WAVE FIELD SYNTHESIS IN A LISTENING ROOM

Dissertation
zur
Erlangung des akademischen Grades
Doktor-Ingenieurin (Dr.-Ing.)
der Fakultät für Informatik und Elektrotechnik
der Universität Rostock

vorgelegt von
Vera Erbes, geb. am 23.02.1984 in Karlsruhe
aus Berlin

Berlin, 24.09.2020

https://doi.org/10.18453/rosdok_id00002763

Dieses Werk ist lizenziert unter einer
Creative Commons Namensnennung 4.0 International Lizenz.

Datum der Einreichung: 11.05.2020
Datum der Verteidigung: 21.08.2020

Gutachter: Prof. Dr.-Ing. Sascha Spors
Institut für Nachrichtentechnik
Forschungsgruppe Signaltheorie und digitale Signalverarbeitung
Universität Rostock

Prof. Dr. phil. Stefan Weinzierl
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation
Technische Universität Berlin

This work is licenced under CC BY 4.0 except where noted otherwise.

Acknowledgements

I would like to thank

my dissertation adviser Prof. Dr.-Ing. Sascha Spors for his continuous support, inspirational advice, establishing an enjoyable working environment and for giving me the space and freedom to truly dive into my research.

Prof. Dr. Stefan Weinzierl who introduced me to the field of spatial audio in an inspirational and truly multidisciplinary team.

my former colleagues in Rostock and Berlin for their support and humour during long days of thinking and long nights of measurements.

all participants in the listening experiments for their time and dedication.

Andreas Fleck for just making things work and being a long-term friend.

my family and friends for their endless support and for believing in me.

my love Christian who always was my solid rock during the tumultuous time it took to write this thesis.

Contents

1 Introduction **1**
- 1.1 Objective of this thesis . . . 2
- 1.2 Structure of this thesis . . . 3

2 Fundamentals of Wave Field Synthesis **5**
- 2.1 Theory of Wave Field Synthesis . . . 5
  - 2.1.1 Fundamental concept of sound field synthesis . . . 5
  - 2.1.2 Simplification to 2.5D synthesis . . . 8
  - 2.1.3 Derivation of the Wave Field Synthesis driving function for a virtual point source with a linear array . . . 9
  - 2.1.4 Practical limitations . . . 12
    - 2.1.4.1 Discretisation of the secondary source distribution . 12
    - 2.1.4.2 Truncation of the secondary source distribution . . 20
    - 2.1.4.3 Reflective environment . . . 20
- 2.2 Perception of Wave Field Synthesis . . . 23
  - 2.2.1 Perceptual dimensions of spatial audio . . . 23
    - 2.2.1.1 Scope and completeness of vocabularies . . . 23
    - 2.2.1.2 Adequacy of reporting methods . . . 24
    - 2.2.1.3 Applicability of room acoustical parameters . . . 25
    - 2.2.1.4 Conclusions for the investigations in this thesis . . . 25
  - 2.2.2 Spatial perception of Wave Field Synthesis . . . 25
  - 2.2.3 Timbral perception of Wave Field Synthesis . . . 27

3 Methods **29**
- 3.1 Objectivity, validity and reliability . . . 29
- 3.2 Investigation through simulation . . . 31
  - 3.2.1 Binaural synthesis . . . 32
    - 3.2.1.1 Fundamentals of binaural synthesis . . . 32
    - 3.2.1.2 Employed head-related transfer functions and their processing . . . 37
  - 3.2.2 Room acoustical simulation . . . 38
    - 3.2.2.1 Choice of room acoustical simulation . . . 38
    - 3.2.2.2 Parameters of the image source method . . . 40
  - 3.2.3 Generating impulse responses of Wave Field Synthesis systems 42
- 3.3 Evaluation of the reporting method for the localisation experiment . 43
  - 3.3.1 Apparatus for the localisation experiment . . . 43
  - 3.3.2 Conditions . . . 44
  - 3.3.3 Realisation of the listening experiment . . . 45
    - 3.3.3.1 Experimental procedure . . . 45
    - 3.3.3.2 Participants . . . 45

3.3.4 Procedure for data analysis . . . 46
3.3.4.1 Data correction . . . 46
3.3.4.2 Statistical model for data analysis . . . 47
3.3.4.3 Comparison of standard deviations . . . 48
3.3.5 Results . . . 48
3.3.5.1 Comparison of the two tracking systems . . . 48
3.3.5.2 Localisation error . . . 48
3.3.5.3 Elapsed time per trial . . . 50
3.3.6 Discussion . . . 51
3.3.7 Summary . . . 52
3.4 Conducting colouration experiments . . . 52
3.4.1 Binaural synthesis in colouration experiments . . . 52
3.4.2 Experimental design for the investigation of colouration . . . 53

4 Spatial perception of Wave Field Synthesis in a listening room 55
4.1 Conditions . . . 56
4.2 Realisation of the listening experiment . . . 58
4.2.1 Experimental procedure . . . 58
4.2.2 Participants . . . 58
4.3 Procedure for data analysis . . . 59
4.3.1 Data correction . . . 59
4.3.2 Statistical model for data analysis . . . 60
4.4 Results . . . 61
4.4.1 Direction of localisation error . . . 61
4.4.2 Accuracy of localisation . . . 63
4.4.3 Difficulty of localisation . . . 63
4.5 Discussion . . . 64
4.6 Summary . . . 66

5 Timbral perception of Wave Field Synthesis in a listening room 67
5.1 Conditions . . . 68
5.2 Realisation of the listening experiment . . . 69
5.2.1 Experimental design and sample size estimation . . . 69
5.2.2 Experimental procedure . . . 70
5.2.3 Participants . . . 71
5.3 Data processing . . . 71
5.4 Results . . . 71
5.4.1 Rating results . . . 71
5.4.2 Influence of the listening room . . . 74
5.5 Discussion . . . 75
5.6 Summary . . . 80

6 Conclusions 81

Appendices 85
Appendix A – Coordinate System . . . 85
Appendix B – Conventions for the Fourier transform . . . 86
Appendix C – List of Open Science contributions . . . 87

List of abbreviations **89**

Bibliography **91**

Abstract **109**

Zusammenfassung **111**

Selbstständigkeitserklärung **113**

Curriculum vitae **115**

List of Figures

1.1 Illustration of the available listening area (shaded in blue) for two-channel stereophony and for SFS. Figure taken from [Wie14, Fig. 1.4], licence: CC BY 3.0 DE. The figure of the stereophony setup is a modified version of [Ahr12, Fig. 1.1]. 2

1.2 SFS in a listening room that is sending reflections back into the listening area (shaded in blue). 3

2.1 Geometry for the discussion of the fundamental concept of SFS for the desired sound field $S(\mathbf{x}, \omega)$ of a virtual source. Explanation of other symbols in text. Figure taken from [Wie14, fig. 2.1], licence: CC BY 3.0 DE. 6

2.2 Cross section of the sound pressure of monochromatic virtual sources synthesised with 3D WFS and a spherical secondary source distribution (black line). The driving functions are given in [SRA08, eq. (17)] and [SRA08, eq. (19)] for the virtual plane wave and the virtual point source, respectively. The reference point of the driving function is $\mathbf{x}_\text{ref} = (0, 0, 0)^\text{T}$ m, cf. eq. (2.18), $f = 1$ kHz. The sound fields are normalised to the origin of coordinates. 7

2.3 Sound pressure (upper row) and sound pressure level (lower row) of a real (left column) and a virtual (middle column) monochromatic point source. The point source is located at $\mathbf{x}_\text{ps} = (0, 1, 0)^\text{T}$ m and has a frequency of $f = 1$ kHz. The virtual source is synthesised with 2.5D WFS and a linear secondary source array (black line) with the driving function in eq. (2.18) and $\mathbf{x}_\text{ref} = (0, -1, 0)^\text{T}$ m. The right plot compares the levels of the real and virtual point source over distance from the secondary source array along the line $\mathbf{x} = (0, y, 0)^\text{T}$. 10

2.4 Considered geometry for the derivation of the WFS driving functions for a planar secondary source distribution in the x, z-plane and for a linear secondary source distribution along the x-axis. The listening area is shaded in blue. The contribution of all secondary sources along the red line can be approximated by the contribution of the secondary source on the intersection of this line with the x-axis by the stationary phase approximation. The figure is adapted from [Ver97, fig. 2.6]. . . 11

2.5 Principle of sampling, aliasing and reconstruction. Left: Magnitude spectra $|X|$ of a time signal and its time-sampled version $|X_\mathrm{s}|$ with sampling rate f_s. H_rec is the ideal reconstruction filter. Right: Wave number spectra $|\tilde{D}|$ of the 2.5D WFS driving function for a virtual point source and its spatially sampled version $|\tilde{D}_\mathrm{s}|$ sampled every Δx_0 along a linear array of secondary sources. $|\tilde{G}_{0,\mathrm{3D}}|$ is the reconstruction filter in the wave number domain when using spherical monopoles as secondary sources. Shaded areas denote a higher magnitude of the non-bandlimited functions. 13

2.6 Magnitude of the wave number spectrum of the Green's function in eq. (2.3) with $\mathbf{x}_0 = (0,0,0)^\mathrm{T}$ m evaluated at $y = 1$ m and $z = 0$. The spectrum is calculated according to [AS10, eq. (52)]. 14

2.7 Sound pressure of a virtual monochromatic point source at $\mathbf{x}_\mathrm{ps} = (0,1,0)^\mathrm{T}$ synthesised with 2.5D WFS with driving function eq. (2.18) and a linear secondary source array with $\Delta x_0 = 0.2$ m and $\mathbf{x}_\mathrm{ref} = (0,-1,0)^\mathrm{T}$ m. Left: $f = 600$ Hz, middle: $f = 1.2$ kHz, right: $f = 2.4$ kHz. The sound fields are normalised to the reference position $\mathbf{x}_\mathrm{ref}$. 14

2.8 Magnitude of the frequency responses in dB of 2.5D WFS with driving function eq. (2.18) at $\mathbf{x}_\mathrm{ref}$ with a linear secondary source array and four different Δx_0 as given in the figure. The spectra are normalised and shifted for better discriminability. The aliasing frequency for each Δx_0 according to eq. (2.21) is marked with a dashed line. 15

2.9 Level of the sound pressure of a point source at $\mathbf{x}_\mathrm{ps} = (0,1,0)^\mathrm{T}$ m emitting a broadband impulse synthesised by 2.5D WFS with driving function eq. (2.18) and $\mathbf{x}_\mathrm{ref} = (0,-1,0)^\mathrm{T}$ m at the time $t = 8.7$ ms. Left: secondary source distance $\Delta x_0 = 0.5$ cm, middle: $\Delta x_0 = 20$ cm, right: $\Delta x_0 = 50$ cm. The sound fields are normalised to the maximum value in the plots. The slightly raised level before the first wave front is caused by the pre-equalisation filter, that in practical implementations is chosen as a shelving filter, cf. end of section 2.1.4.1. 15

2.10 Impulse responses of a real (left) and a virtual point source with $\Delta x_0 = 50$ cm (middle) and $\Delta x_0 = 20$ cm (right) synthesised by 2.5D WFS with driving function eq. (2.18) and a linear array of 10 m length along the x-axis. The virtual point sources are at $\mathbf{x}_\mathrm{ps} = (0,1,0)^\mathrm{T}$ m, the impulse responses are given for $\mathbf{x}_\mathrm{ref} = (0,-1,0)^\mathrm{T}$ m. The impulse response of the real source is given for the same relative distance of 2 m. 16

2.11 Sound pressure in Pa (upper row) and its level in $\mathrm{dB_{SPL}}$ (lower row) of a real monochromatic point source at $\mathbf{x}_{\mathrm{ps}} = (0, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ (left column) and virtual monochromatic point sources synthesised by 2.5D WFS with $\Delta x_0 = 20\,\mathrm{cm}$, that are positioned close to the linear secondary source array (black dots) and either right behind a secondary source at $\mathbf{x}_{\mathrm{ps}} = (0, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ (middle column) or between two secondary sources at $\mathbf{x}_{\mathrm{ps}} = (0.1, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ (right column). The driving function for WFS reproduction is derived by SDM [SA10a, eq. (24)]. The reference line for correct synthesis is marked by the dashed line. The frequency of the point sources is $f = 350\,\mathrm{Hz}$. The amplitude of the real prototype point source is chosen to yield $94\,\mathrm{dB_{SPL}}$ at $\mathbf{x} = (0, -1, 0)^{\mathrm{T}}\,\mathrm{m}$ on the reference line. At the same point, the virtual sources exhibit a sound pressure level of $104\,\mathrm{dB_{SPL}}$ in the middle column and $83\,\mathrm{dB_{SPL}}$ in the right column. 17

2.12 Sound field components according to table 2.1 for a virtual monochromatic point source with $f = 350\,\mathrm{Hz}$ at $\mathbf{x}_{\mathrm{ps}} = (0, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ synthesised by 2.5D WFS and a linear array with $\Delta x_0 = 20\,\mathrm{cm}$. The driving function for WFS reproduction is derived by SDM [SA10a, eq. (24)]. Sound field components are calculated by [EWS15, eq. (5)–(8)]. All four components together form the sound field in fig. 2.11, middle column. 18

2.13 Sound field components according to table 2.1 for a virtual monochromatic point source with $f = 350\,\mathrm{Hz}$ at $\mathbf{x}_{\mathrm{ps}} = (0.1, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ synthesised by 2.5D WFS and a linear array with $\Delta x_0 = 20\,\mathrm{cm}$. The driving function for WFS reproduction is derived by SDM [SA10a, eq. (24)]. Sound field components are calculated by [EWS15, eq. (5)–(8)]. All four components together form the sound field in fig. 2.11, right column. 19

2.14 Sound pressure of a monochromatic plane wave in direction $\mathbf{n}_{\mathrm{pw}} = (0, -1, 0)^{\mathrm{T}}$ synthesised by 2.5D WFS with driving function [SRA08, eq. (27)] and a truncated linear secondary source distribution (black line). $\mathbf{x}_{\mathrm{ref}} = (0, -1, 0)^{\mathrm{T}}\,\mathrm{m}$, $f = 1\,\mathrm{kHz}$. Left: without tapering window, right: with raised-cosine shaped tapering window applied over 30% of the array length. The sound fields are normalised to the centre of the plots. 20

2.15 Raised-cosine shaped tapering window as applied in fig. 2.14, right-hand side. The outer secondary sources covering 30% of the array length, i.e. 15% on each side, are faded out towards the edges. . . . 21

2.16 Impulse responses of a real point source (upper row) and a virtual point source (lower row) in free field (left column) and inside a rectangular listening room (right column). The virtual source is synthesised by 2.5D WFS with driving function eq. (2.18). The geometry of the setup is depicted in fig. 4.1 with a linear array of 15 secondary sources with $\Delta x_0 \approx 20\,\mathrm{cm}$, impulse responses are evaluated at the centre listening position $\mathbf{x}_{\mathrm{ref}}$. The room is simulated with the image source method (see section 3.2.2) and a constant reflection factor of $\beta = 0.7$. 22

2.17 Absolute value of the magnitude spectrum $X(f)$ of the signal $x(t) = \delta(t) + \delta(t - t_0)$ illustrating a comb filter. The delay is $t_0 \approx 0.23\,\mathrm{ms}$. The resulting distance between minima of $|X(f)|$ is $\frac{1}{t_0} = 4410\,\mathrm{Hz}$. . . 27

3.1 Illustration of the test-retest reliability in a scatter plot. In this example Pearson's correlation coefficient is approx. 0.96, as the blue data points are close to the ideal black line. This indicates a very good test-retest reliability of the underlying test method. 30

3.2 Mean HpTF of the FABIAN HATS with headphones type AKG K601, averaged out of 21 measurements (12 measurements for left and right channel each, three outliers removed), headphone compensation filter and compensated result. The ripple in the frequency response of the compensated result is caused by the design of the target bandpass as a linear-phase FIR filter with a least-squares approach and windowing [EGWS17]. 34

3.3 HRTFs and head-related impulse responses (HRIRs) of a KEMAR manikin for 0° azimuth in the horizontal plane, left ear, before and after low-frequency correction. 35

3.4 Linear interpolation in the frequency domain of two HRTFs for 0° and 20° azimuth to an HRTF at 10° azimuth. The ground truth at 10° azimuth is given as well. Left: HRTFs with corrupt phase at low frequencies, right: HRTFs with low-frequency correction. 36

3.5 Image that is warped to a simple spherical grid as visual virtual reality (VR) presented on the HMD in the listening experiment on localisation, the frontal direction is in the centre of the right half. 44

3.6 Location of virtual sources (loudspeaker symbols) and listener in the evaluation experiment for localisation, only source positions in black are used. 45

3.7 Listener in the experimental setup for the localisation experiment. . 46

3.8 Comparison of azimuthal data of the two utilised tracking systems in the evaluation experiment for localisation during rotation of the head of the listener from 90° to −90° azimuth, resting in between. 48

3.9 The rows show the arithmetic mean (left column) and standard deviation (right column) of the signed localisation error for the three compared studies on evaluation of localisation. Blue markers show results of individual subjects (10 subjects per study), black markers show results for all subjects together with the 95% confidence interval. The black lines in the left column represent the mixed models with the best fit for the signed localisation errors. 49

3.10 Normalised histogram for the elapsed time per trial in the evaluation experiment for localisation. 50

4.1 Geometry of sources and receivers in listening room (black border) in the localisation listening experiment. The grey dot indicates the position of the real or virtual source, the black dots the secondary sources for WFS reproduction. The secondary source distances depending on the used number of secondary sources in a condition are: approx. 20 cm for all 15 secondary sources, 41 cm for 8 sources and 143 cm for 3 sources. The black crosses mark the listener positions labelled L1 to L3. Room height is 3 m. All sources and receivers are positioned 1.59 m above the floor. For free field conditions, no room is present. For conditions with a real source, no secondary sources are present. Figure taken from [ES20], © 2020 IEEE. 57

4.2 IC-weighted histograms of the ITD (red) for frequencies below 1.4 kHz and the ILD (light blue) and ITD of the envelope (dark blue) above this threshold for an auralisation of the 0° azimuth/0° elevation FABIAN HRTFs with white noise. 59

4.3 Average results of the listening experiment on localisation. The grey dots indicate the positions of the real or virtual sources, the black dots in the case of WFS reproduction the secondary sources. The first line of plots without secondary sources displays the case of real sources. In columns, the results are ordered by amount of reverberation, ranging from free field up to a rectangular room with a reflection factor of $\beta = 0.9$ for all boundaries. At each listener position, an arrow is pointing in the direction of the average auditory event. The colour of the arrow visualises the localisation error as difference of the average auditory event from the real or virtual source position as indicated by the colour bar. The localisation error is also stated below each listener position in the top line. The grey rays starting at each listener position display the 95% confidence interval which is also stated below each listener position in the lower line. Figure taken from [ES20], © 2020 IEEE. 62

5.1 Geometry of the WFS setup in the colouration listening experiment. The grey dot indicates the position of the virtual source, the black dots the secondary sources. The black cross marks the listener position. The black borders show the walls of the listening room. Room height is 3 m. All sources and the receiver are positioned 1.59 m above the floor. 68

5.2 Graphical user interface in the rating phase of the listening experiment on colouration. Text translated from originally German. 70

5.3 Medians and 96.4% confidence intervals of colouration ratings for noise (left) and speech (right) as audio content. The WFS conditions are named according to their reflective environments. 72

5.4 Magnitude of the long-term spectrum of the speech signal used in the colouration experiment (blue) in comparison to the magnitude of the single-channel transfer function of the WFS system in free field to the listener position (red). The spectrum of the speech signal is normalised to its maximum, the transfer function of the WFS system is shifted for better discriminability. The vertical black line denotes the lower limit spatial aliasing frequency of the WFS system according to eq. (2.21). 75
5.5 Magnitude responses of RIRs corresponding to the six conditions of the listening experiment. The WFS responses are smoothed in third-octave bands, all responses are shifted along the y-axis for better discriminability. The shaded areas enclose the mean and mark the ± standard deviation (STD) borders of the WFS magnitude responses. 76
5.6 Magnitude responses of a virtual point source at the receiver position for different absorber configurations in the Audio Lab at the University of Rostock and in free field as given by the geometry in fig. 5.7a. The absorber configurations have been converted to uniform reflection coefficients as stated in the legend. The magnitude responses are smoothed in third-octave bands and shifted along the y-axis for better discriminability. 77
5.7 WFS system in the Audio Lab at the University of Rostock. (a) Employed geometry of the system. The grey dot indicates the position of the virtual source, the black dots the secondary sources (exact positions are not equidistantly spaced as depicted here, cf. [EGWS15]). The black cross marks the listener position. The black borders show the walls of the listening room. Room height is 3 m. All sources and the receiver are positioned 1.59 m above the floor. (b) Loudspeaker array in the Audio Lab. In this acoustic condition, the walls are equipped with broadband absorbers in black in the upper and pyramid-shaped foam in the lower half. 78
5.8 Low frequency magnitude responses of virtual sources for the WFS setup in fig. 5.7a. The virtual point source is positioned as depicted in fig. 5.7a, the virtual plane wave is travelling perpendicular to the array. The responses have been normalised to the lowest mode at approx. 61 Hz. 79
5.9 Medians and 95.9% (group 1) and 96.1% (group 2) confidence intervals of colouration ratings for speech as audio content divided in two groups. The WFS conditions are named according to their reflective environments. 79

A.1 Coordinate system used in this thesis. 85

List of Tables

2.1 Four spatial aliasing components of a synthesised sound field as analysed by [SA09, EWS15]. 17

3.1 Estimated latencies for the binaural system used in the listening experiments on localisation. The sampling rate is $f_\mathrm{s} = 44.1\,\mathrm{kHz}$. 38

3.2 Schroeder frequencies calculated for the rooms simulated in the listening experiments with $V = 210\,\mathrm{m}^3$ according to eq. (3.3). $\alpha = 1 - \beta^2$ is the absorption coefficient for all boundaries. 41

3.3 Descriptive statistics on elapsed time per trial in s in the evaluation experiment for localisation compared to previous studies. 51

4.1 Slope coefficients for fixed effects of the mixed-effects model with best fit for predicted variable *'signed localisation error'* together with p-values and estimated effect sizes f^2 in the localisation experiment. Table taken from [ES20], © 2020 IEEE. 63

4.2 Slope coefficients for fixed effects of the mixed-effects model with best fit for predicted variable *'absolute deviation from mean perceived direction'* together with p-values and estimated effect sizes f^2 in the localisation experiment. Table taken from [ES20], © 2020 IEEE. . . 64

4.3 Slope coefficients for fixed effects of the mixed-effects model with best fit for predicted variable *'elapsed time per trial'* together with p-values and estimated effect sizes f^2 in the localisation experiment. Table taken from [ES20], © 2020 IEEE. 64

5.1 Undershot p-values and effect sizes r_C for the matched-pairs Wilcoxon signed rank test for *noise* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold. 73

5.2 Undershot p-values and effect sizes r_C for the matched-pairs Wilcoxon signed rank test for *speech* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold. 73

5.3 p-values and differences of medians (value of row minus value of column) for the Conover post-hoc test for *noise* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold. 74

5.4 *p*-values and differences of medians (value of row minus value of column) for the Conover post-hoc test for *speech* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold. 74

1 Introduction

Sound field synthesis (SFS) techniques such as Wave Field Synthesis (WFS) [Ber88, Ver97, SRA08] represent the most advanced methods in the field of spatial audio [Rum01]. After monophonic audio reproduction has become possible, it appeared desirable to extend the auditive experience to include spatial features as well. In natural sound fields, these spatial aspects are inherently present. They include the position and dimension of a source as well as reflections that give a sense of the surrounding environment. These spatial aspects become especially important in venues for musical performance where the piece played by musicians on a stage is enhanced by the acoustics of the concert hall [Ber04]. A recording of such a performance that is reproduced should ideally convey the same auditory spatial impression that a listener experienced sitting in the audience [Rum02]. As of today, state-of-the-art loudspeaker-based reproduction techniques can at least come close to this goal.

The first spatial audio technique that is still omnipresent today, probably also due to its simplicity, is two-channel stereophony which Blumlein was granted a patent for as early as 1931 [Blu31]. Stereophonic reproduction relies on two loudspeakers only that are ideally arranged in an equilateral triangle with the listener [DIN08]. The sound field that arises from the superposition of the outputs of the loudspeakers, that are fed with the same source signal, creates the impression of a so-called phantom source for the listener, that appears in-between the loudspeakers [Bla97]. The exact position depends on the time or level differences (so-called ‘panning’) of the two channels. Since then, the stereophonic method has been extended to using more channels, distinct milestones being quadraphony [Woo77] that has become popular in the 1970s and Vector Base Amplitude Panning (VBAP) [Pul97] which generalises the stereophonic panning concept to arbitrary 2D and 3D loudspeaker arrangements using up to three loudspeakers to create a phantom source. The most well-known representative today is 5.1 surround sound, where five loudspeakers are arranged around the listener and an additional subwoofer is used for low-frequency reproduction [ITU12].

All stereophonic reproduction methods and the traditional Ambisonics have in common that they have an ideal listener position, usually termed the ‘sweet spot’ where the auditory impression is perceived as desired. In contrast, SFS techniques aim at creating an extended listening area as they try to physically correctly (re-)construct an entire sound field [Wit07], cf. fig. 1.1. Unfortunately, this goal can only be reached in mathematical theory. The most severe limitations come from the requirement to have an infinitely dense loudspeaker distribution surrounding the volume where the sound field is to be synthesised in. In practice, this is prohibited by the spatial extent of real-world transducers. This constitutes spatial sampling, which leads to artefacts in the synthesised sound field known as spatial aliasing [SR06].

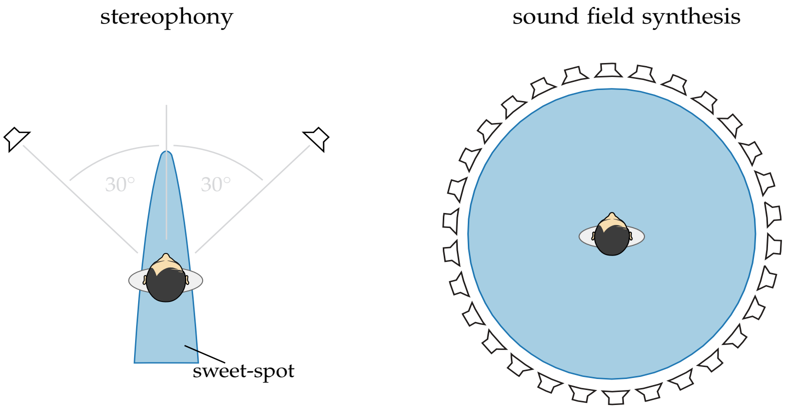


Figure 1.1: Illustration of the available listening area (shaded in blue) for two-channel stereophony and for SFS. Figure taken from [Wie14, Fig. 1.4], licence: CC BY 3.0 DE. The figure of the stereophony setup is a modified version of [Ahr12, Fig. 1.1].

Theory of SFS also assumes an anechoic environment [Wil99], but as loudspeaker arrays are typically installed inside listening rooms, reflections are unavoidable and change the desired synthesised sound field as is illustrated in fig. 1.2. To reduce reflections, the listening room can be equipped with acoustic absorbent material. In the most extreme case, an anechoic chamber is used as the listening room. Still, there are many application areas that do not permit this strategy, such as cinemas, museums or multi-purpose rooms. The largest WFS system in the world for example is installed inside a lecture hall at the Technische Universität Berlin [MGM+07]. This hall should therefore fulfil both the requirement of SFS theory to be as anechoic as possible and also support speech intelligibility by early reflections [BAM07]. Mobile systems for SFS are even installed inside different rooms, where acoustical properties cannot be changed by constructional modifications. A different approach to reduce reflections is taken by active compensation algorithms which is an ongoing research topic, e.g. [Spo06, Cor06, GB07, SK12]. Due to a finite number of microphones and loudspeakers used to monitor and control the sound field, these methods cannot provide a perfect solution either [CN03, SRR05].

As has been described above, the practical limitations of SFS can only be amended to some extent. Therefore, artefacts and physical deviations from the desired sound field are unavoidable. The recipient of the reproduced sound is not a physically accurate measurement device, though, but a human listener. His perception might be sensitive to some artefacts, while others go by unnoticed. It is therefore important in a first step to know how the deviations in the desired sound field are perceived and to what extent artefacts can be tolerated.

1.1 Objective of this thesis

This thesis investigates the influence of the listening room on sound fields synthesised by WFS, both in the physical and the perceptual domain. The provided insights help to improve realisations of this spatial audio reproduction method. The focus lies on the perceptual aspects of colouration and localisation as these appear to be of

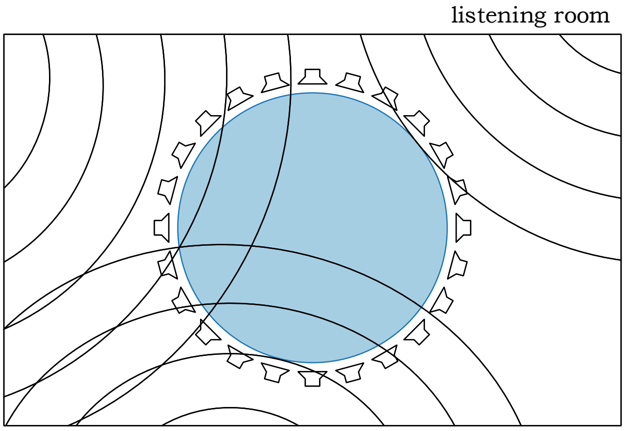


Figure 1.2: SFS in a listening room that is sending reflections back into the listening area (shaded in blue).

highest importance for the overall quality of spatial audio [RZKB05]. Special care is invested in the choice of methods to preserve the objectivity, validity and reliability of the research results.

1.2 Structure of this thesis

Following this introduction, the theory of SFS is reviewed with a focus on WFS in chapter 2. The practical limitations of WFS and their consequences on the synthesised sound fields are demonstrated. Concluding the chapter, the findings in the research literature on the perceptual dimensions of spatial audio and especially on spatial and timbral perception in WFS are summarised.

Chapter 3 gives a detailed account of the chosen methods for investigation. The methodological approach in this thesis is integrated into the three fundamental concepts from classical test theory: objectivity, validity and reliability. The decision to perform listening experiments based on simulated stimuli is motivated. The fundamentals and chosen parameters of the simulation methods binaural synthesis and room acoustical simulation are described. The reporting method and apparatus for the listening experiment on localisation is evaluated in a preliminary listening experiment. The use of binaural synthesis and the experimental design in colouration experiments are discussed.

The evaluation of spatial and timbral perception of WFS in a listening room is reported in chapter 4 and 5, respectively. Spatial perception is investigated in terms of azimuthal localisation. The results are analysed with a linear mixed-effects model that additionally provides quantitative effects of the investigated aspects of localisation. Timbral perception is evaluated as difference in timbre compared to a reference. Chapter 6 summarises this thesis and gives indications for future research.

2 Fundamentals of Wave Field Synthesis

This chapter summarises the theory and the state of the art in research of Wave Field Synthesis. This includes both the mathematical foundations as well as perceptual aspects of this reproduction method. Section 2.1 first reviews the mathematical foundations of WFS as a specific variant of sound field synthesis and then demonstrates the limitations that WFS faces in practical realisations from a physical point of view. It is these limitations that make research on perceptual aspects of WFS necessary, which are treated in section 2.2.

2.1 Theory of Wave Field Synthesis

2.1.1 Fundamental concept of sound field synthesis

Sound field synthesis is based on the Kirchhoff-Helmholtz integral equation (KHI). It states that the pressure inside a volume V, that is free of sources and objects that reflect or scatter sound, is completely determined by the sound pressure $P(\mathbf{x}_0,\omega)$ and the velocity in normal direction to the boundary, which is proportional to the directional gradient of the pressure $\frac{\partial}{\partial \mathbf{n}}P(\mathbf{x}_0,\omega)$ on its surface $\partial V = A(\mathbf{x}_0)$. The KHI reads [Wil99, eq. (8.15)]

$$\oint\limits_{\partial V} -\frac{\partial}{\partial \mathbf{n}}P(\mathbf{x}_0,\omega)\cdot G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) + P(\mathbf{x}_0,\omega)\cdot\frac{\partial}{\partial \mathbf{n}}G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega)\,\mathrm{d}A(\mathbf{x}_0) = \begin{cases} P(\mathbf{x},\omega) & \forall\,\mathbf{x}\in V \\ \frac{1}{2}P(\mathbf{x},\omega) & \forall\,\mathbf{x}\in\partial V \\ 0 & \forall\,\mathbf{x}\notin V \end{cases} \tag{2.1}$$

with $\mathbf{n}$ the normal vector on the surface directed inside the volume, $\omega = 2\pi f$ the angular frequency in $\frac{\mathrm{rad}}{\mathrm{s}}$ and f the frequency in Hz. $\mathbf{x} = (x,y,z)^{\mathrm{T}}$ in $(\mathrm{m},\mathrm{m},\mathrm{m})^{\mathrm{T}}$ is a field point and $\mathbf{x}_0 = (x_0,y_0,z_0)^{\mathrm{T}}$ in $(\mathrm{m},\mathrm{m},\mathrm{m})^{\mathrm{T}}$ a point on the surface $A(\mathbf{x}_0)$. The normal derivative of the sound pressure is defined as

$$\frac{\partial}{\partial \mathbf{n}}P(\mathbf{x}_0,\omega) = \nabla_{\mathbf{x}}P(\mathbf{x},\omega)\Big|_{\mathbf{x}=\mathbf{x}_0}\cdot\mathbf{n}. \tag{2.2}$$

Fig. 2.1 gives an illustration of the underlying geometry. The coordinate system used in this thesis is defined in appendix A. $G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega)$ denotes the three-dimensional free-field Green's function

$$G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) = \frac{\mathrm{e}^{-\mathrm{j}k|\mathbf{x}-\mathbf{x}_0|}}{4\pi|\mathbf{x}-\mathbf{x}_0|} \tag{2.3}$$

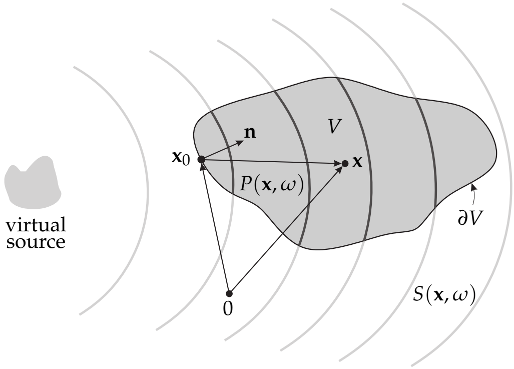


Figure 2.1: Geometry for the discussion of the fundamental concept of SFS for the desired sound field $S(\mathbf{x},\omega)$ of a virtual source. Explanation of other symbols in text. Figure taken from [Wie14, fig. 2.1], licence: CC BY 3.0 DE.

with the wave number $k = \frac{\omega}{c}$ in $\frac{\mathrm{rad}}{\mathrm{m}}$ and c the speed of sound in $\frac{\mathrm{m}}{\mathrm{s}}$. $G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega)$ is the solution to the inhomogeneous Helmholtz equation [Wil99, eq. (8.4)]

$$\nabla^2 G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) + k^2 G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) = -\delta(\mathbf{x}-\mathbf{x}_0) \tag{2.4}$$

where $\delta(\cdot)$ denotes the Dirac delta function. As eq. (2.1) states further, the field outside V is 0 and the pressure on the surface $A(\mathbf{x}_0)$ is $\frac{1}{2}P(\mathbf{x},\omega)$.

The three-dimensional free-field Green's function in eq. (2.3) can be interpreted as a monopole point source and its derivative

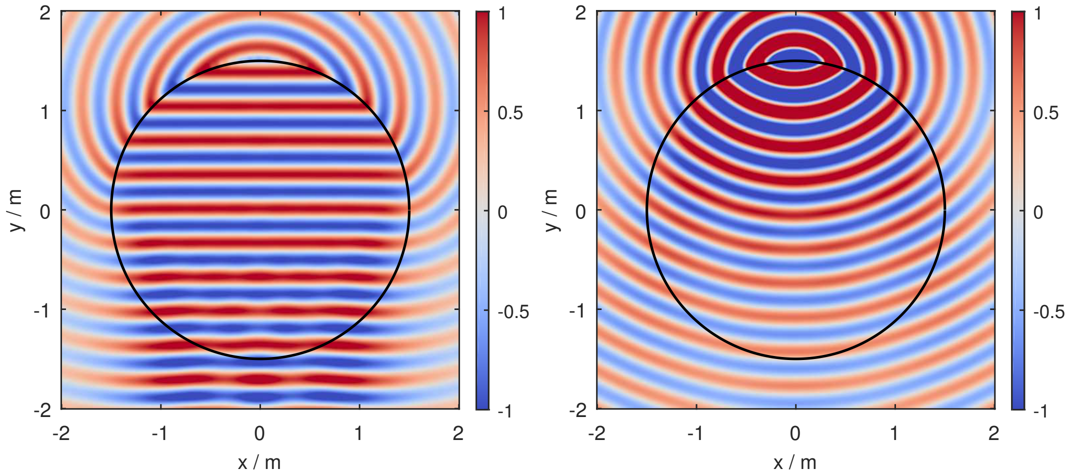
$$\frac{\partial}{\partial \mathbf{n}} G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) = \nabla_{\mathbf{x}_0} G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) \cdot \mathbf{n} \tag{2.5}$$

$$= G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega) \cdot \frac{1+\mathrm{j}\frac{\omega}{c}r}{r} \cdot \cos\varphi_r \tag{2.6}$$

as a dipole point source, with $r = |\mathbf{x}-\mathbf{x}_0|$ in m, φ_r the angle between $\mathbf{x}-\mathbf{x}_0$ and $\mathbf{n}$ in rad. If these sources on the surface $A(\mathbf{x}_0)$ are driven according to the velocity-related term $-\frac{\partial}{\partial \mathbf{n}}P(\mathbf{x}_0,\omega)$ and the pressure $P(\mathbf{x}_0,\omega)$ on $A(\mathbf{x}_0)$, the right-hand side of the KHI is achieved. In terms of SFS, this means that any desired sound field that is emitted by a so-called virtual source outside V can be synthesised inside the volume V as long as there is an infinitely dense distribution of so-called secondary sources – a transferred term from the domain of scattering problems [CK12] – on $A(\mathbf{x}_0)$ that is appropriately driven. The virtual source is a notional source that emits the desired sound field.

