# Synthetic reproduction of head-related transfer functions by using microphone arrays

**Eugen Rasumow**

# Synthetic reproduction of head-related transfer functions by using microphone arrays

Der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von
**Eugen Rasumow**
geboren am 5. März 1983
in Schachtinsk (Kasachstan)

Betreuer: Prof. Dr. ir. Simon Doclo, Prof. Dr.-Ing. Matthias Blau

Erstgutachter: Prof. Dr. ir. Simon Doclo
Zweitgutachter: Prof. Dr.-Ing. Matthias Blau
Externe Gutachterin: Prof. Dorte Hammershøi, Ph.D

Tag der Disputation: 09.03.2015

# Glossary

## Mathematical notation

| | |
|---|---|
| $a$ | Scalar $a$ |
| $\mathbf{a}$ | Vector $\mathbf{a}$ |
| $\mathbf{A}$ | Matrix $\mathbf{A}$ |
| $a^*$ | Complex conjugate of scalar $a$ |
| $\mathbf{a}^T$ | Transpose of vector $\mathbf{a}$ |
| $\mathbf{a}^H$ | Hermitian transpose of vector $\mathbf{a}$ |
| $\mathbf{A}^{-1}$ | Inverse of matrix $\mathbf{A}$ |
| $\otimes$ | Convolution |
| $\mathcal{F}\{\cdot\}$ | Fourier transform |
| $\mathcal{F}^{-1}\{\cdot\}$ | Inverse Fourier transform |
| $\Im\{\cdot\}$ | Imaginary part |
| $\Re\{\cdot\}$ | Real part |
| $\log_{10}(\cdot)$ | Logarithm with base 10 (decadic logarithm) |
| $\ln(\cdot)$ | Logarithm with base $e$ (natural logarithm) |

# Fixed symbols

| | |
|---|---|
| $\mathrm{B}$ | Number of considered frequency bins for the synthesis error measure $\bar{\epsilon}_f$ in equation 4.4 |
| $\mathrm{B}_{\mathrm{W}}$ | Bandwidth for smoothing in the spatial domain |
| $c$ | Speed of sound propagation |
| $\mathbf{d}(f,\theta)$ | Complex-valued $\mathrm{N} \times 1$-dimensional steering vector, containing the transfer functions between a source from direction $\theta$ and the N microphones for frequency $f$ |
| $\widetilde{\mathbf{d}}\,(f,\theta)$ | Disturbed steering vector according to equation 3.24 |
| $d'$ | Sensitivity index used in signal detection theory |
| $d_n$ | Distance between $n$-th microphone and center of microphone array |
| $\mathrm{D}(f,\theta)$ | Desired spatial directivity pattern of beamformer for frequency $f$ and direction $\theta$ |
| $e$ | Euler's number approximately equal to 2.71828 |
| $f$ | Frequency |
| $f_c$ | Cutoff-frequency for phase substitution at higher frequencies |
| $f_s$ | Sampling frequency |
| $\mathrm{F}(f,\theta)$ | Weighting function for the used cost function |
| $g(f,\theta)$ | Weighting function for the white noise gain (cf. equation 3.7) |
| $g_w^t(f,\theta)$ | Alternative weighting function for the white noise gain (cf. equation 3.22) |
| $G$ | Order of an one-dimensional Golomb series |
| $\mathrm{H}(f,\theta)$ | Resulting spatial directivity pattern of a beamformer for frequency $f$ and direction $\theta$ |

| | |
|---|---|
| $\mathbf{I}_{\mathrm{N}}$ | N × N-dimensional unity matrix |
| $\mathrm{J}_{\mathrm{K}}$ | Least squares cost function for a joint optimization for multiple sets of steering vectors |
| $\mathrm{J}_{\mathrm{LS}}$ | Least squares cost function |
| $\mathrm{J}_{\mathrm{m}}$ | Least squares cost function with a constraint on the mean white noise gain for all considered directions |
| $\mathrm{J}_{\mathrm{v}}$ | Least squares cost function incorporating multiple frequencies in $\boldsymbol{\Omega}^t$ |
| $\mathrm{J}_{\mathrm{vm}}$ | Least squares cost function incorporating multiple frequencies in $\boldsymbol{\Omega}^t$ and a constraint on the mean white noise gain over all considered frequencies |
| $\mathrm{J}_{\mathrm{sd}}$ | Least squares cost function with a constraint on the white noise gain for a single direction |
| K | Number of considered sets of steering vectors within a joint optimization (cf. equation 3.1) |
| $l_{hrir}$ | Length of the head-related impulse responses in samples |
| $l_{win}$ | Length of the tapered Hann window (descending flank) in samples |
| L | Number of frequency bins within a critical band |
| $\mathrm{L}^t$ | Number of frequency bins in frequency vector $\boldsymbol{\Omega}^t$ for the $t$-th band |
| $L_\phi$ | Mixing gain weighting different phases in equation 5.1 |
| $L_{mod}$ | Model for the individual IPD-discrimination ability in equation 5.4 |
| N | Number of microphones |
| $p_{l/r}(t,\theta)$ | Sound pressure signals arriving at the left and the right ear for the azimuthal direction $\theta$ and time $t$ |

| | |
|---|---|
| $p_0(t,\theta)$ | Sound pressure signal at the center of the head when the head is absent for the azimuthal direction $\theta$ and time $t$ |
| P | Number of directions |
| $\mathbf{r}$ | Random disturbance vector (cf. equation 3.24) |
| $S(f,\theta)$ | Source signal for frequency $f$ and direction $\theta$ |
| $s_{l/r}^{hp}(t)$ | Electrical signal driving the headphone for the left and right ear |
| $t$ | Time variable |
| $w_n(f)$ | Complex-valued filter coefficient for the $n$-th microphone and frequency $f$ |
| $\mathbf{w}(f)$ | Complex-valued $\mathrm{N} \times 1$-dimensional vector, containing all N filter coefficients for frequency $f$ |
| $Y_\mathrm{n}(f,\theta)$ | $n$-th microphone signal for frequency $f$ and direction $\theta$ |
| $\overline{Y}(f,\theta)$ | Signal for frequency $f$ and direction $\theta$ arriving at the center of the array |
| $\mathbf{Y}(f,\theta)$ | $\mathrm{N} \times 1$-dimensional vector containing microphone signals for the N microphones |
| $\mathrm{Z}(f,\theta)$ | Output signal of the microphone array for frequency $f$ and direction $\theta$ |

## Greek letters

| | |
|---|---|
| $\alpha_\mathrm{inv}$ | Regularization parameter for inversion of HPTFs |
| $\beta$ | Minimum desired value for $\mathrm{WNG}(\mathbf{w}(f),\theta_d)$ |
| $\beta_\Gamma$ | Maximum desired value for the square norm of the filter coefficients $\Gamma(f)$ |
| $\beta_\mathrm{m}$ | Minimum desired value for $\mathrm{WNG}_\mathrm{m}$ |
| $\beta_\mathrm{v}$ | Minimum desired value for $\mathrm{WNG}_\mathrm{v}$ |

| | |
|---|---|
| $\mathit{\Gamma}(f)$ | Square norm of filter coefficients $\mathbf{w}(f)$ |
| $\delta$ | Elevation angle |
| $\mathit{\Delta}_{\mathrm{dB}}(f,\theta)$ | Absolute dB-error between the synthesized directivity pattern $\mathrm{H}(f,\theta)$ and the desired directivity pattern $\mathrm{D}(f,\theta)$ for frequency $f$ and direction $\theta$ (cf. equation 1.13) |
| $\mathit{\Delta}x_{max}$ | Maximum absolute values for random shift in $x$-direction |
| $\epsilon(f_c,\theta)$ | Synthesis error as a function of the direction $\theta$ and the center frequency $f_c$ of the associated band |
| $\epsilon_f(f_c)$ | Synthesis error averaged over all directions and frequencies within the associated band as a function of its center frequency $f_c$ |
| $\zeta(f)$ | Spatial dynamic range of an HRTF for frequency $f$ in dB (cf. equation 5.3) |
| $\zeta'(f)$ | Reduced spatial dynamic range of an HRTF for frequency $f$ |
| $\theta$ | Azimuthal angle in the horizontal plane |
| $\theta_d$ | Angle for a desired look direction |
| $\mathbf{\Theta}$ | All considered directions $\theta$ |
| $\kappa_{top}$ | Total number of tested microphone topologies |
| $\mu$ | Lagrange multiplier |
| $\nu_k$ | Normalization factor for the $k$-th set of steering vectors (cf. equation 3.23) |
| $\nu_{mod}$ | Normalization constant for the individual IPD-discrimination ability in equation 5.4 |
| $\tau_n(\theta)$ | Delay for $n$-th microphone and direction $\theta$ |
| $\phi_{\mathrm{JND}}(f)$ | Interaural time difference threshold as a function of frequency $f$ |

| | |
|---|---|
| $\phi_{lin}(f)$ | Linearized HRTF phase as a function of frequency $f$ |
| $\phi_{orig}(f)$ | Original HRTF phase as a function of frequency $f$ |
| $\phi_{test}(f)$ | Test phase integrating $\phi_{lin}$ and $\phi_{orig}$ as a function of frequency $f$ (cf. equation 5.1) |
| $\Psi$ | Model psychometric function |
| $\Psi_0$ | Base function for the psychometric function |
| $\boldsymbol{\Omega}^t$ | Frequency vector for the t-th frequency band centered around $f_c^t$, incorporating $\mathrm{L}^t$ frequency bins, with $\boldsymbol{\Omega}^t = f_1^t \ldots f_c^t \ldots f_{\mathrm{L}^t}^t$ |

## List of acronyms and abbreviations

| | |
|---|---|
| AFC | Alternative forced choice |
| CB | Critical band |
| DH | Dummy head |
| ERB | Equivalent rectangular bandwidth |
| FEC | Free air equivalent coupling of a headphone to the ear |
| FIR | Finite impulse response |
| FFT | Fast Fourier transform |
| GUI | Graphical user interface |
| HPTF | Headphone transfer function |
| HRTF | Head-related transfer function |
| hrir | Head-related impulse response |
| ILD | Interaural level difference |
| IPD | Interaural phase difference |

| | |
|---|---|
| ITD | Interaural time difference |
| JND | Just noticeable difference |
| NFFT | Number of frequency bins for FFT |
| SNR | Signal to noise ratio |
| VAH | Virtual artificial head |
| WNG | White noise gain (cf. equation 3.3) |
| $\mathrm{WNG}_{\mathrm{m}}$ | Mean white noise gain (cf. equation 3.7) |
| $\mathrm{WNG}_{\mathrm{v}}$ | White noise gain with a variable bandwidth (cf. equation 3.16) |

# Contents

**Glossary** v

Mathematical notation . . . . . . . . . . . . . . . . . . . . . . . v
Fixed symbols . . . . . . . . . . . . . . . . . . . . . . . . . . . vi
List of acronyms . . . . . . . . . . . . . . . . . . . . . . . . . x

**List of Figures** xvii

**List of Tables** xxv

**1 General introduction and main objectives** 9

1.1 Introduction and main objectives . . . . . . . . . . . . . . . 10
1.2 Spatial hearing and head-related transfer functions . . . . . 11
1.3 HRTF measurements . . . . . . . . . . . . . . . . . . . . . . 16
1.4 Binaural reproduction . . . . . . . . . . . . . . . . . . . . . 18
1.5 Spatial filtering using microphone arrays . . . . . . . . . . . 21
1.6 Regularization . . . . . . . . . . . . . . . . . . . . . . . . . 25
1.7 Modeling aspects of spatial hearing . . . . . . . . . . . . . . 26
1.8 Outline of this thesis . . . . . . . . . . . . . . . . . . . . . 28

**I Optimizing microphone arrays to synthesize individual head-related directivity patterns** 29

**2 Microphone positioning for the VAH** 33

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 34
2.2 Calculation of filter coefficients . . . . . . . . . . . . . . . . 34
2.3 Random Sampling method . . . . . . . . . . . . . . . . . . . 35
2.4 Golomb method . . . . . . . . . . . . . . . . . . . . . . . . 36
2.5 Regularization . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.6 Numerical Simulation . . . . . . . . . . . . . . . . . . . . . 38
2.7 Results and discussion . . . . . . . . . . . . . . . . . . . . . 39
  2.7.1 Effect of positioning errors . . . . . . . . . . . . . . 39
  2.7.2 Effect of regularization . . . . . . . . . . . . . . . . . 40
  2.7.3 Best-case scenario: optimal number of microphones per critical band . . . . . . . . . . . . . . . . . . . . 43

2.8 Conclusions . . . . . . . . . . 43

3 Regularization approaches for synthesizing head-related transfer functions with microphone arrays 45
3.1 Introduction . . . . . . . . . . 46
3.2 Least squares beamformer design . . . . . . . . . . 47
3.3 Regularization by joint optimization for multiple steering vectors . . . . . . . . . . 49
3.4 Regularization using a WNG constraint . . . . . . . . . . 50
3.4.1 White noise gain . . . . . . . . . . 50
3.4.2 Mean white noise gain over all directions . . . . . . 52
3.4.3 Optimization and white noise gain for multiple frequencies . . . . . . . . . . 53
3.5 Experimental results . . . . . . . . . . 57
3.5.1 Joint optimization for multiple steering vectors . . . 58
3.5.2 White noise gain constraint . . . . . . . . . . 60
3.5.3 Mean white noise gain constraint . . . . . . . . . . 61
3.5.4 Constraint on the white noise gain for multiple frequencies . . . . . . . . . . 63
3.5.5 Robustness of the different regularization approaches 63
3.6 Conclusion . . . . . . . . . . 66

II Smoothing individual HRTFs prior to the synthesis using the virtual artificial head 69

4 Psychoacoustically appropriate truncation of individual hrirs 73
4.1 Introduction . . . . . . . . . . 74
4.2 Test preparation . . . . . . . . . . 75
4.2.1 HRTF measurement . . . . . . . . . . 75
4.2.2 Reference setting . . . . . . . . . . 75
4.3 Listening tests - material and methods . . . . . . . . . . 76
4.4 Listening tests - results . . . . . . . . . . 78
4.5 Impact of information reduction on the performance of the VAH . . . . . . . . . . 81
4.6 Summary and further steps . . . . . . . . . . 82

5 Smoothing HRTFs in the frequency and spatial domains 83
5.1 Introduction . . . . . . . . . . 84
5.2 HRTF smoothing operations considered in this study . . . . 89
5.3 Methods . . . . . . . . . . 91
5.3.1 Subjects . . . . . . . . . . 91
5.3.2 HRTF measurements . . . . . . . . . . 92
5.3.3 Procedure and stimuli . . . . . . . . . . 93

5.4 Results and discussion . . . . . . . . . . . . . . . . . . . . . . . 99
  5.4.1 Substitution of the individual HRTF phase by a mixture of the original phase and a linear phase . . . . . 99
  5.4.2 Substitution of the individual HRTF phase by a linear phase above a certain cutoff-frequency . . . . . . 101
  5.4.3 Complex-valued smoothing of the individual HRTFs into constant relative bandwidths . . . . . . . . . . . 103
  5.4.4 Limiting the spatial dynamic range of the individual HRTFs . . . . . . . . . . . . . . . . . . . . . . . . . . 105
5.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
  5.5.1 Perceptual effects of phase linearization . . . . . . . 106
  5.5.2 Perceptual effects of smoothing HRTFs in the frequency domain . . . . . . . . . . . . . . . . . . . . . 107
  5.5.3 Influence of the test signal . . . . . . . . . . . . . . . 109
  5.5.4 Impact of the proposed smoothing methods on the HRTFs . . . . . . . . . . . . . . . . . . . . . . . . . . 109
  5.5.5 Impact of the proposed smoothing methods on the accuracy of the VAH . . . . . . . . . . . . . . . . . . 110
5.6 Summary and conclusions . . . . . . . . . . . . . . . . . . . 113

**III Evaluation of individualized binaural reproduction using the virtual artificial head 115**

**6 Perceptual impact of the WNG 119**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
6.2 Regularized least squares cost function . . . . . . . . . . . . 121
  6.2.1 Influence of the WNG-constraint on the VAH-synthesis 122
6.3 Used microphone arrays . . . . . . . . . . . . . . . . . . . . 123
6.4 Experimental procedure . . . . . . . . . . . . . . . . . . . . 125
  6.4.1 Material . . . . . . . . . . . . . . . . . . . . . . . . . . 125
  6.4.2 Test stimulus . . . . . . . . . . . . . . . . . . . . . . . 125
  6.4.3 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . 126
6.5 Results and discussion - Perceptual evaluation . . . . . . . . 127
  6.5.1 Localization . . . . . . . . . . . . . . . . . . . . . . . . 130
  6.5.2 Sensor noise . . . . . . . . . . . . . . . . . . . . . . . . 130
  6.5.3 Overall performance . . . . . . . . . . . . . . . . . . . 132
  6.5.4 Spectral coloration . . . . . . . . . . . . . . . . . . . . 133
6.6 Conclusions and further work . . . . . . . . . . . . . . . . . 134

**7 Individualized binaural reproduction using a VAH 135**
7.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 136
7.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137
  7.2.1 Head-related transfer functions (HRTFs) . . . . . . . 137

7.2.2 Synthesis using the virtual artificial head . . . . . . 139
7.2.3 Traditional artificial heads . . . . . . . . . . . . . . . . 140
7.2.4 Headphone transfer functions (HPTFs) . . . . . . . . 141
7.2.5 Procedure and stimuli . . . . . . . . . . . . . . . . . . 142
7.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 145
7.3.1 Explicitly considered directions . . . . . . . . . . . . 145
7.3.2 Intermediate directions . . . . . . . . . . . . . . . . . 147
7.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . 149
7.5 Conclusion and future work . . . . . . . . . . . . . . . . . 152

**8 General Conclusions 153**
8.1 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . 154
8.2 Suggestions for future research . . . . . . . . . . . . . . . 158

**A Estimation of the delay and the resulting linear phase $\phi_{lin}(f)$ 161**

**B Applied method for the spectral smoothing of HRTFs 165**

**C Measuring apparatus 169**

**D Causality of optimized filter coefficients 171**

**E Imposing a constraint on the least squares cost function 175**

**Bibliography 179**

**Acknowledgments 195**

**Curriculum Vitae 197**

# List of Figures

1.1 Exemplary illustration of sound arriving at the ears in the horizontal plane varying with the azimuthal direction $\theta$. . . 11
1.2 HRTFs for the left ear (blue) and the right ear (red) in dB as a function of frequency for three exemplary directions ($\theta = 0°$, $\theta = 90°$ and $\theta = 225°$). Note that the HRTFs are offset from each other by 35 dB (between $\theta = 0°$ and $\theta = 90°$) and 55 dB (between $\theta = 90°$ and $\theta = 225°$) . . . . 13
1.3 HRTFs in dB re. 1 for the left ear (blue lines) and the right ear (red lines) as a function of direction $\theta$ in the horizontal plane for frequencies of $f \approx 500$ Hz (left), $f \approx 1500$ Hz (center) and $f \approx 7000$ Hz (right). . . . . . . . . . . . . . . . 14
1.4 HRTFs in dB as a function of frequency for three exemplary directions ($\theta = 0°$, $\theta = 90°$ and $\theta = 225°$) and for three exemplary subjects ($S_1$, $S_2$ and $S_3$) for the left ear. Note that the HRTFs are offset from each other by 35 dB (between $\theta = 0°$ and $\theta = 90°$) and 55 dB (between $\theta = 90°$ and $\theta = 225°$). . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
1.5 Rough summary of direction-dependent (directional) and direction-independent (non-directional) components of the HRTFs. The frequency ranges most likely affected by body parts are indicated. Illustration adapted from Begault (2000) 15
1.6 Exemplary course of the Hilbert envelope of hrirs for the left ear (blue lines) and the right ear (red lines) as a function of time on the x-axis and of azimuthal direction $\theta$ on the y-axis. 16
1.7 Knowles FG-23329 miniature electret microphones mounted in an individual earmold (left) and in a foam earplug (right). Illustration adapted from Raufer *et al.* (2013). . . . . . . . . 17
1.8 Binaural reproduction of the monaural signal s(t) by convolution ($\otimes$) with the hrirs and with the equalization of the HPTFs for both ears separately. . . . . . . . . . . . . . . . . 18
1.9 Exemplary HPTFs of the AKG-240-K Studio headphones measured 10 times subsequently (the headphone was taken off and on before each repetition) on an artificial head (Head Acoustics HMS-II). . . . . . . . . . . . . . . . . . . . . . . 20

1.10 Exemplary illustration of spatial filtering (using a filter-and-sum structure) with a line array (Illustration adapted from Doclo and Moonen (2003a)). . . . . . . . . . . . . . . . . . 22
1.11 HRTFs as the desired directivity patterns with an exemplary microphone array with N = 4 microphones and the complex-valued weights $\mathbf{w}(f) = [wl_1(f) \dots wl_4(f)]^T$ for the left ear and $\mathbf{w}(f) = [wr_1(f) \dots wr_4(f)]^T$ for the right ear. . . . . . 24

2.1 Procedure to derive a set of valid nodes according to DE 10 2010 012 388 A1. The dotted black lines indicate a two-dimensional grid originating from a Golomb-ruler with order $G = 10$. Black crosses indicate chosen nodes and the red lines show a shifted copy of the chosen nodes. . . . . . . . . 36
2.2 Simulated synthesis error $\epsilon_f$ of the virtual artificial head, as a function of frequency and the number of microphones in the array. The left column shows results for the Random Sampling method, the right column for the Golomb method. **Top row:** no positioning error, no regularization, **middle row:** random positioning error $|\Delta x_{max}|, |\Delta y_{max}| = 0.1$ mm, no regularization, **bottom row:** random positioning error $|\Delta x_{max}|, |\Delta y_{max}| = 0.1$ mm, regularization with $\beta_\Gamma = 10$. . 41
2.3 Simulated synthesis error $\epsilon_f$ of the virtual artificial head, for optimal numbers of microphones per critical band. The left column shows results for the Random Sampling method, the right column for the Golomb method. **Top row:** random positioning error $|\Delta x_{max}|, |\Delta y_{max}| = 0.1$ mm, no regularization, **bottom row:** random positioning error $|\Delta x_{max}|, |\Delta y_{max}| = 0.1$ mm, regularization with $\beta_\Gamma = 10$. Note that the ordinate scale is smaller than in the top row. The number of microphones leading to the minimal error (per critical band) is indicated by the color. . . . . . . . . 42

3.1 Schematic diagram of a filter-and-sum beamformer with N = 4 microphones and the associated filter weights $\mathbf{w}(f) = [w_1(f) \dots w_4(f)]^T$. Desired directivity pattern $\mathrm{D}(f, \theta)$ and exemplary resulting directivity pattern $\mathrm{H}(f, \theta)$. . . . . . . . . . . . . . . . . . . . 48
3.2 Mean white noise gain $\mathrm{WNG_m}$ for a left ear synthesis as a function of the Lagrange multiplier $\mu$ for three frequencies. 58
3.3 Mean (black line) and standard deviation (error bar) of the variability (absolute values) among the differently measured steering vectors (please note the linear frequency scaling). . 59

3.4 $\mathrm{WNG_m}$ of filter coefficients $\mathbf{w}_\mathrm{K}$ resulting from the (unconstrained) joint optimization when increasing the number of considered sets of steering vectors K. For the sake of clarity, the range of the depicted $\mathrm{WNG_m}$ is limited between $-60\ \mathrm{dB} \leq \mathrm{WNG_m} \leq -5\ \mathrm{dB}$. . . . . . . . . . . . . . . . . . . . . . 60

3.5 Desired HRTFs (black solid lines) and synthesis associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ (blue dot-dashed lines, $\beta = -5$ dB) and on $\mathrm{WNG_m}$ (red dashed lines, $\beta_\mathrm{m} = -5$ dB) for three directions ($\theta = 0°,\ 90°,\ 225°$). . . . . . . . . . . . 61

3.6 Error $\epsilon$ as a function of frequency and azimuthal direction for the synthesis of a left ear HRTF associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ ($\beta = -5$ dB) in the left panel and associated with a constraint on $\mathrm{WNG_m}$ ($\beta_\mathrm{m} = -5$ dB) in the right panel. The illustration of $\epsilon$ is limited to 8 dB for the sake of clarity. . . . . . . . . . . . . . . . . . . . . . . . 62

3.7 Error $\epsilon$ as a function of frequency and azimuthal direction for the synthesis of a left ear HRTF associated with a constraint on $\mathrm{WNG_v}$ ($\beta_\mathrm{v} = -5$ dB) with a weighting function $g(f,\theta) = \frac{1}{\mathrm{P}}$ in the left panel and the weighting function $g_w^t(f,\theta)$ according to equation 3.22 in the right panel. . . . 62

3.8 Mean error $\tilde{\epsilon}$ for 100 randomly disturbed analytic steering vectors $\tilde{\mathbf{d}}$ as a function of frequency and azimuthal direction for the synthesis of a left ear HRTF associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$. . . . . . . . . . . . . . . . . 66

4.1 Example hrir (left ear, azimuth=90°) with the reference length $l_{hrir} = 512$ samples and a tapered Hann-window with the fixed length $l_{win} = 50$ samples for the descending flank. 76

4.2 Sketch of the envelope of the test signal as a function of time. Each noise burst (350 Hz $\leq f \leq$ 18000 Hz) lasted 0.15 seconds and was windowed with 0.01 seconds onset-offset ramps. The pauses in between the noise bursts lasted 0.15 seconds as well. . . . . . . . . . . . . . . . . . . . . . . . . . 77

4.3 Exemplary psychometric function $\Psi(x)$ estimated from two supposed arguments $x_{80} = 200$ and $x_{40} = 400$. According to equation 4.1 and 4.2, this yields an argument at threshold ($\Psi(x_{50}) = 0.5$) of $x_{50} \approx 328$. . . . . . . . . . . . . . . . . . 79

4.4 Means and standard deviations of thresholds in the truncation experiment for 80% score (green) and 40% score (blue) for four subjects (○, □, ◇ and ✡). Red symbols indicate the calculated 50% scores and the dashed lines indicate their mean across subjects. . . . . . . . . . . . . . . . . . . . . . 80

5.1 Illustration of the interrelationship between the spatial complexity of HRTFs and the number of microphones needed to accurately synthesize them. The example HRTF set was taken from the IRCAM database (right ear of subject #1002), the original directivity is shown as solid line, the synthesis as dashed line. Left: $f$ = 1 kHz N = 8 microphones, center: $f$ = 11 kHz N = 8 microphones, right: $f$ = 11 kHz N = 24 microphones. . . . . . . . . . . . . . . . . . . . . . . . 85
5.2 Magnitude in dB (top) and group delay in samples (bottom) for an exemplary HRTF of subject S2 for the left ear from azimuth $\theta = 210°$. The black dot-dashed lines show the original HRTF (NFFT = 512), the blue dashed lines show the resulting HRTF after complex-valued smoothing into third octave bands, the solid red lines show the smoothed HRTF (third octave bands) when the phase of the HRTF is substituted by a linear phase above 5 kHz before smoothing. . . . 87
5.3 **Top:** Temporal course of the used test signal consisting of three noise bursts with a total length of 0.75 seconds. **Bottom:** Spectral shape of the test signal with a constant magnitude for frequencies 200 Hz $\leq f \leq$ 18000 Hz. . . . . . 94
5.4 Substitution of the original HRTF phase ($\phi_{\mathrm{orig}}$) by a linear phase ($\phi_{\mathrm{lin}}$) above a variable cutoff-frequency $f_{\mathrm{c}}$ (black circle): The resulting test phase $\phi_{\mathrm{test}}$ is equal to $\phi_{\mathrm{orig}}$ for $f \leq f_{\mathrm{c}}$, and equal to $\phi_{\mathrm{lin}}$ for $f > f_{\mathrm{c}}$. The slope of $\phi_{\mathrm{lin}}$ is computed from the maximum of the Hilbert envelope of the original *hrir*. . . . . . . . . . . . . . . . . . . . . . . . . . . 96
5.5 Spatial dynamic range $\zeta$ of all left ear HRTFs (24 directions in the horizontal plane, all eight subjects), as a function of frequency. The minimum and maximum values of all subjects are shown as red lines, the range in between (characterizing the remaining dynamic ranges) as a gray area. Similar $\zeta$-values also result for HRTFs of the right ear (not shown). 97
5.6 Exemplary reduction of the spatial dynamic range of a normalized right ear HRTF magnitude in the horizontal plane from $\zeta \approx 43$ dB (original) to $\zeta' = 20$ dB (test situation). . . 98
5.7 Threshold data for experiment 1: means and standard deviations over three to six runs of the the mixing gain $L_\phi$ at threshold, as a function of subjects and direction of incidence $\theta$. Smaller values of $L_\phi$, i.e. a smaller proportion of the linear phase, indicate a higher sensitivity to the broadband modification of the individual HRTF phase and vice versa. . . . . . . . . . . . . . . . . . . . . . . . . . . 100

5.8 Experimentally observed mean values of $L_\phi$, as a function of the individual discrimination ability $L_{\mathrm{mod}}$ (equation 5.4). This relation can be characterized by a linear regression $L_{\mathrm{reg}}(L_{\mathrm{mod}}) = -1.0077 \cdot\ L_{\mathrm{mod}} + 0.8627$ (red line), with a linear correlation coefficient of $\rho(L_{\mathrm{mod}}, L_\phi) = -0.83$. 101
5.9 Threshold data for experiment 2: means and standard deviations over three to six runs of the cutoff-frequency $f_{\mathrm{c}}$ at threshold between the original (for $f \leq f_{\mathrm{c}}$) and the linear phase (for $f > f_{\mathrm{c}}$). Lower $f_{\mathrm{c}}$ indicate that a larger proportion of the original phase can be substituted with the linear phase still yielding a discrimination at threshold and vice versa. . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
5.10 Exemplary course of a measured *hrir* (red solid line) and a manipulated *hrir* (blue dashed line) for the left ear of subject S8 and $\theta = 105^\circ$ as a function of time. The phase of the latter *hrir* is linearized for $f \geq 1$ kHz, which results in a rather symmetrical envelope of the corresponding *hrir* in the time domain. Note that the delay of the peak magnitude remains unchanged. . . . . . . . . . . . . . . . . . . . . . . 103
5.11 Threshold data for experiment 3: means and standard deviations over three to six runs of the relative bandwidth $\mathrm{B_W}$ at threshold expressed as fractions of one octave when complexly smoothing HRTFs in the frequency domain. Prior to smoothing, the HRTF phase (for $f > 5$ kHz) was substituted with the linear phase $\phi_{lin}$. Higher $\mathrm{B_W}$ indicate more smoothing at threshold and thus less sensitive thresholds and vice versa. . . . . . . . . . . . . . . . . . . . . . . . . . 104
5.12 Threshold data for experiment 4: means and standard deviations over three to six runs of the spatial dynamic range $\zeta'$ as a function of azimuthal direction $\theta$. Note the non-equidistant spacing. Here the individual subjects are depicted as various symbols. Smaller spatial dynamic ranges $\zeta'$ indicate larger manipulations of the directivity patterns and thus less sensitive thresholds and vice versa. . . . . . 105

5.13 Effects of the proposed smoothing methods, **top**: Exemplary HRTF (subject S2, right ear) as function of frequency in kHz on the x-axis and as a function of azimuthal direction $\theta$ on the y-axis, **middle:** the same HRTF, processed through a complex-valued smoothing as described in section 5.3.3.3 with a bandwidth of $\mathrm{B_W} = \frac{1}{5}$ octave, **bottom:** the same HRTF, with the spatial dynamic range limited to $\zeta' = 29$ dB as described in section 5.3.3.4. The latter manipulation is primarily apparent for contralateral directions ($0^\circ \leq \theta \leq 180^\circ$) where the spatial notches (light areas) are adjusted. In this illustration, levels were limited between -5 and -35 dB to highlight the differences of the particular smoothing methods. . . . . . . . . . . . . . . . . . . . . . . . 108
5.14 Desired $\mathrm{HRTF_{des}}$ (solid lines) and synthesized $\mathrm{HRTF_{synth}}$ using the VAH (dashed lines) are depicted for the original (red) and for the spectrally smoothed case (blue) with $\mathrm{B_W} = \frac{1}{5}$ for an exemplary HRTF (subject S2, right ear, $\theta = 210^\circ$), as a function of frequency. . . . . . . . . . . . . . 111
5.15 The $\mathrm{HRTF_{des}}$ (solid lines) and their synthesis $\mathrm{HRTF_{synth}}$ using the VAH (dashed lines) are depicted for the original $\mathrm{HRTF_{des}}$ (red) and for the $\mathrm{HRTF_{des}}$ with reduced spatial dynamic range ($\zeta' = 29$ dB, blue lines) for an exemplary HRTF (subject S2, right ear, $f \approx 13.5$ kHz). . . . . . . . . . 112
5.16 Effects of the proposed smoothing methods on the accuracy of the VAH. Plotted is the VAH-error $\epsilon_f$ (equation 5.5) of the right ear HRTF of subject S2, as a function of frequency, for the original $\mathrm{HRTF_{des}}$ (red solid line), the spectrally smoothed $\mathrm{HRTF_{des}}$ ($\mathrm{B_W} = \frac{1}{5}$ octave, blue solid line) and the $\mathrm{HRTF_{des}}$ with reduced spatial dynamic range ($\zeta' = 29$ dB, green-dashed line). . . . . . . . . . . . . . . . . 113

6.1 Magnitude of the desired HRTF ($\theta = 90^\circ$) for the left ear of subject $\mathrm{S_1}$ (black line) and VAH-synthesis with various $\mathrm{WNG_m}$ (dashed lines) for $\mathrm{array_2}$ as a function of frequency. 123
6.2 Two used microphone arrays with 24 KE-4 microphones ($\mathrm{array_1}$, left) and 24 sensors composed of 48 MEMS microphones ($\mathrm{array_2}$, right) with the same planar microphone topology according to the procedure described in Rasumow *et al.* (2011b) or **Chapter 2**. . . . . . . . . . . . . . . . . 124

6.3 Perceptual ratings for array$_1$ (left block) and array$_2$ (right block). The aspects of evaluation are aligned in separate rows (first row: localization, second row: sensor noise, third row: overall performance and fourth row: spectral coloration) and the direction of arrival $\theta$ is aligned in three columns ($\theta = 90°$ in the left column, $\theta = 0°$ in the middle column and $\theta = 225°$ in the right column). The individual ratings (mean and standard deviation over three randomized presentations) are depicted as a function of the WNG$_\mathrm{m}$ in dB. The colors and markers indicate the four subjects (S$_1$, S$_2$, S$_3$ and S$_4$). . . . . . . . . . . . . . . . . . . . . . . . . . . 129
6.4 Perceptual ratings averaged over all subjects for array$_1$ (upper block) and array$_2$ (lower block) are depicted as the mean and the standard deviation for the four evaluated aspects (localization, overall performance, sensor noise and coloration). . . . . . . . . . . . . . . . . . . . . . . . . . . . . 131
6.5 Exemplary course of the Lagrange multiplier $\mu$ (cf. equation 3.11) for array$_1$ and array$_2$ (blue and red lines, respectively) and WNG$_\mathrm{m}$ of 0 dB and -6 dB (solid and dashed lines, respectively) as a function of frequency of the left-ear synthesis for S$_1$. . . . . . . . . . . . . . . . . . . . . . . . . . 132

7.1 Illustration of the directions which are explicitly considered in the optimization of the VAH-filters (white and blue loudspeaker symbols), the tested directions for three explicitly considered directions (blue) and three intermediate directions (red symbols). . . . . . . . . . . . . . . . . . . . . . . . . 138
7.2 Temporal envelope of the noise bursts with an additional stationary white noise of 40 dB SPL. . . . . . . . . . . . . . . 142
7.3 Used paradigm with the test signal associated with one of the five methods (Test device) played back via headphone in the test setting and via spectrally-equalized loudspeakers (1/LS) in the reference setting. . . . . . . . . . . . . . . . . . 142
7.4 Used headphones seated on the DH$_3$ with a toggle button (cf. red arrow) that allows to switch between the test and the reference setting. . . . . . . . . . . . . . . . . . . . . . . . 143
7.5 GUI for evaluating the five methods with regard to loudspeaker presentation for each direction and perceptual quality. 144
7.6 Aggregated ratings (y-axis) for all six subjects regarding the perception of localization for the five tested methods on the x-axis and three explicitly considered directions (left, middle and right panel). . . . . . . . . . . . . . . . . . . . . . . . . . 145

7.7 Aggregated results for all six subjects regarding the perception of spectral coloration for the five tested methods and three explicitly considered directions (left, middle and right panel). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 146
7.8 Aggregated results for all six subjects regarding the overall performance for the five tested methods and three explicitly considered directions (left, middle and right panel). . . . . . 147
7.9 Aggregated ratings (y-axis) for all six subjects regarding the perception of localization for the five tested methods on the x-axis and three intermediate directions ($\theta = 7.5°,\ 97.5°$ & $232.5°$).147
7.10 Aggregated results for all six subjects regarding the perception of spectral coloration for the five tested methods and three intermediate directions. . . . . . . . . . . . . . . . . . . 148
7.11 Aggregated results for all six subjects regarding the overall performance for the five tested methods and three intermediate directions. . . . . . . . . . . . . . . . . . . . . . . . . . 149

A.1 An exemplary *hrir* is depicted (dashed blue line) as a function of time in samples on the x-axis. The red solid line depicts the corresponding Hilbert envelope of this *hrir*. The blue asterisk is the maximum of the Hilbert envelope, characterizing the estimated delay $\tau$ in samples. . . . . . . . . . 162

B.1 Block diagram characterizing the procedure for the applied complex-valued smoothing of HRTFs. The process can be divided into five separate steps: In step I the delay $\tau$ is estimated as described in Appendix A. The resulting linear phase is applied to the HRTF for frequencies $f \geq 5$ kHz in step II. In step III the delay $\tau$ is removed from the measured HRTF. In step IV the preconditioned HRTF is smoothed complexly according to the algorithm proposed in Hatziantoniou and Mourjopoulos (2000). In step V the initial delay $\tau$ is reconstructed for the smoothed HRTF. . . 167

C.1 Circular loudspeaker array in the anechoic room at the Institut für Hörtechnik und Audiologie in Oldenburg . . . . . 170

D.1 Steering vectors for N=4 microphones (top), the desired hrir for $\theta = 0°$ and the resulting noncausal filter coefficients (bottom) as a function of time in samples ($f_s = 44100$ Hz). . . . 172
D.2 Steering vectors for N=4 microphones (top), the delayed desired hrir for $\theta = 0°$ and the resulting causal filter coefficients (bottom) as a function of time in samples ($f_s = 44100$ Hz). 173

# List of Tables

3.1 Mean error $\bar{\epsilon}$ for the presented regularization approaches without and with an additional random deviation of the analytic and measured steering vectors. . . . . . . . . . . . . . 64

6.1 p-values (rounded to 3 digits) according to the Friedman test regarding localization, overall performance, sensor noise and coloration for the three tested directions separately. p-values indicating significantly different ratings when varying the $\mathrm{WNG}_{\mathrm{m}}$ (p$\leq \frac{0.05}{24} = 0.0021$) are depicted as bold numbers.128

7.1 p-values resulting from the Friedman test. Values associated with a significant effect of the tested methods on the ratings ($p \leq \frac{0.05}{18}$) are indicated as bold numbers. . . . . . . . . . . 150

# Abstract

Spatial hearing for human listeners is based on the interaural as well as on the monaural analysis of the signals arriving at both ears, enabling the listeners to assign certain spatial components to these signals. This spatial aspect gets lost when the signals are reproduced via headphones without considering the acoustical influence of the head and torso, i.e. head-related transfer function (HRTFs). A common procedure to take into account spatial aspects in a binaural reproduction is to use so-called artificial heads. Artificial heads are replicas of a human head and torso with average anthropometric geometries and built-in microphones in the ears. Although, the signals recorded with artificial heads contain relevant spatial aspects, binaural recordings using artificial heads often suffer from front-back confusions and the perception of the sound source being inside the head (internalization). These shortcomings can be attributed to the missing individualization of the binaural recordings in a static scenario without visual cues. Alternatively, the desired frequency-dependent directivity pattern of individual HRTFs can also be synthesized by using a microphone array with individually optimized filter coefficients (referred to as virtual artificial head, VAH), which is the main goal of this thesis. The main advantages of a VAH are the possibility of adjusting the filter coefficients to HRTFs of different listeners (individualization) and to different look directions (orientation), the possibility of employing head tracking in the reproduction stage and a better flexibility and manageability due to the smaller size/weight of the device.

This thesis deals with the individual aspects of human spatial hearing, with the measurement and the perception of the associated HRTFs and the synthesis of the associated directivity patterns by using microphone arrays. This thesis is thematically subdivided into three parts: 1. The optimization of the beamformer filter coefficients to synthesize the desired directivity patterns, 2. The imperceptible simplification of individual HRTFs prior to the optimization in order to synthesize only perceptually relevant aspects of HRTFs, and 3. The evaluation of the resulting VAH-synthesis in comparison to binaural recordings using traditional artificial heads.

In the first part of this thesis, a mathematically motivated method to derive appropriate microphone topologies for HRTF-synthesis using a VAH is introduced. In a subsequent study, different regularization strategies to improve the robustness of the VAH-synthesis against errors in the microphone characteristics are presented and numerically evaluated. It is shown to be advantageous for the regularization to take into account all directions and to adapt the bandwidth of the optimization and regularization according to the frequency grouping of the human auditory system.

In the second part of this thesis, it is examined to which extent individual HRTFs may be smoothed without causing a detectable perceptual difference compared to a chosen reference condition (binaural reproduction with head-related impulse responses (hrirs) truncated to approximately 12 ms). The main reason behind this investigation is to synthesize only the perceptually-relevant aspects of individual HRTFs and hence to improve the accuracy and the robustness of the synthesis. It turns out that individual hrirs may be truncated to approximately 6 ms in the time domain, and the individual phase response of HRTFs may be substituted by a linear phase response for frequencies $f \geq 1$ kHz. Based on these findings, the complex-valued HRTFs can be smoothed in relative bandwidths after substituting the original phase by a linear phase for higher frequencies. The bandwidth of this complex-valued smoothing can be increased up to $\frac{1}{5}$ octave without yielding a detectable difference. Furthermore, it is shown that spatial notches in the frequency-dependent directivity pattern do not need to be retained in detail if they are less than 29 dB below the maximum value. It is found that such an imperceptible smoothing of the HRTFs prior to optimizing the beamformer filter coefficients improves the VAH-synthesis.

