ response as a reference, an error metric can be calculated as the spectral difference between the log magnitude response of the reference and the log magnitude response of each of the tested conditions.

Figure 4.5a shows the absolute spectral difference for (a) the complete binaural response and (b) the ILD (Xie, 2013, Equation 1.12). Each curve was constructed by taking the median spectral difference across all subjects for each frequency bin, and smoothing the result with a third-octave window. As anticipated, NoHpEQ conditions showed the largest deviations from the reference, with errors of the order of 20 dB, particularly above 2 kHz. Taking into account the variability of the results across individuals, and the error introduced by the generic BRIR, the differences between IndHpEQ and GenHpEQ are relatively small. The performance of each tested condition on the subjective evaluation may be predicted by observing these results, e.g. large errors may be perceived by listeners as severe colouration changes.

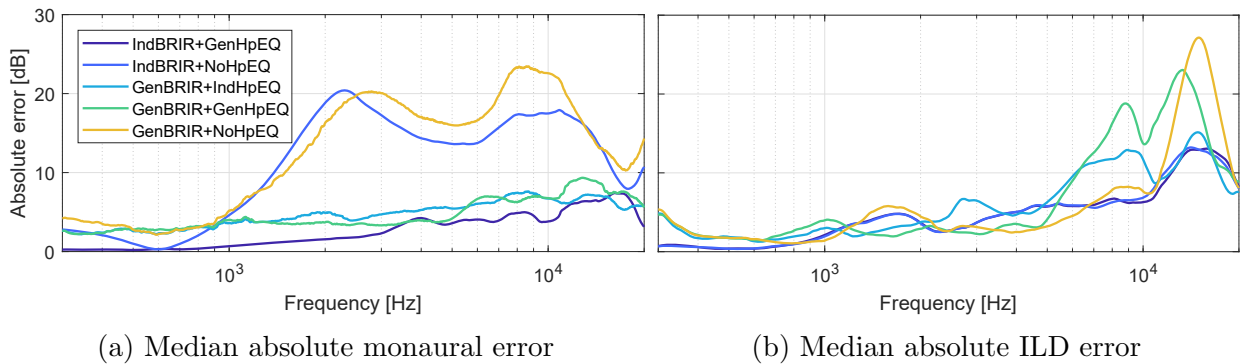


Figure 4.5: Error curves for all test conditions with respect to the reference condition: (a) absolute monaural error; (b) absolute ILD error. Curves were calculated by convolving the BRIR and HpEQ filter, both dependent on the test condition, then subtracting the reference curve (IndBRIR+IndHpEQ), and finally taking the median value across subjects for each frequency bin (final result smoothed with a third-octave window).

A similar spectral difference error metric can be calculated for the ILD, by subtracting the ILD of each tested condition from the reference IndBRIR+IndHpEQ. Given that filters were

independently designed for left and right ears, it was fair to assume that ILDs changed from one condition to another. Figure 4.5b shows the absolute ILD error of each condition, taking IndBRIR+IndHpEQ as a reference. A higher error was observed above 6 kHz for GenBRIR than for IndBRIR conditions, with GenBRIR+GenHpEQ and GenBRIR+NoHpEQ having the largest error (above 11 kHz). This result can be explained by the HRTF variability across different subjects, which is generally dominant at higher frequencies.

Interaural time difference (ITD) analysis was considered as well, following the recommendations by Katz and Noisternig (2014). However, impulse response onset detection was found to be problematic on the contra-lateral ear, perhaps due to the non-anechoic conditions of the measurements not providing a high enough signal-to-noise ratio. It was therefore decided to leave ITD analysis for future follow-up studies, where other methods such as cross-correlation analysis may be used.

4.4 Subjective evaluation

4.4.1 Method

A MUSHRA test, as defined in by ITU-R (2015a), was used to perform the perceptual evaluation. The MUSHRA paradigm was chosen because it allows the listener to compare all the test conditions to one another, in addition to the reference, which makes it easier to detect small and intermediate differences between them.

Three different dry audio signals and four perceptual attributes were assessed, making a total of 12 trials per participant. Each subject performed the experiment in a single session, which lasted approximately 45 minutes and included around 5 minutes for training and 40 minutes for evaluation and ranking. All audio signals were presented from the position of the loudspeaker, as described in Subsection 4.2.1.

Assessed attributes

A list of relevant perceptual attributes had to be chosen so that subjects could be asked to systematically compare across the different listening conditions (IndHpEQ, GenHpEQ, NoHpEQ). To keep the length of the experiment reasonable, four attributes were chosen as the most important ones based on informal listening tests performed by the authors, which are the following:

1. *overall similarity (OVS)*: any and all detected differences between the reference and the tested signal,
2. *colouration (COL)*: any and all detected differences in timbral impression and tonalness between the reference and the tested signal,
3. *distance (DIS)*: whether the tested signal is perceived at the same distance as the reference, regardless of the direction of incidence, and
4. *direction (DIR)*: whether the tested signal's angle of incidence (azimuth/elevation) is the same as for the reference.

This attribute choice is similar to the findings of Brinkmann, Lindau, and Weinzierl (2017), who found that the most relevant attributes when rating binaural renderings were *difference*, *high frequency colour*, *brightness*, *pitch*, and *distance*. It was also similar to the recommendation by the ITU-R (2015a), which suggests using *basic audio quality*, *timbral quality*, *localisation quality* and *environment quality*. Participants were provided an informative sheet with attribute definitions and examples, extracted from the Spatial Audio Quality Inventory (SAQI) by Lindau, Erbes, et al. (2014).

Training stage

Participants performed a training phase, where they were exposed to all the signals which they would later experience during the test. A screen was presented, where all the combinations of

audio material and test conditions were available as buttons, and could be played as many times as needed by clicking them. Buttons were unlabeled and their order was randomised to avoid introducing bias. Participants were encouraged to spend as long as they needed to familiarise themselves with the test material.

Evaluation stage

After the training, participants proceeded to the evaluation stage. This consisted of 12 trials, one for each combination of audio material and attribute. The evaluation was divided into four blocks, one for each attribute. The first block was always OVS, whereas the other three were presented in a random order. Within each block, the order of the trials (one per audio material) was also randomised. In each trial, subjects were presented an audio material and could switch at will between all the test conditions (BRIR/HpEQ types). A reference was provided, which was always the IndBRIR+IndHpEQ condition, and subjects were asked to rate the similarity of each condition against the said reference, according to the current attribute (OVS/COL/DIS/DIR), with a score from 0 to 100. A rating of 100 would mean that the signal and the reference were identical for that particular attribute.

Subjects were informed that one of the signals was a hidden reference, and therefore they were required to give at least one rating of 100. All signals could be played as many times as needed. It was possible to switch between signals during the playback; a 2 ms raised cosine fade in/out with no crossfade was applied in the transition between audio signals. A practice trial, which did not count towards the results, was presented at the beginning of the evaluation stage so the subjects could familiarise themselves with the rating procedure.

Although in MUSHRA tests it is recommended to use at least one low-passed version of the reference signal as a low-quality anchor (ITU-R, [2015a](#)), in the proposed experiment it was not trivial to define a proper anchor that worked similarly for attributes such as DIR or DIS. For instance, a low-passed version of the reference (IndBRIR+IndHpEQ) could be a valid anchor for the attribute COL, but not for attributes related to spatial perception, such as DIR or DIS given that the binaural and monaural cues of the individualised BRIR would be still preserved and

the perceived position of the sound source would remain unaffected. Similarly, it was considered to employ a diotic presentation as a DIS anchor and an ILD-shifted version of the stimuli as a DIR anchor, respectively. However, it was ultimately found that all these potential anchors led to undesirable compression of the data points during the perceptual evaluation. Therefore, it was decided not to use any anchor in this study.

4.4.2 Results

The subjective evaluation was conducted for a total of 10 subjects, which was estimated to be a high enough quantity to produce statistically relevant data for the purposes of this study, and is in line with previous similar work such as by Brinkmann, Lindau, and Weinzierl (2017). Figure 4.6 provides an overview of the subjective evaluation results across all subjects, divided by test condition and audio material. Two individuals were excluded in post-screening because they failed to rate the reference above 90 points for more than 15% of the trials (ITU-R, 2015a). Interquartile ranges are indicated with boxes, medians with circles, and the most extreme data points with thin lines. BRIR and audio material are denoted by background and line colours, respectively. A quick inspection reveals that IndBRIR+IndHpEQ was correctly identified as the hidden reference condition by most listeners, as it was rated consistently close to 100. Data for other conditions shows considerable spread, which makes it difficult to extract conclusions from the boxplot alone. This spread is caused by the individual bias caused by each subject’s own rating criteria. Thus, the appropriate way to analyse the data is through a statistical method for dependent samples, where all ratings by the same subject are considered as dependent on each other—which in this case would be a repeated measures analysis of variance (RM-ANOVA), as recommended by the ITU-R (2015a). A three-way RM-ANOVA was run for each of the four attributes (OVS/COL/DIR/DIS), where HpEQ, BRIR and audio material were the between-subject factors. A significance level of 0.05 was used. Results are reported by means of the F statistic returned by the RM-ANOVA and the t statistic in the case of post-hoc dependent samples t -tests.

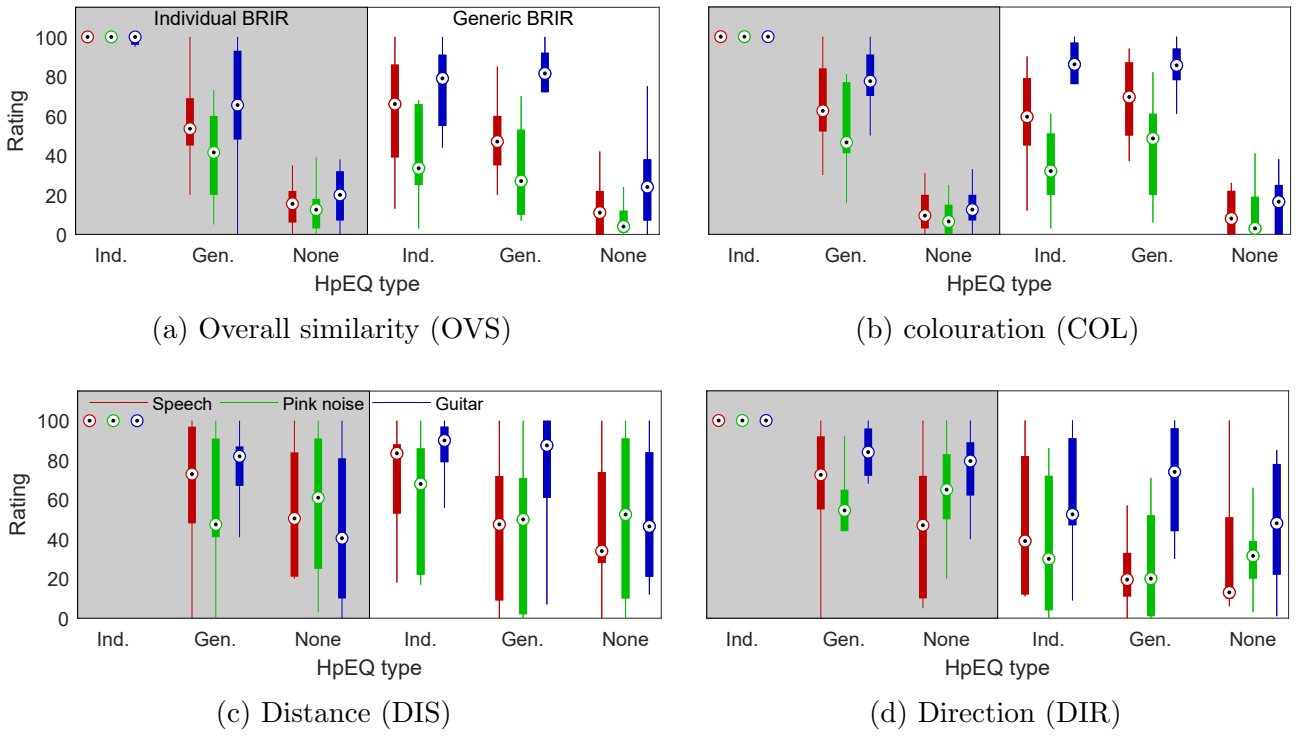


Figure 4.6: Results of the MUSHRA test for each of the four assessed attributes. Interquartile ranges are indicated with boxes, medians with circles, and the most extreme data points with thin lines. Background colour (grey/white) denotes the BRIR type. Audio materials are, from left to right, *speech* (red), *pink noise* (green) and *guitar* (blue).

Effect of BRIR

A significant effect of BRIR type was found on all attributes [$F(1, 9) > 11.91, p < 0.008$], as well as a strong interaction between BRIR and HpEQ type [$F(2, 18) > 5.58, p < 0.02$]. This interaction was largely caused by the IndBRIR+IndHpEQ (reference) condition obtaining very high and consistent ratings, therefore biasing the results of the RM-ANOVA.

When considering only GenHpEQ and NoHpEQ data, the effect of the BRIR was found to be significant only on DIR [$F(1, 9) = 14.49, p = 0.004$]. Post hoc dependent samples *t*-tests show that for this attribute, individual BRIRs got higher ratings than the generic BRIR, independently of the audio material and HpEQ type. No significant effect of the BRIR was found on OVS, COL or DIS in the same analysis.

Effect of HpEQ

A significant effect of the HpEQ type was found for all attributes [$F(2, 18) > 6.05, p < 0.01$]. As mentioned above, the significant interaction between HpEQ and BRIR made it necessary to perform separate analyses for the different conditions.

The difference between IndHpEQ and GenHpEQ was found to be significant on IndBRIR conditions for all attributes and audio materials [$F(1, 9) > 15.23, p < 0.004$], and non significant on GenBRIR conditions.

On the other hand, GenHpEQ ratings were found to be significantly higher than NoHpEQ ones on OVS and COL [$F(1, 9) > 30.69, p < 0.001$], but not on DIS or DIR. Dependent samples t -tests showed that on OVS and COL the difference was significant for all audio materials and both BRIR types [$t(19) > 4.82, p < 0.001$]. Further, the effect on DIS, the effect was found to be significant for *guitar* [$t(19) = 3.36, p = 0.003$] and not significant for *speech* or *pink noise*.

4.5 Discussion

Results of the subjective evaluation indicate that BRIR type was the dominant factor in perceived direction, which agrees with the common consensus that an individualised HRTF should always yield more accurate sound localisation than a generic one (Møller, Sørensen, et al., 1996). This result is commensurate with the observations made in the objective evaluation, where GenBRIR conditions were found to have a higher ILD error than IndBRIR ones, which probably caused a larger error in the perceived source's azimuth.

Another finding was that colouration was mainly affected by the HpEQ type, and the largest differences are perceived in the NoHpEQ condition. This result could also have been anticipated from the objective evaluation, given that the largest monaural errors were observed on unequalised conditions.

The differences between IndHpEQ and GenHpEQ seem to be very evident when using the individual BRIR, which can be explained by the fact that the reference is mostly rated 100,

whereas the other condition has the subject's individual bias within it, but are less accentuated when using the generic BRIR. Actually, results suggest that listeners did not perceive any improvement when using their own HpEQ filter on GenBRIR conditions, not only in terms of overall similarity to a reference (which was previously shown by Lindau and Brinkmann, 2012) but also for specific attributes such as colouration, distance and direction. The latter two are particularly interesting because they imply that the spatial perception of a non-individualised rendering is not significantly altered if a generic HpEQ is used instead of an individualised one.

It is noteworthy that the correlation between OVS and COL is higher (Pearson correlation coefficient $\rho = 0.76$) than the correlation between OVS and DIR ($\rho = 0.52$) or OVS and DIS ($\rho = 0.48$), which may indicate that, when rating overall similarity, subjects were paying more attention to colouration changes than to spatial features. This would be supported by the results from Brinkmann, Lindau, and Weinzierl (2017), where colouration and timbre-related attributes were the ones given the most weight by listeners when comparing binaural content.

Interestingly, DIS, which was chosen as an intuitive attribute to evaluate externalisation, is partially affected by the HpEQ type, as GenHpEQ obtained higher ratings than NoHpEQ for some audio materials. This trend may indicate that subjects externalised the equalised audio content better than the non-equalised one. Several subjects claimed in informal discussions after the experiment that it was harder for them to externalise the broadband noise than the other audio materials. Such statements would support the observed trend, given that *pink noise* stimuli received lower DIS ratings than *speech* or *guitar* ones.

Finally, the lack of evidence of any effect of HpEQ on DIR indicates that (a) the HpEQ filter did not alter the interaural cues enough to be perceivable, which is supported by the objective evaluation, where ILD error was not found to be significantly affected by HpEQ type, and (b) perception of elevation did not change even though monaural cues were altered by the HpEQ filter. This might be partially explained by the fact that the loudspeaker was visible to the subjects at all times, so it is possible that a 'ventriloquist effect' (Alais and Burr, 2004) was taking place, effectively pulling the sounds towards the loudspeaker as the most plausible visible source. In hindsight, this could have been prevented by employing an acoustically transparent

curtain to block the view of the loudspeaker. It would have been interesting to test more source directions. However, due to the relatively high number of independent variables, testing more directions is suggested for future experiments. One possible follow-up to this study would be to evaluate the same set of HpEQ types in a sound localisation task, which would provide a better understanding of the impact of a fixed headphone compensation filter on elevation perception for a static listener.

Overall, results suggest that generic headphone EQ offers perceptual benefits over unequalised content in terms of overall quality, colouration and, for certain audio contents, perceived distance. Thus, generic EQ may have a positive effect, both in the perceived quality and in the accuracy of the spatial features of binaural content, and therefore it might be beneficial to use:

1. when presenting a non-individualised binaural rendering, or
2. when presenting an individualised binaural rendering and an individual HpEQ is not available.

4.6 Conclusions

In this study, the effect of generic headphone compensation on binaural renderings was explored. Individual BRIRs and HpTFs were measured for several listeners and a ‘generic’ dummy head and open-ear headphones. Then, an objective analysis and a perceptual evaluation were performed to compare different combinations of generic and individual BRIRs and HpEQ filters.

These results suggest that headphone compensation should be applied to minimise any potential degradation caused by the headphones’ frequency response, in order to optimise the perceived quality of binaural renderings. Ideally, such compensation should be applied through individualised EQ filters that are measured for each user. However, it has been shown that a generic filter can also provide perceptual benefits compared to unequalised reproduction, including a reduction in colouration, an increase in overall quality, and an improved perception of distance when using open-ear headphones. Furthermore, it was found that, for binaural renderings

based on non-individualised BRIRs, the benefits provided by generic EQ were not statistically significantly different to those provided by individualised EQ for all the tested attributes (overall similarity, colouration, distance and direction), for the tested open-ear headphones. It is worth noting that this study was limited to testing a single source position and, even though it is speculated that the results may not necessarily vary if a different position is used (e.g. outside the horizontal plane), this should be specifically tested in follow-up studies.

Chapter 5

Binaural Rendering Methods: a Model-Based Assessment

The central component of AAR is the auralisation or binaural rendering of virtual sound sources. In practice, this process will likely be done in real time in order to react to potential movements of the listener or the sources; therefore computational cost must be considered. For an efficient binaural rendering, particularly if it includes reverberation, it is often beneficial to employ the Ambisonics framework to efficiently manipulate the sound field and potentially reduce its spatial resolution to alleviate computational costs. This chapter reviews and compares a selection of state-of-the-art Ambisonics-based binaural rendering methods, including a novel one, looking at their ability to produce accurate auralisations from low-order Ambisonics signals.

This chapter was adapted from a manuscript by the present author and others (Engel, Goodman, et al., [2021](#)), which was submitted to the journal Acta Acustica in March 2021 and is under review at the time of writing this Thesis. Copyright on articles published in said journal is retained by the authors under a Creative Commons Attribution 4.0 International License (CC BY).

Abstract

Binaural rendering of Ambisonics signals is often used to reproduce spatial audio content.

Processing Ambisonics signals at low spatial orders is desirable in order to reduce complexity, although it may degrade the perceived quality. Several methods have been proposed to alleviate this issue (including one introduced in this study), but they have not been thoroughly compared yet. Nine state-of-the-art binaural Ambisonics rendering methods were compared for spatial orders 1 to 44 by estimating localisation performance, externalisation and speech reception using perceptual hearing models. This assessment was supported by numerical analyses of HRTF interpolation errors, interaural differences, perceptually-relevant spectral differences, and loudness stability. Models predicted that the binaural renderings' accuracy increased with spatial order, as expected. A notable effect of the rendering method was observed: whereas all methods performed similarly at the highest spatial orders, some were considerably better at lower orders. The proposed method, bilateral magnitude least squares (BiMagLS), displayed the best performance overall. This study provides a review of existing binaural Ambisonics rendering methods using binaural models as the main mean of assessment. The results, which were in line with previous literature, indirectly validate the models' ability to predict listeners' responses in a consistent and explicable manner.

5.1 Introduction

5.1.1 Binaural rendering and Ambisonics

Binaural rendering allows to present auditory scenes through headphones while preserving spatial cues, so the listener perceives the simulated sound sources at precise locations outside their head (Wightman and Kistler, 1989a; Møller, 1992; Begault and Trejo, 2000). Traditionally, this is achieved by convolving an anechoic audio signal with a HRIR (Cuevas-Rodríguez et al., 2019). Typically, HRIRs are measured or simulated for a set of directions on a specific listener. Convolution signals with HRIRs is a convenient method to simulate a limited number of sound sources in anechoic conditions, but it cannot be easily used to accurately render reverberation or 'scene-based' spatial audio formats, e.g. recorded with spherical microphone arrays. Furthermore, the implementation of rotations, in order to allow the listener to turn their head and keep the sources fixed relative to the surrounding space, can be relatively inconvenient when using

(anechoic) HRIRs, given that a separate convolution with a HRIR should be performed for produce each single reflection, potentially resulting in an impractically large number of real-time convolutions. For such applications and features, it is common to employ Ambisonics instead.

Ambisonics, first introduced by Gerzon (1973), is an audio signal processing framework that allows to conveniently record, represent, post-process and reproduce spatial audio (Zotter and Frank, 2019). It is the de facto standard format for three-dimensional sound fields recorded with spherical microphone arrays (Bernschütz, 2016; Nowak and Klockgether, 2017; Zaunschirm, Frank, et al., 2018; Engel, Henry, Amengual Garí, Robinson, Poirier-Quinot, et al., 2019; Ahrens and Andersson, 2019; Lübeck, Pörschmann, et al., 2020; Lübeck, Arend, et al., 2020) and also finds use in acoustic simulations for VR and AR, such as Facebook’s (formerly Oculus’) audio engine (Schissler, Stirling, et al., 2017) and Google’s Resonance Audio (Gorzel et al., 2019). Although it was initially intended for loudspeaker playback, Ambisonics has recently found a niche in binaural audio reproduction, mostly due to an increased interest in VR and AR.

In essence, Ambisonics allows the representation of a three-dimensional sound field by projecting it on a hypothetical sphere surrounding the listener. Then, the signal can be conveniently manipulated through a mathematical framework known as spherical harmonics (SH)—an excellent introduction for its usage in acoustics is given by Rafaely (2015, Chapter 2). When a sound field is ‘encoded’ into the Ambisonics domain, it is assigned an inherent spatial order ($N \in \mathbb{N}$), also known as truncation order, which dictates its spatial resolution. As a general rule, lower orders will lead to coarser spatial resolution with wide or blurry sound sources, whereas higher orders will produce finer resolution with narrower and better-localised sources (Avni et al., 2013; Zotter and Frank, 2019). The spatial order of an Ambisonics signal is often constrained by the application, e.g. commercial microphone arrays typically operate at order four or lower, while real-time audio engines benefit from working with low orders, as it reduces computational costs (Schissler, Stirling, et al., 2017).

For binaural playback, an Ambisonics signal must be ‘decoded’ to two channels (left and right ears) by combining it with an HRTF, which is how we refer to an HRIR set when expressed in the frequency-domain. This has traditionally been done with the virtual loudspeaker method

(McKeag and McGrath, 1996), although recent studies have suggested to instead encode the HRTF in the SH domain and operate there directly, as this allows to preprocess the HRTF in various ways to improve the quality of the output binaural signals (Bernschütz, Vázquez Giner, et al., 2014; McKenzie et al., 2018; Zaunschirm, Schörkhuber, et al., 2018; Schörkhuber et al., 2018). In addition to this added versatility, the SH-based approach allows to implement the virtual loudspeaker method itself, as shown by Ben-Hur, Sheaffer, et al. (2018) and discussed in Section 5.2, therefore providing a more general solution to the binaural rendering problem. For these reasons, the SH-based approach is employed here.

Since HRTFs are typically measured or simulated offline, it is safe to assume that they can be provided with high spatial resolution. In fact, high-quality, densely sampled generic HRTFs are already available, e.g. as in (Bernschütz, 2013; Brinkmann, Lindau, Weinzierl, et al., 2017), and there is a good amount of ongoing research on the production of individual HRTFs of similar quality (see Guezenoc and Segui, 2018, for a review) and on the spatial upsampling of sparse HRTFs (Pörschmann et al., 2019; Ben-Hur, Alon, Mehra, et al., 2019).

Therefore, in practice, it is common to encounter situations where a binaural rendering must be obtained by combining a low-order Ambisonics signal and a spatially dense HRTF. This mismatch can cause a loss of relevant information from the HRTF due to order truncation (as demonstrated in Section 2.5), which leads to audible artefacts in the binaural signals, such as spectral colouration, loudness instability across directions and localisation blur (Avni et al., 2013; Ben-Hur, Alon, Rafaely, et al., 2019). Several techniques (hereafter, ‘rendering methods’) have been proposed to mitigate these artefacts. A selection of relevant rendering methods is reviewed in Section 5.2.

5.1.2 Research question and contributions

Finding the best binaural Ambisonics rendering method, i.e. the one that best mitigates artefacts related to the spatial resolution mismatch between Ambisonics signal and HRTF, is an active research topic. Previous studies have compared different rendering methods through listening tests (Lübeck, Helmholz, et al., 2020; Lübeck, Arend, et al., 2020; McKenzie et al., 2019a; Lee

et al., 2019; Ben-Hur, Alon, Mehra, et al., 2020; Ben-Hur, Alon, Mehra, et al., 2021) but the complexity and time-consuming nature of such experiments heavily limits the amount of test conditions. Ideally, one would compare all state-of-the-art rendering methods through a variety of metrics, e.g. localisation performance, speech reception in noise, or externalisation, for a wide range of spatial orders. However, most of the aforementioned studies only assessed one perceptual metric (usually, similarity to a reference signal) or considered just a few spatial orders in their evaluation.

Binaural models, which offer a computational simulation of binaural auditory processing and, in certain cases, allow also to predict listeners' responses to binaural signals, are an invaluable tool that could help overcome such limitations. Using them, it is possible to rapidly perform comprehensive evaluations that would be too time-consuming to implement as actual auditory experiments, e.g. as in (Brinkmann and Weinzierl, 2018). Additionally, model-based evaluations could be extremely useful when access to human subjects is limited, such as in times of pandemic. It is likely that models will not provide the exact same results as an auditory experiment and therefore will not lead to assess fine differences between stimuli at a high level of accuracy, e.g. near to the zone of perfect reproduction, but it is reasonable to expect them to provide broadly correct predictions for larger errors. This means that they could be particularly useful in our particular case for comparing between low-order rendering methods, and possibly providing insights on overall trends.

The aim of the present study is twofold: first, to propose a systematic evaluation procedure for binaural Ambisonics rendering methods based on auditory models; and second, to find which state-of-the-art rendering method performs best for a wide range of spatial orders and perceptual metrics. In particular, three different models from the auditory modeling toolbox (AMT), developed by Sørensgaard and Majdak (2013), were employed in the assessment: the localisation model by Reijniers et al. (2014), the externalisation model by Baumgartner and Majdak (2020), and the speech reception in noise model by Jelfs et al. (2011). Furthermore, this evaluation is complemented by numerical analyses in order to relate the models' predictions to objective metrics.

All in all, the contributions of the present study can be summarised as follows:

1. a review of the state of the art in binaural Ambisonics rendering methods;
2. a comparison of relevant rendering methods through numerical analyses and perceptual models (localisation performance, externalisation, speech perception);
3. a novel rendering method, BiMagLS, which combines two state-of-the-art methods to produce more accurate binaural signals; and
4. an indirect validation of the perceptual models' ability to predict user responses to binaural signals, by being able to replicate results from previous auditory experiments.

This chapter is structured as follows: Section 2.5 provides some theoretical background on the Ambisonics framework and the issue of order truncation in binaural rendering; Section 5.2 presents the different binaural Ambisonics rendering methods under evaluation and introduces the novel BiMagLS; Section 5.3 describes the evaluation procedure, including numerical analyses and perceptual models; Section 5.4 presents the results; Section 5.5 discusses them; and Section 5.6 summarises the outcomes and concludes the chapter.