The problem at hand can be simplified if it is sufficient to control the sound field inside the volume, but not necessary to eliminate the field outside V. If either the monopoles or the dipoles are left out, it is still possible to synthesise any desired sound field inside V. The according equations are either the double layer potential (DLP) with dipoles only, or the single layer potential (SLP) with monopoles only [CK12]. Typically, SFS is only dealing with the SLP,

$$P(\mathbf{x},\omega) = \oint_{\partial V} D(\mathbf{x}_0,\omega) \cdot G_{0,\mathrm{3D}}(\mathbf{x},\mathbf{x}_0,\omega)\,\mathrm{d}A(\mathbf{x}_0), \tag{2.7}$$



(a) Plane wave in direction $\mathbf{n}_{\mathrm{pw}} = (0, -1, 0)^{\mathrm{T}}$ **(b)** Point source at $\mathbf{x}_{\mathrm{ps}} = (0, 2, 0)^{\mathrm{T}}$ m

Figure 2.2: Cross section of the sound pressure of monochromatic virtual sources synthesised with 3D WFS and a spherical secondary source distribution (black line). The driving functions are given in [SRA08, eq. (17)] and [SRA08, eq. (19)] for the virtual plane wave and the virtual point source, respectively. The reference point of the driving function is $\mathbf{x}_{\mathrm{ref}} = (0, 0, 0)^{\mathrm{T}}$ m, cf. eq. (2.18), $f = 1$ kHz. The sound fields are normalised to the origin of coordinates.

as in practice point monopoles can be approximated by readily available loudspeakers in a closed-box design whereas dipole loudspeakers are not widely in use. In eq. (2.7), $D(\mathbf{x}_0, \omega)$ is the so-called driving function for the secondary sources. As a consequence of this simplification, the outer field is not equal to 0 any more, as is illustrated in fig. 2.2 for the cases of a plane wave and a point source as desired sound fields synthesised by a spherical secondary source distribution of point monopoles with 3D WFS. The sound pressure in fig. 2.2 and in the following figures is obtained by using only the real part of the complex notation as is usual for convenience in mathematical calculations [GRS01]. These plots as well as the following sound field plots in this chapter have been created with the Sound Field Synthesis Toolbox [WS12], release 2.5.0[1]. The scripts to create the figures are publicly available[2].

To determine how the secondary sources have to be driven, $D(\mathbf{x}_0, \omega)$ in eq. (2.7) has to be specified. There exist different approaches to solve the SLP, also depending on the desired sound field and the geometry of the secondary source distribution, which can be three- or two-dimensional. The desired sound field can either be determined from measured data of a real sound field or constitute simple source models such as plane waves and point sources. The data-based approach is subject to a number of additional limitations like a finite spatial resolution and a limited number of microphones [AAG+13]. The focus in the present thesis is on model-based rendering of simple point sources.

Solutions of the SLP are classified in explicit solutions, that directly solve eq. (2.7) for the driving function $D(\mathbf{x}_0, \omega)$ delivering closed-form solutions, and implicit solutions. An explicit solution of eq. (2.7) for planar or linear secondary source dis-

---

[1] `http://doi.org/10.5281/zenodo.2597212`

[2] `http://doi.org/10.5281/zenodo.3745986`

tributions is given by the Spectral Division Method (SDM) [AS10]. The SLP formulated for, e.g., a linear secondary source distribution positioned along the x-axis of a Cartesian coordinate system at $\mathbf{x}_0 = (x_0, 0, 0)^\mathrm{T}$, as it is used in the listening experiments in this thesis, reads

$$P(\mathbf{x}, \omega) = \int_{-\infty}^{\infty} D(\mathbf{x}_0, \omega) G_{0,\mathrm{3D}}(\mathbf{x}, \mathbf{x}_0, \omega) \, \mathrm{d}x_0. \tag{2.8}$$

Eq. (2.8) constitutes a convolution along the x-axis as $G_{0,\mathrm{3D}}$ is depending on $\mathbf{x} - \mathbf{x}_0$ (cf. eq. (2.3)). With a spatial Fourier transform with respect to x as defined in appendix B, this can be expressed as a multiplication of the respective spectra

$$\tilde{P}(k_x, y, z, \omega) = \tilde{D}(k_x, \omega) \cdot \tilde{G}_{0,\mathrm{3D}}(k_x, y, z, \omega) \tag{2.9}$$

where k_x in $\frac{\mathrm{rad}}{\mathrm{m}}$ is the x-component of the wave number vector $\mathbf{k} = (k_x, k_y, k_z)^\mathrm{T}$ in $(\frac{\mathrm{rad}}{\mathrm{m}}, \frac{\mathrm{rad}}{\mathrm{m}}, \frac{\mathrm{rad}}{\mathrm{m}})^\mathrm{T}$. The spectra in eq. (2.9) marked with a tilde are called wave number spectra and are obtained after a temporal and a spatial Fourier transform. Eq. (2.9) can then easily be solved for $\tilde{D}(k_x, \omega)$. Inverse spatial and temporal Fourier transform finally deliver the driving function $D(\mathbf{x}_0, \omega)$ and the driving signal $d(\mathbf{x}_0, t)$ for the secondary sources, respectively, if the wave number spectrum of the desired sound field is known. Similarly, for spherical or circular secondary source distributions, there exists Near-Field-Compensated Higher Order Ambisonics (NFC-HOA) [Dan03] as the explicit solution. In contrast, traditional WFS originally started with the implicit solution of the DLP for a plane of secondary sources [Ber88]. Later formulations treated the case of monopoles as secondary sources [Vog93, Sta97, Ver97].

As this short overview indicates, there emerged a variety of approaches over the years leading to a number of sometimes only slightly different driving functions. Only recently, the connections between these approaches and their presumed inconsistencies have been resolved [Ahr12, Sch16] including showing that WFS is the high-frequency/far-field approximation of the SDM solution. As this thesis is only examining WFS with a linear array of secondary sources synthesising a virtual point source, solely the derivation of the used WFS driving function for this case is given in section 2.1.3.

Besides the summarised analytic approaches to SFS, there also exist methods that try to control the synthesised sound field at discrete points inside a specified listening area [KN93, Pol05]. These approaches are typically based on solving an inverse problem with least square error techniques. A special case of SFS is represented by Local Sound Field Synthesis [SA10b, HWS16] which is reducing the influence of spatial aliasing (cf. section 2.1.4.1) by targeting only the synthesis in a small listening area. Perceptual properties of Local Sound Field Synthesis have been investigated by [Win19]. Recent work in SFS has also presented a more general referencing scheme for WFS [FFSS17], that is necessary as driving functions depend on field points in the reproduction plane.

2.1.2 Simplification to 2.5D synthesis

3D WFS with sources surrounding a volume requires an enormous effort regarding material and space in practice. Therefore, practical setups are usually restricted to synthesis in a plane with secondary source geometries that are typically circular, linear or piece-wise linear. The correct secondary source type for this 2D synthesis is a line monopole, but this would also involve considerable effort in practice. Moreover, line sources themselves are in practice typically approximated by an array of point sources. Therefore, WFS is usually realised with point monopoles only. As a perfect monopole point source is not available in reality, a good approximation can be found by using closed-box loudspeakers that at least in the lower frequency range exhibit an omnidirectional radiation characteristic. An alternative are, e.g., specially designed loudspeaker panels that try to optimise the radiation characteristics at low and high frequencies [GMMW07].

Due to this so-called secondary source type mismatch – using secondary sources with a 3D characteristic in a 2D secondary source geometry [AS09] – the resulting SFS variant is termed 2.5D synthesis [Sta97]. As point sources do not radiate energy in the same way as line sources, the consequence of this mismatch are amplitude deviations in the synthesised sound fields. Regarding the control of the listening area, a degree of freedom is missing, resulting in at maximum lines or curves where the synthesis of the desired sound field is correct in amplitude. These positions of correct synthesis are termed reference positions [FFSS17]. Fig. 2.3 illustrates the difference between the desired sound field of a point source and the sound field synthesised with a linear secondary source distribution of point monopoles. The level of the virtual point source should be decreasing according to the inverse square law, but due to the secondary source type mismatch, the amplitude decay is faster. Only at the chosen reference point at $\mathbf{x}_{\mathrm{ref}} = (0, -1, 0)^{\mathrm{T}}$ m a correct amplitude of the synthesised sound field is achieved.

2.1.3 Derivation of the Wave Field Synthesis driving function for a virtual point source with a linear array

The experiments in this thesis are using 2.5D WFS with a linear array of point monopoles. Therefore, the driving function used for this geometry is derived in the following.

The simplification of the KHI in eq. (2.1) to using only point monopoles, leaving out the addend including the normal derivative of the Green's function in the integral, implies that the Neumann boundary condition must be fulfilled for the Green's function at $\mathbf{x}_0$

$$\frac{\partial}{\partial \mathbf{n}} G_{\mathrm{N}}(\mathbf{x}, \mathbf{x_0}, \omega) = 0. \tag{2.10}$$

For the simple geometry of $A(\mathbf{x}_0)$ as a plane, e.g. the x, z-plane as illustrated in fig. 2.4, the Neumann Green's function is known as [Wil99, eq. (8.83)]

$$G_{\mathrm{N}}(\mathbf{x}, \mathbf{x}_0, \omega) = 2 \cdot G_{0,\mathrm{3D}}(\mathbf{x}, \mathbf{x}_0, \omega). \tag{2.11}$$

Eq. (2.11) only holds true for a Green's function located on the boundary, i.e. the plane $A(\mathbf{x}_0)$. Consequently the SLP for the case of a planar secondary source distribution in the x, z-plane reads

$$P(\mathbf{x}, \omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} -\frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0, \omega) \cdot 2 \cdot G_{0,\mathrm{3D}}(\mathbf{x}, \mathbf{x}_0, \omega) \, \mathrm{d}x_0 \, \mathrm{d}z_0. \tag{2.12}$$

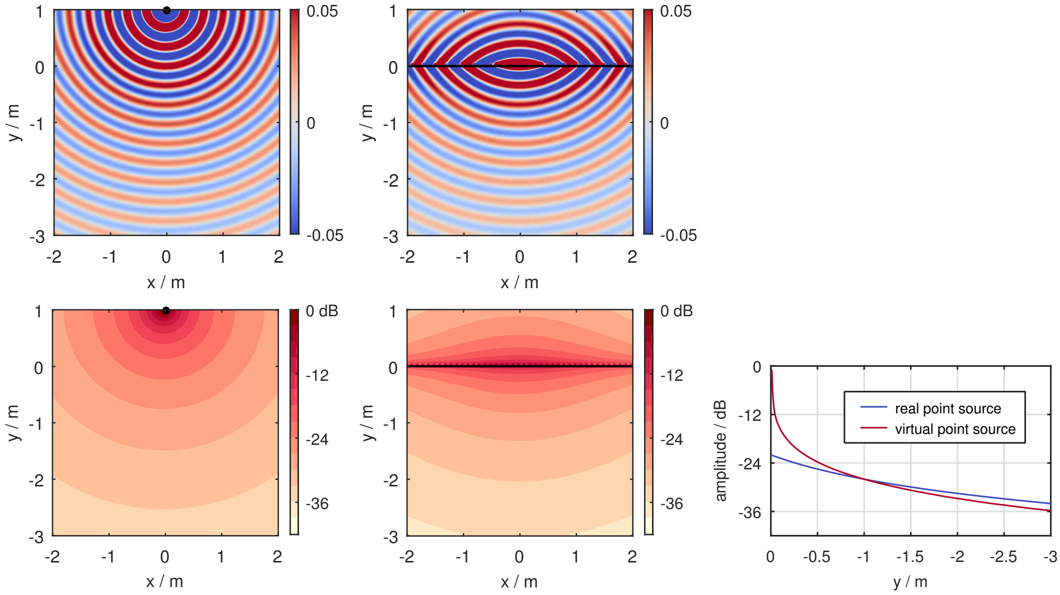


Figure 2.3: Sound pressure (upper row) and sound pressure level (lower row) of a real (left column) and a virtual (middle column) monochromatic point source. The point source is located at $\mathbf{x}_{\mathrm{ps}} = (0, 1, 0)^{\mathrm{T}}$ m and has a frequency of $f = 1$ kHz. The virtual source is synthesised with 2.5D WFS and a linear secondary source array (black line) with the driving function in eq. (2.18) and $\mathbf{x}_{\mathrm{ref}} = (0, -1, 0)^{\mathrm{T}}$ m. The right plot compares the levels of the real and virtual point source over distance from the secondary source array along the line $\mathbf{x} = (0, y, 0)^{\mathrm{T}}$.

with $S(\mathbf{x}, \omega)$ the sound field of the desired virtual source, $\frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0, \omega)$ its normal derivative on the surface $A(\mathbf{x}_0)$ and $\mathbf{x_0} = (x_0, 0, z_0)^{\mathrm{T}}$. Eq. (2.12) is also known as Rayleigh's first integral equation [Wil99, eq. (8.84)]. The driving function can now be directly calculated with

$$D(\mathbf{x}_0, \omega) = -2 \frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0, \omega). \tag{2.13}$$

For the desired field of a point source at position $\mathbf{x}_{\mathrm{ps}}$

$$S_{\mathrm{ps}}(\mathbf{x}, \omega) = \frac{\mathrm{e}^{-\mathrm{j}\frac{\omega}{c}|\mathbf{x}-\mathbf{x}_{\mathrm{ps}}|}}{4\pi|\mathbf{x}-\mathbf{x}_{\mathrm{ps}}|}, \tag{2.14}$$

the driving function thus reads

$$D(\mathbf{x}_0, \omega) = 2 \cdot \frac{1 + \mathrm{j}\frac{\omega}{c}|\mathbf{x_0} - \mathbf{x}_{\mathrm{ps}}|}{|\mathbf{x_0} - \mathbf{x}_{\mathrm{ps}}|} \cdot \cos\varphi_{\mathrm{s}} \cdot S_{\mathrm{ps}}(\mathbf{x}_0, \omega) \tag{2.15}$$

with φ_{s} in rad the angle between $\mathbf{n}$ and $\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}$. i.e.

$$\cos\varphi_{\mathrm{s}} = \frac{\langle \mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}, \mathbf{n} \rangle}{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|}. \tag{2.16}$$

If the aim is not the synthesis in 3D space, but only in the x, y-plane with the virtual source in the positive x, y-half-plane and the receiver in the negative x, y-half-plane, the driving function eq. (2.15) can be simplified with the stationary phase

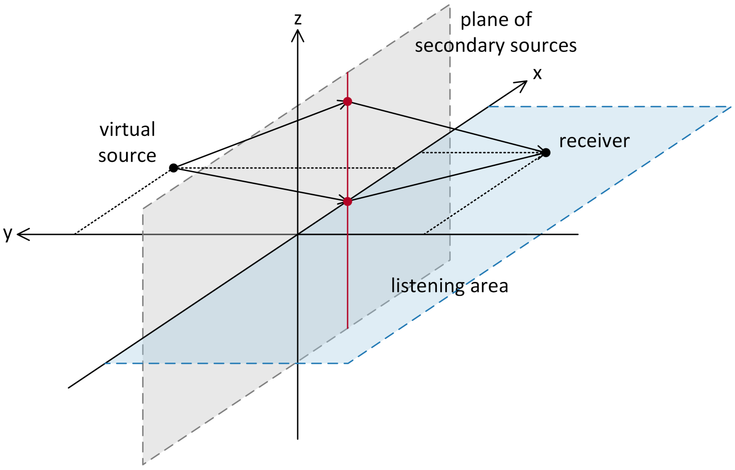


Figure 2.4: Considered geometry for the derivation of the WFS driving functions for a planar secondary source distribution in the x,z-plane and for a linear secondary source distribution along the x-axis. The listening area is shaded in blue. The contribution of all secondary sources along the red line can be approximated by the contribution of the secondary source on the intersection of this line with the x-axis by the stationary phase approximation. The figure is adapted from [Ver97, fig. 2.6].

approximation [Ver97, section 2.3.1]: All secondary sources on a straight line with a fixed x_0 (marked as the red line in fig. 2.4) can be approximated by the contribution of the secondary source in $(x_0, 0, 0)^{\mathrm{T}}$ alone (marked as the red dot on the x-axis in fig. 2.4). Thus, only secondary sources on a straight line along the x-axis are necessary for the synthesis instead of secondary sources on the whole x,z-plane. For additional simplification, the far-field approximation

$$\frac{\omega}{c}|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}| \gg 1 \tag{2.17}$$

is introduced, requiring high frequencies and/or the virtual source to be distant to the secondary source distribution. With these approximations, the driving function for a virtual point source synthesised by a linear array in $\mathbf{x}_0 = (x_0, 0, 0)^{\mathrm{T}}$ is [Ver97, eq. (2.22a)]

$$D(x_0, \omega) = \sqrt{\mathrm{j}\frac{\omega}{c}} \cdot \sqrt{\frac{1}{2\pi}} \cdot \sqrt{\frac{|\mathbf{x}_{\mathrm{ref}} - \mathbf{x}_0|}{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}| + |\mathbf{x}_{\mathrm{ref}} - \mathbf{x}_0|}} \cdot \frac{\langle \mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}, \mathbf{n} \rangle}{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|} \cdot \frac{\mathrm{e}^{-\mathrm{j}\frac{\omega}{c}|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|}}{\sqrt{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|}}. \tag{2.18}$$

As this driving function is dependent on the field point in the reproduction plane, $\mathbf{x}$ has been replaced by $\mathbf{x}_{\mathrm{ref}}$, denoting a reference point where correct synthesis is achieved. Inverse Fourier transform with respect to time yields the driving signal

$$\begin{aligned} d(\mathbf{x}_0, t) = \sqrt{\frac{1}{2\pi}} \cdot \sqrt{\frac{|\mathbf{x}_{\mathrm{ref}} - \mathbf{x}_0|}{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}| + |\mathbf{x}_{\mathrm{ref}} - \mathbf{x}_0|}} \cdot \frac{\langle \mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}, \mathbf{n} \rangle}{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|} \cdot \frac{1}{\sqrt{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|}} \\ \cdot \mathcal{F}^{-1}\left\{\sqrt{\mathrm{j}\frac{\omega}{c}}\right\} * \delta\left(t - \frac{|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ps}}|}{c}\right) \end{aligned} \tag{2.19}$$

where $\mathcal{F}^{-1}\left\{\sqrt{\mathrm{j}\frac{\omega}{c}}\right\}$ is the inverse Fourier transform of the so-called pre-equalisation filter $\sqrt{\mathrm{j}\frac{\omega}{c}}$. The driving function and driving signal in eq. (2.18) and (2.19), respectively, are used for all experiments in this thesis. An additional second stationary phase approximation on eq. (2.18) delivers a modified version of the driving function that yields correct synthesis on a reference line parallel to x-axis instead of only a single reference point [Ver97, eq. (2.27)]. A more general referencing scheme for WFS is presented by [FFSS17].

2.1.4 Practical limitations

The theory of WFS is based on a number of ideal assumptions that cannot be met in practice. Consequently, synthesised sound fields differ in reality from the desired prototype field. The following sections treat these practical limitations and their influences on the synthesised sound field.

2.1.4.1 Discretisation of the secondary source distribution

The most prominent violation of a theoretic assumption in SFS is the discretisation of the continuous secondary source distribution by single loudspeakers surrounding the listening area leading to spatial aliasing [SR06]. In typical setups, the loudspeaker distance cannot be made less than 15–20 cm due to the size of the transducers. This results in spatial aliasing impacting frequencies from above approx. 1 kHz, clearly in the range of human hearing.

Fig. 2.5 compares spatial aliasing in WFS to the more well-known aliasing in the time domain. Sampling in the time domain shown on the left-hand side of fig. 2.5 with the sampling rate f_s leads to spectral repetitions of the signal in the frequency domain. The repetitions appear with a distance of f_s. If the Nyquist frequency $f_\mathrm{Nyquist} = \frac{f_\mathrm{s}}{2}$ is lower than the highest frequency contained in the original signal, the spectral repetitions overlap and temporal aliasing occurs. To recover the original continuous signal, a reconstruction filter is necessary, that filters out the spectral repetitions, which is not possible if aliasing is present due to overlapping spectral repetitions. The right-hand side of fig. 2.5 demonstrates these principles for the case of spatial sampling of a linear array in $\mathbf{x}_0 = (x_0, 0, 0)^\mathrm{T}$ synthesising a virtual point source. The driving function $D(x_0, \omega)$ is sampled spatially by employing discrete loudspeakers every Δx_0, assuming equidistant sampling. The corresponding spectrum of the driving function is the two-dimensional wave number spectrum $\tilde{D}(k_x, \omega)$ where spatial and temporal frequencies are connected by the dispersion relation

$$\left(\frac{\omega}{c}\right)^2 = k_x^2 + k_y^2 + k_z^2. \tag{2.20}$$

Due to spatial sampling, the wave number spectrum $\tilde{D}(k_x, \omega)$ is repeated with a period of $\frac{2\pi}{\Delta x_0}$. The wave number spectrum of the Green's function $\tilde{G}_{0,\mathrm{3D}}(k_x, y, z = 0, \omega)$ as depicted in fig. 2.6 for $y = 1$ serves as reconstruction filter. While this constitutes a filter that is not bandlimited, only the portion of the spectrum inside the triangle delimited by the black lines following the relation $|k_x| = \frac{\omega}{c}$, termed the propagating part, contributes to a propagating sound field [SA09]. The remaining portion of $\tilde{G}_{0,\mathrm{3D}}(k_x, y, z = 0, \omega)$ is termed the evanescent part and is indicated by

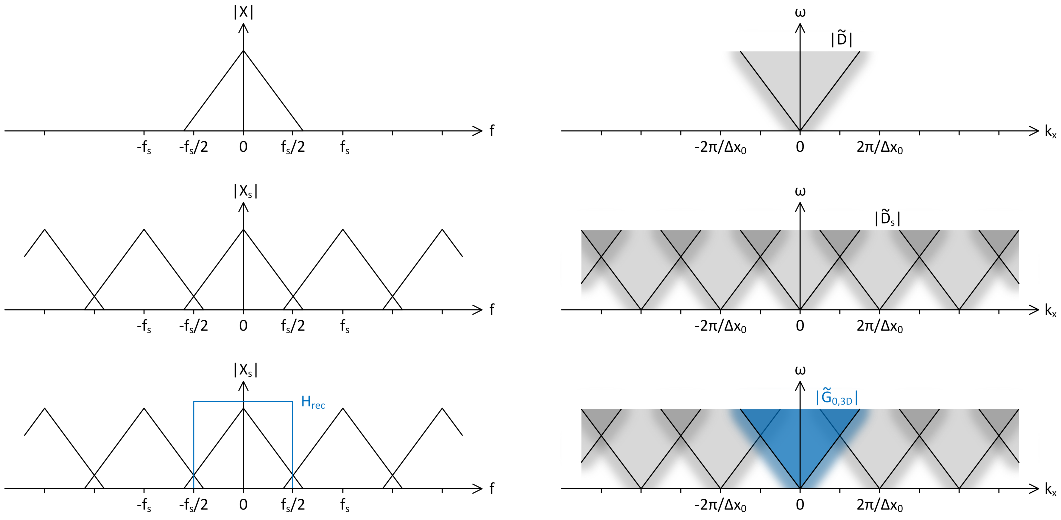


Figure 2.5: Principle of sampling, aliasing and reconstruction. Left: Magnitude spectra $|X|$ of a time signal and its time-sampled version $|X_\mathrm{s}|$ with sampling rate f_s. H_rec is the ideal reconstruction filter. Right: Wave number spectra $|\tilde{D}|$ of the 2.5D WFS driving function for a virtual point source and its spatially sampled version $|\tilde{D}_\mathrm{s}|$ sampled every Δx_0 along a linear array of secondary sources. $|\tilde{G}_{0,\mathrm{3D}}|$ is the reconstruction filter in the wave number domain when using spherical monopoles as secondary sources. Shaded areas denote a higher magnitude of the non-bandlimited functions.

the shaded fade-out outside the triangle formed by $|k_x| = \frac{\omega}{c}$ in fig. 2.5. It is only generating a sound field that is rapidly decaying with increasing distance from the secondary sources, but still oscillating in one direction. In the same way, $\tilde{D}(k_x, \omega)$ can be divided in a propagating and an evanescent part as well. As can be seen in fig. 2.5, the propagating part of the wave number spectrum of the Green's function is multiplied with spectral repetitions of the driving function containing substantial amounts of energy, leading to spatial aliasing components being propagated throughout the listening area. Propagating spatial aliasing is only present above the so-called aliasing frequency [AS10]

$$f_\mathrm{alias} = \frac{c}{2\Delta x_0}. \tag{2.21}$$

This lower corner frequency can be derived from fig. 2.5 from the intersection point of the lines delimiting the propagating part of $\tilde{G}_{0,\mathrm{3D}}(k_x, y, z = 0, \omega)$, i.e. $\omega = c \cdot k_x$ on the positive half of the spectrum, and the propagating part of the first repetition of the wave number spectrum of the driving function, i.e. $\omega = -c(k_x - \frac{2\pi}{\Delta x_0})$ on the left side of the spectrum. Repetitions of $\tilde{D}(k_x, \omega)$, where at least the propagating parts are not overlapping in the range of human hearing, which can be considered to be as high as 20 kHz, i.e. $f_\mathrm{alias} > 20\,\mathrm{kHz}$, would require a transducer spacing of less than 1 cm. This is not attainable in practice. Therefore, spatial aliasing in WFS typically occurs in the range of human hearing.

The impact of spatial aliasing is illustrated in fig. 2.7 that shows the sound field of a virtual point source reproduced by WFS with a linear array at three different frequencies. Below the spatial aliasing frequency in eq. (2.21), which is $f_\mathrm{alias} \approx$

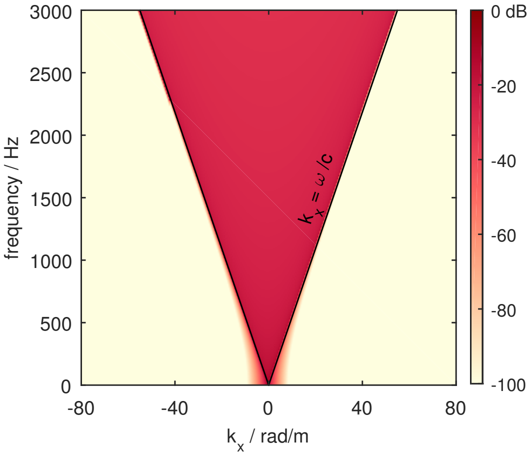


Figure 2.6: Magnitude of the wave number spectrum of the Green's function in eq. (2.3) with $\mathbf{x}_0 = (0,0,0)^\mathrm{T}$ m evaluated at $y = 1$ m and $z = 0$. The spectrum is calculated according to [AS10, eq. (52)].

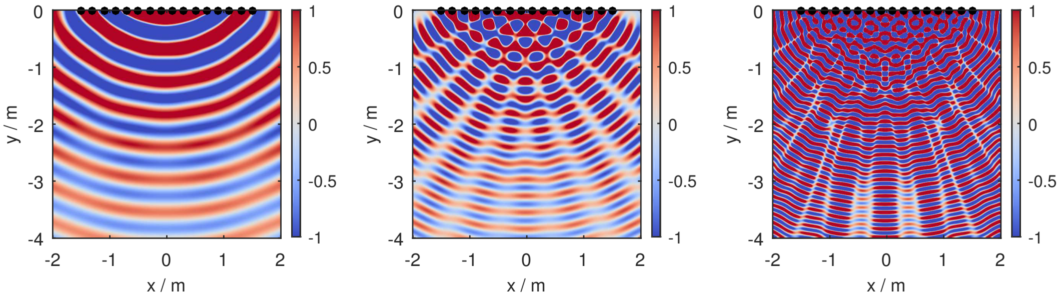


Figure 2.7: Sound pressure of a virtual monochromatic point source at $\mathbf{x}_\mathrm{ps} = (0,1,0)^\mathrm{T}$ synthesised with 2.5D WFS with driving function eq. (2.18) and a linear secondary source array with $\Delta x_0 = 0.2\,\mathrm{m}$ and $\mathbf{x}_\mathrm{ref} = (0,-1,0)^\mathrm{T}$ m. Left: $f = 600\,\mathrm{Hz}$, middle: $f = 1.2\,\mathrm{kHz}$, right: $f = 2.4\,\mathrm{kHz}$. The sound fields are normalised to the reference position $\mathbf{x}_\mathrm{ref}$.

858 Hz for the given setup of a linear array with $\Delta x_0 = 0.2\,\mathrm{m}$, the reproduced sound field in the left column of fig. 2.7 with $f = 600\,\mathrm{Hz}$ displays the shape of point source, only with the amplitude deviations arising from the secondary source type mismatch, cf. section 2.1.2. For frequencies above f_alias, e.g $f = 1.2\,\mathrm{kHz}$ in the middle column of fig. 2.7, spatial aliasing artefacts begin to spread over the listening area. The effect becomes more severe with higher frequencies, cf. right column of fig. 2.7 with $f = 2.4\,\mathrm{kHz}$. Magnitude responses of the reproduced sound field at a fixed position in the listening area, coinciding with the reference point $\mathbf{x}_\mathrm{ref}$ of the driving function in eq. (2.18), are given in fig. 2.8. Above f_alias, spatial aliasing artefacts are visible as strong fluctuations in the magnitude responses. Eq. (2.21) is only setting a lower limit for propagating spatial aliasing artefacts. As can be observed in fig. 2.7, the occurrence of these artefacts depends on the position in the listening area. It also depends on the type and position of the virtual source. To predict where in the listening area spatial aliasing artefacts appear, [WFSS19] has developed a geometric model that is not only taking 2.5D WFS into account, but other 2.5D SFS methods as well.

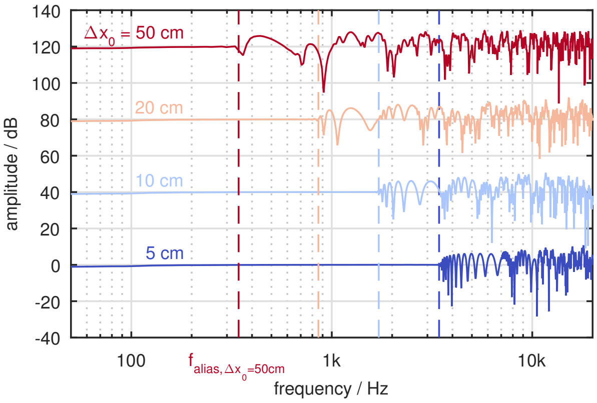


Figure 2.8: Magnitude of the frequency responses in dB of 2.5D WFS with driving function eq. (2.18) at $\mathbf{x}_{\text{ref}}$ with a linear secondary source array and four different Δx_0 as given in the figure. The spectra are normalised and shifted for better discriminability. The aliasing frequency for each Δx_0 according to eq. (2.21) is marked with a dashed line.

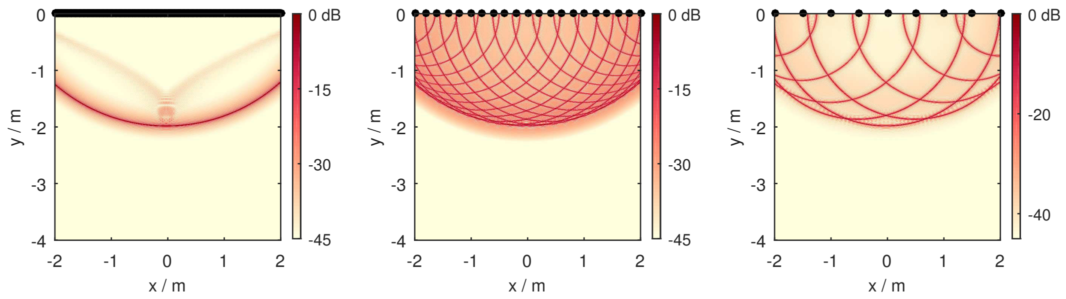


Figure 2.9: Level of the sound pressure of a point source at $\mathbf{x}_{\text{ps}} = (0, 1, 0)^{\text{T}}$ m emitting a broadband impulse synthesised by 2.5D WFS with driving function eq. (2.18) and $\mathbf{x}_{\text{ref}} = (0, -1, 0)^{\text{T}}$ m at the time $t = 8.7$ ms. Left: secondary source distance $\Delta x_0 = 0.5$ cm, middle: $\Delta x_0 = 20$ cm, right: $\Delta x_0 = 50$ cm. The sound fields are normalised to the maximum value in the plots. The slightly raised level before the first wave front is caused by the pre-equalisation filter, that in practical implementations is chosen as a shelving filter, cf. end of section 2.1.4.1.

In the time domain and when considering a broadband impulse as signal of the virtual source, spatial aliasing manifests as additional wave fronts trailing the first wave front that is correctly synthesised by WFS. Fig. 2.9 demonstrates this effect for a linear array synthesising a point source emitting a broadband impulse with three different spacings Δx_0 between the secondary sources. For a close to continuous array of secondary sources with $\Delta x_0 = 0.5$ cm in the left column of fig. 2.9, there is only the desired impulse wave front visible, but with increasing Δx_0 additional wave fronts after the first one occur, cf. middle and right column of fig. 2.9. Examining the impulse responses of reproduced sound fields at one specific point in the listening area in fig. 2.10, the additional wave fronts appear as trailing pulses, one for each secondary source involved in the synthesis. In a perfectly symmetric setup, the contributions from pairs of secondary sources coincide and create impulses that even surpass the first impulse in amplitude.

A special form of spatial aliasing can be observed when a virtual source is located

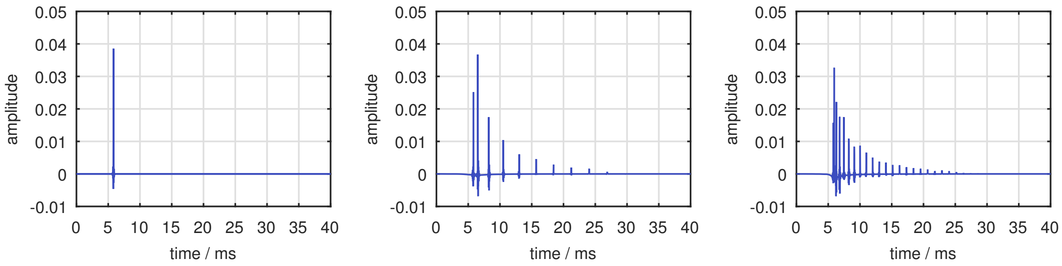


Figure 2.10: Impulse responses of a real (left) and a virtual point source with $\Delta x_0 = 50\,\mathrm{cm}$ (middle) and $\Delta x_0 = 20\,\mathrm{cm}$ (right) synthesised by 2.5D WFS with driving function eq. (2.18) and a linear array of 10 m length along the x-axis. The virtual point sources are at $\mathbf{x}_\mathrm{ps} = (0, 1, 0)^\mathrm{T}\,\mathrm{m}$, the impulse responses are given for $\mathbf{x}_\mathrm{ref} = (0, -1, 0)^\mathrm{T}\,\mathrm{m}$. The impulse response of the real source is given for the same relative distance of 2 m.

very close to the discrete secondary sources as is illustrated in fig. 2.11. In this and the following fig. 2.12 and 2.13, the sound fields are shown in dB_SPL to have the same reference value for comparison. If the virtual source is positioned right behind a secondary source in a linear array (middle column of fig. 2.11), the synthesised amplitude is too high throughout the entire sound field and in particular also not correct at the reference point compared to the desired field of a real point in the left column of fig. 2.11. In contrast, for a virtual source positioned close, but in-between two secondary sources, the resulting sound field exhibits an amplitude that is too low (right column of fig. 2.11). The reason for this can be found by conducting an analysis as in [SA09], where the synthesised sound field is segmented into four components resulting from the overlap of propagating and evanescent parts of both the wave number spectrum of the Green's function $\tilde{G}_{0,3\mathrm{D}}(k_x, y, z = 0, \omega)$ and the spectral repetitions of the wave number spectrum of the driving function $\tilde{D}_\mathrm{s}(k_x, \omega)$, cf. fig. 2.5. These four components are denoted as given in table 2.1 and depicted in fig. 2.12 and 2.13 for the two cases of the virtual source right behind and in-between secondary sources. As for virtual sources close to the secondary source array the evanescent part of the driving function contains much more energy than for more distant virtual sources, these evanescent parts get propagated by the Green's function as so-called evanescent aliasing and thus make a substantial contribution to the resulting sound field. For the case of a virtual source right behind a secondary source in fig. 2.12, this aliasing component $P_{\mathrm{S,pr2}}(\mathbf{x}, \omega)$ is in phase with the desired sound field and thus amplifies it. For the case of a virtual source close to the secondary sources and between two of them as in fig. 2.13, $P_{\mathrm{S,pr2}}(\mathbf{x}, \omega)$ is out of phase with the desired sound field, leading to an overall amplitude which is too low. If such a close point source is moved along the secondary source contour, this would result in an amplitude modulation. This effect has been analysed by the author of this thesis together with the calculation of the individual aliasing components [EWS15]. This paper also includes the derivation of an approximative criterion for the distance of a virtual source to the secondary sources beyond which the amplitude deviation from evanescent aliasing are not relevant. The examples shown in [EWS15] and in this thesis in fig. 2.11–2.13 are based on the driving function derived by the spectral division method as given in [SA10a, eq. (24)]. However, the demonstrated effect of a too high or too low amplitude for point sources close to the secondary sources can

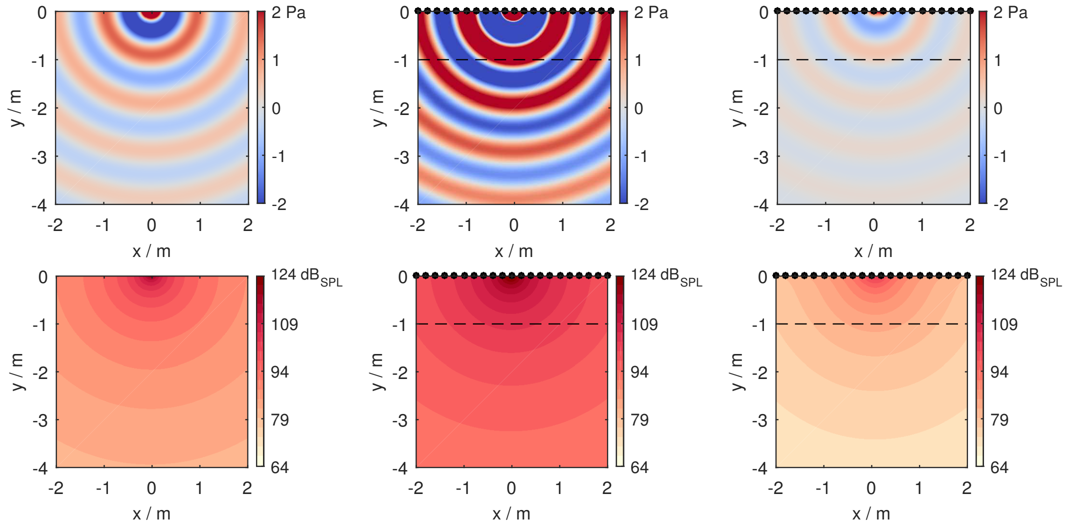


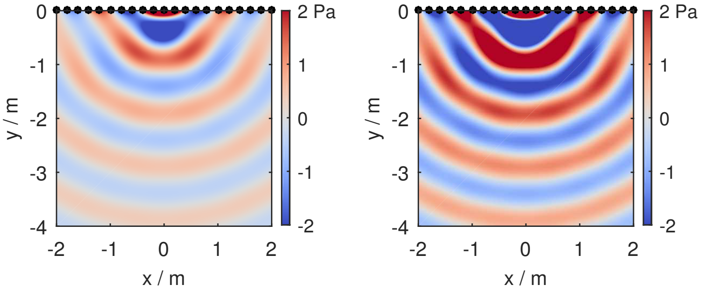
Figure 2.11: Sound pressure in Pa (upper row) and its level in $\mathrm{dB_{SPL}}$ (lower row) of a real monochromatic point source at $\mathbf{x}_\mathrm{ps} = (0, 0.02, 0)^\mathrm{T}$ m (left column) and virtual monochromatic point sources synthesised by 2.5D WFS with $\Delta x_0 = 20$ cm, that are positioned close to the linear secondary source array (black dots) and either right behind a secondary source at $\mathbf{x}_\mathrm{ps} = (0, 0.02, 0)^\mathrm{T}$ m (middle column) or between two secondary sources at $\mathbf{x}_\mathrm{ps} = (0.1, 0.02, 0)^\mathrm{T}$ m (right column). The driving function for WFS reproduction is derived by SDM [SA10a, eq. (24)]. The reference line for correct synthesis is marked by the dashed line. The frequency of the point sources is $f = 350$ Hz. The amplitude of the real prototype point source is chosen to yield 94 $\mathrm{dB_{SPL}}$ at $\mathbf{x} = (0, -1, 0)^\mathrm{T}$ m on the reference line. At the same point, the virtual sources exhibit a sound pressure level of 104 $\mathrm{dB_{SPL}}$ in the middle column and 83 $\mathrm{dB_{SPL}}$ in the right column.

also be shown for the WFS driving function in eq. (2.18) with a reference point or the WFS driving function with a reference line as given by [Ver97, eq. (2.27)]. For these WFS driving functions, the effect is only slightly modified by the additional far field/high frequency approximations used.