In the third part of this thesis, it is shown that the perceptual evaluation of the VAH-synthesis depends on, e.g., the desired regularization but also on the used microphone array and the associated sensor noise. In general, microphone arrays with a lower sensor noise yield better properties for the synthesis. In a subsequent study, the individual VAH-synthesis and the binaural reproduction using traditional artificial heads are perceptually evaluated using listening experiments in comparison to free field presentation. It is found that individuality plays an important role when evaluating binaural reproductions. On average, the VAH-synthesis results in good to excellent perceptual ratings for explicitly considered directions, mainly with better perceptual ratings for the VAH-synthesis in comparison to traditional artificial heads. Perceptual ratings range between fair and good for intermediate, i.e. not explicitly considered directions, and are roughly at the level (or slightly better in terms of the overall performance) of the best ratings associated with a traditional artificial head. In summary, the perceptual ratings confirm the validity of synthesizing HRTFs using the VAH and emphasize the advantages associated with individualized binaural reproduction.

# Zusammenfassung

Das räumliche Hören des Menschen basiert auf der interauralen und monauralen Auswertung der beiden Signale, die an den jeweiligen Ohren ankommen. So ist dem Menschen mithilfe einer internen Verarbeitung möglich, diesen zwei Ohrsignalen eine räumliche Komponente zuzuordnen. Dieser räumliche Aspekt geht jedoch verloren, wenn die Signale ohne die akustischen Veränderungen durch den Kopf und Torso (d.h. kopfbezogene Übertragungsfunktion, englisch: head-related transfer function (HRTF)) direkt über Kopfhörer wiedergegeben werden. Ein übliches Vorgehen, um eine räumliche Komponente bei der binauralen Wiedergabe zu ermöglichen, ist die Aufnahme mit so genannten Kunstköpfen. Diese stellen eine Nachbildung des menschlichen Oberkörpers inklusive Kopfes mit mittleren anthropometrischen Geometrien dar und haben eingelassene Mikrofone anstelle der Ohren. Über Kunstköpfe aufgenommene Signale erhalten dadurch relevante räumliche Aspekte, wobei die Darbietung dieser binauralen Aufnahmen oft mit Vorne-Hinten-Vertauschungen oder der Wahrnehmung der Schallquellen im Kopf (Internalisierung) verbunden ist. Diese Nachteile können auf eine fehlende Individualisierung der binauralen Aufnahmen innerhalb eines statischen Szenarios ohne visuelle Reize zurückgeführt werden. Alternativ lassen sich die Richtcharakteristiken individueller HRTFs auch mithilfe von Mikrofonarrays mit individuell optimierten Filterkoeffizienten (weiterhin bezeichnet als virtual artificial head, VAH) synthetisieren, was das Hauptziel dieser Dissertation darstellt. Die Hauptvorteile eines VAH sind die Möglichkeit zur Anpassung der Filterkoeffizienten an HRTFs verschiedener Hörer (Individualisierung) und an verschiedene Blickrichtungen (Ausrichtung), als auch die Möglichkeit zur Einbindung von Head-Tracking auf Ebene der Reproduktionsstufe und die bessere Flexibilität und Handhabbarkeit aufgrund der geringeren Größe und des geringeren Gewichts des VAH.

Diese Dissertation befasst sich mit den individuellen Aspekten des räumlichen Hörens beim Menschen, mit der Messung und der Wahrnehmung der damit verbundenen kopfbezogenen Übertragungsfunktionen und der Synthese dieser individuellen Richtcharakteristiken mithilfe von Mikrofonarrays. Die Arbeit ist thematisch in drei Abschnitte unterteilt: 1. Die Optimierung eines beamformers, das die gewünschten Richtcharakteris-

tiken synthetisiert, 2. Die nicht wahrnehmbare Vereinfachung individueller HRTFs vor der Optimierung, um lediglich perzeptiv-relevante Aspekte der HRTFs zu synthetisieren und 3. Die Evaluation der resultierenden VAH-Synthesen im Vergleich zu binauralen Aufnahmen mithilfe traditioneller Kunstköpfe.

Im ersten Teil dieser Dissertation wird eine mathematisch motivierte Methode zur Positionierung der Mikrofone für den VAH vorgestellt, die angemessene Eigenschaften für die Synthese von HRTFs aufweist. In einer weiteren Studie werden unterschiedliche Regularisierungsstrategien zur Verbesserung der Robustheit der VAH-Synthese vorgestellt und numerisch miteinander verglichen. Es zeigen sich eindeutige Vorteile für eine Regularisierung, die alle Richtungen berücksichtigt bzw. eine psychoakustisch motivierte Anpassung der Bandgruppenbreiten innerhalb der Optimierung und Regularisierung ermöglicht.

Im zweiten Teil der Dissertation wird untersucht, inwieweit sich individuelle HRTFs vereinfachen bzw. glätten lassen, ohne dass diese Glättung einen detektierbaren Unterschied relativ zu einer gewählten Referenzbedingung zur Folge hat. Die Grundidee dieser Untersuchungen ist es, lediglich perzeptiv-relevante Aspekte individueller HRTFs zu synthetisieren, um somit die Präzision der Synthese zu verbessern. Es zeigt sich, dass die Impulsantworten individueller HRTFs im Zeitbereich auf ca. 6 ms gekürzt werden können (im Vergleich zu ca. 12 ms in der Referenzbedingung) und der individuelle Phasengang der HRTFs oberhalb von ca. 1 kHz durch einen linearen Phasengang ersetzt werden kann. Darauf basierend können die HRTFs nach einer hochfrequenten Phasenlinearisierung komplexwertig in relativen Bandbreiten geglättet werden, wobei die Bandbreite der Glättung auf bis zu $\frac{1}{5}$ Oktaven erhöht werden kann, ohne dass dieses einen detektierbaren Unterschied bewirkt. Weiterhin zeigt sich, dass Einbrüche im Betrag der räumlichen Richtcharakteristiken der HRTFs (pro Frequenz) nicht im Detail berücksichtigt werden brauchen, falls diese mehr als 29 dB unterhalb des lautesten Anteils dieser Richtcharakteristik liegen. Es wird gezeigt, dass eine derartige nicht-detektierbare Glättung der HRTFs vor der Optimierung der individuellen VAH-Filterkoeffizienten eine Verbesserung der Synthesen ergibt.

Im dritten Teil der Dissertation kann gezeigt werden, dass die subjektive Beurteilung der Synthesen von der gewünschten Regularisierung und auch von dem verwendeten Mikrofonarray und dem zugehörigen Grundrauschen abhängt. Im Allgemeinen ergeben Mikrofonarrays mit einem geringen Grundrauschen bessere Eigenschaften für die Synthese. In einer weiteren Studie werden individuelle Synthesen des VAH zusammen mit den binauralen Aufnahmen traditioneller Kunstköpfe im Vergleich zur Freifelddarbietung evaluiert. Es zeigt sich dabei, dass individuelle Aspekte bei der binauralen Reproduktion eine große Rolle in der Beurteilung dieser Reproduktionen einnehmen. Die Evaluationen der VAH-Synthesen für explizit berücksichtigte Richtungen ergeben im Mittel gute bis sehr gute Beurteilungen, größtenteils mit besseren Beurteilungen der VAH-Synthesen im Vergleich zu traditionellen Kunstköpfen. Für zwischenliegende und somit nicht explizit berücksichtigte Richtungen ergeben sich mittlere bis gute Beurteilungen, welche sich grob auf dem Niveau (bzw. leicht besser bezüglich der Gesamtperformance) des am besten bewerteten traditionellen Kunstkopfes befinden. Insgesamt bestätigen die Evaluationen die Validität der Synthese mithilfe des VAH und unterstreichen die Vorteile einer individualisierten binauralen Synthese.

# 1

# General introduction and main objectives

## 1.1 Introduction and main objectives

Hearing is an important social sense of human beings (Kohlrausch *et al.* (2013)) and enables the acquisition of information at various levels. Sound, as for instance speech, may transfer the semantic information between two or multiple communication partners. In addition, the perception of an acoustical event is also associated with the information about the acoustical environment and the spatial information (e.g. source direction, reflections and reverberation). Primarily, the spatial information is acquired by evaluating the time and level-differences of the signals arriving at the two ears (cf. Figure 1.1) and their characteristic properties (e.g. spectral coloration, temporal course etc.), which is further explained in section 1.2. This spatial information typically is lost when sounds are recorded without considering the acoustical influence of the listener (i.e. diffraction, reflection etc.), which is disadvantageous for an authentic reproduction of this recording via headphones. This aspect is usually taken into account by using so-called artificial heads (cf. Mellert (1972) or Paul (2009) for an extensive review), which are replicas of a real human head and torso with average anthropometric geometries and built-in microphones in the ears. Using artificial head recordings, it is possible to capture spatial aspects of sound and mimic a virtual perception from a certain direction when such recordings are reproduced via headphones (binaural reproduction, cf. section 1.4). However, binaural reproduction with artificial heads is also known to result in perceptual shortcomings as, e.g., front-back confusions (cf. Hill *et al.* (2000), Wenzel *et al.* (1993)) and the perception of acoustical events inside the head (internalization, cf. Hartmann and Wittenberg (1996)). This can primarily be attributed to the geometrical deviations of the artificial head in comparison to the listener in a static scenario without visual cues. The perceptual authenticity of a non-individualized binaural reproduction can be enhanced notedly when the individual acoustical properties (diffraction and reflection due to the listener) are considered (cf. Wenzel *et al.* (1993) and Masiero (2012)). However, a traditional artificial head with individualized anthropometric geometries would result in a considerable increase in labor costs, financial expense and a limited variability.

The main motivation and objective of this thesis is to synthesize individual acoustical properties, i.e. the individual directivity patterns of the listeners' head and torso in the horizontal plane with microphone arrays using appropriate beamformer techniques as a substitute for traditional artificial heads. Such a device is further referred to as a virtual artificial head (VAH). In this context, the desired directivity patterns are synthesized using individually optimized filter coefficients as opposed to the (fixed) geometrical filtering of traditional artificial heads. The main advantages of the VAH are the possibility to (post-hoc) individualize binaural reproductions by simply replacing the filter coefficients. Furthermore, recordings with the VAH enable the individualized reproduction for multiple listeners by using the same recording or allow the virtual illusion of movement of the listener's head by adapting the filter coefficients (e.g. using a head-tracker) for a static recording. In addition, a VAH would be easier to handle in comparison to rather bulky devices associated with traditional artificial heads. The conducted studies and experiments are aimed to obtain novel insights into the field of spatial hearing and into the synthesis of multi-directional desired directivity patterns by using microphone arrays.

## 1.2 Spatial hearing and head-related transfer functions

In contrast to the visual system, the auditory system of humans processes the acoustical information from all directions (front, back, top etc.) combining it to a spatial acoustical image (Seeber (2003)). In the horizontal plane (i.e. without considering elevation), this spatial information is primarily acquired by evaluating the interaural differences (i.e. the time and level-differences of the signals arriving at the two ears). This is exemplarily depicted in Figure 1.1 for a sound source on the right side of the

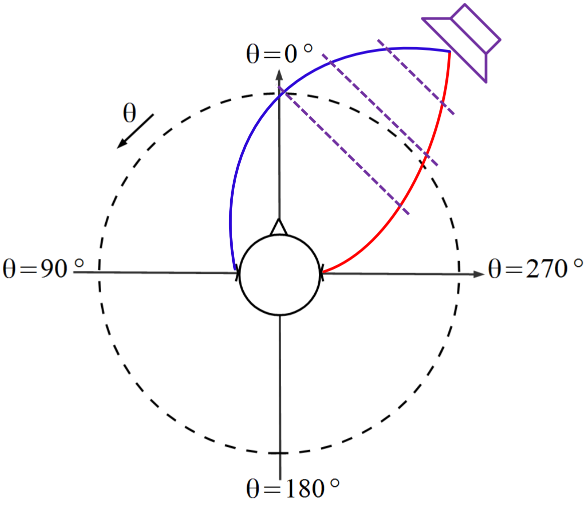


**Figure 1.1:** Exemplary illustration of sound arriving at the ears in the horizontal plane varying with the azimuthal direction $\theta$.

listener in the horizontal plane. A plane wave (under far-field assumptions) radiating from a sound source on the right hemisphere first arrives at the right ear and subsequently at the left ear. The temporal difference is usually termed the interaural time difference (ITD) (cf. Figure 1.1) and the phase difference of the two signals is frequently termed as the interaural phase difference (IPD). Additionally, the head of the listener acts as a frequency-dependent acoustic obstacle, yielding a perceptually relevant interaural level difference (ILD).

A frequency-dependent combination of these interaural cues to perceive a spatial acoustic image was already postulated in the Duplex-theory by Lord Rayleigh in 1907 and further investigated and expanded in the last century (e.g. Licklider (1951), Wightman and Kistler (1992)). It has been shown that ITDs are primarily dominant with regard to localization at lower frequencies ($f < 640$ Hz, cf. Seeber (2003)), where ILDs can be considered negligible. ILDs, on the other hand, are dominant with regard to localization at higher frequencies. However, such assumptions do not take into account the localization of sounds where interaural differences do not contain all relevant spatial information, as for instance for elevated sources or directions with ambiguous ILDs and ITDs (cone of confusion, cf. Searle *et al.* (1976), Blauert (1997), Moore (2003)). It is known that listeners with one occluded ear are still able to assign different source directions to broadband stimuli (cf. Belendiuk and Butler (1975), Wightman and Kistler (1997)), which indicates that the monaural properties of sounds arriving at the ears also contain essential information with regard to localization. Furthermore, only considering the individually-shaped spectral information at the two ears enables the externalized perception (i.e. a perception outside the head) of signals presented via headphones, while considering *only the interaural difference spectrum is inadequate* (Hartmann and Wittenberg (1996)).

The complex-valued transfer function of signals arriving at the ears, usually termed as the head-related transfer function (HRTF), is defined (cf. Mellert *et al.* (1974), Mehrgardt and Mellert (1977) or Seeber (2003)) as

$$\mathrm{HRTF}_{l/r}(f,\theta) = \frac{\mathcal{F}\{p_{l/r}(t,\theta)\}}{\mathcal{F}\{p_0(t,\theta)\}}, \tag{1.1}$$

with $p_l(t,\theta)$ and $p_r(t,\theta)$ denoting the sound pressure signals arriving at the left and the right ear and $p_0(t,\theta)$ the sound pressure signal at the center of the head when the listener is absent as a function of time $t$ and direction $\theta$. The HRTF is the quotient of the spectra (Fourier transform $\mathcal{F}\{\,\cdot\,\}$) of the pressure signals, describing the complex-valued transfer function for frequency $f$ of a sound arriving from direction $\theta$. Analogously, the

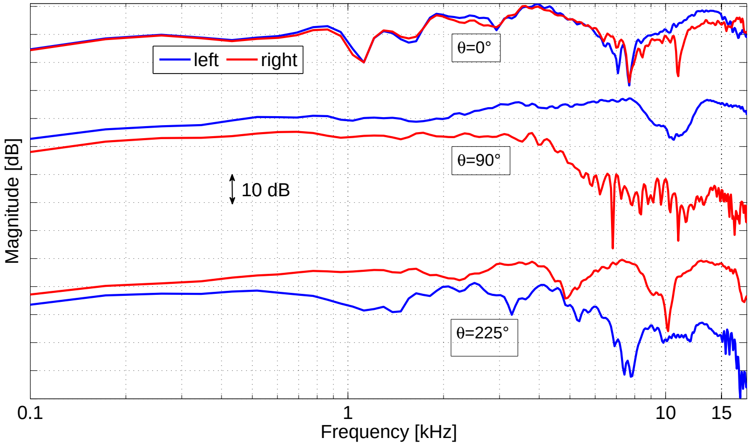


**Figure 1.2:** HRTFs for the left ear (blue) and the right ear (red) in dB as a function of frequency for three exemplary directions ($\theta = 0°$, $\theta = 90°$ and $\theta = 225°$). Note that the HRTFs are offset from each other by 35 dB (between $\theta = 0°$ and $\theta = 90°$) and 55 dB (between $\theta = 90°$ and $\theta = 225°$)

associated head-related impulse responses (hrir) in the time domain are given as

$$\mathrm{hrir}_{\mathrm{l/r}}(t,\theta) = \mathcal{F}^{-1}\{\mathrm{HRTF}_{l/r}(f,\theta)\}, \tag{1.2}$$

with $\mathcal{F}^{-1}\{\cdot\}$ denoting the inverse Fourier transform. It is worth noting that the magnitude of the HRTFs (cf. Figure 1.2, 1.3 and 1.4) as well as the phase response (primarily perceptually relevant at lower frequencies) vary both with direction $\theta$ but are also different for different subjects. This is due to the individual geometry and consequently to the individual effect of diffraction and reflection off the head and torso of human listeners, which results in individually-shaped HRTFs for different listeners (cf. Figure 1.4).

The magnitude of the HRTFs as a function of frequency for three exemplary directions is depicted in Figure 1.2. The HRTFs for the left ear (blue lines) and the right ear (red lines) are quite similar for the frontal direction ($\theta = 0°$) and deviate clearly for $\theta = 90°$ and $\theta = 270°$. It can be observed that the ILDs for these directions are rather moderate for lower frequencies and increase with frequency, resulting from a more efficient shadowing effect of the head for smaller wavelengths. This effect is even more apparent when depicting the directivity patterns of HRTFs (in Figure 1.3) for three exemplary frequencies, with the frequency increasing from

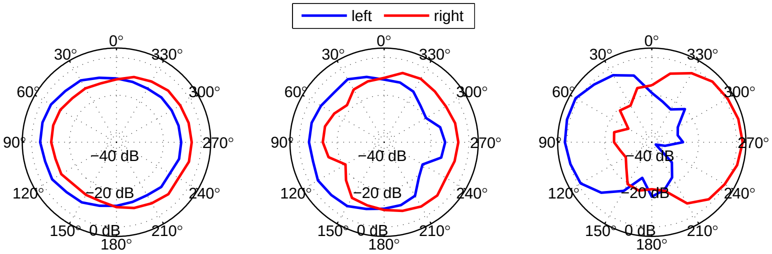


**Figure 1.3:** HRTFs in dB re. 1 for the left ear (blue lines) and the right ear (red lines) as a function of direction $\theta$ in the horizontal plane for frequencies of $f \approx 500$ Hz (left), $f \approx 1500$ Hz (center) and $f \approx 7000$ Hz (right).

$f \approx 500$ Hz (left) to $f \approx 1500$ Hz (center) and $f \approx 7000$ Hz (right). From this figure, it can be observed that the head shadowing for contralateral directions increases with frequency, altering the HRTFs from rather smooth (e.g. $f \approx 500$ Hz) to rather peaky directivity patterns (e.g. $f \approx 7000$ Hz). In addition, it seems plausible that the characteristic spectral shapes of

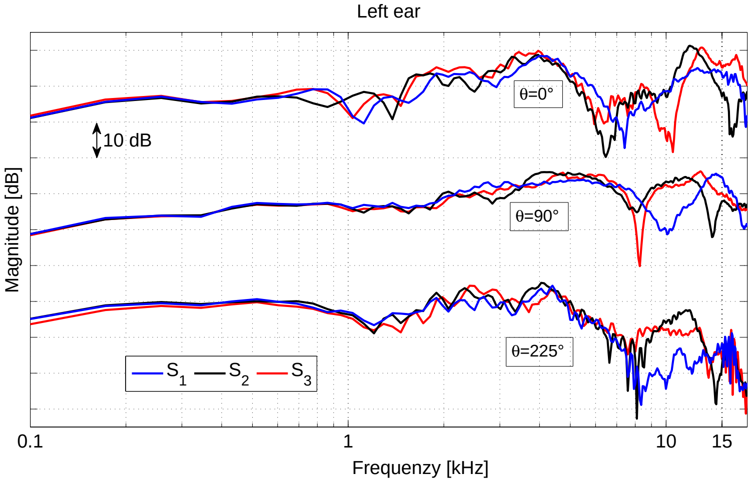


**Figure 1.4:** HRTFs in dB as a function of frequency for three exemplary directions ($\theta = 0°$, $\theta = 90°$ and $\theta = 225°$) and for three exemplary subjects ($S_1$, $S_2$ and $S_3$) for the left ear. Note that the HRTFs are offset from each other by 35 dB (between $\theta = 0°$ and $\theta = 90°$) and 55 dB (between $\theta = 90°$ and $\theta = 225°$).

HRTFs and the associated directivity patterns depend on the individual anthropometric geometries. This is exemplarily depicted in Figure 1.4. From this figure it can be observed that, e.g., the characteristic notches as well as the general spectral shape of the HRTFs vary clearly for the three depicted subjects (three exemplary subjects $S_1$, $S_2$ and $S_3$). Again, larger deviations in the HRTF magnitude are mainly evident at higher frequencies, whereas the HRTF magnitudes are quite similar at lower frequencies. This individual character is also evident for the phase of HRTFs and hence also for IPDs and ITDs (cf. Kulkarni *et al.* (1999), Ziegelwanger and Majdak (2014)).

A rough summary of which frequency ranges are assumed to be most likely affected by different body parts (according to Begault (2000)) is depicted in Figure 1.5. The blocks on the left side are assumed to alter the HRTFs differently for different directions while the blocks on the right side summarize the direction-independent aspects. According to this rough classification, mainly the head and torso and parts of the ear outside the ear canal seem to affect the signals at the ear and hence the HRTFs.

In summary, individual HRTFs and hrirs for both ears include the essential (monaural) spectral properties of signals arriving at the ears (cf. Figure 1.2-1.4) but also the individual interaural characteristics, such as IPDs, ITDs and ILDs, when comparing the HRTFs and hrirs between the two ears. This is exemplarily illustrated in Figure 1.6 for a measured set of hrirs in the horizontal plane. The temporal course of the envelope of hrirs for

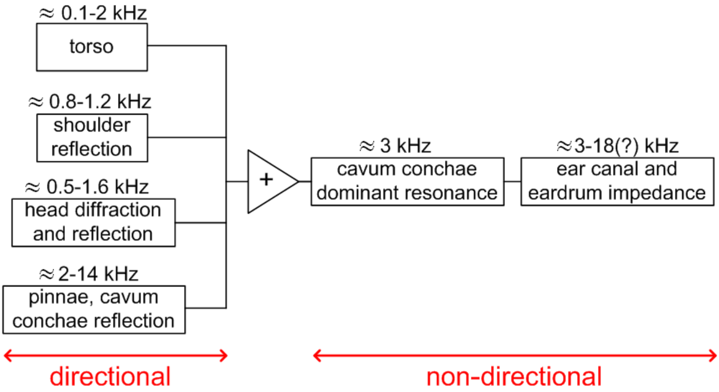

≈ 0.1-2 kHz
torso
≈ 0.8-1.2 kHz
shoulder reflection
≈ 0.5-1.6 kHz
head diffraction and reflection
≈ 2-14 kHz
pinnae, cavum conchae reflection
+
≈ 3 kHz
cavum conchae dominant resonance
≈3-18(?) kHz
ear canal and eardrum impedance
directional
non-directional

**Figure 1.5:** Rough summary of direction-dependent (directional) and direction-independent (non-directional) components of the HRTFs. The frequency ranges most likely affected by body parts are indicated. Illustration adapted from Begault (2000)

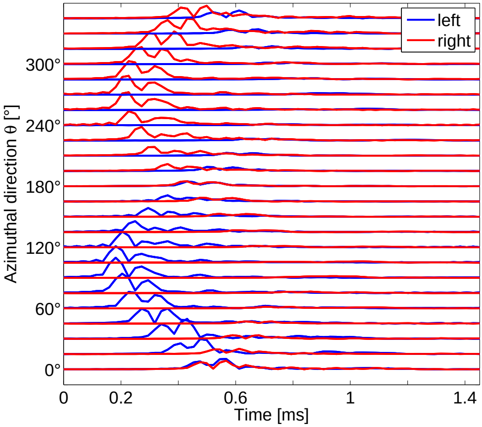


**Figure 1.6:** Exemplary course of the Hilbert envelope of hrirs for the left ear (blue lines) and the right ear (red lines) as a function of time on the x-axis and of azimuthal direction $\theta$ on the y-axis.

the left ear (blue lines) and for the right ear (red lines) are depicted for different azimuthal directions $\theta$ on the y-axis. The varying delay between the hrirs illustrates the effects of IPDs/ITDs, with a minimum interaural delay for $\theta = 0^\circ$ and $\theta = 180^\circ$ and a maximum interaural delay for $\theta = 90^\circ$ and $\theta = 270^\circ$. Furthermore, the hrirs in Figure 1.6 also indicate the different ILDs since the magnitude of the hrirs is consistently larger for the ipsilateral side (i.e. $0^\circ \leq \theta < 180^\circ$ for left side and $180^\circ \leq \theta < 360^\circ$ for the right side) compared to the contralateral side.

## 1.3 HRTF measurements

In order to measure individual HRTFs as defined in equation 1.1, one needs to capture the sound pressure signals $p_l(t, \theta)$ and $p_r(t, \theta)$ for all desired directions $\theta$, thus including the relevant spatial information that is needed for the perception of a spatial acoustical image. It may seem plausible to measure the pressure signals directly at the eardrum since this is the pressure signal which is further processed by the middle and inner ear (cf. Hammershøi and Møller (1996), Schmidt *et al.* (2011) and Blau *et al.* (2013)). However, Hammershøi and Møller (1996) have shown that sound transmission in the ear canal is largely independent of direction for the major part of the audio frequency range (this is also indicated by Begault

(2000), cf. Figure 1.5), which is why the authors of that study recommend to measure individual HRTFs at the blocked entrance of the ears. In doing so, miniature microphones are usually mounted to the measuring apparatus, e.g., using fixed tubes attached to the seat of the listener (cf. Dobrucki *et al.* (2010)) or using custom-made earmolds (cf. Kulkarni and Colburn (1998) and the left picture in Figure 1.7) that at the same time block the entrance of the ear. Alternatively, the microphones can also be mounted into conventional foam earplugs (cf. Møller *et al.* (1995) and the right picture in Figure 1.7), which was shown to be equally suitable compared to custom-made earmolds by Raufer *et al.* (2013) and hence will also be used in the following studies. As a rule of thumb resulting from the acquired experience in measuring individual HRTFs, the microphones should be mounted behind the tragus in the medial direction (i.e. in the direction to the eardrum) to capture the relevant spatial aspects of HRTFs and to enable an appropriate binaural reproduction.

Having an appropriate mounting for the microphones, individual HRTFs can be measured using various techniques and broadband stimuli, such as e.g. sweeps (cf. Majdak *et al.* (2007), Köhler *et al.* (2014)), impulses (cf. Mellert *et al.* (1974)) or broadband noise (cf. Enzner (2009), Köhler *et al.* (2014)), only to name a few. In the following studies all transfer functions were estimated using white noise signals and standard FFT-based techniques. Unless otherwise stated, an $H_1$ estimate (cf. Mitchell (1982)) with a 8192-point Hann window, 50% overlap, 52 averages at a sampling frequency of $f_s = 44100$ Hz was used. The white noise stimulus was chosen since noise signals are usually associated with less annoyance for the subject compared to sweeps and impulses.

All HRTFs were measured in the anechoic room of the *Institut für Hörtechnik und Audiologie* at the University of Applied Sciences (Jade Hochschule) in Oldenburg. The subject was seated in the center of a circular loud-

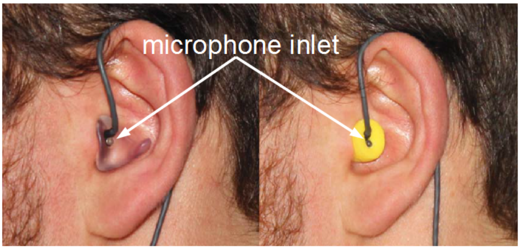


**Figure 1.7:** Knowles FG-23329 miniature electret microphones mounted in an individual earmold (left) and in a foam earplug (right). Illustration adapted from Raufer *et al.* (2013).

speaker array consisting of 24 uniformly distributed loudspeakers (equidistant 15° spacing) at a radius of 1.25 m in the horizontal plane (cf. Appendix C for further details). In order to limit the risk of reflections by the experimental apparatus, small loudspeakers with a diameter of 7 cm (custom-made one-way loudspeakers) and a very light support structure were used. All individual HRTFs in the following studies are measured in the horizontal plane (i.e. with an elevation angle of $\delta = 0°$ relative to the interaural axis).

The free field pressure signal $p_0(t, \theta)$ was measured with the same loudspeakers (for each loudspeaker individually) and a calibration microphone (G.R.A.S. Microphone Type 40AF pointed towards each of the measured loudspeakers with $\delta = 0°$ elevation) at the position of the center of the head without the subject being present.

## 1.4 Binaural reproduction

Since HRTFs/hrirs incorporate the relevant aspects of the sound transmission from a certain direction to the ears of a listener, the virtual reproduction of the associated spatial sound perception from this direction may be synthesized by convolving a (monaural) signal s(t) with the hrirs and playing back the resulting signals via headphones. The reproduction of binaural signals via headphones is *straightforward as headphones are able to deliver each binaural channel independently to each ear* (Masiero (2012)), in comparison to a scenario where loudspeakers (including crosstalk cancellation, cf. Kirkeby and Nelson (1999), Parodi (2010)) are used to reproduce binaural signals, for instance. In a binaural reproduction (cf. Figure 1.8), the spatial information is gathered in the associated HRTFs for a fixed direction. Alternatively, instead of using fixed directions in the binaural reproduction, one could use an artificial head for binaural recordings which inherently would incorporate the spatial information associated with the

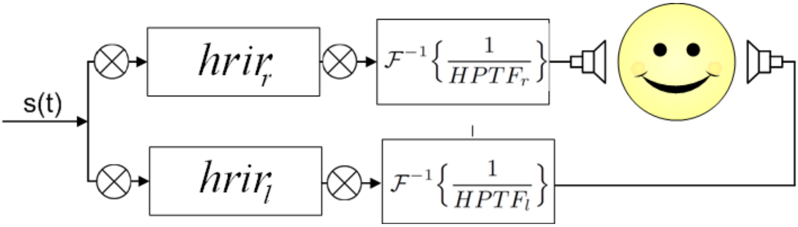

s(t)
$hrir_r$
$hrir_l$
$\mathcal{F}^{-1}\left\{\frac{1}{HPTF_r}\right\}$
$\mathcal{F}^{-1}\left\{\frac{1}{HPTF_l}\right\}$

**Figure 1.8:** Binaural reproduction of the monaural signal s(t) by convolution (⊗) with the hrirs and with the equalization of the HPTFs for both ears separately.

HRTFs of the artificial head used.

Binaural headphones presentation inherently introduces a certain spectral coloration that not only depends on the headphone, but also on the subject (cf. Møller *et al.* (1995), Schmidt (2009)) and on the position of the headphone (cf. Kulkarni and Colburn (2000) and Figure 1.9). Hence, the individual spectral characteristics of the used headphones need to be equalized in order to authentically synthesize a binaural reproduction (cf. Kim and Choi (2005)), as illustrated schematically in Figure 1.8. To equalize the characteristics of the headphones one has to design a headphone equalization filter, which has the inverse characteristics of the transfer function of the headphones, when measured in the physical point that was also used for the HRTF measurement (cf. Hammershøi and Hoffmann (2011)).

Analogously to the HRTFs, the (direction-independent) headphone transfer function (HPTF) is given as

$$\mathrm{HPTF}_{l/r}(f) = \frac{\mathcal{F}\{p^{hp}_{l/r}(t)\}}{\mathcal{F}\{s^{hp}_{l/r}(t)\}}, \tag{1.3}$$

with $p^{hp}_{l/r}(t)$ denoting the sound pressure signal at the ear when the headphone is driven with the electrical signal $s^{hp}_{l/r}(t)$. Subsequently after measuring the HRTFs using the setup depicted in Figure 1.7, the sound pressure signals $p^{hp}_{l/r}(t)$ and hence the HPTFs can be measured with the microphones left in place. By doing so, individual scattering and/or reflection effects due to the microphone mounting is included in both transfer functions (HRTFs and HPTFs) and hence equalized with the binaural reproductions (cf. Figure 1.8).

It has to be mentioned that the sound pressure measured outside the blocked ear is (only) sufficient to describe the transfer characteristics of headphones with a free air equivalent coupling (FEC) to the ear, i.e. headphones that have a coupling to the ear similar to the coupling to free air (Møller *et al.* (1995)). Whereas the simplification when designing an equalization filter with the inverse characteristics of the headphone transfer function (measured in the physical point of the blocked ear where the HRTFs were measured) *presumes that the acoustical coupling of the headphones and ear canal is similar to the coupling of the ear canal to free air* (Hammershøi and Hoffmann (2011)). This has been shown to be largely the case for several headphones (cf. Møller *et al.* (1995), Ćirić and Hammershøi (2006), Masiero (2012)) and will be further assumed to be valid for the used headphones (AKG-240-K Studio) in the following studies.

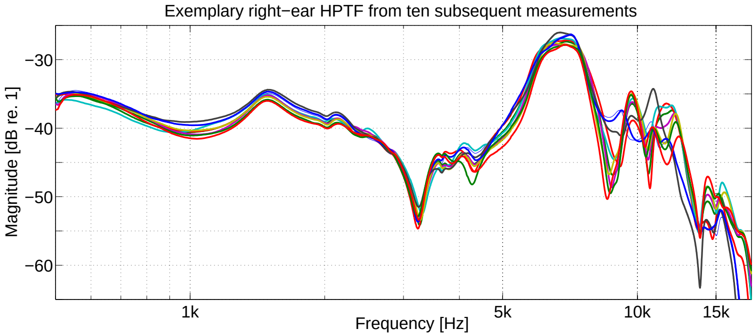


**Figure 1.9:** Exemplary HPTFs of the AKG-240-K Studio headphones measured 10 times subsequently (the headphone was taken off and on before each repetition) on an artificial head (Head Acoustics HMS-II).

In order to illustrate the intra-individual variability of HPTFs when slightly varying the position of the headphone, HPTFs were measured repeatedly with the headphone taken off and on before each repetition (cf. Figure 1.9). HPTFs were measured in ten repetitions, using AKG-240-K Studio headphones put on an artificial head (Head Acoustics HMS-II). Obviously, this resulted in a large variability in the magnitude of the HPTFs. It is apparent that primarily at higher frequencies the spectral notches vary clearly in amplitude but also in frequency. Consequently, it seems reasonable to assume that the inversion of these HPTFs would result in a significantly different perception (cf. Paquier *et al.* (2011)), partly dominated by spectral peaks. To avoid excessively large peaks in the inversion of HPTFs and at the same time to achieve a certain robustness to small fitting variations, a straightforward inversion of measured HPTFs is unfavorable. Instead, an appropriate preselection of measured HPTFs and/or a regularized inversion of measured HPTFs seems advantageous (cf. Møller *et al.* (1995), Kirkeby *et al.* (1998), Kirkeby and Nelson (1999), Schärer and Lindau (2009) and Masiero (2012)).

In this thesis, the HPTFs were measured repeatedly until a dynamic range of less than 25 dB for both ears and frequencies 2 kHz $\leq f \leq$ 16 kHz was obtained. Additionally to this rather heuristic preselection, the inversion was accomplished using the regularized inversion method described in Kirkeby and Nelson (1999), where the regularization parameter $\alpha_{\mathrm{inv}}$ determines the robustness of the resulting equalization filter. Suitable values for $\alpha_{\mathrm{inv}}$, however, depend on the overall magnitude of the HPTFs, the spectral notches etc. Hence, specific values of the regularization parameter $\alpha_{\mathrm{inv}}$ will be provided in the respective sections, if the regularized inversion procedure was applied.

## 1.5 Spatial filtering using microphone arrays

Microphone arrays, consisting of multiple microphones, enable to capture a sound field at different positions. By linearly combining the microphone signals (also known as beamforming, cf. Van Veen and Buckley (1988), Brandstein and Griebel (2001), McCowan (2004), Benesty *et al.* (2008), Doclo *et al.* (2010)), it is possible to achieve a certain spatial selectivity. The main principle is based on constructive and destructive interference of the signals arriving at the microphone array, depending on the different direction of incidence. Beamforming is often used to enhance a signal from a specific direction by attenuating signals from other directions.

Spatial filtering is typically achieved using a filter-and-sum structure, depicted in Figure 1.10. Consider a line array with N microphones (with $d_n$ the distance between the $n$-th microphone and the center of the microphone array) and a source $\mathrm{S}(f,\theta)$ impinging on the array from an angle/direction $\theta$. Each microphone signal $Y_n(f,\theta)$, $n = 1\ldots\mathrm{N}$, is weighted with the complex-valued filter coefficient[1] $w_n^*(f)$ and summed to yield the output signal $\mathrm{Z}(f,\theta)$, which is given as

$$\mathrm{Z}(f,\theta) = \sum_{n=1}^{\mathrm{N}} w_n^*(f) Y_n(f,\theta) = \mathbf{w}^H(f)\mathbf{Y}(f,\theta), \tag{1.4}$$

with the N×1-dimensional vectors $\mathbf{w}(f) = [w_1(f)\; w_2(f)\ldots w_{\mathrm{N}}(f)]^T$ and $\mathbf{Y}(f,\theta) = [Y_1(f,\theta)\; Y_2(f,\theta)\ldots Y_{\mathrm{N}}(f,\theta)]^T$. The spatial directivity pattern of the filter-and-sum beamformer in Figure 1.10 is given as

$$\mathrm{H}(f,\theta) = \frac{Z(f,\theta)}{\overline{Y}(f,\theta)} = \mathbf{w}^H(f)\mathbf{d}(f,\theta), \tag{1.5}$$

with $\overline{Y}(f,\theta)$ denoting the signal arriving at the center of the microphone array and the N×1-dimensional steering vector $\mathbf{d}(f,\theta)$ denoting the frequency- and direction-dependent transfer functions between a source from direction $\theta$ and the N microphones.

Assuming far-field conditions (i.e. planar wave propagation) and an equal attenuation for all microphones, the relative delay between the $n$-th microphone and the center of the microphone array is given by

$$\tau_n(\theta) = \frac{d_n \cdot \cos(\theta)}{c}, \tag{1.6}$$

---

[1] To simplify notation it is conventional to use conjugate filter coefficients (cf. Van Veen and Buckley (1998)).

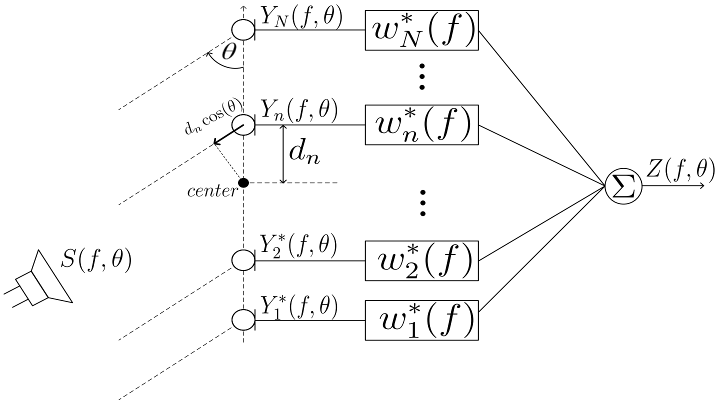


**Figure 1.10:** Exemplary illustration of spatial filtering (using a filter-and-sum structure) with a line array (Illustration adapted from Doclo and Moonen (2003a)).

with $c$ the speed of sound ($c$ mainly depends on the temperature and is approximately 343 $\frac{\mathrm{m}}{\mathrm{s}}$ in air at 20 °C). The steering vector is there given by

$$\mathbf{d}(f,\theta) = \left[\, e^{-2\pi\, f\tau_1(\theta)\, i} \;\; e^{-2\pi\, f\tau_2(\theta)\, i} \;\ldots\; e^{-2\pi\, f\tau_{\mathrm{N}}(\theta)\, i} \,\right]^T . \tag{1.7}$$

In this thesis, both theoretically derived steering vectors $\mathbf{d}(f,\theta)$ under far-field conditions according to equation 1.7 (cf. **Chapter 2** and **3**) as well as measured steering vectors (cf. **Chapter 3** and **6**) will be used.

A frequently used beamformer is the so-called delay-and-sum beamformer (cf. McCowan (2004), Bitzer and Simmer (2001)), where the filter coefficients compensate for the inter-microphone delay for a signal arriving from direction $\phi$, i.e. $w_n(f) = \frac{1}{\mathrm{N}} \cdot e^{-2\pi\, f\tau_n(\phi)\, i}$ . In this case, the signals arriving from direction $\phi$ and $360^\circ - \phi$ are added constructively (i.e. in phase), whereas signals arriving from other directions are added not constructively[2], resulting in a frequency-dependent directivity pattern. Hence, a delay-and-sum beamformer enables to achieve a preferential direction of the output signal $\mathrm{Z}(f,\theta)$. For synthesizing directivity patterns associated

---

[2] Note that if the distance between the microphones is too large, the constructive superposition of the signals will also occur for other angles than $\phi$ and $360^\circ - \phi$, an ambiguity called spatial aliasing.

with individual HRTFs, we will, however, use the more general filter-and-sum structure depicted in Figure 1.10. It should be noted that a linear alignment of the microphones results in directivity patterns that are symmetric with respect to the microphone axis. Hence, the topology of the microphone array (i.e. the spatial positioning of the microphones) significantly affects the achievable directivity patterns and thus also the performance of the beamformer, an aspect that will be further examined in **Chapter 2**.