5.2 Rendering methods

This section presents the binaural Ambisonics rendering methods that were compared in this study. We assume a scenario where a binaural signal must be obtained from an Ambisonics sound field of limited order N (free of spatial aliasing errors) and a high-resolution HRTF. Each rendering method aims to obtain the SH coefficients of the HRTF (SH-HRTF) up to order N , so it can be applied to the sound field in a way that the resulting binaural signals contain minimal truncation-related artefacts. A discussion on the process of obtaining the SH-HRTF and the nature of truncation-related artefacts is provided in Section 2.5. Implementation details are briefly described for each method and the corresponding MATLAB (MathWorks, Natick, MA, United States) code is available at the BinauralSH repository (Engel, [2021](#)). Parametric

methods, which exploit prior knowledge of the sound field (McCormack and Delikaris-Manias, 2019), are not considered in this review.

Table 5.1: Evaluated binaural rendering methods.

Method	Implementation notes
Trunc	Obtain N th order SH-HRTF via discrete SFT (Equation 2.11).
EQ	Apply Trunc, then equalise through HRF (Ben-Hur, Brinkmann, et al., 2017) with frequency-dependent regularisation (Kirkeby and P. A. Nelson, 1999).
Tap	Apply Trunc, then tapering (Hold et al., 2019) with a Hann window for $n \geq (N - 3)$ and $f > f_a$, and finally apply EQ.
TA	Time-align HRTF via phase correction by ear alignment (Ben-Hur, Alon, Mehra, et al., 2019), then apply Trunc.
MagLS	Obtain N th order SH-HRTF via magnitude least squares as in (Zotter and Frank, 2019, Section 4.11.2) with smoothing around the cutoff.
MagLS+CC	Same as MagLS and then apply covariance constraint as in (Zotter and Frank, 2019, Section 4.11.3).
SpSub	Obtain N th order SH-HRTF via spatial subsampling with N^{th} order Gauss grids (Bernschütz, Vázquez Giner, et al., 2014).
SpSubMod	Time-align HRTF below f_a as in (Zaunschirm, Schörkhuber, et al., 2018), then apply SpSub and finally apply Tap (McKenzie et al., 2019a).
BiMagLS	First apply TA to time-align the HRTF and then obtain N th order SH-HRTF via MagLS with cutoff at 3 kHz.

5.2.1 Truncation (Trunc)

This is the baseline method, which does not attempt to mitigate truncation-related artefacts. Here, the N th order SH-HRTF is obtained by simply taking a high-order SH-HRTF and removing all SH coefficients from $N + 1$ onwards. In practice, this is typically done by applying the discrete version of SFT, as defined in Equation 2.11. This method is expected to produce large artefacts in the binaural signals, particularly for frequencies above the so-called aliasing frequency, which is proportional to the truncation order (see Equation 2.10).

5.2.2 Equalisation (EQ)

One of the most distinct effects of order truncation is the spectral roll-off that occurs above the aliasing frequency, leading to an excessive emphasis of low-frequency content in the rendered binaural signals. An easy way to mitigate this effect is applying a global EQ filter to the

SH-HRTF, so that its diffuse field component, i.e. its average magnitude across directions, matches the one of a reference—usually a higher-order version (Ben-Hur, Brinkmann, et al., 2017; McKenzie et al., 2018).

Different EQ methods have been proposed. Ben-Hur, Brinkmann, et al. (2017) discuss the two most popular approaches: the first one calculates the diffuse field component of the HRTF and inverts it, resulting in HRTF-related filters (HRF), whereas the second one employs spherical head filters (SHF) derived from an analytical spherical head model. In that study, the authors argue that HRF is more accurate than SHF, but also more sensitive to noise within the HRTF, e.g. inverting a notch of the diffuse field component could lead to excessive amplification and subsequent ringing artefacts, although both methods produce perceptually similar results.

Implementation: In this study, the EQ method was implemented by first obtaining the truncated SH-HRTF (Equation 2.11) and then applying HRF obtained from a 44th order SH-HRTF, following Equation 14 from the study by Ben-Hur, Brinkmann, et al. (2017). Additionally, frequency-dependent regularisation (Kirkeby and P. A. Nelson, 1999) was employed when calculating the EQ filters to avoid excessive amplification, as implemented in (Engel, Alon, Robinson, et al., 2019; Engel, Alon, Scheumann, et al., 2020). Preliminary tests showed that SHF and HRF performed similarly under these conditions, so only the latter was included in the evaluation for the sake of clarity.

5.2.3 Tapering (Tap)

Another effect of order truncation is the emergence of side lobes in the polar pattern of the SH-HRTF, which often cause sound sources to rapidly change loudness across directions (Ben-Hur, Alon, Rafaely, et al., 2019). These changes can happen for the two ears independently, and can therefore result in alterations of the ILDs, which are an essential cue for sound localisation (McKenzie et al., 2019b). To mitigate this effect, Hold et al. (2019) proposed the tapering (Tap) method, which consists in applying gradually decreasing weights to the SH coefficients of the SH-HRTF. Hold et al. draw the following analogy: the same way that a rectangular window applied to a time-domain signal produces undesired frequency-domain leakage in the form of

side lobes, hard order truncation in the SH domain produces side lobes in the space domain. Following this analogy, they propose to apply a Hann window to the SH coefficients, giving lower weight to the higher spatial orders, which smooths the SH-HRTF's polar pattern.

The tapering method is reminiscent of Max-rE weighting, a technique used to maximise sound field directivity in Ambisonics loudspeaker decoding. This method, proposed by Jérôme Daniel et al. (1998) and Jérôme Daniel (2001), applies scalar weights to the different Ambisonics channels in a way that the sound field's energy vector (rE) from Gerzon's sound localisation model (Gerzon, 1992) is maximised. In essence, the weights are highest for order 0 and decrease monotonically for higher spatial orders, much like the tapering window by Hold et al. Although mostly used for loudspeaker decoding, Max-rE weighting has also been employed in a binaural context in (McKenzie et al., 2018; McKenzie et al., 2019a), where a dual-band approach is employed, applying the weighting only above the aliasing frequency f_a .

Implementation: In this study, the Tap method was implemented by obtaining the truncated SH-HRTF (Equation 2.11) and applying Hann weights following (Hold et al., 2019), except that a shorter Hann window was employed so that only the 3 highest orders [$n \geq (N - 3)$] were tapered in order to avoid excessive attenuation, as suggested in (Lübeck, Helmholz, et al., 2020). Furthermore, a dual-band approach was employed, so weights were only applied above the aliasing frequency (Equation 2.10). Finally, HRF EQ was applied to the tapered SH-HRTF. Informal tests showed that dual-band tapering performed generally better than single-band, which is in agreement with the findings by McKenzie et al. (2019a), whereas Max-rE and Hann weights performed similarly.

5.2.4 Time-alignment (TA)

Previous studies have shown that time-aligning all HRIRs within a dataset, essentially removing the ITDs, substantially reduces the effective spatial order of the resulting SH-HRTF (Evans et al., 1998; Ben-Hur, Alon, Mehra, et al., 2019; Arend et al., 2021)—of course ITDs must be reinserted back when rendering the binaural signals, similarly to what is done in (Cuevas-Rodríguez et al., 2019). This is illustrated in Figure 2.4: whereas a time-aligned HRTF presents a compressed

SH spectrum that can be truncated at $N = 5$ and still preserve 90% of its energy at 10 kHz (Figure 2.4b), the non-aligned version needs up to $N = 17$ to preserve the same amount at that frequency (Figure 2.4a). This is because phase accounts for most of the spatial complexity of an HRTF; therefore, if we remove the HRIRs' onset delays (which vary slightly across directions due to the ears not being at the origin of the coordinate system), we can considerably reduce the effective order of the SH-HRTF (Ben-Hur, Alon, Mehra, et al., 2019).

When the time-alignment method (TA) is used for HRIR interpolation, ITDs can be easily reinserted in the signal without losing information. However, this cannot be done when binaurally rendering Ambisonics signals, which is why TA requires so-called bilateral Ambisonics signals, for which two receivers at the listener's ears' positions are used instead of a single one at the centre of the head Ben-Hur, Alon, Mehra, et al., 2021. This dual-receiver setup is straightforward to implement in an acoustic simulation, but it is worth noting that it will require separate simulations for different head rotations due to the left- and right-ear signals not sharing the same coordinate system, which contrasts with typical Ambisonics rendering in which head rotations can be easily derived Zotter and Frank, 2019, Sec. 5.2.2. An approximate conversion from standard to bilateral Ambisonics signals and vice-versa is possible via sound field translation, as discussed by Ben-Hur, Alon, Mehra, et al. (2021), but it is still unclear if the errors introduced during this process introduce large perceptual degradation in the signals.

Based on an evaluation with auditory models, Brinkmann and Weinzierl (2018) suggested that a time-aligned SH-HRTF truncated to $N = 3$ could produce binaural signals that were not significantly different from a higher-order reference in terms of localisation performance, colouration and interaural cross-correlation coefficient (IACC), whereas a non-aligned one required $N = 19$. This is in agreement with a recent study by Ben-Hur, Alon, Mehra, et al. (2021), who showed that a fourth-order binaural Ambisonics rendering generated with a time-aligned HRTF was rated by listeners as identical to a 41st-order reference in a perceptual test.

Implementation: In this study, time-alignment (TA) was implemented with the 'phase correction by ear alignment' method, as proposed by Ben-Hur, Alon, Mehra, et al. (2019), time-aligning

the HRTF before obtaining the truncated SH-HRTF with Equation 2.11. This approach has been shown to be more robust than methods based on onset detection, e.g. as in (Brinkmann and Weinzierl, 2018), and obtained promising results in recent perceptual studies (Ben-Hur, Alon, Mehra, et al., 2020; Ben-Hur, Alon, Mehra, et al., 2021).

5.2.5 Magnitude least squares (MagLS, MagLS+CC)

Following the idea of TA, Zaunschirm, Schörkhuber, et al. (2018) proposed a perceptually-motivated alternative method where HRIR alignment is performed only above a given frequency cutoff (f_c), whereas ITDs are left intact below it. This frequency-dependent time-alignment (FDTA) method is based on the duplex theory (Strutt, 1907; Hartmann et al., 2016), which establishes that ITDs (and therefore, phase) are perceptually most relevant at low frequencies, whereas ILDs, i.e. magnitude, are dominant at high frequencies. The same authors later presented an improved version called magnitude least squares (MagLS), which achieved superior performance by entirely disregarding phase errors above f_c (Schörkhuber et al., 2018). Figure 2.4c shows how a magnitude-only version of an SH-HRTF displays an even lower effective spatial order than the time-aligned version (Figure 2.4b), which provides an intuition of why MagLS performs better than FDTA at low orders. In the same study, the authors showed that listeners could not perceive phase errors beyond 2 kHz for continuous signals (speech) or 4 kHz if considering envelope ITD, e.g. for pulsed noise.

There exists a variant of MagLS (hereafter, referred to as MagLS+CC) that employs the covariance matrix framework proposed by Vilkamo et al. (2013), applying a global EQ and correcting the interaural coherence of the binaural signal, which is expected to affect important perceptual cues such as source width (Zaunschirm, Schörkhuber, et al., 2018). Zotter and Frank (2019, Section 4.11.3) have recommended to employ this variant for spatial orders equal or lower than 3, but this has not been thoroughly tested yet.

Note that, in contrast to TA, MagLS reduces the effective order of the SH-HRTF while preserving ITDs. Consequently, it is compatible with standard Ambisonics signals and with dynamic simulation of listener’s head rotations.

Implementation: In this study, MagLS was implemented through a simple iterative procedure, as proposed by Zotter and Frank (2019, Section 4.11.2), setting the cutoff to the aliasing frequency ($f_c = f_a$)—the rationale being that, since large phase errors are expected to occur above the aliasing frequency, it is preferable to minimise magnitude errors as much as possible in that range. Furthermore, a smooth transition was applied one half-octave below and above the cutoff to avoid sharp changes in the frequency response and subsequent audible artefacts. The MagLS+CC variant was implemented following (Zotter and Frank, 2019, Section 4.11.3).

5.2.6 Spatial subsampling (SpSub, SpSubMod)

The spatial subsampling (SpSub) method mitigates truncation errors by sampling an HRTF at a reduced number of directions prior to obtaining its SH coefficients (Bernschütz, Vázquez Giner, et al., 2014). This intentionally introduces spatial aliasing errors in the SH-HRTF, effectively shifting high-frequency content towards low spatial orders. Although aliasing is often undesirable, it has been shown that, in this particular case, it compensates for truncation errors to some extent, mainly because aliasing introduces a high-shelf filtering effect that partially cancels out the roll-off caused by truncation (Bernschütz, Vázquez Giner, et al., 2014; Ben-Hur, Alon, Rafaely, et al., 2019).

The SpSub method produces identical output to the popular virtual loudspeakers method, first introduced by McKeag and McGrath (1996) and later employed by Noisternig et al. (2003) and the developers of Google’s Resonance Audio (Gorzel et al., 2019), among others. The equivalence between SpSub and virtual loudspeakers is subject to choosing an appropriate sampling scheme, i.e. the number of virtual loudspeakers and their locations, as shown in (Ben-Hur, Sheaffer, et al., 2018). Common sampling schemes include platonic solids which are only available for $N \leq 3$ (Engel, Henry, Amengual Garí, Robinson, Poirier-Quinot, et al., 2019; Gorzel et al., 2019), Gaussian quadratures (Stroud and Secrest, 1966; Bernschütz, 2016), Lebedev quadratures (Lebedev, 1977; Bernschütz, 2016) and T-designs (Hardin and Sloane, 1996; McKenzie et al., 2018).

McKenzie et al. (2019a) proposed a variant of SpSub, here referred to as modified spatial

subsampling (SpSubMod), which combines SpSub with FDTA, dual-band Max-rE weighting, i.e. tapering, and diffuse field EQ (HRF), which was shown to perform well for orders 1 to 3.

Implementation: In this study, SpSub was implemented by obtaining a high-order ($N_h = 44$) SH-HRTF via discrete SFT, then sampling this SH-HRTF to an N^{th} order Gaussian quadrature via discrete ISFT (Equation 2.12) and finally applying the SFT again to the result, as in Equation 2.11. This procedure was done according to Bernschütz (2016) and with the help of the SOFiA toolbox (Bernschütz, Pörschmann, et al., 2011). Gaussian quadrature, which is a spherical sampling scheme that consists in choosing points at constant angle intervals, were chosen as they perform well for a wide range of truncation orders, compared to other sampling schemes such as Lebedev quadratures, according to Bernschütz (2016). Additionally, the SpSubMod variant was implemented by applying FDTA (as in Zotter and Frank, 2019, Section 4.11.1) prior to SpSub, then applying dual-band Hann tapering and, finally, HRF EQ (McKenzie et al., 2019a).

5.2.7 Bilateral magnitude least squares (BiMagLS)

A novel method is introduced in this study called bilateral magnitude least squares (BiMagLS). This method is presented as an improved version of TA and consists of the following steps:

1. first, the HRTF is time-aligned as in the TA method;
2. for frequencies below a given threshold, the SH-HRTF of order N is obtained by means of least-squares fitting of a high-order HRTF;
3. for frequencies above the threshold, the SH-HRTF of order N is obtained by means least-squares fitting *only the magnitude* of the same high-order HRTF, while phase is estimated with the iterative procedure suggested by Zotter and Frank Zotter and Frank, 2019, Sec. 4.11.2.

In other words, BiMagLS is equivalent to applying MagLS preprocessing to a time-aligned HRTF. Much like TA, this method is only compatible with bilateral Ambisonics due to the

HRTF being time-aligned across the whole frequency spectrum. By combining the accurate phase reconstruction of TA and the accurate magnitude reconstruction of MagLS, this method is expected to outperform TA when rendering bilateral Ambisonics signals.

Implementation: BiMagLS is implemented by first time-aligning the HRTF using phase correction by ear alignment (Ben-Hur, Alon, Mehra, et al., 2019) and then generating the order-limited SH-HRTF via MagLS, as described earlier. The frequency threshold was set to 3 kHz, independently from the truncation order. This cutoff was chosen empirically, as it provided best results in informal tests. A smooth transition is applied one half-octave below and above the cutoff.

5.2.8 Overview

The following nine rendering methods were implemented: truncation (Trunc), EQ, Tap, TA, MagLS, MagLS+CC, SpSub, SpSubMod, and BiMagLS, as summarised in Table 5.1. The method BiMagLS, which combines the qualities of TA and MagLS and is presented as a direct improvement of the former, has been introduced in this work. Of the nine methods, two of them (TA and BiMagLS) assume a time-aligned HRTF and cannot be used directly to binaurally render a standard Ambisonics signal. Even though they are not directly comparable to the other methods, they have been included for the sake of completeness, as they are still valuable for HRTF interpolation and for rendering bilateral Ambisonics signals, i.e. measured at the ears' positions.

A previous perceptual study by Lübeck, Arend, et al. (2020) has already compared Trunc, EQ, Tap, SpSub and MagLS for the binaural rendering of microphone array recordings, using a dummy-head recording as the reference. Their data showed that all methods achieved an increase in quality compared to a low-quality anchor (low-passed diotic signal), but no significant differences were observed among the methods at high orders. One limitation of the said study was that only three spatial orders were evaluated (3, 5 and 7) and only one perceptual metric (similarity to the reference) was evaluated. In the present study, we aim to complement their results by assessing some additional methods, a wider range of spatial orders and several

perceptual metrics. This is achieved thanks to a model-based evaluation and complementary numerical analyses, which are detailed in the next section.

5.3 Evaluation methods

The previous section introduced the nine binaural Ambisonics rendering methods to be assessed. For the evaluation, a publicly available HRTF of the FABIAN dummy head with an upright head-torso orientation (Brinkmann, Lindau, Weinzierl, et al., 2017) was employed. The HRTF was measured for 11950 directions and HRIRs had a length of 2048 samples (zero-padded from 256), sampled at a rate of 44.1 kHz.

SH-HRTFs of orders 1 to 44 were generated with every rendering method as indicated in Section 5.2. Then, from each order-limited SH-HRTF, HRIRs were interpolated to the original 11950 directions via ISFT (Equation 2.12). Some subsets of directions were given special attention: those in the horizontal plane (180 directions), those in the median plane (also 180) and those closest to a 110-point Lebedev grid, which is evenly spaced around the sphere. Finally, the differences between the interpolated and actual HRIRs were evaluated through numerical analysis and auditory models. The procedure is detailed in the following subsections.

5.3.1 Numerical analysis

The first step of the numerical analysis was to obtain *magnitude and phase interpolation errors* for the 110 positions closest to the aforementioned Lebedev grid. Magnitude error was calculated as the absolute difference between the log-magnitude of the original HRIRs and the interpolated ones, averaged across directions. Phase error was calculated as the absolute difference between the interaural phase delay of the original HRIRs and the interpolated ones, averaged across directions. Interaural phase delay was obtained by subtracting phase delay (unwrapped phase, calculated with the `unwrap` function from MATLAB R2020b, divided by frequency) of the right channel from the left one in an HRIR pair. These two metrics would offer a first insight on the

accuracy of a given rendering method. For instance, large magnitude errors are expected to distort monoaural cues and, by extension, externalisation and vertical localisation performance (Baumgartner and Majdak, 2020; Baumgartner, Majdak, and Laback, 2014), as well as ILDs. On the other hand, large phase errors are expected to affect ITDs and low-frequency lateral localisation, being most perceptually relevant below 2 kHz (perhaps 4 kHz, for some stimuli), according to Schörkhuber et al. (2018).

The second step was to estimate the *interaural cues*, namely ITDs and ILDs, for the 180 horizontal-plane directions on both the original and interpolated HRTFs. This would complement the interpolation error data and allowed for a more perceptually-motivated analysis. ITDs were estimated with the MaxIACCe method, after applying a low-pass filter (3 kHz) to the HRIRs, as described by Katz and Noisternig (2014). ILDs were estimated according to McKenzie et al. (2019b), by calculating them for thirty ERBs on high-passed (1.5 kHz) HRIRs and then averaged. Interaural coherence was also initially considered, but preliminary tests showed that it was generally very close to one in most cases, probably because of the anechoic test conditions—although it could be a useful metric in future evaluations including reverberation.

The third step was to analyse how the magnitude interpolation errors varied across different directions. This was expected to provide insights on the spatial leakage effects that methods such as Tap aim to alleviate. Instead of looking at the direction-dependent errors for each frequency bin separately, we opted for ‘collapsing’ the frequency axis by using the model by Calum Armstrong et al. (2018). This model translates magnitude deviations into estimated loudness differences, and performs a weighted average over the full frequency range by means of ERBs (Glasberg and Moore, 1990). As a result, we estimate the magnitude of the HRTF for a given direction as a single scalar measured in sones. The loudness difference between a given interpolated HRTF and a reference is referred to as the *perceptual spectral difference (PSD)*, which quantifies the distance between two HRTFs’ magnitude spectra in a perceptually-motivated way, as shown by McKenzie et al. (2018, Sec. 4.1).

5.3.2 Auditory models

Finally, the interpolated HRIRs were evaluated through binaural models. First, *localisation performance* was estimated using the ideal-observer model by Reijniers et al. (2014), as implemented by Barumerli et al. (2020) in the AMT. The model predicted localisation performance 100 times for each of the 110 Lebedev grid directions, in order to account for the stochastic processes implemented by the model, which aim to replicate the uncertainty by the listener when performing a localisation task. Then, the overall lateral and polar accuracy and precision were calculated. This model estimates sound localisation performance on the whole sphere, which allows for more insightful predictions than previous models like the ones by May, van de Par, et al. (2011, lateral localisation only) or Baumgartner, Majdak, and Laback (2014, sagittal localisation only). A key feature is its Bayesian modelling approach, which allows to predict listener’s uncertainty when assessing the location of a sound source. This is crucial for our purposes, considering that one of the effects of spatial order truncation is localisation blur, or sound sources appearing wider than they should (Avni et al., 2013). A wide sound source (e.g. one that extends over a large surface) and a narrow one (e.g. one that is perceived as providing from a precise point in space) should, on average, be both localised at the correct position (same accuracy), but the narrow source should yield lower localisation variance than the wide one (different precision). Therefore, the localisation precision predicted by the model could be very valuable in this evaluation. For an example of analysis of localisation accuracy and precision, the reader is referred to the work by Majdak et al. (2010).

Second, *externalisation* was predicted for the 180 median-plane directions, using the model by Baumgartner and Majdak (2020), as implemented in the AMT, and then averaged across the said directions to obtain a single value. This model predicts externalisation as a weighted sum of two parameters: monaural spectral similarity and interaural broadband time-intensity coherence. It is worth noting that the model considers a static (non-head-tracked) and unimodal (auditory information only) binaural rendering. Externalisation can be influenced by several factors that have not yet been accounted for in existing binaural models, such as early reflections and reverberation, visual information, listener expectations (see the ‘divergence effect’, Werner

et al., 2016) and dynamic cues, especially when caused by self-movements—an extensive review on externalisation was done by Best et al. (2020). However, these additional factors are not necessarily influenced by the independent variables used in this study (spatial order, binaural rendering method) and, therefore, a static externalisation estimation was considered a valuable metric for our purposes.

Finally, *speech reception in noise* was evaluated with the model by Jelfs et al. (2011), as implemented in the AMT. The model predicted the spatial release from masking (SRM), expressed as the benefit in dB provided by the better-ear and binaural unmasking effects, for one target source and one masker (multiple maskers could have been used as well, but this was not considered beneficial for the purpose of the current study). It was run 180 times per HRTF, changing the masker position between each of the horizontal plane directions, whereas the target was always placed in front of the listener. No reverberation was included and the masker was set to the same level as the target source. Even though the model is intended to assess reverberant signals, it could provide useful insights on perceived source separation in a practical application of anechoic binaural rendering, e.g. a videoconference with spatial audio.

5.4 Results

5.4.1 Numerical analysis

Figure 5.1 shows how magnitude and phase interpolation errors varied with spatial order within the Trunc condition, which was chosen as a baseline for not implementing any mitigation of truncation-related artefacts (see Section 5.2). It can be seen how errors rapidly increase after the aliasing frequency is surpassed, which depends on the order, e.g. 0.6 kHz for $N = 1$, 3 kHz for $N = 5$, etc. For the highest tested order (44), with an aliasing frequency well above the audible range, the average magnitude error is generally below 1 dB and the phase delay error, below 20 μ s.

By comparing these results to those from a perceptual experiment by Ben-Hur, Brinkmann,

et al. (2017), it is expected that magnitude errors will lead to large audible differences for low orders such as $N = 1$ and to more subtle (but still audible) differences for orders up to $N = 30$, for broadband signals such as pink noise. However, for band-limited stimuli like speech, it is expected that perceptual differences will be negligible above order 10, due to interpolation errors being prevalent at high frequencies. Regarding phase, it is expected that errors will be perceived as long as they modify the ITD above the just noticeable difference (JND) for frequencies below 4 kHz (Schörkhuber et al., 2018), which is the case for $N = 1$ and $N = 5$.

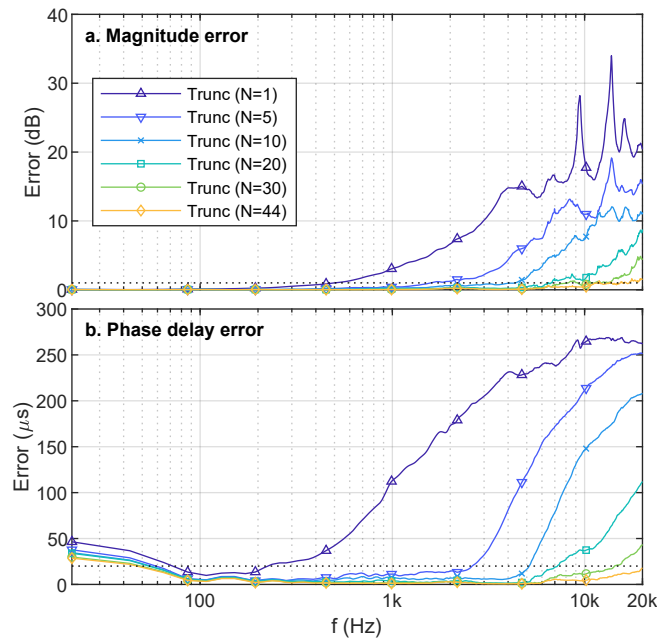


Figure 5.1: Absolute HRTF magnitude (left ear) and interaural phase delay errors, averaged across 110 directions in an approximate Lebedev grid. HRTFs were interpolated from truncated SH-HRTFs (Trunc method) for five different spatial orders (1, 10, 20, 30, 44). The dotted lines indicate an approximation of the just noticeable differences: 1 dB for magnitude and 20 μs for phase delay.

The same interpolated HRTFs are compared in terms of ITD and ILD on the horizontal plane (which is where interaural differences are expected to have the largest impact) in Figure 5.2. It can be seen how the interaural differences for $N = 1$ are quite different from the reference, as it could be anticipated from the large magnitude and phase errors reported earlier. Also, the 44th-order interpolated HRTF obtained very similar results to the reference, which is in agreement with its low interpolation errors. The figure also shows that ITD converges towards the reference quicker than ILD when increasing spatial order, with little improvement beyond $N = 5$. This can be explained by the fact that ITD estimation method mainly considers

frequencies below 3 kHz, whereas the ILD estimation method is mostly influenced by frequencies above 1 kHz (see Section 5.3) and, therefore, is affected more by high-frequency truncation errors..