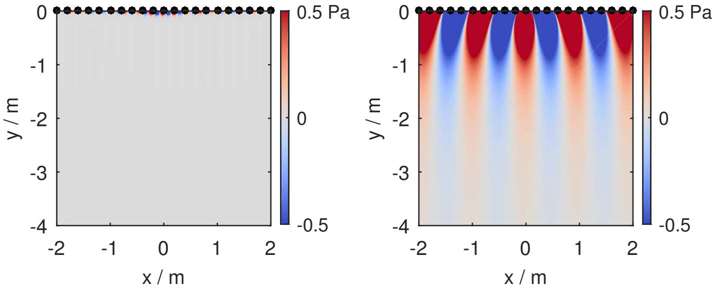
Spatial aliasing also has a consequence for the implementation of the driving function in WFS regarding the pre-equalisation filter. In 2.5D synthesis, the linear array is radiating like a line source, which is corrected with a filter with a 3 dB/octave slope. As was shown by [SA10a], the spatial aliasing artefacts also generate an approx. 3 dB/octave slope above the spatial aliasing frequency in the magnitude response of the reproduced sound field, making it advisable to set the pre-equalisation filter constant above the spatial aliasing frequency. Additionally, the pre-equalisation filter is also set constant below a lower corner frequency where the array does not

Table 2.1: Four spatial aliasing components of a synthesised sound field as analysed by [SA09, EWS15].

		Driving function part	
		propagating	evanescent
Green's function part	propagating	$P_{\mathrm{S,pr1}}(\mathbf{x}, \omega)$	$P_{\mathrm{S,pr2}}(\mathbf{x}, \omega)$
	evanescent	$P_{\mathrm{S,ev1}}(\mathbf{x}, \omega)$	$P_{\mathrm{S,ev2}}(\mathbf{x}, \omega)$

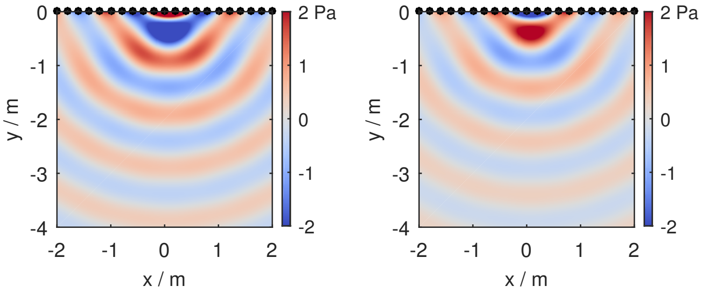


(a) $P_{\mathrm{S,pr1}}(\mathbf{x},\omega)$ **(b)** $P_{\mathrm{S,pr2}}(\mathbf{x},\omega)$

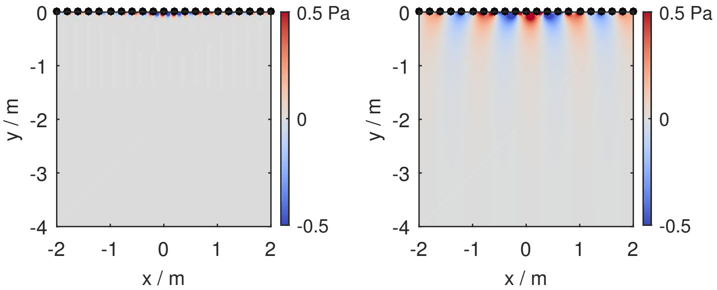


(c) $P_{\mathrm{S,ev1}}(\mathbf{x},\omega)$ **(d)** $P_{\mathrm{S,ev2}}(\mathbf{x},\omega)$

Figure 2.12: Sound field components according to table 2.1 for a virtual monochromatic point source with $f = 350\,\mathrm{Hz}$ at $\mathbf{x}_{\mathrm{ps}} = (0, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ synthesised by 2.5D WFS and a linear array with $\Delta x_0 = 20\,\mathrm{cm}$. The driving function for WFS reproduction is derived by SDM [SA10a, eq. (24)]. Sound field components are calculated by [EWS15, eq. (5)–(8)]. All four components together form the sound field in fig. 2.11, middle column.



(a) $P_{\mathrm{S,pr1}}(\mathbf{x}, \omega)$ **(b)** $P_{\mathrm{S,pr2}}(\mathbf{x}, \omega)$



(c) $P_{\mathrm{S,ev1}}(\mathbf{x}, \omega)$ **(d)** $P_{\mathrm{S,ev2}}(\mathbf{x}, \omega)$

Figure 2.13: Sound field components according to table 2.1 for a virtual monochromatic point source with $f = 350\,\mathrm{Hz}$ at $\mathbf{x}_{\mathrm{ps}} = (0.1, 0.02, 0)^{\mathrm{T}}\,\mathrm{m}$ synthesised by 2.5D WFS and a linear array with $\Delta x_0 = 20\,\mathrm{cm}$. The driving function for WFS reproduction is derived by SDM [SA10a, eq. (24)]. Sound field components are calculated by [EWS15, eq. (5)–(8)]. All four components together form the sound field in fig. 2.11, right column.

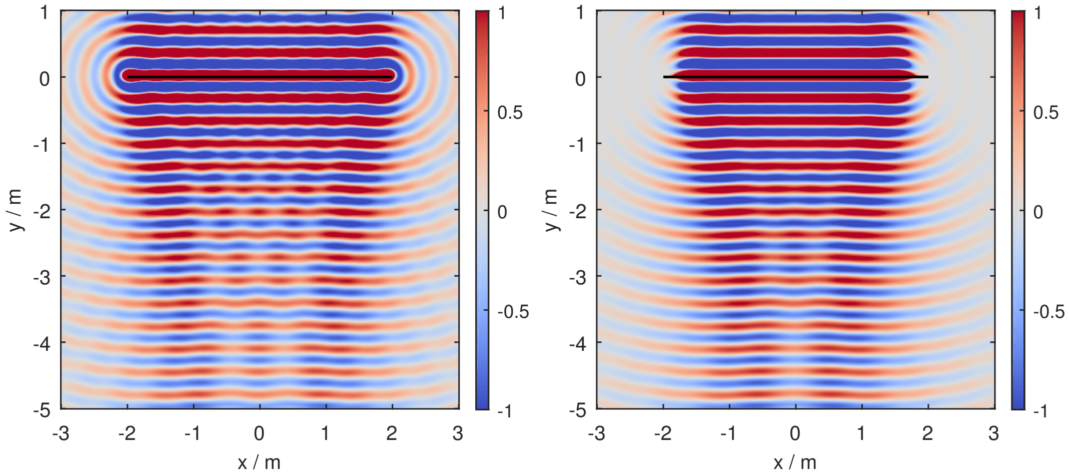


Figure 2.14: Sound pressure of a monochromatic plane wave in direction $\mathbf{n}_{\text{pw}} = (0, -1, 0)^{\text{T}}$ synthesised by 2.5D WFS with driving function [SRA08, eq. (27)] and a truncated linear secondary source distribution (black line). $\mathbf{x}_{\text{ref}} = (0, -1, 0)^{\text{T}}\,\text{m}$, $f = 1\,\text{kHz}$. Left: without tapering window, right: with raised-cosine shaped tapering window applied over 30% of the array length. The sound fields are normalised to the centre of the plots.

exhibit a line source characteristic any more due to its necessarily finite length (more on this truncation of the secondary source distribution is given in the following section 2.1.4.2). The lower corner frequency is determined by the transition when the array length can be considered short compared to the wave length of low frequencies. With these two corner frequencies, the pre-equalisation filter has a shelving filter shape [SESW13].

2.1.4.2 Truncation of the secondary source distribution

The infinitely long linear array required for the driving function in eq. (2.18) has to be truncated in practice, which has consequences for the synthesised sound field. The resulting sound field is a superposition of the desired sound field and two spherical waves from the edges of the truncated array [Ver97]. This is in accordance with diffraction theory. A countermeasure that is easy to implement has been introduced by [Vog93] by applying a so-called tapering window to the secondary source distribution. This constitutes a fade-out towards the edges by multiplying the driving function with the weights of the tapering window.

Fig. 2.14 shows the effect of a tapering window for a linear secondary source distribution synthesising a plane wave. On the left-hand side, no tapering window is applied, leading to clearly visible disturbances in the sound field. On the right-hand side, a raised-cosine shaped tapering window as shown in fig. 2.15 is applied over 30% of the array length, efficiently reducing the diffraction artefacts.

2.1.4.3 Reflective environment

The restriction to using monopoles only for the secondary source distribution and the consequently arising outer field makes an anechoic environment for WFS necessary to avoid reflections that disturb the desired sound field. This is unattainable in

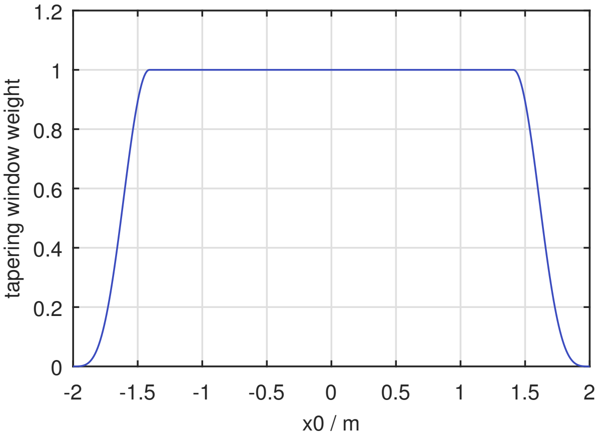


Figure 2.15: Raised-cosine shaped tapering window as applied in fig. 2.14, right-hand side. The outer secondary sources covering 30% of the array length, i.e. 15% on each side, are faded out towards the edges.

practice, as even an anechoic chamber is only an approximation of an anechoic environment. Furthermore, WFS arrays can only be installed inside an anechoic chamber for research purposes. Therefore, applications of WFS always have to deal with a more or less reflective listening room in which a loudspeaker array is situated.

The resulting reverberation is different from the one that would belong to a real source in the listening room. This can be illustrated by examining the impulse responses of a real point source inside a room in comparison to the impulse responses of a virtual point source in free field and inside the room at the same position as the real source in fig. 2.16. The room for these analyses has been simulated with the image source method [AB79]. As can be observed clearly at the beginning of the impulse response, the single impulses constituting early reflections in case of the real point source have made room for a series of impulses caused by spatial aliasing in the case of the virtual source. As a result, the gaps between the early reflections are progressively filled in, leading to an impulse response that is much more dense in the case of a virtual source than for a real source [ES15].

Two- or three-dimensional numerical simulations of SFS in reflective environments have mostly been performed within the scope of compensation algorithms as preliminary studies [GB07, SK12] or for verification purposes [BA05, PSR05]. Measurements of synthesised sound fields in listening rooms have typically been conducted with linear microphone arrays in the horizontal plane to demonstrate the performance of a specific SFS system [SRdV97, Kut03]. Such simulations and measurements can show additional wave fronts stemming from reflections in illustrative figures, but typically do not lend themselves to in-depth analyses apart from e.g. finding a physical measure for successful reduction of reflections by a room compensation algorithm. Data based on measurements is additionally limited in its explanatory power by spatial aliasing caused by the microphone arrays.

The problems arising from a reflective environment are not exclusive to SFS. Conventional stereophonic reproduction faces similar challenges [Too06, Rum08], though restricted to a single listener position, the ‘sweet spot’, while SFS aims at controlling an extended listening area. For the reduction of reflections of a listening room, dif-

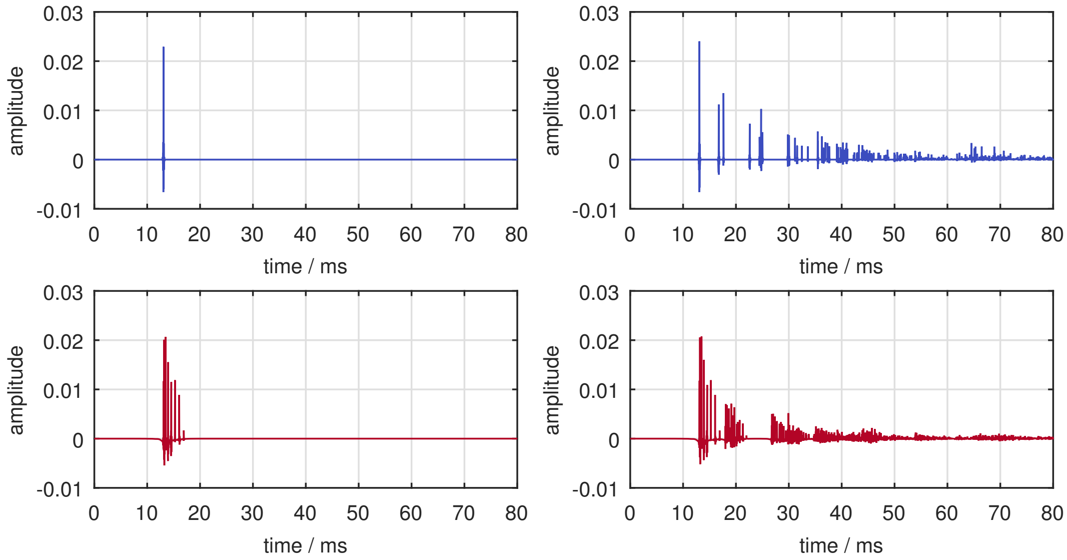


Figure 2.16: Impulse responses of a real point source (upper row) and a virtual point source (lower row) in free field (left column) and inside a rectangular listening room (right column). The virtual source is synthesised by 2.5D WFS with driving function eq. (2.18). The geometry of the setup is depicted in fig. 4.1 with a linear array of 15 secondary sources with $\Delta x_0 \approx 20\,\mathrm{cm}$, impulse responses are evaluated at the centre listening position $\mathbf{x}_{\mathrm{ref}}$. The room is simulated with the image source method (see section 3.2.2) and a constant reflection factor of $\beta = 0.7$.

ferent strategies have been developed. Apart from passive absorbent materials lining the walls, also active methods have been proposed such as the control of the wall impedance by electroacoustical systems on the walls [GKR85]. Other approaches use higher-order loudspeakers to eliminate the outer field that arises from the limitation to use monopoles only (cf. section 2.1.1). This reduces the reflections of the listening room [CJ12, PAS12], but constitutes a step up in complexity back to using the full KHI in eq. (2.1) for SFS. [BAP10] chose a different approach by using directional loudspeakers to exploit reflections of the listening room as image sources that are incorporated in the synthesis of the desired sound field, thus reducing the required number of loudspeakers.

The largest number of countermeasures for reflections of a listening room are made up of algorithms for compensation of reflections, often using adaptive techniques with microphones to control the sound field [Spo06, Cor06, GB07, SK12]. Using the feedback information of microphones is necessary e.g. due to changes in the speed of sound with room temperature, making the listening room a time-variant system for this application. A physically perfect compensation of reflections is not obtainable with multi-channel systems existing in practice as has already been discussed for one-channel systems due to problems in constructing inversion filters [Fie01]. Moreover, in the case of SFS, not a single point has to be equalised, but the whole listening area, making it necessary to analyse and control this area in the entire frequency range of human hearing. This is not possible when using a loudspeaker array that is only arranged in the horizontal plane. Reflections in this plane can be compensated to some degree, but not reflections from the floor or ceiling as the array provides no control in this dimension. Above f_{alias} in eq. (2.21), no active compensation of reflections is possible [CN03, SRR05]. As a complete elimination of room reflections is not feasible, algorithms for compensation of reflections should be perceptually evaluated which is still missing in the literature so far.

2.2 Perception of Wave Field Synthesis

The practical limitations in section 2.1.4 lead to deviations from the desired sound field, which can influence the perception of a listener. Researchers have been investigating different aspects of human perception in the context of WFS to determine if the violation of theoretic assumptions is degrading the quality of reproduction with regard to these perceptual aspects. Especially the influence of spatial aliasing on perception has been studied intensively. Past research focussed mainly on the categories of spatial and timbral perception in SFS, which themselves are made up of different aspects, i.e. spatial perception includes direction and distance of a source, but also the width of a source and the perception of a room. Several studies have tried to untangle the web of perceptual aspects of spatial audio. Section 2.2.1 gives an overview on this topic. The following sections 2.2.2 and 2.2.3 are summarising the state of research concerning spatial and timbral perception in WFS reproduction, respectively. These perceptual aspects have also been investigated in the current thesis on WFS in a listening room.

2.2.1 Perceptual dimensions of spatial audio

There exist several studies with the scope to determine the perceptual aspects relevant for evaluation of spatial audio techniques and systems, e.g. [BR99, KZ01, LEL^{+}14, FBM17]. The following sections discuss the applicability of the determined vocabularies of these studies for the topic at hand.

2.2.1.1 Scope and completeness of vocabularies

The studies differ in their ways of eliciting the attributes of interest, using e.g. the Repertory Grid Technique (RGT) [Kel55] in [BR99], Audio Descriptive Analysis & Mapping (ADAM) [KZ01] in [KZ01, FBM17] or the Focus Group method [SS90] in [LEL^{+}14]. The most defining difference between these methods might be if auditory stimuli are presented (RGT and ADAM) or if instead the study is based on a discussion between experts of the field without using any audio material (Focus Group method). The latter method might bear the risk of leaving out important aspects, while the use of auditory stimuli could cause an unintended limitation of the range of validity of the compiled list of attributes. Other differences in the results of these studies can arise due to the varying choice of spatial audio technologies that are targeted in each study. The study [LEL^{+}14] e.g. included SFS in its scope, while [KZ01] did not mention this type of techniques, though both studies claim to cover the domain of ‘spatial audio’.

Results from the above studies strongly differ in their levels of detail. [BR99] intended to find only spatial attributes in their study. [KZ01] directed their test subjects on finding timbral and spatial attributes, but focussing on spatial attributes. Thus, these two studies did not cover all aspects of perception in spatial audio reproduction. [LEL$^+$14] developed a very detailed set of attributes consisting of 48 items sorted into eight categories (timbre, tonalness, geometry, room, time behaviour, dynamics, artefacts, general impression). In the study by [FBM17], two vocabularies were developed, one with 27 perceptual attributes by experienced listeners and one with 24 attributes by inexperienced listeners, partly overlapping. Attributes comprised spatial and timbral aspects of perception, but also artefacts of reproduction such as distortion or compression and quality-related attributes.

To evaluate the influence of a reflective environment on human perception in WFS in this thesis, it would be desirable to cover all relevant aspects of perception in necessary detail. As outlined above, care has to be taken, when using the results from the above studies to evaluate this topic. This relates to the suitability and completeness of the proposed set of perceptual attributes for the topic of WFS in a listening room as well as to the need of an adequate investigation method for each attribute. Considering the suitability of the proposed vocabularies, it is deemed important that their range of covered spatial audio technologies explicitly includes SFS, which is only the case for [LEL$^+$14]. Concerning completeness in the context of the influence of the listening room, a comparison of [LEL$^+$14] with a set of attributes generated for perceptual evaluation in room acoustics by [WLA18] suggests that attributes related to perception of reverberation might be underrepresented in [LEL$^+$14] with only three such attributes compared to 29 in [WLA18]. Especially attributes such as 'room acoustic suitability' and 'irregularity in sound decay' from [WLA18] appear to be potentially also relevant for WFS in a listening room, considering undesired reflections by the listening room and the atypical structure of an impulse response of WFS in a listening room as illustrated in fig. 2.16.

Other challenges when trying to evaluate all perceptual aspects are posed by attributes that are multidimensional in nature, e.g. the attribute 'colouration' which is a change in timbre compared to a reference. The concept of timbre itself eludes a short and simple definition, it is typically defined implicitly by stating what it is not, cf. [Wie14] for a discussion on this topic. Different studies on perceptual attributes relevant for spatial audio come up with different lists of timbre-related items. For example, [LEL$^+$14] contains eight attributes related to timbre, such as 'low-frequency tone color' or 'comb filter coloration'. [FBM17] only lists two or three such attributes (the authors did not group the attributes), such as 'overall spectral balance' or 'spectral resonances' for experienced listeners, that tend to cover a broader range of perception, and three or four attributes such as 'bass' or 'richness of sound' for inexperienced listeners.

2.2.1.2 Adequacy of reporting methods

Many perceptual attributes found in the studies mentioned above are readily to be judged on a scale with two defined end points, such as 'loudness' on a scale between 'soft' and 'loud', forming a semantic differential as introduced by [OST57]. This way, a large number of perceptual aspects can be evaluated in a relatively short period of time. Apart from different concepts that test subjects might have of the exact expression labelling a scale end, which would introduce additional variance in the data, there are also specific attributes that cannot adequately be judged by providing just a simple scale. Especially spatial attributes related to geometric values of a sound scene, such as direction or distance of a source, require a dedicated reporting method.

2.2.1.3 Applicability of room acoustical parameters

Additionally, in the context of perception in rooms, the traditional room acoustical parameters as defined in [ISO09] exist to predict certain perceptual aspects from the (directional) impulse response in a room. They have been established in concert hall acoustics and it is unclear to what extent they can be applied to small room acoustics [Too06, KBJvW14] or to the special nature of synthesised sound fields. In particular, the temporal separation in early and late reflections at 50 and 80 ms for measures like ‘clarity’ in music and ‘definition’ in speech signals, respectively, are doubted to be adequate for small rooms as due to the differences in room size, modal behaviour is shifted to a different frequency range [KBJvW14]. In the case of synthesised sound fields, the very early and densely spaced pulses emerging from spatial aliasing differ considerably from the usual distribution of early reflections even in small rooms. A measure like the time gap between direct sound and the first reflection for prediction of perceptive attributes such as ‘intimacy’ [Ber04] might lose its meaning in such sound fields. Other energetic measures such as the early decay time (EDT) and the reverberation time can be shown to yield the same values for a real and a virtual point source synthesised by WFS in a listening room at the same position [ES15]. Due to the differences of the sound field of a real and a virtual source in a listening room described above, it is to be doubted if reverberance and perceived room size, that are predicted by the EDT [Bar95, SB95] and the reverberation time [KBJvW14], respectively, are really the same in these situations, though.

2.2.1.4 Conclusions for the investigations in this thesis

Due to the reasons outlined above, this thesis concentrates on a small subset of perceptual attributes relevant for WFS in a listening room instead of evaluating an extensive list of attributes that might not be adequate for the topic. Concluding from the majority of past studies, that spatial and timbral fidelity are the most important aspects of a spatial audio reproduction method, azimuthal localisation (section 4) and colouration (section 5) were investigated as these were considered the most relevant spatial and timbral perceptual aspects in 2.5D WFS in a listening room. The apparatus and reporting method for a localisation experiment on azimuthal source direction have carefully been evaluated before conducting the actual experiment, cf. section 3.3. This way, the suitability and accuracy of the proposed method are ensured. In the light of the different outcomes concerning timbre-related aspects of spatial audio in the above discussion, this thesis investigates colouration as a whole compared to a reference stimulus, cf. section 5.

2.2.2 Spatial perception of Wave Field Synthesis[3]

In free field conditions, early research by [Vog93] showed accurate localisation of a virtual point source synthesised by a linear WFS array for a central listening position.

---

[3] Parts of this section have been published in [ES20].

The azimuthal direction of sources in 2.5D WFS has been more intensively investigated by [Wit07] and by [Wie14] who included changes in listener position. According to their studies, an accurate directional localisation of virtual point sources or plane waves in the horizontal plane is possible as long as the spacing between secondary sources is not too large and the listener is not positioned close to the secondary source array. This is attributed to the fact, that though spatial aliasing is generating additional trailing wave fronts, WFS is capable of creating an accurate first wave front. As perception of direction is dominated by the direction of the first wave front, accurate localisation for virtual sources is possible. This psychoacoustical effect is termed the precedence effect [LCYG99], which is originally the explanation for the ability of humans to localise sources in reflective environments. Sound that reaches the listener after the first wave front, such as reflections or spatial aliasing artefacts, does not influence the perceived direction as long as its delay lies between 2 and 50 ms. For shorter delays, summing localisation would occur. For longer delays, an echo would be perceived with the echo threshold depending strongly on signal type. For large secondary source spacings (1.43 m in [Wie14]), virtual sources are localised in the direction of the nearest loudspeaker. This effect has also been found in an experiment by [LSK$^+$07]. It could also explain the localisation errors for listeners close to the secondary source distribution, as the relative distance, seen from the listener, is quite large. For focused sources, which are virtual point sources inside the listening area, in 2.5D WFS, the localisation accuracy is degraded, possibly because of inconsistent binaural cues caused by, in this case, preceding additional wave fronts due to spatial aliasing [WGS10].

Concerning the perception of distance, [UMW04] found that the curvature of the wave front alone does not provide sufficient cues for distance judgement in WFS. This has been confirmed by the more detailed study by [Wit07] that showed that there is no reliable perception of distance in 2.5D WFS in free field for sources without additional cues provided by e.g. reflections or level differences.

Localisation properties of WFS in a listening room have not been intensively studied so far. [SRdV97] conducted experiments on the perception of direction and distance of a virtual point source in three listening rooms (anechoic chamber, auditorium, concert hall) at different listener positions. The resulting localisation accuracy proved to be fairly good with an RMS error of ±2.8° for a virtual source compared to ±1.8° for a real source with speech as audio content. For noise and music the accuracy was slightly decreased, especially for virtual sources. Distance judgements were also possible to some extent in the experiments by [SRdV97]. The conditions of the three rooms were not directly comparable, though, as different loudspeaker array configurations for 2.5D synthesis were used and the properties of the rooms were not varied systematically. Moreover, the experiments suffered from very low subject numbers (2–7 for the different rooms). [Ver97] compared directional localisation of real and virtual point sources synthesised by a linear WFS array in an anechoic chamber and in a listening room with a short reverberation time and found a slightly larger localisation error (increase of approx. 1°) and increased standard deviation (increase of approx. 0.5°) of reported answers for real as well as virtual sources in the reflective environment. In the experiments of [KS03, SK04], WFS systems in a living room and a cinema were evaluated. Results also show that correct localisation in WFS in a reverberant environment is possible, but the data

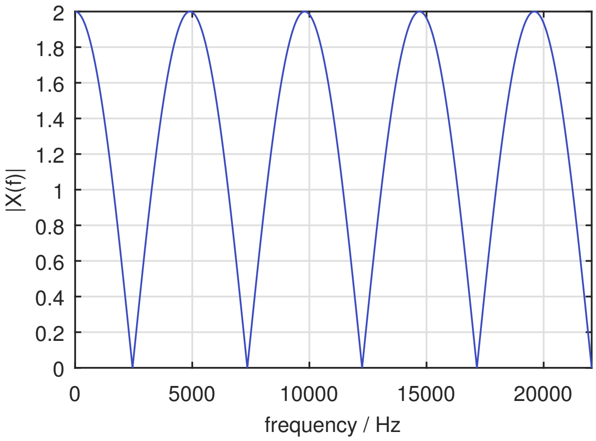


Figure 2.17: Absolute value of the magnitude spectrum $X(f)$ of the signal $x(t) = \delta(t) + \delta(t - t_0)$ illustrating a comb filter. The delay is $t_0 \approx 0.23\,\mathrm{ms}$. The resulting distance between minima of $|X(f)|$ is $\frac{1}{t_0} = 4410\,\mathrm{Hz}$.

is not analysed regarding the influence of the listening rooms specifically.

There are no studies on the spatial perception of 3D WFS known to the author, probably as such systems are not commonly in use. Most likely, 3D WFS would also not work very well due to spatial aliasing in the frequency range of elevation cues. There is also very few research on WFS properties for spatial properties other than direction and distance. [Sta97] reported an experiment in which he compared the perceived source width of real and virtual point sources in the same three listening rooms as in his localisation experiment. Results were unfortunately inconclusive, which might be due to the use of only four test subjects or due to a high complexity of the topic.

2.2.3 Timbral perception of Wave Field Synthesis

A reproduction method can induce colouration as a change in timbre compared to an internal or external reference, which is usually considered an undesired artefact. WFS typically introduces some colouration. It is caused by spatial aliasing artefacts that produce strong ripples in the magnitude spectrum at high frequencies, cf. fig. 2.8. These ripples are often referred to as ‘comb-filter-like’, e.g. [Wie14], due to their similarity in appearance, origin and perception. However, a comb filter is typically characterised by a very regular structure as illustrated in fig. 2.17. In this figure, only one delayed and attenuated version of a signal is added to itself, but a regular comb filter can also arise from the superposition of several delayed signals [Kut17].

The larger the distance between secondary sources is, the lower the spatial aliasing frequency (cf. fig. 2.8) and the stronger the perception of colouration. This has been shown by several studies [dB04, Wit07, Wie14]. The study by [Wie14], who performed the most detailed research with different listener positions and secondary source spacings, found slight colouration even for secondary source spacings as small as 0.3 cm when presenting noise as audio content. The perception of colouration is additionally dependent on the audio content of a virtual source. In [Wie14], colouration ratings were lower for female speech as audio content than for pink noise and music (a section of an instrumental electronic song including cymbals and subtle white noise). This is suspected to be due to the bandlimitation of speech, resulting in a narrower frequency band that is affected by spatial aliasing.

The cited studies on colouration in WFS have only been concerned with WFS systems in an anechoic environment. The presence of reverberation in a listening room conceivably also has an influence on the perception of colouration in WFS. Very few research has been conducted on this topic so far. Reflections of the listening room can possibly generate colouration of any sound emitted in the room highly dependent on the properties of the room and positioning of source and receiver. As long as expectations of the listener for the environment are met, though, the perceived colouration will be less than what the frequency spectrum for a certain position in the room might suggest [Vor98]. Also if the ear signals after turning of the listener's head inside a room are compared, dramatic changes in the frequency responses can be observed, though no change in timbre is perceived. This phenomenon is termed binaural decolouration [Zur79, Brü01].

For colouration of WFS in listening rooms, paired comparison experiments by [Sta97] give a first impression on the topic. Slightly moving real and virtual sources, that do not lead to a change in localisation, but contain varying colouration due to the source-dependence of the spatial aliasing frequency, were compared for WFS conditions in three different rooms (anechoic chamber, auditorium, concert hall) with noise as audio content. Results indicate a reduction of colouration for virtual sources in the reflective environments, but the listening experiment was only conducted with four subjects and the used WFS system was operated with a sampling rate of only $f_\mathrm{s} = 32\,\mathrm{kHz}$. Additionally, [Sta97] included stimuli with a randomisation of the high-frequency delays of the driving function in his listening experiment. This approach should dissolve regular patterns in the synthesised sound field and was thus suspected to reduce colouration, which tendencies in the listening experiment seemed to confirm. Stereophonic reproduction is also subject to comb filtering effects that lead to colouration. [Pul01] could show in a listening experiment that this type of colouration is lessened by reflections of a listening room, but that the reverberation itself might also induce colouration.

3 Methods

In this chapter, the methods employed in this thesis to investigate the influence of the listening room on Wave Field Synthesis are described. All steps necessary to yield reliable and meaningful results are treated. Section 3.1 integrates the intentions for diligent research in this thesis with the test criteria from classical test theory. The experiments in this work rely on simulation methods, which is motivated in section 3.2 together with a discussion of their fundamentals and limitations. Section 3.3 reports on the preceding evaluation experiment of the apparatus for the listening experiment on localisation. The last section 3.4 discusses the realisation of listening experiments on colouration and their limitations.

3.1 Objectivity, validity and reliability

The author of this thesis strongly believes it to be necessary to carefully select one's methods and to pay attention to detail in order to yield meaningful research results. Classical test theory provides three fundamental concepts that have to be taken into account to achieve this: objectivity, validity and reliability [BD06], that are discussed in the context of this thesis.

Objectivity demands an unbiased measurement. It is possible, though, that the experimenter influences the results unintentionally by giving unsuitable, misleading or confusing instructions to the test subjects either via the reporting method or software to obtain the answers of test subjects or in verbal instructions. Therefore, care has to be taken to give the same clear instructions to all test subjects in a listening experiment. To ensure objectivity, an in-depth documentation of the whole process is necessary that should also be available to other researcher that want to repeat the experiment to enable full benefit from the research results. It is not possible to report every detail necessary to exactly repeat a study in a scientific paper, partly due to restrictions concerning the length of a paper and partly for reasons of clarity and comprehensibility. Therefore, a scientific study should be accompanied by additional information, including software and code, datasets and statistical analyses, that need to be documented in detail to be of use to other researchers.

To really measure, what is claimed to be measured, is understood as the principle of validity. If there exist other variables, that are expected to be correlated with the aspect that is desired to be measured, the so-called criterion validity can be quantified by this correlation. Other types of validity, such as content validity, that describes if a test is conceptually adequate for the aspect that is to be measured, cannot be assessed in a quantitative manner [Dro11, Gar16]. To use a technical measure such as the room acoustical parameters defined in [ISO09] to conclude on perception in room acoustics can thus only deliver valid results if the area of applicability of these parameters is heeded. These room acoustical parameters have

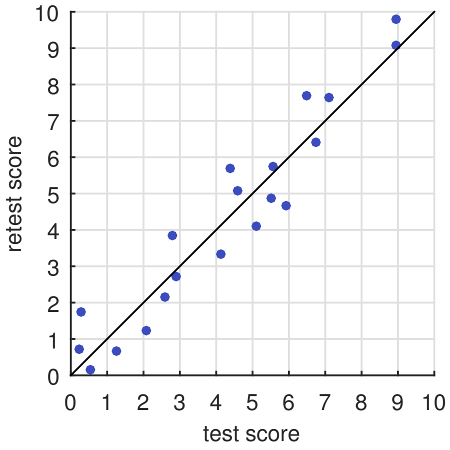


Figure 3.1: Illustration of the test-retest reliability in a scatter plot. In this example Pearson's correlation coefficient is approx. 0.96, as the blue data points are close to the ideal black line. This indicates a very good test-retest reliability of the underlying test method.

been developed for large performance venues and might not be appropriate for small rooms, for example. In contrast, a listening experiment on a perceptual aspect of room acoustics can be inherently more valid. Just as misleading instructions jeopardise the objectivity of an experiment, though, they might also render the results invalid and have to be exercised with care. Lastly, the different conditions in a listening experiment should be produced diligently so that only the intended variations between conditions are present. The stimuli used in this thesis are based on simulations that are susceptible to errors endangering validity. The following section 3.2 therefore treats the steps to generate valid stimuli by simulation.

Finally, research results have to be reliable, demanding a sufficient accuracy of the measurement results. Hence, the reporting method in a listening experiment has to be chosen carefully. An example from this thesis is the reporting method for azimuthal localisation of a source. While a bias in the reporting method can be classified as a threat to the validity of the experiment, a high variance interferes with its accuracy and thus reliability. A basic verbal reporting of an azimuth value in degree by test subjects is possibly not delivering an acceptable accuracy. In the case of localisation experiments, many studies have been concerned with finding superior reporting methods. In this thesis, the reporting method of the localisation experiment has been evaluated in a preceding experiment, cf. section 3.3. The requirement of reliability also includes that the repetition of the experiment under the same conditions yields the same result within acceptable limits. This type of reliability is known as test-retest reliability and it is usually quantified by calculating Pearson's correlation coefficient between the results of the test and the retest [Dro11, Gar16]. The relationship between test and retest results can also be visualised in a scatter plot as is illustrated in fig. 3.1: If the data points are close to the ideal black line, the reliability of the test method is high.

The above mentioned principles supporting objectivity and reproducibility have been incorporated in this thesis by relying on Open Research tools such as the Sound

Field Synthesis Toolbox[1] [WS12] and the SoundScape Renderer[2] [GS12] for generating figures and stimuli and for rendering of stimuli in the listening experiments, respectively. Furthermore, no closed source software has been used for room acoustical simulation. The stimuli, results and statistical analyses of the conducted listening experiments have been made publicly available as well as scripts that created illustrative figures for SFS in this thesis and the implementation of the room acoustical simulation. Additionally, measured headphone transfer functions (HpTFs) and Matlab code for calculating headphone compensation filters as well as for low-frequency correction of head-related transfer functions (HRTFs) have been published. The locations for download and more details of these additional informations are given in the respective sections and are summarised in appendix C. As a final remark, the author of this thesis believes it would be of great value, if more attention and appreciation would be given in the scientific community to the replication of other researcher's findings.