The objective of synthesizing desired directivity patterns associated with human HRTFs using microphone arrays was already described in previous studies either in the spatial domain considering discrete directions (cf. Mellert and Tohtuyeva (1997), Tohtuyeva and Mellert (1999), Kahana *et al.* (1999), Sakamoto *et al.* (2008), Haut (2009)) or using a spherical harmonics decomposition (cf. Zotkin *et al.* (2009), Atkins (2011a,b), Castaneda *et al.* (2013), Sakamoto *et al.* (2013)). In this thesis, the optimization in the spherical harmonics domain will not be considered, since it is known to require a very large number of measurement directions and microphones for appropriate model orders (cf. Castaneda *et al.* (2013)).

For the considered VAH application, the individual (frequency-dependent) HRTFs for the left and the right ear (cf. Figure 1.3) are considered as the desired directivity patterns for a filter-and-sum beamformer. Hence, the beamformer filter coefficients should be designed such that the synthesized directivity patterns (blue and red dashed lines in Figure 1.11) resemble the desired HRTF directivity patterns at the two ears (solid lines in Figure 1.11). The output signals of the VAH may then be considered as binaural recordings associated with these individual HRTFs. Analogous to the described binaural reproduction in Figure 1.8, these output signals (red and blue arrows in Figure 1.11) would further need to be headphone-equalized (right blocks in Figure 1.8) in order to yield authentic binaural reproductions via headphones.

The filter coefficients $\mathbf{w}(f)$ are now computed such that the resulting directivity pattern of the VAH $\mathrm{H}(f,\theta)$ resembles a desired directivity pattern $\mathrm{D}(f,\theta)$ for all $f$ and $\theta$, which is usually achieved by minimizing a cost function, for which different design criteria have been proposed (cf. Nordebo *et al.* (1994), Kajala and Hamaldinen (1999) and Doclo and Moonen (2003a)). In this thesis we will use the narrowband least squares cost function $\mathrm{J}_{\mathrm{LS}}$, i.e. the sum over P directions of the squared absolute differences between $\mathrm{H}(f,\theta)$ and $\mathrm{D}(f,\theta)$, i.e.

$$\boxed{\mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f)) = \sum_{i=1}^{\mathrm{P}} F(f,\theta_i)\left|\mathbf{w}^H(f)\mathbf{d}(f,\theta_i) - \mathrm{D}(f,\theta_i)\right|^2}, \tag{1.8}$$

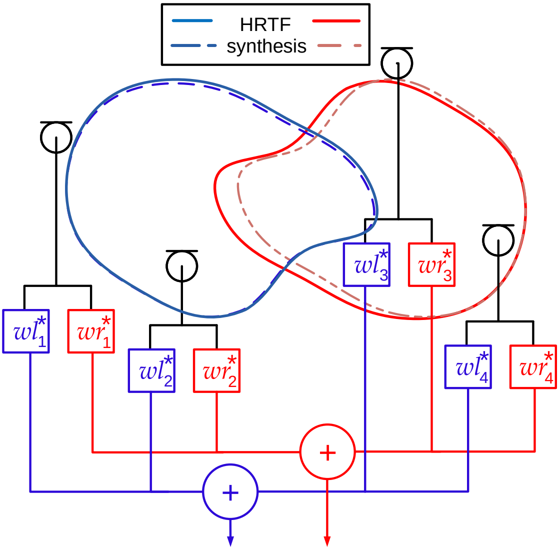


**Figure 1.11:** HRTFs as the desired directivity patterns with an exemplary microphone array with N = 4 microphones and the complex-valued weights $\mathbf{w}(f) = [wl_1(f) \dots wl_4(f)]^T$ for the left ear and $\mathbf{w}(f) = [wr_1(f) \dots wr_4(f)]^T$ for the right ear.

with $F(f, \theta)$ a real-valued positive weighting function that assigns more or less importance to certain directions. Note that it would be ineffective to apply a frequency-dependent weighting within the narrowband cost function in equation 1.8. A frequency-dependent weighting would only be effective for a broadband cost function incorporating multiple frequencies (cf. Doclo and Moonen (2003b)). Such a broadband cost function will, however, not be considered in this thesis due to its unfavorable numerical properties for a large number of microphones and fine spectral resolution. Hence, a decoupled (narrowband) optimization that yields the filter coefficients for a single frequency $f$ will be considered throughout this thesis.

A major advantage of the considered least squares cost function is its closed-form solution. The narrowband least squares cost function in equation 1.8 can be rewritten as

$$\mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f)) = \mathbf{w}^H(f)\mathbf{Q}(f)\mathbf{w}(f) - \mathbf{w}^H(f)\mathbf{a}(f) - \mathbf{a}^H(f)\mathbf{w}(f) + d(f), \quad (1.9)$$

with

$$\mathbf{Q}(f) = \sum_{i=1}^{\mathrm{P}} F(f,\theta_i) \cdot \mathbf{d}(f,\theta_i)\mathbf{d}^H(f,\theta_i),$$

$$\mathbf{a}(f) = \sum_{i=1}^{\mathrm{P}} F(f,\theta_i) \cdot \mathbf{d}(f,\theta_i)\mathrm{D}^*(f,\theta_i), \text{ and}$$

$$d(f) = \sum_{i=1}^{\mathrm{P}} F(f,\theta_i) \cdot \left|\mathrm{D}(f,\theta_i)\right|^2. \tag{1.10}$$

The filter coefficients minimizing the cost function $\mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f))$ can be derived by setting the gradient $\nabla_{\mathbf{w}(f)}\mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f))$ to zero, yielding

$$\boxed{\mathbf{w}(f) = \mathbf{Q}^{-1}(f) \cdot \mathbf{a}(f)}. \tag{1.11}$$

## 1.6 Regularization

It has been shown that the filter coefficients in equation 1.11 may lead to an undesirable amplification of spatially uncorrelated noise at the microphones and may cause large errors of the synthesized directivity patterns even for small deviations from the assumed microphone positions and/or characteristics (cf. Cox *et al.* (1986), Doclo and Moonen (2003b), Doclo and Moonen (2007), Rasumow *et al.* (2011b) or **Chapter 2**, Levin *et al.* (2013)). To increase the robustness of a beamformer it is common practice to impose a certain regularization constraint on the filter coefficients. Due to its physical interest, the output power of the beamformer for the desired acoustic field in comparison to the output power for spatially uncorrelated noise, defined as the white noise gain (WNG), is a common measure for quantifying its robustness (cf. Cox *et al.* (1986), Bitzer and Simmer (2001)). The white noise gain for filter coefficients $\mathbf{w}(f)$ and direction $\theta$ is defined as (see e.g. Bitzer and Simmer (2001)),

$$\mathrm{WNG}(\mathbf{w}(f),\theta) = \frac{\left|\mathbf{w}^H(f)\mathbf{d}(f,\theta)\right|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}. \tag{1.12}$$

In general, a larger WNG is associated with a larger attenuation of spatially uncorrelated noise in comparison to the output power for the considered direction $\theta$ and hence with an increased robustness. It is worth noting that the WNG depends on the filter coefficients $\mathbf{w}(f)$ and on the direction $\theta$. Various regularization strategies to increase the robustness of the VAH synthesis will be presented and discussed in **Chapter 3**, where one chosen regularization strategy will be perceptually evaluated in **Chapter 6**.

## 1.7 Modeling aspects of spatial hearing

Auditory models are an important aspect of hearing research. *An auditory model is a mathematical algorithm that mimics a part of the human auditory system* (Søndergaard and Majdak (2013)). The objective of auditory models is mainly to validate and to explain the characteristics of psychoacoustic data by means of theoretical assumptions and to improve the understanding of the functioning of the auditory system. *Both our general understanding and the fragmental knowledge of details known from hearing research can be reconstructed and tested in the form of functional* (auditory) *models* (Karjalainen (1987)). Furthermore, auditory models can be used to assess the relevance of individual aspects, like the suitability of spatial sound reproductions. In this thesis, auditory models will be used to objectively analyze the VAH-synthesis and to examine the perceptual limits when processing individual HRTFs.

In the context of spatial hearing, many auditory models evaluate the temporal interaural differences and coherent components of both ear signals (Blauert (1997)). A well-known model aiming to explain how temporal cues are extracted for sound localization is the so-called Jeffress-model (cf. Jeffress (1948)). This model basically analyzes the coincidences in the auditory pathways from the left and right ear. An extension of the Jeffress-model to a binaural signal detection model (including a prior peripheral and a posterior central stage) has been proposed in Breebaart *et al.* (2001), where the contralateral inhibition of ipsilateral signals is considered. A comparable model in Le Goff *et al.* (2013) additionally considers a potential hearing impairment of the listener. These models basically combine quantitative (monaural) models for signal processing in the auditory system (cf. Dau *et al.* (1996), Jepsen *et al.* (2008)) with binaural aspects, for instance, associated with the Jeffress-model. Such comprehensive models may be used to mimic the detection and/or discrimination of binaural cues. Also the modeling of localization is often based on the interaural differences at the ears (cf. Dietz *et al.* (2011)), for instance, with an additional learning of the complex patterns resulting from frequency-dependent ITDs and ILDs (cf. May *et al.* (2011)).

Despite the availability of sophisticated auditory models, in this thesis primarily the discrimination ability of chosen cues when smoothing individual HRTFs and the perceptual suitability of a VAH-synthesis are of interest. In line with the mentioned models, a simple model to describe the individual discrimination ability of IPDs is introduced in **Chapter 5**. This model basically emphasizes the importance of IPDs at lower frequencies and supports the approach to simplify the HRTF phase only at higher frequencies, which enables a complex-valued smoothing in relative bandwidths. The discrimination of smoothing HRTFs in relative bandwidths is in good agreement with the model assumption that the human auditory system groups incoming sounds into so-called critical frequency bands (cf. Fletcher (1940)) that broaden with increasing center frequencies.

As an objective measure to quantify the error of the VAH-synthesis, the monaural (absolute) dB-error between the synthesized directivity patterns $\mathrm{H}(f,\theta)$ and the desired directivity pattern $\mathrm{D}(f,\theta)$ for each frequency $f$ and direction $\theta$ will be used, i.e.

$$\Delta_{\mathrm{dB}}(f,\theta) = \left| 20\log_{10}\left(\left|\frac{\mathrm{D}(f,\theta)}{\mathrm{H}(f,\theta)}\right|\right)\right| . \tag{1.13}$$

In addition, the dB-error $\Delta_{\mathrm{dB}}(f,\theta)$ is averaged in frequency bands within equivalent rectangular bandwidths (ERB, cf. Moore and Glasberg (1983)) that mimic the spectral resolution of the auditory system, i.e.,

$$\boxed{\epsilon(f_c,\theta) = \frac{1}{\mathrm{L}}\sum_{l=1}^{\mathrm{L}} \Delta_{\mathrm{dB}}(f_l,\theta)} . \tag{1.14}$$

The mean absolute dB-error $\epsilon(f_c,\theta)$ for direction $\theta$ and center frequency $f_c$ of the associated ERB-band is computed as the mean over all frequencies in this ERB-band, with $f_1$, $f_c$ and $f_{\mathrm{L}}$ denoting the lowest frequency, the center frequency and the highest frequency within this ERB, respectively. The spacing of the ERB-bands was chosen with a 50% overlap for adjacent bands (cf. Rasumow *et al.* (2014c) or **Chapter 5**).

There are several reasons for using a monaural error measure in the context of HRTF-synthesis. The monaural error measure in equation 1.13 and 1.14 was chosen due to its simplicity and since binaural shortcomings of the (monaurally optimized) synthesis are also reflected in the monaural signals (i.e. a perfect monaural synthesis for both ears implies perfect interaural differences, but not vice versa). In addition, based on preliminary listening tests, the main shortcomings associated with the VAH-synthesis can be attributed to spectral coloration rather than to a spatial misalignment. Hence, a monaural measure will be used for objective evaluations, accompanied by perceptual evaluations.

## 1.8 Outline of this thesis

This thesis is roughly divided into three thematically different parts:

Part I mainly deals with signal processing aspects associated with HRTF-synthesis using microphone arrays. In **Chapter 2**, the impact of the used microphone topology on the accuracy and the robustness of the VAH-synthesis when assuming small positioning errors of the microphones is evaluated. **Chapter 3** deals with the impact of different regularization strategies, including a modified optimization and regularization design with a psychoacoustically motivated adaptation of the bandwidth. Different regularization strategies are presented and evaluated objectively.

In part II, the perceptual limits of the auditory system regarding the temporal (**Chapter 4**) and the spectral and spatial resolution (**Chapter 5**) of individual HRTFs is investigated. The aim of these studies is to derive a smoothing procedure that enables to reduce the temporal, spectral and spatial resolution of HRTFs as far as possible without yielding discriminable artifacts. Such imperceptible smoothing procedures of HRTFs prior to their VAH-synthesis are shown to enhance the performance of the VAH in terms of robustness and/or the number of microphones required to achieve a desired accuracy.

Part III deals with the evaluation of individualized synthesis using the VAH. In **Chapter 6**, the perceptual impact when varying the regularization constraint in terms of the desired mean white noise gain (i.e. averaged over all considered directions) of the VAH-synthesis is evaluated. In **Chapter 7**, a perceptual study evaluating individual HRTFs, individualized synthesis with the VAH and non individual synthesis using traditional artificial heads in comparison to free field playback is presented. The resulting perceptual ratings are interpreted with regard to possibilities for further improving the VAH performance.

General conclusions with regard to the performed studies including suggestions for improvement and an outlook for further investigations are presented in **Chapter 8**.

Please note that this thesis is mainly based on single journal and conference publications and hence contains certain redundancies (e.g. in the introductions of the respective chapters) in this merged version.

# Part I

# Optimizing microphone arrays to synthesize individual head-related directivity patterns

In part I, aspects associated with the implementation of appropriate microphone array topologies and the optimization and regularization of the beamformer filter coefficients when synthesizing individual HRTFs using microphone arrays are discussed. This design procedure is related to existing beamformers (e.g. filter-and-sum beamformer, cf. Doclo and Moonen (2003b)). However, several aspects specifically tailored for optimizing and regularizing filter coefficients as well as deriving appropriate microphone topologies for the VAH are discussed in the next two chapters.

The impact of the microphone array topology on the robustness and the accuracy of the VAH-synthesis is examined objectively in **Chapter 2**.

Multiple ways to increase the robustness of the VAH-synthesis against small errors of the microphone characteristics (i.e. gain, phase and/or positioning errors) are presented in **Chapter 3**. Appropriate regularization approaches for synthesizing multi-directional HRTFs are presented and evaluated objectively.

# 2

# Microphone positioning for the virtual artifcial head*

---

*This chapter is based on the publications

- Rasumow, Blau, Doclo, Püschel, Hansen, Mellert, van de Par, and Püschel (2011a), "Zahlentheoretisch motivierte Optimierung von Mikrofonpositionen für einen virtuellen Kunstkopf," in *Fortschritte der Akustik - DAGA 2011*, Düsseldorf, Germany, pp. 637-638.
- Rasumow, Blau, Hansen, Doclo, van de Par, Mellert, and Püschel (2011b), "Robustness of virtual artificial head topologies with respect to microphone positioning errors," in *Proc. Forum Acusticum, Aalborg*, Aalborg, pp. 2251-2256.

*In this chapter, a microphone array is presented that can be used to approximate the frequency-dependent directional characteristics of an artificial head. The desired HRTFs can be realized by a set of appropriate filters. Such a setup may be referred to as a virtual artificial head. Virtual artificial heads are much more flexible than real artificial heads, since e.g. the filters can be adjusted to match an individual set of HRTFs. However, virtual artificial heads are sensitive to small errors in the characteristics and the position of the microphones. In the present study, the relevance of different microphone topologies in terms of robustness with respect to positioning errors of the microphones is investigated. First, a method for optimizing the microphone positions based on a Golomb array topology, which originates from number-theoretic considerations, is introduced. The method successively computes a set of microphone positions with the possibility to vary the number of microphones without changing the general topology. Second, the robustness against positioning errors is improved by applying a norm constraint for the computation of the filter coefficients for a two-dimensional array. It is shown by numerical simulations, that both procedures considerably improve the robustness.*

## 2.1 Introduction

The use of so-called artificial heads, which are a replica of real human heads, is common practice today. Alternatively, the desired frequency-dependent beam pattern of human head-related transfer functions (HRTFs) can also be approximated by a microphone array with appropriate filters (cf. for instance Mellert and Tohtuyeva (1997), Atkins (2011a)). Such a setup may be referred to as a virtual artificial head (VAH). The resulting directivity pattern of the VAH does not only depend on the filter coefficients but also on the number and the topology of the microphones used in the array. Thus, microphone positioning is a crucial step in the setup of a VAH. In this study a method which originates from number-theoretic considerations is introduced for a planar array and compared to traditional methods with respect to accuracy and robustness.

## 2.2 Calculation of filter coefficients

Consider a given (two-dimensional) directivity pattern $\mathrm{D}(f, \theta)$, depending on frequency $f$ and direction of arrival $\theta$. The analytical derivation of the N-dimensional filter coefficients $\mathbf{w}(f)$, with N the number of microphones, can e.g. be performed by minimizing a least squares cost function (cf. section 1.5) that relates the general desired directivity pattern $\mathrm{D}(f, \theta)$ to the resulting directivity pattern $\mathrm{H}(f, \theta)$ (cf. Figure 1.11). The resulting directivity $\mathrm{H}(f, \theta)$ in turn depends on the steering vector $\mathbf{d}(f,\theta)$ of the microphones and the filter coefficients $\mathbf{w}(f)$ (cf. equation 1.5).

In this chapter, the steering vector $\mathbf{d}(f,\theta)$ was derived analytically (according to equation 1.7) assuming omnidirectional microphones, no microphone shielding effects (i.e. the microphones do not alter the directivity of adjacent microphones) and far-field sound propagation.

In general, a chosen cost function can either be minimized for all frequencies simultaneously (broadband optimization) or separately for each frequency bin (narrowband optimization). In the following computations the narrowband least squares cost function $\mathrm{J}_{\mathrm{LS}}$ from equation 1.8 with a uniform weighting $F(f,\theta_i) = 1$ for all $i$ was used. The filter coefficients minimizing the least squares cost function $\mathrm{J}_{\mathrm{LS}}$ from equation 1.8 are given in the equations 1.10 and 1.11.

## 2.3 Random Sampling method

The resulting directivity pattern $\mathrm{H}(f,\theta)$ of the VAH depends on the desired directivity pattern $\mathrm{D}(f,\theta)$ and the number and the steering vector of the microphones. Assuming independent omnidirectional microphones (which is a fair assumption for an array with small electret microphones), the steering vector $\mathbf{d}(f,\theta)$ depends only on the relative position of the microphones (cf. equation 1.6).

Assuming further there is a predetermined grid of possible microphone positions, the best microphone topology could be determined using a brute force approach, i.e. by computing the resulting $\mathrm{H}(f,\theta)$ for each possible combination of microphone positions and choosing the one with the lowest value of the cost function. This eventually would lead to the optimal microphone topology, but even with a moderate number of microphones and grid density, the time required to probe all possible combinations of microphone positions is prohibitive.

Instead of systematically probing each possible topology one could generate topologies at random. This will be referred to as the "Random Sampling" method. In this study, $\kappa_{top} = 10000$ randomly chosen topologies on a $x \times y$ plane within a 10 cm × 10 cm plane were compared, and the one with the lowest least squares error was finally selected. The number of $\kappa_{top} = 10000$ different topologies was chosen as a reasonable compromise between all possible topologies and computational cost.

## 2.4 Golomb method

In general, the microphone topology or, more specifically, the inter-microphone distances determine the frequency-dependent beamwidth of a beamformer. Since the number of microphones is limited, the topology of choice must present a compromise between spectral and directional accuracy. A good basis for an optimal broadband fit to the frequency- and angle-dependent directivity pattern $\mathrm{D}(f,\theta)$ would be a topology with as many as possible different inter-microphone distances in all possible directions. A topology that fulfills the requirement of as many as possible inter-microphone distances in one dimension is the so-called Golomb-ruler (cf. Golomb and Taylor (1982)). A Golomb-ruler is a numerical series that does not contain any inter-mark distance more than once. A possible method to expand this topology to a two- or three-dimensional grid was proposed in DE 10 2010 012 388 A1. The procedure for determining a set of valid nodes (intersec-

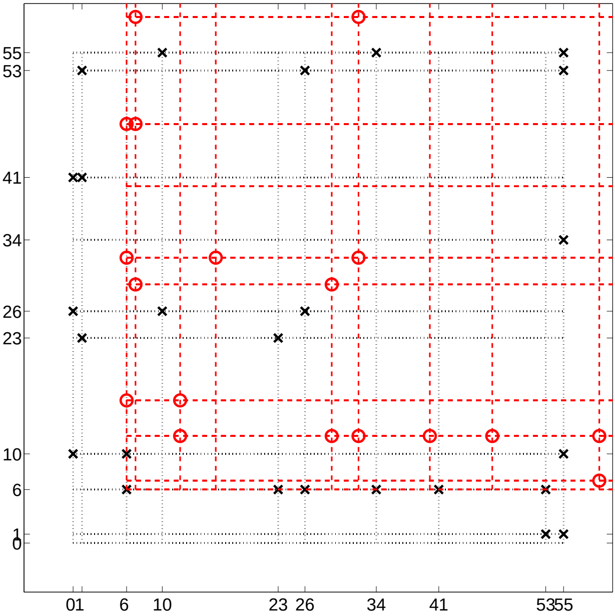


**Figure 2.1:** Procedure to derive a set of valid nodes according to DE 10 2010 012 388 A1. The dotted black lines indicate a two-dimensional grid originating from a Golomb-ruler with order $G = 10$. Black crosses indicate chosen nodes and the red lines show a shifted copy of the chosen nodes.

tion of chosen marks in a two- or three-dimensional topology, cf. × and ○ in Figure 2.1) according to DE 10 2010 012 388 A1 can be subdivided (for a two-dimensional topology) into seven steps:

1. Create a two-dimensional grid using the same one-dimensional Golomb-ruler in both dimensions (see gray structure in Figure 2.1)
2. Select a first reference node on the grid arbitrarily
3. Select a second test node on the grid arbitrarily
4. Copy and shift the topology along both dimensions by all possible distances (see red structure in Figure 2.1)
5. Check condition: Shifted nodes must not coincide with more than one of the previously selected nodes within each shift
6. Keep the test node as an additional reference node if (5) is true. Repeat from (3) if (5) is false.
7. Repeat this procedure until all additional nodes were tested or a desired number of valid nodes N has been reached.

Note that since the reference and test nodes are chosen randomly, the position and number of valid nodes may vary with each new run of the method.

In general, the Golomb method is less computationally expensive compared to the Random Sampling method. Furthermore, a Golomb topology is also valid for microphone numbers $<$ N since the small-number topologies are a subset of the large-number topologies. This feature may be of advantage when varying the number of microphones (e.g. by setting chosen filter coefficients to zero) with a fixed microphone array.

## 2.5 Regularization

When the filter coefficients $\mathbf{w}(f)$ are computed using a non-regularized cost function (cf. equation 1.8), it is known that small changes of the microphone characteristics (gain/phase) and the microphone positions can drastically deteriorate the resulting directivity patterns $\mathrm{H}(f, \theta_i)$ (cf. Cox *et al.* (1986), Doclo and Moonen (2003b, 2007)). Even if the microphone characteristics are known very accurately (e.g. using a proper calibration), there could still be changes in the position of the microphones, e.g. due to environmental parameters. In order to make the procedure more robust, different regularization techniques have been proposed. A well-known regularization technique is the so-called diagonal loading (a special case of Tikhonov regularization, see for instance Cox *et al.* (1987)), in which the cost function in equation 1.8 is replaced by

$$\mathrm{J}_{\Gamma}(\mathbf{w}(f)) = \mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f)) + \mu \cdot \underbrace{\mathbf{w}^{H}(f)\,\mathbf{w}(f)}_{\Gamma(f)} \tag{2.1}$$

where $\Gamma(f)$ represents the square norm of the filter coefficients and $\mu$ is the Lagrange multiplier. The filter coefficients minimizing the cost function in equation 2.1 are given by

$$\mathbf{w}_{\Gamma}(f,\mu) = \left(\mathbf{Q}(f) + \mu \cdot \mathbf{I}_{\mathrm{N}}\right)^{-1} \cdot \mathbf{a}(f), \tag{2.2}$$

where $\mathbf{I}_{\mathrm{N}}$ represents the $\mathrm{N} \times \mathrm{N}$-dimensional unity matrix. The regularization parameter $\mu$ has to be carefully chosen, such that on the one hand it provides robustness against, e.g., microphone position errors and on the other hand the resulting directivity pattern does not strongly deviate from the desired directivity pattern. A common way to determine an appropriate value for $\mu$ is to impose a constraint on the robustness measure of the filter coefficients, i.e. $\Gamma(f,\mu) \leq \beta_{\Gamma}$ with $\beta_{\Gamma}$ the maximum desired value for the square norm of the filter coefficients $\Gamma$. The parameter $\mu$ is derived by minimizing $(\Gamma(f,\mu) - \beta_{\Gamma})^2$, see for instance Dörbecker (1998). $\beta_{\Gamma}$, in contrast, has to be chosen manually according to the expected error of the steering vector. In general, $\beta_{\Gamma}$ should be proportional to the desired accuracy and reciprocal to the resulting robustness. This means that lower $\beta_{\Gamma}$ enhance the robustness at the expense of accuracy and vice versa.

## 2.6 Numerical Simulation

The left-ear HRTF of subject 1005 from the IRCAM database[1] was used as the desired directivity pattern $\mathrm{D}(f,\theta)$ in the horizontal plane for a microphone array with $\mathrm{N} = 4,6,8,10,12,14,16,18,19,20,22,23,24$ microphones using both introduced methods (Random Sampling, Golomb) for microphone positioning. In order to compare the accuracy associated with the positioning methods, the synthesis error $\epsilon_f(f_c)$ was used, with

$$\epsilon_f(f_c) \quad = \quad \frac{1}{\mathrm{P}} \sum_{i=1}^{\mathrm{P}} \epsilon(f_c,\theta_i)\ . \tag{2.3}$$

---

[1] The HRTF-database is available on http://recherche.ircam.fr

Here, $\epsilon(f_c, \theta_i)$ denotes the synthesis error between the resulting directivity pattern $\mathrm{H}(f, \theta)$ and the desired directivity pattern $\mathrm{D}(f, \theta)$ averaged over frequencies in ERB-bands according to equation 1.14. This error measure is further averaged over all P directions, yielding $\epsilon_f(f_c)$ as a function of the center frequencies $f_c$ of the analyzed ERB-bands.

## 2.7 Results and discussion

As described before, even small positioning errors can cause huge deviations in the resulting directivity pattern (cf. for instance Doclo and Moonen (2003b)). In order to evaluate the effect of positioning errors, a uniformly distributed random shift in $x$- and in $y$- direction, both with maximum absolute values $|\varDelta\, x_{max}| = |\varDelta\, y_{max}| = 0.1$ mm, was superimposed onto the true microphone positions. The resulting error $\epsilon_f(f_c)$ with and without one representative realization of the randomized positioning error is shown in Figure 2.2. Note that the synthesis error $\epsilon_f(f_c)$ changed only marginally when randomly varying the positioning error for fixed $|\varDelta\, x_{max}|$ and $|\varDelta\, y_{max}|$. Since the same trends were observed for different realizations of the positioning error, only the synthesis error $\epsilon_f(f_c)$ for one representative realization of the positioning error is shown in Figure 2.2 and 2.3.

From Figure 2.2 (depicting the synthesis error $\epsilon_f$ as a function of frequency and number of microphones used) it can be observed that it seems advantageous to vary the number of microphones with frequencies since the directivity pattern $\mathrm{D}(f, \theta)$ is rather smooth at low and rather peaky at high frequencies. Thus it is of interest to investigate the minimal error $\epsilon_f(f_c)$ per critical band while the microphone number is assumed variable. This characterization will be referred to as the best-case scenario, shown in Figure 2.3.

### 2.7.1 Effect of positioning errors

The resulting synthesis errors when using the Random Sampling method both with unbiased and corrupted steering vectors are shown in the left column of Figure 2.2. Using the unbiased steering vector (top row in Figure 2.2), the resulting error is relatively small ($\approx 1$ dB), except at higher frequencies ($f \geq 3$ kHz) and a small number of microphone ($\mathrm{N} \leq\ 12$). This phenomenon is expected, since the directivity pattern is peakier at higher frequencies, requiring more microphones to accurately synthesize it.

When a (representative) random positioning error with an amplitude of $|\Delta x_{max}|$, $|\Delta y_{max}| = 0.1$ mm is introduced, the synthesis error $\epsilon_f$ increases drastically, especially for frequencies 500 Hz $\leq f \leq$ 5000 Hz and number of microphones N $\geq$ 12 (cf. middle row in Figure 2.2). These synthesis errors clearly show that robustness needs to be improved more when a large number of microphones are used. The deterioration of accuracy due to positioning error is primarily limited to lower frequencies, since a bias in the phase response (which is the main effect of positioning errors) is more critical at lower frequencies.

The resulting synthesis errors using the Golomb method with unbiased and corrupted steering vectors are shown in the right column of Figure 2.2. The synthesis error obtained with the unbiased steering vector is approximately equal to that obtained with the Random Sampling method (top row in Figure 2.2). Thus, the Golomb method only achieves a minor advantage (lower computational cost) compared to the Random Sampling method when using unbiased steering vectors.

However, the synthesis errors of both positioning methods become much larger when the steering vector is assumed to have a random positioning error with $|\Delta x_{max}|, |\Delta y_{max}| = 0.1$ mm (middle row in Figure 2.2). In this case, the synthesized directivity pattern using the Golomb method seems to be more robust compared to the directivity pattern using the Random Sampling method. More specifically, the error $\epsilon_f$ is smaller with the Golomb method for frequencies 500 Hz $\leq f \leq$ 5000 Hz and number of microphones N $\geq$ 12. These regions appear to be particularly affected by positioning errors (for both methods). Moreover, the Golomb method enables a wider frequency range with smaller synthesis errors.

However, the Golomb method seems to be slightly disadvantageous to the Random Sampling method for frequencies $f \geq$ 5 kHz when only a small number of microphones (N $\leq$ 12) are used. This points to the fact that the Golomb topology with N $\leq$ 12 has an unfavorable spatial distribution. In fact, the procedure outlined in section 2.4 has only been used for N $=$ 24. On that score, the error using the Golomb method probably would have been smaller if the spatial distribution of the topology would have been monitored for each single number of microphones successively.

### 2.7.2 Effect of regularization

For both positioning methods an improvement can be achieved by applying regularization as in equation 2.1 and 2.2 (see bottom row in Figure 2.2). It

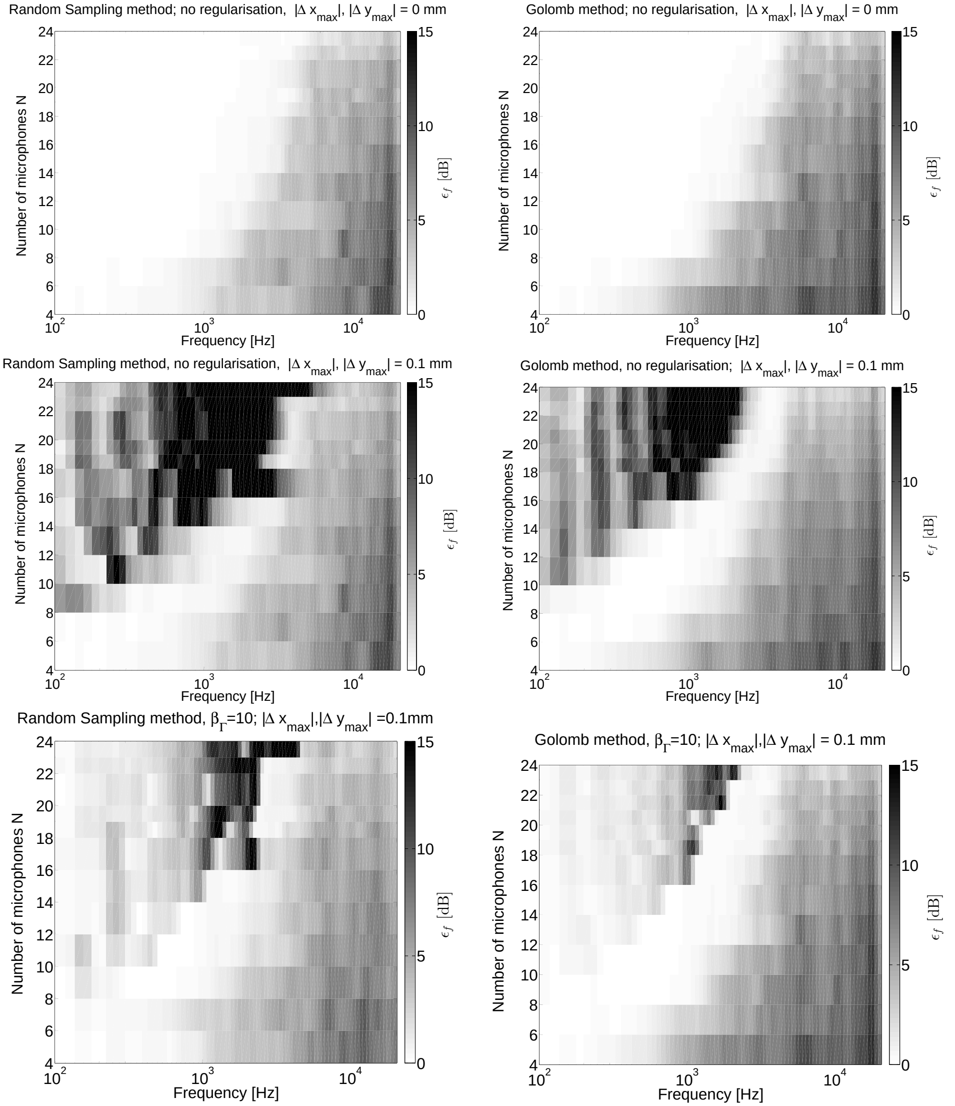


**Figure 2.2:** Simulated synthesis error $\epsilon_f$ of the virtual artificial head, as a function of frequency and the number of microphones in the array. The left column shows results for the Random Sampling method, the right column for the Golomb method. **Top row:** no artificial positioning error, no regularization, **middle row:** random positioning error $|\Delta\, x_{max}|, |\Delta\, y_{max}| = 0.1$ mm, no regularization, **bottom row:** random positioning error $|\Delta\, x_{max}|, |\Delta\, y_{max}| = 0.1$ mm, regularization with $\beta_\Gamma = 10$.

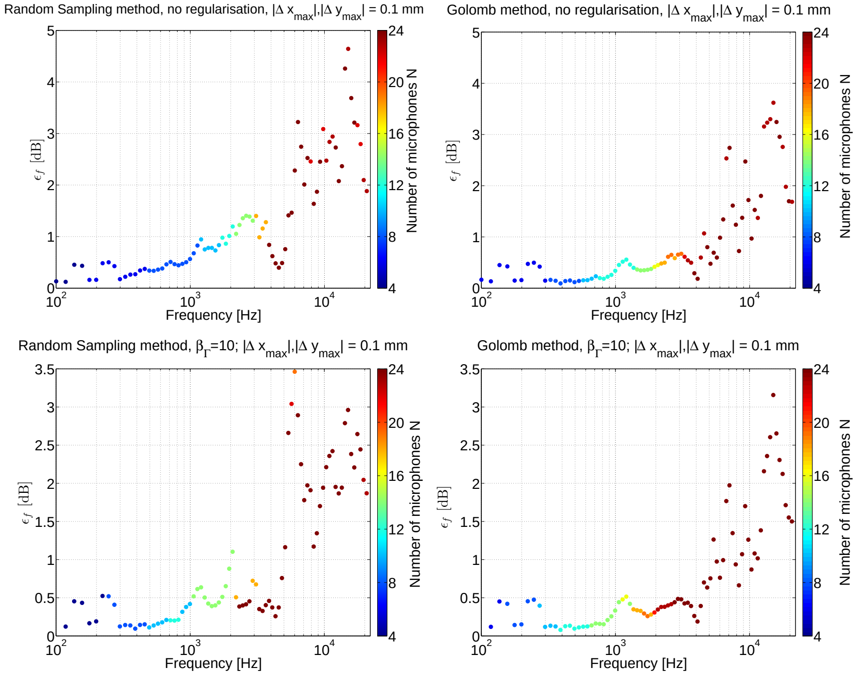


**Figure 2.3:** Simulated synthesis error $\epsilon_f$ of the virtual artificial head, for optimal numbers of microphones per critical band. The left column shows results for the Random Sampling method, the right column for the Golomb method. **Top row:** random positioning error $|\varDelta\, x_{max}|, |\varDelta\, y_{max}| = 0.1$ mm, no regularization , **bottom row:** random positioning error $|\varDelta\, x_{max}|, |\varDelta\, y_{max}| = 0.1$ mm, regularization with $\beta_\varGamma = 10$. Note that the ordinate scale is smaller than in the top row. The number of microphones leading to the minimal error (per critical band) is indicated by the color.

can be observed that regularization reduces the synthesis error, especially for a large number of microphones N $\geq$ 12. For a positioning error of $|\varDelta\, x_{max}|, |\varDelta\, y_{max}| = 0.1$mm, the regularization parameter $\beta_\varGamma = 10$ seems to be appropriate since this regularization improves robustness and does not deteriorate the synthesis accuracy for higher frequencies considerably. For larger positioning errors, robustness may be improved by using smaller $\beta_\varGamma$. Therefore, $\beta_\varGamma$ should be adjusted to the expected errors, in order to optimize the trade-off between robustness and accuracy.

### 2.7.3 Best-case scenario: optimal number of microphones per critical band

As could be seen above, the optimal number of microphones depends on the frequency band: it is lower at low frequencies and higher at high frequencies. This feature can be exploited in a VAH, by choosing the optimal number of microphones for each frequency band. This best-case scenario is illustrated in Figure 2.3.

It should also be noted that only the topologies derived with the Golomb method enable to achieve the depicted minimal synthesis error with a fixed setup (by setting corresponding filter coefficients to zero), because the small-number topologies are a subset of the large-number topologies.

When a positioning error of $|\varDelta x_{max}|, |\varDelta y_{max}| = 0.1$mm is applied, the best-case error is smaller using the Golomb method for the entire frequency range. Especially at frequencies $f \geq$ 1000 Hz, the Golomb-method seems to yield smaller synthesis errors. In general, the difference between the methods is smaller when regularization is applied. As could be expected, the best-case error becomes smaller when regularization with $\beta_\varGamma = 10$ is applied.

## 2.8 Conclusions

A method for optimizing the positioning of microphones for a virtual artificial head, based on a two-dimensional extension of a Golomb-ruler, has been presented. This method is computationally less expensive compared with brute-force methods. Topologies derived using the Golomb method yield a smaller synthesis error compared to the Random Sampling method. A further advantage of the Golomb method for the depicted application within a VAH is the variable number of microphones for a fixed setup.

The robustness of the synthesis can be considerably improved by applying a norm constraint when positioning errors occur. The crucial regularization parameter needs to be optimized for the expected positioning error. However, the regularization in equation 2.1 and 2.2 turned out to be suboptimal since it tends to scale down the magnitude of the optimized filter coefficients for larger $\mu$ and hence distort the output of the VAH. For this reason, more appropriate regularization strategies will be proposed and evaluated in the following chapter (cf. **Chapter 3**).

# 3

# Regularization approaches for synthesizing head-related transfer functions with microphone arrays*

---

*This chapter is based on the publication

- Rasumow, Hansen, van de Par, Püschel, Mellert, Doclo, and Blau (2014d), "Regularization approaches for synthesizing head-related transfer functions with microphone arrays," *Journal paper in preparation for submission to the IEEE Transactions on Audio, Speech and Language Processing.*

*As an alternative to traditional artificial heads, it is possible to synthesize individual head-related transfer functions (HRTFs) using a so-called virtual artificial head (VAH), consisting of a microphone array with an appropriate topology and individually optimized filter coefficients. The resulting spatial directivity pattern of a VAH is known to be sensitive to small deviations of the assumed microphone characteristics, e.g. gain, phase and/or the positions of the microphones. In many beamformer designs, this sensitivity is reduced by imposing a white noise gain (WNG) constraint on the filter coefficients for a single desired look direction. In this study, this constraint is shown to be inappropriate for regularizing the HRTF-synthesis with multiple desired directions and three different regularization approaches are proposed. In the first approach, the measured deviations of the microphone characteristics are taken into account into the filter design. In the second approach, the filter coefficients are regularized using the mean WNG for all directions. The third approach additionally takes into account the information of neighbouring frequency bins into both the optimization and the regularization. The different regularization approaches are compared using measured and analytic transfer functions of the microphones including random deviations. Experimental results show that the optimization and regularization with the third approach using frequency bands mimicking the spectral resolution of the human auditory system yields the best results among the considered regularization approaches.*

## 3.1 Introduction

Spatial information is essential in the appraisal of sound. One important reproduction method which aims at preserving the spatial information is the so-called binaural reproduction, where the goal is to reproduce the sounds at listeners' ears via headphones in the same way as if the listener had been in the real sound field. The recordings needed for such a binaural reproduction are traditionally made using so-called artificial heads. Unfortunately, because of their non-individual character, these recordings often entail perceptual deficiencies, such as front-back confusion and internalization (cf. Wenzel *et al.* (1993),Hartmann and Wittenberg (1996) and Hill *et al.* (2000)).

Concurrently, microphone arrays have long been used for spatial filtering using filter-and-sum beamforming. Even though most popular beamformers aim at steering into one look direction (e.g. superdirective beamformers, cf. Cox *et al.* (1986),Kates (1993) and Bitzer and Simmer (2001)), they can, in principle, be used to synthesize an arbitrary desired directivity pattern, e.g. using different cost functions (cf. Nordebo *et al.* (1994), Kajala and Hamaldinen (1999) and Doclo and Moonen (2003a)). They can therefore also be used to synthesize individual head-related transfer functions (HRTFs), thus mimicking the directivity patterns of artificial or real human heads (cf. Chen *et al.* (1992), Kahana *et al.* (1999), Tohtuyeva and Mellert (1999), Sakamoto *et al.* (2008), Atkins (2011a), Atkins (2011b), Rasumow *et al.* (2011b), Rasumow *et al.* (2013a) and Rasumow *et al.* (2014b)). This approach is referred to as virtual artificial head (VAH). Its main advantages are the possibility of post-hoc adjusting the filter coefficients to HRTFs of different listeners (individualization), the possibility of employing head tracking in the reproduction stage and a better flexibility and manageability due to the smaller size/weight of the device.