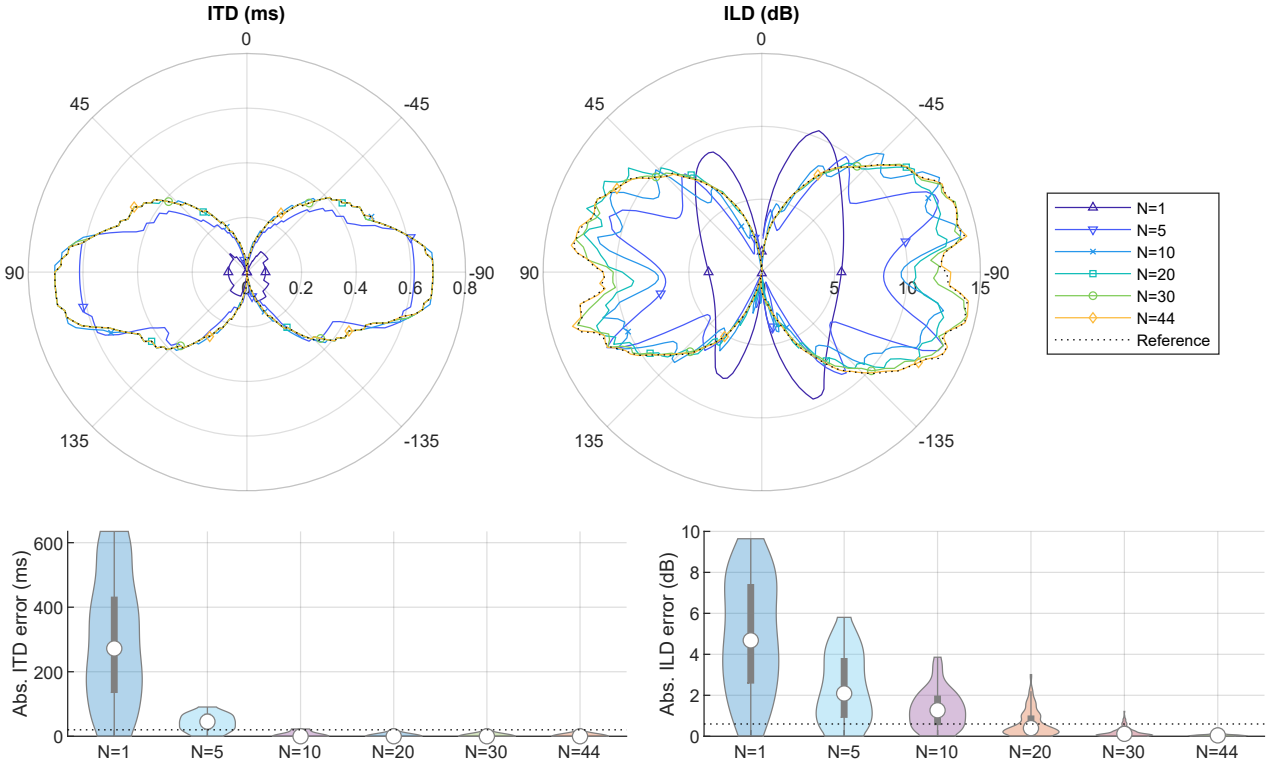


Figure 5.2: Top: ITD in ms and ILD in dB, plotted as a function of azimuth on the horizontal plane for the same HRTFs evaluated in Figure 5.1. Bottom: violin plots showing the absolute ITDs and ILD errors for each HRTF on the horizontal plane, where the dotted lines represent the approximate just noticeable differences in anechoic conditions, according to Klockgether and van de Par (2016).

The nine rendering methods are compared in Figure 5.3 for a spatial order of $N = 3$. It can be seen how magnitude and phase errors increase considerably above the aliasing frequency (marked with a vertical dashed line), as expected. The highest magnitude error was obtained for Trunc and the lowest ones, for MagLS and BiMagLS, which is in agreement with the instrumental evaluation by Lübeck, Helmholz, et al. (2020). In terms of phase, TA and BiMagLS both obtained the smallest errors overall, which was expected since these methods are able to accurately reconstruct ITDs.

Looking at interaural differences, Figure 5.4 shows how TA and BiMagLS are, as expected, the methods with most accurate ITDs (for $N = 3$) by a large margin, whereas other methods performed poorly in comparison, as a consequence of large phase errors. In terms of ILD, the

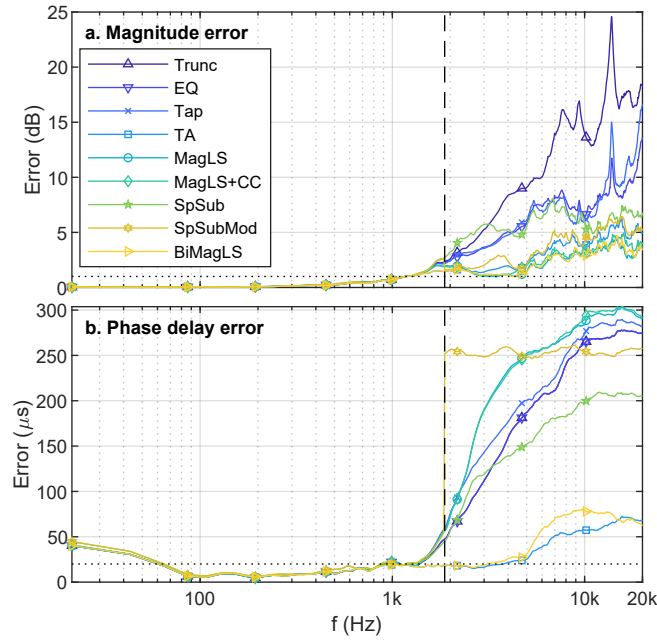


Figure 5.3: Absolute HRTF magnitude (left ear) and interaural phase delay errors, averaged across 110 directions in an approximate Lebedev grid. HRTFs were interpolated with the nine rendering methods and $N = 3$. The horizontal dotted lines indicate an approximation of the just noticeable differences: 1 dB for magnitude and $20 \mu\text{s}$ for phase delay. The vertical dashed line indicates the aliasing frequency.

similarity with the reference seems to agree with the magnitude error data shown earlier, with the largest deviations being produced by Trunc, EQ and SpSub, which showed lower ILDs at lateral directions. This can be explained by the side lobes that appear in the polar pattern of the SH-HRTF generated with said methods. Detailed data on interaural errors is shown in Table 5.2 and Table 5.3.

Table 5.2: Mean absolute ITD error between each method and the reference, per order (in μs).

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	296	231.4	169.3	100.9	40.9	12.6	9.3	14.5	5	5.2	1.6	0.8	0.1	0.5	0.3	0.3	0.1
EQ	319	254.9	189.5	111.6	43.3	12.8	9.7	14.5	5.2	4.9	1.6	0.8	0.3	0.5	0.3	0.3	0.1
Tap	332.5	259.5	223.6	149.4	59.1	13.6	9.6	14.6	5.2	5	1.6	0.8	0.3	0.5	0.3	0.3	0.1
TA	58.1	34.5	30	19.9	12.6	7.3	5.8	6.4	4.9	3	1.1	0.5	0.4	0.1	0.3	0.3	0.1
MagLS	404.5	410.9	290.6	143.5	47.7	12.7	9.2	14.4	5.3	4.9	1.6	0.8	0.3	0.5	0.3	0.3	0.1
MagLS+CC	340	388.9	295.4	149.7	49.3	13.6	8.8	14.5	5.2	5	1.6	0.8	0.1	0.5	0.3	0.3	0.1
SpSub	241.2	192.4	156.6	93.9	42.3	13.4	13.6	12.1	6.9	4.8	1.6	1.1	0.4	0.6	0.3	0.3	0.3
SpSubMod	412.2	363.3	297.8	117.8	33	11.5	13.4	12.2	6.9	4.5	1.6	1.1	0.4	0.5	0.3	0.3	0.3
BiMagLS	56.7	31.6	29.7	20.2	11.8	7.3	5.4	6.4	4.9	2.8	1.3	0.5	0.4	0.1	0.3	0.3	0.1

The issue of side lobes is further explored in Figure 5.5 which shows the estimated left-ear HRTF loudness across directions. The plots show ripples in the cases of Trunc, EQ and SpSub,

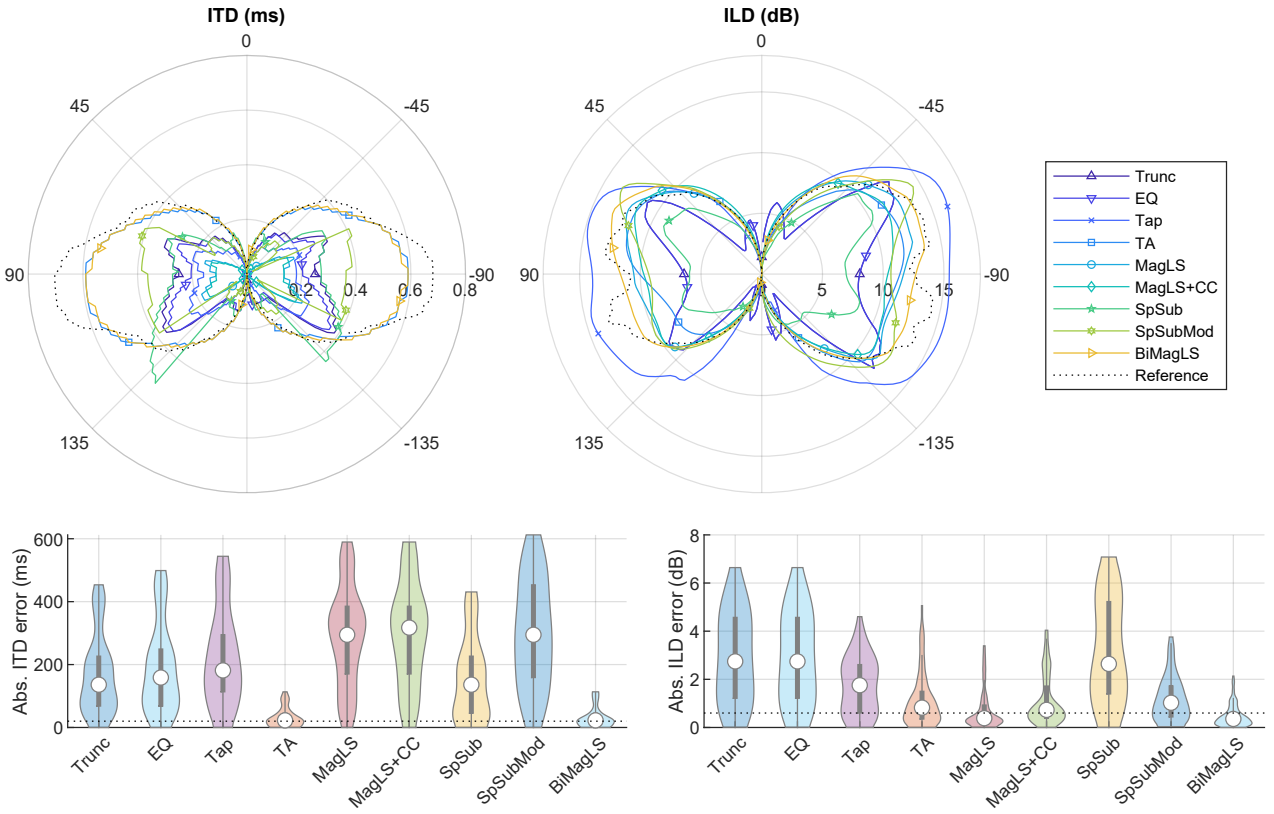


Figure 5.4: Top: ITD in ms and ILD in dB, plotted as a function of azimuth on the horizontal plane for HRTFs interpolated with the nine rendering methods and $N = 3$. Bottom: violin plots showing the absolute ITD and ILD errors for each HRTF on the horizontal plane, where the dotted lines represent the approximate just noticeable differences in anechoic conditions, according to Klockgether and van de Par (2016).

Table 5.3: Mean absolute ILD error between each method and the reference, per order (in dB).

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	4.8	3.3	2.9	2.5	2.3	2.1	1.9	1.7	1.5	1.3	0.8	0.6	0.4	0.2	0.1	0.1	0.1
EQ	4.8	3.3	2.9	2.5	2.3	2.1	1.9	1.7	1.5	1.3	0.8	0.6	0.4	0.2	0.1	0.1	0.1
Tap	4.1	2.6	1.7	1.6	2.8	2.6	2.3	2	1.9	1.7	1.4	0.6	0.3	0.1	0.1	0.1	0.1
TA	1.8	1.4	1.1	0.9	0.5	0.6	0.6	0.6	0.6	0.5	0.5	0.4	0.2	0.1	0.1	0.1	0
MagLS	1.7	1.1	0.7	0.7	0.5	0.5	0.4	0.5	0.5	0.4	0.3	0.3	0.3	0.2	0.1	0.1	0.1
MagLS+CC	2.9	1.6	1.1	0.6	0.6	0.6	0.5	0.5	0.5	0.5	0.3	0.3	0.3	0.2	0.1	0.1	0.1
SpSub	4.6	3.9	3.2	2.8	2.8	2.3	2	1.8	1.7	1.3	1	0.8	0.5	0.2	0.1	0.1	0.1
SpSubMod	2.3	1.7	1.2	1	0.9	0.7	0.6	0.5	0.4	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1
BiMagLS	1.2	1	0.5	0.5	0.4	0.4	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0

suggesting that SH-HRTFs produced with those methods have some degree of loudness instability as described by Ben-Hur, Alon, Rafaely, et al. (2019), also explaining their inaccurate ILDs. In comparison, the other methods display a loudness that is more similar to the reference, with MagLS and BiMagLS displaying the lowest error overall with 0.27 sones on average, whereas the

highest error was obtained for Trunc (1.45 sones on average).

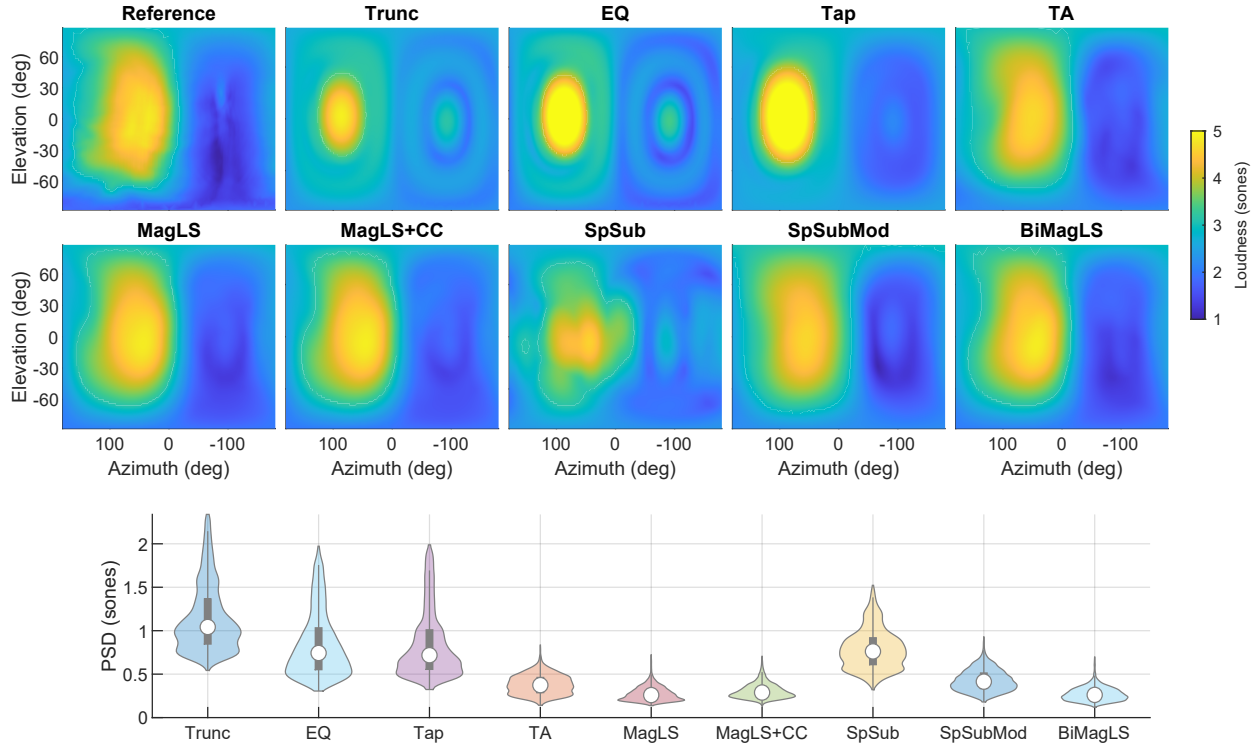


Figure 5.5: Top: Estimated loudness of the left-ear HRTF (in sones) for the 11950 available directions. The top-left plot shows the Reference (original HRTF) and the rest show HRTFs interpolated with the nine rendering methods and $N = 3$. Bottom: violin plots showing the loudness difference between each method and the reference, for all directions.

PSD is shown as a function of spatial order in Figure 5.6 and in Table 5.4. The overall trend seems to be that PSD decreases monotonically with spatial order, as expected. According to this metric, the best performer was BiMagLS, followed by MagLS, MagLS+CC, TA and SpSubMod, whereas the worst one was Trunc. Differences among methods were found to be largest for lower orders and become smaller for higher ones, falling below 0.03 sones for any pair of methods above $N = 30$. For $N \leq 6$, MagLS and BiMagLS obtained the best results, especially if compared with the methods SpSub, EQ and Tap. For $N > 6$, BiMagLS still obtained the best results, while TA performed slightly better than MagLS. It is worth noting that SpSubMod performed overall better than SpSub and also that MagLS+CC did not outperform MagLS, according to this metric.

In agreement with Ben-Hur, Alon, Mehra, et al. (2020) and Ben-Hur, Alon, Mehra, et al. (2021), TA showed overall excellent results, also considering its accurate ITD reconstruction; however, it is worth remembering that this is achieved at the expense of incompatibility with the standard

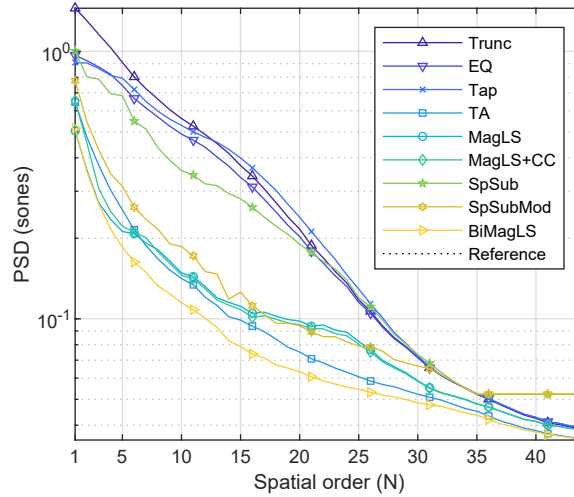


Figure 5.6: Left-ear PSD between the original HRTF and HRTFs that were interpolated with different methods and spatial orders 1 to 44 (in sones; lower is better). The reference is located at minus infinity.

Table 5.4: PSD between each method and the reference, per order (in sones).

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	1.45	1.3	1.16	1.03	0.91	0.8	0.72	0.66	0.61	0.56	0.37	0.22	0.12	0.07	0.05	0.04	0.04
EQ	0.97	0.92	0.87	0.82	0.74	0.67	0.61	0.57	0.53	0.49	0.34	0.2	0.12	0.07	0.05	0.04	0.04
Tap	0.91	0.9	0.86	0.81	0.79	0.72	0.65	0.6	0.56	0.53	0.4	0.24	0.13	0.07	0.05	0.04	0.04
TA	0.65	0.48	0.38	0.3	0.26	0.22	0.19	0.17	0.15	0.14	0.1	0.08	0.06	0.05	0.05	0.04	0.04
MagLS	0.5	0.36	0.27	0.24	0.21	0.21	0.19	0.18	0.16	0.15	0.11	0.1	0.08	0.06	0.05	0.04	0.04
MagLS+CC	0.65	0.45	0.31	0.25	0.22	0.21	0.2	0.18	0.16	0.15	0.11	0.1	0.08	0.06	0.05	0.04	0.04
SpSub	1	0.8	0.79	0.69	0.68	0.55	0.51	0.44	0.39	0.36	0.28	0.19	0.12	0.07	0.05	0.05	0.05
SpSubMod	0.78	0.54	0.43	0.35	0.31	0.26	0.24	0.22	0.19	0.19	0.13	0.09	0.08	0.07	0.05	0.05	0.05
BiMagLS	0.52	0.35	0.27	0.22	0.19	0.16	0.15	0.13	0.12	0.11	0.08	0.06	0.05	0.05	0.04	0.04	0.04

Ambisonics framework, as explained in Section 5.2. The proposed method BiMagLS performed similarly to TA in terms of ITD reconstruction but outperformed it for other metrics thanks to its smaller magnitude errors. Among the methods that are compatible with standard Ambisonics, MagLS provided the most accurate binaural signals according to the evaluated metrics.

5.4.2 Auditory models

Lateral precision, defined as the circular standard deviation of localisation estimates in the lateral dimension (Majdak et al., 2010), is shown as a function of spatial order in Figure 5.7 and in Table 5.5. Compared to the other metrics, it seems to converge quite early, with all

methods displaying an error below 2° for $N = 20$. This is likely due to the strong influence of ITDs in lateral localisation and the fact that ITDs converge at a relative early order (see Figure 5.2) due to not being much affected by high-frequency truncation errors. BiMagLS and TA showed the best performance overall, probably because of their small phase errors. Other methods performed poorly for $N < 5$, likely due to inaccurate ITDs, e.g. as reported in the numerical analysis. For $N \geq 5$, when ITDs become more accurate, all methods perform similarly well except Trunc, EQ and SpSub; this is attributed to their higher ILD errors, reported in Figure 5.4.

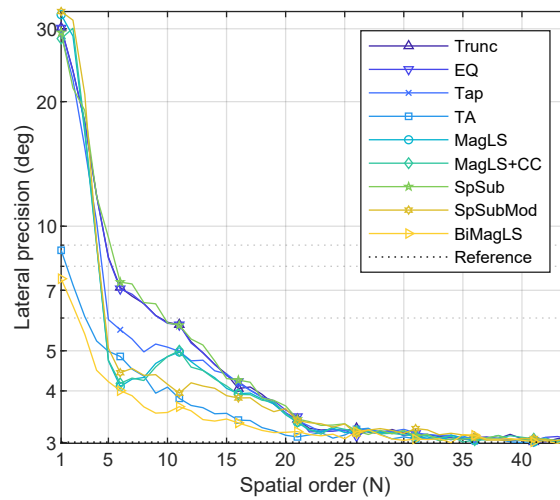


Figure 5.7: Predicted lateral localisation precision (in degrees; lower is better) for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference data (black dotted line) was obtained from the original HRTF.

Table 5.5: Lateral precision per method and order (in degrees). Reference: 3° .

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	30.2	23.7	17.9	11.9	8.4	7.1	6.8	6.5	6.1	5.9	4.4	3.5	3.2	3.2	3.1	3.1	3.1
EQ	30.1	23.7	17.8	11.9	8.4	7.1	6.9	6.5	6.1	5.8	4.4	3.6	3.2	3.1	3.1	3	3.1
Tap	29.1	22.6	15.5	10.3	5.9	5.6	5.3	5	5.2	5.1	4.4	3.5	3.2	3.2	3.1	3	3.1
TA	8.7	7.2	6	5.3	5	4.8	4.5	4.3	3.9	4.1	3.5	3.1	3.1	3.1	3.1	3	3
MagLS	32.4	28.8	16.8	9	4.7	4.1	4.3	4.3	4.7	4.8	4.1	3.5	3.2	3.2	3.1	3	3.1
MagLS+CC	28.4	30.1	17.3	9.1	4.7	4.2	4.3	4.2	4.6	4.8	4.1	3.5	3.2	3.1	3	3.1	3.1
SpSub	29.4	21.6	19	11.8	9.4	7.3	7.3	6.5	6.5	5.8	4.3	3.7	3.3	3.2	3.1	3	3.1
SpSubMod	33	31.5	20.8	9.1	5.1	4.4	4.5	4.4	4.4	4.2	3.9	3.6	3.3	3.2	3.1	3.1	3.1
BiMagLS	7.5	6.4	5.5	4.5	4.2	4	3.9	3.7	3.5	3.5	3.4	3.2	3.1	3	3.1	3	3

Polar precision, defined as the circular standard deviation of localisation estimates in the polar dimension (Majdak et al., 2010), is shown as a function of spatial order in Figure 5.8 and in

Table 5.6. In this case, errors were relatively large for all methods at low orders and converged between orders 20 and 25. BiMagLS and TA displayed the best performance in general, followed by SpSubMod, MagLS and MagLS+CC, whereas the rest showed larger errors in comparison.

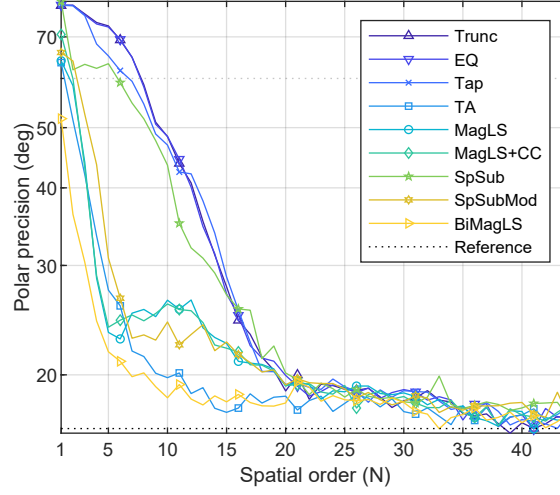


Figure 5.8: Predicted polar localisation precision (in degrees; lower is better) for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference data (black dotted line) was obtained from the original HRTF.

Table 5.6: Polar precision per method and order (in degrees). Reference: 16.39°.

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	78.6	78.7	76.2	73.8	72.9	69.2	64.9	58.6	51	48.4	27.7	18.8	19	18.2	17.3	16.7	17
EQ	78.8	78.7	75.9	73.4	72.6	69.3	64.9	58.1	50.6	48.4	27	20	19.1	18.9	16.9	17.3	17.2
Tap	78.8	78.5	75.7	68.2	65.2	61.8	59.4	54.6	48.8	46.9	28.8	19.3	19.1	18.9	17.8	16.9	17.3
TA	63.6	51.4	42	33.5	27.4	25.8	21.8	21.4	20.1	19.8	17.4	18.5	18	17.4	17.8	17.2	17.8
MagLS	64.1	58.4	43.9	28.6	23.4	22.8	25.1	25.5	24.8	26.4	22.5	19.2	18.8	18.7	17.8	17.5	16.9
MagLS+CC	70.6	59.9	43.1	28.9	23.9	24.5	25	24.4	25.1	26	22.1	19.2	19.2	18.5	17.4	17.2	17.9
SpSub	79.8	62	63.2	62.3	63.4	59.1	54.8	51.7	47.6	43.5	26.9	20.1	18.4	18.1	17.7	17.7	17
SpSubMod	66.1	64	52.5	43.6	30.8	26.6	22.9	23.2	22.7	24.3	22.6	19.3	19	18.1	17.5	18.1	17
BiMagLS	51.7	36.2	30.4	24.4	21.8	21	19.9	20.1	19.1	18.4	18.1	18	18.2	18.1	17.3	16.7	17.4

Note that lateral and polar accuracy, i.e. mean localisation error, were also assessed but no important differences among methods or spatial orders were found, so they were not reported for the sake of brevity.

Externalisation (Figure 5.9; Table 5.7) was computed as a scalar between zero (sound is perceived completely inside the head) and one (sound is perceived completely outside the head), where intermediate values denote the degree to which a sound source is well externalised. The results seemed to follow a very similar trend to those of PSD, with the methods MagLS, BiMagLS and

MagLS+CC obtaining the best performance overall, with values above 0.9 for orders as low as 3, indicating that binaural signals produced under this conditions will be perceived well outside the head. Like with PSD, the methods Trunc, EQ, Tap and SpSub displayed comparatively worse performance than the rest, approaching values of 0.5 for orders below 5, indicating that binaural signals will be perceived closer to being inside the head than outside. This similarity in trends between externalisation and PSD is attributed to the fact that the externalisation model assigns a considerable weight to monaural spectral similarity, which is highly related to the PSD metric (Baumgartner and Majdak, 2020).

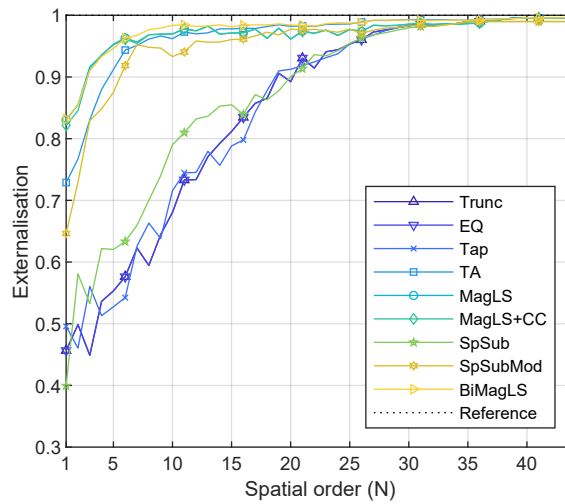


Figure 5.9: Predicted externalisation (between zero and one; higher is better) for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference is located at 1.

Table 5.7: Externalisation per method and order. Reference: 1.