3.2 Investigation through simulation[3]

The study of WFS in a listening room requires the variation of several parameters related to the room, the array of secondary sources and the listener position to allow for generalisation of the results. While measurements of setups with these variations can be done sequentially, they are challenging to realise in a listening experiment where immediate comparisons and randomisation of stimuli are desired, which contradicts a static setup. To place listeners reliably in the same positions inside or in front of a loudspeaker array in steps of only a few ten centimetres and letting the listeners move between these positions during one listening experiment is difficult as has already been discussed by [WSR12b]. Making changes to the array of loudspeakers is only possible within strong limitations by using only parts of the existing loudspeakers in the array, whereas changing the array geometry in an instant is not feasible. Moreover, due to the dimensions of real loudspeakers it is not possible to make comparisons to arrays with very densely spaced secondary sources that are free from spatial aliasing in the frequency range of human hearing. The most problematic part in such listening experiments are variations related to the listening room. It is impracticable to modify the wall properties of the room during the listening experiment and changing the room or the room geometry is not possible at all. All of these more time-consuming modifications of the experimental setup could therefore only be compared in separate listening experiment sessions. This would not only take up a lot of time, but also prohibit the revelation of not so obvious differences between stimuli that could only be reliably judged by test subjects in direct comparisons.

To make all comparisons within one listening experiment possible and to allow for a broader range of variations, a simulation approach has been chosen in this thesis. Information about a reverberant environment that can be analysed or used in listening experiments can be represented as room impulse responses. They constitute the answer of a system to an impulse as input. As the system in this case is formed

---

[1] `http://www.sfstoolbox.org`

[2] `http://www.spatialaudio.net/ssr`

[3] Small parts of this section have been published in [EGWS17, ES20].

out of room, source and receiver, room impulse responses depend not only on the room and its wall characteristics, but also on positions and directivities of source and receiver. A lot of information about a certain setup can already be found in a single-channel room impulse response (RIR) that is captured by an omnidirectional receiver. When using a human head with microphones at the entrance or in the ear canals, binaural room impulse responses (BRIRs) can be recorded. They can be used to let test subjects listen to a certain auditive environment. This process is termed an auralisation and in this thesis it is performed by (dynamic) binaural synthesis. Room impulse responses that represent real and virtual sources in a reverberant environment have been acquired by room acoustical simulation. With this approach it is possible to compare real and virtual sources in free field or inside rooms in listening experiments freely.

There are also drawbacks in relying on simulation tools. The use of room acoustical simulation or non-individual binaural synthesis (using a dummy head for recording BRIRs instead of the head of the individual listener) might introduce a certain degree of ‘unnaturalness’ to the presented stimuli, often related to an unexpected timbre, which could introduce a bias in the results. Binaural synthesis is in general not a completely transparent simulation tool. Its technical realisation contains challenges evoked e.g. by the need to have impulse responses for discrete source or listener positions that have to be adapted to the listeners position or head direction. The technical fundamentals as well as limitations of binaural synthesis are given account of in section 3.2.1. In sections 3.3.1 and 3.4.1, the application of binaural synthesis for the investigation of localisation and colouration is discussed, respectively. The main challenges for room acoustical simulation are centred around the goal to perfectly replicate the acoustics of a real room. This includes the difficulties to incorporate all relevant wave phenomena such as scattering and diffraction as well as to measure the acoustic properties of materials and objects in a room. Section 3.2.2 therefore discusses the choice of room acoustical simulation in this thesis and describes the parameters of the used simulation method in detail.

Employing a simulation is not solely a compromise that has to be made, it also bears advantages. Additionally to the above mentioned possibilities of simulation of aliasing-free WFS, accurate listener placement and immediate comparisons, the influence of the directivities of real loudspeakers on the synthesis can be separated from the influence of the room as the simulation can incorporate perfectly omnidirectional sources. Different loudspeaker directivities can also be introduced if desired. Furthermore, a simulation enables full control over the wall characteristics that are typically difficult to vary in practice in predefined steps with the usually available acoustic treatment or even to determine by measurement [Bor05a, Vor13].

3.2.1 Binaural synthesis

3.2.1.1 Fundamentals of binaural synthesis

Binaural synthesis is an auralisation method based on the principle that it is possible to recreate any desired auditive environment for a listener if the sound pressure at the ear drum matches the sound pressure at the ear drum in the real sound field. This way, also non-existing auditive environments can be auralised by synthesising the proper ear signals. To achieve this, the directional filters stemming from pinnae, head and torso have to be taken into account. In free field, this filter set is termed head-related transfer functions (HRTFs), whereas in a reverberant environment the time-domain equivalent is typically used termed binaural room impulse responses (BRIRs). They are usually measured with miniature microphones at the entrance of or in the ear canals [Møl92]. As there exist anthropometric differences between individuals, individual HRTFs or BRIRs are required. To reduce effort, it is usual to employ HRTFs or BRIRs measured with a dummy head or a head and torso simulator (HATS), though. This so-called non-individual binaural synthesis is also capable of creating the impression of a three-dimensional auditive environment, but several perceptual aspects can be impaired or modified compared to the real prototype sound field [MSJH96, Mas12, BLW17a].

To synthesise the correct ear signals for a listener in binaural synthesis, all transfer functions from the measurement and reproduction equipment have to be compensated. For most elements of the equipment, this is a straightforward task. When measuring BRIRs, the directivity of the measurement loudspeaker cannot be compensated for, though. Moreover, the last step in the chain of binaural reproduction poses difficulties. The ear signals in binaural synthesis can be presented to the listener either by loudspeakers, which then requires crosstalk cancellation techniques [MFV11], or by headphones, which is used in this thesis. When headphones are used for reproduction, the transfer function of the headphones including the path to the entrances of the ear canals – together termed headphone transfer function (HpTF) – has to be compensated for by inverting the HpTF, which yields the headphone compensation filters. Though this, too, is a matter of individual differences, using non-individual compensation filters is easier to realise. Furthermore, results from [BL10] suggest that combining non-individual HRTFs with non-individual headphone compensation filters leads to superior results regarding colouration. As HpTFs typically belong to non-minimum phase systems, direct inversion is not possible [KN99]. Different methods have been proposed to handle this problem. [SL09] compared several approaches perceptually, revealing the method of regularisation in the frequency domain [NSL04, KNHOB98] as the best choice which has also been applied in this thesis. The straightforward implementation of this filter generation method used in this thesis has been made publicly available[4] by the author [EGWS17]. To calculate the compensation filters, the underlying HpTF is typically taken as the mean out of several measurements with repositioning the headphones on the HATS between measurements. This allows for averaging out differences due to slightly varying positions of the headphones. Fig. 3.2 shows the HpTF of the FABIAN HATS [LW09] with headphones type AKG K601 averaged out of 21 measurements (12 measurements per channel with three outliers removed) as well as the compensation result. The measured HpTFs are publicly available[5] [BLW$^+$17b]. The high-frequency range is only weakly compensated as a regularisation filter with high-shelf characteristics has been used. This way ringing artefacts are avoided that could arise from slight inter-individual shifts of narrow-band notches in the high-frequency range. When not using HpTFs of the same headphone that is used for reproduction, it is recommended to use filters that are averaged out of measurements of the left and right channel for one or more headphones of the same type. When measuring with the

---

[4] `http://doi.org/10.5281/zenodo.401042`

[5] `http://doi.org/10.14279/depositonce-5718.3`

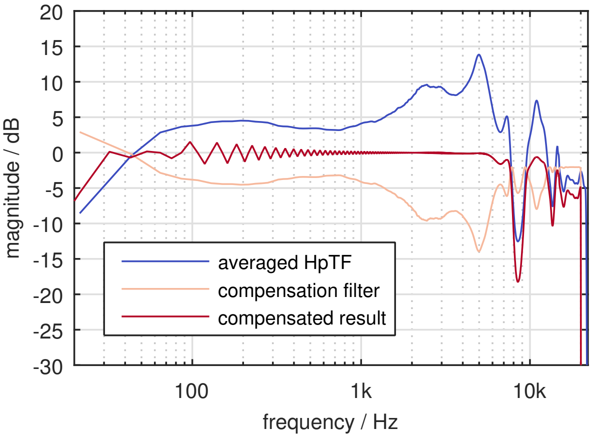


Figure 3.2: Mean HpTF of the FABIAN HATS with headphones type AKG K601, averaged out of 21 measurements (12 measurements for left and right channel each, three outliers removed), headphone compensation filter and compensated result. The ripple in the frequency response of the compensated result is caused by the design of the target bandpass as a linear-phase FIR filter with a least-squares approach and windowing [EGWS17].

widely used KEMAR manikin type 45BA, this has the additional advantage that the slight level difference of approx. 0.5 dB between the left and right HpTF due to different installation depths of the microphones is not falsely compensated for. The shape of the KEMAR manikin is on purpose not exactly symmetric as it is intended as a model for a real human head. The different microphone installation depths of 3 mm result from differing recesses for the pinna on the left and right side. This leads to a level differences due to the inverse square law which is only of importance for very close sources which is the case when measuring HpTFs. A false compensation of this slight level difference might especially influence results of localisation experiments as it is associated with a lateralisation slightly outside the median plane [Bla97, ZF05]. Lateralisation is the lateral displacement of an auditory event inside the head on a line between the ears as experienced in headphone presentation that does not rely on externalisation cues as e.g. in binaural synthesis.

A headphone suitable for binaural synthesis should apart from excellent reproduction quality including a high spectral bandwidth and signal-to-noise ratio exhibit a frequency response that is robust with respect to repositioning on the head of a listener. Additionally, for in-situ comparisons with real sound fields, it should comply to the ‘free air equivalent coupling’ criterion [MHJS95], i.e. it should approach the acoustic impedance of free air as seen from the ear canal entrances. This can be achieved by a so-called extraaural headphone where the transducers are placed with a distance of a few centimetres to the pinnae of the listener [ESLW12]. Alternatively, an open headphone is typically used.

As common measurement loudspeakers lack energy at very low frequencies, measured HRTFs contain random results in this frequency range due to measurement noise, e.g. from microphones, that should be corrected [Xie09]. Different strategies have been proposed to achieve this. The missing information can be completed by means of geometric models of head and torso [ADD$^+$02], with results from the boundary element method (BEM) [GODZ10] or by cross-over filtering with an ad-

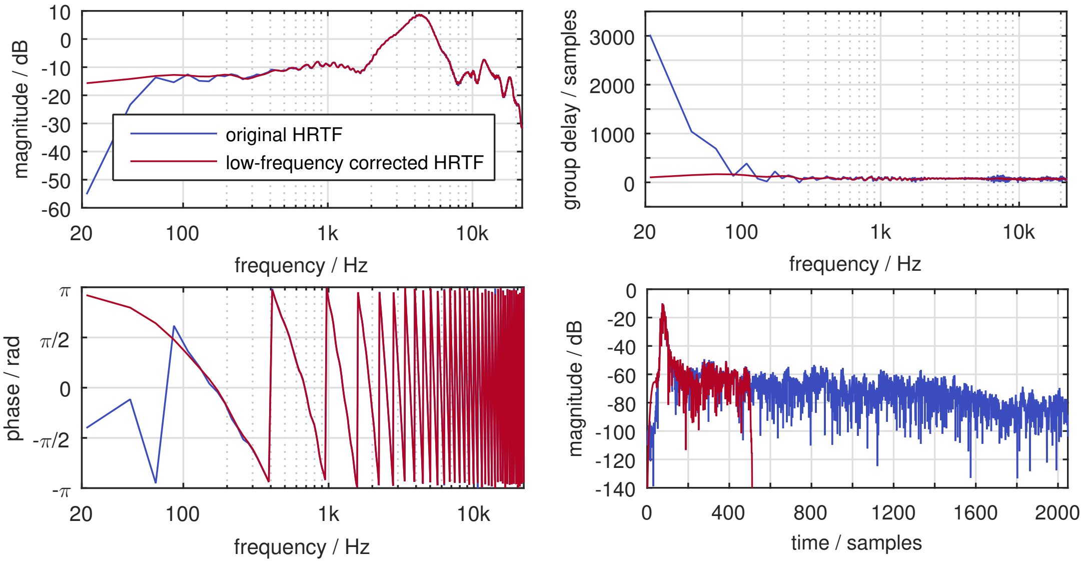


Figure 3.3: HRTFs and head-related impulse responses (HRIRs) of a KEMAR manikin for 0° azimuth in the horizontal plane, left ear, before and after low-frequency correction.

equate low-frequency response [Ber13]. All methods have in common that they lead to a monotone function for the phase and an approximately linear magnitude response at low frequencies. This can be expected as a dummy head or HATS does not present an obstacle to sound waves with large wavelengths. This finding directly translates into the approach by [Xie09] where the magnitude response is set to a constant value and the phase is extrapolated linearly. Fig. 3.3 shows an example of a low-frequency corrected HRTF measured on a KEMAR manikin. The original HRTF set[6] [WGRS11] as well as the low-frequency corrected version accompanied by the code used for the correction[7] [EGWS17] are publicly available. The magnitude below 100 Hz is replaced by the mean value from 100–300 Hz and the phase is linearly extrapolated from the same frequency range under the assumption that the data in this frequency range is reliable. The correction of the low-frequency range of the HRTFs bears the additional advantage that the implausibly high group delays at low frequencies are decreased, cf. fig. 3.3. Thus, the HRTFs can be truncated to 512 samples which saves storage space and makes usage computationally more efficient. The truncation comes with a slight deviation of the magnitude response at low frequencies that is now not perfectly linear any more. The low-frequency correction and subsequently possible truncation of the HRTFs is not only a matter of computational efficiency but also crucial to allow for interpolation of the HRTFs in the frequency domain, separately for magnitude and phase. As random phase values at low frequencies prohibit a proper phase unwrapping, interpolation of HRTFs with uncorrected low-frequency content can yield unexpected results, e.g. the inversion of the interpolated impulse response as is demonstrated in fig. 3.4. The HRTF for the left ear and an azimuthal direction of 10° in the horizontal plane (i.e. source slightly to the left) is linearly interpolated out of the HRTFs for 0° and 20° azimuth. The in-

---

[6] `http://doi.org/10.5281/zenodo.55418`

[7] `http://doi.org/10.5281/zenodo.401041`

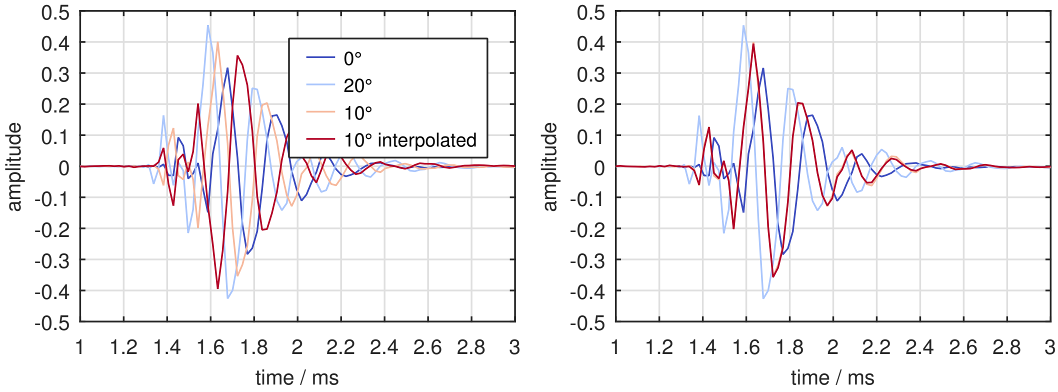


Figure 3.4: Linear interpolation in the frequency domain of two HRTFs for 0° and 20° azimuth to an HRTF at 10° azimuth. The ground truth at 10° azimuth is given as well. Left: HRTFs with corrupt phase at low frequencies, right: HRTFs with low-frequency correction.

terpolation is performed with the help of the Sound Field Synthesis Toolbox [WS12], release 2.5.0[8]. As can be seen, the interpolation based on the original data with the corrupt phase leads to an inverted impulse response for the 10° direction while the interpolation based on the low-frequency corrected HRTFs shows the expected sign.

Binaural synthesis can be static, where a single pair of ear signals is presented to the listener, or dynamic, where the ears signals are adapted according to the position and orientation of the head of the listener in space, improving also externalisation in binaural synthesis – the localisation of sources clearly outside the head of the listener [MOCM01, HSM$^+$17]. Typically only the head rotation of the listener is tracked and the ear signals are adapted accordingly. Due to the immense effort in required data, translational movements of a listener are less often tracked. [WSS14, NJNN18] investigated approaches to make translational movements possible with reduced effort. [LMW08] investigated the necessary resolution of HRTF and BRIR data sets for head movements and found a grid resolution of 2° for both horizontal and vertical head movements inaudible even for noise as critical audio content. For BRIRs in general, even coarser resolutions are acceptable. When the resolution of the available filter set is not sufficient for the intended task in dynamic binaural synthesis with tracking of head rotation, directional interpolation of HRTFs is necessary. According to [MPC05] a 4° resolution is sufficient to interpolate inaudibly except for directions from below the subject. HRTFs can be interpolated in the time domain with extraction and separate interpolation of the time of arrival (TOA) or in the frequency domain with separate interpolation of magnitude and phase. Other approaches perform the interpolation in the spherical harmonics domain [DZG04]. Several studies have been concerned with finding a (perceptually) superior method for interpolation that differ in details such as the method of TOA extraction, e.g. [RBW95, HBS99, BRLW15]. Less attention has been given to the selection of HRTFs that are the basis for the interpolation to a new direction. This appears to be especially important when the available HRTF grid is coarse. Promising approaches are the Delaunay triangulation [Gam13] or the Voronoi interpolation [Sib81] that are both implemented in the Sound Field Synthesis

---

[8] `http://doi.org/10.5281/zenodo.2597212`

Toolbox.

The rendering in dynamic binaural synthesis inevitably produces latency between the moment the listener turns his head and the play-back of the proper ear signals. Sources for latency include the head tracker update rate, parameters of the convolution engine for rendering such as the block size, delays of linear-phase headphone compensation filters and time of flight determined by the distance between source and receiver. The maximum latency in binaural synthesis that is still unnoticed by the listener has been investigated by several studies. Applying a strict criterion of inaudibility, the threshold latency appears to partly depend on the audio content [BSK05, Mac04, YIS06], but larger differences have also been found between individuals or when using BRIRs [Lin09]. For noise, [BSK05] found the strictest threshold of 60 ms. Fortunately, not every perceptual aspect is impaired by a greater latency. Localisation is only slightly influenced by it, resulting merely in a higher response time of subjects [Bro95, San96, Wen01, BSM$^+$04].

To evaluate the overall quality of a binaural simulation, a differentiation between plausibility and authenticity can be made. A binaural simulation is plausible when it is perceptually identical to an internal reference of a listener, while for authenticity an external reference is given. Plausibility and authenticity have to be judged for the whole system for binaural synthesis, including the HRTFs/BRIRs, headphones and headphone compensation filters and the renderer. [BLW17a] tested a system with individual dynamic binaural synthesis on authenticity in different anechoic and reverberant environments. Differences between simulation and reality could only be detected with noise as audio content, while the system proved to be authentic using a speech stimulus. The anechoic environment was shown to be more demanding for achieving authenticity in binaural synthesis, resulting in higher detection rates of differences. The percepts that were rated as most different when comparing reality and simulation were several aspects related to colouration as well as distance. These differences do not seem to play a major role when only plausibility of a binaural system is desired, as [LW12] found binaural synthesis to be able to achieve a high degree of plausibility.

3.2.1.2 Employed head-related transfer functions and their processing

For the synthesis of the stimuli for the listening experiments in this thesis, a three-dimensional HRTF dataset was necessary in high resolution. The used dataset stems from the FABIAN HATS [BLW$^+$17b]. It is full-spherical with HRTFs for 11,950 directions with a dense spatial resolution (2° in elevation, 2° great circle distance in azimuth) measured for different head-above-torso orientations and is publicly available[9]. Only the dataset for 0° head-above-torso orientation was used. The lower part of the sphere, which could not be acquired by measurement, as well as inconsistent information in the HRTFs at low frequencies and for the TOA has been filled in by numerical simulation with BEM.

Though [LMW08] state a resolution of 2° for HRTFs and BRIRs to be inaudible, the HRTF dataset was still considered too coarse for a listening experiment targeting localisation. This holds especially true for direct sounds in the horizontal plane, as the localisation accuracy of human hearing can be as good as 1° for frontal sound

---

[9] `http://doi.org/10.14279/depositonce-5718.3`

Table 3.1: Estimated latencies for the binaural system used in the listening experiments on localisation. The sampling rate is $f_\mathrm{s} = 44.1\,\mathrm{kHz}$.

Source for latency	Latency	
	evaluation experiment	main experiment
block size renderer	$\frac{2\cdot 256\,\mathrm{samples}}{f_\mathrm{s}} = 11.6\,\mathrm{ms}$	$\frac{2\cdot 512\,\mathrm{samples}}{f_\mathrm{s}} = 23.2\,\mathrm{ms}$
linear-phase HpTF compensation filter	$\frac{\frac{2048-1}{2}\ \mathrm{samples}}{f_\mathrm{s}} = 23.2\,\mathrm{ms}$	
headtracker Polhemus Patriot	update rate $60\,\mathrm{Hz} \mathrel{\hat{=}} 16.7\,\mathrm{ms}$ + latency $18.4\,\mathrm{ms}$	
time of flight in impulse responses	$5.9\,\mathrm{ms}$	$8.1\,\mathrm{ms}$
overall latency	$\approx 76\,\mathrm{ms}$	$\approx 87\,\mathrm{ms}$

incidence [Bla97]. As interpolation from HRTF grids of up to 4° step size have been proven to be inaudible [MPC05], the FABIAN HRTF dataset was interpolated to a $1° \times 1°$ Gauss grid, as no other full-spherical HRTF dataset was publicly available to the knowledge of the author. This saves computation time in the calculation of the image source method employed for room acoustical simulation in this thesis, cf. section 3.2.2.2. The interpolation was performed in the time domain with estimation of the TOA by cross-correlation and HRTF selection according to Delaunay triangulation as implemented in the Sound Field Synthesis Toolbox.

Headphone compensation filters were generated by regularisation in the frequency domain. The Matlab code for generating the headphone compensation filters is documented in [EGWS17] and publicly available[10]. The HpTFs of the FABIAN HATS for different headphones including headphones type AKG K601, that have been used in the listening experiments in this thesis, can be found in the FABIAN HRTF database [BLW+17b].

The latency of the whole system in dynamic binaural synthesis for the localisation experiments was estimated by taken the sources with the largest contribution to the overall latency into account as listed in table 3.1. The overall latency of approx. 76 ms in the evaluation experiment (described in section 3.3) and 87 ms in the main experiment on localisation (described in chapter 4) was considered sufficient for the localisation task according to literature, cf. section 3.2.1.1.

3.2.2 Room acoustical simulation

3.2.2.1 Choice of room acoustical simulation

Room acoustical simulation aims at recreating the phenomena of sound propagation inside reflective environments. The current models mostly rely on hybrid approaches that combine the image source method [AB79] for early reflections with a stochastic ray tracing [KSS68] for late reverberation [Vor08]. This accounts for the fact that in room acoustics only early reflections are purely geometrical while later reflections are more diffuse due to more components stemming from scattering and diffraction. As ray tracing is also far more computationally efficient compared to the image source method, these hybrid models are less time consuming than simulations relying solely

---

[10] `http://doi.org/10.5281/zenodo.401042`

on the image source method. The hybrid model RAVEN of the Institute of Technical Acoustics of the RWTH Aachen University [Sch11] is even capable of doing real-time auralisations when parameters are chosen appropriately. Hybrid models also strive to be more adapted to human perception where the exact level and direction of single reflections only matters in the early part of reverberation, but the late, more diffuse part has an impact as a whole [Vor08]. Modal wave behaviour, which is an important part of the acoustics of small rooms, is mostly not included in current room acoustical simulation software, e.g. in RAVEN it is only included in an experimental version [PAV11]. This makes the hybrid models only valid above the Schroeder frequency [Sch87, SK62]. There exist entirely modal models [SK02, Sav10], but these are very time consuming to calculate over the whole audible frequency range.

The physical and perceptual accuracy of current room acoustical software is still subject of ongoing research. A series of round robin studies [Vor95, Bor00, Bor05a, Bor05b] investigated if real rooms could be recreated by different software packages by comparing if room acoustical parameters were met. The simulation of the sound field in a real room requires as input not only geometrical data in appropriate detail, but also exact data of the acoustical properties of all surface materials. The latter are hard to come by as in-situ measurements are difficult [Bor05a, Vor13] and the usually employed laboratory measurements can only deliver insufficient information. For example in the case of the absorption coefficient, either the method in the reverberation room according to [ISO03] is used resulting in an absorption coefficient for diffuse sound incidence, or a measurement in the Kundt's tube that yields the absorption coefficient for normal sound incidence [Kut17]. Both are not appropriate for the partially diffuse sound field in a real room. Moreover, results from different measurement teams vary caused by different measurement equipment and by different algorithms for the calculation of the room acoustical parameters [Bor05a]. Thus, parts of the input data for the simulation have to be estimated and there is no clear ground truth for comparison, leading to the problem that differences of simulation results to the measurements do not reveal how close a simulation software is to reality, but also how good the acoustical input data has been estimated and/or measured. These influences cannot be separated in the results making it difficult to judge the quality of the simulation software. Moreover, results in the round robin studies did not only differ between simulation programs, but also when one program was operated by different users due to varying levels of experience or insufficient instructions [Bor00, Bor05b].

To gain more insight into the perceptual accuracy of room acoustical simulation, a round robin on auralisation including five current simulation models has been conducted [BAA^{+}19], delivering this valuable information for the first time. The simulation results for different acoustic environments were physically and perceptually compared to ground truth data from measurements revealing that plausible, but not authentic auralisations could be achieved for most simulation models. The perceived differences were mostly related to timbral differences and differences in the position of the source. The round robin on auralisation also suffered from the lack of exact acoustical input as described above, though, making it very difficult to match the measured ground truth data. As some of the participating simulation models were commercial software packages, the results had to be published anonymously. This way, unfortunately no recommendations for the selection of a certain simulation algorithm can be extracted from the publication for a user of room acoustical simulation. Therefore, the perceptual appropriateness of the current simulation software may still have to be investigated further.

More sophisticated software packages are either proprietary, such as EASE [AF92], ODEON [Nay93] and CATT-Acoustic [Dal95], or closed source software such as RAVEN. Due to insufficient documentation, these software packages have to be considered 'black boxes' which because of their complexity and in combination with the insufficient state of research concerning their physical and perceptual accuracy makes them unideal research tools and inappropriate for the goal of Open Research [SGW17].

Due to the reasons discussed above, a simplified but also controllable approach has been chosen by employing the image source method [AB79] for simulation in this thesis. The exact parameters to achieve suitable room acoustical simulation for the research tasks are described in the next section 3.2.2.2. The image source method has already successfully been applied for various research studies, including evaluation of algorithms for speech recognition [PBW04] and blind source separation [IM01] or in a modified approach for the prediction of sound propagation in long enclosures [LI04].

3.2.2.2 Parameters of the image source method

To make a simple model like the image source method valid within its model-specific restrictions, parameters have to be chosen appropriately. The implementation described in this section, that has been used for generating the stimuli in the listening experiments of this thesis, has been made publicly available[11]. The model has been calculated according to [AB79] with correction of equations as given in [BEW18]. For simplicity and to reduce the number of degrees of freedom in this thesis, only rectangular rooms are simulated. By this choice, computationally costly tests for visibility of image sources and obstructions are not necessary [Bor84]. Constant reflection factors $\beta = 0.7$, 0.8 and 0.9 were chosen for all boundaries, resulting in reverberation times of 0.27, 0.39 and 0.74 s, respectively, according to Sabine's formula [Kut17]

$$T_{60} = 0.161 \cdot \frac{V}{A} \cdot \frac{\mathrm{s}}{\mathrm{m}} \tag{3.1}$$

with

$$A = \sum_i \alpha_i S_i \tag{3.2}$$

where V is the volume of the room in m^3, A the equivalent absorption area summed up over all faces of the room in m^2 and $\alpha_i = 1 - \beta_i^2$ and β_i the unit-less absorption coefficient and the reflection factor of the i-th face S_i in m^2, respectively. The chosen reflection factors cover the typically assumed range for validity of the image source method, that is only a good approximation for absorptive walls if the angle of sound incidence is not close to grazing incidence [Vor08].

---

[11] `http://doi.org/10.5281/zenodo.3745990`

Table 3.2: Schroeder frequencies calculated for the rooms simulated in the listening experiments with $V = 210\,\mathrm{m}^3$ according to eq. (3.3). $\alpha = 1 - \beta^2$ is the absorption coefficient for all boundaries.

β	α	T_{60} / s	$f_{\mathrm{Schroeder}}$ / Hz
0.7	0.51	0.27	72
0.8	0.36	0.39	86
0.9	0.19	0.74	118

To avoid simulating a case with extreme symmetry, where many reflections are coincidental, the positions of loudspeakers and receiver inside the rectangular room have been chosen slightly asymmetrical, cf. fig. 4.1 and 5.1 for the setup for the localisation and colouration experiment, respectively. Apart from the additionally varied listener position and secondary source distance for the investigation of localisation, the setups for both experiments are the same. The room size has been chosen such that the Schroeder frequency [SK62, eq. (2)]

$$f_{\mathrm{Schroeder}} = 2000 \cdot \sqrt{\frac{T_{60}}{V}} \cdot \sqrt{\frac{\mathrm{m}^3}{\mathrm{s}^3}} \tag{3.3}$$

that determines the lower frequency end of valid simulation is at maximum 118 Hz for $\beta = 0.9$, cf. table 3.2 for the Schroeder frequencies for all used reflection factors. More severe are the restrictions in validity concerning the distances of sources and receivers to the surrounding walls. These positions were chosen to keep some distance to the walls, but still constitute a realistic setup. The image source method is only valid for distances of sources and receivers to walls of several wave lengths [Vor08]. Applying this rule of thumb to the geometry in fig. 4.1 and 5.1 renders the model valid for frequencies above approx. 500 Hz. i.e. two times the wave length of the minimum distance of sources and receivers to the walls, which is 1.41 m to the ceiling.

As suggested by [BM09], the regular grid formed by the image sources was modified for image sources of order greater than 3 by varying their positions randomly according to a uniform distribution by up to ±1 m along all three axes of the Cartesian coordinate system to achieve a more naturally sounding reverberation. For all stimuli, the same positional variation was used.

It has been shown that a peak-to-noise ratio (maximum value of impulse response in relation to power of noise) of approx. 60 dB leads to a noise floor below the threshold of hearing in measured BRIRs at a sound pressure level of approx. 60 $\mathrm{dB_{SPL}}$ [HES19]. This means that all information (in this case just the noise floor) after a decay of the impulse response by 60 dB is irrelevant for an auralisation. This finding can be transferred to the case of simulated room impulse responses, where the simulation can stop after the decay of the impulse response of 60 dB. For playback levels of on average 65 $\mathrm{dB_{SPL}}$ and 70 $\mathrm{dB_{SPL}}$ for the localisation (cf. section 4.1) and colouration (cf. section 5.1) experiment, respectively, the lengths of the simulated room impulse responses were chosen to fade out until a level of at least 65 dB below the peak of the impulse responses. This is only critical for simulations with a reflection factor of $\beta = 0.9$. The reflection with the highest image source order contained in the BRIRs is of order 91. With this choice it is assumed, that the simulations contain all perceptible reverberation.

For the synthesis of BRIRs, the incidence angle of each reflection has been taken into account and the appropriate HRTFs from the FABIAN HRTF set (cf. section 3.2.1.2) have been selected to be added up with the according delay in the resulting impulse responses for the left and right ear. For direct sound and early reflections with maximum image source order of 3, the HRTFs for the desired direction were interpolated in the time domain with estimation of the TOA by cross-correlation and Delaunay triangulation for the selection of HRTFs, that were used as basis for the interpolation of the new direction. To this end, the Delaunay triangulation forms a network of triangles for the points on a sphere that correspond to the source directions encoded in the HRTFs. The triangles are chosen such that a circle running through the three corner points of a triangle is not encompassing another point on the sphere. For image source orders greater than 3, the selected HRTFs were determined by a nearest-neighbour search from the original 1° × 1° Gauss grid of the HRTF dataset to reduce computation time.

To increase the temporal accuracy of the simulation and avoid accumulation of energy in one sample when using integer samples, impulse responses were calculated during all simulation steps with a 10 times higher sampling rate of 441 kHz than the sampling rate used for the auralisation playback. In the end, the impulse responses were sampled down to 44.1 kHz. Thus, a fractional delay has been used.

3.2.3 Generating impulse responses of Wave Field Synthesis systems

The previous sections described how to acquire monaural or binaural impulse responses for single sources in free field or in a reverberant environment. As for sound pressure levels below the pain threshold of humans sound propagation can be approximated as a linear system [Mös09], the impulse responses for multiple sources can simply be superposed. If the impulse response of a WFS array is desired, the impulse response of each secondary source has to be convolved with the appropriate driving signal before superposition, cf. section 2.1.3. This yields the same results as acquiring the impulse responses from the array as a whole as long as ideal omnidirectional sources are used which has been the case in this thesis. If the synthesis is based on real loudspeakers with specific directivities, the orientation of the loudspeakers for each position in the array as well as the influence of loudspeakers that are positioned closely together on each other comes into play as well. The differences arising from these two approaches have been investigated by [Wie14] who found deviations especially for high frequencies above 7 kHz, where the frequencies of real loudspeakers arranged in a linear array were attenuated compared to the simulation.

3.3 Evaluation of the reporting method for the localisation experiment[12,13]

While some perceptual attributes are suitable to be evaluated on a simple scale between two carefully chosen endpoints, the directional localisation of a source requires a dedicated research method. The choice of the reporting method for perceived direction is crucial in a localisation experiment and it has been investigated by several studies, e.g. [LDE00, See03, BCNW16]. Existing methods range from graphical indications on maps [SWA$^+$14] to pointing with a hand-held device [FMSZ10] or with the head [Bro95] in direction of the auditory event. More technically advanced methods use tracking of eye movements [SSJW10] or provide visual feedback in virtual reality environments [MGL10, PKV14]. Best results could be achieved with methods that combine pointing with the head supported by visual feedback of the head direction as it reduces errors induced by interaction with the locomotor system, which can lead to undershoots of reported directions for laterally positioned sources [LDE00]. This approach has been adopted in this thesis and the chosen experimental apparatus has been evaluated in a listening experiment regarding its accuracy in the frontal horizontal plane as described in the following sections. The method should exhibit an accuracy that is at least as high as the human localisation accuracy of about 1° for frontal sound incidence [Bla97]. The results of the listening experiment are compared to results from the literature and to two previous studies [WSR12a, WWS17] that conducted similar experiments.

3.3.1 Apparatus for the localisation experiment

The chosen approach combines dynamic binaural synthesis with the reporting method of pointing with the head supported by visual feedback of the head direction with a head-mounted display (HMD), similarly to the study by Majdak et al. [MGL10]. In [MGL10], static binaural synthesis was used, though, where subjects had to indicate the direction of the auditory event after presentation of the stimulus.

In this thesis, non-individual dynamic binaural synthesis has been used for localisation in the horizontal plane (cf. section 3.3.2) which has been shown to impair localisation mainly for elevated sources or cause front-back confusions if the synthesis is static [WAKW93, MSJH96, HRO98]. Moreover, shifts of a source during head rotation originating from different head widths and therefore different interaural time differences (ITDs) of dummy head and listener [AAD01] are not problematic as the ITD approaches zero when a subject directly faces a source. It was shown by [WSR12a] that non-individual dynamic binaural synthesis is a transparent tool for localisation studies by comparing human localisation performance for real loudspeakers in a room with binaural synthesis with HRTFs and BRIRs.

The rendering of the stimuli in dynamic binaural synthesis over headphones type AKG K601 was carried out by the SoundScape Renderer [GS12], commit 2b11775[14].

---

[12] Major parts of this section have been published in [EFS19, ES20], but the statistical analysis has been revisited in this thesis.

[13] The stimuli and results of the listening experiment and the revisited statistical analysis are publicly available at `http://doi.org/10.5281/zenodo.3520127`.

[14] `http://github.com/SoundScapeRenderer/ssr`

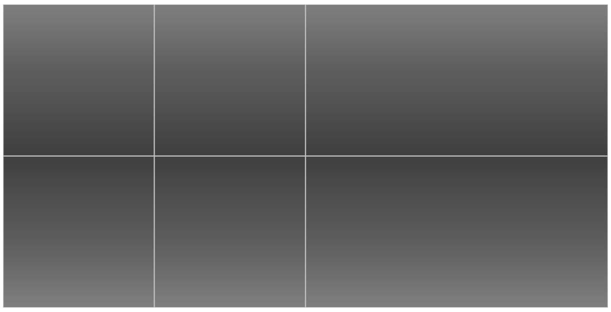


Figure 3.5: Image that is warped to a simple spherical grid as visual virtual reality (VR) presented on the HMD in the listening experiment on localisation, the frontal direction is in the centre of the right half.

Head movements in azimuthal directions were provided by a head tracker type Polhemus Patriot. The sensor of this electromagnetic tracker was attached to the top of the headphones with the source emitting a electromagnetic dipole field about 1 m behind the head of the subject. The HMD type Oculus Rift CV1 provided a visual feedback of the head direction by an orange circle in a very simple spherical grid. Fig. 3.5 shows the image that is warped to the spherical grid in the HMD. In the frontal direction only the horizon was visible except for the calibration phase at the start of the experiment where the 0° direction was marked to synchronise the independently running head tracking systems of the HMD and the electromagnetic tracker. The simplified visual environment was chosen to avoid possible anchor effects, where subjects associate the direction of an auditory event with a prominent visual mark, like e.g. additional grid lines [Rec09]. Both head tracking devices were connected to a Windows system, so the timestamps for comparison were generated on the same machine. On a Linux system, the software for executing the listening experiment received the data from the electromagnetic tracker over network and passed it on to the rendering software. The latency introduced by forwarding the data over network were negligible compared to the estimated latencies caused by other sources, cf. table 3.1. The sensor of the HMD was placed right in front of the subject.