However, the synthesis of spatial directivity patterns with many microphones and rather small inter-microphone distances is also known to be sensitive to small deviations of the assumed microphone characteristics (e.g. gain, phase, positions, temperature changes and/or drifting microphone characteristics, cf. Cox *et al.* (1986), Doclo and Moonen (2003b), Doclo and Moonen (2007), Rasumow *et al.* (2011b) or **Chapter 2**, Levin *et al.* (2013)). To improve robustness, regularization is usually employed.

In this study, three regularization approaches are proposed: In the first approach, the filter coefficients are regularized by taking into account measured deviations of the microphone characteristics (i.e. steering vectors) within a least squares optimization procedure (cf. section 3.3). This is comparable to taking into account the probability density function of the microphone characteristics (cf. Doclo and Moonen (2003b), Doclo and Moonen (2007)), but using measured data. If such measured deviations of the steering vectors are not available, it is common practice to impose a so-called white noise gain (WNG) constraint on the filter coefficients (cf. Cox *et al.* (1986), Bitzer and Simmer (2001)), however typically only for one desired look direction. In this study, it is shown in section 3.5.2 that this constraint is inappropriate for synthesizing HRTFs with multiple desired directions. In section 3.4, different WNG constraints are hence proposed: In the second approach, all directions are included into the weighting of the WNG (cf. section 3.4.2). In the third approach, the information of neighbouring frequency bands is additionally considered in the optimization and regularization procedure (cf. section 3.4.3). In section 3.5, these different approaches to increase the robustness of the VAH-synthesis are evaluated with measured and analytic steering vectors, with and without adding small random deviations.

## 3.2 Least squares beamformer design

In general, for synthesizing a desired directivity pattern using a microphone array, the filter coefficients of a filter-and-sum beamformer (cf. Figure 3.1) can be computed by minimizing a cost function (e.g. least squares, total least squares or non-linear cost functions, cf. Kajala and Hamaldinen (1999), Doclo and Moonen (2003a)), either for all frequencies jointly (broadband design) or for each frequency bin independently (narrowband design). In this study, a narrowband design procedure is used because of its better numerical stability and since previous studies have already shown its adequacy to synthesize HRTFs using a VAH (cf. Rasumow *et al.* (2013a), Rasumow *et al.* (2014b) or **Chapter 6**).

The synthesized spatial directivity pattern $\mathrm{H}(f,\theta)$ of the VAH can be expressed as a vector product of $\mathbf{w}(f)$ and $\mathbf{d}(f,\theta)$ (cf. equation 1.5), with $\mathbf{d}(f,\theta)$ the N-dimensional steering vector, describing the acoustic transfer functions between a sound source from angle[1] $\theta$ to the N microphones at frequency $f$, and $\mathbf{w}(f)$ an N-dimensional vector containing the complex-valued filter coefficients. In order to synthesize a desired directivity pattern $\mathrm{D}(f,\theta)$, e.g. an individual frequency- and angle-dependent HRTF (cf. Figure 3.1), the filter coefficients $\mathbf{w}(f)$ can be computed by minimizing the weighted narrowband least squares cost function $\mathrm{J}_{\mathrm{LS}}$ given in equation 1.8.

---

[1] In this study only the azimuthal angle $\theta$ is considered, but the proposed procedure can also be straightforwardly extended to a three-dimensional design including elevation angles.

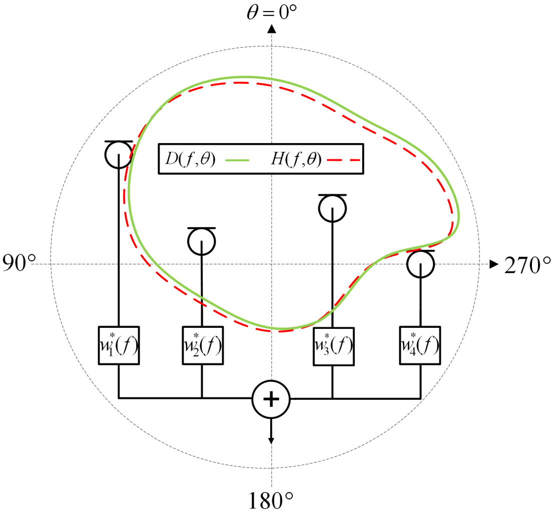


**Figure 3.1:** Schematic diagram of a filter-and-sum beamformer with $\mathrm{N} = 4$ microphones and the associated filter weights $\mathbf{w}(f) = [w_1(f) \ldots w_4(f)]^T$. Desired directivity pattern $\mathrm{D}(f,\theta)$ and exemplary resulting directivity pattern $\mathrm{H}(f,\theta)$.

The least squares cost function $J_{LS}$ is a weighted sum over P discrete directions of the squared absolute difference between the synthesized directivity pattern $H(f,\theta)$ and the desired directivity pattern $D(f,\theta)$. The weighting function $F(f,\theta)$ enables to assign more or less importance to certain directions. The least squares cost function in equation 1.8 will be used in the remainder of this study because it has a closed-form solution and since previous studies have already shown its suitability to synthesize HRTFs (cf. Rasumow *et al.* (2013a), Rasumow *et al.* (2014b) or **Chapter 6**).

The filter coefficients minimizing this cost function can be obtained by setting the gradient $\nabla_{\mathbf{w}(f)} J_{LS}(\mathbf{w}(f))$ to zero, leading to the solution in equation 1.11 with the variables given in 1.10.

When using many microphones with rather small inter-microphone distances, the synthesized directivity pattern $H(f,\theta)$ is known to be highly sensitive to small deviations of the assumed steering vectors $\mathbf{d}(f,\theta)$. In the next sections, different regularization approaches will be discussed, either by using multiple measured steering vectors or by imposing an appropriate constraint.

## 3.3 Regularization by joint optimization for multiple steering vectors

When multiple sets of steering vectors are available, e.g. measured under different conditions (e.g. temperature), the robustness can be increased by jointly optimizing the least squares cost function in equation 1.8 over the sets of steering vectors, i.e.

$$\boxed{J_K(\mathbf{w}(f)) = \sum_{k=1}^{K} \sum_{i=1}^{P} F_k(f,\theta_i) \left| \mathbf{w}^H(f)\mathbf{d}_k(f,\theta_i) - D(f,\theta_i) \right|^2} \tag{3.1}$$

with K the total number of measured sets of steering vectors $\mathbf{d}_k$ and $F_k(f,\theta)$ the weighting function for the $k$-th set of steering vectors. Analogous to equation 1.11, the filter coefficients minimizing equation 3.1 are equal to

$$\mathbf{w}_K(f) = \mathbf{Q}_K^{-1}(f) \cdot \mathbf{a}_K(f), \tag{3.2}$$

with

$$\mathbf{Q}_{\mathrm{K}}(f) = \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{P}} F_k(f,\theta_i) \cdot \mathbf{d}_k(f,\theta_i)\mathbf{d}_k^H(f,\theta_i)$$

$$\mathbf{a}_{\mathrm{K}}(f) = \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{\mathrm{P}} F_k(f,\theta_i) \cdot \mathbf{d}_k(f,\theta_i)\mathrm{D}^*(f,\theta_i).$$

This joint optimization can be interpreted as a regularization using measured data and is comparable to incorporating assumed deviations of the microphone characteristics, e.g. in terms of probability density functions (cf. Doclo and Moonen (2003b), Doclo and Moonen (2007)). In contrast to assuming probability density functions, here the measured deviations of the steering vectors may be considered as an empirical estimate for the expected deviations of the microphone characteristics.

## 3.4 Regularization using a WNG constraint

When only one set of measured steering vectors is available (i.e. $\mathrm{K} = 1$), it is common practice to increase the robustness of the beamformer by imposing a constraint on the filter coefficients. Due to its physical interest, the output power of the beamformer for spatially uncorrelated noise in comparison to the output power for the desired acoustic field, defined as the white noise gain (WNG), is a common measure for quantifying its robustness (cf. Cox *et al.* (1986), Bitzer and Simmer (2001)). However, the most appropriate formulation of the desired acoustic field may vary with the application and hence may have a significant influence on the resulting performance. The advantages and disadvantages of various WNG variants with regard to their application for a VAH are discussed in the following sections.

### 3.4.1 White noise gain

The frequency- and angle-dependent white noise gain is defined as (see e.g. Bitzer and Simmer (2001))

$$\mathrm{WNG}(\mathbf{w}(f),\theta) = \frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}. \tag{3.3}$$

For many beamformers, e.g. superdirective beamformers (cf. Cox *et al.* (1986), Kates (1993), Bitzer and Simmer (2001)), the WNG is considered only for the look direction $\theta_d$. In this case, $\mathrm{WNG}(\mathbf{w}(f), \theta_d)$ relates the output power for the look direction $\theta_d$ to the output power for spatially uncorrelated noise. In general, a larger WNG is associated with a larger attenuation of spatially uncorrelated noise in comparison to the output power for the look direction $\theta_d$ and hence with an increased robustness.

When imposing an inequality constraint on the WNG for direction $\theta_d$, the constrained optimization problem can be written as

$$\boxed{\min_{\mathbf{w}(f)} \mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f)) \quad \text{subject to} \quad \mathrm{WNG}(\mathbf{w}(f), \theta_d) \geq \beta} \tag{3.4}$$

with $\beta$ the minimum desired value for $\mathrm{WNG}(\mathbf{w}(f), \theta_d)$. The Lagrange function associated with this constrained optimization problem is equal to[2]

$$\begin{aligned} \mathrm{J}_{\mathrm{sd}}(\mathbf{w}(f), \mu) &= \mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f)) \\ &\quad + \mu \left( \mathbf{w}^H(f)\mathbf{w}(f) - \frac{1}{\beta} \left| \mathbf{w}^H(f)\mathbf{d}(f, \theta_d) \right|^2 \right), \end{aligned} \tag{3.5}$$

with $\mu$ the Lagrange multiplier. Analogous to equation 1.11 and 3.2, the filter coefficients minimizing the cost function in equation 3.5 can be obtained by setting the gradient to zero, leading to

$$\mathbf{w}_{\mathrm{sd}}(f, \mu) = \left( \mathbf{Q}(f) + \mu \Big( \mathbf{I}_{\mathrm{N}} - \frac{1}{\beta}\, \mathbf{d}(f, \theta_d)\mathbf{d}^H(f, \theta_d) \Big) \right)^{-1} \mathbf{a}(f), \tag{3.6}$$

with $\mathbf{Q}(f)$ and $\mathbf{a}(f)$ given in equation 1.10 and $\mathbf{I}_{\mathrm{N}}$ the $\mathrm{N} \times \mathrm{N}$-dimensional unity matrix. The Lagrange multiplier $\mu$ in equation 3.6 needs to be determined, e.g. using an iterative procedure, such that the inequality constraint $\mathrm{WNG}(\mathbf{w}_{\mathrm{sd}}(f, \mu), \theta_d) \geq \beta$ is satisfied.

Note that if the beamformer response in the look direction $\theta_d$ is $\mathrm{H}(f, \theta_d) = \mathbf{w}^H(f)\mathbf{d}(f, \theta_d) = 1$, which is typically the case for superdirective beamforming (cf. Cox *et al.* (1986), Kates (1993), Bitzer and Simmer (2001)), the WNG is equal to the inverse of the square norm of the filter coefficients, i.e.

$$\mathrm{WNG}(\mathbf{w}(f), \theta_d) = \frac{1}{\mathbf{w}^H(f)\mathbf{w}(f)}.$$

In this case, the term $\frac{1}{\beta} \left| \mathbf{w}^H(f)\mathbf{d}(f, \theta_d) \right|^2$ will vanish in the gradient of the

---

[2] A paradigmatic derivation of the constrained cost function and the filter coefficients minimizing this cost function is given in Appendix E

cost function given in equation 3.5, and thus the term $\frac{1}{\beta}\,\mathbf{d}(f,\theta_d)\mathbf{d}^H(f,\theta_d)$ will vanish in equation 3.6.

The experimental results in section 3.5 will show that imposing a WNG constraint for one single look direction $\theta_d$ is inappropriate for synthesizing HRTFs, where the objective is to synthesize a desired directivity pattern for *all* considered directions. Hence, a direction-dependent weighting for $\mathrm{WNG}(\mathbf{w}(f),\theta)$ will be presented in the following section.

### 3.4.2 Mean white noise gain over all directions

When synthesizing HRTFs using a VAH, the accuracy and the robustness of the synthesized directivity pattern needs to be assured for all considered directions. Hence, instead of constraining the WNG in equation 3.3 for only one direction, we propose to constrain the mean white noise gain over all directions, defined as

$$\mathrm{WNG_m}(\mathbf{w}(f)) = \sum_{i=1}^{\mathrm{P}} g(f,\theta_i)\ \mathrm{WNG}(\mathbf{w}(f),\theta_i), \tag{3.7}$$

where the weighting function $g(f,\theta)$ enables to assign more or less importance to certain directions (cf. Rasumow *et al.* (2014b) or **Chapter 6**). The $\mathrm{WNG_m}$ in equation 3.7 can be rewritten as

$$\mathrm{WNG_m}(\mathbf{w}(f)) = \frac{\mathbf{w}^H(f)\mathbf{Q}_\mathrm{m}(f)\ \mathbf{w}(f)}{\mathbf{w}^H(f)\mathbf{w}(f)}, \tag{3.8}$$

with

$$\mathbf{Q}_\mathrm{m}(f) = \sum_{i=1}^{\mathrm{P}} g(f,\theta_i)\cdot\mathbf{d}(f,\theta_i)\mathbf{d}^H(f,\theta_i).$$

In the remainder of this study, we will use a constant weighting for all directions, i.e. $g(f,\theta_i) = \frac{1}{\mathrm{P}},\ i = 1\ldots\mathrm{P}$. In this case, $\mathrm{WNG_m}(\mathbf{w}(f))$ relates the mean output power from all P considered directions to the output power for spatially uncorrelated noise. Note that this regularization approach is evaluated perceptually in Rasumow *et al.* (2014b) or **Chapter 6**, whereas in this study it is objectively compared with other regularization approaches (cf. section 3.5.5).

When imposing a constraint on the mean white noise gain, the constrained optimization problem can be written as

$$\boxed{\min_{\mathbf{w}(f)} \mathrm{J_{LS}}(\mathbf{w}(f)) \quad \text{subject to} \quad \mathrm{WNG_m}(\mathbf{w}(f)) \geq \beta_\mathrm{m}} \tag{3.9}$$

with $\beta_{\mathrm{m}}$ the minimum desired value for $\mathrm{WNG}_{\mathrm{m}}$. The Lagrange function associated with this constrained optimization problem is equal to

$$\begin{aligned}\mathrm{J}_{\mathrm{m}}(\mathbf{w}(f),\mu) &= \mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f)) \\ &+ \mu\left(\mathbf{w}^{H}(f)\mathbf{w}(f) - \frac{1}{\beta_{\mathrm{m}}}\,\mathbf{w}^{H}(f)\mathbf{Q}_{\mathrm{m}}(f)\mathbf{w}(f)\right). \qquad (3.10)\end{aligned}$$

Analogous to equation 3.6, the filter coefficients minimizing the cost function in equation 3.10 are equal to

$$\mathbf{w}_{\mathrm{m}}(f,\mu) = \left(\mathbf{Q}(f) + \mu\left(\mathbf{I}_{\mathrm{N}} - \frac{1}{\beta_{\mathrm{m}}}\,\mathbf{Q}_{\mathbf{m}}(f)\right)\right)^{-1}\mathbf{a}(f). \qquad (3.11)$$

Note that the only difference between the solutions in equation 3.6 and 3.11 is the exchange of the rank-1 matrix $\mathbf{d}(f,\theta_d)\mathbf{d}^{H}(f,\theta_d)$ in equation 3.6 by the rank-P (assuming independent steering vectors from P directions) matrix $\mathbf{Q}_{\mathrm{m}}(f)$ in equation 3.11.

### 3.4.3 Optimization and white noise gain for multiple frequencies

It is a well-known phenomenon that the human auditory system groups incoming sounds into so-called critical frequency bands that broaden with increasing center frequencies (cf. Fletcher (1940)). Hence, defining a cost function and a WNG constraint that incorporate the grouping of frequencies within a perceptually-relevant bandwidth seems plausible. Consequently, the filter optimization at each frequency $f$ may be formulated within equivalent rectangular bandwidths (ERB), corresponding to human auditory filters (cf. Moore and Glasberg (1983)), with $f$ as its center frequency.

Let us consider the $t$-th frequency band centered around $f_c^t$, incorporating $\mathrm{L}^t$ frequency bins. We now define the frequency vector $\mathbf{\Omega}^t = f_1^t \ldots f_c^t \ldots f_{\mathrm{L}^t}^t$, where $f_1^t$ and $f_{\mathrm{L}^t}^t$ denote the first and the last frequency bin, and the $\mathrm{N}\cdot\mathrm{L}^t$-dimensional stacked filter vector $\mathbf{w}_{\mathrm{v}}(\mathbf{\Omega}^t)$ for the $t$-th frequency band

$$\mathbf{w}_{\mathrm{v}}(\mathbf{\Omega}^t) = \begin{bmatrix}\mathbf{w}(f_1^t)\\ \vdots \\ \mathbf{w}(f_c^t)\\ \vdots \\ \mathbf{w}(f_{\mathrm{L}^t}^t)\end{bmatrix}. \qquad (3.12)$$

The cost function for the $t$-th frequency band can be defined as the sum of the least squares cost functions in the $\mathrm{L}^t$ frequency bins, i.e.

$$\begin{aligned}\mathrm{J}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) &= \sum_{l=1}^{\mathrm{L}^t} \mathrm{J}_{\mathrm{LS}}(\mathbf{w}(f_l^t)) \qquad (3.13)\\ &= \sum_{l=1}^{\mathrm{L}^t}\sum_{i=1}^{\mathrm{P}} F(f_l^t,\theta_i)\cdot\left|\mathbf{w}^H(f_l^t)\mathbf{d}(f_l^t,\theta_i)-\mathrm{D}(f_l^t,\theta_i)\right|^2,\end{aligned}$$

which can be rewritten using the stacked filter vector in equation 3.12 as

$$\begin{aligned}\mathrm{J}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) = {}&\mathbf{w}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{Q}_{\mathrm{v}}(\boldsymbol{\Omega}^t)\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)-\mathbf{w}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{a}_{\mathrm{v}}(\boldsymbol{\Omega}^t)\\ &-\mathbf{a}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)+d_{\mathrm{v}}(\boldsymbol{\Omega}^t), \qquad (3.14)\end{aligned}$$

with

$$\mathbf{Q}_{\mathrm{v}}(\boldsymbol{\Omega}^t) = \begin{bmatrix}\mathbf{Q}(f_1^t) & & & & \\ & \ddots & & & \\ & & \mathbf{Q}(f_c^t) & & \\ & & & \ddots & \\ & & & & \mathbf{Q}(f_{\mathrm{L}^t}^t)\end{bmatrix},$$

$$\mathbf{a}_{\mathrm{v}}(\boldsymbol{\Omega}^t) = \begin{bmatrix}\mathbf{a}(f_1^t)\\ \vdots \\ \mathbf{a}(f_c^t)\\ \vdots \\ \mathbf{a}(f_{\mathrm{L}^t}^t)\end{bmatrix},$$

$$d_{\mathrm{v}}(\boldsymbol{\Omega}^t) = \sum_{l=1}^{\mathrm{L}^t}\sum_{i=1}^{\mathrm{P}} F(f_l^t,\theta_i)\cdot\left|\mathrm{D}(f_l^t,\theta_i)\right|^2. \qquad (3.15)$$

The WNG for the $t$-th frequency band can be defined similarly to equation 3.8 as

$$\mathrm{WNG}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) = \frac{\sum_{l=1}^{\mathrm{L}^t}\mathbf{w}^H(f_l^t)\mathbf{Q}_{\mathrm{m}}(f_l^t)\mathbf{w}(f_l^t)}{\sum_{l=1}^{\mathrm{L}^t}\mathbf{w}^H(f_l^t)\mathbf{w}(f_l^t)}, \qquad (3.16)$$

which relates the mean output power from all P directions summed over the $\mathrm{L}^t$ considered frequency bins in $\boldsymbol{\Omega}^t$ to the output power for spatially uncorrelated noise over the $\mathrm{L}^t$ considered frequency bins. The WNG in equation 3.16 can be rewritten using the stacked filter vector in equation 3.12 as

$$\mathrm{WNG}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) = \frac{\mathbf{w}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{Q}_{\mathrm{vm}}(\boldsymbol{\Omega}^t)\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)}{\mathbf{w}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)}, \tag{3.17}$$

with

$$\mathbf{Q}_{\mathrm{vm}}(\boldsymbol{\Omega}^t) = \begin{bmatrix} \mathbf{Q}_{\mathrm{m}}(f_1^t) & & & & \\ & \ddots & & & \\ & & \mathbf{Q}_{\mathrm{m}}(f_c^t) & & \\ & & & \ddots & \\ & & & & \mathbf{Q}_{\mathrm{m}}(f_{\mathrm{L}^t}^t) \end{bmatrix}. \tag{3.18}$$

When imposing a constraint on the WNG for the $t$-th frequency band, the constrained optimization problem for the $t$-th frequency band can be written as

$$\boxed{\min_{\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)} \mathrm{J}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) \quad \text{subject to} \quad \mathrm{WNG}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) \geq \beta_{\mathrm{v}}} \tag{3.19}$$

with $\beta_{\mathrm{v}}$ the minimum desired value for $\mathrm{WNG}_{\mathrm{v}}$. The Lagrange function associated with this constrained optimization problem is equal to

$$\begin{aligned} &\mathrm{J}_{\mathrm{vm}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t), \mu) = \mathrm{J}_{\mathrm{v}}(\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)) \\ &+ \mu\Big(\mathbf{w}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t) - \frac{1}{\beta_{\mathrm{v}}}\mathbf{w}_{\mathrm{v}}^H(\boldsymbol{\Omega}^t)\mathbf{Q}_{\mathrm{vm}}(\boldsymbol{\Omega}^t)\mathbf{w}_{\mathrm{v}}(\boldsymbol{\Omega}^t)\Big). \end{aligned} \tag{3.20}$$

Analogous to equation 3.11, the filter coefficients minimizing $\mathrm{J}_{\mathrm{vm}}$ are equal to

$$\mathbf{w}_{\mathrm{vm}}(\boldsymbol{\Omega}^t, \mu) = \left(\mathbf{Q}_{\mathrm{v}}(\boldsymbol{\Omega}^t) + \mu\left(\mathbf{I}_{\mathrm{v}} - \frac{1}{\beta_{\mathrm{v}}}\mathbf{Q}_{\mathrm{vm}}(\boldsymbol{\Omega}^t)\right)\right)^{-1} \mathbf{a}_{\mathrm{v}}(\boldsymbol{\Omega}^t), \tag{3.21}$$

with $\mathbf{I}_{\mathrm{v}}$ the $\mathrm{N} \cdot \mathrm{L}^t \times \mathrm{N} \cdot \mathrm{L}^t$-dimensional unity matrix. It is important that we solve the optimization problem in equation 3.19 for all frequency bins (and not only for the center frequencies of the critical auditory bands, cf. Fletcher (1940)), i.e. for each frequency we consider the ERB with frequencies $\boldsymbol{\Omega}^t$ around that frequency as its center frequency $f_c^t$. The solution in equation 3.21 then yields an $\mathrm{N} \cdot \mathrm{L}^t$-dimensional vector with filter coefficients $\mathbf{w}(f_1^t) \ldots \mathbf{w}(f_{\mathrm{L}^t}^t)$, where we only consider the filter coefficients $\mathbf{w}(f_c^t)$ at the center frequency as the solution for that frequency bin. This procedure can hence be interpreted as taking neighbouring frequencies into account both for the cost function in equation 3.13 as well as for the regularization in equation 3.16. It is worth noting that when a Lagrange multiplier $\mu = 0$ is used, the optimization for a single frequency band according to equation 3.11 and for multiple frequency bands according to equation 3.21 yield very similar numerical results.

In contrast to the WNG for a single frequency, for the WNG in equation 3.16 the weighting function $g(f, \theta)$ in equation 3.8 can be chosen to assign more or less importance to certain directions *and* frequencies[3]. By setting the weighting function equal to $g(f, \theta_i) = \frac{1}{\mathrm{P}}$, $i = 1 \ldots \mathrm{P}$, the mean WNG over all considered directions and frequencies is used. Another possibility would be to use a weighting function that depends on the angle- and frequency-dependent desired directivity pattern $\mathrm{D}(f, \theta)$. A psychoacoustically reasonable weighting would be, for instance, to assign more importance to the louder parts of the desired directivity pattern than to spatial and/or spectral notches, i.e. a weighting

$$g_w^t(f, \theta) = \frac{1}{\mathrm{P}} \left( g_n^t(f, \theta) - \overline{g_n^t} + 1 \right), \tag{3.22}$$

with

$$\begin{aligned}
\overline{g_n^t} &= \frac{1}{\mathrm{L}^t \cdot \mathrm{P}} \sum_{l=1}^{\mathrm{L}^t} \sum_{i=1}^{\mathrm{P}} g_n(f_l^t, \theta_i), \\
g_n^t(f, \theta) &= g_{\mathrm{pos}}^t(f, \theta) / \max_{\boldsymbol{\Omega}^t, \boldsymbol{\Theta}}(g_{\mathrm{pos}}^t(f, \theta)), \\
g_{\mathrm{pos}}^t(f, \theta) &= g_{\mathrm{dB}}(f, \theta) - \min_{\boldsymbol{\Omega}^t, \boldsymbol{\Theta}}(g_{\mathrm{dB}}(f, \theta)), \\
g_{\mathrm{dB}}(f, \theta) &= 20 \cdot \log_{10}\left(|\mathrm{D}(f, \theta)|\right).
\end{aligned}$$

Here, $g_n^t(f, \theta)$ denotes a positive and normalized weighting function according to the dB-values of the desired directivity pattern $g_{\mathrm{dB}}(f, \theta)$ and $\overline{g_n^t}$ is the mean value over all frequencies in the $t$-th frequency band $\boldsymbol{\Omega}^t$ and all directions $\boldsymbol{\Theta}$. The function $g_n^t(f, \theta)$ is further normalized by the number of directions P and subtracted to have a mean value of 1 for frequencies $\boldsymbol{\Omega}^t$ to enable a fair comparison with the other discussed regularization approaches.

---

[3] We chose to use a weighting of the regularization term only. Another approach would be to also use a similar weighting function $F(f, \theta)$ for the cost function $\mathrm{J}_{\mathrm{vm}}$ in equation 3.13. However, this will not be considered in this study.

## 3.5 Experimental results

In this section, the proposed optimization and regularization approaches from sections 3.3 and 3.4 will be compared with respect to their accuracy and robustness synthesizing a measured HRTF in the horizontal plane (left ear, subject $\mathrm{S}_1$, for details we refer to Rasumow *et al.* (2014c)). The HRTFs (i.e. the desired directivity pattern $\mathrm{D}(f,\theta)$) were measured using the blocked ear method (cf. Hammershøi and Møller (1996)) where microphones (Knowles FG-23329 miniature electret microphones) were flush mounted in individual earmolds that blocked the ear entrance. The steering vectors $\mathbf{d}(f,\theta)$ were measured with a planar microphone array (cf. $\text{array}_2$ from Figure 6.2) with a Golomb-based topology, which is a mathematically motivated method for deriving microphone topologies (cf. Rasumow *et al.* (2011b) or **Chapter 2**). The microphone array contained $\mathrm{N} = 24$ microphones (each consisting of two Analog Devices ADMP 504 Ultralow Noise sensors) on a 20 cm × 20 cm plate covered with additional absorbing material. The steering vectors and the HRTFs were measured in an anechoic room in the horizontal plane with an azimuthal resolution of $\varDelta\theta = 15°$ (i.e. $\mathrm{P} = 24$ directions), since this resolution is assumed to be sufficiently fine as suggested by previous evaluations (cf. Rasumow *et al.* (2013a), Rasumow *et al.* (2014b) or **Chapter 6**) and preliminary listening tests. Both the steering vectors and the HRTFs were measured with white noise stimuli using the $\mathrm{H}_1$ estimate (cf. Mitchell (1982)) with a 8192-point Hann window, 50% overlap and 27 averages and were truncated in the time domain to a length of 256 samples ($\approx$ 5.8 ms at a sampling frequency of 44.1 kHz). Furthermore, the steering vectors were normalized by a constant factor

$$\nu_k = \sqrt{\frac{1}{\mathrm{N}\cdot\mathrm{L}\cdot\mathrm{P}}\sum_{l=1}^{\mathrm{L}}\sum_{i=1}^{\mathrm{P}}\mathbf{d}_k^H(f_l,\theta_i)\mathbf{d}_k(f_l,\theta_i)}\ , \tag{3.23}$$

with $f_1 = 500$ Hz, $f_\mathrm{L} = 1000$ Hz and L the number of intermediate frequency bins. This normalization was performed to achieve unit power on average over all directions, microphones and frequencies 500 Hz $\leq f \leq$ 1000 Hz.

For each of the discussed regularization approaches, the Lagrange multiplier $\mu$ in (3.6), (3.11) and (3.21) was determined numerically. To guarantee the smallest possible $\mu$ that complies with the desired constraint (resulting in the most accurate synthesis), $\mu$ was increased logarithmically until the resulting WNG reached the desired WNG within an accuracy of 0,05 dB. Figure 3.2 exemplarily depicts $\mathrm{WNG}_\mathrm{m}$ for three frequencies, where it can be observed that the WNG does not always increase monotonically with

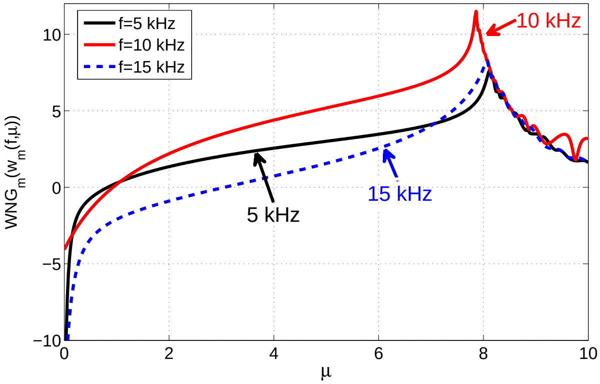


**Figure 3.2:** Mean white noise gain $\mathrm{WNG_m}$ for a left ear synthesis as a function of the Lagrange multiplier $\mu$ for three frequencies.

increasing $\mu$.

In order to quantify the accuracy of the synthesized directivity pattern, we used the synthesis error $\epsilon(f_c, \theta)$ according to equation 1.14. The synthesis error $\epsilon(f_c, \theta)$ is the mean error between the desired directivity pattern $\mathrm{D}(f, \theta)$ and the resulting directivity patterns $\mathrm{H}(f, \theta)$ for each direction $\theta$ averaged over all frequencies in ERB-bands, with $f_c$ denoting the center frequency of the associated ERB-band. Note that equation 1.14 is used in the next sections as the accuracy measure (cf. sections 3.5.2 to 3.5.4) instead of the least squares error from equation 1.8 because it is better suited to represent the perceptually relevant error of the synthesis (cf. Rasumow *et al.* (2014c) or **Chapter 5**). This measure was, however, not used as a cost function because no closed-form solution for minimizing equation 1.14 exists. In order to quantify the robustness of the different approaches, we used the mean white noise gain $\mathrm{WNG_m}$ (cf. section 3.5.1) and the mean error $\bar{\epsilon}$ over directions and frequencies (cf. section 3.5.5).

### 3.5.1 Joint optimization for multiple steering vectors

To demonstrate the effect of a joint optimization, as discussed in section 3.3, multiple sets ($\mathrm{K} = 4$) of steering vectors were measured with the same measuring apparatus (with approximately 24 hours lying between the measurements). In Figure 3.3, the mean and standard deviation of the variability among the sets of measured steering vectors (absolute values) is illustrated.

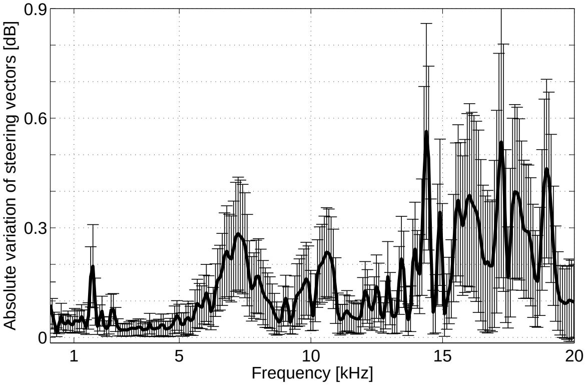


**Figure 3.3:** Mean (black line) and standard deviation (error bar) of the variability (absolute values) among the differently measured steering vectors (please note the linear frequency scaling).

Although the positioning and the environmental influences of the measuring apparatus were kept as constant as possible, the absolute value of the measured transfer functions clearly varied between the measurements. Explanations for these deviations may be, for instance, drifting characteristics of the measuring apparatus and/or slightly differing environmental conditions, e.g. temperature (cf. Yasuno and Ohga (2005)). These measured deviations may be considered as an empirical estimate of the expected deviations of the steering vectors. Hence, the four sets of slightly different steering vectors were integrated into a joint optimization according to equation 3.1 with a uniform weighting $F_k(f, \theta) = 1$.

As a measure for robustness, the associated mean white noise gain $\mathrm{WNG_m}$ for the (unconstrained) joint optimization including $\mathrm{K} = 1 \dots 4$ sets of steering vectors is shown in Figure 3.4. It is apparent that the filter coefficients resulting from an unconstrained optimization with one set of steering vectors ($\mathbf{d}_1(f, \theta)$) yield very low $\mathrm{WNG_m}$ (corresponding to a low robustness), down to $\mathrm{WNG_m} \approx -85$ dB at lower frequencies ($f \leq 4$ kHz). However, the $\mathrm{WNG_m}$ clearly increases with increasing K. Hence, it is possible to regularize the filter coefficients by using multiple sets of measured steering vectors.

The performance of the associated filter coefficients in terms of robustness against randomly disturbed steering vectors will be compared to the other regularization approaches in section 3.5.5.

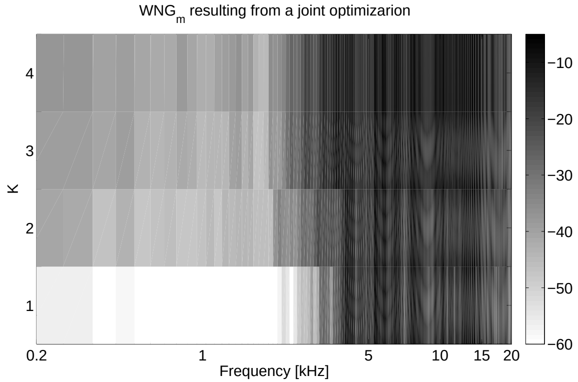


**Figure 3.4:** WNG$_\mathrm{m}$ of filter coefficients $\mathbf{w}_\mathrm{K}$ resulting from the (unconstrained) joint optimization when increasing the number of considered sets of steering vectors K. For the sake of clarity, the range of the depicted WNG$_\mathrm{m}$ is limited between $-60$ dB $\leq$ WNG$_\mathrm{m}$ $\leq -5$ dB.

### 3.5.2 White noise gain constraint

The desired HRTFs and the synthesis associated with a constraint on the white noise gain for a single direction $\mathrm{WNG}(\mathbf{w}(f), 0°)$ are depicted in Figure 3.5 for three directions ($\theta = 0°,\ 90°,\ 225°$). The synthesis associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ results in a large deviation from the desired HRTF for $\theta = \theta_d = 0°$ (cf. 5 kHz $\leq f \leq$ 10 kHz and $f \geq$ 15 kHz) and for $\theta = 225°$ (cf. $f \geq 7$ kHz), while the synthesis for $\theta = 90°$ approximates the desired HRTF quite well. Furthermore, the error $\epsilon$ associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ is depicted in the left panel of Figure 3.6 as a function of frequency and direction. The largest errors $\epsilon$ primarily occur for the frontal direction $\theta = \theta_d = 0°$ and for contralateral directions ($180° \leq \theta \leq 360°$) and at higher frequencies. Interestingly, the largest errors exactly occur for the direction $\theta_d$, which is used for the WNG constraint. This has also been confirmed when changing $\theta_d$ to other directions then 0°. Thus, a constraint on $\mathrm{WNG}(\mathbf{w}(f), \theta_d)$ yields a direction-dependent impact on the synthesis, which is clearly an undesirable effect and results in inappropriate perceptual appraisals (cf. Rasumow *et al.* (2013a)).

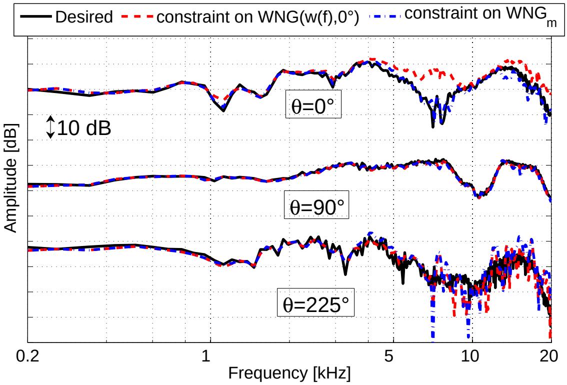


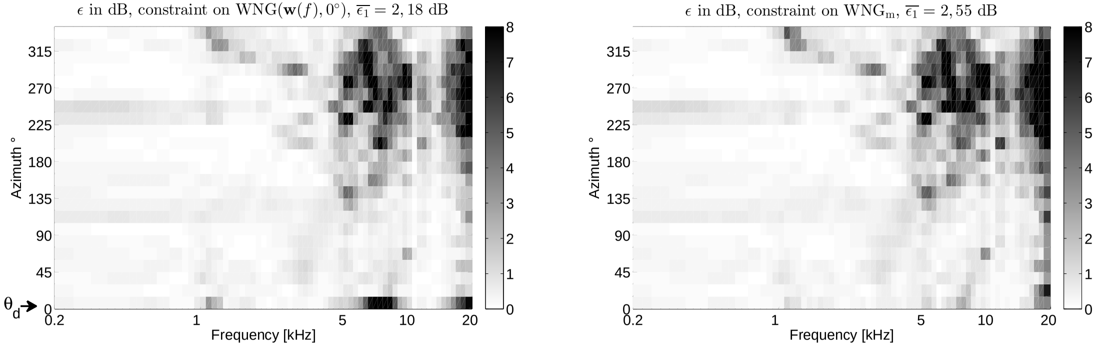
**Figure 3.5:** Desired HRTFs (black solid lines) and synthesis associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ (blue dot-dashed lines, $\beta = -5$ dB) and on $\mathrm{WNG_m}$ (red dashed lines, $\beta_\mathrm{m} = -5$ dB) for three directions ($\theta = 0°,\ 90°,\ 225°$).

### 3.5.3 Mean white noise gain constraint

The synthesis associated with a constraint on the mean white noise gain $\mathrm{WNG_m}$ shows a clearly smaller deviation from the desired HRTF for $\theta = 0°$ (cf. Figure 3.5). However, this synthesis also exhibits larger deviations from the desired HRTF for the contralateral direction $\theta = 225°$ at frequencies $f \geq 4$ kHz.

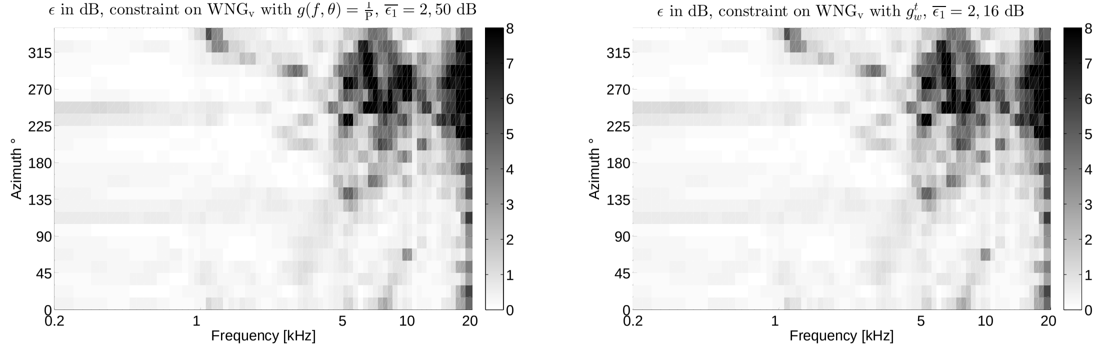
This is also apparent in the right panel of Figure 3.6 (cf. $\theta = 0°$). In general, the synthesis associated with a constraint on $\mathrm{WNG_m}$ yields a rather homogeneous and small error $\epsilon$ for ipsilateral directions. This, however, comes at the cost of a slightly higher error $\epsilon$ for contralateral directions ($180° \leq \theta \leq 360°$) compared to the synthesis associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$, especially at frequencies $5 \text{ kHz} \leq f \leq 10$ kHz and $f \geq 15$ kHz.

Overall, the filter optimization associated with a constraint on $\mathrm{WNG_m}$ results in a much better synthesis compared to the filter optimization associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), \theta_d)$, in particular for direction $\theta_d$.


$\epsilon$ in dB, constraint on $\mathrm{WNG}(\mathbf{w}(f), 0^\circ)$, $\overline{\epsilon_1} = 2,18$ dB

$\epsilon$ in dB, constraint on $\mathrm{WNG}_\mathrm{m}$, $\overline{\epsilon_1} = 2,55$ dB

**Figure 3.6:** Error $\epsilon$ as a function of frequency and azimuthal direction for the synthesis of a left ear HRTF associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0^\circ)$ ($\beta = -5$ dB) in the left panel and associated with a constraint on $\mathrm{WNG}_\mathrm{m}$ ($\beta_\mathrm{m} = -5$ dB) in the right panel. The illustration of $\epsilon$ is limited to 8 dB for the sake of clarity.