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	0.46	0.5	0.45	0.54	0.55	0.58	0.62	0.59	0.64	0.68	0.81	0.89	0.95	0.98	0.99	1	1
EQ	0.46	0.5	0.45	0.54	0.55	0.58	0.62	0.59	0.64	0.68	0.81	0.89	0.95	0.98	0.99	1	1
Tap	0.5	0.46	0.56	0.51	0.53	0.54	0.63	0.66	0.64	0.72	0.79	0.91	0.95	0.98	0.99	1	1
TA	0.73	0.77	0.83	0.88	0.91	0.94	0.95	0.96	0.97	0.96	0.98	0.98	0.99	0.99	0.99	1	1
MagLS	0.83	0.85	0.92	0.93	0.95	0.96	0.96	0.97	0.97	0.97	0.97	0.96	0.98	0.99	0.99	1	1
MagLS+CC	0.82	0.85	0.91	0.94	0.95	0.96	0.95	0.97	0.97	0.97	0.97	0.96	0.98	0.99	0.99	1	1
SpSub	0.4	0.58	0.53	0.62	0.62	0.63	0.66	0.7	0.74	0.79	0.85	0.9	0.95	0.98	0.99	0.99	0.99
SpSubMod	0.65	0.73	0.83	0.85	0.88	0.92	0.95	0.95	0.95	0.93	0.96	0.98	0.98	0.98	0.99	0.99	0.99
BiMagLS	0.83	0.86	0.91	0.93	0.95	0.96	0.97	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	1	1

Finally, *SRM* (Figure 5.10; Table 5.8) also seemed to display a strong dependence on spatial order, but all methods quickly converged towards the reference as the order increased. The

methods BiMagLS and TA showed good performance at low orders, being generally within 1 dB from the reference, followed closely by MagLS and SpSubMod. On the other hand, Trunc, EQ and SpSub displayed comparatively worse performance up to $N = 15$ where all methods converge within 0.1 dB from the reference.

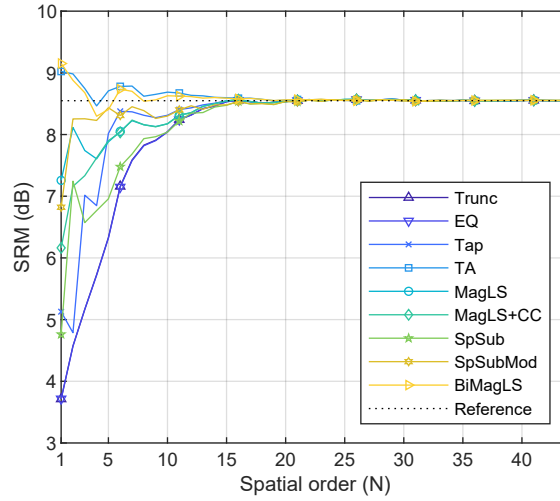


Figure 5.10: Predicted SRM (in dB), averaged for 180 masker positions, for HRTFs that were interpolated with different methods and spatial orders 1 to 44, according to the binaural model. Reference data (black dotted line) was obtained from the original HRTF.

Table 5.8: SRM per method and order (in dB). Reference: 8.6 dB.

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	44
Trunc	3.7	4.6	5.2	5.7	6.3	7.2	7.6	7.8	7.9	8	8.5	8.5	8.6	8.6	8.6	8.6	8.6
EQ	3.7	4.6	5.2	5.7	6.3	7.1	7.6	7.8	7.9	8	8.5	8.5	8.6	8.6	8.6	8.6	8.6
Tap	5.1	4.8	7	6.8	8	8.4	8.4	8.3	8.3	8.3	8.5	8.5	8.6	8.6	8.6	8.6	8.6
TA	9	9	8.7	8.5	8.7	8.8	8.8	8.6	8.7	8.7	8.6	8.6	8.6	8.5	8.6	8.6	8.6
MagLS	7.3	8.1	7.7	7.6	7.9	8.1	8.2	8.2	8.1	8.2	8.5	8.5	8.6	8.6	8.6	8.6	8.6
MagLS+CC	6.2	7.2	7.3	7.6	7.9	8	8.2	8.2	8.1	8.2	8.5	8.5	8.6	8.6	8.6	8.6	8.6
SpSub	4.8	7.2	6.6	6.8	7	7.5	7.7	7.9	8	8	8.5	8.5	8.6	8.6	8.6	8.6	8.6
SpSubMod	6.8	8.3	8.3	8.2	8.4	8.3	8.5	8.4	8.3	8.3	8.5	8.5	8.6	8.6	8.6	8.6	8.6
BiMagLS	9.2	8.9	8.7	8.3	8.4	8.7	8.7	8.5	8.6	8.6	8.6	8.6	8.6	8.5	8.6	8.6	8.6

5.5 Discussion

5.5.1 Comparing rendering methods

The binaural models' output mostly agreed with the initial analysis. For instance, magnitude interpolation errors were shown to correlate with the disruption of monaural spectral cues, loudness stability and ILDs, which translated to lower localisation precision, externalisation and speech intelligibility in the presence of maskers. As a consequence, methods that achieved smaller magnitude errors, such as MagLS, BiMagLS or TA, displayed better results according to those metrics. The same can be said about phase errors correlating to lateral precision, given that TA and BiMagLS outperformed other methods in this aspect. Similarly, increasing spatial order led to better performance, regardless of the preprocessing method.

Among methods that do not assume a time-aligned HRTF and, thus, are compatible with standard Ambisonics signals, MagLS displayed the best performance in terms of PSD and externalisation (e.g. for $N = 3$ it showed a PSD of 0.27sones, which is approximately 0.5sones lower than Tap and SpSub, and an externalisation rating of 0.92, indicating that binaural signals will be substantially more externalised than Tap or SpSub, both below 0.6). The method modified spatial subsampling (SpSubMod) also displayed relatively good results in PSD and externalisation, but was overall slightly worse than MagLS (e.g. 0.16sones and 0.09 rating, respectively, for $N = 3$). On a different note, magnitude least squares with covariance constraint (MagLS+CC) did not display superior results to MagLS, indicating that the covariance constraint did not provide an obvious benefit. However, future evaluations with reverberant sound fields may lead to different results, as MagLS+CC is expected to restore interaural coherence more accurately than other methods, which is an important feature for accurately rendering reverberant binaural signals Leclère et al., 2019. For lateral and polar precision, the best results were often disputed between MagLS, MagLS+CC and SpSubMod, depending on the spatial order, with no method being clearly superior overall. For SRM, most methods performed well since relatively low orders, with the best performance again being shared between MagLS and SpSubMod.

Overall, the data suggests that the choice of rendering method has a rather small impact on the perceived quality for spatial orders beyond 20 (perhaps smaller) but it can definitely be impactful for the lowest orders. Note that, in practice, an hypothetical threshold of $N = 20$ is well beyond what current spherical microphone arrays can capture without using multiple measurements, but could be relevant for computer which can work with an arbitrarily high order (although lower orders may still be preferred due to efficiency, as the number of channels of an Ambisonics signal increases quadratically with the order, as $M = (N + 1)^2$). Among the tested methods, MagLS performed well across the board and can be recommended as a good option for the binaural rendering of Ambisonics signals of any spatial order. For orders below 5, MagLS displayed higher ITD errors (up to five times the JND) than other methods such as SpSub, but these do not seem to have negatively impacted lateral localisation precision, according to the auditory models. Regardless, this recommendation should be validated by listening tests, e.g. comparing MagLS and SpSub in a lateral sound localisation task.

On the other hand, two of the methods (TA and BiMagLS) assumed a different rendering scenario in which the Ambisonics signal is measured bilaterally at the ears' positions, which is why they are discussed separately here. For these methods, the validity of the results is subject to the bilateral signal being properly obtained, so that phase is reconstructed accurately. Under this assumption, these two methods outperform most of the alternatives across most spatial orders, with BiMagLS being the best performing method overall for the tested metrics. This would confirm the hypothesis that BiMagLS is a direct upgrade over TA, on which it is based, due to its more accurate magnitude reconstruction, which led to better results for all metrics and spatial orders without compromising ITDs. However, a perceptual comparison of TA and BiMagLS should be performed to formally confirm that the predicted differences between the two methods are perceptually relevant.

5.5.2 Validity of the model-based assessment

The models' predictions were generally in line with results from previous perceptual experiments, namely:

1. EQ and SpSub were more similar to a reference than Trunc in terms of timbre, i.e. PSD, but not so much in terms of localisation performance for orders 3 and 6, as reported by Sheaffer and Rafaely (2014);
2. SpSubMod was more similar to a reference than SpSub for orders 1 to 3, as reported by McKenzie et al. (2019a);
3. SpSub showed more loudness stability (lower PSD, also see Figure 5.5) than Trunc for orders 2, 4 and 10, as reported by Ben-Hur, Alon, Rafaely, et al. (2019);
4. MagLS was more similar to a reference than SpSub for orders 1 to 5, as reported by Lee et al. (2019);
5. TA achieved better lateral localisation performance than MagLS for orders below 5 while being similar across other metrics, which could result in an overall more accurate rendering, as found by Ben-Hur, Alon, Mehra, et al. (2021)—note that the relatively low MagLS ratings reported in that study may have been caused by artefacts around the cutoff frequency, which here were avoided by smoothing the frequency response—;
6. TA at order 2 was more similar to a reference than Tap at order 6, as reported by Ben-Hur, Alon, Mehra, et al. (2020); and
7. MagLS, SpSub, Tap, EQ were all more similar to the reference than Trunc for orders 3, 5 and 7, as reported by Lübeck, Helmholz, et al. (2020) and Lübeck, Arend, et al. (2020).

These similarities support the argument that binaural models could be valuable as a tool for evaluations as the present one, and might be a valid alternative to real listening experiments. However, it is important to also point out the limitations of this model-based assessment. First of all, the models may not always be perfectly calibrated. For instance, the localisation model may have over- or underestimated the listener’s uncertainty, resulting in a biased estimation of localisation precision (Barumerli et al., 2020). However, even if models show some bias compared to the real world, they could still be useful for relative comparisons such as the one performed here, particularly to detect overall trends within a large set of test conditions, being much faster to run than a listening experiment.

Perhaps a more important limitation of this evaluation was the lack of dynamic listening conditions (allowing movements of sources or listener), which are possible in real listening experiments, but are not supported by current binaural models. Dynamic conditions could potentially affect the perception of externalisation (Best et al., 2020) and of the smoothness of the sound field (Engel, Henry, Amengual Garí, Robinson, and Picinali, 2021). We can get some insights by looking at Figure 5.5, which suggests that MagLS will provide a smoother rendering than Trunc, for instance. However, proper evaluation of dynamic conditions are left for future work, when appropriate auditory models become available, e.g. in the line of the work carried out by May, Ma, et al. (2015).

5.5.3 Future work

Future assessments could employ different HRTFs and include reverberant conditions, e.g. as in the work by Lübeck, Pörschmann, et al. (2020) and Engel, Henry, Amengual Garí, Robinson, and Picinali (2021). In such case, it would be interesting to look at additional binaural metrics such as interaural coherence, which has been linked to externalisation in reverberant scenarios (Leclère et al., 2019). As a consequence, perhaps MagLS+CC will outperform other methods like MagLS due to its more accurate reconstruction of interaural coherence.

More importantly, the natural next step would be to validate the models' outputs through an actual listening experiment, assessing the same perceptual metrics that were modelled in this work. Since auditory models do not typically account for cognitive processes (which can influence localisation and other metrics), a perceptual evaluation should provide more meaningful data. For such future evaluation it might not be necessary to include all test conditions such as the 44 spatial orders. Instead, it would be more efficient to employ an adaptive procedure (such as the up-down method, or perhaps one informed by artificial intelligence) with the current results as a starting point, e.g. to find the minimum spatial order at which some perceptual effect becomes apparent. This could open up an interesting avenue in auditory perception research, where not only experimental data is used to inform models, but also the other way around.

5.6 Conclusions

The present study presented a comparison of a selection of state-of-the-art binaural Ambisonics rendering methods. Results suggested that, among the reviewed methods that work with standard (not bilateral) Ambisonics signals, MagLS displayed the best results across the evaluated metrics and most of the tested spatial orders, e.g. 0.16 sones lower PSD than the second best method at $N = 3$, and can therefore be recommended for the binaural rendering of order-limited standard Ambisonics signals. Additionally, the novel BiMagLS method was proposed as an improved version of TA to render bilateral Ambisonics signals. According to the results, the proposed method outperformed TA across all metrics, e.g. 0.11 sones lower PSD and 11.6° better polar precision at $N = 3$, and can therefore be recommended for the rendering of bilateral Ambisonics signals of any spatial order. However, these recommendations are subject to further evaluations including sound fields with reverberation or spatial aliasing errors.

In the context of AAR, the outcomes of this study are helpful to decide which method should be used to binaurally render low-order Ambisonics signals (e.g. generated by a sound field simulation, as mentioned in Section 3.1). In an AAR application, the sound field is expected to change over time as the listener and sound sources move around, meaning that the room simulation must be updated in real-time to account for these changes. Therefore, the main benefit of standard Ambisonics, which is the efficient implementation of head rotations when rendering a static sound field, becomes irrelevant, since the head orientation can be considered as just another variable when re-computing the sound field. This means that bilateral Ambisonics become a valid alternative, especially considering their superior performance when working with low-order signals (Ben-Hur, Alon, Mehra, et al., 2021). The real-time constraint also means that efficiency becomes a priority, and thus it is preferable to work with low-order audio simulations, since the number of channels of an Ambisonics signal increases quadratically with the spatial order. The results of this study indicate that the method with the best performance at low spatial orders, considering both standard and bilateral categories, was BiMagLS. For $N \leq 3$, said method outperformed TA by more than 0.1 sones in PSD and 10° in polar precision, and MagLS by more than 10° in lateral precision. It is worth noting that bilateral Ambisonics

simulations required by BiMagLS (and TA) utilise two receivers rather than the single one used in standard Ambisonics, but whether this factor affects the choice of rendering method in an actual real-time implementation is a question left for future studies. Finally, further work is required to investigate the absolute thresholds in terms of the minimum truncation order per rendering method, which is discussed in Chapter 6.

Regarding the evaluation methods, it was shown that the auditory models' predictions were consistent with previous perceptual data. This makes a strong point in favour of model-based evaluations in future auditory perception research, considering that they require a fraction of the time and effort of actual listening experiments, while providing reproducible results. In practice, listening experiments will still likely be carried out, as they provide the 'ground-truth' data of auditory research, but model-based evaluations will be valuable as preliminary tests to identify 'areas of interest' (e.g. where two test conditions obtain similar results) and discard those that are likely to be uninteresting (e.g. where one condition clearly outperforms the rest).

To ensure the reproducibility of this study, all the MATLAB code required to implement the evaluated methods has been made available online at the BinauralSH repository by the present author (Engel, 2021), and the figures can be reproduced with the script `exp_engel2021` from the Auditory Modeling Toolbox (<https://amtoolbox.org/>, last viewed 28th August 2021).

Chapter 6

Binaural Reverberation: Comparing Ambisonics-Based Approaches

Reverberation, produced as the result of the interaction between a sound source and its acoustic environment, plays a crucial role in AAR. In order for virtual sources to blend with the real environment, the acoustics of the room must be simulated with enough accuracy, which can be computationally costly to do in real time. Therefore, it is relevant to investigate the perception of spatial reverberation in order to propose computationally efficient models that can help to solve this problem. A promising approach to simplify the simulation process would be to reduce the spatial resolution of the reverberant sound field through the Ambisonics framework. This chapter explores the perceptual impact of this and other simplifications for processing binaural reverberation within the Ambisonics framework.

This chapter was adapted from the article by the present author and others (Engel, Henry, Amengual Garí, Robinson, and Picinali, [2021](#)), which was peer-reviewed and published in the Journal of the Acoustical Society of America under a Creative Commons Attribution 4.0 International License (CC BY).

Abstract

Reverberation is essential for the realistic auralisation of enclosed spaces. However, it can be

computationally expensive to render with high fidelity and, in practice, simplified models are typically used to lower costs while preserving perceived quality. Ambisonics-based methods may be employed to this purpose as they allow us to render a reverberant sound field more efficiently by limiting its spatial resolution. The present study explores the perceptual impact of two simplifications of Ambisonics-based binaural reverberation that aim to improve efficiency. First, a ‘hybrid Ambisonics’ approach is proposed in which the direct sound path is generated by convolution with a spatially dense head related impulse response set, separately from reverberation. Second, the reverberant virtual loudspeaker method (RVL) is presented as a computationally efficient approach to dynamically render binaural reverberation for multiple sources with the potential limitation of inaccurately simulating listener’s head rotations. Numerical and perceptual evaluations suggest that the perceived quality of hybrid Ambisonics auralisations of two measured rooms ceased to improve beyond the third order, which is a lower threshold than what was found by previous studies in which the direct sound path was not processed separately. Additionally, reverberant virtual loudspeaker method (RVL) is shown to produce auralisations with comparable perceived quality to Ambisonics renderings.

6.1 Introduction

Digital reverberation (reverb) was first conceived by Schroeder and Logan (1961) and has undergone continuous evolution ever since (Välimäki et al., 2016). For the most part, research in this area has been driven by the music and acoustic architecture industries in which efficiency often comes second to fidelity when producing room auralisations. More recently, however, the emergence of virtual and augmented reality has increased the demand for highly realistic interactive audiovisual experiences. The real-time requirements of such applications mean that acoustic modelling must often be simplified in favour of efficiency, even more so if audio-dedicated computational resources are limited, which is likely the case in portable devices.

A common approach to simplify the computation of a reverberant sound field is to encode it in the SH domain, popularly known as Ambisonics in the context of audio production (Gerzon, 1985; Zotter and Frank, 2019). In this encoding process, a specific SH order (hereafter, spatial or Ambisonics order) may be chosen, which dictates the spatial resolution of the reproduced

sound field as well as the computational load and memory requirements (Avni et al., 2013). This can be useful for applications which demand real-time dynamic room auralisations. For instance, Schissler, Stirling, et al. (2017) describe a practical implementation which simulates sound propagation through physical and geometrical models and processes the resulting sound field in the Ambisonics domain at different spatial orders, depending on its directivity at each time instant.

Arguably, the minimum spatial order required for a binaural Ambisonics rendering is mostly dictated by the direct sound path (anechoic) from the source to the listener, rather than the reverb (including rearly reflections and late reverberation), as the former is generally more directive than the latter and, therefore, needs finer spatial resolution to be simulated accurately (Engel, Henry, Amengual Garí, Robinson, Poirier-Quinot, et al., 2019; Lübeck, Pörschmann, et al., 2020; Schissler, Stirling, et al., 2017). In this study, a ‘hybrid Ambisonics’ method is proposed where the direct path is rendered through convolution with HRIRs sampled in a dense spatial grid, while the reverb is processed in the SH domain. This contrasts with the more straightforward traditional method, here referred to as ‘standard Ambisonics’, in which the sound field is rendered as a whole in the SH domain (Zotter and Frank, 2019; Ahrens and Andersson, 2019). Even though the goal of this study is not to perform a direct comparison between hybrid and standard Ambisonics, it is expected that the former will require lower spatial order than the latter to produce renderings of similar perceived quality as suggested in a preliminary study by the present authors (Engel, Henry, Amengual Garí, Robinson, Poirier-Quinot, et al., 2019). The evaluation of the proposed method will give insight on the minimum spatial resolution required to render Ambisonics-based reverb, assuming that the direct sound path is simulated separately and with enough accuracy. This could, in turn, lead to the development of more efficient rendering methods.

At the same time, a practical limitation of Ambisonics rendering has to do with the number of sound sources that can be processed dynamically, i.e. in an interactive environment where sources or listener change their position with time. Typically, when multiple sources are dynamically rendered through either standard or hybrid Ambisonics, it is necessary to perform separate convolutions with RIRs for each source. These RIRs must be either precomputed, which can

become memory-intensive if each source-listener position pair is considered, or calculated in real time, which quickly becomes costly as the number of sources or the RIR length increase (Schissler, Mehra, et al., 2014). In practice, RIRs need not be modified to simulate listener head rotations, e.g. as shown by Noisternig et al. (2003), given that the Ambisonics sound field can be easily manipulated via rotation matrices (Zotter and Frank, 2019) but they must still be recomputed for translational movements of sources or listener. In this study, the reverberant virtual loudspeaker method (RVL) is presented as a way to binaurally render an arbitrary number of sources in a reverberant space with a relatively low computational cost while allowing for the dynamic addition and translational movement of the sources. This is achieved by making several assumptions such as the listener having low sensitivity to the directionality of reverb—for instance, Lindau, Kosanke, et al. (2012) showed that there exists a so-called mixing time, a hypothetical instant within the RIR after which reflections become numerous enough for a listener not to perceive their directional features. The main drawback of RVL is that listener’s head rotations are approximated by having the room ‘locked’ to the head as explained in more detail in Subsection 6.2.3. This allows the algorithm to be highly efficient at simulating a large number of sources, but whether and by how much such simplifications negatively affect the realism of the rendering is something that remains to be investigated.

The general goal of this study is to explore the perception of Ambisonics-based binaural reverb with aims to make recommendations for efficient rendering techniques. More concretely, two main research questions are tackled through numerical analyses and perceptual evaluation:

1. What is the perceptual impact of decreasing the spatial order of hybrid Ambisonics binaural reverberation? (experiment 1), and
2. how does RVL compare to a more accurate method in terms of subjective preference, given its approximate simulation of head rotations? (experiment 2).

In terms of computational complexity, the number of audio channels required to process a binaural Ambisonics sound field via convolution with multiple RIRs is $2M(N + 1)^2$ where M is the number of sound sources and N is the spatial order. Therefore, it is expected that

decreasing the spatial order in Ambisonics reverberation rendering could lead to exponential computational savings in the best case. On the other hand, RVL eliminates the dependence on the number of sources, providing even greater savings in cases where multiple sources are rendered simultaneously, as described the following Sections.

The rest of this chapter is structured as follows. Section 6.2 provides a literature review; Section 6.3 describes the methods, including the measurements and binaural rendering procedure; Section 6.4 presents numerical analyses of the methods under comparison; Section 6.5 describes the listening tests performed to perceptually evaluate the methods; Section 6.6 discusses the results and potential future work, and Section 6.7 summarises the findings and concludes the chapter.

6.2 Background and motivations

6.2.1 Reverb perception: a summary

Reverberation comes as a result of pairing an acoustic source with an environment. As a sound wave propagates from the source, it interacts with its surroundings, leading to reflection and diffraction. Consequently, filtered replicas of the original wavefront arrive at a receiver through different paths at distinct times. As time passes, the echo density increases as the wave continues to interact with the room, eventually resulting in a diffuse reverberant sound field. This process highly depends on the geometry of the room and the acoustic properties of the materials therein.

The effects of room acoustics on auditory perception have long been an active research topic. The precedence effect establishes that the direct sound allows the listener to localise the source, whereas later reflections are generally not perceived as separate auditory events (Wallach et al., 1949; Litovsky et al., 1999; Brown et al., 2015). However, strong specular early reflections can shift the perceived position of a source, broaden its apparent width (Olive and Toole, 1989) and modify its spectrum due to phase cancellations and subsequent comb-filtering (Bech, 1996). This can affect the perception of the actual space. For instance, Barron and Marshall (1981)

state that the timing, direction and spectrum of early lateral reflections contribute to the room envelopment. Similarly, the time delay between the direct sound and the first perceptually distinct early reflection has been shown to affect the perception of presence and environment dimensions in small rooms (Kaplanis et al., 2014) and the intimacy of concert halls (Beranek, 2008). As the temporal density of the reflections increases, perception is governed less by temporal characteristics and more by statistical properties of the reverberant tail. Research done by Yadav et al. (2013) suggests that reverberation time (RT) contributes to the perception of size most significantly in large rooms, whereas early reflections are of greater importance in small rooms. With respect to binaural rendering, it has been shown that reverb improves the externalisation of sound sources in the binaural domain, even if only early reflections are used (Begault, Wenzel, et al., 2001). Also, it has been found that the congruence between presented virtual sounds and the acoustic properties of the actual listening space contribute to the level of externalisation (Werner et al., 2016).

Based on the aforementioned research, various reverb-rendering methods which try to achieve high fidelity at reasonable costs have been proposed throughout the years. According to a comprehensive review by Valimaki et al. (2012), these methods can be generally classified in three categories: delay networks, such as feedback delay networks (Jot and Chaigne, 1991; Jot, 1997) or Schroeder reverberators (Schroeder and Logan, 1961); convolution algorithms in which a dry input signal is convolved with an omnidirectional or Ambisonics RIR; and computational acoustics, which encompass geometry-based simulations, such as the image source method (Allen and Berkley, 1979), and wave-based methods, such as the finite-difference time-domain method (Botteldooren, 1995). In practice, these categories overlap; for instance, an RIR used for convolution may be generated through computational acoustics (Pelzer et al., 2014; Schissler, Stirling, et al., 2017).

The present work focuses on convolution methods based on Ambisonics RIRs, measured with a spherical microphone array, and on BRIRs, measured on a head and torso simulator. The measurement process and rendering methods will be explained in more detail in Section 6.3.

6.2.2 Spatial order perception and hybrid Ambisonics reverb

Standard Ambisonics binaural rendering typically involves convolving a dry audio signal with an Ambisonics RIR (either measured or simulated) for each rendered sound source. Then, all the resulting Ambisonics signals may be accumulated into a single sound field, which is decoded to a binaural signal by means of a free-field HRTF and a method of choice, e.g. virtual loudspeakers (McKeag and McGrath, 1996; Bernschütz, Vázquez Giner, et al., 2014) or by convolution of the sound field and the HRTF in the SH domain (Schörkhuber et al., 2018; Zaunschirm, Schörkhuber, et al., 2018). A straightforward way of reducing the cost of this process is by decreasing the spatial order of the Ambisonics signals, but this can alter the perception of the resulting auralisations as previous studies have shown. First, Avni et al. (2013) performed listening tests with simulated room renderings of varying order, showing that listeners mainly relied on perception of spaciousness and timbre to discriminate them, with higher orders producing spatially sharper and brighter sounds. Later, Bernschütz (2016, Section 5.6.1) observed that renderings became generally indistinguishable from each other for spatial orders of 11 and above, obtaining ‘excellent’ results for orders as low as 5 in noncritical scenarios. He also reported that perceptual differences between spatial orders were more accentuated for direct sound and early reflections than they were for diffuse reverb. More recently, Ahrens and Andersson (2019) reported that order 8 was sufficient for lateral sources when compared to auralisations based on measured BRIRs, but slight spectral differences were detected up to 29th order for frontal sources in a discrimination task.

The studies above suggest that realistic standard Ambisonics auralisations may be achievable if a sufficiently high spatial order is employed. However, these may be computationally costly and their feasibility in practice is limited as commercially available microphone arrays are generally of order four and lower. A promising alternative is to render the direct sound (and, possibly, some early reflections) by convolution with a spatially dense HRIR dataset, while computing the rest of the RIR in the Ambisonics domain. The rationale is that if the direct sound path is rendered accurately, sources should still be well localised because of the precedent effect (Wallach et al., 1949; Brown et al., 2015), minimising perceptual degradation caused by spatial

order reduction. This approach is here referred to as hybrid Ambisonics and has been previously employed by Picinali et al. (2017) and Engel, Henry, Amengual Garí, Robinson, Poirier-Quinot, et al. (2019). Due to the reasons stated above, it is hypothesised that hybrid Ambisonics could potentially achieve a level of quality that is comparable to that of standard Ambisonics, without the need of employing high orders. This would in turn reduce computational requirements and the need for costly high-order microphone arrays.

Promising results have recently been reported by Lübeck, Pörschmann, et al. (2020), who showed through perceptual tests that the minimum required spatial order for early reflections and late reverb was significantly lower than it was for the direct sound path, for auralisations based on sparse BRIR grids. An important difference between that study and the present study is that they generated their sparse BRIR set by means of spatial subsampling (Bernschütz, Vázquez Giner, et al., 2014), which introduces both aliasing and truncation error in the signals (Ben-Hur, Alon, Rafaely, et al., 2019), whereas in this work, Ambisonics RIRs were directly truncated in the SH domain. Whether the findings of Lübeck, Pörschmann, et al. (2020) can be extended to the rendering of order-truncated (rather than spatially subsampled) Ambisonics sound fields is a question that the present work aims to answer.

Of course, the level of resolution needed to produce perceptually convincing reverberation depends on the relative weight of reverberation in a particular scenario. For instance, in a near-anechoic room, reverberation is likely to have a much smaller perceptual weight than in a space filled with reflective surfaces. It is therefore reasonable to expect that the findings of this work will show dependency on the characteristics of the rendered room, not to mention other factors such as the proximity of the listener to a wall or to the source itself. This issue will be further addressed when discussing the numerical analyses and the results of the perceptual evaluations, particularly looking at the direct to reverberant ratio (DRR), i.e. the ratio between direct and reverberant sound energy, which is expected to be a highly important parameter.

6.2.3 Multiple-source rendering and RVL

In both standard and hybrid Ambisonics renderings, the cost of the convolution stage increases (at least) linearly with the number of simulated sources as the dry audio signal of each source must typically be convolved with a separate Ambisonics RIR (Schissler, Stirling, et al., 2017). In a low-cost scenario, this can limit the number of sources that can be rendered dynamically, e.g. allowing the addition of a new source or the changing of a source’s position in real time. RVL—previously used by Picinali et al. (2017) and Engel, Henry, Amengual Garí, Robinson, Poirier-Quinot, et al. (2019) and natively implemented by the 3D Tune-In Toolkit (Cuevas-Rodríguez et al., 2019)—is proposed here as an alternative computationally efficient approach to dynamically render multiple sources. Its main feature is that the number of convolutions needed to produce a reverberant sound field is independent of the number of rendered sources.