3.3.2 Conditions

Subjects had to report the direction of the auditory events for 11 source directions ϕ_{Source} synthesised by HRTFs recorded with a KEMAR manikin type 45BA with 1° resolution in the horizontal plane. The HRTF database is described in [WGRS11] and is freely available for download[15]. Fig. 3.6 shows the chosen source positions as the black loudspeakers, that coincided with the position of real loudspeakers in [WSR12a]. Therefore, the synthesised source directions stem from measurement of the loudspeaker positions and thus differ slightly from the idealised geometry in fig. 3.6. The directions range in between approx. ±42°. For source directions in between measured HRTFs, linear interpolation was applied. The distance-induced level differences between the 11 sources were compensated for. The source content was 100 s of statistically independent white noise pulses with a pulse length 700 ms

---

[15] `http://doi.org/10.5281/zenodo.55418`

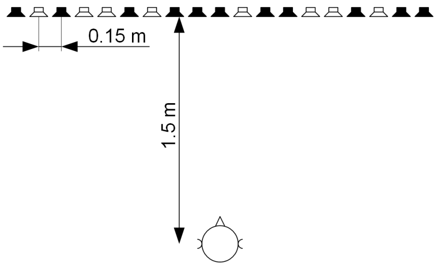


Figure 3.6: Location of virtual sources (loudspeaker symbols) and listener in the evaluation experiment for localisation, only source positions in black are used.

with cosine-shaped 20 ms fade-in/fade-out and a pause length of 300 ms played back in a loop. Each pulse was bandpass filtered with a 4th order Butterworth filter from 125 Hz–20 kHz. These stimuli were the same as in the two previous studies [WSR12a, WWS17].

3.3.3 Realisation of the listening experiment

3.3.3.1 Experimental procedure

The subjects sat on a revolving chair wearing the headphones and the HMD with a keypad in their hands. After the calibration phase for the head trackers, which included adjusting the HMD to the individual's head, subjects were instructed to point their head in the direction of an auditory event ignoring the vertical dimension. It was possible for subjects to complete this movement by both turning the head as well as turning on the revolving chair. Subjects were encouraged to perform oscillating head movements to help determining the direction. When the subjects found the direction of the auditory event, they pressed a button on the keypad and the mean of the last 5 values of the Polhemus Patriot tracking data were saved as result. After pressing the button, the next trial started. The 11 conditions had to be repeated 5 times by each subject leading to 55 trials presented in a randomised order with a preceding training of 11 trials (each condition once) also in randomised order.

Fig. 3.7 shows the setup of the experiment performed at the Audio Lab of the Institute of Communications Engineering at the University of Rostock. The Audio Lab is a small rectangular room (5 m × 5.75 m floor, 3 m in height) with several broadband absorbers at walls and ceiling, resulting in a mid-frequency reverberation time of approx. 0.3 s. Though the playback room can possibly impair the localisation performance in binaural synthesis [WK14], it was assumed that the influence of the experimental environment was negligible as subjects were wearing the HMD during the experiment.

3.3.3.2 Participants

10 subjects with an average age of 33 years participated in the listening experiment. 6 had home or professional experience in the field of audio, 8 had participated in listening experiments before. These included experiments involving non-individual

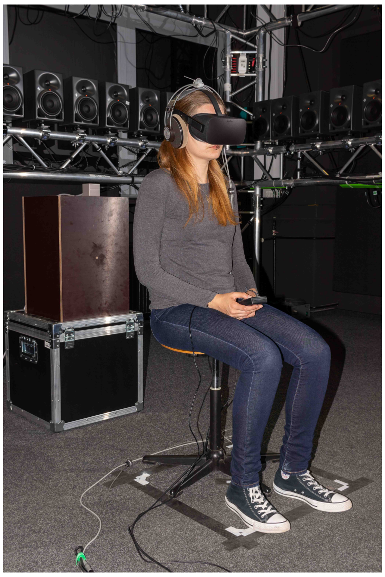


Figure 3.7: Listener in the experimental setup for the localisation experiment.

binaural synthesis, but the test subjects could not be considered specifically trained for the used HRTF dataset, which could have increased their localisation performance [HRO98].

3.3.4 Procedure for data analysis

The necessary steps for data correction are described in the following section 3.3.4.1. As accuracy of the method the absolute maximum of the mean localisation error over the 11 conditions is determined. Additionally, it is investigated if the signed localisation error depends on the azimuthal source directions with a linear mixed-effects model. The according data analysis is described in section 3.3.4.2. Statistics on comparing the standard deviations of the evaluation experiment and of the previous studies from [WSR12a] and [WWS17] are reviewed in section 3.3.4.3.

3.3.4.1 Data correction

Despite the calibration procedure, there is still a cause for possible lateral bias present in the apparatus of the reporting method: Depending on the positioning of the HMD on the head of the subject, the visual feedback cursor indicating the head direction of the subject can be shifted. This leads to a bias to the left or right in the reported directions as subjects were instructed to point with the feedback cursor in direction of the auditory event. Therefore, the reported directions from the localisation experiment were corrected by subtracting the mean localisation errors per subject over all conditions from all reported directions. This is justified as it has been shown that accurate localisation of HRTF-based stimuli just as for real sound sources can be expected [WSR12a].

3.3.4.2 Statistical model for data analysis

To analyse the dependency of the signed localisation error on azimuthal source direction, a linear mixed-effects model with multiple levels was fitted to the data [FBK14, QvdB04]. This model can account for both fixed and random effects and assumes a linear relationship between effects and outcome variables. The outcome variables predicted by the model in the evaluation experiment is the localisation error as difference between the direction of a synthesised source and the reported auditory event. The fixed effect that is tested as predictor for the localisation error is the azimuth of the direction of a synthesised source.

The repeated-measures design of the listening experiment with 10 subjects reporting answers for all 11 conditions 5 times constitutes a multi-level framework where repetitions are nested in subjects. To avoid over-parametrisation of the model, only the intercepts of the random effects are modelled allowing for offset differences in localisation results between subjects and within repetitions of one subject. The random effects are assumed to be normally distributed and independent from another. A general correlation structure for the covariance matrix is chosen, because the order of presentation of the conditions was randomised. Hence, there are no strong relationships between reported answers of a subject to be expected as e.g. in longitudinal studies. The calculations for the mixed-effects models were performed in R [R C17] with restricted maximum likelihood estimation (REML).

In the so-called null model without any predictor variables, the response Y_{ij} (e.g. the signed localisation error, cf. section 3.3.5.2) for trial i ($i = 1, ..., I_j$) within subject j ($j = 1, ..., J$) on level 1 of the model can be regarded as a deviation from the mean B_j of the jth subject, i.e.

$$Y_{ij} = B_j + e_{ij}. \tag{3.4}$$

The residuals e_{ij} are assumed to be normally distributed with zero mean and variance $\sigma^2_{e_{ij}}$. The mean response B_j for subject j can be regarded as a deviation from the grand mean γ_{00} on level 2, i.e.

$$B_j = \gamma_{00} + u_{0j}. \tag{3.5}$$

The residuals u_{0j} are also assumed to be normally distributed with zero mean and variance $\sigma^2_{u_{0j}}$. In the special case of the evaluation experiment for localisation, $u_{0j} = 0$ due to the data correction in section 3.3.4.1. Level 1 and level 2 residuals, e_{ij} and u_{0j}, respectively, are assumed to be uncorrelated. Substitution of eq. (3.5) in eq. (3.4) yields the multi-level null model

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{3.6}$$

With an additional fixed effect, the models reads

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{\text{predictor}} + u_{0j} + e_{ij} \tag{3.7}$$

with γ_{10} the fixed effect slope and $x_{\text{predictor}}$ the independent predictor variable for the fixed effect.

A fixed effect is only included in the model if a better fit is reached according to the Bayesian information criterion (BIC) that assesses goodness of fit while penalising

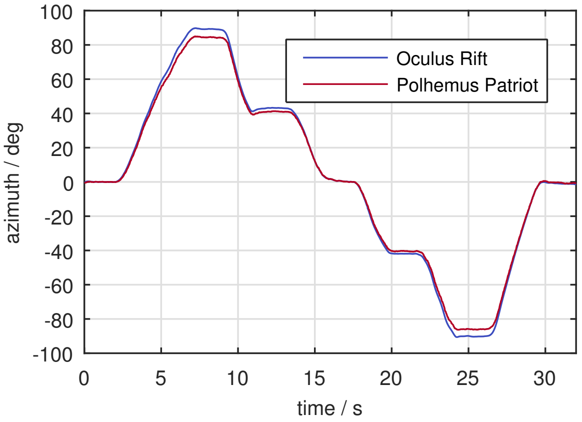


Figure 3.8: Comparison of azimuthal data of the two utilised tracking systems in the evaluation experiment for localisation during rotation of the head of the listener from 90° to −90° azimuth, resting in between.

the number of estimated parameters. The BIC was chosen over the also widely used Akaike information criterion (AIC) as the AIC tends to favour overly complex models [MSW13]. A fixed effect is included in a model if the BIC value for this model is smaller than for the model tested without this fixed effect.

3.3.4.3 Comparison of standard deviations

To compare more than two standard deviations, a Bartlett test on homogeneity of variances is performed to determine if a difference exists [Bar37]. A post-hoc pairwise F-test with Bonferroni correction for multiple testing can then provide the information which standard deviations are different.

3.3.5 Results

3.3.5.1 Comparison of the two tracking systems

Fig. 3.8 compares the data of the two independently running tracking systems with an exemplary movement of a human head ranging from approx. +90° to −90° azimuth while resting in between, similar to a movement in the listening experiment. There exist slight differences between the tracking systems that become larger for lateral head directions. The Oculus Rift tracking data appears to be a bit smoother presumably caused by the incorporated forecast of the trajectory, which can also lead to slight overshoots for more abrupt movements. The differences between the two systems in the azimuth range of the presented virtual sources between ±42° do not exceed 2°, which was tolerated in this listening experiment.

3.3.5.2 Localisation error

The localisation error is defined as difference between the directions of the synthesised sources and the reported auditory events of the subjects. Fig. 3.9 shows an overview of the results after data correction as described in section 3.3.4.1 in comparison to the previous studies by [WSR12a, WWS17]. These studies used exactly the

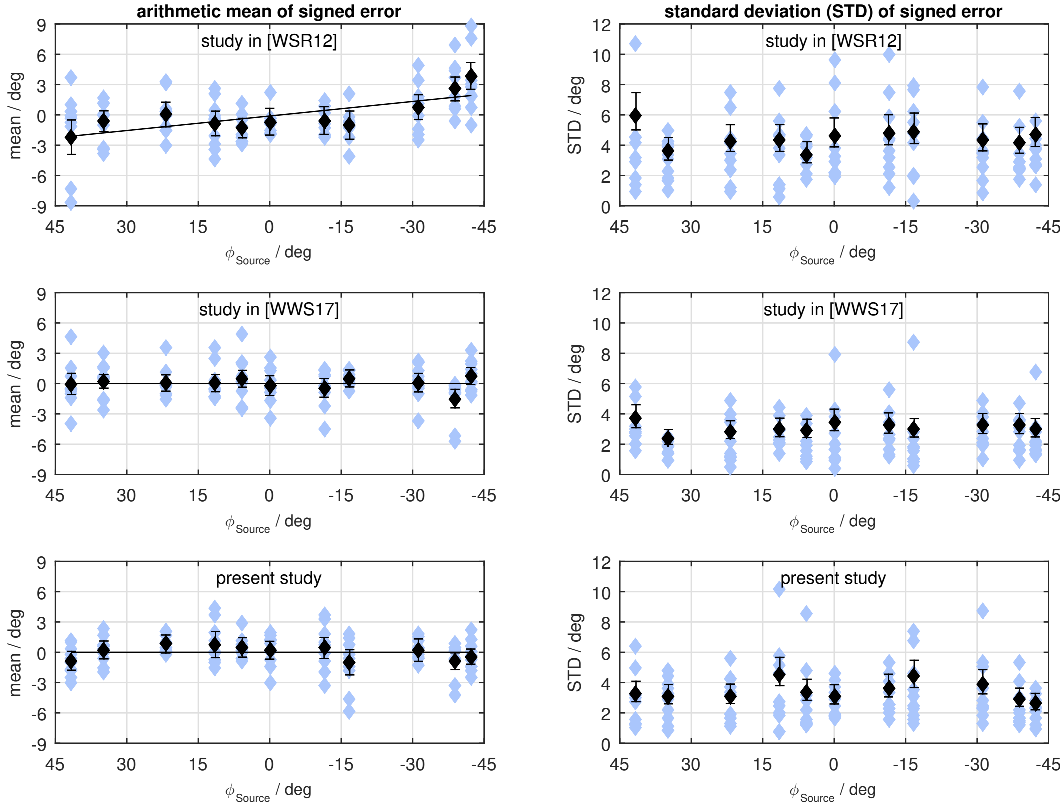


Figure 3.9: The rows show the arithmetic mean (left column) and standard deviation (right column) of the signed localisation error for the three compared studies on evaluation of localisation. Blue markers show results of individual subjects (10 subjects per study), black markers show results for all subjects together with the 95% confidence interval. The black lines in the left column represent the mixed models with the best fit for the signed localisation errors.

same stimuli and experimental design, but different head tracking devices (Polhemus Fastrak in [WSR12a] and NaturalPoint OptiTrack in [WWS17]) and especially a different way of providing the visual feedback: Both [WSR12a] and [WWS17] used a laser pointer attached to the headphones on the subjects' heads projecting on a curtain. While [WSR12a] used a straight curtain, [WWS17] used a circular curtain around the listener. Both previous studies have been submitted to the same data correction as the results from the listening experiment in this thesis as it was not possible to mount the laser pointer on the headphones so that it pointed exactly in the frontal direction. This can introduce a similar bias as is caused by differing positions of the HMD on the head of the subjects. Both previous studies were performed with 11 subjects, but excluded one subject each from the analysis as the subject exhibited a twice as high standard deviation compared to the maximum standard deviation of the other participants. Thus, all three compared studies have evaluable results from 10 subjects.

The left column of fig. 3.9 shows the individual mean signed localisation error of the 10 subjects as blue markers and the overall mean localisation errors as black markers together with the 95% confidence intervals depending on the direction of the

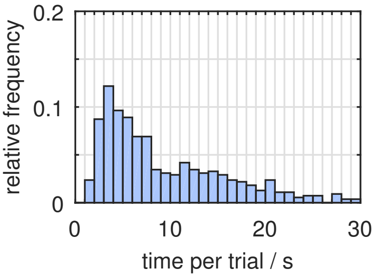


Figure 3.10: Normalised histogram for the elapsed time per trial in the evaluation experiment for localisation.

synthesised sources. The evaluation experiment in this thesis shows with means not exceeding ±1.0° good results compared to the previous studies (±3.9° for [WSR12a], ±1.5° for [WWS17]). To evaluate the localisation error depending on condition, a mixed model as described in section 3.3.4.2 was fitted to the data. According to the BIC, the direction of the synthesised source is only a suitable predictor for the results from [WSR12a], the models for the two other studies are only consisting of the intercept expressing the mean localisation error over all directions, which is 0° in both studies due to the data correction from section 3.3.4.1. The model outcomes are represented as black lines in the left column of fig. 3.9. This shows that the apparatus from [WSR12a] is provoking undershoots of the reported answers for more lateral source directions while the present study and [WWS17] do not suffer from this limitation.

The right column of fig. 3.9 shows the individual standard deviations of the signed localisation error as blue markers and the overall standard deviations together with the 95% confidence intervals as black markers depending on the direction of the synthesised sources. The overall standard deviations of the three studies are $s_{\text{[WSR12a]}} = 4.8°$, $s_{\text{[WWS17]}} = 3.1°$ and $s_{\text{present study}} = 3.5°$. To compare the overall standard deviations of the three studies, a Bartlett test on homogeneity of variances is performed. The test confirms differences between the studies ($\alpha = 0.05$) and a follow-up pairwise F-test with Bonferroni correction for multiple testing on the overall variances reveals the following ranking of the standard deviations with a confidence level of 95%:

$$s_{\text{[WSR12a]}} > s_{\text{present study}} > s_{\text{[WWS17]}}. \tag{3.8}$$

3.3.5.3 Elapsed time per trial

Subjects needed on average 9 minutes to complete the whole experiment (without training), minimum and maximum durations ranged from about 3.5 up to 20 minutes. The elapsed time per trial is not normally distributed as the histogram in fig. 3.10 indicates. Table 3.3 shows a descriptive statistic compared to the results from the previous studies. The elapsed time per trial was shorter for the study by [WSR12a] than for the other two studies that showed a median of about 7 s.

Table 3.3: Descriptive statistics on elapsed time per trial in s in the evaluation experiment for localisation compared to previous studies.

Elapsed time per trial	Study by [WSR12a]	Study by [WWS17]	Present study
arithmetic mean	5.6	9.0	9.8
5th percentile	2.2	2.6	2.3
median	4.6	6.8	7.1
95th percentile	13.6	24.2	24.7

3.3.6 Discussion

Humans are able to localise real broadband sound sources with an accuracy of up to 1° in the frontal horizontal plane [Bla97]. For virtual sources based on non-individual HRTFs, this accuracy can be degraded [See03]. A reporting method for the localisation of (virtual) auditory events should be at least as accurate as human localisation, which is the case for the presented apparatus with a maximum localisation error of 1.0°. This makes it suitable for localisation experiments in the frontal horizontal plane. Also the standard deviation appears to be acceptable compared to the previous studies and to the averaged median-to-quartile distance of 2.9° in [See03] achieved for non-individual, but pre-selected HRTFs with the Proprioception Decoupled Pointer method (9 test subjects).

The undershoots in [WSR12a] are suspected to be caused by the straight curtain where even in a darkened room the limits of the projection plane were visible, seducing subjects to localise sources more towards the centre of the curtain as has been pointed out by [WWS17]. In contrast, the projection planes of [WWS17] and the experiment in this thesis are rotationally invariant.

The independently working tracking systems exhibited deviations for more lateral directions as shown in fig. 3.8. Though this could have been a source of error, there is no dependence of the mean localisation error on the source direction. This appears also not to be the case for the standard deviation, cf. fig. 3.9.

The performance of the individual test subjects differed considerably in time devoted to solving the task and in standard deviations of the localisation error. There seemed to be no obvious relation between these two observations, though. Also, the performance of the subjects did not seem to depend on previous experience in listening experiments or the field of audio or on age of the subjects. Possibly, a longer training phase as investigated by [MGL10] is necessary to familiarise all subjects with the unusual task in a virtual environment with an HMD. The unfamiliar task could also be the reason for the medium standard deviation compared to the two previous studies.

It has to be noted, that the data correction from section 3.3.4.1 is not only compensating the bias caused by inaccurate positioning of the laser pointer, but also any other source of bias, e.g. a test subject with a slight hearing loss on one ear, a randomly occurring bias or bias due to non-individual HRTFs, thus decreasing the measured standard deviation.

3.3.7 Summary

The presented reporting method for localisation tasks has been shown to be accurate enough for localisation experiments with virtual sources exhibiting a maximum mean localisation error in the frontal horizontal plane of 1.0°. Varying performances of individual subjects suggest a need for an intensified training phase. The method can also be applied for experiments with elevated sources, but an investigation of the optimal visual environment with a trade-off between orientation marks and potential visual anchor effects should be performed prior to that.

3.4 Conducting colouration experiments

The investigation of colouration in listening experiments is a challenging task, as colouration is a multidimensional perceptual concept. Additionally, the use of binaural synthesis as reproduction method for the stimuli is apt to introduce colouration itself. The following sections therefore discuss the applicability of binaural synthesis for listening experiments on colouration (section 3.4.1) and a suitable experimental design for evaluation of colouration (section 3.4.2).

3.4.1 Binaural synthesis in colouration experiments

As discussed in section 3.2.1.1, binaural synthesis is not a completely transparent reproduction method, not even for individual HRTFs [BLW17a]. This holds especially true for the preservation of timbre. [Wie14] argues, though, that evaluation of colouration in listening experiments with binaural synthesis is possible as the colouration introduced by the reproduction method influences all systems under test in the same way. He also measured the spectral differences in the ear signals of a listener between the reproduction with real loudspeakers and their binaural representation for different loudspeaker array systems synthesising a virtual point source with WFS. Results show that the linear distortion introduced by binaural synthesis that leads to colouration is independent of the array setup. Still, deviations of up to 15 dB have been found that can possibly mask or emphasise timbral differences between systems, as has also been discussed by [Win19]. As these deviations are dependent on the individual listener, it is assumed in this thesis, that their influence averages out in a listening experiment with a higher number of test subjects (34 subjects have participated in the listening experiment on colouration in this thesis, cf. section 5.2.3) if the investigated colouration differences are not too small.

Additional confirmation that binaural synthesis can be employed as a simulation tool in listening experiments can be found in a study by [OWM07] comparing the use of binaural synthesis in a listening experiment to the presentation of stimuli with real loudspeakers. It was shown that preference ratings for different loudspeakers models are independent from the presentation mode. Furthermore, [Wit07, Wie14, WWH$^+$18] all found plausible results in their colouration experiments using binaural synthesis for comparison of different SFS systems.

As discussed in section 3.2, the necessary direct comparisons of stimuli in a listening experiment on colouration with WFS systems in different listening rooms are not possible in situ and have to be performed with simulation and representation of stimuli by binaural synthesis. On basis of the above mentioned studies, it is assumed that this way at least relative evaluation of colouration is feasible if not absolute judgements. Furthermore, as dynamic binaural synthesis based on non-individual HRTFs can cause changes in colouration depending on the head direction, only static binaural synthesis was used in the colouration experiment described in chapter 5. A drawback of the static simulation is that the interesting connection of WFS in a listening room with the phenomenon of binaural decolouration (cf. section 2.2.3) cannot be investigated.

3.4.2 Experimental design for the investigation of colouration

The investigation of a multidimensional perceptual concept such as colouration in a listening experiment is challenging and requires a suitable experimental design. An often chosen style of experiment is a modified Multiple Stimulus Test with Hidden Reference and Anchor (MUSHRA) [ITU15] design, in the following just referred to as MUSHRA test, as employed by [Wit07, Wie14, WWH^{+}18] which is also used in this thesis. A MUSHRA test is originally intended to evaluate the quality of audio systems with an intermediate audio quality. Several stimuli representing the different audio systems are to be compared simultaneously by test subjects to a reference stimulus on a scale from ‘bad’ to ‘excellent’ with additional subdivisions of the scale that are also labelled. In the pool of stimuli that are to be compared, the hidden reference and a low and – since the version of the ITU recommendation of 2014 – mid-range anchor are also included. The anchors should exhibit similar artefacts as the audio systems under test. The hidden reference and the anchors serve to set the limits for the rating scale to enable pooling of the responses of the test subjects.

In the modified MUSHRA design in this thesis, colouration instead of quality was to be rated on a scale ranging from ‘no difference’ to ‘large difference’ without subdivisions of the scale, cf. section 5.2.1. Only a low and no mid-range anchor was included. The MUSHRA style design appears to be suitable as quality is also an inherently multidimensional percept just as colouration. Still, there are limitations of this design and a careful application of suitable statistical tests is needed. This has been discussed in recent publications [ZHHR07, SLS09, MDM18] and specific challenges of the MUSHRA design have been revealed in a preliminary run of the experiment in this thesis. In this experiment, 33 test subjects had to compare the timbre of virtual point sources synthesised by WFS with a discrete secondary source array in free field and in different listening rooms as well as real point sources in different listening rooms to a real point source in free field. The high-pass (corner frequency 2 kHz, 2nd order Butterworth) filtered reference stimulus resembling the timbre evoked by spatial aliasing artefacts served as low anchor. While the hidden reference was reliably detected and the anchor with high colouration compared to the reference was rated as such by the subjects, the ratings for the remaining stimuli differed strongly between subjects. Therefore, the experiment was redesigned slightly, reducing sources for this wide spread of ratings, that generate noise in the acquired data, and avoiding the need for a very large number of test subjects. This resulted in the experiment that is reported in detail in chapter 5. These sources for a wide spread of ratings were identified as follows:

- Subjects have difficulties to compare too many stimuli with one another, especially if the percept that is to be evaluated is multidimensional in nature. Therefore, the number of stimuli was reduced from nine to six (including hidden reference and anchor) in the second run of the experiment.

- Subjects use scales differently [SLS09] and might have individual perceptions of colouration. These sources for differences in ratings are hard to separate. While the first one is undesired, the second one is part of the research question. The complex task might necessitate the use of so-called expert listeners, but the definition for these is lacking. To at least control for inconsistencies in the ratings of one subject, the MUSHRA runs were conducted twice per subject in the second experiment and subjects that did not reach the test-retest reliability found by [MDM18] for a MUSHRA test were excluded from the subsequent analysis, cf. section 5.3.
- The task might be interpreted differently by subjects. In the first experiment, subjects were instructed to rate the overall colouration of the direct sound component and potential reverberation. This might have led to different decisions between subjects concerning the distinction of 'different timbre' and 'stronger/longer reverberation'. This presumption is supported by the high number of subjects reporting this decision to be difficult for them after the experiment. In the second experiment, subjects were instructed more clearly to rate the colouration of the direct sound component alone and ignore potential reverberation and its timbre.

Furthermore, [ZHHR07] found that the results of different MUSHRA runs cannot be directly compared as the rating of a stimulus is relative to the other stimuli presented within one run of the experiment even if the same anchors are presented in the runs. Therefore, only the rating results for stimuli presented in the same run are meaningful relative to each other and different MUSHRA runs cannot be pooled.

For the statistical analysis of a MUSHRA test, [ITU15] proposes the analysis of variance (ANOVA). According to [MDM18], this conventional statistic test method is not suitable for MUSHRA data as it is prone to violate all assumptions of an ANOVA including normality of data, homogeneity of variances and independence of observations. While there exist corrections for the ANOVA when single assumptions are violated, these corrections are not approved for the violation of several assumptions at once. As the observations in the conducted experiment are not independent, only a repeated measures ANOVA could be eligible. However, two of its necessary assumptions – the normality of data and sphericity – are violated for the data collected in the experiment in this thesis. The normality of data has been tested with a Kolmogorov-Smirnov test [Mas51]. For the test on sphericity, the Mauchly test [Mau40] has been applied. Both tests have been documented and made publicly available together with the experimental results[16]. As a consequence, [MDM18] recommend non-parametric statistical tests, in particular the Wilcoxon signed rank test [Wil45] that has also been applied in this thesis, cf. section 5.2.1 and 5.4.1. Additionally, a Friedman test [Fri37] followed by a Conover post-hoc test [Con99] has been performed to investigate the data further.

---

[16] `http://doi.org/10.5281/zenodo.4036228`

4 Spatial perception of Wave Field Synthesis in a listening room

Reflections from a listening room impose changes on the sound field synthesised by Wave Field Synthesis as the theoretic assumption of an anechoic environment for reproduction is violated. This chapter investigates if these changes of the physical sound field also influence the perception of the direction of virtual sources in a listening experiment with the apparatus evaluated in section 3.3. Parts of this chapter have been published in [ES20]. The stimuli and results of the listening experiment and the statistical analysis are publicly available[1]. Only the aspect of perception of direction, not distance, is explored. As this thesis focuses on the most common case of 2.5D synthesis only azimuthal direction is investigated. In the following 'azimuthal directional localisation' is just referred to as 'localisation'. Restrictions had also to be made regarding the degrees of freedom of the topic at hand: Only a linear secondary source array in a rectangular room was considered with a point source as desired sound field. Wall properties, listener position and the distance between secondary sources were varied in a full-factorial design.

This thesis extends the work of [Wie14] on localisation in WFS in free field to the case of reflective environments and explicitly studies the following three different aspects of localisation:

- The direction of the localisation error: Which of the varied degrees of freedom causes a localisation error in a specific direction?
- The accuracy of localisation: Which of the varied degrees of freedom have an influence on the certainty of a perceived direction (or the so-called localisation blur)?
- The difficulty of localisation: Which of the varied degrees of freedom have an influence on how difficult the localisation of a source is?

As discussed by [Win19], there exists no uniformly used measure for the accuracy of answers in a localisation task. Primarily, the two terms 'minimum audible angle' (MAA) and 'localisation blur' have been established. The MAA has been defined as the smallest perceivable difference between the azimuth angles of two sources, that are otherwise identical [Mil58]. Blauert defines this as the localisation blur [Bla97]. In contrast, other researchers used the term localisation blur for measures that quantify the spread of perceived directions around the mean perceived direction and are related to the standard deviation, e.g. [Ver97, Wie14, SFTW^{+}19]. The different concepts of measuring the accuracy in a localisation task can lead to similar results [Har83], but cannot be understood as equivalent. In this thesis, the accuracy of localisation is quantified as the absolute deviations from the mean

---

[1] `http://doi.org/10.5281/zenodo.3358956`

perceived direction, which is related to the standard deviation. This measure fits in the same framework for statistical analysis as it is used for the other investigated aspects of localisation, i.e. the direction of the localisation error and the difficulty of localisation.

In different areas of research, the response time in experiments is taken as an indication of how difficult a task is for the test subjects. [BCNW16] compared different pointing methods for localisation of auditory events in 3D space in listening experiments. The response time of subjects was measured to conclude on difficulties related to the human motor control. The reaction time is also used in research on audio-visual perception, that found a faster identification of visual stimuli through coincident auditory stimuli. This decrease in difficulty of the task is called the 'intersensory facilitation effect' [CA01, See03]. Therefore, to investigate the difficulty of localisation of real and virtual sources in a listening room in this thesis, the response time or elapsed time per trial has been evaluated.

For the three investigated aspects of localisation listed above, the influence of the varied degrees of freedom (wall properties, listener position and secondary source distance) was examined. Section 4.1 describes the conditions of the listening experiment, followed by an account of the realisation of the experiment in section 4.2. The employed statistical analysis is described in section 4.3. The results in section 4.4 are subdivided according to the three key research aspects listed above. Section 4.5 discusses the findings of the listening experiment and the relevant auditory mechanisms. The chapter closes with the summary in section 4.6.

4.1 Conditions

The azimuthal localisation of virtual point sources synthesised by linear WFS arrays with three different secondary source distances were compared to the localisation of real point sources. The geometry of the setup is given in fig. 4.1. For free field conditions, no room is present. For conditions with a real source, no secondary sources are present. For conditions with virtual sources synthesised by WFS, 15, 8 and 3 secondary sources were equidistantly spaced over an array length of 2.85 m, resulting in distances of approx. $\Delta x_0 = 20$, 41 and 143 cm between the sources. These discretisations of the secondary source distribution were chosen to allow for the comparison of the results with the findings from [Wie14] who investigated the localisation properties of WFS for these array setups in free field. Each of these cases was set up in free field and in a rectangular room with three different wall properties simulated by the image source method with constant reflection factors $\beta = 0.7$, 0.8 and 0.9 for all boundaries, resulting in reverberation times of 0.27, 0.39 and 0.74 s, respectively, according to Sabine's formula eq. (3.1). Details for the room acoustical simulation are given in section 3.2.2.2. With three listener positions for each case, this results in 48 conditions in the listening experiment.

For WFS reproduction, the driving function given in [Ver97, eq. 2.22a] with a point as reference location was used. The reference point was chosen to be the listener position $\mathbf{x}_{\mathrm{L3}}$ for all conditions. The upper corner frequency of the pre-equalisation filter was set to the aliasing frequency in Hz, which is the lower limit for the occurrence of spatial aliasing in the sound field. It was calculated according to eq. (2.21) resulting in $f_{\mathrm{alias}} = 842$, 421 and 120 Hz for $\Delta x_0 = 20$, 41 and 143 cm,

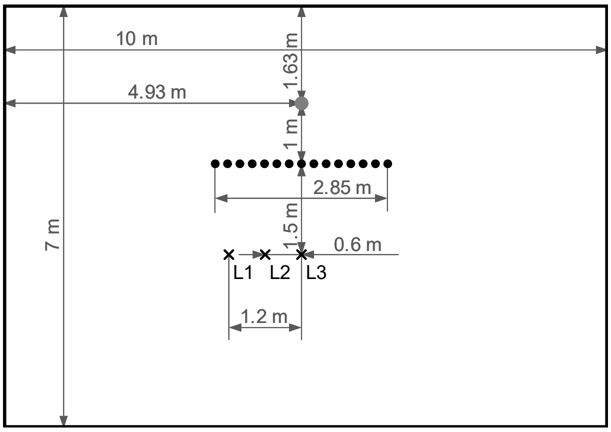


Figure 4.1: Geometry of sources and receivers in listening room (black border) in the localisation listening experiment. The grey dot indicates the position of the real or virtual source, the black dots the secondary sources for WFS reproduction. The secondary source distances depending on the used number of secondary sources in a condition are: approx. 20 cm for all 15 secondary sources, 41 cm for 8 sources and 143 cm for 3 sources. The black crosses mark the listener positions labelled L1 to L3. Room height is 3 m. All sources and receivers are positioned 1.59 m above the floor. For free field conditions, no room is present. For conditions with a real source, no secondary sources are present. Figure taken from [ES20], © 2020 IEEE.

respectively. The lower corner frequency of the pre-equalisation filter was adapted to the array length and set to 50 Hz. For the secondary source arrays with $\Delta x_0 = 20\,\text{cm}$ and 41 cm, a raised-cosine tapering window over 30% of the array length was used to reduce diffraction artefacts from the edges of the secondary source array.

The stimuli were simulated with dynamic binaural synthesis as described in section 3.2.1.2. As the present study focuses on azimuthal localisation in 2.5D synthesis, only head movements in the horizontal plane were tracked. The binaural impulse responses with an azimuthal resolution of 1° were generated with the help of the Sound Field Synthesis Toolbox [WS12], release 2.5.0[2]. As the room simulations were time-consuming, the impulse responses for the back semicircle, where no virtual sources were placed and due to dynamic binaural synthesis no front-back confusion were to be expected [Wal40, WK99], were only calculated in 5° steps.

As audio content, 100 s of statistically independent Gaussian white noise pulses with a duration of 700 ms per pulse followed by a pause of 300 ms played back in a loop were used. Each pulse was windowed with a half-sided Hann window of 20 ms length at the start and end. The signal was bandpass filtered with a 4th order Butterworth filter between 125 Hz and 20 kHz. This noise signal has also been used in the evaluation experiment in section 3.3.

To avoid both loudness differences between free field and room conditions as well as correlation of loudness with listener position, a loudness model [ANS07, MGB97] was used to adjust the loudness of all conditions. For the loudness estimation, white noise was used. The sound level at the ear of the listeners was determined with an ear coupler G.R.A.S. RA0039 Ear Simulator IEC 60318.1 according to the standard [IEC09]. The average sound level for all conditions was $65\,\text{dB}_\text{SPL}$.

---

[2] `http://doi.org/10.5281/zenodo.2597212`

4.2 Realisation of the listening experiment

4.2.1 Experimental procedure

The reporting method for the azimuthal localisation of a source described and evaluated in section 3.3 has been used in the listening experiment. The method has been tested to exhibit a localisation accuracy of about 1° for sources in the frontal horizontal plane for free field sources simulated with dynamic binaural synthesis and is thus considered suitable for the present localisation study. Subjects sat on a revolving chair wearing headphones type AKG K601 and the HMD type Oculus Rift CV1 with a keypad in their hands. Stimuli were rendered with the SoundScape Renderer [GS12], commit 2b11775[3]. After a calibration phase, subjects were instructed to point with the visual feedback cursor in the direction of an auditory event ignoring the vertical dimension. It was possible for subjects to complete this movement by both turning the head as well as turning on the revolving chair. Subjects were encouraged to perform oscillating head movements to help determining the direction. It was also recommended to subjects to perform larger head movements if subsequent stimuli appeared to come from the same or a very similar direction. When the subjects found the direction of the auditory event, they pressed a button on the keypad and the mean of the last 5 values of the audio tracking data (from the head tracking device Polhemus Patriot) were saved as result. After pressing the button, the next trial started. The 48 conditions had to be repeated 5 times by each subject leading to 240 trials presented in randomised order with a preceding training of 20 trials also in randomised order. The training conditions were chosen as being representative for the whole range of conditions. The number of trials in the training has been doubled compared to the evaluation experiment in section 3.3 to give subjects a better opportunity to familiarise with the task. To introduce a larger variety of directions, a randomised offset was added to the head tracking for audio rendering, turning the scene around the listener. To avoid fatigue, the experiment was split in two sessions with half of the trials each. Subjects were encouraged to stand up and take a break in between sessions.

4.2.2 Participants

10 subjects with an average age of 33 years participated in the experiment. All subjects self-reported normal hearing. 4 had home or professional experience in the field of audio. All but one had participated in at least one listening experiment before. As for the evaluation experiment in section 3.3, these included experiments involving non-individual binaural synthesis, but the test subjects could not be considered specifically trained for the used HRTF dataset, which could have increased their localisation performance [HRO98].

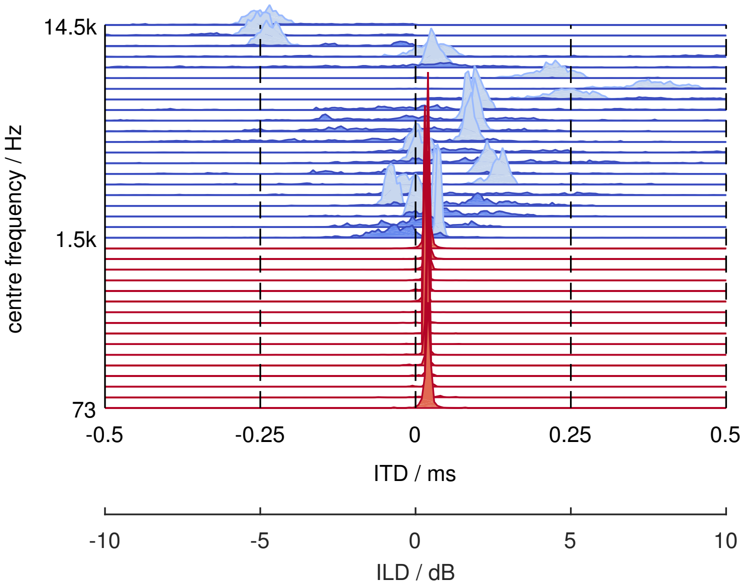


Figure 4.2: IC-weighted histograms of the ITD (red) for frequencies below 1.4 kHz and the ILD (light blue) and ITD of the envelope (dark blue) above this threshold for an auralisation of the 0° azimuth/0° elevation FABIAN HRTFs with white noise.