$\epsilon$ in dB, constraint on $\mathrm{WNG}_\mathrm{v}$ with $g(f,\theta) = \frac{1}{\mathrm{P}}$, $\overline{\epsilon_1} = 2,50$ dB

$\epsilon$ in dB, constraint on $\mathrm{WNG}_\mathrm{v}$ with $g_w^t$, $\overline{\epsilon_1} = 2,16$ dB

**Figure 3.7:** Error $\epsilon$ as a function of frequency and azimuthal direction for the synthesis of a left ear HRTF associated with a constraint on $\mathrm{WNG}_\mathrm{v}$ ($\beta_\mathrm{v} = -5$ dB) with a weighting function $g(f,\theta) = \frac{1}{\mathrm{P}}$ in the left panel and the weighting function $g_w^t(f,\theta)$ according to equation 3.22 in the right panel.

### 3.5.4 Constraint on the white noise gain for multiple frequencies

The error $\epsilon$ associated with a constraint on $\mathrm{WNG_v}(\mathbf{w_v}(\boldsymbol{\Omega}^t))$, incorporating the grouping of frequencies within ERBs as discussed in section 3.4.3, with the weighting function $g(f,\theta) = \frac{1}{\mathrm{P}}$ and $\beta_\mathrm{v} = -5$ dB is depicted in the left panel of Figure 3.7. As can be expected, the error $\epsilon$ is smoother and decreases in comparison to the error $\epsilon$ associated with a constraint on $\mathrm{WNG_m}$ and on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ (cf. Figure 3.6). This effect can mainly be observed for contralateral directions and higher frequencies which are associated with broader ERBs.

The error $\epsilon$ associated with a constraint on $\mathrm{WNG_v}$ considering frequencies within ERBs with a frequency- and direction-dependent weighting according to $g_w^t(f,\theta)$ in equation 3.22 and $\beta_\mathrm{v} = -5$ dB is depicted in the right panel of Figure 3.7. It can be observed that a psychoacoustically motivated weighting of $\mathrm{WNG_v}$ yields an even smoother error $\epsilon$, especially for contralateral directions where the weighting related to the desired directivity pattern $\mathrm{D}(f,\theta)$ is generally lower.

### 3.5.5 Robustness of the different regularization approaches

In the sections 3.5.2 to 3.5.4 and in Figures 3.6-3.7, only the error $\epsilon$ of the synthesis for one set of measured steering vectors was considered and discussed. However, it is in addition very relevant to consider the error $\epsilon$ associated with the different regularization approaches for disturbed steering vectors, i.e. to investigate the robustness of the synthesis against deviations of the steering vectors. To this end, synthesized directivity patterns were calculated for the measured but also for analytic steering vectors, both incorporating an additional random deviation of the steering vectors $\mathbf{d}(f,\theta)$. The analytic steering vectors were calculated assuming far-field conditions according to equation 1.7, i.e. pure delays with the same microphone topology as the measured steering vectors. Independent normally distributed vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ with zero-mean and a variance of $\sigma^2 = 1 - 10^{\frac{-0.05}{10}}$ were added to the real and imaginary parts of the steering vector $\mathbf{d}$ for each $\theta$ and $f$ independently (cf. Bendat and Piersol (1986)), resulting in the disturbed steering vector $\overset{\sim}{\mathbf{d}}\,(f,\theta)$,

$$\overset{\sim}{\mathbf{d}}\,(f,\theta)=\big(\Re\{\mathbf{d}(f,\theta)\}+\mathbf{r}_1(f,\theta)\big)+\big(\Im\{\mathbf{d}(f,\theta)\}+i\cdot\mathbf{r}_2(f,\theta)\big), \qquad (3.24)$$

with $\Re\{\mathbf{d}\}$ and $\Im\{\mathbf{d}\}$ denoting the real and the imaginary part of the (undisturbed) steering vector $\mathbf{d}$, respectively.

A Monte-Carlo simulation with 100 realizations of the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$

**Table 3.1:** Mean error $\bar{\epsilon}$ for the presented regularization approaches without and with an additional random deviation of the analytic and measured steering vectors.

| $\overline{\epsilon_1}$ in dB<br>5 kHz $\leq f \leq$ 18 kHz | analytic | | measured | |
|---|---|---|---|---|
| | $\mathbf{d}$ | $\tilde{\mathbf{d}}$ | $\mathbf{d}$ | $\tilde{\mathbf{d}}$ |
| non regularized | 0 | 9,98 | 0 | 7,45 |
| joint optimization with K=4 | / | / | 2,40 | 5,39 |
| constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ | **2,69** | 4,94 | 2,18 | 4,68 |
| constraint on $\mathrm{WNG_m}$ | 3,16 | 4,38 | 2,55 | 4,26 |
| constraint on $\mathrm{WNG_v}$ with $g = \frac{1}{\mathrm{P}}$ | 3,03 | **4,35** | 2,50 | **4,21** |
| constraint on $\mathrm{WNG_v}$ with $g_w^t$ | 2,73 | 4,47 | **2,16** | 4,41 |
| $\overline{\epsilon_2}$ in dB<br>0 Hz $\leq f \leq$ 20 kHz | analytic | | measured | |
| | $\mathbf{d}$ | $\tilde{\mathbf{d}}$ | $\mathbf{d}$ | $\tilde{\mathbf{d}}$ |
| non regularized | 0,12 | 42,84 | 0 | 18,67 |
| joint optimization with K=4 | / | / | **1,06** | 13,86 |
| constraint on $\mathrm{WNG}(\mathbf{w}(f), 0°)$ | **1,52** | 2,97 | 1,07 | 2,76 |
| constraint on $\mathrm{WNG_m}$ | 1,70 | 2,76 | 1,22 | 2,64 |
| constraint on $\mathrm{WNG_v}$ with $g = \frac{1}{\mathrm{P}}$ | 1,65 | **2,75** | 1,20 | **2,63** |
| constraint on $\mathrm{WNG_v}$ with $g_w^t$ | 1,54 | 2,85 | 1,08 | 2,75 |

was performed. In Figure 3.8 the mean error $\tilde{\epsilon}$ (for 100 realization of disturbed analytic steering vectors $\tilde{\mathbf{d}}$) associated with a constraint on the white noise gain for a single direction $\mathrm{WNG}(\mathbf{w}(f), 0°)$ is shown. Although this error is quite similar (cf. $\theta = 0°$) to the error $\epsilon$ for undisturbed measured steering vectors $\mathbf{d}$ in the left panel of Figure 3.6, the mean error $\tilde{\epsilon}$ in Figure 3.8 is significantly higher, primarily for contralateral directions.

To obtain a single error measure, the mean error $\bar{\epsilon}$ over all realizations, all directions and all frequencies was calculated for two different frequency ranges, shown in Table 3.1. In the upper Table, the mean error $\overline{\epsilon_1}$ for frequencies 5 kHz $\leq f \leq$ 18 kHz is shown since the largest errors primarily occur at frequencies $f \geq$ 5 kHz and are of less relevance for $f >$ 18 kHz. In the lower Table, the mean error $\overline{\epsilon_2}$ for frequencies 0 Hz $< f \leq$ 20 kHz is shown.

Both errors $\overline{\epsilon_1}$ and $\overline{\epsilon_2}$ increase when using the disturbed steering vectors $\tilde{\mathbf{d}}$ compared to the undisturbed steering vectors $\mathbf{d}$ for analytical as well for measured steering vectors and for all (regularization) approaches. Especially for the synthesis associated with a non-regularized filter optimization, the errors $\overline{\epsilon_1}$ and $\overline{\epsilon_2}$ increase drastically when the steering vectors are disturbed, which was to be expected.

For the synthesis associated with a joint optimization (using $\mathrm{K}=4$ sets of measured steering vectors) the increase of the mean synthesis errors for the disturbed steering vectors is slightly smaller than for the non-regularized optimization. Interestingly, the increase of the error for the disturbed steering vectors is considerably larger for $\overline{\epsilon_2}$ than for $\overline{\epsilon_1}$. Note also that the error $\overline{\epsilon_2}$ is very small for undisturbed steering vectors, which is also reflected in the low $\mathrm{WNG_m}$ of the associated filter coefficients for lower frequencies in Fig. 3.4. Overall, the joint optimization with $\mathrm{K}=4$ only slightly enhances the robustness of the synthesis, which may presumably be enhanced by incorporating more sets of measured steering vectors.

The optimization associated with a constraint on the white noise gain for a single direction $\mathrm{WNG}(\mathbf{w}(f), 0^\circ)$ yields the smallest mean synthesis errors when assuming undisturbed analytic steering vectors and the smallest $\overline{\epsilon_2}$ when assuming undisturbed measured steering vectors. This may be explained by the direction-dependent regularization which primarily influences only the direction $\theta_d$. When considering disturbed steering vectors, a constraint on the white noise gain for a single direction yields smaller mean synthesis errors than the joint optimization with $\mathrm{K}=4$ (but larger mean synthesis errors compared to constraints on $\mathrm{WNG_m}$ and $\mathrm{WNG_v}$). However, it should be kept in mind that a white noise gain for a single direction yields a direction-dependent impact on the synthesis (cf. section 3.5.2) and is hence undesirable for a multi-directional synthesis.

For the synthesis associated with constraints on the mean white noise gains $\mathrm{WNG_m}$, $\mathrm{WNG_v}$ with $g(f,\theta)=\frac{1}{\mathrm{P}}$ or with $g_w^t(f,\theta)$, the mean errors $\overline{\epsilon_1}$ and $\overline{\epsilon_2}$ further decrease for the disturbed steering vectors $\tilde{\mathbf{d}}$ compared to the synthesis associated with a constraint on $\mathrm{WNG}(\mathbf{w}(f), 0^\circ)$ and a joint optimization with $\mathrm{K}=4$. This holds for the analytic as well as for the measured steering vectors, showing that these proposed regularization approaches further increase the robustness compared to a constraint on the white noise gain for a single direction $\mathrm{WNG}(\mathbf{w}(f), 0^\circ)$. The highest robustness is obtained by the constraint on the mean white noise gain $\mathrm{WNG_v}$ taking into account multiple frequencies. The weighting of the $\mathrm{WNG_v}$ with $g_w^t(f,\theta)$ seems only beneficial for undisturbed steering vectors and disadvantageous (compared to the weighting $g(f,\theta)=\frac{1}{\mathrm{P}}$ and $\mathrm{WNG_m}$) when assuming deviations of the steering vectors.

In conclusion, the proposed regularization approaches constraining the mean white noise gains $\mathrm{WNG_m}$ and $\mathrm{WNG_v}$ seem to be appropriate for regularizing the synthesis of head-related transfer functions using a virtual artificial head. The multi-directional synthesis and its robustness is clearly enhanced when considering all directions for the white noise gain ($\mathrm{WNG_m}$)

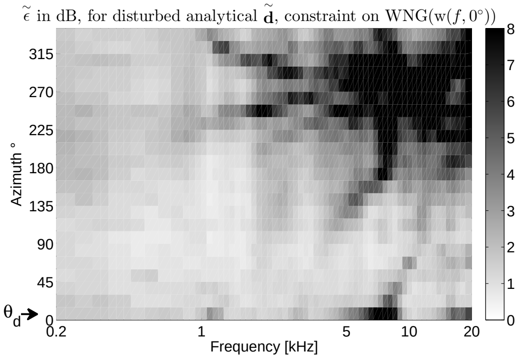


**Figure 3.8:** Mean error $\tilde{\epsilon}$ for 100 randomly disturbed analytic steering vectors $\tilde{\mathbf{d}}$ as a function of frequency and azimuthal direction for the synthesis of a left ear HRTF associated with a constraint on WNG($\mathbf{w}(f)$, 0°).

and when incorporating a psychoacoustically-motivated frequency grouping ($\text{WNG}_\text{v}$).

## 3.6 Conclusion

In this study, several approaches to increase the robustness of filter-and-sum beamformers for synthesizing spatial directivity patterns were presented, with a focus on synthesizing HRTFs. Firstly, a design procedure incorporating multiple measured steering vectors into the optimization was shown to enhance the robustness against random deviations of the steering vectors compared to a non-regularized optimization. However, a joint optimization for only 4 differently measured sets of steering vectors resulted in a lower robustness at lower frequencies compared to using a constraint on the WNG for one look direction.

Secondly, a mean weighting of the WNG over all directions was presented and shown to substantially outperform the WNG for a single look direction when synthesizing spatial directivity patterns with multiple desired directions (e.g. HRTFs).

Thirdly, a design for incorporating neighbouring frequency bins in the optimization and regularization, inspired by the frequency grouping of the human auditory system, was presented. It was shown that the approach incorporating frequencies within ERBs resulted in the best robustness of the synthesis for measured as well as for analytic steering vectors. This shows the suitability of the presented regularization approaches for synthesizing HRTFs using microphone arrays. The proposed weighting of the WNG within frequency bands was shown to be beneficial only when assuming no/small deviations of the steering vectors.

# Part II

# Smoothing individual HRTFs prior to the synthesis using the virtual artificial head

In part II, the psychoacoustical aspects when smoothing individual HRTFs prior to the synthesis using the VAH are investigated. This smoothing as a preprocessing step is motivated by the assumption that spatially smoother directivity patterns may be synthesized with less microphones and/or enhance the performance of the VAH in terms of robustness. At the same time it is known from literature that measured HRTFs may be smoothed to certain degrees without significantly deteriorating the spatial perception (cf. Kulkarni and Colburn (1998), Kulkarni *et al.* (1999), Huopaniemi *et al.* (1999) and Xie and Zhang (2010)). Hence, a reasonable smoothing of the individual HRTFs (desired directivity patterns) which does not cause any perceptible degradations of the HRTFs may improve the performance of the VAH when applied to the desired HRTFs prior to the synthesis. To this end, it is examined in the next two chapters which type and amount of smoothing is imperceptible when smoothing individual HRTFs.

In **Chapter 4**, smoothing is conducted by truncating the hrirs in the time domain, assuming that peaky directional characteristics of the HRTFs are also reflected in the time and frequency domain. However, it is shown that truncating the hrirs in the time domain to psychoacoustically reasonable lengths only marginally enhances the VAH-synthesis.

Hence, an imperceptible phase-linearization of HRTFs for higher frequencies and a subsequent complex-valued smoothing of the HRTFs (with a linearized phase for higher frequencies) is examined in **Chapter 5**. Moreover, the perceptual limits for a spatial smoothing of HRTFs by reducing the spatial dynamic range are presented. It is found that such an imperceptible smoothing of the HRTFs prior to optimizing the beamformer filter coefficients improves the VAH-synthesis.

# 4

# Psychoacoustically appropriate truncation of individual hrirs[*]

---

[*] This chapter is based on the publication

- Rasumow, Blau, Hansen, Doclo, van de Par, Püschel, and Mellert (2012b), "Smoothing head-related transfer functions for a virtual artificial head," in *Acoustics 2012*, Nantes, France, pp. 1019-1024

*The frequency-dependent directivity pattern of a human head (head-related transfer functions, HRTFs) can be synthesized with a microphone array and digital filtering (cf. Rasumow* et al. *(2011b)), which may be referred to as a virtual artificial head (VAH). However, in order to synthesize the peaky directivity patterns (especially at high frequencies) with a sufficiently high accuracy, a large number of microphones is required, resulting in an increased sensitivity to gain, phase and positioning errors. Therefore, it is beneficial to smooth the HRTFs in such a manner that they become (spatially) as smooth as possible while still yielding an unmodified individual perception. Since peaky directional characteristics of the HRTF are reflected in the frequency response, smoothing is conducted by truncating the head-related impulse responses (hrirs) in the time domain. The aim of this study is to investigate the limits of this smoothing operation in terms of the truncation length. Using a 3-AFC listening test, the truncation length was determined as a function of the angle of incidence in the horizontal plane. The conducted experiment indicates that a truncation of hrirs from 512 to about 250 samples (corresponding to approximately 5.7 ms at a sampling frequency of* $f_s = 44100$ *Hz) yield a discrimination ability approximately at 50% correct score.*

## 4.1 Introduction

The usual way to incorporate (subjective) binaural spatial information into recordings is to use so-called artificial heads, which are a replica of real human heads. A comprehensive summary of previous binaural recording techniques can be found in Paul (2009). Alternatively, the frequency-dependent directivity patterns of HRTFs can also be approximately synthesized using a set of spatially distributed microphones with appropriate filters (cf. Chen *et al.* (1992), Kahana *et al.* (1999), Tohtuyeva and Mellert (1999), Sakamoto *et al.* (2008), Rasumow *et al.* (2011b), Atkins (2011a), Atkins (2011b)). Such a setup may be referred to as a virtual artificial head (VAH). The performance of the VAH depends (inter alia) on the microphone topology, the number of microphones (cf. Rasumow *et al.* (2011b) or **Chapter 2**) and the applied cost function (cf. Nordebo *et al.* (1994), Doclo and Moonen (2003a), Rasumow *et al.* (2013a)) but also on the desired directivity pattern.

In general, a microphone array achieves a more accurate fit to the desired directivity pattern if this pattern is spatially smooth. One indirect way of obtaining spatially smooth HRTFs is to smooth them in the frequency domain (or equivalently, truncate the corresponding hrirs). Therefore, the following listening experiment explored the ability to discriminate between (individual) reference-HRTFs and HRTFs containing reduced information. The experiment was supposed to answer the following research question:

- What is the appropriate truncation length for the direction-dependent head-related impulse responses (hrir) that leads to an unaltered perception compared to a chosen reference setting?

## 4.2 Test preparation

### 4.2.1 HRTF measurement

Prior to each test session, individual HRTFs and HPTFs were measured in an anechoic room with a circular loudspeaker array (radius 1.25 m, custom-made loudspeaker boxes).

The HRTF measurements were carried out with one loudspeaker signal and two ear signals using the blocked ear method (cf. Hammershøi and Møller (1996)) with custom-made ear shells and Knowles FG-23329 miniature electret microphones. The transfer functions were estimated using white noise test signals and standard FFT-based techniques ($H_1$ estimate with 8192-point Hann window, 50% overlap, 52 averages at a sampling frequency of $f_s = 44100$ Hz). The measured transfer functions were divided by the free field transfer function (cf. equation 1.1) obtained from a measurement (G.R.A.S. Microphone Type 40AF) at the position of the center of the head without the acoustic influence of the subject.

In order to provide a rough overview of a potential direction-dependency of the hrir-length ($l_{hrir}$) at threshold and to limit the number of experiments to a manageable amount, three different source positions (out of the 24 measured) in the horizontal plane with azimuth angles 0° (front), 90° (left) and 225° (back right) were chosen.

### 4.2.2 Reference setting

In order to obtain perceptual limits for reducing information from individual head-related impulse responses, it is crucial to define a reference setting that represents the basis for comparison within the discrimination experiments. Ideally, this reference setting should contain all desired aspects/information regarding the individual HRTF, whereas all disturbances (e.g., measurement noise, non-linear distortion, reflections due to the measuring apparatus etc.) should be omitted to avoid potential misleading cues in the evaluation. The available measuring apparatus (especially the loudspeakers, microphones and anechoic room) enables a signal-to-noise ratio (SNR) of about 50-55 dB. The decaying hrir reaches this noise floor

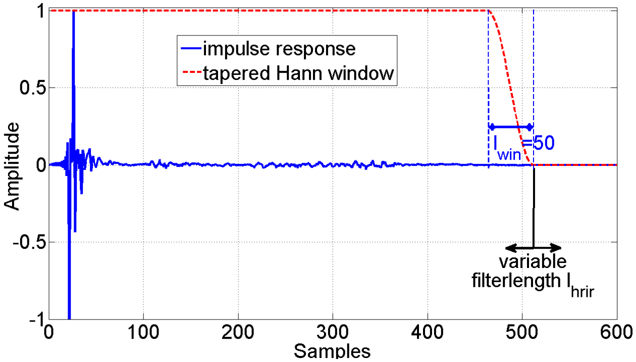


**Figure 4.1:** Example hrir (left ear, azimuth=90°) with the reference length $l_{hrir} = 512$ samples and a tapered Hann-window with the fixed length $l_{win} = 50$ samples for the descending flank.

after approximately $400 - 500$ samples, corresponding to approximately $9 - 11$ milliseconds (at a sampling frequency of $f_s = 44100$ Hz). It is partly for this reason that the hrir-length $l_{hrir}$ in the reference-setting was chosen as $l_{hrir} = 2^9 = 512$ samples ($\approx 11.6$ ms) for the subsequent experiment.

All used hrirs were flanked with a tapered Hann-window with a descending flank of $l_{win} = 50$ samples. An exemplary hrir (blue line) and its associated window (red dashed line) with the previously explained parameters are shown in Figure 4.1.

As a first control, white noise stimuli filtered with individual HRTF-filters with $l_{hrir} = 512$ from 24 directions (equidistant 15°-spacing) in the horizontal plane were played dichotically to each subject via headphones. All of the tested subjects could assign the corresponding various source positions and perceive the presented signals outside the head.

## 4.3 Listening tests - material and methods

### Generic Procedure and Stimuli

In order to cover a wide frequency range and include temporal cues, the test signal contained short bursts of white noise with a spectral content of 350 Hz $\leq f \leq$ 18000 Hz. Each signal was constructed out of three segments, each with a noise burst of 0.15 seconds with 0.01 seconds raised-

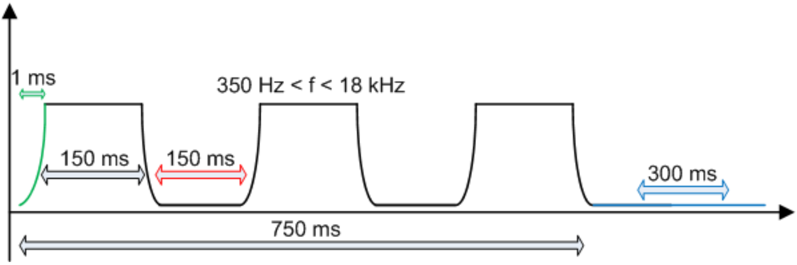


**Figure 4.2:** Sketch of the envelope of the test signal as a function of time. Each noise burst (350 Hz $\leq f \leq$ 18000 Hz) lasted 0.15 seconds and was windowed with 0.01 seconds onset-offset ramps. The pauses between the noise bursts lasted 0.15 seconds as well.

cosine onset and offset ramps, followed by silence of 0.15 seconds, resulting in a total signal length of 0.75 seconds. A three-alternative forced-choice (3-AFC) paradigm was employed to determine threshold values for the hrir-length. Three intervals of the filtered signals were presented to the subjects in randomized order, each separated by 0.3 seconds silence. One out of the three signals was filtered with the modified hrirs that were truncated in the time domain and two with the reference hrirs. The subjects were instructed to indicate the odd of the three intervals. Feedback for the correct answer was presented after each trial. The variable of the test signal was modified according to the weighted up-down method as to estimate the 80%-, as well as the 40%-correct score. These particular values of the psychometric function were targeted by adaptively modifying the step size according to Kaernbach (1991). The initial variable was a filter length of $l_{hrir} = 256$. The variable $l_{hrir}$ was varied adaptively in the familiarization phase until a minimal step of 3 samples was reached. As soon as this minimal step was reached, the measurement phase started and the variable was not reduced any more. The threshold of the examined variable was defined as the median of 6 following reversals within the measurement phase.

Four male and experienced subjects participated in this experiment in which each combination of source direction, desired correct score and subject was repeated at least three times.

All experiments were designed using *psylab*[1] (cf. Hansen (2006)), which is a set of MATLAB–scripts for the design of various psychoacoustical detection and discrimination experiments. The signals were generated digitally

---

[1] *Psylab* is available at `http://www.hoertechnik-audiologie.de/psylab/`

and presented via a RME ADI-8 DA-converter and AKG K 240 Studio headphones (individually equalized) at an overall sound pressure level of 78 dB SPL to the subject, seated in an anechoic room. The system was calibrated using the Artificial Ear Type 43AA (G.R.A.S.) with the preamplifier Type 27A (G.R.A.S.).

The variable $l_{hrir}$ was varied adaptively, whereas in each run, the target correct score (either 40% or 80%) and the source direction (one of $\theta = 0^\circ$, $\theta = 90^\circ$ and $\theta = 225^\circ$) were kept constant. It should be noted that the length of the fading window was set to $l_{win} = 50$ samples for all values of $l_{hrir}$ (see Figure 4.1). In the case that a particular length of the impulse response $l_{hrir}$ was shorter than 50 samples, $l_{win}$ was set to this particular $l_{hrir}$.

## 4.4 Listening tests - results

Equation 4.1 describes the used model psychometric function $\Psi(x)$ which is controlled by the parameters $x_m$ and $m$, where $x_m$ is the midpoint of the base function $\Psi_0(x)$, i.e. $\Psi_0(x_m) = \frac{1}{2}$, and $m$ is the slope of $\Psi_0(x)$ at $x_m$, i.e. $\Psi_0{}'(x_m) = m$. The additional value $\frac{1}{3} = p_{\text{guess}}$ results from the 3-AFC method.

$$\Psi_0(x) = \frac{1}{1 + e^{-4m(x-x_m)}}, \qquad \Psi(x) = \frac{1}{3} + \frac{2}{3} \cdot \Psi_0(x) \tag{4.1}$$

The experiments yielded the 40% and 80% correct points, $x_{40}$ and $x_{80}$, on the psychometric function for each condition. The parameters $x_m$ and $m$ could therefore be expressed in a closed-form solution as a function of $x_{40}$ and $x_{80}$ as:

$$x_m = \left(-\frac{\ln(3/7)}{\ln(9)} \cdot x_{40} + x_{80}\right) \cdot \left(\frac{1}{1 - \frac{\ln(3/7)}{\ln(9)}}\right) \tag{4.2}$$

$$m = \frac{\ln(9)}{4 \cdot (-x_{40} + x_m)} \tag{4.3}$$

Subsequently, also the 50% threshold $x_{50}$, defined by $\Psi(x_{50}) = \frac{1}{2}$, could then be computed. This was deemed particularly interesting since it was chosen to represent the just noticeable difference (JND) of the investigated variable (i.e. $l_{hrir}$).

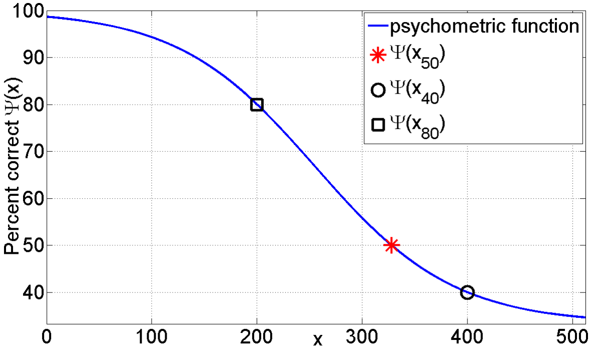


**Figure 4.3:** Exemplary psychometric function $\Psi(x)$ estimated from two supposed arguments $x_{80} = 200$ and $x_{40} = 400$. According to equation 4.1 and 4.2, this yields an argument at threshold ($\Psi(x_{50}) = 0.5$) of $x_{50} \approx 328$.

## Results

The results from the experiment are shown in Figure 4.4. The mean values and standard deviation of the 80% scores and 40% scores, $l_{hrir,80}$ (green symbols) and $l_{hrir,40}$ (blue symbols), and the interpolated 50% threshold $l_{hrir,50}$ (red symbols) are indicated by individual symbols for each subject. The inter-subject mean value for $l_{hrir,50}$ is shown as a red dashed line. The results are separated into three panels according to the source direction ($\theta = 0°$, $\theta = 90°$, and $\theta = 225°$).

It is evident for the left ($\theta = 0°$) and the middle panel ($\theta = 90°$) that the standard deviation for the 80% threshold is smaller than for the 40% threshold. One possible explanation may be the fact that the acoustical cue to be distinguished is more present for the shorter $l_{hrir}$ at 80% score (cf. Figure 4.3). A 40% score on the other hand corresponds to rather long $l_{hrir}$ and consequently less audible cues which may introduce a greater variance in the individual judgement process. Furthermore, also a large standard deviation of the 80% score is observed for $\theta = 225°$ for three of the four subjects. Although each subject received a training phase of at least one complete run, three of four subjects might have altered their judgement successively (not shown). This may suggest that some subjects experienced a learning effect or potentially evaluated different acoustical cues in successive runs. However, due to the small population and a deviating focus, this aspect will not be investigated further in the present study.

For the source direction $\theta = 0°$, the inter-subject mean $l_{hrir,50}$ at threshold

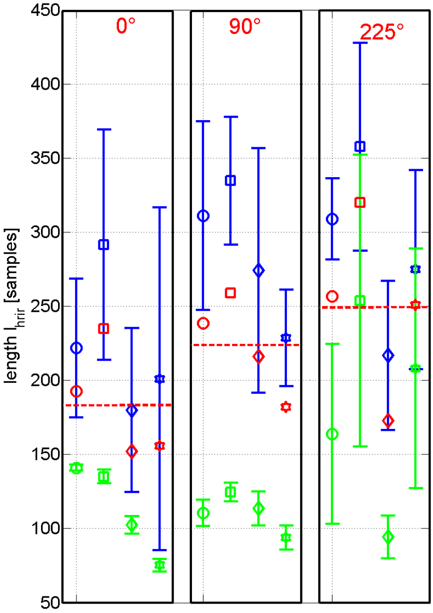


**Figure 4.4:** Means and standard deviations of thresholds in the truncation experiment for 80% score (green) and 40% score (blue) for four subjects (○, □, ◇ and ✡). Red symbols indicate the calculated 50% scores and the dashed lines indicate their mean across subjects.

was 184 samples, meaning that for the mean performance the individual hrir from $\theta = 0°$ could be truncated to this length, in order to yield a 50% score for discrimination relative to the reference setting (cf. section 4.2.2). For the most sensitive subject, this threshold was at $l_{hrir,50} = 235$ samples. For the source direction $\theta = 90°$, the mean $l_{hrir,50}$ at threshold increased to 224 samples and to 259 samples for the most sensitive subject. The longest $l_{hrir,50}$ could be observed for source direction $\theta = 225°$. The mean $l_{hrir,50}$ at threshold for this condition was 250 samples, with the most sensitive subject yielding 320 samples.

To summarize, the $l_{hrir,50}$ at threshold seems to depend on the subject, the source direction and in some cases on a potential learning effect (source direction of $\theta = 225°$). For the source directions $\theta = 0°$ and $\theta = 90°$, a truncation of the hrirs from $l_{hrir} = 512$ samples to $l_{hrir} \approx 250$ samples (corresponding to approximately 5.7 ms) seems to retain all essential information, since the last part of the hrirs does not seem to alter the discrimination noticeably. For the source direction $\theta = 225°$, it appears that

$l_{hrir,50}$ increases. Especially for subjects who repeated the experiments several times (more than 3 runs plus training phase) the $l_{hrir,50}$ at threshold seems to increase steadily. However, one must bear in mind that the discrimination in the conducted experiments is based on all feasible aspects of the hrirs (as opposed to e.g. localization only). Since the learning effect may also be due to a reflection or noise in the HRTF (generally HRTFs from $\theta = 225°$ exhibit the worst SNR of the tested source directions), this effect requires further investigation with a larger population and a finer directional grid.

## 4.5 Impact of information reduction on the performance of the VAH

The results from the truncation experiment reveal the opportunity to truncate the hrirs and thus to yield a smoother, but psychoacoustically equivalent frequency response. For the VAH application, this raises the question whether the smoother frequency responses enable a more precise fit of the resulting directivity pattern of the beamformer to the desired HRTFs. In order to compare the synthesis of the VAH with respect to $l_{hrir}$, the synthesis error $\epsilon_f(f_c)$ was used, with

$$\overline{\epsilon}_f = \frac{1}{\mathrm{B}\cdot\mathrm{P}}\sum_{b=1}^{\mathrm{B}}\sum_{i=1}^{\mathrm{P}}\epsilon(f_b,\theta_i) , \tag{4.4}$$

where $\epsilon(f_b,\theta_i)$ denotes the synthesis error between the resulting directivity pattern $\mathrm{H}(f,\theta)$ and the desired directivity pattern $\mathrm{D}(f,\theta)$ averaged over frequencies in ERB-bands according to equation 1.14, with $f_b$ as its center frequency and azimuthal direction $\theta_i$. This error measure is further averaged in equation 4.4 over all directions and center frequencies $0\ \mathrm{Hz} < f \leq 16000\ \mathrm{Hz}$, yielding the mean synthesis error $\overline{\epsilon}_f$. The left-ear hrirs of subject ○ were applied to the VAH with the same settings as in **Chapter 2** and Rasumow *et al.* (2011b) (i.e. Golomb-topology, 24 microphones, theoretical steering vectors according to equation 1.7 and no artificial positioning error). The length of the head-related impulse responses that were used as the desired directivity patterns D for the VAH-synthesis was either $l_{hrir} = 512$ samples or $l_{hrir} = 300$ samples, assuming that the finding from section 4.4 approximately holds for all directions in the horizontal plane.

It appears from the simulations that the advantage of truncating hrirs depends on the regularization of the VAH-filters. Regularization was applied by imposing a norm constraint ($\beta_\Gamma \geq \mathbf{w}^{\mathbf{H}}\mathbf{w}$) on the filter coefficients (see equation 2.1 and 2.2 in **Chapter 2** or equation 6 and 7 in Rasumow *et al.* (2011b)). Hardly no advantage is achieved if no regularization is applied, whereas the improvement of the VAH still stays very small for reasonable regularization parameters. The mean fitting error decreases from $\bar{\epsilon}_f = 1.47$ dB with $l_{hrir} = 512$ samples to $\bar{\epsilon}_f = 1.42$ dB with $l_{hrir} = 300$ samples when a $\beta_\Gamma = 0$ dB is used, for instance. For a desired $\beta_\Gamma = 5$ dB, the mean fitting error decreases from $\bar{\epsilon}_f = 1.04$ dB with $l_{hrir} = 512$ samples to $\bar{\epsilon}_f = 0.98$ dB with $l_{hrir} = 512$ samples. This analysis implies that truncating the desired head-related impulse responses to psychoacoustically reasonable lengths causes only very small analytical improvements of the VAH-synthesis.

## 4.6 Summary and further steps

The present study deals with the ability to distinguish between variously truncated hrirs. It was found that on average the tested subjects could distinguish between hrirs with $l_{hrir} = 512$ and $l_{hrir} \approx 250$ at a 50% correct score. For some hrirs (source direction of 225°) there appears to be a learning effect in recognizing truncated hrirs which needs further investigation.

It should be noted that all the reviewed aspects only depict tendencies based on experiments for three source directions. Further investigations with more subjects and source directions are needed in order to provide general statements.

It could be shown that the reduction of information of hrirs in the time-domain revealed no considerable enhancement regarding the application to a VAH. For this reason further steps may be for instance a psychoacoustically reasonable smoothing using constant *relative* bandwidths. Moreover, a major advantage regarding the VAH may be a reasonable spatial smoothing of the desired directivity patterns. Hence, further experiments investigating the perceptual limits for a spectral and spatial smoothing of HRTFs are presented in **Chapter 5**.

# 5

# Smoothing HRTFs in the frequency and spatial domains[*]

---

[*] This chapter is based on the publications

- Rasumow, Blau, Doclo, Hansen, Mellert, Püschel, and van de Par (2012a), "Psychoakustisch motivierte Glättung von kopfbezogenen Übertragungsfunktionen: Hörbarkeit der Linearisierung von Phasengängen," in *Fortschritte der Akustik - DAGA 2012*, Darmstadt, Germany, pp. 633-634.
- Rasumow, Blau, van de Par, Hansen, Doclo, Püschel, and Mellert (2013b), "Subjective importance of individual HRTF phase," in *Proc. Annual Conference on Acoustics (AIA-DAGA)*, Merano, Italy, pp. 604-607.
- Rasumow, Blau, Hansen, van de Par, Doclo, Mellert, and Püschel (2014c), "Smoothing individual head-related transfer functions in the frequency and spatial domains," *Journal of the Acoustical Society of America* 135(4), pp. 2012-2025.

*When synthesizing individual head-related transfer functions (HRTFs) with a microphone array, smoothing HRTFs spectrally and/or spatially prior to the computation of appropriate microphone filters may improve the synthesis accuracy. In this study, the limits of the associated HRTF modifications, until which no perceptual degradations occur, are explored.*

*First, complex-valued spectral smoothing of HRTFs into constant relative bandwidths was considered. As a prerequisite to complex-valued smoothing, the HRTF phase spectra were substituted by linear phases, either for the whole frequency range or above a certain cut-off frequency only. The results indicate that a broadband phase linearization of HRTFs can be perceived for certain directions/subjects and that the thresholds can be predicted by a simple model. HRTF phase spectra can be linearized above 1 kHz without being detectable. After substituting the original phase by a linear phase above 5 kHz, HRTFs may be smoothed complexly into constant relative bandwidths of 1/5 octave, without introducing noticeable artifacts.*

*Second, spatially smoother HRTF directivity patterns were obtained by levelling out spatial notches. It turned out that spatial notches do not have to be retained if they are less than 29 dB below the maximum level in the directivity pattern.*

## 5.1 Introduction

Spatial information is a major factor in the perception and appraisal of sounds. A typical way to include spatial information into recordings and measurements is to use so-called artificial heads, which are reproductions of real human heads with microphones placed in the ear canals (cf. Paul (2009) for a recent overview). Alternatively, the direction- and frequency-dependent head-related transfer functions (HRTFs) can be approximately synthesized using a set of spatially distributed microphones with appropriate digital filtering (cf. Chen *et al.* (1992), Tohtuyeva and Mellert (1999), Kahana *et al.* (1999), Atkins (2011a) and Rasumow *et al.* (2011b, 2013a)). Such a device is referred to as a virtual artificial head (VAH).

The performance of such a VAH not only depends on the microphone array (topology, number of microphones, calibration, mechanical stability) and the filter design procedure but also on the desired directivity patterns, i.e. the HRTFs to be synthesized. In general, more microphones are needed when the directivity patterns exhibit more spatial detail, or equivalently, the accuracy decreases given a fixed number of microphones.

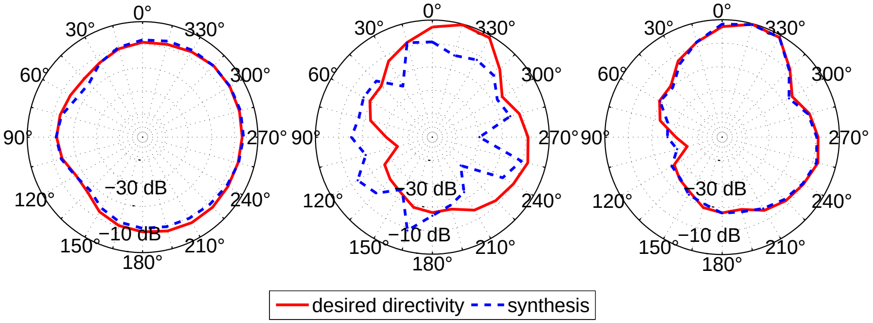


**Figure 5.1:** Illustration of the interrelationship between the spatial complexity of HRTFs and the number of microphones needed to accurately synthesize them. The example HRTF set was taken from the IRCAM database (right ear of subject #1002), the original directivity is shown as solid line, the synthesis as dashed line. Left: $f = 1$ kHz N = 8 microphones, center: $f = 11$ kHz N = 8 microphones, right: $f = 11$ kHz N = 24 microphones.

This is illustrated in Figure 5.1 where a sample HRTF set (IRCAM database[1] subject #1002, right ear) is synthesized using different numbers of microphones (the steering vectors of which were approximated by pure delays, cf. equation 1.7) and the methodology described in Rasumow *et al.* (2013a). At $f = 1$ kHz (left diagram) the HRTF directivity is spatially smooth and can accurately be synthesized with N = 8 microphones, whereas at $f = 11$ kHz it becomes spatially more detailed and the synthesis with N = 8 microphones fails completely (center diagram). If the same directivity is synthesized with N = 24 microphones, an accurate fit is obtained again (right diagram).

For the VAH, one seeks to minimize the number of microphones because this not only reduces the cost of the system but also its sensitivity to e.g. microphone gain, phase and position errors (cf. Rasumow *et al.* (2011b) or **Chapter 2**).

Hence, given the observation made above, a preprocessing step involving spatial smoothing of the HRTFs could improve the concept of a VAH in terms of the number of microphones needed. It needs to be assured then of course that the preprocessing does not cause perceptible degradations

---

[1] The IRCAM HRTF-database is available at `http://recherche.ircam.fr/equipes/salles/listen` (date last viewed 15/1/14)

of the HRTFs. Therefore, we will investigate in this paper what type and amount of smoothing is still imperceptible.

Several types of smoothing of HRTFs will be considered in this study. One reasonable and well investigated way to smooth HRTFs is to reduce their spectral resolution, which has the indirect effect of obtaining spatially smoother directivity patterns. Spectral smoothing of HRTFs is often done by truncating the length of the corresponding *hrir*s (head-related impulse responses). Typically, filter lengths of 512 taps (corresponding to about 11.6 ms at a sampling frequency of $f_s = 44.1$ kHz) are considered to represent the individual cues of the HRTFs sufficiently well in order to get an externalized virtual directional perception (cf. Kulkarni *et al.* (1999), Huopaniemi *et al.* (1999) and Algazi *et al.* (2001)). Truncating the length of the *hrirs* corresponds to smoothing the HRTFs into constant absolute bandwidths. From a psychoacoustic point of view, smoothing into constant relative bandwidths is preferable (cf. Kohlrausch and Breebaart (2001)) since it is a well known phenomenon that the human ear groups incoming sounds into frequency bands that broaden with increasing center frequencies ("critical bands", cf. Fletcher (1940), Patterson and Nimmo-Smith (1980) and Moore (2003)). Also, compared to smoothing into constant absolute bandwidths, smoothing into constant relative bandwidths automatically results in more smoothing at higher frequencies, which in turn will result in smoother directivity patterns per (constant bandwidth) frequency bin at higher frequencies and is therefore beneficial to the VAH. The concept of smoothing HRTFs into constant relative bandwidths was for instance applied by Kohlrausch and Breebaart (2001) and Breebaart *et al.* (2010), who found that smoothing into approximately one critical band was acoustically transparent for spatial audio coding applications using non-individual HRTFs, or by Xie and Zhang (2010) who even proposed to smooth into up to 3.5 equivalent rectangular bandwidths (ERB) at higher frequencies ($f > 5000$ Hz) for the contralateral ear. However, smoothing into such broad bandwidths is assumed to lead to discriminable artifacts for the ipsilateral ear and generally at lower frequencies ($f < 5000$ Hz).