RVL is inspired by the classic virtual Ambisonics approach first outlined by McKeag and McGrath (1996) and later used by Noisternig et al. (2003). In the original method, one or more anechoic sound sources are encoded in an Ambisonics sound field, which is then decoded to a virtual loudspeaker grid distributed around the listener, and the resulting signals are finally convolved with the corresponding HRIRs to produce the binaural output. To implement reverb, Noisternig et al. (2003) suggested computing early reflections as additional sources and late reverb through a delay network. In RVL, the procedure is analogous to anechoic virtual Ambisonics except that BRIRs are used in the place of HRIRs, effectively integrating the room acoustics in the binaural rendering. Also, the direct sound path is rendered separately from the reverb through convolution with discrete HRIRs (like in hybrid Ambisonics), as will be explained in Section 6.3.

Even though RVL is devised as a binaural reproduction method, it might be helpful to provide an intuition of the algorithm in a loudspeaker reproduction context. In essence, RVL is equivalent to reproducing a number of anechoic sound sources over an Ambisonics loudspeaker setup within a reverberant room, rather than in an anechoic space. This means that the reverberant characteristics of the listening room will be integrated in the sound field, even though the reflections therein do not coincide with those that would be generated by an accurate room

acoustics simulation.

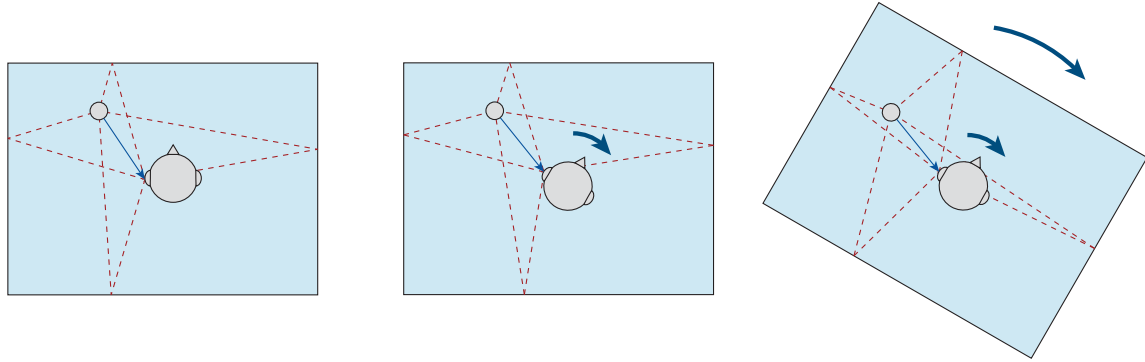


Figure 6.1: Direct sound path and first-order early reflections as they reach the left ear of a listener in three scenarios: (left) before any head rotation; (middle) canonical rendering after a head rotation of 30 degrees clockwise; and (right) RVL rendering after the same head rotation. Note how in (c) the direct sound path is accurate, whereas the room is head-locked, affecting the incoming direction of reflections.

Because the convolutions with the BRIRs happen at the Ambisonics decoding step, once all sources have been blended into a single sound field, the number of required real-time convolutions is always equal to the number of Ambisonics channels multiplied by two (to account for both ear signals), resulting in $2(N + 1)^2$, where N is the Ambisonics order, independently of the number of sources. This feature makes RVL highly efficient at dynamically rendering multiple sources as shown in Figure 6.2. Also, it allows for the simulation of virtual sound sources at any position in a sphere around the listener from a reduced set of measured BRIRs, e.g. six BRIRs for first order. Therefore, it requires fewer measurements and memory usage than traditional convolution-based methods, such as standard Ambisonics, which need separate RIRs or BRIRs for every possible source-receiver pair location. The main limitation of RVL is that although the relative position of the sound sources can be changed in the Ambisonics domain, the room is head-locked due to the set of BRIRs being fixed. This means that a rotation of the listener's head is simulated by translating all sound sources in the opposite direction, which may produce inaccurate reflections as depicted in Figure 6.1. The assumption of RVL is that such approximations will not be noticeable by listeners, e.g. directionality of late reverb may not generally be perceived by listeners, according to Lindau, Kosanke, et al. (2012) or, at least, will not lead to implausible renderings (Lindau and Weinzierl, 2012).

A previous study by Picinali et al. (2017) evaluated the perceived quality of RVL auralisations

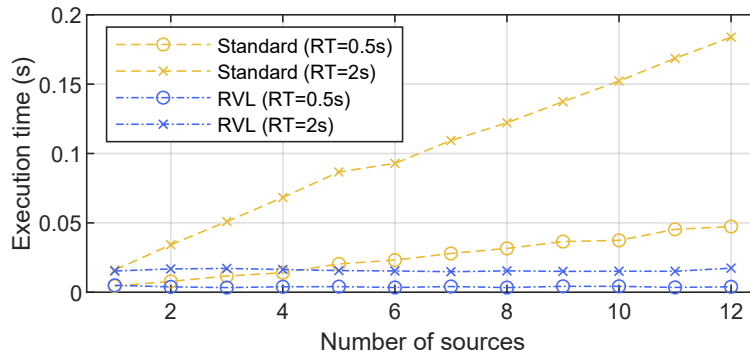


Figure 6.2: Comparison between the average execution time of the convolution stage in standard Ambisonics and RVL binaural rendering, as a function of the number of rendered sources, for two different RTs. A random input signal with a length of 1024 samples was used as input. Simulations were done in MATLAB using the overlap-add method (Oppenheim et al., 2001), running on a quad-core processor at 2.8 GHz.

of different spatial orders while the direct sound path was rendered identically for all conditions. The results suggested that first-order RVL was able to produce room auralisations which were indistinguishable from higher order simulations for a particular room geometry. This study was web-based and had some limitations, namely, that it was carried out with uncontrolled hardware in an uncontrolled environment and did not implement head tracking, thus, the perceptual effect of approximated head rotations could not be evaluated. Furthermore, renderings were simulation based instead of measurement based, only one room with a particular direct to reverberant ratio (DRR) (not reported) was tested and it lacked a comparison with a benchmark method, all of which limited the scope of the findings.

6.2.4 Contributions

The goal of this study is to explore the perceptual effect of applying practical simplifications to binaural Ambisonics reverb—first, by investigating the perceptual impact of varying spatial order on the proposed hybrid Ambisonics approach, which is expected to be smaller than the impact on standard Ambisonics that has been reported in previous studies (Avni et al., 2013; Bernschütz, 2016; Ahrens and Andersson, 2019) and, second, by comparing the proposed computationally efficient RVL to a more accurate approach in terms of subjective preference. Two different rooms were measured, and dynamic binaural renderings were produced through

hybrid Ambisonics up to order four, and through first-order RVL. The renderings were compared through numerical analyses and perceptual evaluations.

6.3 Methods

This section describes how binaural signals were produced with both hybrid Ambisonics and RVL. First, the procedure to generate RIRs for either method is detailed. This is followed by a description of the binaural rendering process for either method. Finally, a description of the audio material employed in the experiments is provided. It is important to emphasise that the direct sound path was identically rendered across all hybrid Ambisonics and RVL conditions, hence, the output binaural signals differed only in the reverb. Non-individualised HRIRs and BRIRs from a KEMAR head and torso simulator were employed throughout. It is worth noting that, even though RVL was first described in detail in this work, it was originally developed and presented by in the preliminary study by Picinali et al. (2017).

6.3.1 Measurements and RIR generation

Two rooms were measured as shown in Figure 6.3. The first room (*library*) was a large, open space with a carpeted floor, high ceilings and furniture, including chairs, desks, and bookshelves. The second space (*trapezoid*) was a small meeting room also with a carpeted floor, with four slightly asymmetrical walls (two of them made of glass), and with no furniture except for some chairs. Acoustic measurements of the rooms are reported in Table 6.1.

In each room, three RIRs were measured using the sine sweep technique (Farina, 2007), with the receiver placed at the centre of the room and sources at relative azimuths $\varphi = [-30^\circ, 0^\circ, 30^\circ]$, an elevation of $\theta = 0^\circ$ and a distance of $r = 1.2$ m (trapezoid) or $r = 1.5$ m (library). A 32-capsule fourth-order spherical microphone array (Eigenmike, mh acoustics, Summit, NJ, United States) acted as receiver and a Genelec 8030 loudspeaker acted as the source. Note the latter is a two-way non-coaxial loudspeaker, meaning that the source elevation was frequency-dependent,



Figure 6.3: The two measured rooms. (Left) The library is shown during a BRIR measurement. (Right) The trapezoid is shown during a RIR measurement.

Table 6.1: Acoustic parameters of the two measured rooms, including RT per octave band, EDT per octave band and broadband DRR, calculated according to Zahorik (2002).

	f (Hz)	250	500	1000	2000	4000	8000
Library	RT (s)	1.47	1.35	1.16	0.98	0.73	0.52
	EDT (s)	1.21	1.11	1.08	0.57	0.37	0.22
	DRR (dB)	10.09					
Trapezoid	RT (s)	0.78	0.63	0.53	0.48	0.49	0.46
	EDT (s)	0.70	0.49	0.43	0.36	0.34	0.28
	DRR (dB)	4.36					

with high frequencies being approximately 5° higher up than low frequencies. However, this was assumed to be perceptually irrelevant for the purpose of this study, since it did not affect the direct sound path (it was obtained separately, from an HRIR) and given that elevation perception was already distorted by using a non-individual HRTF (Begault, Wenzel, et al., 2001). From the 32-channel RIRs, zeroth- to fourth-order Ambisonics RIRs were generated using the Eigenstudio software package (mh acoustics). RIRs of orders 0–3 were obtained by truncating the fourth-order signals. According to the manufacturer specifications, EQ was applied such that all Ambisonics channels had a nominally flat magnitude response up to the spatial Nyquist frequency (approximately 8 kHz) and down to the lowest operating frequency that ensures a reasonable amount of system self-noise of each Ambisonics channel during the encoding process, namely, 30 Hz for orders zero and one, 400 Hz for order two, 1 kHz for order three and 1.8 kHz for order four (mh acoustics, 2016).

For hybrid Ambisonics renderings, all RIRs had the direct sound path removed by replacing

the first 4.32 ms (trapezoid) or 3.88 ms (library) after the onset with silence and applying a Hanning window for the RIR fade-in. This time was calculated analytically by subtracting the propagation time of the direct sound path from that of the first reflection minus a safety window of 30 samples (0.68 ms). Finally, RIRs were windowed at the corresponding RT and applied a de-noising procedure to remove the noise floor (Cabrera et al., 2011).

For RVL renderings, BRIRs were measured with a KEMAR head and torso simulator from six directions (front, back, left, right, up, down) using the same loudspeaker at the same distance as in the RIR measurements and applied identical post-processing, i.e. removing direct sound, windowing and de-noising. Additionally, frontal BRIRs were used as a reference to equalise Ambisonics RIRs with a series of second-order filters, similarly to Ahrens and Andersson (2019), with the goal of minimising spectral error due to the spatial order limitation of the signal (Avni et al., 2013).

6.3.2 Binaural rendering

For hybrid Ambisonics, head-tracked binaural renderings were generated in real time as follows. The direct sound path was rendered by convolving each source’s dry audio signal with an HRIR generated through barycentric interpolation from the three closest available directions, selected from a set of 8802 HRIRs measured on a KEMAR head and torso simulator (Cal Armstrong et al., 2018). HRIRs were aligned prior to interpolation, and ITDs were restored assuming a nominal head radius of 8.8 cm. Reverb was rendered in the Ambisonics domain at spatial orders 0–4 using the virtual loudspeaker approach (McKeag and McGrath, 1996), which is equivalent to the spatial subsampling method (see Section 5.2). To do so, the measured Ambisonics RIRs (with the direct sound removed) were convolved offline with the dry audio signals and then decoded to loudspeaker signals using a sampling decoder (Zotter and Frank, 2019, Section 4.9.1). Depending on the spatial order N , an appropriate number of virtual loudspeakers $M_N \geq (N+1)^2$ was used and placed at the vertices of a platonic solid (regular and convex polyhedron) when possible: octahedron ($M_{0,1} = 6$), icosahedron ($M_2 = 12$) and dodecahedron ($M_3 = 20$). Because no platonic solid exists with 25 or more vertices, a quasi-regular pentakis-dodecahedral layout

($M_4 = 32$), which is the same one used for the capsule placement on the Eigenmike microphone array, was used for $N = 4$. Finally, the virtual loudspeaker signals were convolved with the corresponding interpolated HRIRs.

For RVL, head-tracked binaural renderings were also generated in real time as detailed in Subsection 6.2.3. The direct sound path was rendered identically to hybrid Ambisonics. The reverb was generated by encoding all sources' dry audio signals in a single first-order Ambisonics sound field, which was then decoded to six virtual loudspeaker signals using a sampling decoder and a octahedral grid (same as first-order hybrid Ambisonics), and these were finally convolved with the six measured KEMAR BRIRs (also without the direct sound).

The gain of the reverberant sound fields was adjusted offline so that the binaural renderings' DRR matched that of the frontal KEMAR recordings. The 3D Tune-In Toolkit (Cuevas-Rodríguez et al., 2019) was used as the spatial audio engine, taking care of HRIR interpolation and real-time convolutions. Also, it enabled head tracking by means of an inertial measurement unit (IMU)-based tracker (EdTracker Pro Wireless, Wokingham, United Kingdom). Informal tests showed that the end-to-end tracking latency was low enough to not be noticeable during the perceptual evaluation.

6.3.3 Audio material

Two different types of audio material were used in the perceptual evaluation, each being an auditory scene comprising one or more spatialised sound sources. Source positions were chosen after a pilot study in which they were verified to provide good degree of perceived spatial separation and externalisation. All sources were presented at a relative elevation of $\theta = 0^\circ$ and a distance of 1.5 m (library) or 1.2 m (trapezoid):

1. *Music*: a performance of 'Take Five' by Paul Desmond consisting of dry recordings of piano, drum kit, and saxophone, spatialised as three different sound sources at azimuths $\varphi = [-30^\circ, 0^\circ, 30^\circ]$, respectively. The audio tracks were recorded separately in near-anechoic conditions and had a length of 47 s.

2. *Speech*: dry recording of a single female speaker (Hansen and Munch, 1991) at azimuth $\varphi = -30^\circ$. The audio track had a length of 47 s.

6.4 Numerical Analyses

This section contains numerical analyses of the signals used for the perceptual evaluation. First, a descriptive analysis of the Ambisonics RIRs is performed. Then, the effect of changing the spatial order is evaluated on the synthesised BRIRs through different metrics. Finally, the effects of rendering reverb statically are explored.

6.4.1 Descriptive analysis of Ambisonics RIR

Figure 6.4 illustrates the differences in spatial structure of Ambisonics RIRs of both rooms when rendered at different spatial orders. The time axis is split in three segments that will be simply referred to as *direct sound* or HRIR ($0 < t < \tau_{\text{dir}}$), *early reflections* ($\tau_{\text{dir}} < t < \tau_{\text{mix}}$) and *late reverberation* ($t > \tau_{\text{mix}}$), following typical room acoustics nomenclature. For simplicity, early reflections and late reverb as a whole may also be referred to simply as ‘reverb’. The time instant τ_{dir} separates the direct sound and reverb and was calculated analytically as 3.88 ms for the library and 4.32 ms for the trapezoid as mentioned in Section 6.3. Note that the direct sound is represented by an approximate spatial delta in Figure 6.4, indicating that it was processed through convolution with an HRIR instead of along with the Ambisonics RIR. The mixing time (separation between early reflections and late reverb) was defined at $\tau_{\text{mix}} = 40$ ms, according to Olive and Toole (1989) and considering the RT of the rooms. This time approximately coincides with the precedence effect threshold for speech and music (Wallach et al., 1949; Moore, 2012), meaning that reflections arriving before then are likely not to be perceived as separate auditory events but to shift the perception of the leading stimulus in terms of the colouration and source width (Olive and Toole, 1989; Bech, 1996). Note that τ_{mix} does not intend to follow the more rigorous definition of perceptual mixing time from Lindau, Kosanke, et al. (2010), but this was not critical to the experiment as it was just used for visualisation purposes.

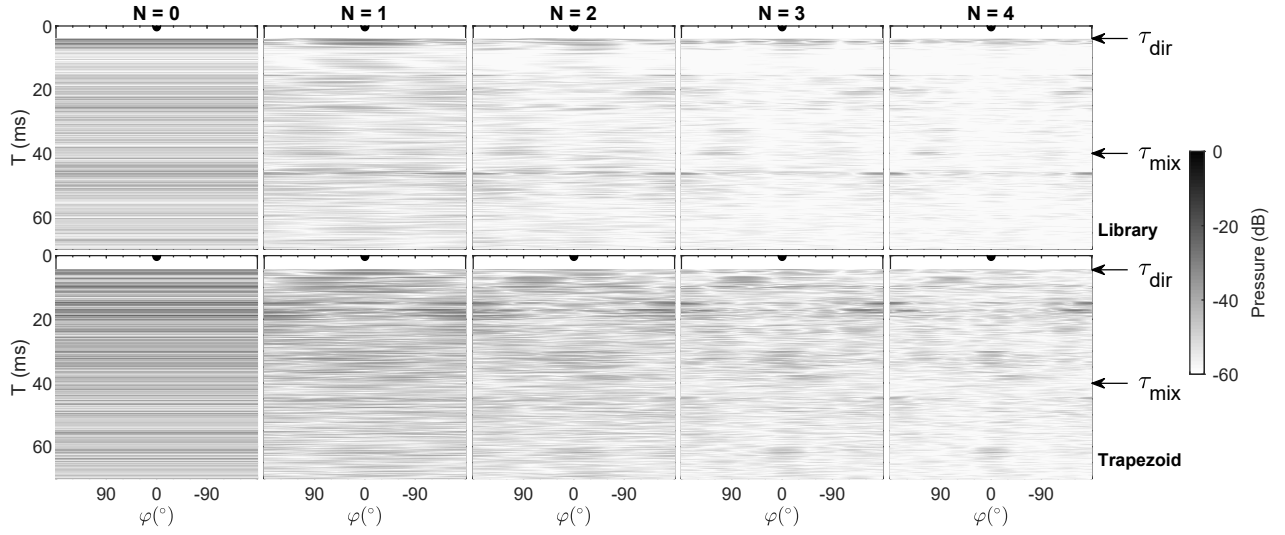


Figure 6.4: Spatial RIR of zeroth to fourth spatial orders (N) for a source placed in front of a listener in the library (top) and trapezoid (bottom), as a function of time and azimuth. The division between direct sound and reverb (τ_{dir}) and the perceptual mixing time (τ_{mix}) are indicated. The direct sound ($t < \tau_{\text{dir}}$) is represented by a discrete spatial delta at ($t = 0\text{ s}, \varphi = 0^\circ$) to indicate that it was excluded from the Ambisonics rendering and generated through convolution with an HRIR.

Given that the direct sound was unchanged throughout the different conditions, the early reflections constitute the segment of the RIR where spatial resolution is most critical. By observing Figure 6.4, it is evident that reflections become more diffuse in the azimuth axis, i.e. less directional, at lower spatial orders, with $N = 0$ being the extreme case in which the signal becomes isotropic. It can also be seen how individual reflections are generally less salient in the library than in the trapezoid, which may lead to a lower requirement in terms of spatial order.

6.4.2 Objective binaural metrics

BRIRs were synthesised for zeroth- to fourth-order hybrid Ambisonics and first-order RVL. This was done for a source placed at ($\varphi = 30^\circ, \theta = 0^\circ$) in both rooms. These BRIRs were analysed to quantify the expected perceived quality of each condition. First, the IACC was calculated, which is an objective metric commonly used to predict spatial perception from binaural content (Okano et al., 1998; Beranek, 2008; Nowak and Klockgether, 2017). As a rule of thumb and per the aforementioned studies, a lower IACC often translates to higher perceived spatial quality. Figure 6.5 (left) shows how the IACC of the direct sound is similar across all conditions, which

was expected given that the HRIR was not modified. For early reflections and late reverb, differences become larger, with zeroth-order renderings showing the highest IACC. This was also expected given that they contained essentially isotropic reverb which produced highly correlated binaural signals. For the rest of the spatial orders, differences seem to increase above 1 kHz, with higher orders generally showing lower IACCs. Unexpectedly, RVL reverb mostly obtained lower values than did the hybrid Ambisonics conditions, including the higher order ones.

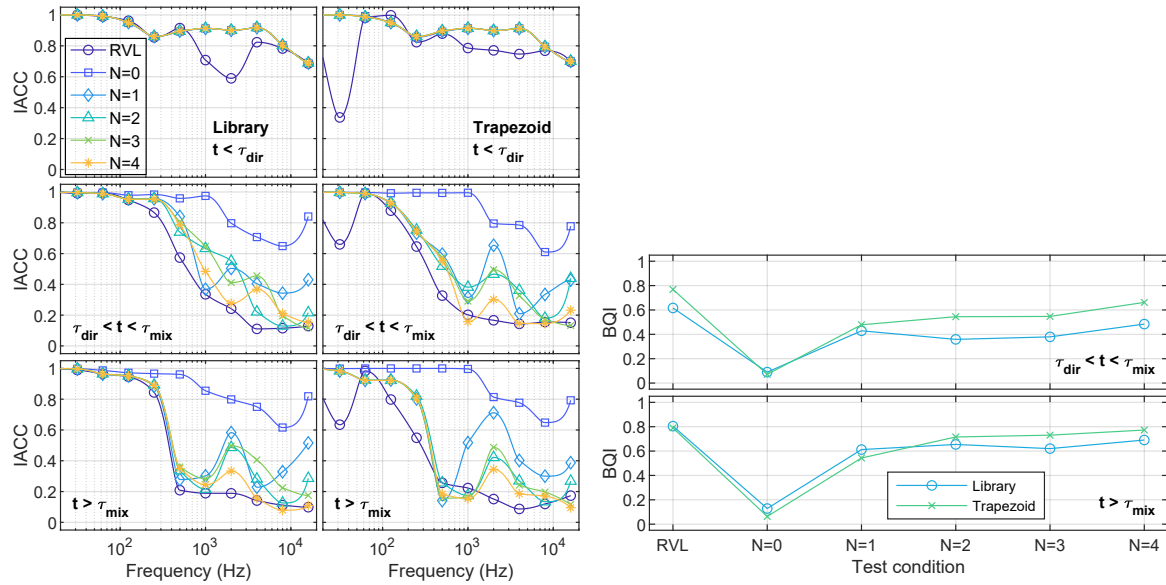


Figure 6.5: Objective binaural metrics of left-ear BRIRs at $(\varphi = 30^\circ, \theta = 0^\circ)$ for different rooms and test conditions. (Left) The IACC per octave band, separated per BRIR segment. (Right) The BQI for early reflections ($\tau_{dir} < t < \tau_{mix}$) and late reverb ($t > \tau_{mix}$).

From the IACC, other more easily interpretable metrics may be derived. One which is typically used in room acoustics studies to estimate room spatial quality is the binaural quality index (BQI), which was shown by Beranek (2008) to be correlated to subjective acoustical quality of concert halls as rated by listeners. According to Nowak and Klockgether (2017), the BQI may be calculated as

$$\text{BQI} = 1 - \frac{\text{IACC}_{500} + \text{IACC}_{1000} + \text{IACC}_{2000}}{3}, \quad (6.1)$$

where IACC_c represents the IACC for the octave band centred at $f = c$. It is evident from Figure 6.5 (right) that the early reflections obtained lower overall BQI values than did the late reverb and also showed more variance across conditions, supporting the idea that they

are the more perceptually critical part of the RIR. These results are consistent with previous studies, which showed higher BQI for late reverb with values close to 0.8 for the best performing conditions (Nowak and Klockgether, 2017). Consistently with the IACC analysis, the zeroth-order BRIR obtained the lowest BQI values, therefore, predicting a low perceived spatial quality. As expected, higher orders produced higher BQI values, although slight differences were observed between the rooms. Whereas in the trapezoid the trend was preserved until order 4, the BQI in the library seemed to plateau between orders one and two. When comparing to previous studies, the range of early BQI values for $N \geq 1$ seems to be lower here (0.22) than what was reported (0.5) by Nowak and Klockgether (2017), which may be explained by the fact that the direct sound path was removed here. Finally, RVL obtained the highest BQI values overall for both rooms, predicting a higher spatial quality that, again, was unexpected given the lower complexity of the method when compared to higher order hybrid Ambisonics.

6.4.3 Spectral analysis

Next, spectral differences across the synthesised BRIRs were explored by convolving them with test signals (speech audio material and drum kit track from music audio material) and analysing the long-term averaged spectra of the results. Spectra were calculated as the average power spectral density obtained from a series of overlapping 4096-sample discrete Fourier transforms (DFTs) after applying 1/3-octave Gaussian smoothing and are shown in Figure 6.6 (left). Note that frame-based comparisons were not considered to keep the analysis simple and because the spectra were not found to display important variations over time. The absolute deviation of each condition from a reference averaged across 42 ERBs is shown in Figure 6.6 (right). The results were similar for the left and right channels of the BRIRs, therefore, only the former are presented for brevity. The condition $N = 4$ was chosen as the reference because it has the highest available spatial order. For the hybrid Ambisonics conditions, it can be seen how the differences are largest for $N = 0$ and decrease for higher orders as expected. In the case of RVL, deviations are clearly larger in the trapezoid than in the library, which may be related to its limitations in accuracy when rendering early reflections. Overall, the range of spectral

differences was observed to be larger for the trapezoid (up to 2.4 dB) than it was for the library (up to 0.8 dB), which might be explained by the lower DRR of the trapezoid. The type of audio material did not seem to have a noticeable effect on the results.

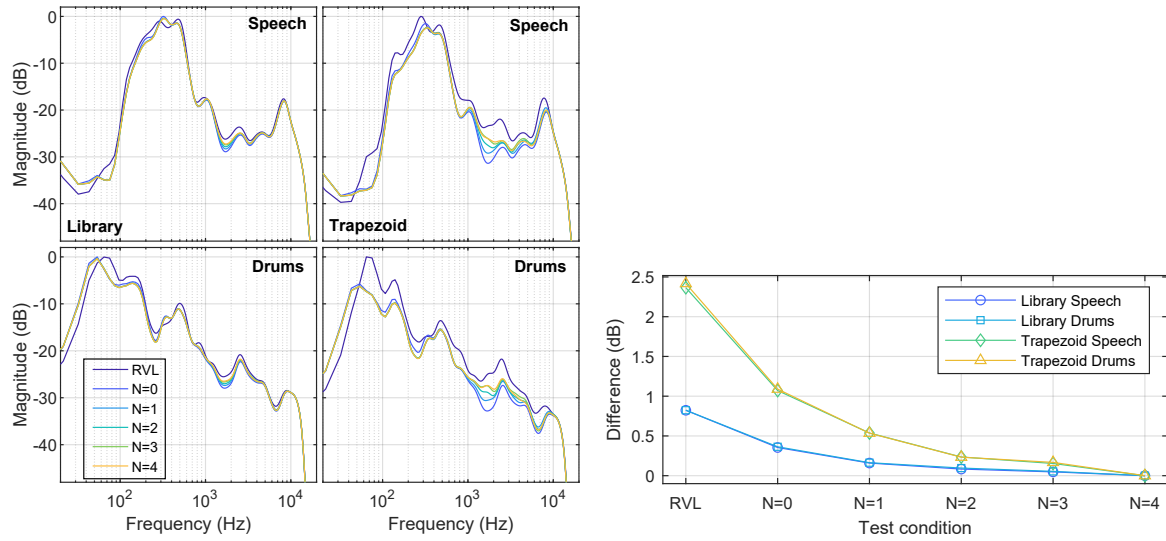


Figure 6.6: (Left) Long-term average spectra of two test signals (speech and drums) convolved with left-ear BRIRs, corresponding to the different test conditions, for a source at ($\varphi = 30^\circ$, $\theta = 0^\circ$). (Right) The absolute difference between each spectrum and the reference ($N = 4$), averaged across 42 ERBs.

6.4.4 Loudness stability

As mentioned in Subsection 6.2.3, one of the limitations of RVL is the way it approximates head rotations by having the room rotate with the listener’s head, which may have perceptual implications. On the one hand, if a strong reflection is perceived as coming from the wrong direction, it may lead to decreased externalisation, which is generally not desired. On the other hand, the loudness of the auditory scene may change more smoothly across head orientations, which might be perceived as preferable. This is particularly relevant for low-order Ambisonics renderings, which suffer from poor loudness stability across head orientations (Ben-Hur, Alon, Rafaely, et al., 2019). Figure 6.7 shows the predicted loudness (ITU-R, 2015b) of the reverberant sound field in first-order hybrid Ambisonics and RVL renderings. Additionally, a ‘static’ version of the first-order hybrid Ambisonics rendering for which reverb was not updated with head rotations was also evaluated. Loudness was estimated after convolving pulsed pink noise with BRIRs of different directions across the horizontal plane. It can be seen how the static condition

(indicated with an ‘S’) shows constant loudness across the different azimuth angles, whereas the dynamic condition is less smooth. RVL, which renders reverb in a semi-static way in which the room is head-locked, falls somewhere in the middle of the other two as it is smoother than the dynamic rendering but not constant like the static rendering.