4.3 Procedure for data analysis

4.3.1 Data correction

As described in section 3.3, there remains a possibility for lateral bias in the apparatus of the reporting method depending on the positioning of the HMD on the head of the subject. This bias has to be corrected before proceeding with the data analysis. Additionally, analysis of the perceived directions in the listening experiment revealed an overall bias to perceive sources of approx. 2–3° farther to the right even for sources that were only composed of a single pair of HRTFs. This bias can be attributed to the used HRTF dataset of the FABIAN dummy head, which is not symmetric with respect to the 0° azimuth source direction as it is a cast of a real human head. Thus, the ear axis is shifted and twisted and not exactly parallel/orthogonal to the anatomical planes of the head. For this dummy head, the frontal viewing direction during measurements cannot be aligned by establishing an ITD of zero for sources in the median plane. Moreover, the setup for the measurement of the HRTFs with two loudspeakers on two rotatable arcs of the Oldenburg Two Arc Source Positioning (TASP) system [Ott01] is not perfectly symmetrical either. This leads to additional inaccuracies that are difficult to correct when using an asymmetric dummy head [BLW^{+}13]. Though the TOAs of the HRTFs have been fitted to BEM simulations of the HRTFs of the FABIAN dummy head that do not suffer from these inaccuracies of the measurement setup, there still remains a bias to perceive sources farther from the right as can be observed in fig. 4.2. In this figure, the binaural cues relevant for localisation of a source have been extracted from the binaural signals. The ITD for frequencies below 1.4 kHz and the interaural level

---

[3] `http://github.com/SoundScapeRenderer/ssr`

difference (ILD) and ITD of the envelope above this threshold were calculated with the localisation model from Dietz et al. [DEH11] as implemented in the Auditory Modeling Toolbox[4] [SM13], release 0.9.9, for an auralisation of the 0° azimuth/0° elevation HRTFs with white noise. The binaural cues were calculated sample-wise over time and are displayed as histograms for each auditory frequency band from 50 Hz–16 kHz. For the histogram, each sample was weighted by its corresponding interaural coherence (IC) value to account for its reliability [FM04]. While the binaural cues at high frequencies do not convey perfectly consistent information, the ITD histograms at low frequencies are not centred around 0 ms, but shifted to 0.02 ms. As the perception of direction for broadband signals is dominated by the ITD below 1.4 kHz [WK92], this indicates a localisation shift to the right.

To account for these two types of bias, the reported directions from the localisation experiment were corrected by subtracting the mean localisation error for the three purely HRTF-based stimuli from all reported directions for each subject. These stimuli could be used as calibration data as it has been shown that accurate localisation by subjects just as for real sound sources can be expected [WSR12a].

4.3.2 Statistical model for data analysis

To analyse differences in localisation results between conditions, three linear mixed-effects models with multiple levels as described in section 3.3.4.2 were fitted to the data. In contrast to the evaluation experiment described in section 3.3, the residuals per subjects u_{0j} are not equal to zero as the data correction was only performed on the basis of trials with a real point source. The outcome variables predicted by the models are either

- the signed localisation error (section 4.4.1),
- the absolute deviation from the mean perceived direction as a measure for the accuracy of localisation (section 4.4.2) or
- the elapsed time per trial until an answer was reported by a subject as a measure of how difficult localisation for a certain condition is (sec. 4.4.3).

There are three fixed effects that could possibly serve to predict the outcome variables, namely

- ‘room’: amount of reverberation from free field with reflection factor $\beta = 0$ to rooms with reflection factors $\beta = 0.7$, 0.8 and 0.9,
- ‘method’: real source corresponding to $\Delta x_0 = 0\,\mathrm{cm}$ and WFS with $\Delta x_0 = 20\,\mathrm{cm}$, 41 cm and 143 cm and
- ‘position’: listener positions from in front of the source with $x_{\mathrm{L3}} = 0\,\mathrm{m}$ to $x_{\mathrm{L2}} = 0.6\,\mathrm{m}$ and $x_{\mathrm{L1}} = 1.2\,\mathrm{m}$ shifted to the left side.

The three two-way interactions ‘room * method’, ‘room * position’ and ‘method * position’ and the three-way interaction ‘room * method * position’ of the fixed effects are considered in the models as well.

---

[4] `http://amtoolbox.sourceforge.net`

The repeated-measures design of the listening experiment with 10 subjects reporting answers for all 48 conditions 5 times constitutes a multi-level framework where repetitions are nested in subjects. A stepwise procedure was performed to choose the model with the best fit according to the BIC while adding fixed effects one at a time to test for their adequacies to predict an outcome variable. The range of conditions in the listening experiment was chosen to, first, replicate the findings of [Wie14], that for WFS reproduction in free field with large secondary source distances localisation is dominated by the nearest secondary source. In a second consideration, these findings should be complemented for the case of reflective environments. Therefore, the first fixed effect added to the mixed-models was the interaction 'method * position', followed by 'room' and subsequently the other main fixed effects and interactions.

For all fixed effects that were included in a model according to the BIC, the slope coefficients and the according p-values are reported together with the effect sizes f^2 [Lor18]. The effect size f^2 is a measure for the explained variance of a certain effect in relation to the unexplained variance in the model [AW91]. According to [Coh92], effect sizes of 0.02, 0.15 and 0.35 can be considered as small, medium and large, respectively. No test power is given, as validated methods for estimating test power are so far only available for a limited class of mixed-effects models [Sni05, GLGM13].

4.4 Results

The average perceived directions in the listening experiment after data correction are illustrated in fig. 4.3. Major deviations from the desired source directions can be observed only for WFS reproduction with larger secondary source distances, especially for the case with only 3 loudspeakers in the lower line of fig. 4.3. Differences in localisation blur associated with the 95% confidence interval (width of grey rays from each listener position) can also be observed with a tendency to increase slightly with more reverberation. The following sections analyse the results in detail.

4.4.1 Direction of localisation error

The localisation error in this work is defined as the difference between the desired real or virtual source direction and the direction of the auditory event. Thus, a negative localisation error means that a source has been perceived farther to the right than desired. The signed localisation error is used to analyse if there are differences between the desired and perceived source directions for the different conditions. The mean signed localisation error for every condition can be found in fig. 4.3 in the top line under each arrow depicting the mean perceived direction. It ranges from $-0.0°$ to $-13.2°$. Table 4.1 shows a summary of the slope coefficients of the fixed effects for the mixed-effects model with the best fit for the signed localisation error. All three main effects as well as the interaction 'method * position' have been included in the model and exhibit low p-values. The interaction constitutes a medium-size effect ($f^2 = 0.17$) with a negative slope coefficient, meaning that higher values of 'method' (larger Δx_0) as well as higher values of 'position' (farther to the side) shift the perceived direction farther to the right and, thus, increase the absolute localisation error. The same holds true for the effect 'method' alone, which only constitutes a small effect ($f^2 = 0.01$). The effect 'position' is also a small effect

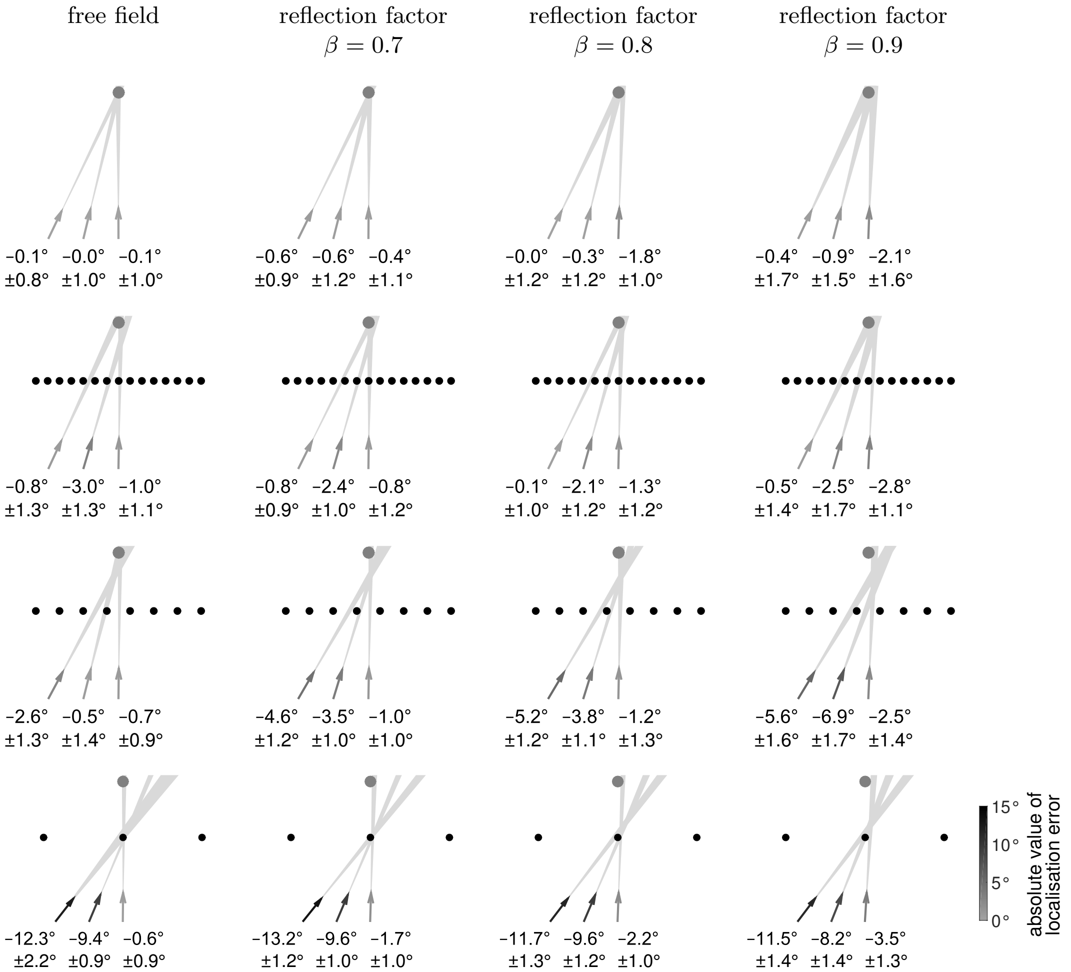


Figure 4.3: Average results of the listening experiment on localisation. The grey dots indicate the positions of the real or virtual sources, the black dots in the case of WFS reproduction the secondary sources. The first line of plots without secondary sources displays the case of real sources. In columns, the results are ordered by amount of reverberation, ranging from free field up to a rectangular room with a reflection factor of $\beta = 0.9$ for all boundaries. At each listener position, an arrow is pointing in the direction of the average auditory event. The colour of the arrow visualises the localisation error as difference of the average auditory event from the real or virtual source position as indicated by the colour bar. The localisation error is also stated below each listener position in the top line. The grey rays starting at each listener position display the 95% confidence interval which is also stated below each listener position in the lower line. Figure taken from [ES20], © 2020 IEEE.

Table 4.1: Slope coefficients for fixed effects of the mixed-effects model with best fit for predicted variable *'signed localisation error'* together with p-values and estimated effect sizes f^2 in the localisation experiment. Table taken from [ES20], © 2020 IEEE.

Fixed effect	Slope	p	f^2
method * position	-0.07	$< .001$	0.17
room	-1.38	$< .001$	0.01
method	-0.01	$< .001$	0.01
position	1.21	$< .001$	0.01

($f^2 = 0.01$) and leads to perception of the auditory event farther to the left. Care has to be taken when interpreting the size of the slope coefficients for each effect, as these depend on the magnitude of the values of each effect, e.g. 0 to 143 (Δx_0 in cm) for 'method' versus only 0 to 1.2 (in m) for 'position'. Therefore, the effect 'position' is easily overcompensated by 'method' and 'method * position' leading overall to localisation of sources farther to the right. The results appear to be consistent with fig. 4.3: WFS reproduction with greater secondary source distances in combination with a listener position farther to the side leads to a greater absolute localisation error, in the tested geometry the perceived direction shifts to the right. This is most obvious for the case of $\Delta x_0 = 143\,\mathrm{cm}$ with only 3 secondary sources in the last line of fig. 4.3: The perceived source direction appears to coincide with the direction of the nearest secondary source. These results replicate the findings by [Wie14], who showed this for anechoic environments, and extends them as they also hold true for WFS reproduction in listening rooms. An additional effect exists in the case of reflective environments: The predictor 'room' also has a small effect ($f^2 = 0.01$) on the signed localisation error. Higher reverberation leads to an increased localisation error, in this setup this means localisation farther to the right.

4.4.2 Accuracy of localisation

To evaluate the accuracy of localisation for WFS reproduction in listening rooms, the absolute deviations from the mean perceived direction in the conducted experiment were analysed. The only effect included in the mixed-effects model was 'room' with a low p-value and a small effect size ($f^2 = 0.01$) for a positive slope coefficient, cf. table 4.2. Thus, higher reverberation increases the localisation blur slightly. This result is consistent with findings in the literature, as reverberation increases the perceived width of the source [Kut17], which also appears to happen for virtual sources in WFS reproduction as shown by [Sta97]. A number of subjects also reported a greater uncertainty in determining the direction of the auditory event when higher reverberation was present. Overall, the accuracy of localisation is good for all conditions as is also expressed by the confidence intervals for the mean perceived directions in fig. 4.3 stated in the lower lines under the arrows depicting the mean perceived directions. The confidence intervals range from $\pm 0.8°$ to $\pm 2.2°$.

4.4.3 Difficulty of localisation

As an indication for the difficulty to localise a source in a certain condition, the elapsed time per trial was also analysed in a mixed-effects model. A summary of the

Table 4.2: Slope coefficients for fixed effects of the mixed-effects model with best fit for predicted variable *'absolute deviation from mean perceived direction'* together with p-values and estimated effect sizes f^2 in the localisation experiment. Table taken from [ES20], © 2020 IEEE.

Fixed effect	Slope	p	f^2
room	0.66	$< .001$	0.01

Table 4.3: Slope coefficients for fixed effects of the mixed-effects model with best fit for predicted variable *'elapsed time per trial'* together with p-values and estimated effect sizes f^2 in the localisation experiment. Table taken from [ES20], © 2020 IEEE.

Fixed effect	Slope	p	f^2
room	1.50	$< .001$	< 0.01
position	1.15	$< .001$	< 0.01

slope coefficients is shown in table 4.3. Only 'room' and 'position' have been included in the model with low p-values, but the effect sizes are very small ($f^2 < 0.01$). In the case of 'elapsed time per trial', a larger effect was found between subjects, indicating differences in the speed of solving the task for different test subjects. Also within-subjects times per trial varied considerably, as they are also influenced by the randomised succession of presented source directions evoking different extents of head rotations. Subjects needed on average 26 minutes to complete each session (without training). Minimum and maximum durations ranged from 6 up to 48 minutes per session. The overall mean elapsed time per trial was 13 s and ranged from below 1 to 85 s.

4.5 Discussion

Localisation of real sources in free field relies on binaural cues extracted from the ear signals. For frequencies below 1.4 kHz, the ITD is used, for higher frequencies the ILD and the ITD of the signal envelope. These binaural cues are evaluated in auditory frequency bands [Bla97]. For the case of a real source in a reflective environment, these cues are less reliable and additional features of the auditory system are needed, most importantly the precedence effect [LCYG99]. Moreover, the ear signals become less similar with reverberation present. This can be expressed by the IC, which is the maximum of the absolute normalised cross-correlation function of the two ear signals. A decrease of the IC has been associated with an increased perceived width of the source [BM81].

For the case of WFS in free field, it is also the precedence effect that allows for accurate localisation. [Wie14] could successfully model localisation of the human auditory system by evaluating only the ITD up to 1.4 kHz, so inconsistent binaural cues above the spatial aliasing frequency are disregarded. This is in accordance with the finding of [WK92], that for broadband signals the ITD below 1.4 kHz is the dominating cue for the direction of an auditory event.

The image source model used is only valid above approx. 500 Hz (cf. section 3.2.2.2), restricting the frequency range of reliable ITD cues. As the results for real sources in a reflective environment are in accordance with the state of research, it is assumed that the findings in this study on WFS in a listening room are valid as well.

The applied statistical analysis is revealing linear dependencies in the acquired data. These relations have been the target of the present study. The sample size was large enough to find even small effects and interactions. If an interaction is missing in the posed mixed-effects models, it can be concluded that there is no or only a very small difference in the data regarding this aspect. Conducting pair-wise comparisons between every condition in the study would have required a very large sample size to account for alpha error accumulation. It is assumed that no additional relations could be found this way that are of relevance in practice.

The listening experiment in the present study has shown, that, first, the localisation ability is only slightly degraded by reverberation for WFS and, second, the same effect holds true for both real and virtual sources in reflective environments. It has to be kept in mind, though, that this behaviour is undesired in a virtual acoustic scene. These findings agree with the experimental results by [SRdV97, Ver97]. The comparison with real sources helps to understand that the same mechanisms of the auditory system seem to come into effect for the localisation of real and virtual sources in rooms. That means that the precedence effect is enabling the listener to determine the direction of the auditory event by the direction of the first wave front, that is correctly synthesised by WFS, while suppressing additional reflections and focusing on consistent low-frequency binaural cues where no spatial aliasing is present.

The accuracy of localisation might be taken as an indicator for the perceived source width, as a broader source cannot be localised at a precise point and therefore causes a greater uncertainty in the responses of subjects. The results from section 4.4.2 on the accuracy of localisation could then be interpreted as follows:

1. Room reflections increase the perceived width of a source,
2. spatial aliasing does not increase the perceived width of a source.

This implicates that the sometimes suspected source broadening effect of spatial aliasing in WFS (cf. [Sta97]) can at most be a minor effect compared to the source broadening that arises from room reflections even in a strongly damped room.

The mixed-effects models given in sections 4.4.1–4.4.3 not only explain which effects are influencing the localisation properties of WFS in free field and in a listening room, they can also be used to declare quantitative models for the investigated aspects of localisation in the given setup. Taking also the grand mean γ_{00} into account, the following models can be used to calculate:

- the signed localisation error y_{error} in deg:

$$\begin{aligned} y_{\text{error}} = -0.53^\circ - 0.07^\circ \frac{1}{\text{cm} \cdot \text{m}} \cdot x_{\text{method}} \cdot x_{\text{position}} \\ - 1.38^\circ \cdot x_{\text{room}} - 0.01^\circ \frac{1}{\text{cm}} \cdot x_{\text{method}} + 1.21^\circ \frac{1}{\text{m}} \cdot x_{\text{position}}, \end{aligned} \tag{4.1}$$

- the absolution deviation from the mean perceived direction $y_{\text{deviation}}$ in deg:

$$y_{\text{deviation}} = 2.85^\circ + 0.66^\circ \cdot x_{\text{room}}, \tag{4.2}$$

and

- the elapsed time per trial y_{time} in s,

$$y_{\text{time}} = 11.43\,\text{s} + 1.50\,\text{s} \cdot x_{\text{room}} + 1.15\,\text{s}\frac{1}{\text{m}} \cdot x_{\text{position}}, \tag{4.3}$$

where the variables x_{room} (unit-less), x_{method} in cm and x_{position} in m take the values utilised for the fixed effects in the experiment. It has to be noted, that these models are only valid for the given geometry (cf. fig. 4.1) and should not be extrapolated, i.e. the variables for the fixed effects should only take values from within the range of the conducted experiment.

4.6 Summary

In this chapter, the influence of the reflections of a listening room on the azimuthal localisation of real and virtual sources by a listener was investigated. A listening experiment based on the reporting method evaluated in section 3.3 was conducted. The compared conditions consisted of real and virtual point sources synthesised by WFS both in free field and in rectangular rooms with different reflection factors simulated with the image source method. The secondary source distance for the WFS array was varied from 20 to 143 cm.

The results confirm the findings from [Wie14] that localisation in WFS in free field is only accurate for small secondary source distances ($\Delta x_0 = 20\,\text{cm}$ in the experiment). For large secondary source distances ($\Delta x_0 = 143\,\text{cm}$ in the experiment) the virtual source is localised in direction of the nearest secondary source. These rules also apply for WFS arrays in a reflective environment. Moreover, depending on the geometry of the setup, the reflections can cause additional small localisation errors. The results furthermore show that higher reverberation leads to a larger localisation error with a greater localisation blur, but this constitutes only a small effect that might be of minor importance in most practical applications. In comparison to the localisation of real sources, the same degradation of localisation through reverberation occurs. This effect is undesired in a virtual acoustic scene. There was also no difference between WFS with different secondary source distances concerning the influence of the listening room found in the experiment. The results suggest, that the precedence effect is the dominant factor in both localisation for WFS in free field as well as in reflective environments. The auditory system appears to ignore inconsistent binaural cues above the spatial aliasing frequency and relies on the existing consistent low-frequency ITD cues for determining the direction of a virtual source.

A larger localisation blur for higher reverberation indicates the broadening of sources through reverberation. No large effect could be found for spatial aliasing increasing the localisation blur. For the geometry given in the listening experiment, quantitative models were formulated to predict the signed localisation error, the absolute deviation from the mean perceived direction and the elapsed time per trial in the experiment.

5 Timbral perception of Wave Field Synthesis in a listening room

Though spatial sound reproduction methods have their emphasis on providing an experience of sources positioned in 2D or 3D space, possibly also with the impression of a virtual reflective environment, the perceived quality of such methods seems to be dominated by timbral and not spatial fidelity. This has been shown by [RZKB05] in a study on surround sound. According to [Oli04], timbre is also the dominant perceptual dimension in preference judgements of loudspeakers, further emphasising the importance of this percept in the field of audio. This chapter investigates colouration, the change in timbre compared to a reference, in Wave Field Synthesis in a listening room. The stimuli and results of the listening experiment and the statistical analysis are publicly available[1]. Under these conditions, colouration can be generated by spatial aliasing artefacts (cf. section 2.2.3) and also by the reverberation of the listening room, which both arise as a consequence of violating theoretic assumptions of WFS. Though it certainly is desirable to research colouration separated into its single dimensions, this thesis aims to assess the influence of the listening room on overall colouration in WFS in a first step (cf. section 2.2.1 for a discussion on the difficulty to break down the dimensions of timbre in the literature). As for the investigation of spatial perception in chapter 4, the wide range of degrees of freedom of the research topic had to be restricted. A linear secondary source array in a rectangular room is studied with a synthesised point source and a fixed listener position. The wall properties are varied as well as the audio content.

The following aspects of colouration in WFS in a reflective environment are targeted in this thesis:

- Is the perception of colouration caused by spatial aliasing in WFS influenced by the reflections of a listening room?
- Is the perception of colouration dependent on the audio content (pink noise or speech)?

Section 5.1 contains the details on the conditions in the listening experiment. The realisation of the experiment is described in section 5.2. The necessary data processing of the experimental results is given in section 5.3. The rating results can be found in section 5.4. Section 5.5 discusses limitations of the listening experiment and the linkage of the results to a spectral analysis of WFS in a listening room, followed by a summary in section 5.6.

---

[1] `http://doi.org/10.5281/zenodo.4036228`

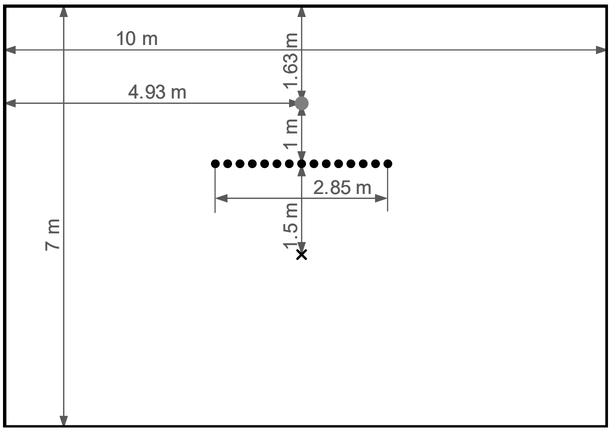


Figure 5.1: Geometry of the WFS setup in the colouration listening experiment. The grey dot indicates the position of the virtual source, the black dots the secondary sources. The black cross marks the listener position. The black borders show the walls of the listening room. Room height is 3 m. All sources and the receiver are positioned 1.59 m above the floor.

5.1 Conditions

To evaluate the influence of the listening room on colouration in WFS, a virtual point source synthesised by a linear WFS array with 15 secondary sources equidistantly spaced with approx. 20 cm in between sources was simulated. The secondary source distance is typical for a realistic setup. The array was placed in free field and in a rectangular room with three different wall properties and compared to a point source in free field at the same position that served as the reference. A high-pass filtered version of the reference with a corner frequency of 2 kHz and a 2nd order Butterworth characteristic was included as the low anchor of the MUSHRA test design (cf. section 3.4.2 and 5.2.1). Only a low and no mid-range anchor was included, leading to 6 conditions in the listening experiment. The anchor was chosen to be similar to the artefacts evoked by spatial aliasing. A high-pass filtered reference stimulus has already successfully been used by other studies on colouration in SFS [Wie14, Win19]. The rectangular room was simulated by the image source method with constant reflection factors $\beta = 0.7$, 0.8, and 0.9 for all boundaries. According to Sabine's formula eq. (3.1), this yields reverberation times of 0.27, 0.39 and 0.74 s, respectively. Section 3.2.2.2 gives more details on the employed room acoustical simulation. The listener was positioned in front of the array facing the real or virtual point source. Fig. 5.1 shows the geometry of the setup.

To synthesise the virtual WFS point sources, the driving function given in [Ver97, eq. (2.22a)] was used with the listener position chosen as the reference point. The upper corner frequency of the pre-equalisation filter was set to the lower limit spatial aliasing frequency according to eq. (2.21) resulting in $f_{\text{alias}} = 842\,\text{Hz}$. The lower corner frequency of the pre-equalisation filter was chosen adapted to the array length as 50 Hz. To reduce diffraction artefacts from the edges of the secondary source array, a raised-cosine tapering window over 30% of the array length was applied.

The stimuli were simulated with binaural synthesis as described in section 3.2.1.2. To avoid colouration changes with turning of the head of a listener, only static binaural synthesis was employed, cf. section 3.4.1. The stimuli were created with the help of the Sound Field Synthesis Toolbox [WS12], release 2.5.0[2].

As audio content, 4 s of female speech and a sequence of four pink noise pulses were used. The pink noise pulses had a duration of 900 ms per pulse and a pause of 500 ms between pulses. Each pulse was windowed with a half cosine window with a length of 50 ms at the start and end. The noise pulses were played back in a loop. The signals have already been successfully used in prior studies on colouration in spatial audio rendering techniques [WHSR14, WWH+18].

A loudness model [ANS07, MGB97] was used to adjust the loudness of all conditions in order to avoid additional distracting cues between the conditions. As audio content for the loudness estimation of the model, white noise was used. The sound level at the ear of the subjects in the experiment was measured with an ear coupler G.R.A.S. RA0039 Ear Simulator IEC 60318.1 [IEC09]. The average sound level for all conditions was approx. 70 $\mathrm{dB_{SPL}}$ for all but one subject, who preferred to turn the volume slightly down.

5.2 Realisation of the listening experiment

5.2.1 Experimental design and sample size estimation

To study the influence of the listening room on the colouration in WFS, a modified MUSHRA [ITU15] test design was chosen, cf. section 3.4.2. Instead of audio quality as defined in the ITU recommendation, the colouration of the presented stimuli compared to the reference was rated on a scale ranging from 'no difference' to 'large difference' (translated from originally German 'kein Unterschied' and 'starker Unterschied') without subdivisions of the scale.

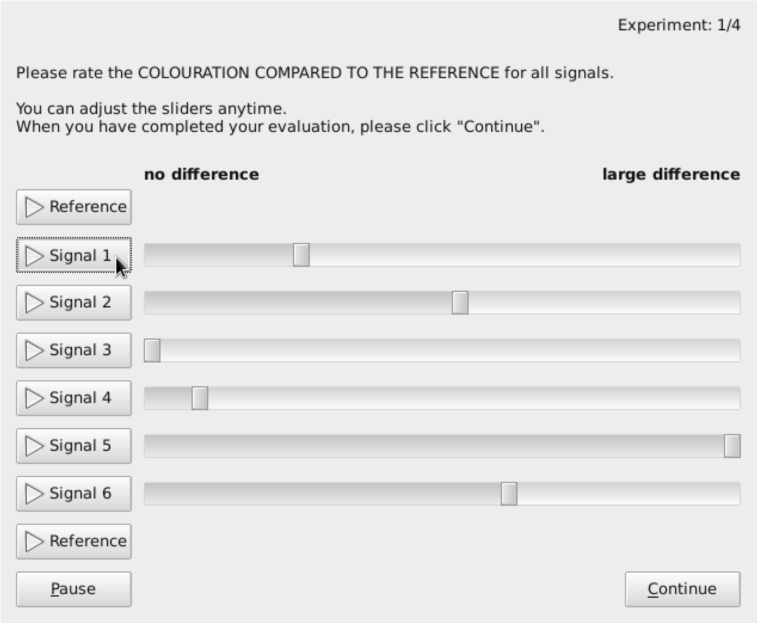
The rating results are first 7compared pairwise by the Wilcoxon signed rank test [Wil45] for matched pairs as suggested by [MDM18]. This non-parametric test evaluates the differences in ratings of two stimuli and takes both the sign and the magnitude of the differences in terms of ranks in account. This test therefore appears to be suitable for a MUSHRA-style test where ratings cannot be assumed to be normally distributed and subjects might use the scale differently. Additionally, a Friedman test [Fri37] followed by a Conover post-hoc test [Con99] is performed to investigate the data further.

Due to the challenges of a MUSHRA-style experiment as described in section 3.4.2, it is to be expected that only large differences can be detected with a moderate sample size. The necessary minimum sample size has been calculated with G*Power [FELB07], version 3.1.9.4, for a two-tailed Wilcoxon signed rank test with an effect size of 0.8 constituting a large effect according to the classification by [Coh88] and a desired power of 0.8. As multiple tests per MUSHRA run are conducted, the chosen type I error probability of $\alpha = 0.05$ has to be corrected. Excluding the anchor, this leaves $N_\text{conditions} = 5$ conditions to be compared pair-wise ($N_\text{compared} = 2$), resulting in

$$\binom{N_\text{conditions}}{N_\text{compared}} = \binom{5}{2} = 10 \tag{5.1}$$

---

[2] `http://doi.org/10.5281/zenodo.2597212`


Experiment: 1/4

Please rate the COLOURATION COMPARED TO THE REFERENCE for all signals.

You can adjust the sliders anytime.
When you have completed your evaluation, please click "Continue".

Figure 5.2: Graphical user interface in the rating phase of the listening experiment on colouration. Text translated from originally German.

comparisons. With Bonferroni correction this yields a corrected $\alpha_{\text{corr}} = 0.005$ and results in a minimum sample size of 26 subjects. Please note that for the sample size estimation the distribution of the ratings has to be approximated with a normal distribution.

5.2.2 Experimental procedure

Before the start of the experiment, subjects were instructed about the experimental task. Each signal should be rated regarding its colouration compared to the reference on the scale from ‘no difference’ to ‘large difference’. The concept of colouration was explained by giving examples of the different dimensions of colouration. It was highlighted that the colouration of only the direct sound component in comparison to the reference was to be evaluated while ignoring other possible differences such as loudness or length and timbre of reverberation. Subjects were also instructed to assign the signal(s) with the strongest colouration compared to the reference to the end of the scale.

Stimuli were presented over headphones type AKG K601. Rendering of stimuli in static dynamic binaural synthesis was performed by the SoundScape Renderer [GS12], commit 2b11775[3]. Subjects were guided through the experiment by a graphical user interface as depicted in fig. 5.2. Overall, four MUSHRA runs had to be completed (noise and speech as audio content, each twice). The order of the four runs as well as the order of the signals within each run was randomised. Before the main part of the experiment, one MUSHRA run with the same conditions, but with a short Cello excerpt as audio content, served as training.

---

[3] `http://github.com/SoundScapeRenderer/ssr`

5.2.3 Participants

34 subjects with an average age of 32 years participated in the experiment. The experiment was conducted partly at the University of Rostock and partly at the Technische Universität Berlin (20 and 14 subjects, respectively). 32 subjects self-reported normal hearing, the remaining two had a minor tinnitus that they considered as not impairing them in the experimental task. 26 subjects had home or professional experience in the field of audio. 26 had participated in at least one listening experiment before.

5.3 Data processing

The rating scale was coded with numerical values from 0 ('no difference') to 1 ('large difference'). The ratings of subjects that did not assign the signal(s) with the largest colouration compared to the reference to the upper end of the scale were rescaled to contain at least one maximum value of 1.

The repetitions of the MUSHRA runs with noise and speech were used to detect subjects that did not reliably rate the stimuli. To determine the test-retest reliability of the repetitions, the rank-based Spearman correlation coefficient was calculated. According to [MDM18], only a moderate correlation of .65 is to be expected in a MUSHRA test design. As subjects could not assign a perfectly equal slider position to stimuli they perceived as equally strongly coloured compared to the reference, the ratings were quantised to 51 steps. This avoids the assignment of different ranks for effectively equal ratings. All ratings of subjects that did not yield a Spearman correlation coefficient of at last .65 in both the noise and speech MUSHRA runs were discarded. This was the case for 6 subjects. For the remaining 28 subjects the results of the repetitions were averaged per condition.

5.4 Results

5.4.1 Rating results

The subjects needed an average time of 16 minutes to complete the main part of the experiment. Fig. 5.3 shows the medians of the rating results and the 96.4% confidence intervals for the medians for noise and speech, respectively, to give a first impression of the data. The confidence intervals are calculated with a distribution-free method [KV07] where the targeted 95% confidence intervals cannot be determined exactly. Therefore, the next larger interval is chosen. As can be seen from fig. 5.3, subjects were able to identify the hidden reference and rated the low anchor as the most coloured stimulus for both the noise and the speech MUSHRA runs. All remaining conditions are also rated as coloured compared to the reference. For noise as audio content, the colouration of WFS in free field is on average rated higher than for WFS in a reverberant environment. For speech, the results appear vice versa. A slight tendency for increasing colouration with reverberation can be observed both for noise and speech as audio content.

To determine whether the observed differences can be regarded as statistically significant, Wilcoxon signed rank tests for matched pairs were performed first, cf. sec-

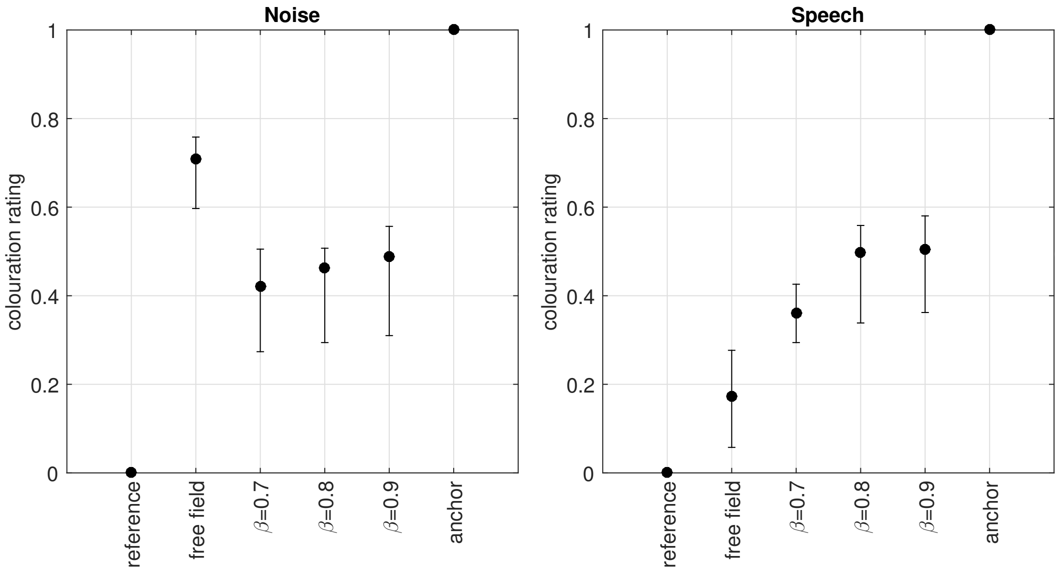


Figure 5.3: Medians and 96.4% confidence intervals of colouration ratings for noise (left) and speech (right) as audio content. The WFS conditions are named according to their reflective environments.

tion 5.2.1. This test takes into account, whether each subject rated one condition as more coloured than another. This relation cannot be deduced from the confidence intervals in fig. 5.3 as the intra-subject relations between the ratings are missing. Therefore, not overlapping confidence intervals cannot be taken as an indication for a significant difference. Tables 5.1 and 5.2 list the resulting undershot p-values taken from the table in [McC65] for each pair of conditions along with the matched-pairs rank biserial correlation coefficient r_C as a measure of effect size [KRM11] for noise and speech as audio content, respectively. Due to multiple testing as described in section 5.2.1, the difference of two medians of the ratings for a pair of conditions is in this study considered as statistically significant (marked in bold in tables 5.1 and 5.2) with error probability $\alpha = 0.05$ if the corresponding p-value is equal or smaller than 0.005. Tables 5.1 and 5.2 show that all conditions are perceived as more coloured than the reference with high effect sizes. Additionally, in the case of noise as audio content, all conditions of WFS in a reflective environment are perceived as less coloured than WFS in free field with medium to high effect sizes. For speech as audio content, there is a significant difference between the WFS conditions with $\beta = 0.7$ and 0.8 as reflection factors for the listening room with a medium effect size. All remaining comparisons are regarded as not significant.

To gain further insights into the acquired data, a Friedman test was performed showing that significant differences between the medians of the ratings are present in the data for both noise ($\chi^2(4) = 73.8$, $p < .0001$) and speech ($\chi^2(4) = 60.9$, $p < .0001$) as audio content. A Conover post-hoc test was then performed with Bonferroni-Holm correction for multiple testing, which constitutes a more liberal method of testing than the Wilcoxon signed rank test with Bonferroni correction. The results for the Conover test are reported in table 5.3 and 5.4 with p-values and the differences of the medians of the ratings as measure for the effect sizes.