Although smoothing into constant relative bandwidths is psychoacoustically appealing, it may also produce inconsistencies between magnitude and phase spectra. For instance, when a complex-valued spectral smoothing is applied at frequencies above about 5 kHz, the noisy and/or steep phase characteristics of measured HRTFs result in notches in the magnitude of the complexly smoothed HRTF, which are not plausible, cf. Figure 5.2 (blue dashed line). Most often this problem is circumvented by smoothing the magnitude and the phase of measured HRTFs separately (cf. Kulkarni and Colburn (1998), Kohlrausch and Breebaart (2001) and Breebaart *et al.* (2010)) or by smoothing the magnitude only and supplementing it with a minimum-phase reconstruction (cf. Huopaniemi and Karjalainen (1996)). However, a separate manipulation of magnitude and phase may result in unwanted interaural cues. As an alternative, we propose to first simplify the phase spectrum in a perceptually transparent way and to subsequently smooth magnitude and phase spectra *simultaneously* (complex-valued smoothing) into constant relative bandwidths.

In order to obtain perceptually transparent phase simplifications of original HRTFs, the minimum-phase-plus-delay approach has been proposed (cf. Mehrgardt and Mellert (1977) and Kulkarni *et al.* (1999)). This approach is now widely used, most often (if not always) motivated by the listening tests carried out by Kulkarni *et al.* (1999). Interestingly though, the results obtained by Kulkarni *et al.* (1999) already indicate that a linear phase might be a perceptually better choice than a minimum phase and that *at low frequencies* it is worthwhile "making the overall low-frequency ITD in the model HRTFs to be the same as the overall low-frequency ITD in the empirical HRTFs", which in fact questions the validity of a broadband phase substitution. Therefore, it appears to be necessary to re-investigate the perceptual limits of a broadband phase substitution with regard to the

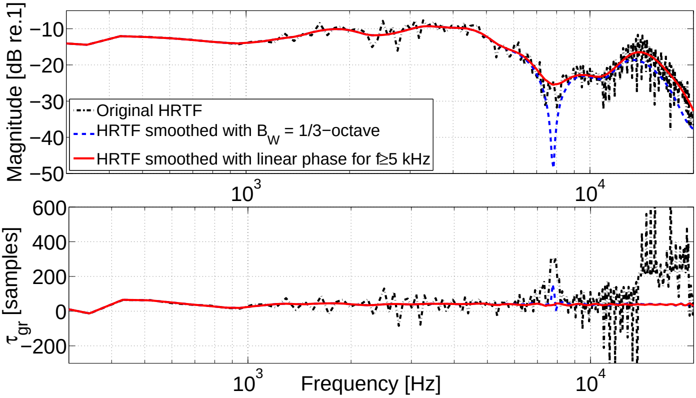


**Figure 5.2:** Magnitude in dB (top) and group delay in samples (bottom) for an exemplary HRTF of subject S2 for the left ear from azimuth $\theta = 210°$. The black dot-dashed lines show the original HRTF (NFFT = 512), the blue dashed lines show the resulting HRTF after complex-valued smoothing into third octave bands, the solid red lines show the smoothed HRTF (third octave bands) when the phase of the HRTF is substituted by a linear phase above 5 kHz before smoothing.

VAH.

As stated above, we hypothesize that primarily the *spatially* smoothness of HRTFs determines the effort needed to synthesize HRTFs with a VAH. Hence, a direct smoothing in the spatial domain may be even more profitable than a smoothing in the frequency domain, which would lower the spatial dynamic range only indirectly.

One well-known method to obtain spatially smoother HRTFs is to model the HRTFs through a spherical or cylindrical harmonic decomposition of finite order (cf. Duraiswaini *et al.* (2004), Zotkin *et al.* (2009) and Atkins (2011a)). This, however, has been shown to require an extremely large number of measurement directions and microphones for appropriate model orders (cf. Zotkin *et al.* (2009) and Castaneda *et al.* (2013)). Even if this additional effort is disregarded, the question of perceptual relevance remains and is more fundamental than the question of an appropriate spatial model structure. Therefore, we address the perceptual relevance without assuming any spatial model structure in terms of spherical or cylindrical harmonics.

Instead, we assume that the (spatial) dynamic range of narrowband directivity patterns can be limited by appropriately levelling out spatial notches. This in turn is motivated by assuming that spatial notches in the directivity patterns are of less perceptual relevance than spatial peaks, which seems a reasonable starting point given that spatial notches in measured HRTFs may become as low as -80 dB compared to the dominant direction (cf. Figure 5.5). More specifically, spatial notches mainly occur at contralateral directions and may therefore be masked by stronger ipsilateral components. Hence, a selective limitation of the spatial dynamic range for directions associated with lower levels in the directivity pattern may introduce only slight artifacts while decreasing the number of microphones needed in the VAH. We will therefore investigate to what extent the spatial notches can be reduced in depth without perceptual effects.

In summary, a successful preprocessing of HRTFs for a VAH should reduce the spatial complexity of narrowband directivity patterns associated with the HRTFs, without being perceptually distinguishable from the original HRTFs. This can be obtained indirectly by smoothing the HRTFs in the frequency domain or directly by limiting the spatial dynamic range of narrowband directivity patterns (which in turn is facilitated by a limited spectral resolution). In the following four experiments, the associated HRTF manipulations are evaluated with the aim of finding limits until which the original HRTFs could be altered without perceptible effects. In order to guarantee the highest flexibility in future VAH applications, all possible cues that the auditory system is able to access, including coloration, loudness, timbre, localization, etc., are considered. It should be noted that this is the most stringent criterion possible to discrimination tasks, which in some other applications might be relaxed (cf. Kulkarni and Colburn (1998) and Romigh *et al.* (2013)).

Section 5.2 introduces and explains the performed experiments investigating the reduction of the spatial dynamic range and the complex-valued smoothing as well as two experiments examining the discriminability of phase linearizations. The applied stimuli and methods as to investigate these smoothing methods and phase manipulations are presented in section 5.3. The results of these experiments are presented in section 5.4 and discussed in section 5.5, including the consequences of the examined types of smoothing for the synthesis using the VAH. The main findings and conclusions for an acoustically transparent preprocessing of HRTFs according to the VAH are summarized in section 5.6.

## 5.2 HRTF smoothing operations considered in this study

Firstly, complex-valued smoothing of HRTFs in the frequency domain into constant relative bandwidths was investigated. As discussed above, complex-valued smoothing into constant relative bandwidths requires a prior treatment of phase spectra. Inspired by the results from Kulkarni *et al.* (1999), we chose a linear phase model as a substitute for the original HRTF phase spectra. The linear phase $\phi_{lin}$ was in this case calculated from the delay of the maximum of the Hilbert envelope of the respective *hrir*, cf. Appendix A. Informal listening tests indicated that a broadband substitution of the original phase spectra by the such computed linear phase can indeed be perceptually distinguished from the original HRTFs. Therefore, in a first experiment, we investigated the sensitivity of listeners to a broadband substitution of the original phase by a linear phase (cf. section 5.3.3.1). More specifically, we tested to which extent listeners are sensitive to a broadband change of the original phase spectrum when fading the (unwrapped) original phase $\phi_{\mathrm{orig}}$ into the linear phase $\phi_{\mathrm{lin}}$ using a variable mixing gain $L_\phi$ between 0 and 1 such that

$$\phi_{\mathrm{test}}(f) = L_\phi \cdot \phi_{\mathrm{lin}}(f) + (1 - L_\phi) \cdot \phi_{\mathrm{orig}}(f). \tag{5.1}$$

The threshold value of the mixing gain $L_\phi$ then provides a means of characterizing the sensitivity of listeners to a broadband phase linearization of HRTFs: If listeners were insensitive to a broadband phase linearization, then $L_\phi$ should approach one whereas a low value of $L_\phi$ would indicate that listeners are sensitive to a broadband phase linearization.

Assuming that the results of the informal listening tests were confirmed in the first experiment (i.e. that listeners are sensitive to a broadband phase linearization), one may go a step further and, in line with the arguments by Kulkarni *et al.* (1999), apply the phase substitution at higher frequencies above a certain cutoff-frequency $f_\mathrm{c}$ only. Therefore, in the second experiment, the original phase $\phi_\mathrm{orig}$ was maintained below a variable cutoff-frequency $f_\mathrm{c}$ and substituted by a linear phase $\phi_\mathrm{lin}$ above $f_\mathrm{c}$, resulting in

$$\phi_\mathrm{test}(f) = \begin{cases} \phi_\mathrm{orig}(f), & \text{for } f \leq f_\mathrm{c} \\ \phi_\mathrm{lin}(f), & \text{for } f > f_\mathrm{c} \end{cases} . \tag{5.2}$$

If the cut-off frequency $f_\mathrm{c}$ is chosen too low, it may lead to audible changes. The threshold value of the cut-off frequency $f_\mathrm{c}$ as determined in this experiment then provides information about the range of cut-off frequencies that are allowable in the third experiment where complex-valued spectral smoothing was applied.

In the third experiment, the noticeability of complex-valued smoothing into constant relative bandwidths with a variable bandwidth $\mathrm{B_W}$ was investigated (cf. section 5.3.3.3). As a prerequisite the original phase was linearized for frequencies $f \geq f_\mathrm{c}$. Then, broader bandwidths $\mathrm{B_W}$ yield smoother HRTFs but also result in more audible artifacts and vice versa. The threshold value of the bandwidth $\mathrm{B_W}$ can then be used to formulate a smoothing rule for the VAH. The complex-valued smoothing into constant relative bandwidths was implemented according to a method proposed by Hatziantoniou and Mourjopoulos (2000). This manipulation is comparable to smoothing into equivalent rectangular bandwidths (ERB) for frequencies above approximately 1 kHz (cf. Glasberg and Moore (1990)), which are the frequencies of primary interest in VAH applications. A detailed description of the applied method for a complex-valued smoothing of HRTFs is given in Appendix B.

Secondly, we investigated the audibility of limiting the spatial dynamic range $\zeta$ directly in the spatial domain, i.e. over the directions of incidence. The spatial dynamic range $\zeta$ is given by

$$\zeta(f) = 20 \cdot \log_{10} \left( \frac{\max\limits_{\theta} |\mathrm{HRTF}(f, \theta)|}{\min\limits_{\theta} |\mathrm{HRTF}(f, \theta)|} \right) \text{ dB}. \tag{5.3}$$

The directivity patterns of the measured HRTFs often exhibit large spatial dynamic ranges $\zeta$ of up to about 80 dB (cf. Figure 5.5). As discussed above, it is unlikely that such a large dynamic range is really exploited by the human auditory system. Therefore, in the fourth experiment, spatial notches were, frequency bin by frequency bin, reduced in depth such that the resulting spatial dynamic range $\zeta'$ would become lower than that of the original HRTF, cf. section 5.3.3.4. The smaller the resulting spatial dynamic range, the better a VAH configuration with a fixed number of microphones could synthesize the HRTF, but at the same time, the more likely it is for subjects to spot a difference in comparison to the original HRTF. Hence, the discriminability of the artifacts introduced by a reduction of the spatial dynamic range was investigated in this experiment. The resulting threshold value of the spatial dynamic range $\zeta'$ could then be used to preprocess HRTFs prior to the design of appropriate VAH filters. In order to evaluate the effect of the reduction of the spatial dynamic range, we exploited the fact that any modification of the spatial dynamic range will cause modifications in the time and frequency domain for the affected directions. For a chosen direction, one can then evaluate the difference between the modified and the original HRTF. The extent of the difference depends on the individual subject, frequency and direction. Since only a fixed number of directions could be tested, the directions with the biggest subjective differences were identified individually for each subject by a preliminary listening test, see section 5.3.3.4.

In the following, listening tests aimed at evaluating the modifications outlined above are described in more detail.

## 5.3 Methods

All experiments were performed with a fixed set of subjects, stimuli and individual HRTFs. The various experiments only differed in the variables to be investigated as described in sections 5.3.3.1 to 5.3.3.4.

### 5.3.1 Subjects

A total of eight normal hearing subjects (four male, four female, aged between 21 and 46 years) participated in the experiments. Four of the subjects were members of the scientific staff of the Institut für Hörtechnik und Audiologie. They had extensive experience with psychoacoustical experiments and participated voluntarily. Three of them are among the authors of this study. The remaining four subjects were students who were paid for their participation. All subjects completed at least one run of each experiment as familiarization, succeeded by three to six subsequent runs which were used for the evaluation. The performance of all tests lasted approximately five hours for each subject, with each session not lasting longer than 90 minutes.

All experimental procedures were approved by the ethics committee at the Carl-von-Ossietzky-Universität Oldenburg.

### 5.3.2 HRTF measurements

Individual HRTFs were measured in an anechoic room. The subject was seated in the center of a circular loudspeaker array consisting of 24 uniformly distributed loudspeakers (equidistant 15° spacing) at a radius of 1.25 m in the horizontal plane. In order to limit the risk of reflections by the experimental apparatus, small loudspeakers with diameter of 7 cm and a very light support structure were used.

Binaural HRTFs were measured for each direction using the blocked ear method (cf. Hammershøi and Møller (1996)) with custom-made ear shells and Knowles FG-23329 miniature electret microphones. The transfer functions were estimated using white noise signals and standard FFT-based techniques ($H_1$ estimate with 8192-point Hann window, 50% overlap, 52 averages, cf. Mitchell (1982)). The measured transfer functions were subsequently divided by the free field transfer function (cf. equation 1.1) derived from a measurement with the same loudspeakers and a calibration microphone (G.R.A.S. Microphone Type 40AF pointed towards 90° elevation) at the position of the center of the head, in order to obtain the (free field related) HRTFs. The latter step was also important to ensure that small differences between the loudspeakers frequency responses at different positions were compensated for. The corresponding *hrirs* were truncated to 512 samples (corresponding to about 11.6 ms at a sampling frequency of $f_s = 44.1$ kHz). Furthermore, the tails of the *hrirs* were flanked with a one-sided tapered Hann window with a descending flank of 50 samples ($\approx$ 1.1 ms). The delay for sound propagation between the loudspeaker and the head was removed so that the obtained impulse responses had minimal initial delays. Immediately after the HRTF measurements, the headphone transfer functions (HPTF) were measured with the ear shells remaining in place. The HPTF measurements were repeated up to ten times per subject (the headphone was taken off and on before each repetition) until a dynamic range of less than 25 dB for frequencies 2 kHz $\leq f \leq$ 16 kHz was obtained. The inverted version of this particular HPTF was then used as the individual HPTF equalization filter, implemented as finite impulse response (FIR) filters with a length of 512 samples ($\approx$ 11.6 ms).

As a first check, white noise stimuli filtered with the individual HRTFs from 24 directions in the horizontal plane were played dichotically via headphones after individual headphone equalization to each subject. All subjects were able to perceive the presented stimuli outside the head and correctly assigned the corresponding direction.

In order to limit the number of experiments to a manageable amount, four directions were chosen for experiments 1-3, with azimuth angles: $\theta = 0°$ (front), $\theta = 90°$ (left), $\theta = 225°$ (back right) and $\theta = 315°$ (front right). These directions were chosen to cover the main variability of HRTFs, including very small ($\theta = 0°$) and rather large ($\theta = 90°$) interaural differences and the varying monaural cues for the lateral front ($\theta = 315°$) and back ($\theta = 225°$). The directions for the 4th experiment (investigating the spatial dynamic range directly) were selected individually, cf. section 5.3.3.4.

### 5.3.3 Procedure and stimuli

All experiments were designed using psylab[2], a set of MATLAB–scripts for designing various psychoacoustical detection, discrimination and matching experiments. The signals were presented binaurally via a D/A-converter (ADI-8 DS, RME Audio) and headphones (K-240 Studio, AKG Acoustics) at an overall sound pressure level of 78 dB SPL, calibrated with the binaural filters for the frontal direction $\theta = 0°$ using an artificial ear (type 43AA, G.R.A.S. Sound & Vibration). During the experiments, the subjects were seated in the same anechoic room that was used for the HRTF measurement.

In order to cover a wide frequency range and to include temporal cues, the digitally generated test signal consisted of short bursts of frozen white noise with a spectral content of 150 Hz $< f <$ 18050 Hz. The band limitation was achieved by a multiplication with a tapered Hann window (between 150 and 200 Hz and 18000 and 18050 Hz) in the frequency domain (cf. lower graph in Figure 5.3). This window was used to avoid pitch cues due to sharp edges in the frequency domain. The test signal consisted of three segments, each with a noise burst of 0.15 seconds with 0.001 seconds onset-offset ramps followed by silence of 0.15 seconds, leading to a total length of the test signal within each interval of 0.75 seconds (cf. upper graph in Figure 5.3).

A 3-AFC (three alternative forced choice) paradigm was applied to determine threshold values for the tested variables. Three intervals of filtered

---

[2] Psylab is available at `http://www.hoertechnik-audiologie.de/psylab` (date last viewed 29/11/14)

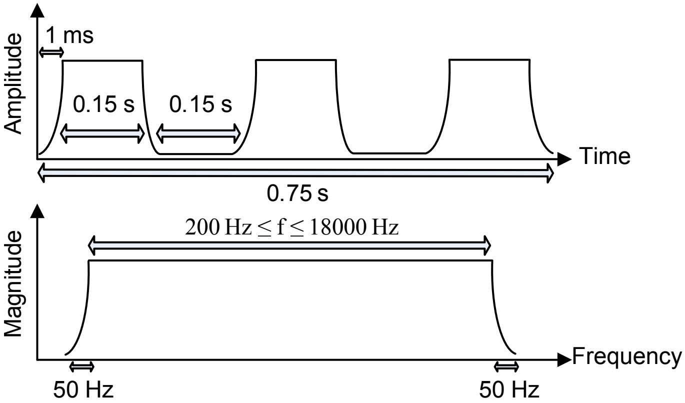


**Figure 5.3:** **Top:** Temporal course of the used test signal consisting of three noise bursts with a total length of 0.75 seconds. **Bottom:** Spectral shape of the test signal with a constant magnitude for frequencies 200 Hz $\leq f \leq$ 18000 Hz.

signals (each separated by 0.3 seconds silence) were presented to the subjects in a randomized order, where one interval contained the test signal filtered with the modified HRTFs and two intervals contained the test signal filtered with the original HRTFs (reference). Subjects were instructed to indicate the odd one of the three intervals. Feedback was presented after each trial. The modification of the variable parameter was adjusted adaptively according to the 1 up - 1 down method, converging to a 50%-correct value on the psychometric function. This particular value on the psychometric function was chosen to represent the JND in all experiments. According to signal detection theory this threshold corresponds to a $d'$ of 0.58 (cf. Hacker and Ratcliff (1979)). The initial value, as well as the initial step size of the tested variable, were chosen for each experiment separately. In the familiarization phase of each experiment the step size of the tested variable was continuously halved at each upper reversal until a minimal step size was reached. As soon as the minimal step size was reached, the measurement phase started and the step size was kept constant. The threshold of the tested variable was defined as the median of six following reversals within the measurement phase. The specific details for each experiment are given in the four subsequent sections 5.3.3.1-5.3.3.4.

#### 5.3.3.1 Substitution of the individual HRTF phase by a mixture of the original phase and a linear phase

In this experiment, the original HRTF phase was substituted by $\phi_{\text{test}}(f)$, a mixture of the original phase $\phi_{\text{orig}}(f)$ and a linear phase $\phi_{\text{lin}}(f)$, cf. equation 5.1. The slope of the linear phase $\phi_{\text{lin}}(f)$ was computed by determining the delay of the maximum of the Hilbert envelope of the corresponding *hrir* in the time domain (cf. Appendix A). The variable mixing gain $L_\phi$ ranged between $L_\phi = 0$ (original phase only) and $L_\phi = 1$ (linear phase only) and was independent of frequency.

The initial supra threshold value of the mixing gain was set to $L_\phi = 0.5$ and the initial step size to $\Delta L_\phi = 0.2$. The minimal step size was set to $\Delta L_\phi = 0.05$.

The length of the impulse response was kept constant throughout all experiments. No effort was made to compensate for acausalities or other artifacts associated with phase linearization.

#### 5.3.3.2 Substitution of the individual HRTF phase by a linear phase above a certain cutoff-frequency

In this experiment, the original phase spectrum was preserved for frequencies $f \leq f_{\text{c}}$, while the phase for frequencies $f > f_{\text{c}}$ was substituted by a linear phase (cf. equation 5.2). The linear phase $\phi_{lin}$ (constant phase slope) was applied starting at the frequency bin next to $f_{\text{c}}$ (cf. Figure 5.4), without any fading between $\phi_{orig}$ and $\phi_{lin}$ as for instance in Rasumow *et al.* (2012b). Again, the slope of the linear phase was computed by determining the delay of the maximum of the Hilbert envelope of the corresponding *hrir* (cf. Appendix A).

The initial supra threshold value of the cutoff-frequency was set to $f_{\text{c}} = 300$ Hz and the initial step size to $\Delta f_{\text{c}} = 160$ Hz. The minimal step size was set to $\Delta f_{\text{c}} = 20$ Hz.

#### 5.3.3.3 Complex-valued smoothing of the individual HRTFs into constant relative bandwidths

In this experiment, the individual HRTFs were smoothed complexly in the frequency domain using the smoothing algorithm proposed by Hatziantoniou and Mourjopoulos (2000). This smoothing algorithm achieves complex-valued smoothing into constant relative bandwidths by replacing each fre-

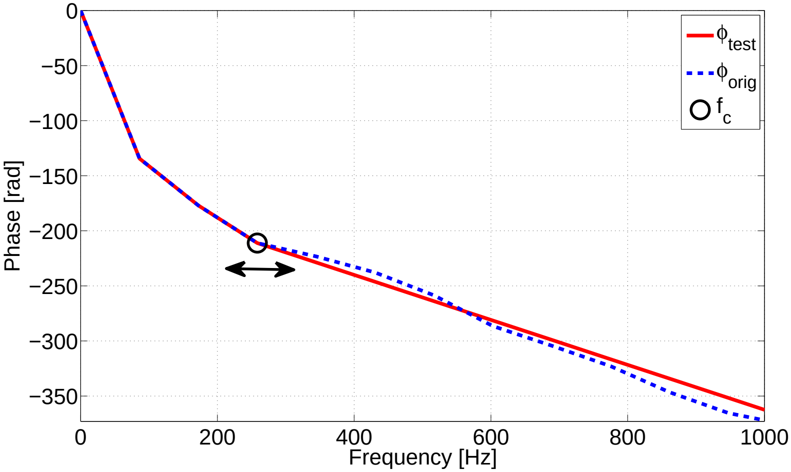


**Figure 5.4:** Substitution of the original HRTF phase ($\phi_{\mathrm{orig}}$) by a linear phase ($\phi_{\mathrm{lin}}$) above a variable cutoff-frequency $f_{\mathrm{c}}$ (black circle): The resulting test phase $\phi_{\mathrm{test}}$ is equal to $\phi_{\mathrm{orig}}$ for $f \leq f_{\mathrm{c}}$, and equal to $\phi_{\mathrm{lin}}$ for $f > f_{\mathrm{c}}$. The slope of $\phi_{\mathrm{lin}}$ is computed from the maximum of the Hilbert envelope of the original *hrir*.

quency bin by a complex average of adjoining frequency bins within the respective bandwidth. Smoothing into constant relative bandwidths is equivalent to a truncation operation in the time domain with a frequency-dependent truncation length. To ensure that the smoothing algorithm processes the main parts of the *hrirs*, it is therefore important to remove the overall delay from the *hrir* before processing. Thus, the estimated overall delay of each *hrir* was removed before smoothing and reconstructed afterwards (cf. Appendix B and Figure B.1).

As discussed above, complex-valued smoothing of HRTFs needs a prior treatment of phase spectra, which we chose to implement as a substitution of the original phase by a linear one above a certain cut-off frequency $f_{\mathrm{c}}$, while maintaining the original phase below $f_{\mathrm{c}}$. For practical reasons, the listening tests for the third experiment started before the second experiment (in which a threshold value for $f_{\mathrm{c}}$ was determined) had been completed. Therefore, a very conservative value of $f_{\mathrm{c}} = 5$ kHz was chosen here, which proved to be much higher than the threshold value eventually determined by the second experiment, while at the same time being low enough to avoid difficulties related to steep and/or noisy phase spectra.

The initial supra threshold value of the relative bandwidth was set to $\mathrm{B_W} = 2/3$ octave (approximately corresponding to the bandwidth of two auditory filters) and the initial step size to $\varDelta \mathrm{B_W} = 1/3$ octave. The minimal step size was set to $\varDelta \mathrm{B_W} = 1/24$ octave.

### 5.3.3.4 Limiting the spatial dynamic range of individual HRTFs

In this experiment, the dynamic range $\zeta$ of the individual HRTFs over all directions of incidence $\theta$ (i.e., in the spatial domain) was limited in each frequency bin. As discussed in sec. 5.2, the spatial dynamic range $\zeta$ is defined as the dB value of the ratio between the largest and the smallest magnitude of a directivity pattern per frequency $f$ (cf. equation 5.3). $\zeta$ will depend on the individual head geometry and the measurement accuracy. For the subjects and the measurement accuracy used in this study ($\mathrm{NFFT} = 512$, $f_s = 44.1$ kHz and 15° resolution in the horizontal plane), the maximum observed spatial dynamic range approached $\zeta_{max} \approx 80$ dB at higher frequencies (cf. Figure 5.5).

In order to limit the spatial dynamic range of the individual HRTFs in every of the 512 frequency bins, low levels in the directivity pattern (i.e. spatial notches) were boosted such that they were not less than $\zeta'$ lower compared to the direction with the highest level. The phase spectra of the manipulated directions were left unchanged. This procedure was applied

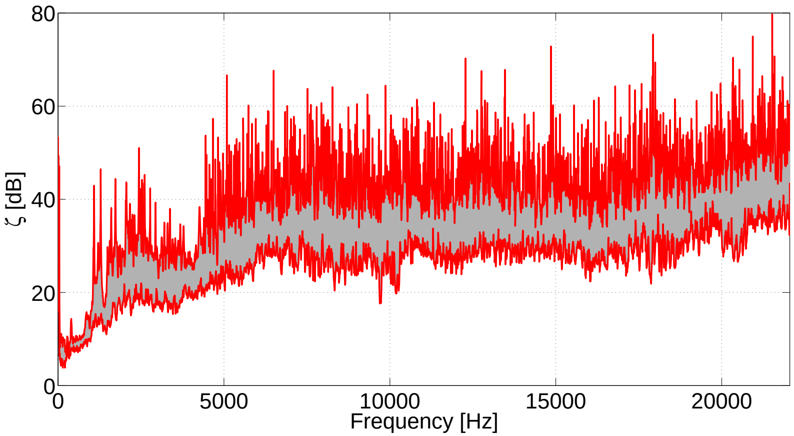


**Figure 5.5:** Spatial dynamic range $\zeta$ of all left ear HRTFs (24 directions in the horizontal plane, all eight subjects), as a function of frequency. The minimum and maximum values of all subjects are shown as red lines, the range in between (characterizing the remaining dynamic ranges) as a gray area. Similar $\zeta$-values also result for HRTFs of the right ear (not shown).

to the directivity pattern in each frequency bin separately, but with the same $\zeta'$.

As discussed above, the impact of limiting $\zeta$ was tested by comparing the altered to the original HRTF at a fixed set of four directions. The four directions were chosen individually such that for each subject the directions with the largest subjective change between original and altered HRTF were included. To this end, the subjects underwent a preliminary listening test in which they had to select those four directions in the horizontal plane where they perceived a reduction of their spatial dynamic range to $\zeta'_{pre} = 25$ dB most saliently. These directions (which differed from subject to subject) were then used in the subsequent experiment.

In the actual experiment, the initial supra threshold spatial dynamic range was set to $\zeta' = 20$ dB, with an initial step size of $\Delta\zeta' = 3$ dB. The minimal step size was set to $\Delta\zeta' = 1$ dB.

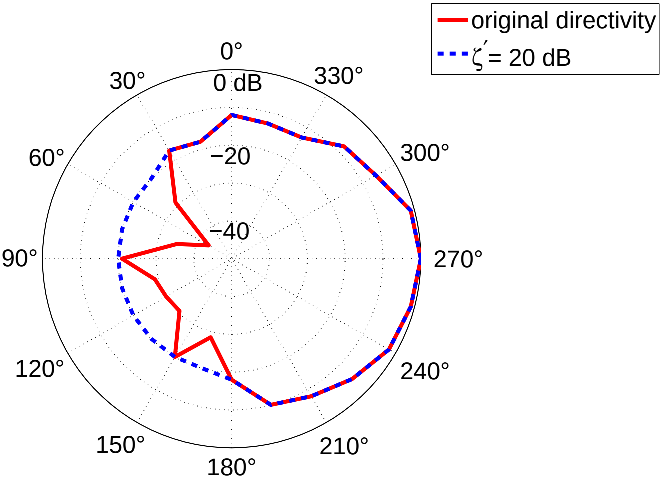


**Figure 5.6:** Exemplary reduction of the spatial dynamic range of a normalized right ear HRTF magnitude in the horizontal plane from $\zeta \approx 43$ dB (original) to $\zeta' = 20$ dB (test situation).

## 5.4 Results and discussion

In the following, the results from eight subjects are shown as means and standard deviations of three to six runs of each experiment.

### 5.4.1 Substitution of the individual HRTF phase by a mixture of the original phase and a linear phase

Means and standard deviations of the the mixing gain $L_\phi$ at threshold (over three to six runs of the experiment), as a function of subjects and direction of incidence $\theta$, are shown in Figure 5.7.

It can be observed that the mean values of $L_\phi$ at threshold seem to vary more or less unpredictably with subject and direction of incidence. Many subjects (excluding subjects S1 and S8) perceived the modification of the individual HRTF phases for the frontal direction ($\theta = 0°$) at high $L_\phi$ (high proportion of the linear phase) only. In contrast, the modification of the HRTF phase for $\theta = 225°$ (green diamonds) was already discriminable at small $L_\phi$ for most subjects. However, in both cases there were subjects who did not follow this trend.

None of the tested subjects showed a complete inability (i.e. $L_\phi = 1$) to discriminate the applied phase modification for the tested directions. Yet, some subjects indicated large $L_\phi$ and hence a pronounced insensitivity to phase modifications for particular directions (e.g. subject S6 at 0° and subject S7 at 0° and 90°).

Can the seemingly inconsistent results be explained by the fact that phase modifications yielded individual (binaural) spectra that were objectively different from the original ones, despite the fact that the same kind of (monaural) modification was applied? In an attempt to do so, we hypothesize that a large alteration of the interaural phase difference (IPD) from the original HRTF pair to the completely manipulated (i.e., 100% linear) HRTF pair corresponds to a rather salient discrimination. This in turn should correspond to small $L_\phi$, and vice versa. In order to quantify the IPD alteration, we introduce a model for the individual discrimination ability $L_{\mathrm{mod}}$ as

$$\begin{aligned} L_{\mathrm{mod}} &= \nu_{mod} \cdot \sum_{\mathrm{f}=150\ \mathrm{Hz}}^{1500\ \mathrm{Hz}} \frac{\Delta\mathrm{f}}{\mathrm{ERB(f)}} \cdot \frac{|\mathrm{IPD_{orig}(f)} - \mathrm{IPD_{lin}(f)}|}{\phi_{\mathrm{JND}}(\mathrm{f})} \\ \text{with} &\sum_{\mathrm{f}=150\ \mathrm{Hz}}^{1500\ \mathrm{Hz}} \left(\frac{\Delta\mathrm{f}}{\mathrm{ERB(f)}}\right) \cdot \nu_{mod} = 1 \end{aligned} \tag{5.4}$$

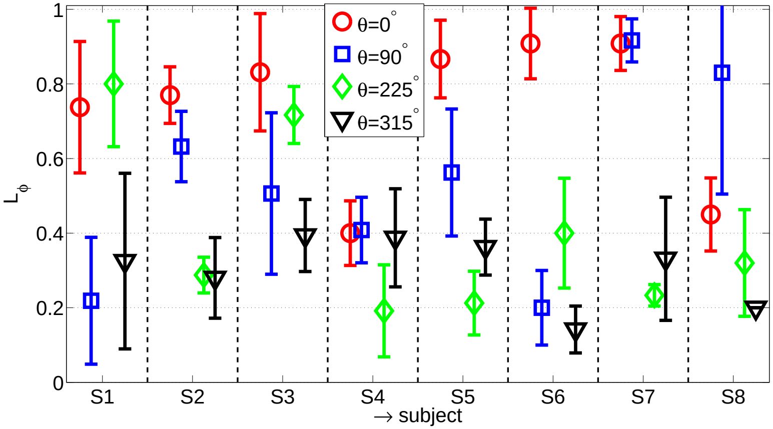


**Figure 5.7:** Threshold data for experiment 1: means and standard deviations over three to six runs of the the mixing gain $L_\phi$ at threshold, as a function of subjects and direction of incidence $\theta$. Smaller values of $L_\phi$, i.e. a smaller proportion of the linear phase, indicate a higher sensitivity to the broadband modification of the individual HRTF phase and vice versa.

where $\nu_{mod}$ is a normalization constant, $\varDelta \mathrm{f} = \frac{f_s}{\mathrm{NFFT}}$ is the frequency resolution of the FFT, given by the sampling frequency $f_s$ and the number of frequency bins NFFT, IPD(f) is the frequency-dependent interaural phase difference at a given frequency, ERB(f) is the frequency-dependent equivalent rectangular bandwidth according to Glasberg and Moore (1990), and $\phi_{\mathrm{JND}}(\mathrm{f})$ is the interaural time difference threshold (taken from Klumpp and Eady (1956)). Note that in the summation of equation 5.4 the difference between the original and the linear phase is weighted with the reciprocal value of the interaural time difference thresholds to account for the frequency-dependence of $\phi_{\mathrm{JND}}(\mathrm{f})$. In addition, the division by ERB(f) ensures that each auditory filter is equally weighted in the summation. Moreover, we assumed that the ability to discriminate phase modifications is dominated by lower frequencies. Therefore, we chose to take into account frequencies between 150 Hz $\leq f \leq$ 1.5 kHz, because interaural phase differences in the stimulus fine-structure are processed by the auditory system up to about 1.5 kHz only (cf. Perrott and Nelson (1969)). The lower limit is determined by the stimulus frequency content (cf. section 5.3.3).

The relationship between the modeled individual discrimination ability $L_{\mathrm{mod}}$ and the measured $L_\phi$ (mean values) is shown in Figure 5.8. As can be seen, high values of the individual discrimination ability correspond to low $L_\phi$ (i.e. to a high sensitivity to phase changes). The linear correlation

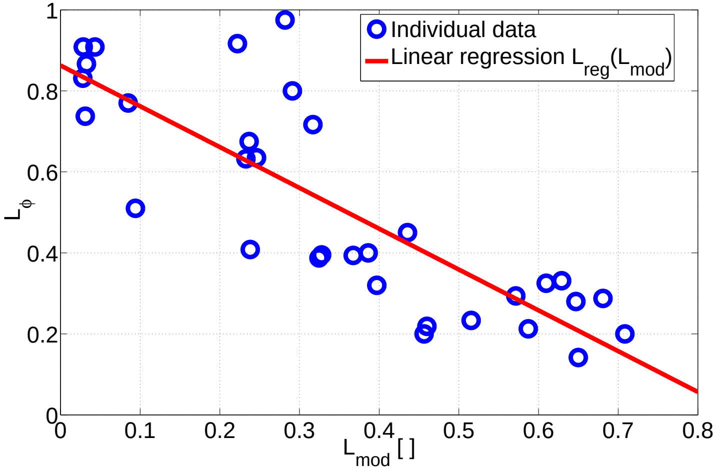


**Figure 5.8:** Experimentally observed mean values of $L_\phi$, as a function of the individual discrimination ability $L_{\mathrm{mod}}$ (equation 5.4). This relation can be characterized by a linear regression $L_{\mathrm{reg}}(L_{\mathrm{mod}}) = -1.0077 \cdot\ L_{\mathrm{mod}} + 0.8627$ (red line), with a linear correlation coefficient of $\rho(L_{\mathrm{mod}}, L_\phi) = -0.83$.

coefficient ($\rho(L_{\mathrm{mod}}, L_\phi) = -0.83$) is rather high for this type of experiment, indicating that $L_{\mathrm{mod}}$ is quite well suited to predict the individual sensitivity to broadband HRTF phase linearizations. On a side note, the observed correlation between $L_{\mathrm{mod}}$ and $L_\phi$ decreases when the upper frequency range of $L_{\mathrm{mod}}$ is extended to higher frequencies (not shown here). This fact again emphasizes the perceptual importance of the individual IPD at lower frequencies, which is well in line with the literature (cf. Kulkarni *et al.* (1999)). Furthermore, it countenances the approach of the second experiment by substituting the phase only at higher frequencies while maintaining the measured phase at lower frequencies.

### 5.4.2 Substitution of the individual HRTF phase by a linear phase above a certain cutoff-frequency

The mean cutoff-frequencies $f_c$ at threshold and corresponding standard deviations are shown in Figure 5.9. Both mean values and standard deviations vary with subject and direction of incidence. The highest $f_c$, i.e. the most sensitive thresholds, are associated with different directions for different subjects.

There is, however, a trend towards larger standard deviations for higher

$f_c$. The most plausible explanation would be that for high $f_c$, the artifacts introduced by the applied phase modification will become so hard to discern that the subjects will have difficulty to judge consistently.

Relating these results to those of the first experiment (section 5.4.1), one would expect that the conditions associated with a poor sensitivity to broadband phase modification (large $L_\phi$ in Figure 5.7) correspond to small cutoff-frequencies $f_c$ in the second experiment. In other words, if a subject is completely unable to discriminate a broadband phase modification, this should also result in a (very) low cutoff-frequency $f_c$. On the other hand, a high $f_c$ is not necessarily related to a high sensitivity to a broadband phase linearization (i.e. a low $L_\phi$ in experiment 1) because for higher $f_c$ the two experiments refer to different phase manipulations.

In fact, when taking a closer look at Figures 5.7 and 5.9 it can be seen that the lowest $f_c$ values (reaching down to $f_c \approx 0$ Hz) for subjects S3 (0°), S5 (0°), S6 (0°) and S7 (0° and 90°) indeed correspond to $L_\phi$ values close to one. In addition, the large $L_\phi$ values, e.g. for S1 (225°) and S8 (90°), imply a lack of discriminability of broadband phase modifications. For these conditions, the according cutoff-frequencies reach down to $f_c = 120$ Hz and $f_c = 205$ Hz, respectively. Considering that the test signal had a steep roll-off for frequencies $f < 200$ Hz, these cutoff-frequencies can also be regarded

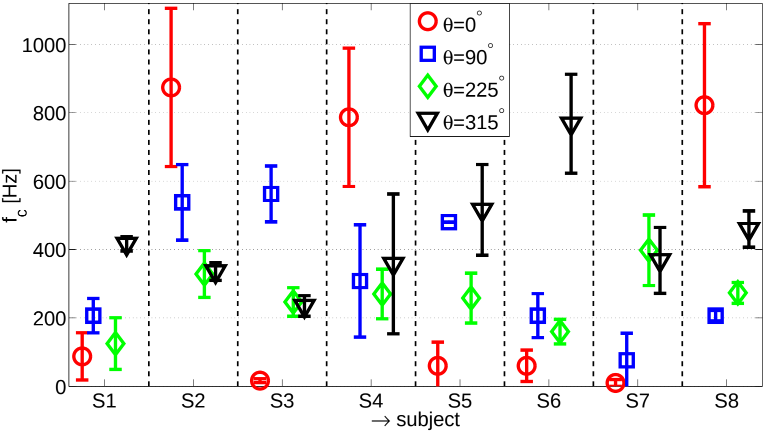


**Figure 5.9:** Threshold data for experiment 2: means and standard deviations over three to six runs of the cutoff-frequency $f_c$ at threshold between the original (for $f \leq f_c$) and the linear phase (for $f > f_c$). Lower $f_c$ indicate that a larger proportion of the original phase can be substituted with the linear phase still yielding a discrimination at threshold and vice versa.

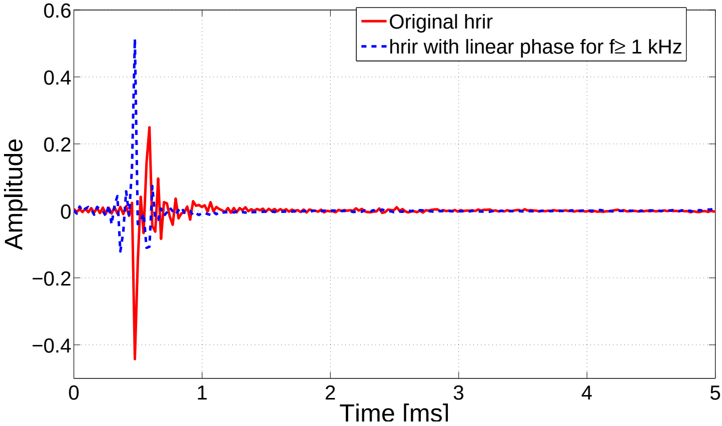


**Figure 5.10:** Exemplary course of a measured *hrir* (red solid line) and a manipulated *hrir* (blue dashed line) for the left ear of subject S8 and $\theta = 105°$ as a function of time. The phase of the latter *hrir* is linearized for $f \geq 1$ kHz, which results in a rather symmetrical envelope of the corresponding *hrir* in the time domain. Note that the delay of the peak magnitude remains unchanged.

as "close to zero Hz". For all other conditions, the relation between $L_\phi$ and $f_c$ does not follow a simple pattern.