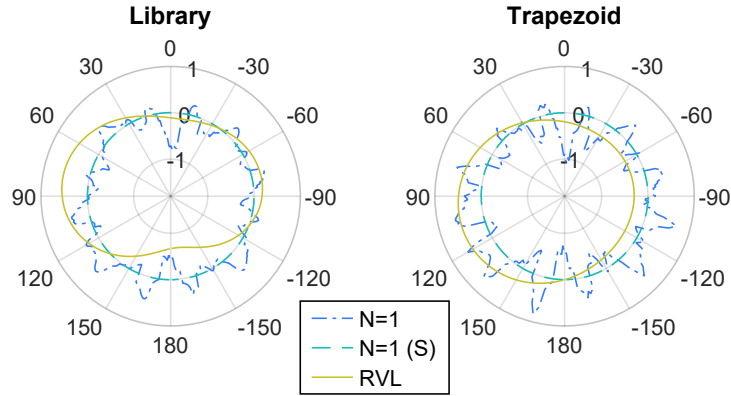


Figure 6.7: Loudness, K-weighted, relative to full scale (in dB) of the reverberant sound field, generated by convolving pulsed pink noise with a BRIR that had the direct sound removed, for different listener head orientations over the horizontal plane. Each line represents a different condition of experiment 2: first-order Hybrid Ambisonics, first-order Hybrid Ambisonics (static) and first-order RVL.

6.4.5 Summary of numerical analyses

An objective evaluation of the reverb-rendering methods being tested has been presented. First, an overview of the RIR characteristics for each room was given through descriptive analysis. Then, different metrics of the synthesised BRIRs were analysed to try to predict their perceived quality. It was observed that the BQI seemed to saturate at an earlier spatial order in the library (it did not vary much for $N \geq 1$) than it did in the trapezoid (it increased monotonically up to $N = 4$). Similarly, spectral differences with respect to the reference ($N = 4$) decreased as the spatial order increased as expected but were smaller for the library (all conditions under 1 dB) than they were for the trapezoid (under 1 dB only for $N \geq 1$) and not much affected by the type of audio material. These results suggest that the room characteristics, such as the DRR, will influence the minimum spatial order needed to achieve a certain subjective quality. Once paired with a perceptual evaluation, these observations may help to identify which objective metrics are more useful to predict the perceived quality of the binaural reverb.

RVL was found more challenging to evaluate objectively against the other conditions due its different nature, i.e. its renderings were based on measured BRIRs, whereas the hybrid Ambisonics renderings were built from Eigenmike measurements, which led to larger spectral deviations from the reference than other conditions. According to BQI data, BRIRs generated with this method were predicted to have a higher spatial quality than for even the highest order hybrid Ambisonics conditions discussed above, which was an unexpected result. Also, it was observed that one of its potential limitations, namely, the way in which head rotations are implemented, caused the rendered sound scene to have smoother loudness variations across different head orientations than did first-order hybrid Ambisonics. However, it is still not clear how this will impact the perceived quality, which should, therefore, be assessed through a perceptual evaluation.

6.5 Perceptual evaluation

The methods under study were perceptually evaluated through two separate experiments in which a total of 32 listeners participated voluntarily. The mean listener age was 32 years old [standard deviation (SD) = 9.6 yr]. Of the 32 listeners, 24 declared to have previous experience in similar listening tests, 31 declared to have no hearing impairments (the remaining one was excluded from both experiments in post-screening), 30 declared to have prior knowledge of highly realistic or binaural audio reproduction, and 18 declared to possess advanced musical knowledge or to have received formal musical education. Listeners were split in two groups: the first one (21 listeners) performed the experiment in an acoustically dampened room with a broadband RT below 0.1 s ('laboratory'), whereas the second one (11 listeners) did so in situ, i.e. in the actual measured rooms (library and trapezoid). The reason for this split was to evaluate the potential 'room divergence' effect or how the listeners' exposure to the actual room acoustics affects their perception of a virtual rendering of the same room (Werner et al., 2016).

6.5.1 Experiment 1: Paradigm

In the first experiment, listeners were asked to rate the quality of sound scenes rendered through hybrid Ambisonics at different spatial orders ($0 \leq N \leq 4$) with the direct sound being rendered through HRIR convolution identically for all conditions, as explained in Section 6.3. A double-blind listening test paradigm was used, based on the MUSHRA by the ITU-R (2015a). The highest order rendering ($N = 4$) was used as the reference and a dry rendering (without reverb) was used as a low-quality anchor. Listeners were asked to rate the similarity of each stimulus to the reference on a scale from 0 to 100, where the latter meant ‘identical to the reference.’ The user interface was implemented in Max 7. Listeners were encouraged to use head movements to explore the scene. Each listener completed one trial per combination of room (library or trapezoid) and type of audio material for a total of four trials. Post-screening was applied to exclude ratings of unreliable listeners from the data analysis, i.e. those who rated the hidden reference lower than 90 points or the anchor higher than 50 points for more than 25% of the trials.

6.5.2 Experiment 1: Results

Results of the first experiment are shown in Figure 6.8. Data for 11 listeners (5 in situ, 6 laboratory) were excluded in post-screening according to the criterion specified in the previous subsection. The relatively high amount of post-screened listeners suggests that the experiment was relatively challenging for the listeners. Listeners took an average time of 186.44 s (SD = 141.48 s) to complete each MUSHRA trial. The displayed data were normalised so that the highest rating of every trial is set to 100 and the lowest rating is set to 0. Note that this was done for the sake of visualisation and all inferential analysis was performed on non-normalised data. Descriptive analysis shows that the ratings were generally higher for higher spatial orders. Dry renderings consistently obtained the lowest ratings, followed by zeroth- and first-order renderings. Mean and median ratings seem to be similar across the higher order conditions, placed close to the top of the rating scale.

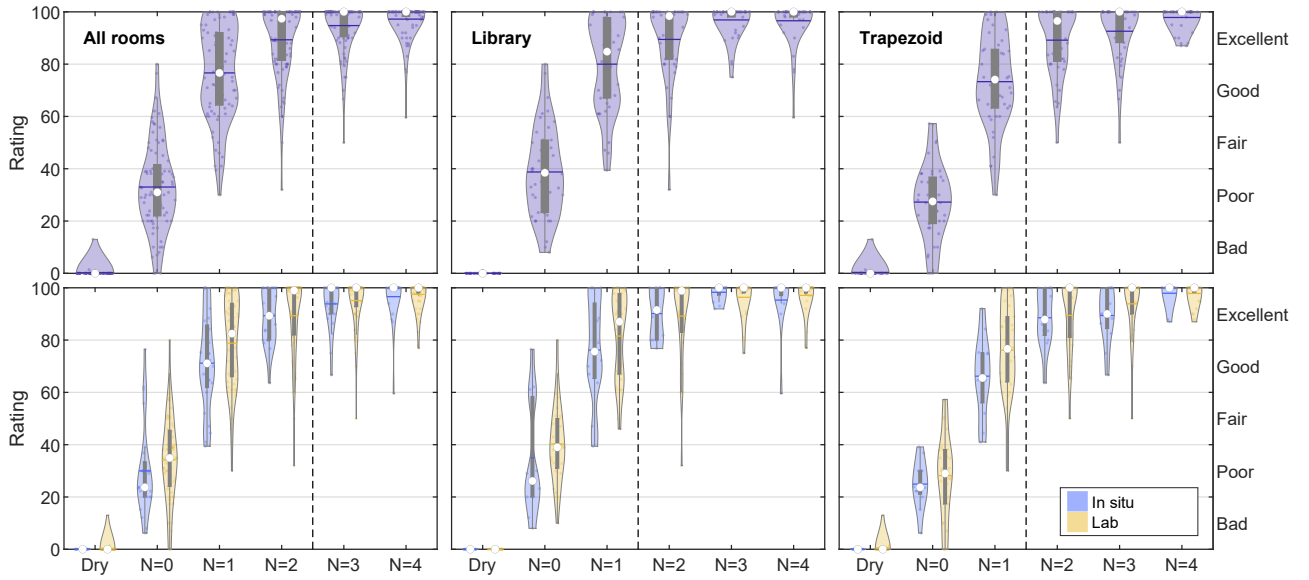


Figure 6.8: Results from experiment 1 represented by violin plots (Hintze and R. D. Nelson, 1998), which show the probability density of the data, median (circle), interquartile range (box), and mean (horizontal line). (Top) Both test locations pooled together. (Bottom) Separated per test location (from left to right, ‘in situ’ and ‘lab’). From left to right, both rooms pooled together, library and trapezoid data are shown. For this visualisation, data were scaled on a per-trial basis by setting the lowest and highest rating of every trial to 0 and 100. The vertical dotted lines indicate that the groups on the left are significantly different ($p < 0.05$) from the groups on the right (before normalisation).

Inferential analysis was performed through a RM-ANOVA. The main dependent variable was the reverb spatial order, but its interactions with other variables such as room, type of audio material and test location (in situ vs laboratory) were investigated as well. Following the MUSHRA recommendation (ITU-R, 2015a), the Huynh-Feldt correction was applied to reduce Type I errors, as the data did not pass the Mauchly sphericity test ($p < 0.001$) and showed a Greenhouse-Geisser epsilon higher than 0.75 ($\epsilon = 0.81$). A significance value of $\alpha = 0.05$ was used.

1. *Effect of spatial order*: The RM-ANOVA found a significant effect of spatial order on listeners’ ratings [$F(5, 380) = 747.17, p < 0.001$]. Post hoc multiple dependent samples t -tests were run using a corrected significance level of $\alpha' = 0.0033$. Significant differences were found between all pairs of conditions [$t(83) \leq -3.65, p < 0.001$] except between the third- and fourth-order conditions [$t(83) = -1.67, p = 0.098$].
2. *Effect of room*: A significant effect was found for the interaction between room and spatial

order [$F(5, 380) = 2.92, p = 0.019$]. Post hoc multiple dependent samples t -tests were run on data separated by rooms using a corrected significance level of $\alpha' = 0.0033$. For the library, differences between the second- and fourth-order conditions [$t(41) = -2.80, p = 0.008$] and between the third- and fourth-order conditions [$t(41) = 0.20, p = 0.846$] were not significant, whereas significant differences were found for all other pairs of conditions [$t(41) \leq -3.34, p \leq 0.002$]. The fact that a significant difference was found between the second- and third-order but not between the second- and fourth-order may seem surprising at first. However, it is worth mentioning that the latter was very close to being significant ($0.008 \not\gg \alpha'$). Also, this result can be explained by the fact that the RM-ANOVA is a parametric analysis which relies on comparisons between means and the fourth-order data presented some outliers which slightly lowered the mean rating, bringing it closer to that of second-order (cf. Figure 6.8, top-middle plot). The outliers are attributed to some listeners having difficulties to discriminate between the hidden reference and other stimuli, given their high similarity. For the trapezoid, on the other hand, differences between the second- and third-order conditions [$t(41) = -1.67, p = 0.103$], and between the third- and fourth-order conditions [$t(41) = -2.39, p = 0.021$] were not significant, whereas all other pairs of conditions showed significant differences [$t(41) \leq -4.15, p < 0.001$].

3. *Other interactions*: No significant interactions were found between the spatial order and test location [$F(5, 380) = 1.28, p = 0.275$], type of audio material [$F(5, 380) = 1.94, p = 0.099$], or any of the three-way interactions ($p > 0.6$).

Listeners performed an average absolute head rotation (in azimuth) of 1202° ($SD = 1672^\circ$) or 6.68 half-circle rotations per MUSHRA trial. Due to the high variance in amount of head movement across subjects, the potential effect of head rotations on MUSHRA ratings was explored. Figure 6.9 shows the MUSHRA relative ratings, i.e. deviations from the reference's rating, plotted against the average head rotation per trial, calculated from the total head movement during the full test for each listener. For each test condition, the Pearson correlation coefficient (ρ) between the average head rotation and the relative ratings was calculated. By inspecting Figure 6.9, it is evident that correlation was small ($|\rho| < 0.1$) for all conditions except

the zeroth- and first-order conditions, which show negative correlation values, indicating that they were rated lower by listeners who employed more head movements. The low correlation on the dry condition indicates that it produced low ratings regardless of the amount of head movement, which was expected as it is the anchor condition. Meanwhile, the low correlations for orders greater than one suggest that listeners did not perceive them as more similar or different to the reference by performing additional head movements. Note that similar trends were observed when separating data per room (library/trapezoid) and test location (in-situ/laboratory), but the analysis is not reported here due to space constraints.

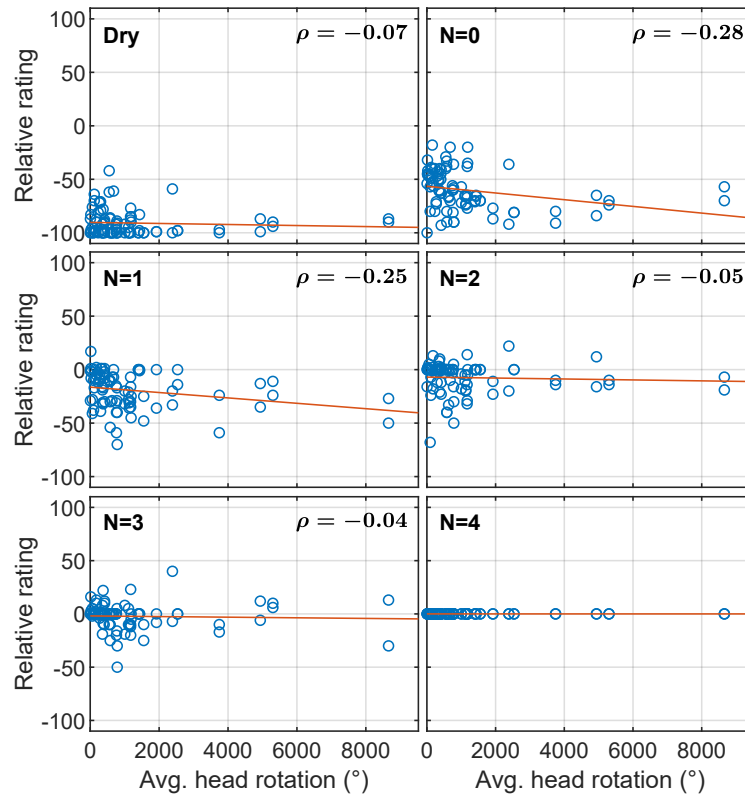


Figure 6.9: Scatterplots showing all relative ratings (absolute ratings minus reference ratings) from experiment 1 and the amount of head rotation (azimuth) that listeners performed for each rating. A linear fit of the data is displayed in red, and the Pearson correlation coefficient is indicated at the top. Data for both rooms were pooled together.

6.5.3 Experiment 2: Paradigm

The second experiment aimed to compare the proposed computationally efficient RVL to a more accurate rendering approach in terms of subjective preference. Because the main limitation

of RVL renderings is the way head rotations are implemented as the room ‘rotates’ with the listener’s head, this experiment focused on that aspect in order to evaluate RVL’s performance in an adverse scenario. Thus, three different rendering methods were evaluated:

1. $N = 1$: head-tracked first order hybrid Ambisonics, i.e. identical to condition $N = 1$ from experiment 1,
2. $N = 1(S)$: hybrid Ambisonics where the direct sound path was head-tracked but the reverb was not, i.e. the reverb did not change according to head movements, and
3. *RVL*: head-tracked first order RVL as described in Subsection 6.2.3, where the direct sound was generated as in the two methods above, and reverb was based on KEMAR BRIRs.

Therefore, RVL was compared to an approach which implemented head rotations properly. Additionally, a ‘static’ method $N = 1(S)$ was introduced as an anchor condition in which head movements did not influence the incoming direction of reverb and, therefore, simulated head rotations with a lower accuracy than RVL. As in the previous experiment, the direct sound path was rendered through HRIR convolution and was head-tracked for all conditions.

Because the methods under comparison were generated from different measurements (Eigenmike RIRs for hybrid Ambisonics and KEMAR BRIRs for RVL) there existed significant spectral differences between the renderings, as shown in Subsection 6.4.3, that were not trivial to compensate through EQ. Preliminary tests with discriminability and MUSHRA tasks showed that listeners focused mainly on these spectral differences rather than other attributes such as reverb directionality. Therefore, a preference task was performed instead in which no reference was provided and listeners evaluated whether the renderings met their internal expectations after seeing a picture of the room or from their own experience, i.e. for those who conducted the test in situ. This ensured that listeners were free to judge which rendering better adjusted to their internal reference without being imposed a particular perceptual metric, be it spectral differences or otherwise.

A double-blind pairwise comparison listening test paradigm was used. In each trial, listeners were shown a picture of the rendered room (library or trapezoid) and a diagram of the sound

scene and were presented two stimuli (A and B). These stimuli were two binaural renderings of the same room, generated with two different methods out of the possible three (except in null pairs, where A and B were identical). Listeners were asked the question, ‘Considering the given room, which example is more appropriate?’. To answer, they would use a continuous rating scale from -2 to $+2$ (with one decimal place) from ‘definitely A’ to ‘definitely B’. Listeners could freely switch between the synchronised stimuli during a trial, and head movements were encouraged to explore the scene. As in experiment 1, the user interface was implemented in Max 7. For each room and type of audio material, listeners evaluated all possible pairs of conditions plus two null pairs, in which A and B were identical (randomly chosen), totalling 16 paired comparisons. Because no reference was provided, there were no correct or wrong answers except for the null pairs, which a reliable listener should rate zero as no audible differences existed in those. Post-screening was applied to exclude listeners who rated the null pairs with an average absolute value higher than 0.25.

Note that the $N = 1(S)$ and RVL conditions were not included in the MUSHRA test (experiment 1) to keep the unidimensionality across test conditions, i.e. spatial order. The reason of using first-order hybrid Ambisonics was to have a fair comparison in the sense that all conditions used the same number of virtual loudspeakers (six). Although the results of this experiment might be insightful regarding the limitations of RVL, it is worth pointing out that they cannot directly be extrapolated to higher-order reproduction conditions.

6.5.4 Experiment 2: Results

Results of the second experiment are shown in Figure 6.10. Data for eight listeners (four in situ, four laboratory) were excluded in post-screening, of which six had also been screened out from experiment 1. Listeners took an average time of 41.33 s (SD = 18.55 s) to complete each paired comparison. Descriptive analysis shows that the mean rating was close to zero for the null pairs, the preference between RVL and hybrid Ambisonics reverb changed depending on the room and type of audio material, and static first-order hybrid Ambisonics renderings were perceived as very similar to the dynamic renderings with a slight trend toward favouring the former. The

fact that RVL was not systematically rated lower than hybrid Ambisonics suggested that the simplifications in the RVL rendering did not significantly impair subjective preference.

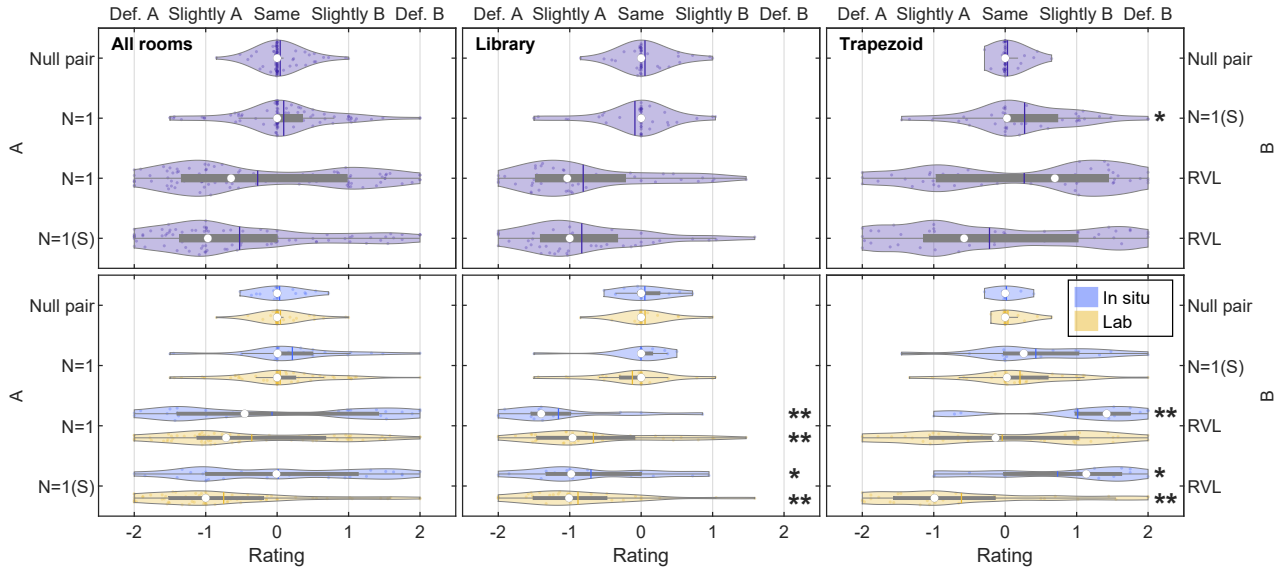


Figure 6.10: Results from experiment 2 represented by violin plots (Hintze and R. D. Nelson, 1998), which show the probability density of the data, median (circle), interquartile range (box) and mean (vertical line). (Top) Both test locations pooled together. (Bottom) Separated per test location (from top to bottom, ‘in situ’ and ‘lab’). From left to right, both rooms pooled together, library and trapezoid data are shown. Ratings range from ‘definitely prefer A’ (–2) to ‘definitely prefer B’ (2). Asterisks indicate whether the mean is significantly different from zero: * for $p < 0.05$ and ** for $p \leq 0.005$.

The inferential analysis tried to find whether listeners had a significant preference on each paired comparison and whether this was affected by factors such as room, type of audio material, and test location. To that end, a RM-ANOVA was first run on the data of each paired comparison to study the effect of the different variables and their interactions. Then, data were grouped accordingly and t -tests were run to evaluate whether the samples deviated from a normal distribution with a mean equal to zero. The Mauchly sphericity test was passed by the $[N = 1$ vs. RVL] and $[N = 1(S)$ vs. RVL] pairs ($p > 0.05$) but not by the $[N = 1$ vs. $N = 1(S)]$ or the null pair, for which the Greenhouse-Geisser correction was applied ($\varepsilon < 0.75$). As in the previous experiment, a significance value of $\alpha = 0.05$ was used.

1. *Null pair*: No significant effect of any variable was found, thus, data were grouped for all types of audio material, rooms, and test locations. A one-sample t -test showed that the mean rating was not significantly different from zero [$t(95) = 1.69$, $p = 0.094$].

2. $[N = 1 \text{ vs. } N = 1(S)]$: A significant effect of the room was found [$F(1, 22) = 4.38$, $p = 0.048$], therefore, the data were separated by rooms. t -tests showed that the trapezoid mean rating was significantly different from zero [$t(47) = 2.57$, $p = 0.013$], with the $N = 1(S)$ renderings being preferred. In the case of the library, the mean rating was not significantly different from zero [$t(47) = -1.15$, $p = 0.254$].
3. $[N = 1 \text{ vs. } RVL]$: Significant effects of the room [$F(1, 22) = 39.18$, $p < 0.001$] and the interaction of room and location [$F(1, 22) = 11.95$, $p = 0.002$] were found, thus, data were separated by room and location (in situ/laboratory). t -tests showed that conditions library–in-situ [$t(13) = -5.30$, $p < 0.001$], library–laboratory [$t(33) = -4.26$, $p < 0.001$] and trapezoid–in-situ [$t(13) = 3.69$, $p = 0.003$] all had mean ratings significantly different from zero. $N = 1$ renderings were preferred for the two library conditions, whereas RVL was preferred for the trapezoid–in-situ condition.
4. $[N = 1(S) \text{ vs. } RVL]$: Significant effects of the room [$F(1, 22) = 10.62$, $p = 0.004$], location [$F(1, 22) = 7.93$, $p = 0.010$], and room/location interaction [$F(1, 22) = 4.93$, $p = 0.037$] were found, therefore, data were again separated by room and location. t -tests showed that all room–location combinations had mean ratings significantly different from zero: library–in-situ [$t(13) = -2.89$, $p = 0.013$], library–laboratory [$t(33) = -6.21$, $p < 0.001$], trapezoid–in-situ [$t(13) = 2.61$, $p = 0.022$] and trapezoid–laboratory [$t(33) = -2.98$, $p = 0.005$]. $N = 1(S)$ renderings were preferred for both library conditions and trapezoid–laboratory, whereas RVL was preferred for the trapezoid–in-situ condition.

Listeners performed an average absolute head rotation (in azimuth) of 449° (SD = 373°) or 2.49 half-circle rotations per paired comparison. This indicates that, as instructed, they employed head movements to inform their ratings. However, given the relative high variance of the head tracking data across subjects, the potential effect of head rotations on the paired comparison ratings was investigated. Figure 6.11 shows the relation between the ratings and the average head rotation per trial, calculated from the total head movement during the full test for each listener. The Pearson correlation coefficient (ρ) between the average head rotation and the ratings was also calculated and indicated in Figure 6.11. Inspection of these data shows that

head rotations had a near-zero correlation with the ratings for the $[N = 1 \text{ vs. } N = 1(S)]$ pair. For the other two pairs, larger correlations between head rotations and ratings were reported in the library than in the trapezoid, particularly for the $[N = 1 \text{ vs. } RVL]$ pair in which a correlation of $\rho = -0.44$ was observed. Correlation values were generally small ($|\rho| < 0.1$) for all trapezoid data, which might suggest that listeners' ratings in this room did not importantly change when additional head movements were performed. Note that similar trends were observed when separating data per test location (in-situ/laboratory), but the analysis is not reported here due to space constraints.

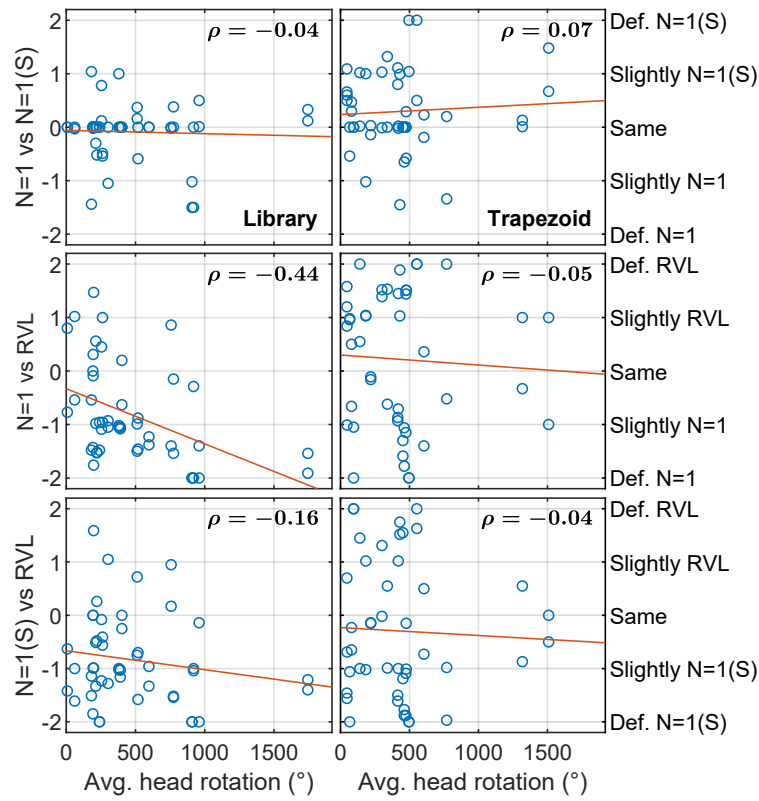


Figure 6.11: Scatterplots showing all paired comparisons' ratings (except null pairs) from experiment 2 and the amount of head rotation (azimuth) that listeners performed for each rating. A linear fit of the data is displayed in red, and the Pearson correlation coefficient is indicated at the top. Data were separated per room (library on the left and trapezoid on the right).

6.6 Discussion

The main goal of this study was to investigate how much spatial reverb rendering can be simplified without degrading perceived quality, assuming that the direct sound path is rendered as accurately as possible. First, the effect of reducing the spatial order of hybrid Ambisonics was investigated. Because the direct sound carries essential information for source localisation and, in some cases, contains most of the energy of the RIR, it was expected that excluding it from the spatial order reduction process would mitigate perceptual degradation. This would imply that provided the direct sound path is rendered separately and with sufficient accuracy, the perceptual impact of reducing the spatial order may be lower than reported in previous studies which used standard Ambisonics (Bernschütz, 2016; Ahrens and Andersson, 2019). Second, the effect of simplifying the implementation of head rotations in dynamic reverb rendering was investigated, by comparing a first-order hybrid Ambisonics rendering against a computationally efficient alternative (RVL), which implemented head rotations in a simplified way, and against a static version where reverb was not head-tracked at all.