Table 5.1: Undershot p-values and effect sizes r_C for the matched-pairs Wilcoxon signed rank test for *noise* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold.

		free field	$\beta = 0.7$	$\beta = 0.8$	$\beta = 0.9$
reference	p	**.0001**	**.0001**	**.0001**	**.0001**
	r_C	**1**	**1**	**1**	**1**
free field	p	-	**.0001**	**.001**	**.001**
	r_C	-	**.81**	**.71**	**.68**
$\beta = 0.7$	p	-	-	1	1
	r_C	-	-	.26	.22
$\beta = 0.8$	p	-	-	-	1
	r_C	-	-	-	.20

Table 5.2: Undershot p-values and effect sizes r_C for the matched-pairs Wilcoxon signed rank test for *speech* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold.

		free field	$\beta = 0.7$	$\beta = 0.8$	$\beta = 0.9$
reference	p	**.0001**	**.0001**	**.0001**	**.0001**
	r_C	**1**	**1**	**1**	**1**
free field	p	-	1	.02	.05
	r_C	-	.28	.52	.44
$\beta = 0.7$	p	-	-	**.005**	.03
	r_C	-	-	**.65**	.48
$\beta = 0.8$	p	-	-	-	1
	r_C	-	-	-	.21

Table 5.3: p-values and differences of medians (value of row minus value of column) for the Conover post-hoc test for *noise* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold.

		free field	$\beta = 0.7$	$\beta = 0.8$	$\beta = 0.9$
reference	p	**<.0001**	**<.0001**	**<.0001**	**<.0001**
	difference of medians	**-.71**	**-.42**	**-.46**	**-.49**
free field	p	-	**<.0001**	**<.0001**	**<.0001**
	difference of medians	-	**.29**	**.24**	**.22**
$\beta = 0.7$	p	-	-	**.0099**	**<.0001**
	difference of medians	-	-	**-.04**	**-.07**
$\beta = 0.8$	p	-	-	-	**.0018**
	difference of medians	-	-	-	**-.02**

Table 5.4: p-values and differences of medians (value of row minus value of column) for the Conover post-hoc test for *speech* as audio content in the colouration experiment. The WFS conditions in the table header row and column are named according to their reflective environments. Statistically significant entries are marked in bold.

		free field	$\beta = 0.7$	$\beta = 0.8$	$\beta = 0.9$
reference	p	**<.0001**	**<.0001**	**<.0001**	**<.0001**
	difference of medians	**-.17**	**-.36**	**-.50**	**-.51**
free field	p	-	.69	**<.0001**	**<.0001**
	difference of medians	-	-.19	**-.32**	**-.33**
$\beta = 0.7$	p	-	-	**<.0001**	**<.0001**
	difference of medians	-	-	**-.14**	**-.14**
$\beta = 0.8$	p	-	-	-	.79
	difference of medians	-	-	-	-.01

Statistically significant entries are marked in bold. According to the Conover test, there also exist significant differences between the condition of WFS in free field and the conditions of WFS in rooms with higher reflection factors for speech as audio content as well as between almost all conditions with different reflection factors for both types of audio content.

5.4.2 Influence of the listening room

The results for noise as audio content show that the reflections of the listening room lessen the colouration of WFS reproduction. With increasing reflection factor, though, the perceived overall colouration increases again indicating that the reverberation of the room introduces colouration itself.

For the speech stimulus, no decolouration effect of the room can be observed. Instead, the colouration due to reverberation is the dominant factor. This can be explained by the fact, that speech contains less high frequency content and thus the influence of spatial aliasing is smaller, as is illustrated in fig. 5.4, where the long-term spectrum of the speech signal is compared to the transfer function of the WFS system. The frequency band corrupted by spatial aliasing contains considerably less energy than the lower frequency band for the speech stimulus. The results

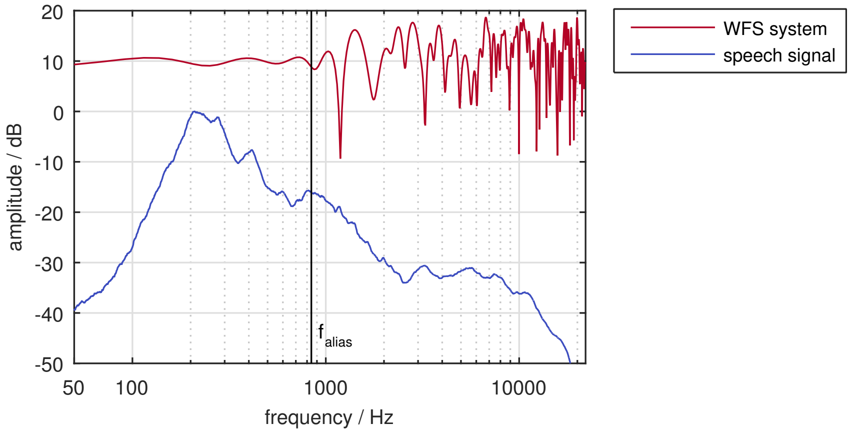


Figure 5.4: Magnitude of the long-term spectrum of the speech signal used in the colouration experiment (blue) in comparison to the magnitude of the single-channel transfer function of the WFS system in free field to the listener position (red). The spectrum of the speech signal is normalised to its maximum, the transfer function of the WFS system is shifted for better discriminability. The vertical black line denotes the lower limit spatial aliasing frequency of the WFS system according to eq. (2.21).

are in accordance with findings in the literature for WFS [Sta97] and stereophonic reproduction in listening rooms [Pul01].

5.5 Discussion

Colouration in WFS is caused by ripples in the magnitude spectrum above the spatial aliasing frequency, cf. section 2.2.3. In a listening room, these spectral fluctuations are lessened by reflections, which serves as an explanation for the reduction of the perceived colouration of WFS in a listening room for broadband stimuli. This is illustrated in fig. 5.5 which depicts the magnitude responses smoothed in third-octave bands of the RIRs corresponding to the conditions in the listening experiment. As quantification of the extent of the spectral fluctuations, the standard deviation of the magnitude spectrum in dB above f_{alias} is used. This measure has also been used to serve as a predictor for colouration in various studies, e.g. for room-in-room scenarios [HvdP15] and for WFS in free field [Wit07]. Alternatively, a measure based on the autocorrelation function has also been proposed [RJ03]. As can be seen, the standard deviations of the magnitude spectra are decreasing with increasing reverberation: 1.67 dB, 1.12 dB, 1.03 dB, and 1.00 dB for WFS in free field and WFS in the simulated listening room with $\beta = 0.7$, 0.8, and 0.9, respectively. The standard deviation measure postulates a further decreasing colouration when more reverberation is present while the listening experiment found the opposite to be true. This can be explained by additional colouration that is introduced by reflections themselves. Especially below f_{alias}, reverberation is causing additional spectral fluctuations as can be seen in fig. 5.5, which are not included in the used standard deviation measure. More sophisticated approaches to model the perception of colouration also rely on spectral differences, evaluated in auditory bands as used for the modelling of colouration of WFS in free field [WER15]. They can be expected to yield results

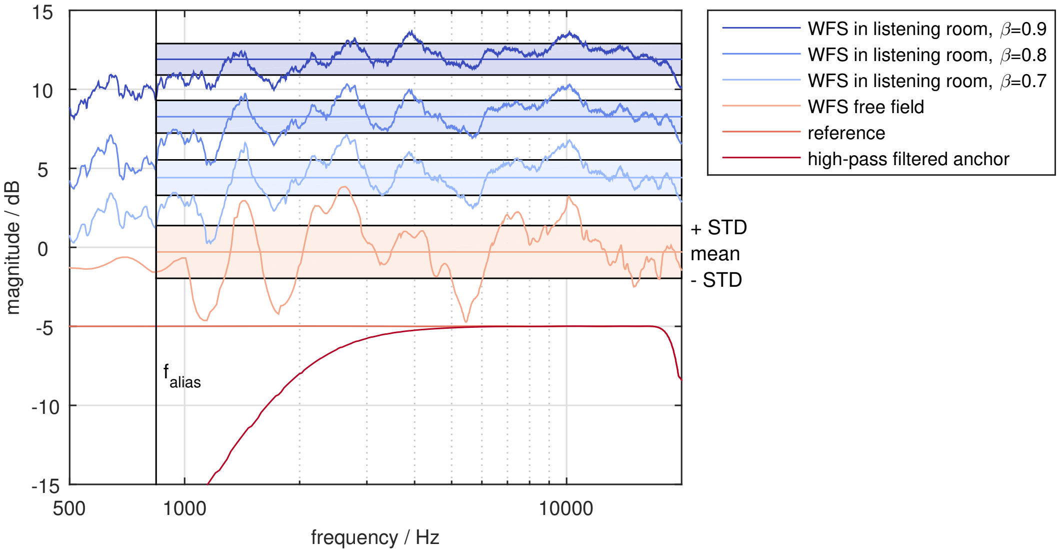


Figure 5.5: Magnitude responses of RIRs corresponding to the six conditions of the listening experiment. The WFS responses are smoothed in third-octave bands, all responses are shifted along the y-axis for better discriminability. The shaded areas enclose the mean and mark the ± standard deviation (STD) borders of the WFS magnitude responses.

similar to the results by the standard deviation measure.

Similar effects of decreasing spectral fluctuations with reverberation can also be shown for the case of measured RIRs of virtual sources synthesised by WFS in a real room [ES17]. Fig. 5.6 shows magnitude responses of a virtual point source in different room acoustic conditions of a rectangular room and in free field similarly to fig. 5.5. The employed RIRs have been measured in the Audio Lab at the University of Rostock for a 64-channel loudspeaker array and are publicly available[4] in the Spatially Oriented Format for Acoustics (SOFA)[5] standardised in [AES15] together with BRIRs for frontal head orientations in 2° steps. Details for the measurements can be found in [EGWS15]. The four room acoustic conditions in fig. 5.6 have been created by applying different amounts of broadband absorbers and pyramid-shaped foam on walls and ceiling of the Audio Lab. The uniform reflection factors given in the legend have been calculated from the measured mid-frequency reverberation times of the four room acoustic conditions via the equivalent absorption area and the reverberation time formula according to Sabine. This simplified calculations serve for comparison with the reflection factors used in the room acoustical simulations in this thesis. The geometry of the setup is given in fig. 5.7a, only a linear subarray of 16 loudspeakers of the 64-channel array has been used. For the measured RIRs, the same trend as in the case of simulated RIRs can be seen above the spatial aliasing frequency: The stronger the reflections, the more the spectral fluctuations caused by spatial aliasing are smoothed although there are some deviations from this finding, e.g. above 10 kHz. This shows, that the decolouration effect of virtual sources synthesised by WFS through reverberation of the listening room as found in the listening experiment is also in effect in real room scenarios. The main difference of

---

[4] `http://doi.org/10.14279/depositonce-87.6`

[5] `http://www.sofaconventions.org`

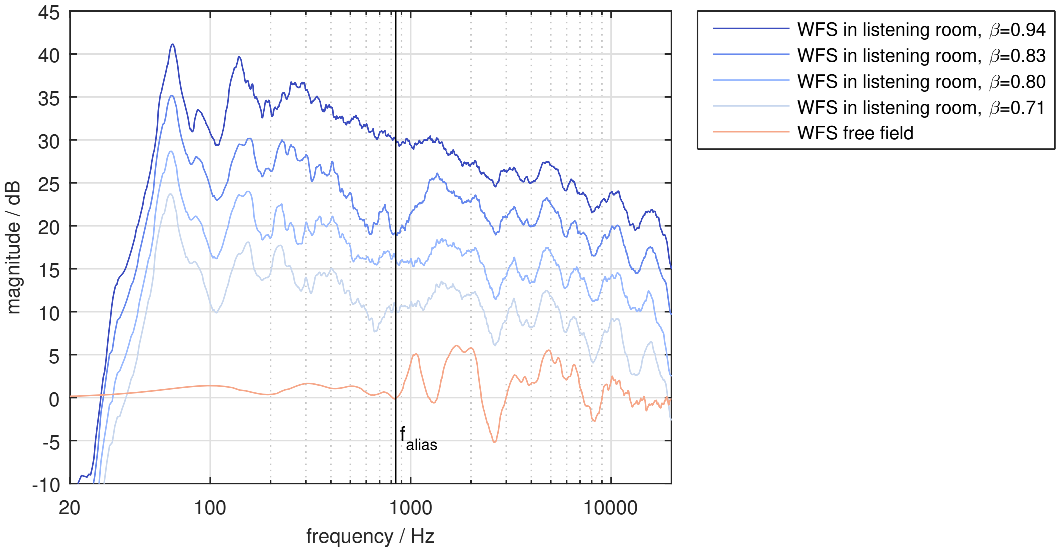
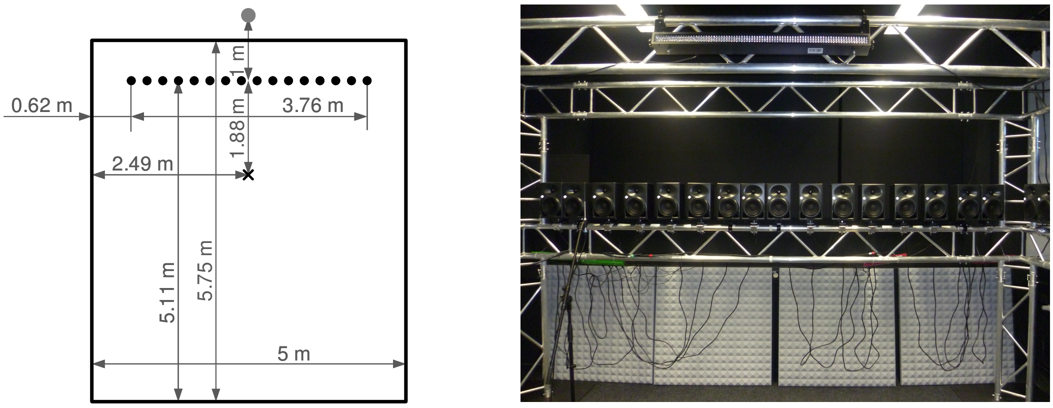


Figure 5.6: Magnitude responses of a virtual point source at the receiver position for different absorber configurations in the Audio Lab at the University of Rostock and in free field as given by the geometry in fig. 5.7a. The absorber configurations have been converted to uniform reflection coefficients as stated in the legend. The magnitude responses are smoothed in third-octave bands and shifted along the y-axis for better discriminability.

the measured RIRs in fig. 5.6 to the simulated cases in fig. 5.5 is the trend to decreasing energy with higher frequencies in the measured RIRs while the simulated RIRs are on average constant over frequency. This is caused by the typical conditions in a real room where absorption increases with frequency as sound energy with smaller wave lengths is more easily absorbed by porous materials inside the room [Kut17]. The frequency independence of the reflection factors in the listening experiment do thus not conform to real room surfaces. With the employed simulation it is shown, though, that the decolouration effect of the listening room in WFS reproduction is not simply due to a lower level of the high frequency content that is corrupted by spatial aliasing.

As in the case of simulated RIRs, a possible additional colouration generated by the reflections themselves is not covered by the analysis in fig. 5.6 when observing only the spectral fluctuations above f_{alias}. In the case of a small room as presented in this example, colouration caused by the room itself is especially to be expected due to low-frequency room modes around 60 Hz. The excitation of room modes in WFS does not only depend on the position of the secondary sources, but also on type and position of the virtual source [ES17] as can be observed in fig. 5.8 which compares the low-frequency magnitude responses of the virtual point source from the setup in fig. 5.7a in the Audio Lab at the University of Rostock and of a virtual plane wave travelling perpendicular to the same array for the condition with the highest reverberation time. As can be seen, the extent of the formation of room modes differs for the two virtual sources. This is due to the fact that the loudspeakers are driven with different gains and delays depending on the type or position of the virtual source and thus the superposition of the loudspeakers differ as well. From this it can be concluded that virtual sources of different types and at



(a) **(b)**

Figure 5.7: WFS system in the Audio Lab at the University of Rostock. (a) Employed geometry of the system. The grey dot indicates the position of the virtual source, the black dots the secondary sources (exact positions are not equidistantly spaced as depicted here, cf. [EGWS15]). The black cross marks the listener position. The black borders show the walls of the listening room. Room height is 3 m. All sources and the receiver are positioned 1.59 m above the floor. (b) Loudspeaker array in the Audio Lab. In this acoustic condition, the walls are equipped with broadband absorbers in black in the upper and pyramid-shaped foam in the lower half.

different positions can be coloured differently by room modes in small rooms.

Evaluation of colouration in room acoustics is also related to adaptation mechanisms of the auditory system and to binaural decolouration. Previous studies derived from their results, that binaural decolouration is not in effect for colouration of WFS in free field conditions although spatial aliasing artefacts are similar to reflections with a very short delay [Wit07, Wie14]. The influence of the combination of spatial aliasing and room reflections on binaural decolouration might therefore be especially interesting. Unfortunately, binaural decolouration and the adaptation to room acoustics are topics that are not easily investigated. From the literature, the role of these auditory mechanisms in binaural synthesis compared to real listening situations is not clear and thus no conclusions can be drawn concerning this aspect.

Additionally, it should be taken into consideration that the perception of colouration could be individual. An indication for this might be found in the reports of subjects after the experiment concerning the speech stimulus in combination with their ratings: While some subjects could perceive no or almost no difference of the free field WFS condition to the reference, 8 subjects rated this condition as strongly coloured compared to the reference. Several of these subjects reported this perceived colouration as a ‘comb filter effect’ that they appeared to be very sensitive for and found disturbing. As demonstration of these differing perceptions, the test subjects have been divided in two groups in fig. 5.9, which shows the medians and approx. 95% confidence intervals for the groups, though this illustration must not be used or regarded as forming a valid statistical analysis. An individual perception of timbre might also make the breakdown of timbre in its single dimensions even more challenging.

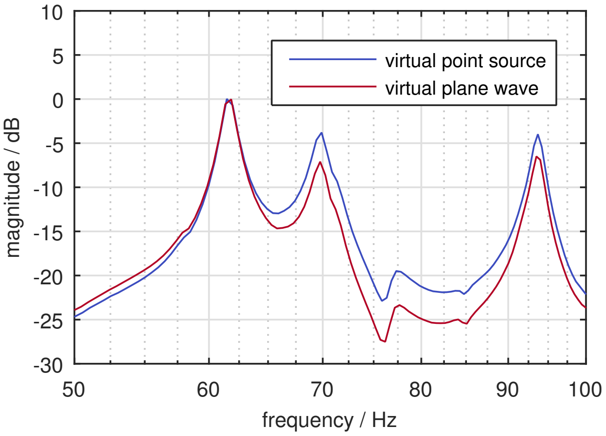


Figure 5.8: Low frequency magnitude responses of virtual sources for the WFS setup in fig. 5.7a. The virtual point source is positioned as depicted in fig. 5.7a, the virtual plane wave is travelling perpendicular to the array. The responses have been normalised to the lowest mode at approx. 61 Hz.

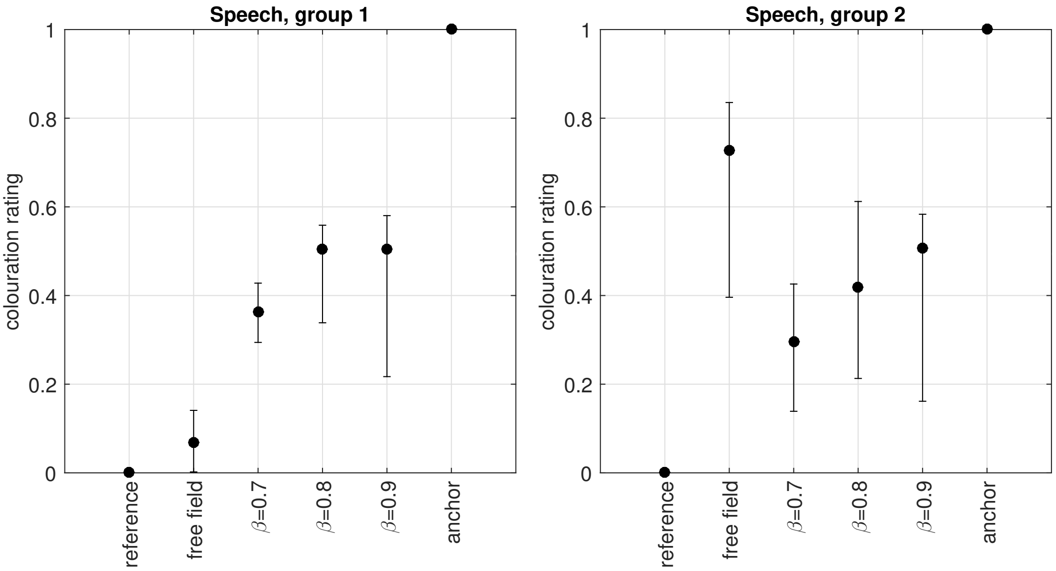


Figure 5.9: Medians and 95.9% (group 1) and 96.1% (group 2) confidence intervals of colouration ratings for speech as audio content divided in two groups. The WFS conditions are named according to their reflective environments.

5.6 Summary

This chapter investigated the influence of the listening room on colouration in WFS that arises due to spatial aliasing. A listening experiment with a modified MUSHRA test design was conducted comparing the timbre of virtual point sources in free field and in a rectangular listening room with different reflection factors to a real source in free field. The listening room in the experiment was simulated with the image source method. Both noise and speech were employed as audio content.

The results show that the reverberation of the listening room leads to less colouration of a virtual source synthesised with WFS for broadband stimuli. At the same time, the colouration increases with increasing reflection factor, though this constitutes a smaller effect. The decolouration effect is not present when using bandlimited audio content such as speech that is affected less by spatial aliasing.

As reflections of the listening room lessen the spectral fluctuations of the reproduced sound field, the standard deviation of the magnitude response of a virtual source above f_{alias} was calculated to quantify the decolouration effect observed in the listening experiment for broadband stimuli, but this measure does not conform with results for different reflection factors. Below the aliasing frequency, colouration by the room itself becomes apparent for measured RIRs in a small room, including the excitation of room modes, which additionally depends on the types and parameters of the virtual source. For a room that induces very strong colouration, this colouration could possibly outweigh the decolouration effect of spatial aliasing artefacts.

In practice, it is therefore advisable to use a listening room with additional absorbent material that is still generating some reverberation above f_{alias}, but not an anechoic chamber if a reduction of the colouration induced by the reproduction system is desired. A low level of reflections should be targeted to not add too much undesired reverberation to the reproduced scene and to minimise additional colouration that could be introduced by reflections, especially room modes.

6 Conclusions

This thesis investigated the influence of the listening room in Wave Field Synthesis. A reverberant environment violates a theoretic assumption of sound field synthesis and, consequently, deviations from the desired sound field arise. These deviations are not only of physical nature, but also influence the perception of a human listener. In this work, both aspects have been analysed with a focus on the spatial and timbral perception of WFS in a listening room which has rarely been treated in research so far.

The perceptual evaluation of localisation and colouration properties in WFS required a thoroughly planned application of methods to yield valid results. Direct comparisons in listening experiments with different reverberant environments necessitate the use of simulated stimuli instead of in-situ experiments. As simulation methods, binaural synthesis and the image source method for room acoustical simulation have been chosen. Both these methods have inherent limitations that put the validity of simulated stimuli at risk, e.g. the use of non-individual HRTFs in binaural synthesis can have an influence on both the spatial and timbral perception of a listener. Therefore, special attention has been given to the application of these methods and how they can be used for listening experiments on the influence of the listening room in WFS. To this end, a comprehensive list of relevant aspects treated in the literature have been revisited, accurately implemented, and experimental designs have been evaluated. Limitations of the simulations have been identified and discussed and the experiments have been conceptualised accordingly to ensure validity of results. Accompanying this work, datasets and code have been published to support reliability of the results and to foster a culture of Open Science. The Open Science contributions created by the author as well as resources provided by other researchers, that have been applied in this thesis, are summarised in appendix C.

Spatial perception has been evaluated in terms of azimuthal localisation, as 2.5D WFS in the horizontal plane is the most typical use case. As a human listener is able to localise sound sources in the frontal horizontal plane with an accuracy of 1° [Bla97], the reporting method for a localisation experiment has to be at least as accurate. A reporting method of pointing with the head while being assisted by visual feedback with an HMD was developed and successfully evaluated regarding its accuracy. Moreover, the method can easily be extended to be used for localisation experiments in 3D. Results of the localisation experiment confirm the findings from [Wie14] for WFS in free field and complement them for the synthesis in reverberant environments: Both in free field and in a listening room, accurate azimuthal localisation in 2.5D WFS is only possible for small secondary source distances, while for larger distances a virtual point source is localised in direction of the nearest secondary source. This has been shown for an extended listening area. These results can be explained by the precedence effect, i.e. the listener determines the direction of a source by the direction of the first wave front which is correctly synthesised by WFS. Furthermore, the reflections of the listening room can cause an additional small localisation error, depending on the geometry of the setup. A small effect of the listening room on the localisation blur has also been found: The localisation blur increases with higher reverberation. This undesired effect is not exclusive to the perception of virtual point sources in WFS, it has been shown to affect real point sources as well and might be linked to broadening of the perceived source width by reverberation.

Timbral perception was investigated as colouration of virtual point sources in different anechoic and reverberant environments compared to a real point source in free field as reference. For broadband stimuli, the reflections lessen the colouration introduced by spatial aliasing artefacts by reducing fluctuations in the magnitude spectrum. For bandlimited audio content which is less affected by spatial aliasing, this decolouration effect could not be observed. Furthermore, colouration is increasing with the amount of reverberation, though this constitutes a smaller effect. The listening room itself can thus also cause colouration, especially below the spatial aliasing frequency and in terms of room modes in small rooms.

The results can serve as guidelines for first recommendations on the design of listening rooms for WFS. In principle, reverberation of the listening room is an undesired change of the reproduced acoustic scene and is prone to degrade localisation accuracy of synthesised sources. On the other hand, reflections of the listening room can alleviate the typical WFS colouration that prevents the use of WFS in high-fidelity applications. As this decolouration effect could already be observed for a room with a low reverberation time of 0.27 s, it is recommended to install WFS arrays in highly damped listening rooms to reduce system-induced colouration while avoiding too much additional reverberation and room-induced colouration as well as degradation of localisation accuracy. Additionally, the influence of the listening room should be taken into account in the mixing and mastering process of content material for WFS systems. This is already an established practice in stereophony where a typical listening room is assumed and the content is treated to overcome to some degree the comb filter effects of this reproduction technique [Too08, WHR18].

This thesis investigated the influence of the listening room only for WFS, but parts of the results appear to be applicable to other SFS techniques such as NFC-HOA [Dan03] as well. The increase in localisation blur might also affect NFC-HOA and the decolouration effect can also hold true for its spatial aliasing induced colouration.

The experiments in this thesis only address some of the possible degrees of freedom of the topic. Future work could expand the experiments to more diverse listening rooms with different geometries, higher reverberation times or frequency-dependent reflection factors of the boundaries as well as extension to other SFS techniques. Listening experiments based on measured instead of simulated room impulse responses should also be included. This requires BRIR measurements of the same array in an anechoic chamber and in rooms with a large spread of different acoustic properties with the same HATS, though, which have not been available at the time of writing of this thesis. As there seems to be a trade-off between the decolouration effect and undesired reverberation with additional colouration of the listening room in WFS reproduction, future research should determine a threshold of listening room reverberation that is still mitigating the colouration of the reproduction method.

Finally, evaluation of colouration in room acoustics is also related to binaural decolouration [Brü01, Zur79]. The influence of the combination of spatial aliasing and room reflections on binaural decolouration is therefore of interest for future research.

Appendices

Appendix A – Coordinate System

Fig. A.1 shows the coordinate system, that is used throughout this thesis. A positional vector $\mathbf{x}$ is either given by its Cartesian coordinates (x, y, z) in a right-handed coordinate system or by its length $R = |\mathbf{x}|$, azimuth angle $\varphi \in [0, 2\pi[$ and elevation angle $\vartheta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. The azimuth is defined in anticlockwise direction starting at the x-axis. The elevation angle is positive for positive values of z.

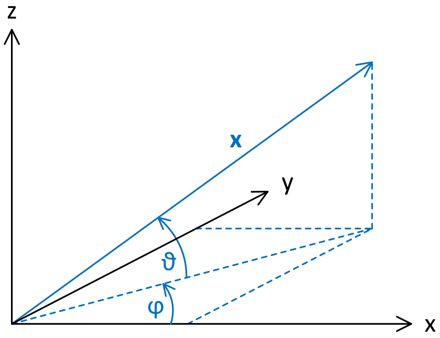


Figure A.1: Coordinate system used in this thesis.

Appendix B – Conventions for the Fourier transform

The Fourier transform and its inverse with respect to time t are defined in this thesis as

$$P(\mathbf{x},\omega) = \mathcal{F}_t\left\{p(\mathbf{x},t)\right\} = \int\limits_{-\infty}^{\infty} p(\mathbf{x},t)\mathrm{e}^{-\mathrm{j}\omega t}\,\mathrm{d}t \tag{B.1}$$

$$p(\mathbf{x},t) = \mathcal{F}_t^{-1}\left\{P(\mathbf{x},\omega)\right\} = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} P(\mathbf{x},\omega)\mathrm{e}^{\mathrm{j}\omega t}\,\mathrm{d}\omega. \tag{B.2}$$

The Fourier transform and its inverse with respect to space, given exemplarily for the coordinate x, are defined as

$$P(k_x,y,z,t) = \mathcal{F}_x\left\{p(\mathbf{x},t)\right\} = \int\limits_{-\infty}^{\infty} p(\mathbf{x},t)\mathrm{e}^{\mathrm{j}k_x x}\,\mathrm{d}x \tag{B.3}$$

$$p(\mathbf{x},t) = \mathcal{F}_x^{-1}\left\{P(k_x,y,z,t)\right\} = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} P(k_x,y,z,t)\mathrm{e}^{-\mathrm{j}k_x x}\,\mathrm{d}k_x. \tag{B.4}$$

The combination of these Fourier transform definitions causes a plane wave

$$p(\mathbf{x},t) = \mathrm{Re}\left\{\mathrm{e}^{-\mathrm{j}\mathbf{k}\mathbf{x}}\mathrm{e}^{\mathrm{j}\omega t}\right\} \tag{B.5}$$

to propagate into the direction of the wave number vector $\mathbf{k} = (k_x, k_y, k_z)^{\mathrm{T}}$.

Appendix C – List of Open Science contributions

In the context of this thesis, several Open Science contributions have been made by the author as is also indicated in the respective sections. The following list summarises these contributions:

- database of headphone compensation filters for the KEMAR manikin, including HpTF measurements and Matlab code to calculate the filters [EGWS17]
  `http://doi.org/10.5281/zenodo.401042`
- low-frequency corrected HRTFs of the KEMAR manikin, including Matlab code to perform the correction [EGWS17]
  `http://doi.org/10.5281/zenodo.401041`
- RIRs and BRIRs of a 64-channel loudspeaker array for different room configurations [EGWS15]
  `http://dx.doi.org/10.14279/depositonce-87.6`
- Matlab code for the used implementation of the image source model, cf. section 3.2.2.2
  `http://doi.org/10.5281/zenodo.3745990`
- stimuli, results and scripts for the statistical analysis of the evaluation experiment for the localisation reporting method [EFS19]
  `http://doi.org/10.5281/zenodo.3520127`
- stimuli, results and scripts for the statistical analysis of the localisation experiment [ES20]
  `http://doi.org/10.5281/zenodo.3358956`
- stimuli, results and scripts for the statistical analysis of the colouration experiment
  `http://doi.org/10.5281/zenodo.4036228`
- figures created for this thesis except for those contained in publications
  `http://doi.org/10.5281/zenodo.3745986`

The following Open Science resources provided by other scientists have been used to conduct the research in this thesis:

- Sound Field Synthesis Toolbox [WS12], release 2.5.0
  `http://doi.org/10.5281/zenodo.2597212`
- SoundScape Renderer [GS12], commit 2b11775
  `http://github.com/SoundScapeRenderer/ssr`
- The Auditory Modeling Toolbox [SM13], release 0.9.9
  `http://amtoolbox.sourceforge.net`
- Spatially Oriented Format for Acoustics (SOFA) [AES15]
  `http://www.sofaconventions.org`
- The FABIAN head-related transfer function data base [BLW^{+}17b]
  `http://doi.org/10.14279/depositonce-5718.3`
- Anechoic HRIRs from the KEMAR manikin with different distances [WGRS11]
  `http://doi.org/10.5281/zenodo.55418`

List of abbreviations

2.5D	2½-dimensional
2D	two-dimensional
3D	three-dimensional
ADAM	Audio Descriptive Analysis & Mapping
AIC	Akaike information criterion
ANOVA	analysis of variance
BEM	boundary element method
BIC	Bayesian information criterion
BRIR	binaural room impulse response
DLP	double layer potential
EDT	early decay time
HATS	head and torso simulator
HMD	head-mounted display
HpTF	headphone transfer function
HRIR	head-related impulse response
HRTF	head-related transfer function
IC	interaural coherence
ILD	interaural level difference
ITD	interaural time difference
ITU	International Telecommunication Union
MUSHRA	Multiple Stimulus Test with Hidden Reference and Anchor
NFC-HOA	Near-Field-Compensated Higher Order Ambisonics
KHI	Kirchhoff-Helmholtz integral equation
RGT	Repertory Grid Technique
RIR	single-channel room impulse response
SDM	Spectral Division Method
SFS	sound field synthesis
SLP	single layer potential
SOFA	Spatially Oriented Format for Acoustics
STD	standard deviation
TASP	Two Arc Source Positioning system
TOA	time of arrival
VR	virtual reality
WFS	Wave Field Synthesis

Bibliography

[AAD01] V. R. Algazi, C. Avendano, and R. O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *J. Audio Eng. Soc.*, 49(6):472–479, 2001.

[AAG$^+$13] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely. Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. *J. Acoust. Soc. Am.*, 133(5):2711–2721, 2013.

[AB79] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, 1979.

[ADD$^+$02] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Am.*, 112(5):2053–2064, 2002.

[AES15] AES69-2015: AES standard for file exchange – Spatial acoustic data file format, Audio Engineering Society, Inc., 2015.

[AF92] W. Ahnert and R. Feistel. EARS Auralization Software. In *Proc. of the 93rd Audio Eng. Soc. Convention*, San Francisco, CA, USA, 1992.

[Ahr12] J. Ahrens. *Analytic Methods of Sound Field Synthesis.* Springer, Berlin, Heidelberg, 2012.

[ANS07] ANSI S3.4-2007: Procedure for the Computation of Loudness of Steady Sounds, American National Standards Institute, 2007.

[AS09] J. Ahrens and S. Spors. On the Secondary Source Type Mismatch in Wave Field Synthesis Employing Circular Distributions of Loudspeakers. In *Proc. of the 127th Audio Eng. Soc. Convention*, New York, NY, USA, 2009.

[AS10] J. Ahrens and S. Spors. Sound Field Reproduction using Planar and Linear Arrays of Loudspeakers. *IEEE Trans. Audio, Speech, Language Process.*, 18(8):2038–2050, 2010.

[AW91] L. S. Aiken and S. G. West. *Multiple Regression: Testing and Interpreting Interactions.* Sage, Newbury Park, London, New Delhi, 1991.

[BA05] T. Betlehem and T. D. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *J. Acoust. Soc. Am.*, 117(4):2100–2111, 2005.

[BAA$^+$19] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl. A round robin on room acoustical simulation and auralization. *J. Acoust. Soc. Am.*, 145(4):2746–2760, 2019.

[BAM07] T. Behrens, W. Ahnert, and C. Moldrzyk. Raumakustische Konzeption von Wiedergaberäumen für Wellenfeldsynthese am Beispiel eines Hörsaals der TU Berlin. In *Proc. of the 33rd German Annual Conference on Acoustics (DAGA)*, Stuttgart, Germany, 2007.

[BAP10] T. Betlehem, C. Anderson, and M. A. Poletti. A Directional Loudspeaker Array for Surround Sound in Reverberant Rooms. In *Proc. of the 20th Int. Congress on Acoustics (ICA)*, Sydney, Australia, 2010.

[Bar37] M. S. Bartlett. Properties of Sufficiency and Statistical Tests. *Proc. R. Soc. A.*, 160(901):268–283, 1937.

[Bar95] M. Barron. Interpretation of Early Decay Times in Concert Auditoria. *Acustica*, 81(4):320–331, 1995.

[BCNW16] H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel. Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments. *Acta Acust. united Ac.*, 102(1):107–118, 2016.

[BD06] J. Bortz and N. Döring. *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* Springer, Heidelberg, 4th, revised edition, 2006.

[Ber88] A. J. Berkhout. A Holographic Approach to Acoustic Control. *J. Audio Eng. Soc.*, 36(12):977–995, 1988.

[Ber04] L. Beranek. *Concert Halls and Opera Houses: Music, Acoustics, and Architecture.* Springer, New York, 2nd edition, 2004.

[Ber13] B. Bernschütz. A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100. In *Proc. of the Int. Conference on Acoustics incl. the 39th German Annual Conference on Acoustics (AIA-DAGA)*, Merano, Italy, 2013.

[BEW18] F. Brinkmann, V. Erbes, and S. Weinzierl. Extending the closed form image source model for source directivity. In *Proc. of the 44th German Annual Conference on Acoustics (DAGA)*, Munich, Germany, 2018.

[BL10] F. Brinkmann and A. Lindau. On the effect of individual headphone compensation in binaural synthesis. In *Proc. of the 36th German Annual Conference on Acoustics (DAGA)*, Berlin, Germany, 2010.

[Bla97] J. Blauert. *Spatial Hearing.* MIT Press, Cambridge, revised edition, 1997.

[Blu31] A. D. Blumlein. Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems. British patent number 394,325, 1931.

[BLW$^+$13] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, and S. van de Par. A high resolution head-related transfer function database including different orientations of head above the torso. In *Proc. of the Int. Conference on Acoustics incl. the 39th German Annual Conference on Acoustics (AIA-DAGA)*, Merano, Italy, 2013.

[BLW17a] F. Brinkmann, A. Lindau, and S. Weinzierl. On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.*, 142(4):1784–1795, 2017.

[BLW$^+$17b] F. Brinkmann, A. Lindau, S. Weinzierl, S. van de Par, M. Müller-Trapet, R. Opdam, and M. Vorländer. A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations. *J. Audio Eng. Soc.*, 65(10):841–848, 2017.

[BM81] M. Barron and A. H. Marshall. Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure. *J. Sound Vib.*, 77(2):211–232, 1981.

[BM09] C. Borß and R. Martin. An Improved Parametric Model for Perception-Based Design of Virtual Acoustics. In *Proc. of the 35th Int. Audio Eng. Soc. Conference on Audio for Games*, London, UK, 2009.

[Bor84] J. Borish. Extension of the image model to arbitrary polyhedra. *J. Acoust. Soc. Am.*, 75(6):1827–1836, 1984.

[Bor00] I. Bork. A Comparison of Room Simulation Software – The 2nd Round Robin on Room Acoustical Computer Simulation. *Acta Acust. united Ac.*, 86(6):943–956, 2000.

[Bor05a] I. Bork. Report on the 3rd Round Robin on Room Acoustical Computer Simulation – Part i: Measurements. *Acta Acust. united Ac.*, 91(4):740–752, 2005.

[Bor05b] I. Bork. Report on the 3rd Round Robin on Room Acoustical Computer Simulation – Part ii: Calculations. *Acta Acust. united Ac.*, 91(4):753–763, 2005.