In conclusion, the experimental results show that for $f >$ 1000 Hz, the HRTF phase of the tested population could be substituted by a suitable linear phase without causing any discriminable artifacts for any of the eight subjects. In order to give a first impression of the consequences of the proposed phase manipulation for the resulting *hrirs*, two exemplary *hrirs* are plotted in Figure 5.10. The red solid graph shows a measured *hrir* and the blue dashed graph shows this *hrir* when its phase is linearized for $f \geq 1$ kHz. It can be seen that the phase linearization alters the envelope of the impulse response, yielding a more symmetrical counterpart. This effect was observed for all *hrirs*. It is worth noting that the peak of the magnitude of the *hrirs* remains at a fixed delay which is encoded in the slope of the linear phase.

### 5.4.3 Complex-valued smoothing of the individual HRTFs into constant relative bandwidths

The mean relative bandwidths $\mathrm{B_W}$ at threshold and corresponding standard deviations when smoothing individual (complex-valued) HRTFs in the frequency domain (with the original phase substituted by a linear phase for $f > 5$ kHz) are shown as fractions of one octave in Figure 5.11. Again, both means and standard deviations seem to vary with subject and direction of incidence in an unpredictable manner. This is not surprising, since HRTFs have individual spectral shapes, and smoothing into relative bandwidths will therefore yield individually varying artifacts.

The lowest relative bandwidths at threshold (most sensitive cases) were at $\mathrm{B_W} \approx 1/5$ octave, i.e. between a major second and a minor third. The highest bandwidth at threshold (least sensitive case) was at $\mathrm{B_W} \geq 1$ octave, this occurred however just for one subject at one direction of incidence (subject S3, 225°). The vast majority of all conditions resulted in bandwidths at threshold between one and two third octaves. These bandwidths approximately correspond to 2 to 4 ERBs at high frequencies (cf. Glasberg and Moore (1990)).

In conclusion the results indicate that the HRTFs of the tested population may be smoothed complexly with a bandwidth of $\mathrm{B_W} \approx 1/5$ octave (corresponding to roughly 1 ERB at high frequencies) without causing a perceptual difference to the original HRTFs. These results also confirm the validity of the proposed phase modifications in section 5.4.2 ($f_c = 5$ kHz), since if the phase linearization was discriminable the subjects

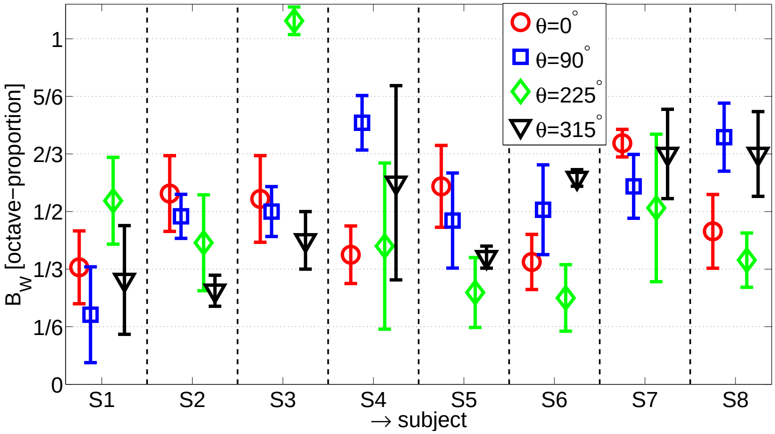


**Figure 5.11:** Threshold data for experiment 3: means and standard deviations over three to six runs of the relative bandwidth $\mathrm{B_W}$ at threshold expressed as fractions of one octave when complexly smoothing HRTFs in the frequency domain. Prior to smoothing, the HRTF phase (for $f > 5$ kHz) was substituted with the linear phase $\phi_{lin}$. Higher $\mathrm{B_W}$ indicate more smoothing at threshold and thus less sensitive thresholds and vice versa.

would have indicated a subjective difference in any case (independent of the bandwidth), which would have resulted in bandwidths at threshold of $\mathrm{B_W} \approx 0$.

### 5.4.4 Limiting the spatial dynamic range of the individual HRTFs

The spatial dynamic ranges $\zeta'$ (means and standard deviations) at threshold are shown in Figure 5.12.

First of all, the directions chosen by the subjects were not equally distributed in the horizontal plane. Instead, they were all in the ranges 90° ± 45° or 270° ± 45°, i.e. at lateral directions. This can be explained by the fact that lateral directions exhibit more spectral notches, which results in larger modifications when the spatial dynamic range is limited.

The experimentally determined spatial dynamic ranges at threshold vary between 19 dB $\leq \zeta' \leq$ 29 dB. The standard deviation varies approximately between 0.1 dB$\leq \sigma(\zeta') \leq$ 3 dB. Again, the variation of individual $\zeta'$ is hypothesized to be due to spectral characteristics of the HRTFs which are

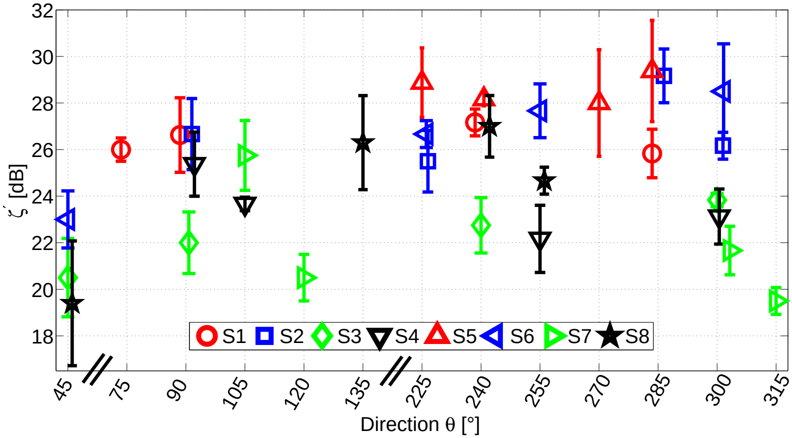


**Figure 5.12:** Threshold data for experiment 4: means and standard deviations over three to six runs of the spatial dynamic range $\zeta'$ as a function of azimuthal direction $\theta$. Note the non-equidistant spacing. Here the individual subjects are depicted as various symbols. Smaller spatial dynamic ranges $\zeta'$ indicate larger manipulations of the directivity patterns and thus less sensitive thresholds and vice versa.

individual for each subject.

As a general conclusion, the spatial dynamic range of the HRTFs in the horizontal plane of the tested population can be reduced to $\zeta' > 29$ dB without leading to a discriminable alteration compared to the original HRTFs.

## 5.5 Discussion

Four different methods for smoothing HRTFs were investigated. The first two addressed phase smoothing which is a preprocessing step required to do complex-valued spectral smoothing. The third method was a complex-valued spectral smoothing method, and the last method was a spatial smoothing method.

### 5.5.1 Perceptual effects of phase linearization

We found that the binaural discriminability of a broadband phase linearization of individual HRTFs (while preserving the delay) largely varies with subject and direction. This agrees well with the investigations by Kulkarni *et al.* (1999). We assumed the individual characteristic of the observed thresholds to be due to the differences in the individual phase spectra that are modified by the broadband linearization. In fact, although the sensitivity to monaural phase differences is rather poor (cf. Kulkarni *et al.* (1999)), the sensitivity to interaural phase differences is high for low frequencies (cf. Klumpp and Eady (1956) and Yost (1974)). We assumed that the discriminability is proportional to the individual IPD alteration at these frequencies. Using a simple model representing the individual discrimination ability (cf. equation 5.4), it was possible to approximately quantify and predict the majority of the individual thresholds (cf. Figure 5.8). More clearly than in Kulkarni *et al.* (1999), the current results indicate a pronounced (individual) sensitivity to a broadband phase linearization for some subjects/directions (cf. Figure 5.7, $\theta = 315°$). Possible explanations for this difference may be the differing HRTF directions ($\theta = 0°, 90°, -90°, 180°$ in Kulkarni *et al.* (1999)) and the higher number of subjects in the current study (eight over four in Kulkarni *et al.* (1999)), which increases the chance to spot individually higher sensitivities. On the basis of these findings, a broadband linearization seems to be an inappropriate preprocessing operation for smoothing complex-valued HRTFs in the frequency domain.

Based on the previous arguments, we assume the phase linearization to be less discriminable when the monaural phase and hence the IPD is preserved for lower frequencies. The observed thresholds (cf. Figure 5.9) indicate that a phase linearization is not discriminated when the original phase is preserved for $f < 1000$ Hz. This cutoff-frequency complies well with literature on interaural phase processing (phase locking) in the auditory system, which is limited up to about 1.5 kHz (cf. Perrott and Nelson (1969)). Moreover, these findings are in line with Kulkarni *et al.* (1999) where no significant phase discrimination was reported with high-pass stimuli which only contained frequencies $f > 2$ kHz.

Interestingly, even the highest cutoff-frequencies at threshold are lower than those observed in a previous study (cf. Rasumow *et al.* (2012b)). This difference turned out to be due to a detail of the phase modification: In the current study, the original phase $\phi_{orig}$ was substituted by a linear phase $\phi_{lin}$ above $f_c$ *without* any fading between the two, whereas in the previous study, the transition between original and linear phase was accompanied by a spectral windowing function extending over 5 frequency bins. This result indicates that a larger proportion of the measured phase (lower $f_c$) can be substituted when *no* fading between the original phase and the linear phase is applied. Hence, we assume the fading between the two phases in Rasumow *et al.* (2012b) to have created cues that are not present when no fading is used.

### 5.5.2 Perceptual effects of smoothing HRTFs in the frequency domain

In the current study, the ability to discriminate complexly smoothed HRTFs with variable relative bandwidths was investigated after applying a linear phase for $f > 5$ kHz (cf. Figure 5.2). In general, the obtained bandwidths at threshold approximately corresponded to one to two third octaves (approximately 2 to 4 ERBs at high frequencies). These results are in good agreement with general investigations on frequency grouping (cf. Patterson and Nimmo-Smith (1980) and Glasberg and Moore (1990)) and comparable with similar investigations (cf. Xie and Zhang (2010)), where the magnitude of HRTFs could be smoothed with bandwidths of 1 to 3.5 ERB for higher frequencies ($f > 5000$ Hz). Furthermore, these results also compare well with findings from Kohlrausch and Breebaart (2001) where unprocessed HRTFs could not be discriminated from those smoothed using a gammatone filter bank (with its bandwidth approximately corresponding to the bandwidth of one ERB). Interestingly, those studies used a separate smoothing of magnitude and phase, whereas in the current study the HRTFs were complexly smoothed after applying a phase linearization above 5 kHz. It thus seems that both approaches are equally well suited to spectrally smooth HRTFs.

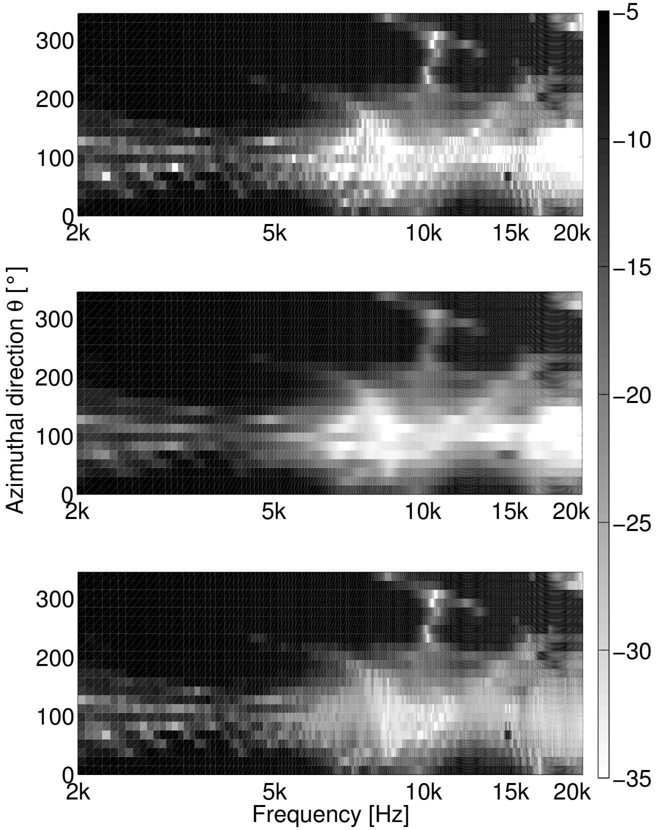


**Figure 5.13:** Effects of the proposed smoothing methods, **top**: Exemplary HRTF (subject S2, right ear) as function of frequency in kHz on the x-axis and as a function of azimuthal direction $\theta$ on the y-axis, **middle:** the same HRTF, processed through a complex-valued smoothing as described in section 5.3.3.3 with a bandwidth of $B_W = \frac{1}{5}$ octave, **bottom:** the same HRTF, with the spatial dynamic range limited to $\zeta' = 29$ dB as described in section 5.3.3.4. The latter manipulation is primarily apparent for contralateral directions ($0° \leq \theta \leq 180°$) where the spatial notches (light areas) are adjusted. In this illustration, levels were limited between -5 and -35 dB to highlight the differences of the particular smoothing methods.

In the most sensitive condition, the relative bandwidth at threshold was at about a minor third, indicating that the individual HRTFs of the tested population could safely be smoothed into constant relative bandwidths of $\mathrm{B_W} = \frac{1}{5}$ octave for any of the eight subjects.

### 5.5.3 Influence of the test signal

In the current study we used broadband noise bursts, which on the one hand have a broadband spectrum and on the other hand a temporal structure. Such signals have successfully been used in a number of discrimination experiments on the effect of HRTF manipulations (cf. Kulkarni *et al.* (1999)). Also, in a study on the effect of various headphone equalizations, noise signals have been shown to give more sensitive thresholds than music stimuli (cf. Lindau (2012)). Since headphone equalization is comparable to spectral HRTF smoothing, these results should apply to the third experiment in a similar manner. Compared to continuous noise stimuli, we assume the discrimination to be still more sensitive when noise bursts are used, which is particularly important for the discrimination of phase manipulations.

In the fourth experiment, we modified HRTFs frequency bin by frequency bin. A test signal which would exhibit dominant components in the particular frequency bin only could then possibly give more salient discrimination cues. Such test signals are, however, extremely artificial and it is hard to imagine a scenario in which they will be relevant for real-world applications of the VAH.

### 5.5.4 Impact of the proposed smoothing methods on the HRTFs

To illustrate the functioning and to give a first impression regarding the two proposed smoothing methods, an exemplary HRTF (subject S2, right ear, top box in Figure 5.13) was complexly smoothed into $\mathrm{B_W} = \frac{1}{5}$ octave bands (second box in Figure 5.13) and the spatial dynamic range of the measured HRTF was reduced to $\zeta' = 29$ dB (third box in Figure 5.13). Comparing the first two boxes, it is apparent that the complex-valued smoothing clearly reduced the spectral resolution of the measured HRTF, especially at higher frequencies ($f \geq 6$ kHz). Furthermore, this example demonstrates that the complex-valued smoothing leads to a smooth magnitude spectrum, also at high frequencies, which would not be the case if no prior phase linearization would have been applied (cf. Figure 5.2). The reduction of the spatial dynamic range in the third box is best visible for contralateral directions ($0° \leq \theta \leq 180°$) where the most spatial notches (light areas) are adjusted per frequency bin. This manipulation is primarily effective for spatial notches which mainly occur for frequencies $f \geq 5000$ Hz.

### 5.5.5 Impact of the proposed smoothing methods on the accuracy of the VAH

The main motivation of this study was to enhance the performance of the VAH by smoothing the desired HRTFs prior to the computation of the VAH filters. In order to illustrate the benefits associated with the complex-valued smoothing of HRTFs in the frequency domain and the reduction of the spatial dynamic range, we computed VAH filters for individual HRTFs using analytical steering vectors (i.e. pure delays, cf. equation 1.7) of 24 microphones positioned according to the topology used in Rasumow *et al.* (2011b) or **Chapter 2**. The filters were calculated by minimizing a least squares cost function linking the synthesized directivity patterns $\mathrm{HRTF}_{\mathrm{synth}}$ to the desired directivity patterns $\mathrm{HRTF}_{\mathrm{des}}$ for each frequency bin separately according to equation (5) in Rasumow *et al.* (2013a), with a desired regularization measure of $\mathrm{WNG}(\mathbf{w}(f)) = 3$ dB (cf. equation (3) in Rasumow *et al.* (2013a)).

First, to illustrate the effect of complexly smoothing HRTFs in the frequency domain an example (subject S2, right ear, $\theta = 210°$) is shown in Figure 5.14. The complex-valued spectral smoothing of the $\mathrm{HRTF}_{\mathrm{des}}$ becomes more and more visible with increasing frequency (cf. blue and red solid lines), which is expected for smoothing into constant relative bandwidths. Furthermore, the synthesis using the VAH clearly changes when the $\mathrm{HRTF}_{\mathrm{des}}$ is smoothed. Especially for frequencies $5 \text{ kHz} \leq f \leq 10 \text{ kHz}$ and $f \geq 16$ kHz the synthesis using the VAH improves considerably when the $\mathrm{HRTF}_{\mathrm{des}}$ is smoothed spectrally prior to the synthesis.

Second, to illustrate the benefits of reducing the spatial dynamic range, an exemplary right-ear $\mathrm{HRTF}_{\mathrm{des}}$ (subject S2, right ear, $f \approx 13.5$ kHz) and its synthesis $\mathrm{HRTF}_{\mathrm{synth}}$ using the VAH are depicted in Figure 5.15 as a function of direction $\theta$. There the spatial notch of the original $\mathrm{HRTF}_{\mathrm{des}}$ at $105°$ is only poorly synthesized by the VAH (red lines, maximum error of $> 20$ dB). The performance of the VAH, however, improves considerably (at $90° \pm 30°$) when the spatial dynamic range of the $\mathrm{HRTF}_{\mathrm{des}}$ is reduced to $\zeta' = 29$ dB (blue lines, maximum error of $\approx$ 5 dB).

In order to quantify the synthesis accuracy for each set of HRTFs (one ear,

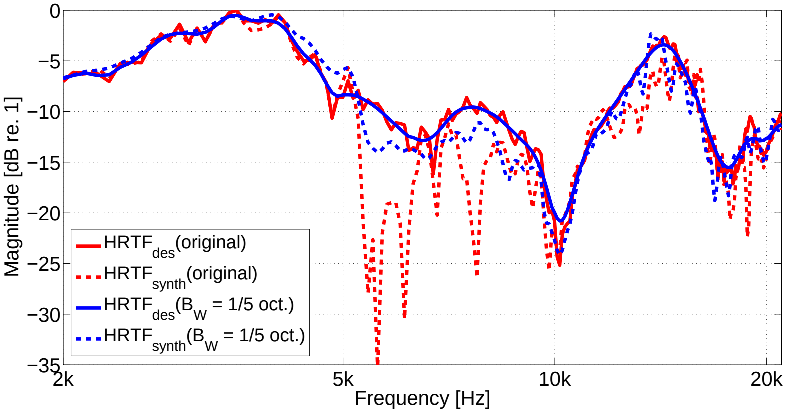


**Figure 5.14:** Desired $\mathrm{HRTF}_{\mathrm{des}}$ (solid lines) and synthesized $\mathrm{HRTF}_{\mathrm{synth}}$ using the VAH (dashed lines) are depicted for the original (red) and for the spectrally smoothed case (blue) with $\mathrm{B_W} = \frac{1}{5}$ for an exemplary HRTF (subject S2, right ear, $\theta = 210°$), as a function of frequency.

P directions of incidence), we used the absolute VAH-error $\epsilon_f$,

$$\epsilon_f(f_c) \quad = \quad \frac{1}{\mathrm{P}} \sum_{i=1}^{\mathrm{P}} \epsilon(f_c, \theta_i) \ , \tag{5.5}$$

where $\epsilon(f_c, \theta_i)$ denotes the synthesis error between the resulting directivity pattern $\mathrm{HRTF}_{\mathrm{synth}}$ and the desired directivity pattern $\mathrm{HRTF}_{\mathrm{des}}$ (for each direction $\theta_i$ and center frequency $f_c$ of the associated ERB-band) averaged over frequencies in ERB-bands according to equation 1.14. This error measure is further averaged over all P directions, yielding $\epsilon_f(f_c)$ as a function of the center frequency $f_c$ of the associated ERB-band.

This error is shown in Figure 5.16 as a function of frequency for an exemplary subject (subject S2, right ear, P=24). As can be seen, the synthesis with the VAH works best for lower frequencies up to $f \approx$ 5 kHz. At low frequencies ($f <$ 2 kHz), the VAH-error $\epsilon_f$ is still lower and therefore not shown here. Larger errors primarily occur at higher frequencies. Compared to the synthesis of the original HRTF, the VAH-error decreases especially in the frequency range of 5 kHz $\leq f \leq$ 10 kHz for the spectrally smoothed HRTF. On the other hand, the error is slightly higher around 4 kHz. This can be explained by the fact that a spectral smoothing alters the $\mathrm{HRTF}_{\mathrm{des}}$ for all directions separately and hence may generate directivity patterns which are harder to synthesize than the original HRTFs. Yet, in general

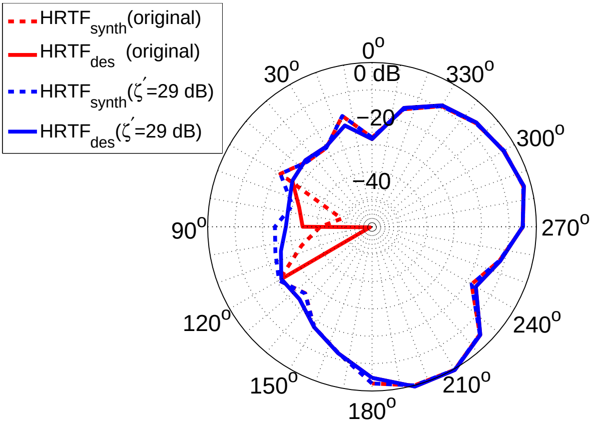


**Figure 5.15:** The $\text{HRTF}_{\text{des}}$ (solid lines) and their synthesis $\text{HRTF}_{\text{synth}}$ using the VAH (dashed lines) are depicted for the original $\text{HRTF}_{\text{des}}$ (red) and for the $\text{HRTF}_{\text{des}}$ with reduced spatial dynamic range ($\zeta' = 29$ dB, blue lines) for an exemplary HRTF (subject S2, right ear, $f \approx 13.5$ kHz).

the VAH-error clearly decreases when the $\text{HRTF}_{\text{des}}$ is smoothed in the frequency domain.

The VAH-error $\epsilon_f$ for the synthesis of the $\text{HRTF}_{\text{des}}$ with a spatially limited dynamic range shows improvements over the synthesis of the original $\text{HRTF}_{\text{des}}$ for frequencies $5\,\text{kHz} \leq f \leq 9\,\text{kHz}$ and $f \geq 10\,\text{kHz}$. At these frequencies, spatial notches occurred frequently and were thus modified by the reduction of the spatial dynamic range. It is worth noting that the VAH-error for the synthesis of the HRTF with limited spatial dynamic range never exceeds the VAH-error for the original HRTF.

To put these synthesis accuracies into one number, the mean error $\overline{\epsilon_f}$ for all directions $\theta$ and frequencies $f$ can be used. It decreased from 3.8 dB to 3.4 dB when the measured HRTFs were smoothed spectrally and from 3.8 dB to 3.0 dB when the spatial dynamic range was reduced according to the obtained thresholds. Although, the benefits of the VAH-synthesis highly depend on the individual HRTFs, the applied optimization strategy etc.

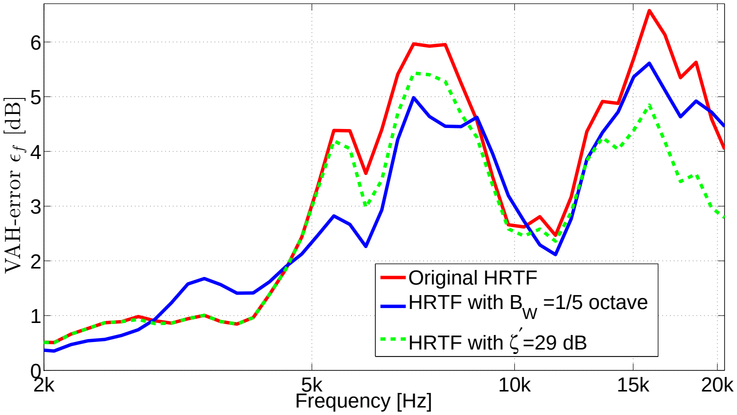


**Figure 5.16:** Effects of the proposed smoothing methods on the accuracy of the VAH. Plotted is the VAH-error $\epsilon_f$ (equation 5.5) of the right ear HRTF of subject S2, as a function of frequency, for the original $\mathrm{HRTF}_{\mathrm{des}}$ (red solid line), the spectrally smoothed $\mathrm{HRTF}_{\mathrm{des}}$ ($\mathrm{B_W} = \frac{1}{5}$ octave, blue solid line) and the $\mathrm{HRTF}_{\mathrm{des}}$ with reduced spatial dynamic range ($\zeta' = 29$ dB, green-dashed line).

## 5.6 Summary and conclusions

In this study we investigated HRTF smoothing in the spectral and spatial domains as a preprocessing step for the virtual artificial head. It was found that:

1. Subjects are sensitive to a broadband phase linearization of HRTFs.
2. The individual sensitivity to a broadband phase linearization can be predicted by a simple model that is based on interaural phase differences at frequencies $f \leq$ 1.5 kHz.
3. The original phase can be substituted by a linear phase above $f >$ 1000 Hz without introducing noticeable artifacts.
4. After substituting the original phase by a linear phase above $f \geq 5$ kHz, HRTFs may be smoothed complexly into constant relative bandwidths of $\mathrm{B_W} \leq 1/5$ octave, without introducing noticeable artifacts.
5. Spatial notches in the directivity pattern do not need to be retained in detail if they are less than 29 dB below the maximum value.

These findings permit to efficiently smooth individual HRTFs in the spectral and spatial domains. It must however be kept in mind that they are limited to HRTFs in the horizontal plane.

In future we will investigate extensions to non-horizontal HRTFs and evaluate the consequences of the proposed preprocessing methods on the synthesis of individual HRTFs using the VAH. In this context the threshold values found in this study may be used as a starting point to optimize the overall performance of the VAH.

# Part III

# Evaluation of individualized binaural reproduction using the virtual artificial head

In part III, individualized binaural synthesis using the VAH and binaural reproductions using traditional artificial heads are perceptually evaluated. The VAH-synthesis is based on the findings described in parts I and II. In more detail, the microphone topology resulting from the findings in **Chapter 2** and the optimization and regularization resulting from **Chapter 3** are used for the VAH-synthesis. Moreover, the desired HRTFs are smoothed according to the limits described in **Chapter 4** and **5** prior to the synthesis using the VAH.

The perceptual evaluation associated with different regularization parameters and two different types of microphones (with different level of sensor noise and the same microphone topology) is presented in **Chapter 6**.

In **Chapter 7**, the suitability of the proposed method to synthesize individualized binaural reproductions is validated using the perceptual evaluation of the VAH-synthesis with respect to free field playback. The evaluations are carried out for explicitly considered and for intermediate directions in the VAH-optimization but also for binaural reproductions obtained with traditional artificial heads, which are currently the state of the art method for non-individualized binaural reproductions.

# 6

# Subjective impact of the mean white noise gain ($\mathrm{WNG_m}$) on the appraisal of binaural sound reproduction*

---

*This chapter is based on the publication

- Rasumow, Blau, Hansen, Doclo, van de Par, Mellert, and Püschel (2014b), "The Impact of the White Noise Gain (WNG) of a Virtual Artificial Head on the Appraisal of Binaural Sound Reproduction," in *Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014, pp. 174-180.*

*As an individualized alternative to traditional artificial heads, individual head-related transfer functions (HRTFs) can be synthesized with a microphone array and digital filtering. This strategy is referred to as "virtual artificial head" (VAH). The VAH filter coefficients are calculated by incorporating regularization to account for small errors in the characteristics and/or the position of the microphones. A common way to increase robustness is to impose a so-called white noise gain (WNG) constraint. The larger the WNG, the more robust the HRTF-synthesis will be. On the other hand, this comes at the cost of decreasing the synthesis accuracy for the given HRTF set. Thus, a compromise between robustness and accuracy must be found, which furthermore depends on the used setup (sensor noise, mechanical stability etc.). In this study, different WNG are evaluated perceptually by four expert listeners for two different microphone arrays. The aim of the study is to derive microphone array-dependent WNG regions which result in appropriate perceptual performance. It turns out that the perceptually optimal WNG varies with the microphone array, depending on the sensor noise and mechanical stability but also on the individual HRTFs and preferences. These results may be used to optimize VAH regularization strategies with respect to microphone characteristics, in particular self noise and stability.*

## 6.1 Introduction

In order to take into account spatial cues within a binaural reproduction, the use of so-called artificial heads, which are a replica of real human heads and pinnae, is common practice today. By this means the signals at the ears receive characteristic spatial information, which encompasses interaural time and level difference cues, but also spectral cues due to the shape of the pinna, for instance. As a disadvantage, artificial heads are inherently bound to non-individual (average) anthropometric geometries and are most often implemented as bulky devices. Alternatively, the individual frequency-dependent directivity patterns of a human head, i.e. head-related transfer functions (HRTFs), can be synthesized with a microphone array and digital filtering (cf. Mellert and Tohtuyeva (1997), Kahana *et al.* (1999), Atkins (2011a)), which will be referred to as a virtual artificial head (VAH). A VAH is more flexible than real artificial heads, since, e.g., the filters can be adjusted post-hoc to match any individual sets of HRTFs. In contrast to approaches in the spherical harmonics domain (i.e. applying a spherical harmonics decomposition, optimization and synthesis, cf. Atkins (2011a), Zotkin *et al.* (2009), Castaneda *et al.* (2013) and Sakamoto *et al.* (2013)), the VAH-synthesis in this study is optimized in the frequency domain for discrete directions in the horizontal plane only, assuming the intermediate directions to be inherently interpolated by the VAH. One advantage of this approach is that much fewer microphones are needed in comparison to e.g. spherical harmonics-based approaches (cf. Castaneda *et al.* (2013) and

Sakamoto *et al.* (2013)). The individual filter coefficients can be calculated by optimizing various cost functions, where a least squares cost function is known to yield good perceptual results (cf. Rasumow *et al.* (2013a)) and is thus used in this study (cf. section 6.2). The robustness of the filter coefficients against small deviations of the microphone characteristics and/or positions is usually assured by imposing a constraint on the so-called white noise gain (WNG). By doing so, the robustness of the filter coefficients increases with larger WNG, while the synthesis accuracy for a given HRTF set decreases and vice versa. Thus, it seems reasonable to find a compromise for the regularization, by assessing the perceptual appraisal of HRTF-synthesis using the VAH as a function of the WNG. Two microphone arrays were used in this study. These arrays were used to measure the steering vectors (as opposed to the application of analytical steering vectors in e.g. Atkins (2011a), Zotkin *et al.* (2009) and in Rasumow *et al.* (2011b) or **Chapter 2**) and to synthesize individual ear signals by individually recalculating pre-recorded signals.

## 6.2 Regularized least squares cost function

Consider the desired directivity pattern $\mathrm{D}(f,\theta)$ as a function of frequency $f$ and azimuthal angle $\theta$, as well as the N×1-dimensional steering vector $\mathbf{d}(f,\theta)$, which represents the frequency- and angle-dependent transfer functions between the source and the N microphones. The synthesized directivity pattern $\mathrm{H}(f,\theta)$ of the VAH for one particular set of steering vectors $\mathbf{d}(f,\theta)$ can be expressed as in equation 1.5, where the N×1-dimensional vector $\mathbf{w}(f)$ contains the complex-valued filter coefficients for each microphone per frequency $f$.

In order to calculate the filter coefficients $\mathbf{w}(f)$, one may minimize the narrowband least squares cost function $\mathrm{J}_{\mathrm{LS}}$, being the sum over P directions of the squared absolute differences between $\mathrm{H}(f,\theta)$ and $\mathrm{D}(f,\theta)$ (cf. equation 1.8). In this study, the VAH filters were optimized to represent individual HRTFs measured in the horizontal plane with an equidistant angular spacing of $\Delta\theta = 15^\circ$, resulting in $\mathrm{P} = 24$ directions. A straightforward minimization of equation 1.8, however, may result in non-robust filter coefficients $\mathbf{w}(f)$, where small errors of the microphone positions and/or characteristics may cause huge errors of the synthesized directivity patterns (cf. Rasumow *et al.* (2011b) or **Chapter 2** and Doclo and Moonen (2003b)) and which may lead to an undesirable amplification of spatially uncorrelated noise at the microphones. More robust filter coefficients can be obtained by imposing a constraint on the filter coefficients. The mean white noise gain ($\mathrm{WNG}_{\mathrm{m}}$), defined in equation 3.8, relates the mean output power from all P considered directions to the output power for spatially uncorrelated noise at the microphones (cf. Simmer *et al.* (2001)). Usually, for beamforming applications the WNG is only considered for a certain direction (steering direction $\theta_d$, cf. equation 3.3) (cf. Bitzer and Simmer (2001), Mabande *et al.* (2009) and Rasumow *et al.* (2013a)), whereas the $\mathrm{WNG_m}$ in equation 3.8 is referred to as the mean WNG over all considered directions $\theta$. This modification of the WNG was applied since a direction-dependent constraint (as is realized in equation 3.3) would yield a direction-dependent regularization, which is not desirable for a multi-directional VAH-synthesis (cf. **Chapter 3**). Hence, the mean white noise gain $\mathrm{WNG_m}$ incorporating all considered directions is used in this study: Positive $\mathrm{WNG_m}$ represent an attenuation of spatially uncorrelated noise, whereas negative $\mathrm{WNG_m}$ represent an amplification relative to the mean output power from the P considered directions. We will apply regularization by imposing the constraint $\mathrm{WNG_m}(f) \geq \beta_m$, where the minimal desired mean white noise gain $\beta_m$ has to be chosen manually according to the expected error of the steering vectors (cf. Rasumow *et al.* (2011b) or **Chapter 2**). The combination of the least squares cost function from equation 1.8 with the constraint incorporating equation 3.8 results in the cost function in equation 3.10. The filter coefficients minimizing the cost function in equation 3.10 are given in equation 3.11.

While the least squares solution of the cost function in equation 1.8 is quite well known in literature (cf. Doclo and Moonen (2003b), Rasumow *et al.* (2013a)), the regularization term in equation 3.11 differs from usual regularization strategies, as for instance diagonal loading (cf. Li *et al.* (2003)), Tikhonov-regularization or similar regularization approaches (cf. Kirkeby and Nelson (1999)). The main difference lies in the dependence of the regularization on the applied steering vectors ($\mathbf{Q_m}(f)$ in equation 3.8) and the desired $\mathrm{WNG_m}$ $\beta_\mathrm{m}$. However, the presented regularization approaches diagonal loading or Tikhonov-regularization for very large $\beta_\mathrm{m}$ (i.e., for the most stringent regularization possible). The optimal $\mu$ in equation 3.11 to satisfy the desired WNG-constraint was determined iteratively. Analogous to the procedure in Rasumow *et al.* (2013a), $\mu$ was increased starting from $\mu = 0$ in steps of $\varDelta\mu = \frac{1}{100}$ until $\mathrm{WNG_m}(f, \mu) \geq \beta_\mathrm{m}$ or $\mu_\mathrm{max} = 100$ were reached for each $f$ (this only occurred at very high frequencies).

### 6.2.1 Influence of the WNG-constraint on the VAH-synthesis

The accuracy of the VAH-synthesis depends on the desired HRTFs, the number of microphones, the topology of the microphone array, the cost function and the Lagrange multiplier $\mu$ (cf. equation 3.11). In general, the desired $\mathrm{WNG_m}$ is approached by gradually increasing $\mu$. This in turn

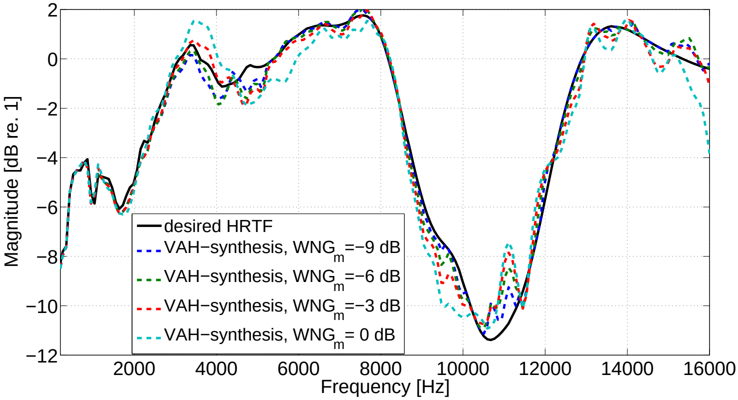


**Figure 6.1:** Magnitude of the desired HRTF ($\theta = 90°$) for the left ear of subject $\mathrm{S}_1$ (black line) and VAH-synthesis with various $\mathrm{WNG}_\mathrm{m}$ (dashed lines) for $\mathrm{array}_2$ as a function of frequency.

will cause increasing deviations of the synthesis from the desired HRTF. The magnitude of the resulting $\mu$ is primarily determined by the desired $\mathrm{WNG}_\mathrm{m}$ $\beta_\mathrm{m}$. On the one hand, the regularization yielding a desired $\mathrm{WNG}_\mathrm{m}$ unavoidably causes distortions of the VAH-synthesis which depends on the desired HRTFs and steering vectors. This aspect is exemplarily depicted in Figure 6.1. On the other hand, higher $\mathrm{WNG}_\mathrm{m}$ are associated with an increased robustness against small changes of the microphone characteristics and with a lower amplification of spatially uncorrelated noise at the microphones.

## 6.3 Used microphone arrays

The main goal of this study is to investigate the perceptually optimal $\mathrm{WNG}_\mathrm{m}$ for different subjects, using different microphone arrays. For this reason, the perceptual evaluation was performed with recordings using two open planar microphone arrays incorporating different kinds of microphones and support structures but using the same number of microphones and an identical topology which was determined according to the procedure described in Rasumow *et al.* (2011b) or **Chapter 2**. The advantage of using open planar arrays over rigid spheres or the like is the opportunity to realize various two-dimensional inter-microphone distances. By this means, a mathematically motivated microphone topology according to Rasumow *et al.* (2011b) or **Chapter 2** was chosen, which is assumed to yield appropriate results regarding the accuracy and robustness of the synthesis.

The first microphone array (array$_1$, left panel in Figure 6.2) consisted of 24 Sennheiser KE 4-211-2 microphones. The individual microphones were mounted on a wooden plate using a solid wire construction. Together with analog preamplifiers the sensor noise of each single microphone signal was approximately 35 dB(A). No absorbent material was used for the support structure of array$_1$.

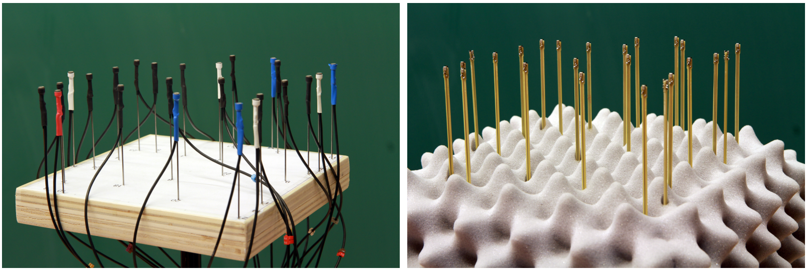


**Figure 6.2:** Two used microphone arrays with 24 KE-4 microphones (array$_1$, left) and 24 sensors composed of 48 MEMS microphones (array$_2$, right) with the same planar microphone topology according to the procedure described in Rasumow *et al.* (2011b) or **Chapter 2**.

For the second array (array$_2$), micro-electromechanical system (MEMS) microphones (Analog Devices ADMP 504 Ultralow Noise Microphone) were used in a custom-made electrical circuit. Here, each sensor is composed of two MEMS microphones, yielding a sensor noise of approximately 27 dB(A), which is quite low for this kind of microphones. The directivity of such a composed sensor can be assumed to be negligible for the frequencies of interest (i.e. $f \lesssim 16$ kHz). For array$_2$, 24 of these sensors (consisting of 48 MEMS microphones) were mounted on a printed circuit board (cf. right panel in Figure 6.2) with the same topology as for array$_1$. In order to reduce the effects of standing waves between the sensors and the board, array$_2$ is covered with absorbent material.

## 6.4 Experimental procedure

### 6.4.1 Material

Prior to the listening experiment, individual HRTFs and headphone (AKG K-240 Studio) transfer functions (HPTFs) were measured for four subjects using the blocked ear method according to Hammershøi and Møller (1996). For measuring the HPTFs, subjects were instructed to reposition the headphone ten times to various realistic carrying positions which successively yielded ten different individual HPTFs. The individual HPTF resulting in the smallest dynamic range of its magnitude for frequencies 300 Hz $\leq f \leq$ 16000 Hz was inverted in the frequency domain and transformed into the time domain. The HRTFs as well as the inverse HPTFs were implemented as finite impulse response (FIR) filters with a filter length of 256 taps, which at a sampling frequency of $f_s = 44100$ Hz corresponds to $\approx 5.8$ ms. This filter length was chosen to incorporate all aspects associated with an appropriate binaural reproduction (cf. Rasumow *et al.* (2012b) or **Chapter 4**). The individual HRTFs as well as the steering vectors $\mathbf{d}(f, \theta)$ for the two microphone arrays were measured in the horizontal plane with an angular spacing of 15°. All HRTFs were smoothed in the frequency and spatial domain prior to the VAH-synthesis according to the perceptual limits derived in the study described in Rasumow *et al.* (2014c) or **Chapter 5**. Moreover, the associated impulse responses of all measured steering vectors $\mathbf{d}(f, \theta)$ were truncated to a filter length of 256 taps in order to achieve smoother transfer functions.