6.6.1 Effect of spatial order

Evaluation of objective metrics, such as the IACC and BQI, predicted a large improvement in the spatial quality when increasing the order from zero to one, but the differences became smaller as the order increased. In fact, early BQI ratings measured in this study did not vary as much across the higher spatial orders as those found by Nowak and Klockgether (2017). This may be explained by the fact that removing the most directional part of the RIR, i.e. the direct sound path, led to a more diffuse sound field, which could be rendered accurately with a lower spatial order. This would also explain why the room with less directional reverb, i.e. the library, produced a lower variance in the BQI values. Similarly, it was observed that spectral differences between each BRIR and the reference (order four) increased as the spatial order decreased with the largest jump happening between orders zero and one and with said differences being under 1 dB for all $N \geq 1$. Also, spectral differences were larger for the trapezoid, which may be

explained by its lower DRR or by its more salient early reflections having a larger impact on the signal spectrum, e.g. due to comb-filtering effects (Bech, 1996).

Results of experiment 1 showed that perceived differences were large between orders zero and one and smaller for higher orders, which was in line with the numerical analysis, and a room dependence effect was observed. Data from third-order renderings are particularly representative in that aspect: although their ratings were not significantly different from the hidden reference for either room, it seems that listeners rated them consistently lower in the trapezoid than they did in the library relative to the reference. In fact, data suggest that in the latter room, third- and fourth-order renderings obtained almost identical ratings. This would agree with the numerical analyses in that the trapezoid, with its lower DRR and less diffuse reverberant sound field, displays larger differences between spatial orders than does the library. This is attributed to the fact that reverberation has a higher perceptual weight relative to the direct sound path in the trapezoid than in the library, as indicated by its lower DRR. From these results, it is speculated that rooms with lower DRR will require higher spatial resolution in order to render reverberation with sufficient perceptual quality. It is also worth noting that the differences in rating among orders equal or higher than two were observed to be similar across listeners regardless of the amount of head movements that they employed to explore the auditory scene.

According to Avni et al. (2013), spaciousness and timbre are the most relevant perceptual attributes that listeners use when evaluating sound fields of varying spatial resolution. The present results suggest that when the direct sound path is rendered accurately, the degradation in both spatial and spectral qualities becomes perceptually less relevant, particularly for more diffuse reverberant sound fields, i.e. large rooms. For the conditions tested here, it was shown that the perceived quality of binaural renderings did not improve beyond an order between two and three. This is notably lower than the eighth-order suggested by Ahrens and Andersson (2019), who included the direct sound path in the Ambisonics rendering—saving the differences in experimental paradigm, which was an A-B comparison with attribute scaling (timbre and spaciousness), rather than a MUSHRA test with a single global attribute. However, spatial orders higher than four, not included here due to limitations of the measurement equipment, should be evaluated to draw a more complete comparison to previous studies. Regardless of

this limitation, the present results are in line with the findings of Lübeck, Pörschmann, et al. (2020), who showed that reverb may be rendered through BRIRs sampled on a spherical grid of a spatial order as low as three without degrading perceived quality. At the sight of this, future work could investigate the effect of spatial resolution on each RIR segment separately, i.e. early reflections and late reverb, and how this may depend on the auralised room in a similar fashion to the work by Lübeck, Pörschmann, et al. (2020), but applied to order-truncated Ambisonics signals instead of spatially subsampled signals. Outcomes could be used to inform perceptual models to evaluate spatial audio quality and to enable efficient parametric reverb rendering, e.g. by determining the amount of early reflections needed to generate plausible virtual scenes (Brinkmann, Gamper, et al., 2020).

6.6.2 Dynamic aspects

Results of experiment 2, which compared dynamic and static first-order hybrid Ambisonics reverb with RVL, were more challenging to interpret. The absence of a reference led to bimodal data distributions in some cases (cf. trapezoid data in Figure 6.10), meaning that listeners could discriminate pairs of conditions, but neither was unanimously preferred. An unexpected outcome was that the static version of hybrid Ambisonics [$N = 1(S)$], which was initially conceived as a low-quality anchor, was found to be preferred, in some cases, to the more accurate dynamic version ($N = 1$). In particular, this was true for the trapezoid but not for the library. Post-hoc informal interviews suggested that this could be due to the dynamic reverb being perceived as less ‘stable’ when head rotations were performed as pointed out in Subsection 6.4.4. This might be explained by the fact that virtual loudspeaker decoding approaches yield angle dependent spectral distortions (Solvang, 2008), which often result in poor loudness stability at low orders (Ben-Hur, Alon, Rafaely, et al., 2019). This is supported by the fact that BRIR-predicted loudness is less smooth in the trapezoid than it is in the library as observed in Figure 6.7. Analysis of head tracking data suggested that even listeners who performed a more exhaustive exploration of the scene through head movements did not rate the dynamic version significantly higher than the static version. This result suggests that, provided that direct sound is rendered

dynamically through convolution with an HRIR, it may actually be detrimental to render reverb dynamically if a low Ambisonics order is used and the loudness stability is not accounted for.

When comparing RVL and hybrid Ambisonics reverb, the spectral analysis showed that strong colouration differences should be expected, which may have led to polarised ratings in the perceptual evaluation. In the case of the trapezoid, RVL was clearly preferred by listeners that performed the test in situ, which suggests that this method captured the room characteristics more accurately than did hybrid Ambisonics. However, the opposite was true for the library, where listeners assigned lower ratings to RVL (even more so when exhaustive head movements were employed during the test), the reasons for which are unclear from the present data and might not be revealed with the proposed experimental paradigm. Perhaps additional insights could be gained by testing a higher number of rooms and configurations (e.g. positions of the listener relative to the source) although this could lead to experiments with an excessive number of variables. In any case, the room divergence effect seemed to play a more important role in the second test than it did in the first test because a reference was not given and listeners provided ratings based on their expectations, which depended on previous exposure to the rendered rooms.

6.7 Conclusions

This study addressed the issue of the trade-off between computational complexity and perceived quality for binaural Ambisonics-based reverb. It introduced the concept of hybrid Ambisonics, or the separation between the direct sound path and the reverb in Ambisonics-based binaural rendering, obtaining the former by convolution with a dense HRIR dataset and encoding the latter in an Ambisonics sound field. It was hypothesised that the perceived quality of the renderings would stop improving at a lower spatial order than in previous studies where the direct sound was not processed separately (Bernschütz, 2016; Ahrens and Andersson, 2019), as the directional information of the signal would be better preserved. Results from the perceptual evaluation suggest that when the direct sound path is computed accurately, an Ambisonics order

of two or three may be enough to render binaural reverb, depending on the room characteristics. For instance, rooms with lower DRR or more salient early reflections are likely to require a higher spatial order than are rooms where reverb is more diffuse. In any case, the scope of this study was limited to two measured rooms and spatial orders up to four and, therefore, further evaluations on different spaces and with higher orders are needed to generalise these results. Additionally, future work could look into rendering the most relevant early reflections (Brinkmann, Gamper, et al., 2020) at a higher spatial order, which could lower the spatial resolution requirements for the diffuse reverb.

Additionally, RVL was formalised as a computationally efficient approach to dynamically render binaural reverb for a large number of sources. It was observed that renderings produced with this method were comparable to (and, in some cases, better than) those obtained through less flexible Ambisonics-based approaches in terms of subjective preference. Considering the advantages of RVL, namely, its efficiency and ease of implementation, this method should be worthy of consideration for convolution-based binaural reverb generation in low-cost scenarios.

Chapter 7

Conclusions

This Chapter summarises the Thesis achievements by recalling the findings of each of the presented studies and discusses their relevance in the context of AAR and current binaural audio techniques. Additionally, a general outlook is provided and future lines of work are proposed.

7.1 Summary of Thesis Achievements

The general aim of this Thesis was to improve the current understanding of spatial audio perception in order to improve future implementations of AAR. This was achieved by conducting a series of studies that made contributions on different topics related to the rendering and reproduction of binaural audio, namely: headphone EQ, binaural rendering of Ambisonics signals, and efficient generation of binaural reverberation. Even though the main focus of the work was put on AAR, its outcomes are relevant for any application that makes use of binaural audio.

An initial pilot study (**Chapter 3**) was conducted to get a first understanding of the challenges that AAR presents. To that end, a hear-through prototype was implemented and evaluated in terms of user adaptation, by having two participants wear the device for several days and conducting periodic control tests. Results suggested that a slight adaptation may have taken

place as participants reported an increase in subjective attributes (e.g. sound externalisation) after wearing the device for an extended period. However, the amount of collected data was insufficient to draw meaningful conclusions and a formal experiment has yet to be conducted to validate these results. Such future experiment could be useful to determine if a hear-through solution is a valid option to achieve acoustic transparency in AAR or if a ‘passive’ solution, i.e. open-ear headphones, is required for an adequate user experience.

The first main study of this Thesis (**Chapter 4**) was related to headphone EQ, which is a crucial tool for the accurate rendering of binaural audio, either in AAR or other applications. A perceptual experiment was conducted which showed that headphone EQ can improve the overall quality of binaural renderings, even when non-individual HpEQ filters are used. Note that the headphones used in this study were open-ear and had a limited reproduction bandwidth when not equalised, and it is yet unclear if these results are generalisable to other types of headphones (e.g. closed or in-ear). Regardless, **these findings are relevant** for binaural audio applications in which acquiring individual headphone measurements is inconvenient, such as most current consumer-oriented AR/VR devices, which often have open-ear audio systems (e.g. Microsoft HoloLens, Oculus Quest). For instance, any of such devices’ binaural audio quality could be easily improved by applying generic headphone EQ, which could be achieved by measuring the device’s HpTF on a large user sample (e.g. > 100 subjects), obtaining the average HpEQ filter, and deploying it on all devices via a software update. This would instantly lead to a better user experience without involving the consumer in the procedure, since individual headphone measurements would not be required.

The second main study (**Chapter 5**) concerned the binaural rendering of Ambisonics signals, which is one of the several possible approaches to obtain a binaural signal from a simulated acoustic environment (another being the convolution of dry audio signals with BRIRs, for instance). In said study, an extensive review of Ambisonics-based binaural rendering methods was performed, unifying previous knowledge and providing an objective comparison between several state-of-the-art such methods in terms of their ability to produce binaural signals from low-order Ambisonics sound fields. Furthermore, BiMagLS was proposed, which was shown to outperform the other tested methods for bilateral Ambisonics reproduction, according to

auditory models. **These findings are relevant** for spatial audio applications that employ low-order Ambisonics signals, either because of practical limitations (e.g. order-limited microphone arrays) or computational efficiency concerns (e.g. AAR system with limited resources). For instance, consider a hypothetical AAR system with a real-time Ambisonics audio engine and a state-of-the-art MagLS binaural renderer that operated at a spatial order of $N = 4$ (25 Ambisonics channels). According to presented evaluation with auditory models, if the rendering method was changed to BiMagLS and the order reduced to $N = 2$ ($2 \times 9 = 18$ channels), it would improve sound localisation while reducing the computational cost.

Finally, **Chapter 6** investigated the issue of efficient reverberation rendering. A study was presented which explored perceptual thresholds within Ambisonics-based reverberation perception, particularly looking at the importance of reverberation's spatial resolution when the direct sound path is reproduced accurately. As part of this study, hybrid Ambisonics was proposed as an efficient method to generate binaural reverberation and is evaluated numerically and perceptually, showing that a second- or third-order Ambisonics reverberation simulation might be enough from a perceptual point of view. Furthermore, the RVL method was formalised, whose main feature is that its computation time is independent from the number of rendered sources, whereas for previous methods it scaled linearly. Numerical and perceptual evaluations showed that RVL is a viable, highly efficient approach to produce binaural reverberation in real time for an arbitrarily large number of sources. **These findings are relevant** to the development of future binaural audio engines. For instance, a sound field may be efficiently processed by using different spatial resolutions in parallel, e.g. high resolution ($N > 30$) for the direct sound path, medium resolution ($N \simeq 5$) for early reflections, and low resolution (RVL, $N < 3$) for late reverberation. Similarly, spatial audio signals could be compressed following a perceptual model of the perceived spatial resolution, e.g. an Ambisonics signal could be analysed by time frames and those that displayed a diffuse pattern would be encoded with a low order, while the more directive ones would be encoded with a high order. This would, in turn, allow for a more efficient generation, transmission, and storage of spatial audio.

7.2 Outlook and Future Work

A crucial issue in AAR that still needs to be solved is the need of **acoustically transparent headphones** that still provide a broad reproduction bandwidth. In this Thesis we have considered two different devices: a hear-through option, which have a sufficient bandwidth but it is unclear if they can provide proper acoustical transparency (Chapter 3), and an open-ear option, which is acoustically transparent but offers a limited bandwidth (Chapter 4). Whether either option is more appropriate for future AAR devices is still an open question. Perhaps the choice can be given to the user by having two different headphone accessories to the AAR device: if their priority is to maximise audio quality for virtual sound sources, they may use the hear-through headphones, while if their priority is to minimise the degradation to real-world sounds, they should opt for the open-ear headphones. Regardless, future work should investigate potential improvements to both these technologies. In the case of hear-through, an important limitation is the undesired amplification of one's own voice and other self sounds (e.g. chewing) due to the ear canal occlusion, which could be addressed via active noise cancellation. In the case of open-ear, the challenge is increasing the reproduction bandwidth while maintaining the open form factor (to not alter the impedance of the ear canal), keep binaural crosstalk low (to not interfere with auditory cues) and also keep a low acoustic radiation to the outside (to prevent other people from hearing the binaural signals and ensure the user's privacy). This may only be achieved by researching new audio drivers and enclosures, potentially considering other headphone technologies such as bone conduction or cartilage conduction (Miller and Mehra, 2019).

Regarding **headphone equalisation**, the perceptual effect of generic EQ filters could be further explored in various ways. First, by proposing generic filters that are calculated as an average of several measurements performed on human listeners, rather than a single measurement on a head and torso simulator, which may not be a good approximation of the average listener's anatomy as shown by Lindau and Brinkmann (2012). Additionally, perceptual studies could be conducted to investigate whether generic HpEQ affects the vertical localisation ability of listeners, and whether the findings presented Chapter 4 will extend to headphones with a wider reproduction

bandwidth. Finally, a more ambitious line of research would be to try and generate individual HpEQ filters through data-driven methods without performing actual in-ear measurements, e.g. by using machine learning techniques to estimate the filters from a set of pictures of the user's ears. A similar approach could be taken to **generate personalised HRTFs** from images of the user's head, torso and ears using artificial intelligence, which would signify a huge leap towards high-quality binaural audio at a consumer level (Zhou et al., 2021).

Regarding the **binaural rendering of Ambisonics signals**, several futures lines of research are proposed. First, the study presented on Chapter 5 could be followed up by an actual listening experiment to validate the results and the proposed BiMagLS method. This experiment should include both anechoic and reverberant conditions; the former to validate the auditory models' predictions since they also assume anechoic signals, and the latter to explore whether such conditions affect the the choice of rendering method (e.g. the MagLS+CC method is designed to optimise interaural coherence, which is especially relevant when rendering reverberant signals).

On a related note, understanding the perceptual thresholds associated to **reverberation** still presents an open set of challenges, of which this Thesis has only partially addressed. For instance, a crucial part of AAR is identifying the listener's environment in order to produce an accurate acoustic simulation. This could rely, for instance, on measuring some key parameters such as the room dimensions, distance to the closest reflecting surfaces, or the approximated RT through the AR system's onboard sensors. Therefore, it is essential to learn to which extent these parameters need to be estimated in order to produce plausible virtual sources, which should be assessed through extensive perceptual experiments or by means of auditory models.

Overall, AAR still poses many scientific and technical challenges at the time of writing this Thesis, which has only scratched the surface of the problem. Years will probably pass until we have the technology ready at a consumer level, not counting the plethora of other scientific and technical challenges related to the non-auditory aspects of AR, such as vision (realistic holograms on a wide field of view), human-computer interfaces (haptic feedback, neural input) and one of the toughest and most important ones: packing all the required technology and processing power in a pair of glasses that is lightweight and does not get too warm so they can

be comfortably worn all day. Whether we will see this technology in our lifetime only time will tell, but we are on the right path.

Bibliography

- Ahrens, Jens and Carl Andersson (2019). ‘[Perceptual Evaluation of Headphone Auralization of Rooms Captured with Spherical Microphone Arrays with Respect to Spaciousness and Timbre](#)’. *The Journal of the Acoustical Society of America* 145.4, pp. 2783–2794 (cited on pp. [59](#), [93](#), [97](#), [101](#), [104](#), [122](#), [123](#), [125](#)).
- Alais, David and David Burr (2004). ‘[The Ventriloquist Effect Results from Near-Optimal Bimodal Integration](#)’. *Current Biology* 14.3, pp. 257–262 (cited on p. [54](#)).
- Albrecht, Robert, Tapio Lokki, and Lauri Savioja (2011). ‘[A Mobile Augmented Reality Audio System with Binaural Microphones](#)’. *Proceedings of Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications*. IwS ’11, pp. 7–11 (cited on pp. [25](#), [26](#), [27](#)).
- Allen, Jont B. and David A. Berkley (1979). ‘[Image Method for Efficiently Simulating Small-room Acoustics](#)’. *The Journal of the Acoustical Society of America* 65.4, pp. 943–950 (cited on pp. [4](#), [12](#), [96](#)).
- Andersson, Carl (2016). ‘[Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays](#)’. Master Thesis. Chalmers University of Technology (cited on p. [16](#)).
- Arend, Johannes M., Fabian Brinkmann, and Christoph Pörschmann (2021). ‘[Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions](#)’. *Journal of the Audio Engineering Society* 69.1/2, pp. 104–117 (cited on p. [65](#)).
- Armstrong, Cal, Lewis Thresh, Damian T. Murphy, and Gavin Kearney (2018). ‘[A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database](#)’. *Applied Sciences* 8.11 (11), p. 2029 (cited on p. [104](#)).

- Armstrong, Calum, Thomas McKenzie, Damian T. Murphy, and Gavin Kearney (2018). ‘[A Perceptual Spectral Difference Model for Binaural Signals](#)’. *Audio Engineering Society Convention 145* (cited on p. 72).
- Avni, Amir, Jens Ahrens, Matthias Geier, Sascha Spors, Hagen Wierstorf, and Boaz Rafaely (2013). ‘[Spatial Perception of Sound Fields Recorded by Spherical Microphone Arrays with Varying Spatial Resolution](#)’. *The Journal of the Acoustical Society of America* 133.5, pp. 2711–2721 (cited on pp. 17, 59, 60, 73, 93, 97, 101, 104, 123).
- Barron, Marshall and Arthur H. Marshall (1981). ‘[Spatial Impression Due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure](#)’. *Journal of Sound and Vibration* 77.2, pp. 211–232 (cited on p. 95).
- Barumerli, Roberto, Piotr Majdak, Jonas Reijniers, Robert Baumgartner, Michele Geronazzo, and Federico Avanzini (2020). ‘[Predicting Directional Sound-Localization of Human Listeners in Both Horizontal and Vertical Dimensions](#)’. *Audio Engineering Society Convention 148* (cited on pp. 73, 87).
- Baumgartner, Robert and Piotr Majdak (2020). ‘[Decision Making in Auditory Externalization Perception](#)’. *bioRxiv* (cited on pp. 21, 61, 72, 73, 83).
- Baumgartner, Robert, Piotr Majdak, and Bernhard Laback (2014). ‘[Modeling Sound-Source Localization in Sagittal Planes for Human Listeners](#)’. *The Journal of the Acoustical Society of America* 136.2, p. 791 (cited on pp. 72, 73).
- Bech, Søren (1996). ‘[Timbral Aspects of Reproduced Sound in Small Rooms. II](#)’. *The Journal of the Acoustical Society of America* 99.6, pp. 3539–3549 (cited on pp. 95, 106, 123).
- Begault, Durand R. and Leonard J. Trejo (2000). ‘[3-D Sound for Virtual Reality and Multimedia](#)’ (cited on pp. 38, 58).
- Begault, Durand R., Elizabeth M. Wenzel, and Mark R. Anderson (2001). ‘[Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source](#)’. *Journal of the Audio Engineering Society* 49.10, pp. 904–916 (cited on pp. 96, 103).
- Ben-Hur, Zamir, David Lou Alon, Ravish Mehra, and Boaz Rafaely (2019). ‘[Efficient Representation and Sparse Sampling of Head-Related Transfer Functions Using Phase-Correction Based](#)

- on Ear Alignment’. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12, pp. 2249–2262 (cited on pp. 17, 18, 19, 20, 60, 63, 65, 66, 70).
- Ben-Hur, Zamir, David Lou Alon, Ravish Mehra, and Boaz Rafaely (2020). ‘Binaural Reproduction Using Bilateral Ambisonics’. *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality* (cited on pp. 61, 67, 79, 87).
- Ben-Hur, Zamir, David Lou Alon, Ravish Mehra, and Boaz Rafaely (2021). ‘Binaural Reproduction Based on Bilateral Ambisonics and Ear-Aligned HRTFs’. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 901–913 (cited on pp. 61, 66, 67, 79, 87, 89).
- Ben-Hur, Zamir, David Lou Alon, Boaz Rafaely, and Ravish Mehra (2019). ‘Loudness Stability of Binaural Sound with Spherical Harmonic Representation of Sparse Head-Related Transfer Functions’. *EURASIP Journal on Audio, Speech, and Music Processing* 2019.1, p. 5 (cited on pp. 60, 64, 68, 78, 87, 98, 110, 124).
- Ben-Hur, Zamir, Fabian Brinkmann, Jonathan Sheaffer, Stefan Weinzierl, and Boaz Rafaely (2017). ‘Spectral Equalization in Binaural Signals Represented by Order-Truncated Spherical Harmonics’. *The Journal of the Acoustical Society of America* 141.6, pp. 4087–4096 (cited on pp. 63, 64, 74).
- Ben-Hur, Zamir, Jonathan Sheaffer, and Boaz Rafaely (2018). ‘Joint Sampling Theory and Subjective Investigation of Plane-Wave and Spherical Harmonics Formulations for Binaural Reproduction’. *Applied Acoustics* 134, pp. 138–144 (cited on pp. 18, 60, 68).
- Beranek, Leo L. (2008). ‘Concert Hall Acoustics—2008’. *Journal of the Audio Engineering Society* 56.7/8, pp. 532–544 (cited on pp. 96, 107, 108).
- Bernschütz, Benjamin (2013). ‘A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100’. *Proceedings of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA)*, pp. 592–595 (cited on p. 60).
- Bernschütz, Benjamin (2016). ‘Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording’. Doctoral Thesis. Technische Universität Berlin (cited on pp. 14, 18, 20, 59, 68, 69, 97, 101, 122, 125).

- Bernschütz, Benjamin, Christoph Pörschmann, Sascha Spors, and Stefan Weinzierl (2011). ‘[SOFiA Sound Field Analysis Toolbox](#)’. *Proceedings of the International Conference on Spatial Audio (ICSA)*, pp. 7–15 (cited on p. 69).
- Bernschütz, Benjamin, Arnau Vázquez Giner, Christoph Pörschmann, and Johannes M. Arend (2014). ‘[Binaural Reproduction of Plane Waves With Reduced Modal Order](#)’. *Acta Acustica united with Acustica* 100.5, pp. 972–983 (cited on pp. 60, 63, 68, 97, 98).
- Best, Virginia, Robert Baumgartner, Mathieu Lavandier, Piotr Majdak, and Norbert Kopčo (2020). ‘[Sound Externalization: A Review of Recent Research](#)’. *Trends in Hearing* 24 (cited on pp. 74, 88).
- Blauert, Jens (1969). ‘[Sound Localization in the Median Plane](#)’. *Acta Acustica united with Acustica* 22.4, pp. 205–213 (cited on p. 14).
- Blauert, Jens (1997). ‘Spatial Hearing: The Psychophysics of Human Sound Localization’. MIT Press (cited on pp. 4, 9, 25, 30).
- Bolaños, Javier Gómez, Aki Mäkilvirta, and Ville Pulkki (2016). ‘[Automatic Regularization Parameter for Headphone Transfer Function Inversion](#)’. *Journal of the Audio Engineering Society* 64.10, pp. 752–761 (cited on pp. 42, 43).
- Botteldooren, Dick (1995). ‘[Finite-difference Time-domain Simulation of Low-frequency Room Acoustic Problems](#)’. *The Journal of the Acoustical Society of America* 98.6, pp. 3302–3308 (cited on p. 96).
- Brinkmann, Fabian, Hannes Gamper, Nikunj Raghuvanshi, and Ivan Tashev (2020). ‘[Towards Encoding Perceptually Salient Early Reflections for Parametric Spatial Audio Rendering](#)’. *Audio Engineering Society Convention 148* (cited on pp. 124, 126).
- Brinkmann, Fabian, Alexander Lindau, and Stefan Weinzierl (2017). ‘[On the Authenticity of Individual Dynamic Binaural Synthesis](#)’. *The Journal of the Acoustical Society of America* 142.4, pp. 1784–1795 (cited on pp. 39, 49, 51, 54).
- Brinkmann, Fabian, Alexander Lindau, Stefan Weinzierl, Steven van de Par, Markus Müller-Trappel, Rob Opdam, and Michael Vorländer (2017). ‘[A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations](#)’. *Journal of the Audio Engineering Society* 65.10, pp. 841–848 (cited on pp. 10, 18, 60, 71).

- Brinkmann, Fabian and Stefan Weinzierl (2018). ‘[Comparison of Head-Related Transfer Functions Pre-Processing Techniques for Spherical Harmonics Decomposition](#)’. *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality* (cited on pp. [61](#), [66](#), [67](#)).
- Brown, Andrew D., G. Christopher Stecker, and Daniel J. Tollin (2015). ‘[The Precedence Effect in Sound Localization](#)’. *JARO: Journal of the Association for Research in Otolaryngology* 16.1, pp. 1–28 (cited on pp. [95](#), [97](#)).
- Brungart, Douglas S., Alex J. Kordik, Cynthia S. Eades, and Brian D. Simpson (2003). ‘[The Effect of Microphone Placement on Localization Accuracy with Electronic Pass-through Earplugs](#)’. *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pp. 149–152 (cited on pp. [24](#), [25](#)).
- Cabrera, Densil, Doheon Lee, Manuj Yadav, and William L. Martens (2011). ‘[Decay Envelope Manipulation of Room Impulse Responses: Techniques for Auralization and Sonification](#)’. *Acoustics 2011* (cited on p. [104](#)).
- Catic, Jasmina, Sébastien Santurette, and Torsten Dau (2015). ‘[The Role of Reverberation-Related Binaural Cues in the Externalization of Speech](#)’. *The Journal of the Acoustical Society of America* 138.2, pp. 1154–1167 (cited on p. [2](#)).
- Cuevas-Rodríguez, María, Lorenzo Picinali, Daniel González-Toledo, Carlos Garre, Ernesto de la Rubia-Cuestas, Luis Molina-Tanco, and Arcadio Reyes-Lecuona (2019). ‘[3D Tune-In Toolkit: An Open-Source Library for Real-Time Binaural Spatialisation](#)’. *PLoS ONE* 14.3 (cited on pp. [10](#), [58](#), [65](#), [99](#), [105](#)).
- Daniel, Jérôme (2001). ‘[Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia](#)’. Doctoral Thesis. Université Paris 6 (cited on p. [65](#)).
- Daniel, Jérôme, Jean-Bernard Rault, and Jean-Dominique Polack (1998). ‘[Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions](#)’. *Audio Engineering Society Convention 105* (cited on p. [65](#)).
- Engel, Isaac (2021). ‘[BinauralSH Library for Matlab](#)’ (cited on pp. [62](#), [90](#)).