[BR99] J. Berg and F. Rumsey. Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods. In *Proc. of the 16th Int. Audio Eng. Soc. Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[BRLW15] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl. Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology. *IEEE J. Sel. Topics Signal Process.*, 9(5):931–942, 2015.

[Bro95] A. W. Bronkhorst. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98(5):2542–2553, 1995.

[Brü01] M. Brüggen. Coloration and Binaural Decoloration in Natural Environments. *Acta Acust. united Ac.*, 87(3):400–406, 2001.

[BSK05] D. S. Brungart, B. D. Simpson, and A. J. Kordik. The detectability of headtracker latency in virtual audio displays. In *Proc. of the 11th Meeting of the Int. Conference on Auditory Displays (ICAD)*, Limerick, Ireland, 2005.

[BSM$^+$04] D. S. Brungart, B. D. Simpson, R. L. McKinley, A. J. Kordik, R. C. Dallman, and D. A. Ovenshire. The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources. In *Proc. of the 10th Meeting of the Int. Conference on Auditory Display*, Sydney, Australia, 2004.

[CA01] H. Colonius and P. Arndt. A two-stage model for visual-auditory interaction in saccadic latencies. *Percept. Psychophys.*, 63(1):126–147, 2001.

[CJ12] J.-H. Chang and F. Jacobsen. Sound field control with a circular double-layer array of loudspeakers. *J. Acoust. Soc. Am.*, 131(6):4518–4525, 2012.

[CK12] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory.* Springer, New York, Heidelberg, Dordrecht, 3rd edition, 2012.

[CN03] E. Corteel and R. Nicol. Listening room compensation for Wave Field Synthesis. What can be done? In *Proc. of the 23rd Int. Audio Eng. Soc. Conference on Signal Processing in Audio Recording and Reproduction*, Helsingør, Denmark, 2003.

[Coh88] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition, 1988.

[Coh92] J. Cohen. A Power Primer. *Psychol. Bull.*, 112(1):155–159, 1992.

[Con99] W. J. Conover. *Pratical nonparametric statistics.* Wiley, New York, 3rd edition, 1999.

[Cor06] E. Corteel. Equalization in an Extended Area Using Multichannel Inversion and Wave Field Synthesis. *J. Audio Eng. Soc.*, 54(12):1140–1161, 2006.

[Dal95] B.-I. Dalenbäck. *A new model for room acoustic prediction and auralization.* PhD thesis, Chalmers University of Technology, Gothenburg, Sweden, 1995.

[Dan03] J. Daniel. Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format. In *Proc. of the 23rd Int. Audio Eng. Soc. Conference on Signal Processing in Audio Recording and Reproduction*, Helsingør, Denmark, 2003.

[dB04] W. P. J. de Bruijn. *Application of Wave Field Synthesis in Videoconferencing.* PhD thesis, Delft University of Technology, Delft, Netherlands, 2004.

[DEH11] M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Commun.*, 53(5):592–605, 2011.

[DIN08] DIN 15996:2008-05: Bild- und Tonbearbeitung in Film-, Video- und Rundfunkbetrieben – Grundsätze und Festlegungen für den Arbeitsplatz, 2008.

[Dro11] E. A. Drost. Validity and reliability in social science research. *Education Research and Perspectives*, 38(1):105–123, 2011.

[DZG04] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov. Interpolation and range extrapolation of HRTFs. In *Proc. of the Int. Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.

[EFS19] V. Erbes, A. Fleck, and S. Spors. Virtual reality based pointing method for localisation experiments in spatial audio. In *Proc. of the 45th German Annual Conference on Acoustics (DAGA)*, Rostock, Germany, 2019.

[EGWS15] V. Erbes, M. Geier, S. Weinzierl, and S. Spors. Database of single-channel and binaural room impulse responses of a 64-channel loudspeaker array. In *Proc. of the 138th Audio Eng. Soc. Convention*, Warsaw, Poland, 2015.

[EGWS17] V. Erbes, M. Geier, H. Wierstorf, and S. Spors. Free database of low-frequency corrected head-related transfer functions and headphone compensation filters. In *Proc. of the 142nd Audio Eng. Soc. Convention*, Berlin, Germany, 2017.

[ES15] V. Erbes and S. Spors. Analysis of a Spatially Discrete Sound Field Synthesis Array in a Reflective Environment. In *Proc. of the 10th European Congress and Exposition on Noise Control Engineering (EuroNoise)*, Maastricht, Netherlands, 2015.

[ES17] V. Erbes and S. Spors. Influence of the Listening Room on Spectral Properties of Wave Field Synthesis. In *Proc. of the 43rd German Annual Conference on Acoustics (DAGA)*, Kiel, Germany, 2017.

[ES20] V. Erbes and S. Spors. Localisation Properties of Wave Field Synthesis in a Listening Room. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 28(1):1016–1024, 2020.

[ESLW12] V. Erbes, F. Schultz, A. Lindau, and S. Weinzierl. An extraaural headphone system for optimized binaural reproduction. In *Proc. of the 38th German Annual Conference on Acoustics (DAGA)*, Darmstadt, Germany, 2012.

[EWS15] V. Erbes, S. Weinzierl, and S. Spors. Evanescent Aliasing of Virtual Sources close to a Wave Field Synthesis Array. In *Proc. of the 41st German Annual Conference on Acoustics (DAGA)*, Nuremberg, Germany, 2015.

[FBK14] W. H. Finch, J. E. Bolin, and K. Kelley. *Multilevel Modeling Using R*. CRC Press, Boca Raton, 2014.

[FBM17] J. Francombe, T. Brookes, and R. Mason. Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences. *J. Audio Eng. Soc.*, 65(3):198–211, 2017.

[FELB07] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*, 39(2):175–191, 2007.

[FFSS17] G. Firtha, P. Fiala, F. Schultz, and S. Spors. Improved Referencing Schemes for 2.5D Wave Field Synthesis Driving Functions. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 25(5):1117–1127, 2017.

[Fie01] L. D. Fielder. Practical Limits for Room Equalization. In *Proc. of the 111th Audio Eng. Soc. Convention*, New York, NY, USA, 2001.

[FM04] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116(5):3075–3089, 2004.

[FMSZ10] M. Frank, L. Mohr, A. Sontacchi, and F. Zotter. Flexible and intuitive pointing method for 3D auditory localization experiments. In *Proc. of the 38th Int. Audio Eng. Soc. Conference on Sound Quality Evaluation*, Piteå, Sweden, 2010.

[Fri37] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, 32(200):675–701, 1937.

[Gam13] H. Gamper. Head-related transfer function interpolation in azimuth, elevation, and distance. *J. Acoust. Soc. Am.*, 134(6):EL547–EL553, 2013.

[Gar16] G. D. Garson. *Validity & Reliability*. Stastistical Associates Publishing, Asheboro, NC, 2016.

[GB07] P.-A. Gauthier and A. Berry. Objective Evaluation of Room Effects on Wave Field Synthesis. *Acta Acust. united Ac.*, 93(5):824–836, 2007.

[GKR85] D. Guicking, K. Karcher, and M. Rollwage. Coherent active methods for applications in room acoustics. *J. Acoust. Soc. Am.*, 78(4):1426–1434, 1985.

[GLGM13] Y. Guo, H. L. Logan, D. H. Glueck, and K. E. Muller. Selecting a sample size for studies with repeated measures. *BMC Med. Res. Methodol.*, 13(100):1–8, 2013.

[GMMW07] A. Goertz, M. Makarski, C. Moldrzyk, and S. Weinzierl. Entwicklung eines achtkanaligen Lautsprechermoduls für die Wellenfeldsynthese. In *Proc. of the 33rd German Annual Conference on Acoustics (DAGA)*, Stuttgart, Germany, 2007.

[GODZ10] N. A. Gumerov, A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin. Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation. *J. Acoust. Soc. Am.*, 127(1):370–386, 2010.

[GRS01] B. Girod, R. Rabenstein, and A. Stenger. *Signals and Systems.* Wiley, New York, 2001.

[GS12] M. Geier and S. Spors. Spatial Audio with the SoundScape Renderer. In *Proc. of the 27th Tonmeistertagung – VDT International Convention*, Cologne, Germany, 2012.

[Har83] W. M. Hartmann. Localization of sound in rooms. *J. Acoust. Soc. Am.*, 74(5):1380–1391, 1983.

[HBS99] K. Hartung, J. Braasch, and S. J. Sterbing. Comparison of different methods for the interpolation of head-related transfer functions. In *Proc. of the 16th Int. Audio Eng. Soc. Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[HES19] W. Hahne, V. Erbes, and S. Spors. On the Perceptually Acceptable Noise Level in Binaural Room Impulse Responses. In *Proc. of the 45th German Annual Conference on Acoustics (DAGA)*, Rostock, Germany, 2019.

[HRO98] P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal. Relearning sound localization with new ears. *Nat. Neurosci.*, 1(5):417–421, 1998.

[HSM$^+$17] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. G. Katz, and C. de Boishéraud. Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *J. Acoust. Soc. Am.*, 141(3):2011–2023, 2017.

[HvdP15] A. Haeussler and S. van de Par. Spectral and perceptual properties of a transfer chain of two rooms. In *Proc. of the 10th European Congress and Exposition on Noise Control Engineering (EuroNoise)*, Maastricht, Netherlands, 2015.

[HWS16] N. Hahn, F. Winter, and S. Spors. Local Wave Field Synthesis by Spatial Band-limitation in the Circular/Spherical Harmonics Domain. In *Proc. of the 140th Audio Eng. Soc. Convention*, Paris, France, 2016.

[IEC09] IEC 60318-1:2009: Electroacoustics – Simulators of human head and ear – Part 1: Ear simulator for the measurement of supra-aural and circumaural earphones, International Electrotechnical Commission, 2009.

[IM01] M. Z. Ikram and D. R. Morgan. A multiresolution approach to blind separation of speech signals in a reverberant environment. In *Proc. of the IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001.

[ISO03] ISO 354: Acoustics – Measurement of sound absorption in a reverberation room, International Organization for Standardization, 2003.

[ISO09] ISO 3382-1: Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces, International Organization for Standardization, 2009.

[ITU12] Recommendation ITU-R BS.775-3: Multichannel stereophonic sound system with and without accompanying picture, International Telecommunication Union, 2012.

[ITU15] Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems, International Telecommunication Union, 2015.

[KBJvW14] N. Kaplanis, S. Bech, S. H. Jensen, and T. van Waterschoot. Perception of Reverberation in Small Rooms: A Literature Study. In *Proc. of the 55th Int. Audio Eng. Soc. Conference on Spatial Audio*, Helsinki, Finland, 2014.

[Kel55] G. A. Kelly. *The Psychology of Personal Constructs.* Norton, New York, 1955.

[KN93] O. Kirkeby and P. A. Nelson. Reproduction of plane wave sound fields. *J. Acoust. Soc. Am.*, 94(5):2992–3000, 1993.

[KN99] O. Kirkeby and P. A. Nelson. Digital Filter Design for Inversion Problems in Sound Reproduction. *J. Audio Eng. Soc.*, 47(7/8):583–595, 1999.

[KNHOB98] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante. Fast Deconvolution of Multichannel Systems Using Regularization. *IEEE Trans. Speech Audio Process*, 6(2):189–194, 1998.

[KRM11] B. M. King, P. J. Rosopa, and E. W. Minium. *Statistical Reasoning in the Behavioral Sciences.* Wiley, Hoboken, NJ, 6th edition, 2011.

[KS03] B. Klehs and T. Sporer. Wave Field Synthesis in the Real World: Part 1 – In the Living Room. In *Proc. of the 114th Audio Eng. Soc. Convention*, Amsterdam, Netherlands, 2003.

[KSS68] A. Krokstad, S. Strøm, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *J. Sound Vib.*, 1968.

[Kut03] H. Kutschbach. Verification for spatial sound systems. In *Proc. of the 24th Int. Audio Eng. Soc. Conference on Multichannel Audio, The New Reality*, Banff, Canada, 2003.

[Kut17] H. Kuttruff. *Room Acoustics.* CRC Press, Boca Raton, 6th edition, 2017.

[KV07] P. H. Kvam and B. Vidakovic. *Nonparametric Statistics with Applications to Science and Engineering.* Wiley, Hoboken, NJ, 2007.

[KZ01] K. Koivuniemi and N. Zacharov. Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. In *Proc. of the 111th Audio Eng. Soc. Convention*, New York, NY, USA, 2001.

[LCYG99] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *J. Acoust. Soc. Am.*, 106(4):1633–1654, 1999.

[LDE00] J. Lewald, G. J. Dörrscheidt, and W. H. Ehrenstein. Sound localization with eccentric head position. *Behav. Brain Res.*, 108(2):105–125, 2000.

[LEL$^+$14] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl. A Spatial Audio Quality Inventory (SAQI). *Acta Acust. united Ac.*, 100(5):984–994, 2014.

[LI04] K. M. Li and K. K. Iu. Propagation of sound in long enclosures. *J. Acoust. Soc. Am.*, 116(5):2759–2770, 2004.

[Lin09] A. Lindau. The Perception of System Latency in Dynamic Binaural Synthesis. In *Proc. of the Int. Conference on Acoustics (NAG/DAGA)*, Rotterdam, Netherlands, 2009.

[LMW08] A. Lindau, H.-J. Maempel, and S. Weinzierl. Minimum BRIR grid resolution for dynamic binaural synthesis. In *Proc. of the Acoustics '08*, Paris, France, 2008.

[Lor18] J. Lorah. Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large-scale Assess, Educ.*, 6(8):1–11, 2018.

[LSK$^+$07] J. Liebetrau, T. Sporer, T. Korn, K. Kunze, C. Mank, D. Marquard, T. Metheja, S. Mauer, T. Mayenfels, R. Möller, M.-A. Schnabel, B. Slobbe, and A. Ueberschaer. Localization in Spatial Audio – from Wave Field Synthesis to 22.2. In *Proc. of the 123rd Audio Eng. Soc. Convention*, New York, NY, USA, 2007.

[LW09] A. Lindau and S. Weinzierl. FABIAN – an instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom. In *Proc. of the 24th Tonmeistertagung – VDT International Convention*, Leipzig, Germany, 2009.

[LW12] A. Lindau and S. Weinzierl. Assessing the Plausibility of Virtual Acoustics Environments. *Acta Acust. united Ac.*, 98(5):804–810, 2012.

[Mac04] P. Mackensen. *Auditive Localization. Head movements, an additional cue in Localization.* PhD thesis, Technische Universität Berlin, Berlin, Germany, 2004.

[Mas51] F. J. Massey. The kolmogorov-smirnov test for goodness of fit. *J. Am. Stat. Assoc.*, 46(253):68–78, 1951.

[Mas12] B. S. Masiero. *Individualized Binaural Technology.* PhD thesis, RWTH Aachen University, Aachen, Germany, 2012.

[Mau40] J. W. Mauchly. Significance test for sphericity of a normal n-variate distribution. *Ann. Math. Statist.*, 11(2):204–209, 1940.

[McC65] R. L. McCornack. Extended tables of the Wilcoxon matched pair signed rank statistic. *J. Am. Stat. Assoc.*, 60(311):864–871, 1965.

[MDM18] C. Mendonça and S. Delikaris-Manias. Statistical tests with MUSHRA data. In *Proc. of the 144th Audio Eng. Soc. Convention*, Milan, Italy, 2018.

[MFV11] B. S. Masiero, J. Fels, and M. Vorländer. Review of the crosstalk cancellation filter technique. In *Proc. of the Int. Conference on Spatial Audio (ICSA)*, Detmold, Germany, 2011.

[MGB97] B. C. J. Moore, B. R. Glasberg, and T. Baer. A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. *J. Audio Eng. Soc.*, 45(4):224–240, 1997.

[MGL10] P. Majdak, M. J. Goupell, and B. Laback. 3-D Localization of Virtual Sound Sources: Effects of Visual Environment, Pointing Method, and Training. *Atten. Percept. Psychophys.*, 72(2):454–469, 2010.

[MGM$^+$07] C. Moldrzyk, A. Goertz, M. Makarski, S. Feistel, W. Ahnert, and S. Weinzierl. Wellenfeldsynthese für einen großen Hörsaal. In *Proc. of the 33rd German Annual Conference on Acoustics (DAGA)*, Stuttgart, Germany, 2007.

[MHJS95] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen. Transfer Characteristics of Headphones Measured on Human Ears. *J. Audio Eng. Soc.*, 1995.

[Mil58] A. W. Mills. On the minimum audible angle. *J. Acoust. Soc. Am.*, 30(4):237–246, 1958.

[MOCM01] P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller. The importance of head movements for binaural room synthesis. In *Proc. of the 7th Int. Conference on Auditory Display (ICAD)*, Espoo, Finland, 2001.

[Møl92] H. Møller. Fundamentals of Binaural Technology. *Appl. Acoust.*, 36(3–4):171–218, 1992.

[Mös09] M. Möser. *Engineering Acoustics. An Introduction to Noise Control.* Springer, Dordrecht, Heidelberg, London, 2nd edition, 2009.

[MPC05] P. Minnaar, J. Plogsties, and F. Christensen. Directional Resolution of Head-Related Transfer Functions Required in Binaural Synthesis. *J. Audio Eng. Soc.*, 53(10):919–929, 2005.

[MSJH96] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammshøi. Binaural Technique: Do We Need Individual Recordings? *J. Audio Eng. Soc.*, 44(6):451–469, 1996.

[MSW13] S. Müller, J. L. Scealy, and A. H. Welsh. Model Selection in Linear Mixed Models. *Stat. Sci.*, 28(2):135–167, 2013.

[Nay93] G. M. Naylor. ODEON – Another hybrid room acoustical model. *Appl. Acoust.*, 38(2–4):131–143, 1993.

[NJNN18] A. Neidhardt, K.-P. Jurgeit, A. Nasrollahnejad, and J. Nowak. Investigating continuous adaptation of binaural reproduction to changing listener position. In *Proc. of the 44th German Annual Conference on Acoustics (DAGA)*, Munich, Germany, 2018.

[NSL04] S. G. Norcross, G. A. Soulodre, and M. C. Lavoie. Distortion Audibility in Inverse Filtering. In *Proc. of the 117th Audio Eng. Soc. Convention*, San Francisco, CA, USA, 2004.

[Oli04] S. E. Olive. A Multiple Regression Model For Predicting Loudspeaker Preference Using Objective Measurements: Part i – Listening Test Results. In *Proc. of the 116th Audio Eng. Soc. Convention*, Berlin, Germany, 2004.

[OST57] C. E. Osgood, G. J. Suci, and P. Tannenbaum. *The measurement of meaning.* University of Illinois Press, Urbana, 1957.

[Ott01] J. Otten. *Factors influencing acoustical localization.* PhD thesis, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany, 2001.

[OWM07] S. E. Olive, T. Welti, and W. L. Martens. Listener Loudspeaker Preference Ratings Obtained in situ Match Those Obtained via a Binaural Room Scanning Measurement and Playback System. In *Proc. of the 122nd Audio Eng. Soc. Convention*, Vienna, Austria, 2007.

[PAS12] M. A. Poletti, T. D. Abhayapala, and P. Samarasinghe. Interior and exterior sound field control using two dimensional higher-oder variable-directivity sources. *J. Acoust. Soc. Am.*, 131(5):3814–3823, 2012.

[PAV11] S. Pelzer, M. Aretz, and M. Vorländer. Quality assessment of room acoustic simulation tools comparing binaural measurements and simulations in an optimized test scenario. In *Poc. of the 6th Forum Acusticum*, Aalborg, Denmark, 2011.

[PBW04] K. J. Palomäki, G. J. Brown, and D. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Commun.*, 43(4):361–378, 2004.

[PKV14] S. Pelzer, M. Kohnen, and M. Vorländer. Evaluation of Loudspeaker-based 3D Room Auralizations using Hybrid Reproduction Techniques. In *Proc. of the 40th German Annual Conference on Acoustics (DAGA)*, Oldenburg, Germany, 2014.

[Pol05] M. A. Poletti. Three-Dimensional Surround Sound Systems Based on Spherical Harmonics. *J. Audio Eng. Soc.*, 53(11):1004–1025, 2005.

[PSR05] S. Petrausch, S. Spors, and R. Rabenstein. Simulation and Visualization of Room Compensation for Wave Field Synthesis with the Functional Transformation Method. In *Proc. of the 119th Audio Eng. Soc. Convention*, New York, NY, USA, 2005.

[Pul97] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, 45(6):456–466, 1997.

[Pul01] V. Pulkki. Coloration of Amplitude-Panned Virtual Sources. In *Proc. of the 110th Audio Eng. Soc. Convention*, Amsterdam, Netherlands, 2001.

[QvdB04] H. Quené and H. van den Bergh. On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Commun.*, 43(1–2):103–121, 2004.

[R C17] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, URL: `http://www.R-project.org`, 2017.

[RBW95] P. R. Runkle, M. A. Blommer, and G. H. Wakefield. A comparison of head related transfer function interpolation methods. In *Proc. of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 1995.

[Rec09] G. H. Recanzone. Interactions of Auditory and Visual Stimuli in Space and Time. *Hear. Res.*, 258(1–2):89–99, 2009.

[RJ03] P. Rubak and L. G. Johansen. Coloration in Natural and Artificial Room Impulse Responses. In *Proc. of the 23rd Int. Audio Eng. Soc. Conference on Signal Processing in Audio Recording and Reproduction*, Helsingør, Denmark, 2003.

[Rum01] F. Rumsey. *Spatial Audio.* Focal Press, Burlington, MA, Abingdon, 2001.

[Rum02] F. Rumsey. Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *J. Audio Eng. Soc.*, 50(9):651–666, 2002.

[Rum08] F. Rumsey. Loudspeakers, Reflections, and Rooms. *J. Audio Eng. Soc.*, 56(5):394–400, 2008.

[RZKB05] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *J. Acoust. Soc. Am.*, 118(2):968–976, 2005.

[SA09] S. Spors and J. Ahrens. Spatial Sampling Artifacts of Wave Field Synthesis for the Reproduction of Virtual Point Sources. In *Proc. of the 126th Audio Eng. Soc. Convention*, Munich, Germany, 2009.

[SA10a] S. Spors and J. Ahrens. Analysis and Improvement of Pre-equalization in 2.5-Dimensional Wave Field Synthesis. In *Proc. of the 128th Audio Eng. Soc. Convention*, London, UK, 2010.

[SA10b] S. Spors and J. Ahrens. Local Sound Field Synthesis by Virtual Secondary Sources. In *Proc. of the 40th Int. Audio Eng. Soc. Conference on Spatial Audio: Sense the Sound of Space*, Tokyo, Japan, 2010.

[San96] J. Sandvad. Dynamic Aspects of Auditory Virtual Environments. In *Proc. of the 100th Audio Eng. Soc. Convention*, Copenhagen, Denmark, 1996.

[Sav10] L. Savioja. Real-time 3D finite-difference time-domain simulation of low- and mid-frequency room acoustics. In *Proc. of the 13th Int. Conference on Digital Audio Effects*, Graz, Austria, 2010.

[SB95] G. A. Soulodre and J. S. Bradley. Subjective evaluation of new room acoustic measures. *J. Acoust. Soc. Am.*, 98(1):294–301, 1995.

[Sch87] M. R. Schroeder. Statistical Parameters of the Frequency Response Curves of Large Rooms. *J. Audio Eng. Soc.*, 35(5):299–306, 1987.

[Sch11] D. Schröder. *Physically Based Real-Time Auralization of Interactive Virtual Environments.* PhD thesis, RWTH Aachen University, Aachen, Germany, 2011.

[Sch16] F. Schultz. *Sound Field Synthesis for Line Source Array Applications in Large-Scale Sound Reinforcement.* PhD thesis, University of Rostock, Rostock, Germany, 2016.

[See03] B. Seeber. *Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode.* PhD thesis, Technical University of Munich, Munich, Germany, 2003.

[SESW13] F. Schultz, V. Erbes, S. Spors, and S. Weinzierl. Derivation of IIR prefilters for soundfield synthesis using linear secondary source distributions. In *Proc. of the Int. Conference on Acoustics incl. the 39th German Annual Conference on Acoustics (AIA-DAGA)*, Merano, Italy, 2013.

[SFTW$^+$19] J. Seebacher, A. Franke-Trieger, V. Weichbold, P. Zorowka, and K. Stephan. Improved interaural timing of acoustic nerve stimulation affects sound localization in single-sided deaf cochlear implant users. *Hear. Res.*, 258(1–2):89–99, 2019.

[SGW17] S. Spors, M. Geier, and H. Wierstorf. Towards Open Science in Acoustics: Foundations and Best Practices. In *Proc. of the 43rd German Annual Conference on Acoustics (DAGA)*, Kiel, Germany, 2017.

[Sib81] R. Sibson. *A brief description of natural neighbor interpolation*, chapter in: Interpreting Multivariate Data, edited by V. Barnett, pages 21–36. Wiley, Chicester, 1981.

[SK62] M. R. Schroeder and K. H. Kuttruff. On Frequency Response Curves in Rooms. Comparison of Experimental, Theoretical, and Monte Carlo Results for the Average Frequency Spacing between Maxima. *J. Acoust. Soc. Am.*, 34(1):76–80, 1962.

[SK02] U. P. Svensson and U. R. Kristiansen. Computational modelling and simulation of acoustic spaces. In *Proc. of the 22nd Int. Audio Eng. Soc. Conference on Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland, 2002.

[SK04] T. Sporer and B. Klehs. Wave Field Synthesis in the Real World: Part 2 – In the Movie Theatre. In *Proc. of the 116th Audio Eng. Soc. Convention*, Berlin, Germany, 2004.

[SK12] M. Schneider and W. Kellermann. Adaptive listening room equalization using a scalable filtering structure in the wave domain. In *Proc. of the 37th Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[SL09] Z. Schärer and A. Lindau. Evaluation of Equalization Methods for Binaural Signals. In *Proc. of the 126th Audio Eng. Soc. Convention*, Munich, Germany, 2009.

[SLS09] T. Sporer, J. Liebetrau, and S. Schneider. Statistics of MUSHRA revisited. In *Proc. of the 127th Audio Eng. Soc. Convention*, New York, NY, USA, 2009.

[SM13] P. Søndergaard and P. Majdak. *The Auditory Modeling Toolbox*, chapter in: The Technology of Binaural Hearing, edited by J. Blauert, pages 33–56. Springer, Berlin, Heidelberg, 2013.

[Sni05] T. A. B. Snijders. *Power and Sample Size in Multilevel Linear Models*, volume 3, chapter in: Encyclopedia of Statistics in Behavioral Science, edited by B. Everitt and D. Howell, pages 1570–1573. Wiley, Chicester, 2005.

[Spo06] S. Spors. *Active Listening Room Compensation for Spatial Sound Reproduction Systems.* PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, 2006.

[SR06] S. Spors and R. Rabenstein. Spatial Aliasing Artifacts Produced by Linear and Circular Loudspeaker Arrays used for Wave Field Synthesis. In *Proc. of the 120th Audio Eng. Soc. Convention*, Paris, France, 2006.

[SRA08] S. Spors, R. Rabenstein, and J. Ahrens. The Theory of Wave Field Synthesis Revisited. In *Proc. of the 124th Audio Eng. Soc. Convention*, Amsterdan, Netherlands, 2008.

[SRdV97] E. W. Start, M. S. Roovers, and D. de Vries. In Situ Measurements on a Wave Field Synthesis System for Sound Enhancement. In *Proc. of the 102nd Audio Eng. Soc. Convention*, Munich, Germany, 1997.

[SRR05] S. Spors, M. Renk, and R. Rabenstein. Limiting Effects of Active Room Compensation using Wave Field Synthesis. In *Proc. of the 118th Audio Eng. Soc. Convention*, Barcelona, Spain, 2005.

[SS90] D. W. Stewart and P. N. Shamdasani. *Focus Groups: Theory and Practice.* Sage, Newbury Park, 1990.

[SSJW10] S. Spors, R. Schleicher, D. Jahn, and R. Walter. On the Use of Eye movement in Acoustic Source Localization Experiments. In *Proc. of the 36th German Annual Conference on Acoustics (DAGA)*, Berlin, Germany, 2010.

[Sta97] E. W. Start. *Direct sound enhancement by wave field synthesis.* PhD thesis, Delft University of Technology, Delft, Netherlands, 1997.

[SWA$^+$14] M. Schoeffler, S. Westphal, A. Adami, H. Bayerlein, and J. Herre. Comparison of a 2D- and 3D-based graphical user interface for localization listening test. In *Proc. of the EAA Joint Symp. on Auralization and Ambisonics*, Berlin, Germany, 2014.

[Too06] F. E. Toole. Loudspeakers and Rooms for Sound Reproduction – A Scientific Review. *J. Audio Eng. Soc.*, 54(6):451–476, 2006.

[Too08] F. E. Toole. *Sound Reproduction. Loudspeakers and Rooms.* Focal Press, Burlington, Oxford, 2008.

[UMW04] J. Usher, W. L. Martens, and W. Woszczyk. The influence of the presence of multiple sources on auditory spatial imagery as indicated by a graphical response technique. In *Proc. of the 18th Int. Congress on Acoustics (ICA)*, Kyoto, Japan, 2004.

[Ver97] E. Verheijen. *Sound Reproduction by Wave Field Synthesis.* PhD thesis, Delft University of Technology, Delft, Netherlands, 1997.

[Vog93] P. Vogel. *Application of Wave Field Synthesis in room acoustics.* PhD thesis, Delft University of Technology, Delft, Netherlands, 1993.

[Vor95] M. Vorländer. International round robin on room acoustical computer simulations. In *Proc. of the 15th Int. Congress on Acoustics*, Trondheim, Norway, 1995.

[Vor98] M. Vorländer. Objective Characterization of Sound Fields in Small Rooms. In *Proc. of the 15th Int. Audio Eng. Soc. Conference on Audio, Acoustics & Small Spaces*, Copenhagen, Denmark, 1998.

[Vor08] M. Vorländer. *Auralization.* Springer, Berlin, 2008.

[Vor13] M. Vorländer. Computer simulations in room acoustics: Concepts and uncertainties. *J. Acoust. Soc. Am.*, 133(3):1203–1213, 2013.

[WAKW93] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94(1):111–123, 1993.

[Wal40] H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *J. Exp. Psychol.*, 27(4):339–368, 1940.

[Wen01] E. M. Wenzel. Effect of increasing system latency on localization of virtual sounds with short and long duration. In *Proc. of the 7th Int. Conference on Auditory Display (ICAD)*, Espoo, Finland, 2001.

[WER15] H. Wierstorf, C. Ende, and A. Raake. Klangverfärbung in der Wellenfeldsynthese – Experimente und Modellierung. In *Proc. of the 41st German Annual Conference on Acoustics (DAGA)*, Nuremberg, Germany, 2015.

[WFSS19] F. Winter, G. Firtha, F. Schultz, and S. Spors. A Geometric Model for Prediction of Spatial Aliasing in 2.5D Sound Field Synthesis. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 27(6):1031–1046, 2019.

[WGRS11] H. Wierstorf, M. Geier, A. Raake, and S. Spors. A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. In *Proc. of the 130 Audio Eng. Soc. Convention*, London, UK, 2011.

[WGS10] H. Wierstorf, M. Geier, and S. Spors. Reducing Artifacts of Focused Sources in Wave Field Synthesis. In *Proc. of the 129th Audio Eng. Soc. Convention*, San Francisco, CA, USA, 2010.

[WHR18] H. Wierstorf, C. Hold, and A. Raake. Listener preference for wave field synthesis, stereophony, and different mixes in popular music. *J. Audio Eng. Soc.*, 66(5):385–396, 2018.

[WHSR14] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake. Coloration in Wave Field Synthesis. In *Proc. of the 55th Int. Audio Eng. Soc. Conference on Spatial Audio*, Helsinki, Finland, 2014.

[Wie14] H. Wierstorf. *Perceptual Assessment of Sound Field Synthesis.* PhD thesis, Technische Universität Berlin, Berlin, Germany, 2014.

[Wil45] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bull.*, 1(6):80–83, 1945.

[Wil99] E. G. Williams. *Fourier Acoustics. Sound Radiation and Nearfield Acoustical Holography.* Academic Press, London, San Diego, 1999.

[Win19] F. Winter. *Local sound field synthesis.* PhD thesis, University of Rostock, Rostock, Germany, 2019.

[Wit07] H. Wittek. *Perceptual differences between wavefield synthesis and stereophony.* PhD thesis, University of Surrey, Surrey, UK, 2007.

[WK92] F. L. Wightman and D. J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, 91(3):1648–1661, 1992.

[WK99] F. L. Wightman and D. J. Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.*, 105(5):2841–2853, 1999.

[WK14] S. Werner and F. Klein. Influence of context dependent quality parameters on the perception of externalization and direction of an auditory event. In *Proc. of the 55th Int. Audio Eng. Soc. Conference on Spatial Audio*, Helsinki, Finland, 2014.

[WLA18] S. Weinzierl, S. Lepa, and D. Ackermann. A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI). *J. Acoust. Soc. Am.*, 144(3):1245–1257, 2018.

[Woo77] J. G. Woodward. Quadraphony – A Review. *J. Audio Eng. Soc.*, 25(10/11):843–854, 1977.

[WS12] H. Wierstorf and S. Spors. Sound Field Synthesis Toolbox. In *Proc. of the 132nd Audio Eng. Soc. Convention*, Budapest, Hungary, 2012.

[WSR12a] H. Wierstorf, S. Spors, and A. Raake. Perception and evaluation of sound fields. In *Proc. of the 59th Open Seminar on Acoustics*, Boszkowo, Poland, 2012.

[WSR12b] H. Wierstorf, S. Spors, and A. Raake. Psychoakustik der Wellenfeldsynthese: Vor- und Nachteile binauraler Simulation. In *Proc. of the 38th German Annual Conference on Acoustics (DAGA)*, Darmstadt, Germany, 2012.

[WSS14] F. Winter, F. Schultz, and S. Spors. Localization Properties of Data-based Binaural Synthesis including Translatory Head-Movements. In *Proc. of the 7th Forum Acusticum*, Kraków, Poland, 2014.

[WWH$^+$18] F. Winter, H. Wierstorf, C. Hold, F. Krüger, A. Raake, and S. Spors. Colouration in Local Wave Field Synthesis. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 26(10):1913–1924, 2018.

[WWS17] F. Winter, H. Wierstorf, and S. Spors. Improvement of the reporting method for closed-loop human localization experiments. In *Proc. of the 142nd Audio Eng. Soc. Convention*, Berlin, Germany, 2017.

[Xie09] B. Xie. On the low frequency characteristics of head-related transfer function. *Chin. J. Acoust.*, 28(2):1–13, 2009.

[YIS06] S. Yairi, Y. Iwaya, and Y. Suzuki. Investigation of system latency detection threshold of virtual auditory display. In *Proc. of the 12th Int. Conference on Auditory Display (ICAD)*, London, UK, 2006.

[ZF05] E. Zwicker and H. Fastl. *Psychoachoustics. Facts and Models.* Springer, Berlin, Heidelberg, 2005.

[ZHHR07] S. Zieliński, P. Hardisty, C. Hummersone, and F. Rumsey. Potential Biases in MUSHRA Listening Tests. In *Proc. of the 123rd Audio Eng. Soc. Convention*, New York, NY, USA, 2007.

[Zur79] P. M. Zurek. Measurements of binaural echo suppression. *J. Acoust. Soc. Am.*, 66(6):1750–1757, 1979.

Abstract

This thesis investigates the influence of the listening room on sound fields synthesised by Wave Field Synthesis. The theory and the practical limitations of Wave Field Synthesis are reviewed and the state of research concerning the human perception of Wave Field Synthesis is summarised. With a detailed analysis of the published literature, methods are developed that allow for investigation of the spatial and timbral perception of Wave Field Synthesis in a reverberant environment using listening experiments based on simulation by binaural synthesis and room acoustical simulation. For the evaluation of localisation, a reporting method is developed and successfully tested for its accuracy. The results of the listening experiment on localisation confirm for reverberant environments that accurate localisation in Wave Field Synthesis is only possible for small secondary source distances and that reflections slightly increase the localisation blur. For the perception of timbre, it is shown that the typical Wave Field Synthesis colouration caused by spatial aliasing is alleviated by reflections from the listening room. The results can serve as guidelines for the design of listening rooms for Wave Field Synthesis.

Zusammenfassung

Diese Dissertation untersucht den Einfluss des Wiedergaberaums auf Schallfelder, die mit Wellenfeldsynthese synthetisiert werden. Die Theorie sowie Einschränkungen in der Praxis für Wellenfeldsynthese werden aufgearbeitet und der Forschungsstand zur menschlichen Wahrnehmung von Wellenfeldsynthese wird zusammengefasst. Aufbauend auf einer detaillierten Analyse der veröffentlichten Literatur werden Methoden zur Untersuchung von räumlicher und klangfarblicher Wahrnehmung von Wellenfeldsynthese in einer reflektierenden Umgebung mittels Hörversuchen entwickelt, die auf Simulation mit Binauralsynthese und raumakustischer Simulation beruhen. Für die Evaluierung von Lokalisation wird eine Anzeigemethode entwickelt und erfolgreich auf ihre Genauigkeit getestet. Die Ergebnisse des Hörversuchs zur Lokalisation bestätigen für hallige Umgebungen, dass akkurate Lokalisation in der Wellenfeldsynthese nur bei kleinen Sekundärquellenabständen möglich ist und dass Reflexionen zu einer leichten Erhöhung der Lokalisationsunschärfe führen. Für die Wahrnehmung von Klangfarbe wird gezeigt, dass die typische Klangverfärbung der Wellenfeldsynthese, die durch räumliches Aliasing verursacht wird, durch die Reflexionen des Wiedergaberaums abgeschwächt wird. Die Ergebnisse können als Richtlinien zur Gestaltung von Wiedergaberäumen für Wellenfeldsynthese dienen.

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich diese der Universität Rostock vorgelegte Dissertation mit dem Titel „Wave Field Synthesis in a listening room“ selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die in den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Vera Erbes
Berlin, 11.04.2020 & Berlin, 24.09.2020

Curriculum vitae

Dipl-Ing. Vera Erbes

Date of birth: 23rd of February 1984

Place of birth: Karlsruhe, Germany

Academic career

2015–2019	Research associate at University of Rostock, Germany Institute of Communications Engineering Signal Theory and Digital Signal Processing Group
2011–2015	Research associate at Technische Universität Berlin, Germany Institut für Sprache und Kommunikation Audio Communication Group

University education

2004–2011	Dipl.-Ing. in industrial engineering with specialisation in communications engineering and electronics at Technische Universität Berlin, Germany

School education

1990–2003	Elementary and secondary school in Karlsruhe, Germany (graduation: Abitur)