- Engel, Isaac, David Lou Alon, Philip W. Robinson, and Ravish Mehra (2019). ‘[The Effect of Generic Headphone Compensation on Binaural Renderings](#)’. *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio* (cited on pp. [7](#), [37](#), [64](#)).
- Engel, Isaac, David Lou Alon, Kevin Scheumann, and Ravish Mehra (2020). ‘[Listener-Preferred Headphone Frequency Response for Stereo and Spatial Audio Content](#)’. *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality* (cited on p. [64](#)).
- Engel, Isaac, Dan F. M. Goodman, and Lorenzo Picinali (2021). ‘Perceptual Models as a Mean of Assessment of Ambisonics-Based Binaural Rendering Methods’. *Submitted to Acta Acustica in March 2021 and currently under review* (cited on pp. [7](#), [57](#)).
- Engel, Isaac, Craig Henry, Sebastià V. Amengual Garí, Philip W. Robinson, and Lorenzo Picinali (2021). ‘[Perceptual Implications of Different Ambisonics-Based Methods for Binaural Reverberation](#)’. *The Journal of the Acoustical Society of America* 149.2, pp. 895–910 (cited on pp. [7](#), [88](#), [91](#)).
- Engel, Isaac, Craig Henry, Sebastià V. Amengual Garí, Philip W. Robinson, David Poirier-Quinot, and Lorenzo Picinali (2019). ‘[Perceptual Comparison of Ambisonics-Based Reverberation Methods in Binaural Listening](#)’. *EAA Spatial Audio Signal Processing Symposium*, pp. 121–126 (cited on pp. [59](#), [68](#), [93](#), [98](#), [99](#)).
- Engel, Isaac and Lorenzo Picinali (2017). ‘[Long-Term User Adaptation to an Audio Augmented Reality System](#)’. *24th International Congress on Sound and Vibration* (cited on pp. [7](#), [23](#)).
- Evans, Michael J., James A. S. Angus, and Anthony I. Tew (1998). ‘[Analyzing Head-Related Transfer Function Measurements Using Surface Spherical Harmonics](#)’. *The Journal of the Acoustical Society of America* 104.4, pp. 2400–2411 (cited on p. [65](#)).
- Farina, Angelo (2007). ‘[Advancements in Impulse Response Measurements by Sine Sweeps](#)’. *Audio Engineering Society Convention 122* (cited on pp. [11](#), [27](#), [43](#), [102](#)).
- Gerzon, Michael A. (1973). ‘[Periphony: With-Height Sound Reproduction](#)’. *Journal of the Audio Engineering Society* 21.1, pp. 2–10 (cited on pp. [3](#), [59](#)).

- Gerzon, Michael A. (1985). ‘[Ambisonics in Multichannel Broadcasting and Video](#)’. *Journal of the Audio Engineering Society* 33.11, pp. 859–871 (cited on p. 92).
- Gerzon, Michael A. (1992). ‘[General Metatheory of Auditory Localisation](#)’. *Audio Engineering Society Convention 92* (cited on p. 65).
- Glasberg, Brian R. and Brian C. J. Moore (1990). ‘[Derivation of Auditory Filter Shapes from Notched-Noise Data](#)’. *Hearing Research* 47.1, pp. 103–138 (cited on pp. 20, 72).
- Gorzel, Marcin, Andrew Allen, Ian Kelly, Julius Kammerl, Alper Gungormusler, Hengchin Yeh, and Francis Boland (2019). ‘[Efficient Encoding and Decoding of Binaural Sound with Resonance Audio](#)’. *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio* (cited on pp. 59, 68).
- Guezenoc, Corentin and Renaud Segulier (2018). ‘[HRTF Individualization: A Survey](#)’. *Audio Engineering Society Convention 145* (cited on p. 60).
- Hansen, Villy and Gert Munch (1991). ‘[Making Recordings for Simulation Tests in the Archimedes Project](#)’. *Journal of the Audio Engineering Society* 39.10, pp. 768–774 (cited on pp. 44, 106).
- Hardin, Ronald H. and Neil J. A. Sloane (1996). ‘[McLaren’s Improved Snub Cube and Other New Spherical Designs in Three Dimensions](#)’. *Discrete & Computational Geometry* 15.4, pp. 429–441 (cited on p. 68).
- Härmä, Aki, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Jarmo Hiipakka, and Gaëtan Lorho (2004). ‘[Augmented Reality Audio for Mobile and Wearable Appliances](#)’. *Journal of the Audio Engineering Society* 52.6, pp. 618–639 (cited on pp. 4, 24, 25, 26, 29).
- Hartmann, William M., Brad Rakerd, Zane D. Crawford, and Peter Xinya Zhang (2016). ‘[Transaural Experiments and a Revised Duplex Theory for the Localization of Low-Frequency Tones](#)’. *The Journal of the Acoustical Society of America* 139.2, pp. 968–985 (cited on p. 67).
- Hiipakka, Marko, Marko Takanen, Symeon Delikaris-Manias, Archontis Politis, and Ville Pulkki (2012). ‘[Localization in Binaural Reproduction with Insert Headphones](#)’. *Audio Engineering Society Convention 132* (cited on pp. 25, 27).
- Hintze, Jerry L. and Ray D. Nelson (1998). ‘[Violin Plots: A Box Plot-Density Trace Synergism](#)’. *The American Statistician* 52.2, pp. 181–184 (cited on pp. 114, 119).

- Hoffmann, Pablo F., Flemming Christensen, and Dorte Hammershøi (2013). ‘[Insert Earphone Calibration for Hear-Through Options](#)’. *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones* (cited on pp. 25, 27).
- Hold, Christoph, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J. Tashev (2019). ‘[Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation](#)’. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265 (cited on pp. 63, 64, 65).
- ITU-R (2015a). ‘[BS.1534: Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems](#)’ (cited on pp. 40, 48, 49, 50, 51, 113, 114).
- ITU-R (2015b). ‘[BS.1770: Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level](#)’ (cited on p. 110).
- Jelfs, Sam, John F. Culling, and Mathieu Lavandier (2011). ‘[Revision and Validation of a Binaural Model for Speech Intelligibility in Noise](#)’. *Hearing Research* 275.1, pp. 96–104 (cited on pp. 21, 61, 74).
- Jot, Jean-Marc (1997). ‘[Efficient Models for Reverberation and Distance Rendering in Computer Music and Virtual Audio Reality](#)’. *ICMC: International Computer Music Conference*, pp. 236–243 (cited on p. 96).
- Jot, Jean-Marc and Antoine Chaigne (1991). ‘[Digital Delay Networks for Designing Artificial Reverberators](#)’. *Audio Engineering Society Convention 90* (cited on p. 96).
- Kaplanis, Neofytos, Søren Bech, Søren Holdt Jensen, and Toon van Waterschoot (2014). ‘[Perception of Reverberation in Small Rooms: A Literature Study](#)’. *Audio Engineering Society Conference: 55th International Conference: Spatial Audio* (cited on p. 96).
- Karjalainen, Matti, Ville Riikonen, and Miikka Tikander (2008). ‘[An Augmented Reality Audio Mixer and Equalizer](#)’. *Audio Engineering Society Convention 124* (cited on pp. 24, 25, 26, 27, 29).
- Katz, Brian F. G. and Markus Noisternig (2014). ‘[A Comparative Study of Interaural Time Delay Estimation Methods](#)’. *The Journal of the Acoustical Society of America* 135.6, pp. 3530–3540 (cited on pp. 48, 72).

- Kirkeby, Ole and Philip A. Nelson (1999). ‘[Digital Filter Design for Inversion Problems in Sound Reproduction](#)’. *Journal of the Audio Engineering Society* 47.7/8, pp. 583–595 (cited on pp. [41](#), [42](#), [63](#), [64](#)).
- Klockgether, Stefan and Steven van de Par (2016). ‘[Just Noticeable Differences of Spatial Cues in Echoic and Anechoic Acoustical Environments](#)’. *The Journal of the Acoustical Society of America* 140.4, EL352–EL357 (cited on pp. [76](#), [78](#)).
- Kohlrausch, Armin and Jeroen Breebaart (2001). ‘[Perceptual \(Ir\)Relevance of HRTF Magnitude and Phase Spectra](#)’. *Audio Engineering Society Convention 110* (cited on p. [44](#)).
- Langendijk, Erno H. A. and Adelbert W. Bronkhorst (1999). ‘[Fidelity of Three-Dimensional-Sound Reproduction Using a Virtual Auditory Display](#)’. *The Journal of the Acoustical Society of America* 107.1, pp. 528–537 (cited on pp. [39](#), [40](#), [41](#), [45](#)).
- Lebedev, Vyacheslav Ivanovich (1977). ‘[Spherical Quadrature Formulas Exact to Orders 25–29](#)’. *Siberian Mathematical Journal* 18.1, pp. 99–107 (cited on p. [68](#)).
- Leclère, Thibaud, Mathieu Lavandier, and Fabien Perrin (2019). ‘[On the Externalization of Sound Sources with Headphones without Reference to a Real Source](#)’. *The Journal of the Acoustical Society of America* 146.4, pp. 2309–2320 (cited on pp. [85](#), [88](#)).
- Lee, Hyunkook, Matthias Frank, and Franz Zotter (2019). ‘[Spatial and Timbral Fidelities of Binaural Ambisonics Decoders for Main Microphone Array Recordings](#)’. *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio* (cited on pp. [60](#), [87](#)).
- Lindau, Alexander and Fabian Brinkmann (2012). ‘[Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings](#)’. *Journal of the Audio Engineering Society* 60.1/2, pp. 54–62 (cited on pp. [38](#), [39](#), [54](#), [130](#)).
- Lindau, Alexander, Vera Erbes, Steffen Lepa, Hans-Joachim Maempel, Fabian Brinkman, and Stefan Weinzierl (2014). ‘[A Spatial Audio Quality Inventory \(SAQI\)](#)’. *Acta Acustica united with Acustica* 100.5, pp. 984–994 (cited on p. [49](#)).
- Lindau, Alexander, Linda Kosanke, and Stefan Weinzierl (2010). ‘[Perceptual Evaluation of Physical Predictors of the Mixing Time in Binaural Room Impulse Responses](#)’. *Audio Engineering Society Convention 128* (cited on p. [106](#)).

- Lindau, Alexander, Linda Kosanke, and Stefan Weinzierl (2012). ‘Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses’. *Journal of the Audio Engineering Society* 60.11, pp. 887–898 (cited on pp. 12, 94, 100).
- Lindau, Alexander and Stefan Weinzierl (2012). ‘Assessing the Plausibility of Virtual Acoustic Environments’. *Acta Acustica united with Acustica* 98.5, pp. 804–810 (cited on pp. 29, 100).
- Litovsky, Ruth Y., H. Steven Colburn, William A. Yost, and Sandra J. Guzman (1999). ‘The Precedence Effect’. *The Journal of the Acoustical Society of America* 106.4, pp. 1633–1654 (cited on p. 95).
- Lübeck, Tim (2019). ‘Perceptual Evaluation of Mitigation Approaches of Errors Due to Spatial Undersampling in Binaural Renderings of Spherical Microphone Array Data’. Master Thesis. Chalmers University of Technology (cited on p. 20).
- Lübeck, Tim, Johannes M. Arend, Christoph Pörschmann, Hannes Helmholtz, and Jens Ahrens (2020). ‘Perceptual Evaluation of Mitigation Approaches of Impairments Due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data: Dry Acoustic Environments’. *International Conference on Digital Audio Effects 2020* (cited on pp. 59, 60, 70, 87).
- Lübeck, Tim, Hannes Helmholtz, Johannes M. Arend, Christoph Pörschmann, and Jens Ahrens (2020). ‘Perceptual Evaluation of Mitigation Approaches of Impairments Due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data’. *Journal of the Audio Engineering Society* 68.6, pp. 428–440 (cited on pp. 60, 65, 76, 87).
- Lübeck, Tim, Christoph Pörschmann, and Johannes M. Arend (2020). ‘Perception of Direct Sound, Early Reflections, and Reverberation in Auralizations of Sparsely Measured Binaural Room Impulse Responses’. *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality* (cited on pp. 59, 88, 93, 98, 124).
- Majdak, Piotr, Matthew J. Goupell, and Bernhard Laback (2010). ‘3-D Localization of Virtual Sound Sources: Effects of Visual Environment, Pointing Method, and Training’. *Attention, Perception, & Psychophysics* 72.2, pp. 454–469 (cited on pp. 73, 80, 81).

- Marentakis, Georgios and Rudolfs Liepins (2014). ‘[Evaluation of Hear-through Sound Localization](#)’. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’14, pp. 267–270 (cited on p. 33).
- May, Tobias, Ning Ma, and Guy J. Brown (2015). ‘[Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues](#)’. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2679–2683 (cited on p. 88).
- May, Tobias, Steven van de Par, and Armin Kohlrausch (2011). ‘[A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End](#)’. *IEEE Transactions on Audio, Speech, and Language Processing* 19.1, pp. 1–13 (cited on p. 73).
- McCormack, Leo and Symeon Delikaris-Manias (2019). ‘[Parametric First-Order Ambisonic Decoding for Headphones Utilising the Cross-Pattern Coherence Algorithm](#)’. *EAA Spatial Audio Signal Processing Symposium*, pp. 173–178 (cited on p. 63).
- McKeag, Adam and David S. McGrath (1996). ‘[Sound Field Format to Binaural Decoder with Head Tracking](#)’. *Audio Engineering Society Convention 6r* (cited on pp. 60, 68, 97, 99, 104).
- McKenzie, Thomas, Damian T. Murphy, and Gavin Kearney (2018). ‘[Diffuse-Field Equalisation of Binaural Ambisonic Rendering](#)’. *Applied Sciences* 8.10 (10), p. 1956 (cited on pp. 60, 64, 65, 68, 72).
- McKenzie, Thomas, Damian T. Murphy, and Gavin Kearney (2019a). ‘[An Evaluation of Pre-Processing Techniques for Virtual Loudspeaker Binaural Ambisonic Rendering](#)’. *EAA Spatial Audio Signal Processing Symposium*, pp. 149–154 (cited on pp. 60, 63, 65, 68, 69, 87).
- McKenzie, Thomas, Damian T. Murphy, and Gavin Kearney (2019b). ‘[Interaural Level Difference Optimisation of First-Order Binaural Ambisonic Rendering](#)’. *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio* (cited on pp. 64, 72).
- Mendonça, Catarina, Guilherme Campos, Paulo Dias, José Vieira, João P. Ferreira, and Jorge A. Santos (2012). ‘[On the Improvement of Localization Accuracy with Non-Individualized HRTF-Based Sounds](#)’. *Journal of the Audio Engineering Society* 60.10, pp. 821–830 (cited on pp. 25, 30).

- Mh acoustics (2016). ‘[Eigenbeam Datasheet](#)’ (cited on p. 103).
- Miller, Antonio J. and Ravish Mehra (2019). ‘[Cartilage Conduction Audio System for Eyewear Devices](#)’. US10231046B1 (cited on p. 130).
- Møller, Henrik (1992). ‘[Fundamentals of Binaural Technology](#)’. *Applied Acoustics* 36.3, pp. 171–218 (cited on pp. 4, 38, 58).
- Møller, Henrik, Dorte Hammershøi, Clemen Boje Jensen, and Michael Friis Sørensen (1995). ‘[Transfer Characteristics of Headphones Measured on Human Ears](#)’. *Journal of the Audio Engineering Society* 43.4, pp. 203–217 (cited on pp. 11, 39).
- Møller, Henrik, Michael Friis Sørensen, Clemen Boje Jensen, and Dorte Hammershøi (1996). ‘[Binaural Technique: Do We Need Individual Recordings?](#)’ *Journal of the Audio Engineering Society* 44.6, pp. 451–469 (cited on p. 53).
- Moore, Brian C. J. (2012). ‘An Introduction to the Psychology of Hearing’. Brill (cited on p. 106).
- Noisternig, Markus, Alois Sontacchi, Thomas Musil, and Robert Holdrich (2003). ‘[A 3D Ambisonic Based Binaural Sound Reproduction System](#)’. *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality* (cited on pp. 68, 94, 99).
- Nowak, Johannes and Stefan Klockgether (2017). ‘[Perception and Prediction of Apparent Source Width and Listener Envelopment in Binaural Spherical Microphone Array Auralizations](#)’. *The Journal of the Acoustical Society of America* 142.3, pp. 1634–1645 (cited on pp. 59, 107, 108, 109, 122).
- Oberem, Josefa, Bruno Masiero, and Janina Fels (2016). ‘[Experiments on Authenticity and Plausibility of Binaural Reproduction via Headphones Employing Different Recording Methods](#)’. *Applied Acoustics* 114, pp. 71–78 (cited on p. 39).
- Okano, Toshiyuki, Leo L. Beranek, and Takayuki Hidaka (1998). ‘[Relations among Interaural Cross-Correlation Coefficient \(IACCE\), Lateral Fraction \(LFE\), and Apparent Source Width \(ASW\) in Concert Halls](#)’. *The Journal of the Acoustical Society of America* 104.1, pp. 255–265 (cited on p. 107).

- Olive, Sean E. and Floyd E. Toole (1989). ‘[The Detection of Reflections in Typical Rooms](#)’. *Journal of the Audio Engineering Society* 37.7/8, pp. 539–553 (cited on pp. 2, 95, 106).
- Olive, Sean E., Todd Welti, and Elisabeth McMullin (2013). ‘[Listener Preferences for Different Headphone Target Response Curves](#)’. *Audio Engineering Society Convention 134* (cited on p. 39).
- Oppenheim, Alan V., John R. Buck, and Ronald W. Schafer (2001). ‘Discrete-Time Signal Processing. Vol.2’. Upper Saddle River, N.J. : Prentice Hall (cited on p. 101).
- Pelzer, Sönke, Lukas Aspöck, Dirk Schröder, and Michael Vorländer (2014). ‘[Interactive Real-Time Simulation and Auralization for Modifiable Rooms](#)’. *Building Acoustics* 21.1, pp. 65–73 (cited on p. 96).
- Picinali, Lorenzo, Alexander Wallin, Yuli Levtoy, and David Poirier-Quinot (2017). ‘[Comparative Perceptual Evaluation between Different Methods for Implementing Reverberation in a Binaural Context](#)’. *Audio Engineering Society Convention 142* (cited on pp. 98, 99, 100, 102).
- Poletti, Mark (2009). ‘[Unified Description of Ambisonics Using Real and Complex Spherical Harmonics](#)’. *Ambisonics Symposium 2009* (cited on p. 16).
- Pörschmann, Christoph, Johannes M. Arend, and Fabian Brinkmann (2019). ‘[Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling](#)’. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.6, pp. 1060–1071 (cited on p. 60).
- Pralong, Danièle and Simon Carlile (1996). ‘[The Role of Individualized Headphone Calibration for the Generation of High Fidelity Virtual Auditory Space](#)’. *The Journal of the Acoustical Society of America* 100.6, pp. 3785–3793 (cited on pp. 5, 11, 38, 39, 46).
- Rafaely, Boaz (2005). ‘[Analysis and Design of Spherical Microphone Arrays](#)’. *IEEE Transactions on Speech and Audio Processing* 13.1, pp. 135–143 (cited on p. 18).
- Rafaely, Boaz (2015). ‘[Fundamentals of Spherical Array Processing](#)’. Springer (cited on pp. 15, 16, 59).

- Rafaely, Boaz and Amir Avni (2010). ‘[Interaural Cross Correlation in a Sound Field Represented by Spherical Harmonics](#)’. *The Journal of the Acoustical Society of America* 127.2, pp. 823–828 (cited on pp. 14, 15, 16).
- Rämö, Jussi and Vesa Välimäki (2012). ‘[Digital Augmented Reality Audio Headset](#)’. *Journal of Electrical and Computer Engineering* 2012, p. 457374 (cited on p. 29).
- Reijnen, Jonas, Dieter Vanderelst, Craig Jin, Simon Carlile, and Herbert Peremans (2014). ‘[An Ideal-Observer Model of Human Sound Localization](#)’. *Biological Cybernetics* 108.2, pp. 169–181 (cited on pp. 20, 61, 73).
- Romigh, Griffin D., Douglas S. Brungart, and Brian D. Simpson (2015). ‘[Free-Field Localization Performance With a Head-Track Virtual Auditory Display](#)’. *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 943–954 (cited on p. 40).
- Rona, Aldo, Martin Cockrill, Lorenzo Picinali, Prakash Panday, Ramesh Tripathi, and Mohammed Ayub (2015). ‘[Vibro-Acoustic Response of Tympanic-Membrane-like Models](#)’. *22nd International Congress on Sound and Vibration* (cited on pp. 25, 27).
- Schärer, Zora and Alexander Lindau (2009). ‘[Evaluation of Equalization Methods for Binaural Signals](#)’. *Audio Engineering Society Convention 126* (cited on p. 42).
- Schissler, Carl, Ravish Mehra, and Dinesh Manocha (2014). ‘[High-Order Diffraction and Diffuse Reflections for Interactive Sound Propagation in Large Environments](#)’. *ACM Transactions on Graphics* 33.4 (cited on p. 94).
- Schissler, Carl, Peter Stirling, and Ravish Mehra (2017). ‘[Efficient Construction of the Spatial Room Impulse Response](#)’. *2017 IEEE Virtual Reality (VR)*, pp. 122–130 (cited on pp. 2, 20, 59, 93, 96, 99).
- Schörkhuber, Christian, Markus Zaunschirm, and Robert Höldrich (2018). ‘[Binaural Rendering of Ambisonic Signals via Magnitude Least Squares](#)’. *Fortschritte Der Akustik–DAGA*, pp. 339–342 (cited on pp. 60, 67, 72, 75, 97).
- Schroeder, Manfred R. and Benjamin F. Logan (1961). ‘[“Colorless” Artificial Reverberation](#)’. *IRE Transactions on Audio* AU-9.6, pp. 209–214 (cited on pp. 92, 96).

- Sheaffer, Jonathan and Boaz Rafaely (2014). ‘Equalization Strategies for Binaural Room Impulse Response Rendering Using Spherical Arrays’. *2014 IEEE 28th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, pp. 1–5 (cited on p. 87).
- Simon, Laurent S. R., Nick Zacharov, and Brian F. G. Katz (2016). ‘Perceptual Attributes for the Comparison of Head-Related Transfer Functions’. *The Journal of the Acoustical Society of America* 140.5, pp. 3623–3632 (cited on p. 32).
- Solvang, Audun (2008). ‘Spectral Impairment of Two-Dimensional Higher Order Ambisonics’. *Journal of the Audio Engineering Society* 56.4, pp. 267–279 (cited on p. 124).
- Søndergaard, Peter L. and Piotr Majdak (2013). ‘The Auditory Modeling Toolbox’. *The Technology of Binaural Listening*. Ed. by Jens Blauert. Modern Acoustics and Signal Processing. Berlin, Heidelberg: Springer, pp. 33–56 (cited on p. 61).
- Stroud, Arthur H. and Don Secrest (1966). ‘Gaussian Quadrature Formulas’. Prentice-Hall (cited on p. 68).
- Strutt, John W. (1907). ‘XII. On Our Perception of Sound Direction’. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, pp. 214–232 (cited on p. 67).
- Tikander, Miikka (2009). ‘Usability Issues in Listening to Natural Sounds with an Augmented Reality Audio Headset’. *Journal of the Audio Engineering Society* 57.6, pp. 430–441 (cited on pp. 24, 25, 26, 27).
- Valimäki, Vesa, Julian D. Parker, Lauri Savioja, Julius O. Smith, and Jonathan S. Abel (2012). ‘Fifty Years of Artificial Reverberation’. *IEEE Transactions on Audio, Speech, and Language Processing* 20.5, pp. 1421–1448 (cited on p. 96).
- Välimäki, Vesa, Julian Parker, Lauri Savioja, Julius O. Smith, and Jonathan Abel (2016). ‘More Than 50 Years of Artificial Reverberation’. *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (cited on p. 92).
- Vilkamo, Juha, Tom Bäckström, and Achim Kuntz (2013). ‘Optimized Covariance Domain Framework for Time–Frequency Processing of Spatial Audio’. *Journal of the Audio Engineering Society* 61.6, pp. 403–411 (cited on p. 67).

- Wallach, Hans, Edwin B. Newman, and Mark R. Rosenzweig (1949). ‘[A Precedence Effect in Sound Localization](#)’. *The Journal of the Acoustical Society of America* 21.4, pp. 468–468 (cited on pp. [95](#), [97](#), [106](#)).
- Wanrooij, Marc M. Van and A. John Van Opstal (2005). ‘[Relearning Sound Localization with a New Ear](#)’. *Journal of Neuroscience* 25.22, pp. 5413–5424 (cited on pp. [25](#), [30](#)).
- Werner, Stephan, Florian Klein, Thomas Mayenfels, and Karlheinz Brandenburg (2016). ‘[A Summary on Acoustic Room Divergence and Its Effect on Externalization of Auditory Events](#)’. *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6 (cited on pp. [2](#), [73](#), [96](#), [112](#)).
- Wightman, Frederic L. and Doris J. Kistler (1989a). ‘[Headphone Simulation of Free-field Listening. I: Stimulus Synthesis](#)’. *The Journal of the Acoustical Society of America* 85.2, pp. 858–867 (cited on pp. [38](#), [58](#)).
- Wightman, Frederic L. and Doris J. Kistler (1989b). ‘[Headphone Simulation of Free-field Listening. II: Psychophysical Validation](#)’. *The Journal of the Acoustical Society of America* 85.2, pp. 868–878 (cited on p. [2](#)).
- Williams, Earl G. (1999). ‘[Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography](#)’. Academic Press (cited on p. [15](#)).
- Xie, Bosun (2013). ‘Head-Related Transfer Function and Virtual Auditory Display’. J. Ross Publishing (cited on p. [47](#)).
- Yadav, Manuj, Densil A. Cabrera, Luis Miranda, William L. Martens, Doheon Lee, and Ralph Collins (2013). ‘[Investigating Auditory Room Size Perception with Autophonic Stimuli](#)’. *Audio Engineering Society Convention 135* (cited on p. [96](#)).
- Zahorik, Pavel (2002). ‘[Direct-to-Reverberant Energy Ratio Sensitivity](#)’. *The Journal of the Acoustical Society of America* 112.5, pp. 2110–2117 (cited on p. [103](#)).
- Zaunschirm, Markus, Matthias Frank, and Franz Zotter (2018). ‘[BRIR Synthesis Using First-Order Microphone Arrays](#)’. *Audio Engineering Society Convention 144* (cited on p. [59](#)).
- Zaunschirm, Markus, Christian Schörkhuber, and Robert Höldrich (2018). ‘[Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness](#)

[Constraint](#)'. *The Journal of the Acoustical Society of America* 143.6, pp. 3616–3627 (cited on pp. [60](#), [63](#), [67](#), [97](#)).

Zhou, Yaxuan, Hao Jiang, and Vamsi K. Ithapu (2021). '[On the Predictability of Hrtfs from Ear Shapes Using Deep Networks](#)'. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 441–445 (cited on p. [131](#)).

Zotter, Franz and Matthias Frank (2019). '[Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality](#)'. Springer Nature (cited on pp. [3](#), [14](#), [15](#), [16](#), [59](#), [63](#), [66](#), [67](#), [68](#), [69](#), [92](#), [93](#), [94](#), [104](#)).

Appendix A

Proof of permission to reproduce published work

12 May 2021

Dear Secretariat of the 24th International Congress of Sound and Vibration (ICSV24),

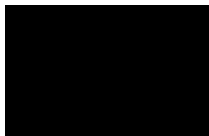
I am completing my PhD thesis at Imperial College London entitled 'Advances in spatial audio for Augmented Reality'.

I seek your permission to reproduce, in my thesis, the published version of a paper that I wrote and was published in the proceedings of ICSV24 with title 'Long-term User Adaptation to an Audio Augmented Reality System'. The contents of the paper would be adapted into a chapter of the thesis and a citation to the conference's proceedings would be provided at the beginning of said chapter.

The thesis will be added to Spiral, Imperial's institutional repository <http://spiral.imperial.ac.uk/> and made available to the public under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

If you are happy to grant me all the permissions requested, please return a signed copy of this letter. If you wish to grant only some of the permissions requested, please list these and then sign.

Yours sincerely,



Isaac Engel

Permission granted for the use requested above:

I confirm that I am the copyright holder of the paper above and hereby give permission to include it in your thesis to be made available, via the internet, for non-commercial purposes under the terms of the user licence.

[please edit the text above if you wish to grant more specific permission]

Signed:  

Name: Linda Canty

Organisation: Institute of Acoustics

Job title: Office Manager

12 May 2021

Dear Audio Engineering Society,

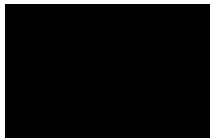
I am completing my PhD thesis at Imperial College London entitled 'Advances in spatial audio for Augmented Reality'.

I seek your permission to reproduce, in my thesis, the published version of a paper that I wrote and was published in the 2019 AES International Conference on Immersive and Interactive Audio, with title 'The Effect of Generic Headphone Compensation on Binaural Renderings'. The contents of the paper would be adapted into a chapter of the thesis and a citation and link to the published version would be provided at the beginning of said chapter.

The thesis will be added to Spiral, Imperial's institutional repository <http://spiral.imperial.ac.uk/> and made available to the public under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

If you are happy to grant me all the permissions requested, please return a signed copy of this letter. If you wish to grant only some of the permissions requested, please list these and then sign.

Yours sincerely,



Isaac Engel

Permission granted for the use requested above:

I confirm that I am the copyright holder of the paper above and hereby give permission to include it in your thesis to be made available, via the internet, for non-commercial purposes under the terms of the user licence.

Signed:



Name: Colleen Harper

Organisation: Audio Engineering Society

Job title: Executive Director