### 6.4.2 Test stimulus

In order to cover a wide frequency range and simultaneously to include temporal cues, the test stimulus for perceptual evaluation consisted of 3 short bursts of pink noise filtered with an eighth-order bandpass filter with cutoff frequencies $f_{\mathrm{low}} = 300$ Hz and $f_{\mathrm{hi}} = 16000$ Hz. The lower bandwidth limitation of the test stimulus $f_{\mathrm{low}}$ was chosen due to the limits of the used loudspeakers. However, since the influence of varying the $\mathrm{WNG_m}$ is primarily evident for frequencies $f \geq 3$ kHz (cf. Figure 6.1) it seems reasonable to assume that this limitation does not have a significant influence on the perceptual evaluation. Each noise burst lasted $\frac{1}{3}$ s with 0.01 s onset-offset ramps followed by silence of $\frac{1}{6}$ s. This test stimulus was intended to facilitate the evaluation of spectral deviations, temporal dispersion but also the influence of the sensor noise. The presented stimuli were calibrated with a G.R.A.S. type 43AA artificial ear to have 70 dB SPL for the frontal direction $\theta = 0°$.

### 6.4.3 Methods

A listening test was carried out with four experienced listeners (two of them are authors of this article). The subjects were instructed to rate four different aspects (localization, sensor noise, overall performance and spectral coloration, cf. section 6.4.3.1) of a test presentation with respect to the reference presentation (binaural reproduction with original individual HRTFs and inverse HPTFs). Since the quality of the reference setting has a major effect on the evaluation, it needed to be assured that the individual binaural reproductions incorporated all essential individual spatial characteristics. For this reason, the individual binaural reproductions used in the reference setting were played to the subjects before the experimental procedure in a preliminary listening test. All subjects were able to perceive the presented stimuli outside of the head and correctly assigned the corresponding directions in the horizontal plane.

Prior to the listening tests, the steering vectors were measured and the test stimuli were recorded using the two microphone arrays (cf. section 6.3) in an anechoic room. Furthermore, the individual VAH filters were optimized to synthesize the individual HRTFs in the horizontal plane with an angular spacing of $\Delta\theta = 15^\circ$. In the test condition, the sum of the filtered stimuli (representing the synthesized ear signals, cf. equation1.5) was also filtered with the inverse HPTF filters (same procedure as in the reference setting) and played to the subject via headphones. In both conditions, the stimuli were played back in an infinite loop with the possibility to switch between the reference and test condition or to stop the playback. To limit the number of experiments to a manageable amount, three directions in the horizontal plane were chosen for evaluation with azimuth angles $\theta = 0^\circ$ (front), $\theta = 90^\circ$ (left) and $\theta = 225^\circ$ (back right) and the $\mathrm{WNG_m}$ was one of $\mathrm{WNG_m}(f)$ = -9 dB, -6 dB, -3 dB or 0 dB for all $f$. These preselected $\mathrm{WNG_m}$ were assumed to roughly cover the area with the best suited $\mathrm{WNG_m}$ based on previous preliminary tests.

The three tested azimuthal directions $\theta$, the two microphone arrays as well as the four $\mathrm{WNG_m}$ were varied in randomized order within one experimental run with three random presentations (retest) for each condition. The true conditions of the signals in the reference and test setting were hidden to the subjects. In total, 216 conditions (presented signal pairs) were evaluated by each subject, whereas one of the tested parameters was eliminated from the analysis in this article in hindsight. Hence, 3 directions $\times$ 2 arrays $\times$ 3 presentations $\times$ 4 $\mathrm{WNG_m} = 72$ individual perceptual ratings (of a total of originally 216 individually gathered ratings) will be analyzed and discussed in section 6.5 and 6.6. Within each condition, subjects were able to switch between the reference and the test setting arbitrarily. The entire experiment was performed using an English category scale, ranging between *bad*, *poor*, *fair*, *good* and *excellent* with four intermediate undeclared steps (cf. Rasumow *et al.* (2013a)). Each session lasted approximately 120-180 minutes, where subjects were able to subdivide the session arbitrarily and take as many breaks as they wanted. Prior to the evaluation each subject had time for familiarization with the various reference and test conditions.

#### 6.4.3.1 Assessed aspects

The subjects were instructed to rate the quality of the test setting with respect to the reference setting for four chosen aspects which are assumed to be significant for the VAH-synthesis:

- localization: The evaluation of localization incorporated the perceived angle of incidence (azimuth and elevation) and the perceived distance in combination.
- sensor noise: Subjects were instructed to evaluate the perceived sensor noise which was primarily apparent in the temporal pauses of the test stimulus.
- overall performance: The evaluation of the perceived overall performance incorporated all feasible aspects depending on the taste and preferences of the individual subject.
- spectral coloration: Subjects were instructed to evaluate the perceived spectral coloration without evaluating the potential deviations of localization or other cues.

## 6.5 Results and discussion - Perceptual evaluation

The mean and the standard deviations (over three randomized presentations) of all individual ratings are depicted in Figure 6.3 as a function of the $\mathrm{WNG_m}$ on the x-axis with the assessed aspects separated in rows, the directions $\theta$ separated in columns and the color indicating the subjects. The average performance (means and standard deviations over subject) is depicted in Figure 6.4, with the color indicating the assessed aspects (see legend).

In general, the perceptual ratings and their variation within repeated trials in Figure 6.3 (standard deviation depicted as error bars) seem to depend on the direction of incidence $\theta$ and the used microphone array, but as well on the subject. This is an effect of individual preferences with individual internal scales and was to be expected according to similar studies (cf. Rasumow *et al.* (2013a)). In order to analyze potential preferences regarding the $\mathrm{WNG_m}$ for the application of a VAH, primarily the relative tendencies of intra- and inter-individual perceptual ratings depending on the $\mathrm{WNG_m}$ are focused on.

**Table 6.1:** p-values (rounded to 3 digits) according to the Friedman test regarding localization, overall performance, sensor noise and coloration for the three tested directions separately. p-values indicating significantly different ratings when varying the $\mathrm{WNG_m}$ ($p \leq \frac{0.05}{24} = 0.0021$) are depicted as bold numbers.

| localization | array$_1$ | array$_2$ | overall | array$_1$ | array$_2$ |
|---|---|---|---|---|---|
| $\theta = 0^\circ$ | 0.164 | 0.445 | $\theta = 0^\circ$ | 0.341 | 0.081 |
| $\theta = 90^\circ$ | 0.004 | 0.006 | $\theta = 90^\circ$ | **0.000** | 0.129 |
| $\theta = 225^\circ$ | 0.147 | 0.933 | $\theta = 225^\circ$ | 0.109 | 0.188 |

| sensor noise | array$_1$ | array$_2$ | coloration | array$_1$ | array$_2$ |
|---|---|---|---|---|---|
| $\theta = 0^\circ$ | 0.004 | 0.049 | $\theta = 0^\circ$ | 0.035 | 0.578 |
| $\theta = 90^\circ$ | **0.000** | 0.340 | $\theta = 90^\circ$ | **0.000** | 0.827 |
| $\theta = 225^\circ$ | **0.000** | 0.079 | $\theta = 225^\circ$ | 0.015 | 0.319 |

Although means and standard deviations were used for illustrating the ratings in Figures 6.3 and 6.4 (for increased clarity), a non-parametric statistical test was applied. The Friedman test was applied to analyze whether the ratings for at least one of the tested $\mathrm{WNG_m}$ (for a fixed direction, array and assessed aspect) was considerably different than the ratings for the other $\mathrm{WNG_m}$. A sufficiently small p-value indicated an effect of the $\mathrm{WNG_m}$ on the ratings. The p-values for the assessed aspects (separate boxes), the applied arrays (columns) and directions (rows) are given in Table 6.1. The p-values for conditions indicating a significant effect of the $\mathrm{WNG_m}$ on the perceptual ratings (considering the Bonferroni correction for 24 repeated tests, a p-value of $p \leq \frac{0.05}{24}$ is assumed to indicate a significant effect of the $\mathrm{WNG_m}$) are depicted as bold numbers. However, due to the rather small number of subjects and the presumably low test power, the p-values in Table 6.1 may primarily be used to highlight tendencies of all ratings for fixed conditions without postulating any statistical (in)significances for the effect of the $\mathrm{WNG_m}$.

In summary, it emerges that the tested $\mathrm{WNG_m}$ mainly seem to have an effect on the ratings for array$_1$ with regard to sensor noise and coloration. The ratings regarding localization primarily seem to be affected by the $\mathrm{WNG_m}$ for $\theta = 90^\circ$ and both arrays. The ratings regarding the overall

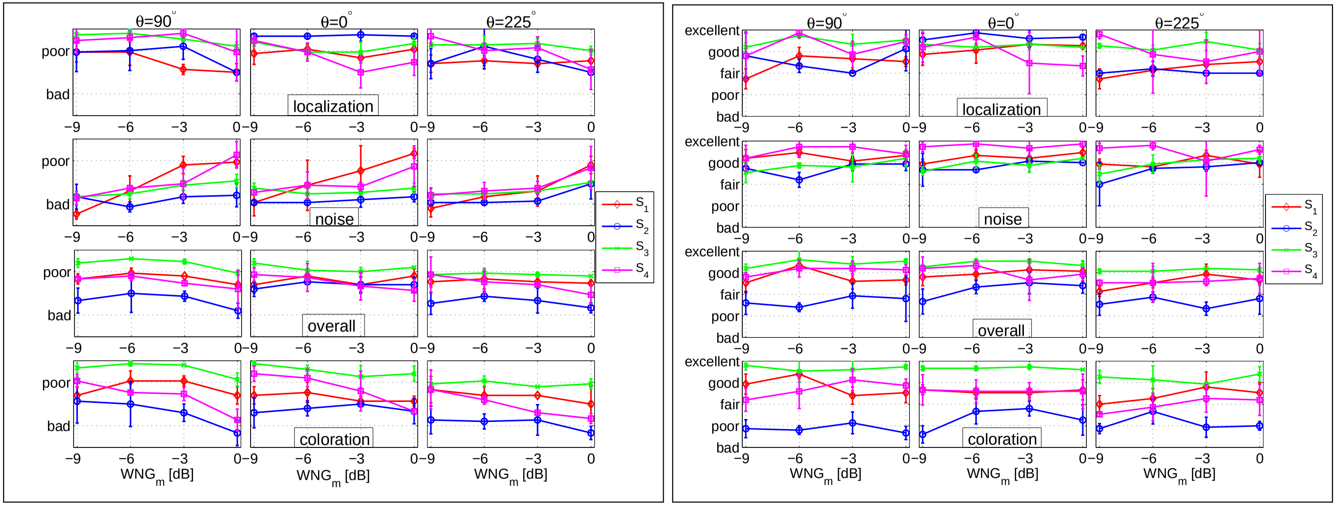


**Figure 6.3:** Perceptual ratings for array$_1$ (left block) and array$_2$ (right block). The aspects of evaluation are aligned in separate rows (first row: localization, second row: sensor noise, third row: overall performance and fourth row: spectral coloration) and the direction of arrival $\theta$ is aligned in three columns ($\theta = 90°$ in the left column, $\theta = 0°$ in the middle column and $\theta = 225°$ in the right column). The individual ratings (mean and standard deviation over three randomized presentations) are depicted as a function of the WNG$_\mathrm{m}$ in dB. The colors and markers indicate the four subjects (S$_1$, S$_2$, S$_3$ and S$_4$).

performance seem to be affected by the $\mathrm{WNG_m}$ mainly for $\mathrm{array}_1$ and $\theta = 90°$.

### 6.5.1 Localization

In general, all subjects concordantly reported the localization in the horizontal plane to be synthesized well by the VAH. However, the aspect localization was also used to evaluate the perceived distance of the sound source (cf. section 6.4.3.1). The perception of distance may vary noticeably when interaural level differences from lateral directions are not synthesized accurately. This may be a possible explanation for the better ratings for $\theta = 0°$, which is especially evident for subject $\mathrm{S_1}$ and $\mathrm{S_2}$ (cf. Figure 6.3).

For subject $\mathrm{S_3}$, the ratings with regard to localization vary hardly with the tested $\mathrm{WNG_m}$ nor with the array. The p-values from Table 6.1 indicate the most notable effect of the $\mathrm{WNG_m}$ on the ratings with regard to localization for $\theta = 90°$ with both arrays. This aspect is also apparent in the averaged ratings (cf. Figure 6.4) for $\mathrm{array}_1$, where the ratings decrease for higher $\mathrm{WNG_m}$. However, there does not seem to be such an unambiguous tendency for the ratings with $\mathrm{array}_2$ and $\theta = 90°$. Moreover, the averaged ratings also seem to decrease slightly with increasing $\mathrm{WNG_m}$ for $\theta = 225°$ and $\mathrm{array}_1$. This slight effect is concordantly associated with a relatively higher p-value from the Friedman test ($p=0.147$), as well indicating a less notable effect of the tested $\mathrm{WNG_m}$.

In sum, the ratings with regard to localization seem to decrease with higher $\mathrm{WNG_m}$ using $\mathrm{array}_1$ and are approximately constant or do not vary in a clearly interpretable way for $\mathrm{array}_2$.

### 6.5.2 Sensor noise

The ratings with regard to the perceived sensor noise for $\mathrm{array}_1$ are considerably different from the ratings for $\mathrm{array}_2$. Especially for lower $\mathrm{WNG_m}$ ($\mathrm{WNG_m} \leq -3$ dB), the sensor noise for $\mathrm{array}_1$ is evaluated worse compared to the ratings for $\mathrm{array}_2$. The ratings improve with increasing $\mathrm{WNG_m}$, especially for subjects $\mathrm{S_1}$ and $\mathrm{S_4}$ where the ratings for $\mathrm{WNG_m}=0$ dB and $\mathrm{array}_1$ are approximately in the range of the ratings for $\mathrm{array}_2$. The ratings for $\mathrm{array}_2$ vary much less with the $\mathrm{WNG_m}$, resulting for subjects $\mathrm{S_1}$ and $\mathrm{S_4}$ in variations of approximately the amount of their standard deviations (over randomized presentations). This effect is also represented by the associated p-values, with relatively small p-values ($p\leq 0.004$) for all di-

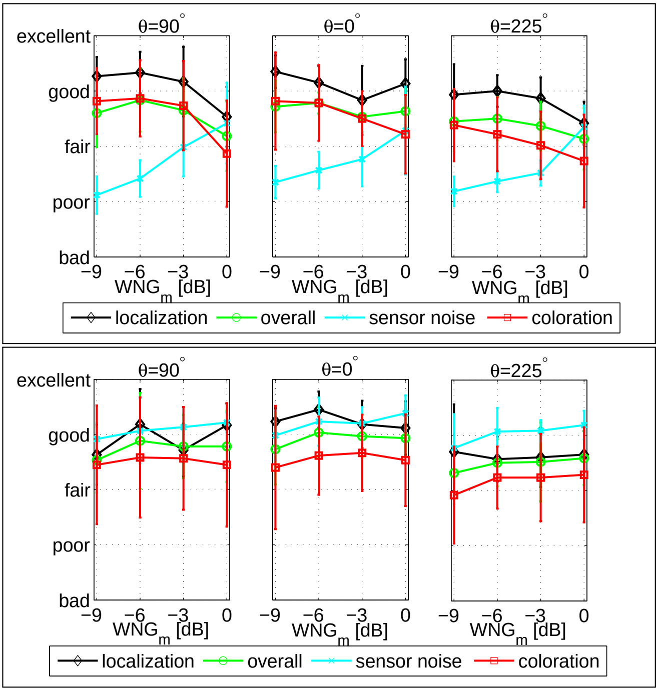


**Figure 6.4:** Perceptual ratings averaged over all subjects for array$_1$ (upper block) and array$_2$ (lower block) are depicted as the mean and the standard deviation for the four evaluated aspects (localization, overall performance, sensor noise and coloration).

rections $\theta$ and array$_1$ and rather high p-values ($p \geq 0.049$) for all directions $\theta$ and array$_2$. On the other hand, there also seems to be a slight trend towards better ratings for higher WNG$_\mathrm{m}$ with array$_2$, with the worst ratings for the lowest WNG$_\mathrm{m}$ of -9 dB (in the averaged ratings in Figure 6.4 as well as for subject S$_2$ and S$_3$ and $\theta = 225°$ in Figure 6.3). This indicates that sensor noise is not negligible for all subjects even with array$_2$. However, the averaged ratings in Figure 6.4 as well as the associated p-values in Table 6.1 indicate that the ratings vary much less with the tested WNG$_\mathrm{m}$ when using array$_2$ compared to array$_1$.

In sum, the perceptually optimal WNG$_\mathrm{m}$ with regard to sensor noise seems to vary with the used microphone array and its inherent sensor noise. The ratings of the sensor noise (if detectable) generally seem to enhance with higher WNG$_\mathrm{m}$, which was to be expected.

### 6.5.3 Overall performance

The largest variations of the ratings with regard to overall performance can be observed across different subjects, while the ratings remain rather constant over different $\mathrm{WNG_m}$, especially for subject $\mathrm{S_3}$ with both microphone arrays. However, there seems to be a slight trend to worse ratings for higher $\mathrm{WNG_m}$ using $\mathrm{array_1}$ (cf. $\theta = 90°$ and $\theta = 225°$) as well as for the lowest $\mathrm{WNG_m}$ of -9 dB (presumably due to the more disturbing sensor noise). This trend is also apparent from the averaged performance using $\mathrm{array_1}$ in Figure 6.4, with the Friedman test indicating the largest effect of the $\mathrm{WNG_m}$ for $\theta = 90°$.

The ratings vary less clearly with the $\mathrm{WNG_m}$ for $\mathrm{array_2}$. There, the best ratings were mostly observed at higher $\mathrm{WNG_m}$ (cf. $\mathrm{S_1}$, $\theta = 225°$ and $\mathrm{S_2}$, $\theta = 0°$) and worsened slightly for the lowest $\mathrm{WNG_m}$ (cf. Figure 6.4). In general, the ratings with regard to overall performance seem to be correlated to the ratings with regard to spectral coloration (cf. section 6.5.4), emphasizing the relevance of spectral coloration for the evaluation of a binaural synthesis with respect to a reference condition. Furthermore, comparing the averaged ratings of the overall performance for both microphone arrays (cf. Figure 6.4) at higher $\mathrm{WNG_m}$, the ratings seem better for $\mathrm{array_2}$ compared to $\mathrm{array_1}$. This aspect is assumed to be a consequence of the lower inherent sensor noise of $\mathrm{array_2}$: To achieve a desired $\mathrm{WNG_m}$, the required $\mu$ is usually lower for $\mathrm{array_2}$ compared to $\mathrm{array_1}$, cf. Figure 6.5. Although not shown here, this tendency has also been observed for the other subjects and $\mathrm{WNG_m}$. The most plausible expla-

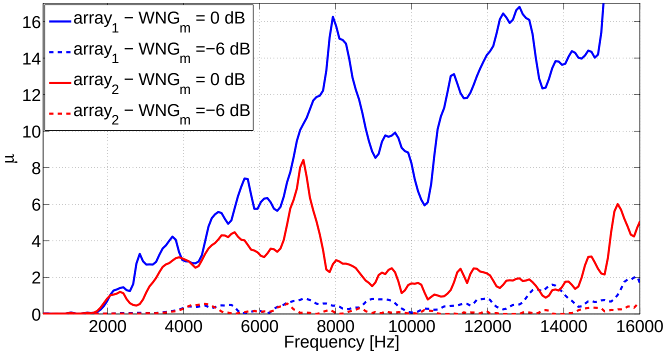


**Figure 6.5:** Exemplary course of the Lagrange multiplier $\mu$ (cf. equation 3.11) for $\mathrm{array_1}$ and $\mathrm{array_2}$ (blue and red lines, respectively) and $\mathrm{WNG_m}$ of 0 dB and -6 dB (solid and dashed lines, respectively) as a function of frequency of the left-ear synthesis for $\mathrm{S_1}$.

nation is that $\mu$ needs to be enlarged more in order to counteract the higher inherent sensor noise of $\text{array}_1$ (resulting in larger random errors on the measured steering vectors) in comparison to $\text{array}_2$. Considering that the accuracy of a synthesis decreases with larger $\mu$, the higher inherent sensor noise of $\text{array}_1$ may therefore be a reasonable explanation for a worse accuracy of the synthesis and subsequently for the worse ratings at $\text{WNG}_\text{m} \geq -3$ dB.

In sum, the ratings with regard to overall performance seem best for $\text{WNG}_\text{m}$=-6 dB and $\text{WNG}_\text{m}$=-3 dB when using $\text{array}_1$ and for $\text{WNG}_\text{m} \geq$=-6 dB when using $\text{array}_2$.

### 6.5.4 Spectral coloration

The ratings with regard to spectral coloration seem to differ considerably for the four subjects. This phenomenon may be partly explained by the fact that the perception and evaluation of spectral coloration is influenced by the perceived localization and the interaction with the perceived sensor noise. This may introduce a certain degree of interpretation to assess this aspect. Furthermore, subjects have individual internal scales and assess individually. This is primarily evident when comparing the ratings of subject $\text{S}_2$ and $\text{S}_3$, for instance. The ratings of subject $\text{S}_3$ vary roughly between good and excellent while the ratings of subject $\text{S}_2$ vary roughly between fair and poor, representing the most negative ratings of this study.

In general, slightly better ratings are evident for the frontal direction $\theta = 0°$ compared with the lateral directions. The averaged ratings in Figure 6.4 as well as the p-values in Table 6.1 indicate that the ratings for $\text{array}_1$ vary considerably across the tested $\text{WNG}_\text{m}$ for all tested directions $\theta$ with decreasing averaged ratings for higher $\text{WNG}_\text{m}$ in Figure 6.4. This tendency does, however, not hold for $\text{array}_2$, with its p-values being relatively high ($p \geq 0.319$) for all directions. This array-dependent difference of ratings may be explained by the differently sized Lagrange multipliers $\mu$ for the two applied arrays (cf. Figure 6.5 and the discussion in section 6.5.3).

In sum, the ratings of the perceived spectral coloration seem to vary with subjects and also with the used microphone arrays. On the one hand, higher $\text{WNG}_\text{m}$ seem to distort the perception of spectral coloration for $\text{array}_1$. On the other hand, the ratings with regard to spectral coloration do not seem to vary considerably with the tested $\text{WNG}_\text{m}$ when using $\text{array}_2$.

## 6.6 Conclusions and further work

In this chapter the effect of regularization on the appraisal of binaural synthesis using the VAH was investigated.

The evaluation of the perceived sensor noise (if noticeable) seems to improve considerably with increasing $\mathrm{WNG_m}$, whereas the explicit presence of sensor noise (primarily at lower $\mathrm{WNG_m}$ with $\mathrm{array}_1$) does not consistently seem to deteriorate the overall performance. This latter observation may be due to the chosen test paradigm - it is conceivable that noise is more disturbing in other scenarios, e.g. when listening to music recordings. Furthermore, the higher sensor noise of $\mathrm{array}_1$ also seems to have caused worse ratings with regard to localization, coloration and overall performance for $\mathrm{WNG_m} \geq -3$ dB. This phenomenon can be explained by the larger Lagrange multipliers $\mu$ that were required for $\mathrm{array}_1$ to comply with a desired $\mathrm{WNG_m}$ (cf. section 6.5.3).

The best compromise with regard to all assessed aspects and the associated robustness can be found at $\mathrm{WNG_m}$ of -6 dB and -3 dB for $\mathrm{array}_1$ and at the highest of the tested $\mathrm{WNG_m}$ of 0 dB for $\mathrm{array}_2$.

In general, the obtained ratings confirm the validity of synthesizing HRTFs using microphone arrays in conjunction with individually suited $\mathrm{WNG_m}$. There is still room for improvement for the calculation and regularization of the filter coefficients, especially with regard to spectral coloration. Thus, one next step may be to devise a more appropriate and frequency-dependent regularization method.

# 7

# Perceptual evaluation of individualized binaural reproduction using a virtual artificial head[*]

---

[*] This chapter is based on the publications

- Rasumow, Blau, Doclo, Hansen, van de Par, Püschel, and Mellert (2014a), "Individualized binaural reproduction using a virtual artificial head," in *Fortschritte der Akustik - DAGA 2014*, Oldenburg, Germany, pp. 26-27.
- Rasumow, Hansen, van de Par, Püschel, Mellert, Doclo, and Blau (2014e), "Perceptual evaluation of individualized binaural reproduction using a virtual artificial head," *Journal paper in preparation.*

*In binaural recordings, spatial information can be captured by using so-called artificial heads, which are a replica of real human heads with ear microphones and average anthropometric geometries. However, because of their non-individual character, such recordings often entail perceptual deficiencies (front-back confusion, internalization etc.). Alternatively, individually measured head-related transfer functions (HRTFs) can be approximately synthesized using a microphone array in conjunction with a filter-and-sum beamformer (referred to as virtual artificial head, VAH). Its main advantage over traditional artificial heads is the possibility to adapt one recording post-hoc to individual HRTFs by an appropriate modification of the directivity pattern of the VAH, but also its smaller size and weight.*

*To validate the suitability of a VAH to synthesize individual HRTFs, the realization of a VAH as a planar microphone array with 24 microphones is presented in this study. Binaural reproductions using the VAH, three traditional artificial heads and individual HRTFs were perceptually evaluated in the horizontal plane with respect to the original free field presentation. Evaluations were conducted for explicitly considered directions in the optimization of the VAH-filters but also for intermediate directions, which are assumed to be interpolated implicitly by the VAH. The ratings confirm the validity of the concept of synthesizing HRTFs using a VAH. It is found that the VAH-synthesis enables sufficiently good binaural reproductions which in general yield better perceptual ratings in comparison to traditional artificial heads for explicitly considered directions and approximately equivalent ratings for intermediate directions.*

## 7.1 Introduction

Spatial information is an important factor for the perception and appraisal of sounds. Spatial information can be introduced into recordings and measurements by using so-called artificial heads, which are reproductions of real human heads with microphones placed in the ear canals (cf. Paul (2009) for an extensive review). Alternatively, the direction- and frequency-dependent head-related transfer functions (HRTFs) can be approximately synthesized using a set of spatially distributed microphones with appropriate digital filtering (cf. Chen *et al.* (1992), Tohtuyeva and Mellert (1999), Kahana *et al.* (1999), Atkins (2011a) and Rasumow *et al.* (2011b, 2013a, 2014d)). Such a device is referred to as a virtual artificial head (VAH). The main advantages of a VAH are the possibility of adjusting the filter coefficients to HRTFs of different listeners (individualization) and to different look directions (orientation), the possibility of employing head tracking in the reproduction stage and a better flexibility and manageability due to the smaller size/weight of the device. One typical feature of the VAH is the fact that the filter coefficients are optimized using the measured steering vectors and HRTFs for discrete directions, assuming that the intermediate directions will be interpolated implicitly by the VAH. This concept is, however, in contrast to other approaches as, for instance, the optimization in the spherical harmonics domain, where spatial information is represented using model assumptions (cf. Atkins (2011a,b) and Zotkin *et al.* (2009)). The previously described optimization using the measured steering vectors and HRTFs for discrete directions was chosen in this study since spherical harmonics based strategies generally need many more HRTF measurements and especially more microphones (cf. Castaneda *et al.* (2013) and Sakamoto *et al.* (2013)). This study is aimed at perceptually assessing the quality of the VAH-synthesis for explicitly considered directions but also for intermediate directions with respect to the free field presentation. In order to enable the best possible synthesis of individual HRTFs with a VAH, various studies were performed in advance to devise an appropriate microphone topology (cf. Rasumow *et al.* (2011b) or **Chapter 2**), an appropriate cost function (cf. Rasumow *et al.* (2013a)), appropriate regularization strategies (cf. Rasumow *et al.* (2014d) or **Chapter 3**), appropriate regularization parameters (cf. Rasumow *et al.* (2014b) or **Chapter 6**) and an appropriate smoothing of HRTFs (cf. Rasumow *et al.* (2012b) and Rasumow *et al.* (2014c) or **Chapter 5**) prior to the synthesis.

In this study, the free field presentation from (single) loudspeakers in an anechoic room served as the reference condition. Binaural presentations via headphones (with headphone equalization) using the following five test methods were perceptually evaluated in comparison to the reference condition: Binaural reproductions using individual HRTFs, individualized synthesis using the VAH and non-individual presentations using three artificial heads ($DH_1$, $DH_2$ and $DH_3$). The evaluations were conducted for three explicitly considered directions in the horizontal plane and as well as for three intermediate directions.

## 7.2 Methods

The procedure for the considered binaural reproductions is described in the following sections (section 7.2.1-7.2.4). The procedure and stimuli for the implementation of the evaluation experiments is described in section 7.2.5.

### 7.2.1 Head-related transfer functions (HRTFs)

Prior to the synthesis using the VAH, individual HRTFs were measured in the horizontal plane for six participating subjects who had extensive expe-

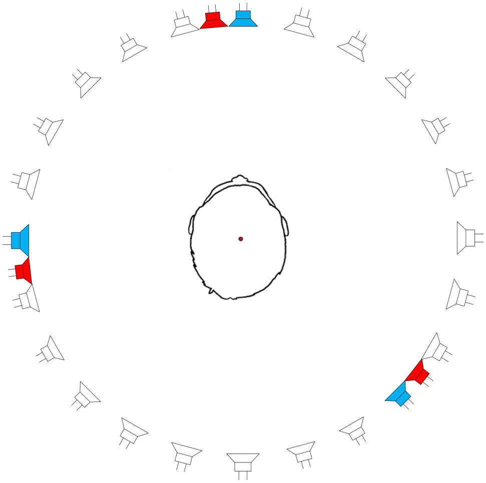


**Figure 7.1:** Illustration of the directions which are explicitly considered in the optimization of the VAH-filters (white and blue loudspeaker symbols), the tested directions for three explicitly considered directions (blue) and three intermediate directions (red symbols).

rience with this kind of psychoacoustical experiments. HRTFs were measured according to the described procedure in section 1.3, using the blocked ear method with the microphones (Knowles FG-23329 miniature electret microphones) embedded in foam earplugs (cf. Møller *et al.* (1995) and Raufer *et al.* (2013)) for 24 equidistantly spaced directions (cf. white and blue symbols in Figure 7.1) in the horizontal plane ($\theta = 0°, 15°, 30° \ldots\ 345°$). All HRTFs were truncated in the time domain to 320 samples ($\approx 7$ ms at a sampling frequency of $f_s = 44100$ Hz) with a tapered Hann-window with a descending flank of 50 samples (cf. Figure 4.1), based on the perceptual limits described in Rasumow *et al.* (2012b) or **Chapter 4**. Based on the findings from Rasumow *et al.* (2014c) or **Chapter 5**, the individual HRTFs were smoothed in the frequency- and spatial domain: The phase response of each HRTF was substituted by a linear phase for frequencies $f > 1$ kHz, which enabled a complex-valued smoothing of the HRTF into constant relative bandwidths of $\mathrm{B_W} = \frac{1}{5}$ octaves in the frequency domain. Furthermore, the spatial notches of the individual directivity patterns (HRTFs as a function of direction) were levelled out per frequency such that the dynamic range of the directivity patterns across azimuth never exceeded 29 dB.

In addition to the HRTFs for these 24 equidistantly spaced directions, HRTFs were measured for three intermediate directions in the horizontal plane ($\theta = 7.5°, 97.5° \& \ 232.5°$, cf. red symbols in Figure 7.1). Analogously to the previous procedure, the HRTFs for the three intermediate directions were truncated in the time domain (320 samples) and smoothed in the frequency domain ($\mathrm{B_W} = \frac{1}{5}$ octave bands) after substituting the measured phases with a linear phase for frequencies $f > 1$ kHz.

The individual HRTFs from the six directions were used to generate individual binaural reproductions by filtering the test signal (cf. section 7.2.5) with the associated HRTFs and subsequent headphone equalization (cf. section 7.2.4). It is to be expected that binaural reproductions using individual HRTFs should result in the best evaluations in the following experiments (cf. Wightman and Kistler (1989), Kim and Choi (2005), Masiero (2012)).

### 7.2.2 Synthesis using the virtual artificial head

Using a filter-and-sum beamformer in conjunction with an appropriate microphone array and filter coefficients, it is possible to approximately synthesize an arbitrary desired directivity pattern. Given the motivation to synthesize individual HRTFs, the filter coefficients of the VAH can be optimized by minimizing a chosen cost function between the desired HRTF and the synthesized directivity pattern of the VAH. In this study, a least squares cost function (cf. equation 3.20) within equivalent rectangular bandwidths (ERB, cf. Moore and Glasberg (1983)) in conjunction with a constraint on the mean white noise gain for all directions and frequencies within ERBs ($\mathrm{WNG_v}$, cf. equation. 3.16) according to the approach described in Rasumow *et al.* (2014d) or **Chapter 3** was used. This procedure can be interpreted as a narrowband optimization in the frequency domain which takes neighbouring frequencies in ERBs into account for the optimization and regularization of the filter coefficients.

Based on the findings from Rasumow *et al.* (2014b) or **Chapter 6**, a planar microphone array with $\mathrm{N} = 24$ sensors ($\mathrm{array}_2$, cf. right panel of Figure 6.2), each composed of 2 MEMS microphones (Analog Devices ADMP 504 Ultralow Noise Microphone), was used for synthesizing the individual binaural reproductions in the horizontal plane. In order to optimize the filter coefficients, steering vectors describing the acoustic transfer functions between a sound source from direction $\theta$ to the microphones were measured in an anechoic room in the horizontal plane for 24 equidistantly spaced ($\Delta\theta = 15°$) directions, with $\theta = 0°, 15°, 30° \ldots\ 345°$. Analogously to the HRTFs, the steering vectors were truncated in the time domain to 320 samples ($\approx 7$ ms at a sampling frequency of $f_s = 44100$ Hz) using a tapered Hann-window with a descending flank of 50 samples. The filter coefficients were calculated using equation 3.21, employing the individual HRTFs and the steering vectors for the 24 equidistantly spaced directions (cf. white and blue loudspeaker symbols in Figure 7.1) with a desired $\mathrm{WNG_v}$=2 dB. This particular value for $\mathrm{WNG_v}$ was chosen based on previous preliminary tests. The appropriate Lagrange multiplier $\mu$ in equation 3.21 was determined by gradually increasing $\mu$ (starting from $\mu = 0$) in steps of $\Delta\mu = \frac{1}{100}$ until the constraint $\mathrm{WNG_v}(f, \mu) \geq 2$ dB was satisfied (for smallest possible $\mu$). This rather heuristic procedure was chosen since the $\mathrm{WNG_v}(f, \mu)$ does not always increase monotonously with increasing $\mu$ (cf. Figure 3.2). It is worth noting that the above described procedure optimizes the filter coefficients only using the measured steering vectors and HRTFs for the 24 equidistantly spaced directions. For the intermediate directions ($\theta = 7.5°$, $97.5°$ & $232.5°$) *no* steering vectors were measured. Instead, it is assumed that the directivity patterns for the intermediate directions are interpolated implicitly by the VAH.

To test this hypothesis, the test signal (cf. section 7.2.5) was recorded with the VAH in an anechoic room for three explicitly considered loudspeaker directions ($\theta = 0°$, $90°$ & $225°$, cf. blue loudspeaker symbols in Figure 7.1) and for three intermediate loudspeaker directions ($\theta = 7.5°$, $97.5°$ & $232.5°$, cf. red loudspeaker symbols in Figure 7.1). Note that the latter three directions exhibit the largest angular deviation from the explicitly considered directions within the horizontal plane and hence are assumed to yield the most salient (possibly negative) perceptual effects due to the interpolation of the VAH.

### 7.2.3 Traditional artificial heads

In addition to the individualized VAH-synthesis and individual binaural HRTF reproductions (cf. section 1.4), three traditional artificial heads (referred to as $\mathrm{DH_1}$, $\mathrm{DH_2}$ and $\mathrm{DH_3}$) were chosen to be evaluated in this experiment. To this end, the test signal arriving from loudspeakers positioned at six directions in the horizontal plane ($\theta = 0°, 7.5°, 90°, 97.5°, 225°$ and $232.5°$, cf. red and blue symbols in Figure 7.1) was recorded with the artificial heads in an anechoic room. Two of the chosen artificial heads were commercially available ($\mathrm{DH_1}$ and $\mathrm{DH_2}$), whereas the third artificial head ($\mathrm{DH_3}$) was a custom-made device (cf. Figure 7.4) consisting of a mannequin with built-in microphones into faithful copies of human ears. In general, the chosen traditional artificial heads are based on two different design principles: The microphones of $\mathrm{DH_2}$ are placed at the entrance of the ear canal, which is a similar position compared to the measurement of HRTFs using the blocked ear method. $\mathrm{DH_2}$ exhibits rather schematically designed pinnae without mimicking distinctive details of the outer ear. On the contrary, the microphones of the other two artificial heads are placed at the ends of approximately 22 mm long ($\mathrm{DH}_1$) and 30 mm long ($\mathrm{DH}_3$) ear canals, and both artificial heads exhibit rather detailed replicas of the average outer ear.

The positioning of the microphones considerably influences the spectral characteristics of the sound arriving at the microphones, which had to be compensated for a binaural reproduction. The characteristics of the acoustical path between the headphones and the microphones of the artificial head were compensated in the course of the equalization of the headphone transfer function (HPTFs), cf. section 7.2.4.

### 7.2.4 Headphone transfer functions (HPTFs)

All binaural test signals were presented via a D/A-converter (ADI-8 DS, RME Audio) and headphones (K-240 Studio, AKG Acoustics) to the subjects. Individual HPTFs were measured for each subject right after measuring the HRTFs with the earplug microphones still left in place (cf. section 7.2.1). It is well known from literature that HPTFs considerably depend on the individual placement/fit of the headphone (cf. Masiero (2012), Völk (2014)) and may lead to audible artifacts (cf. Paquier *et al.* (2011)), when equalizing HPTFs that exhibit narrowband spectral notches. In order to capture the associated variability, subjects were instructed to reposition the headphone ten times to various positions, yielding ten different - yet realistic - individual HPTFs. Out of the ten measured HPTFs, the individual HPTF resulting in the smallest dynamic range for frequencies $300\ \mathrm{Hz} \leq f \leq 16000\ \mathrm{Hz}$ was inverted according to the method given in Kirkeby and Nelson (1999) with the regularization parameter $\alpha_{\mathrm{inv}} = 0.01$. The regularized inversion was carried out after adjusting the root mean square-level of the individual HPTFs to -30 dB re. 1.

Analogously to the measurement and equalization of HPTFs with the headphones positioned on subjects, the described procedure was likewise applied to measure and invert the HPTFs associated with the traditional artificial heads using the built-in microphones (cf. Figure 7.4). It is worth noting that due to the different acoustical paths between the headphone and the microphones of the artificial heads, the HPTFs and consequently their equalization vary considerably with the used artificial heads. Therefore, the described method to achieve robust equalization filters for the HPTFs (tenfold repetition and inversion according to Kirkeby and Nelson (1999)) was also chosen for the three artificial heads to ensure that the quality of the various equalization filters was as constant as possible across all different devices/methods.

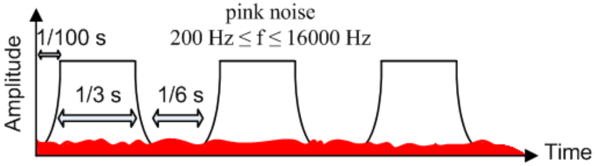


**Figure 7.2:** Temporal envelope of the noise bursts with an additional stationary white noise of 40 dB SPL.

### 7.2.5 Procedure and stimuli

Bursts of pink noise with a spectral content of 200 Hz $\leq f \leq$ 16000 Hz were chosen as the test signal for the evaluation experiment. Each noise burst lasted $\frac{1}{3}$ s with $\frac{1}{100}$ s onset-offset (raised cosine) ramps followed by silence of $\frac{1}{6}$ s (cf. Figure 7.2). This test stimulus was intended to facilitate the evaluation of spectral but also temporal aspects. The test signal was recorded with the VAH and with the traditional artificial heads in an anechoic room using spectrally-equalized loudspeakers placed in the six chosen directions in the horizontal plane. The binaural reproductions and VAH-synthesis presented via headphone were calibrated individually to have 75 dB SPL (G.R.A.S. type 43AA artificial ear) for the ipsilateral side at $\theta = 90^\circ$. During preliminary tests, it became evident that the different test devices exhibit characteristically different sensor noise floors (in level and in spectral coloration), which could potentially be recognized by the subjects in the evaluation of the various devices. To avoid an undesired detection of the devices due to the associated sensor noise, an additional

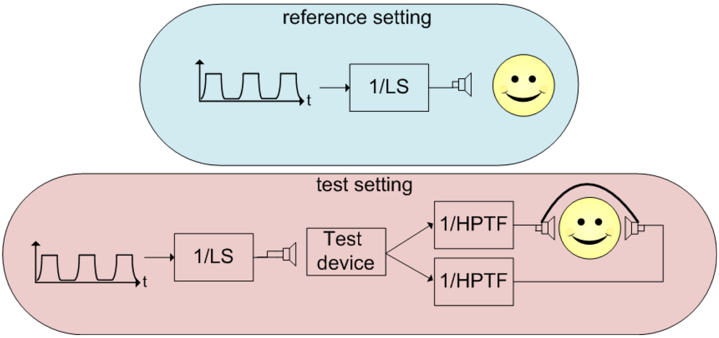


**Figure 7.3:** Used paradigm with the test signal associated with one of the five methods (Test device) played back via headphone in the test setting and via spectrally-equalized loudspeakers (1/LS) in the reference setting.

stationary white noise signal of 40 dB SPL was added to the noise bursts prior to the presentation (cf. red area in Figure 7.2).

In sum, binaural headphone reproductions using five methods were evaluated in the experiment with the free field playback in an anechoic room as the reference condition. The evaluated methods were

| | |
|---|---|
| HRTF | Individual binaural reproduction, where the test signal was filtered with individual HRTFs and equalized using individual HPTFs. |
| VAH | Individualized synthesis using the VAH including individually equalized HPTFs. |
| $DH_1$ | Binaural reproduction using artificial head $DH_1$ and equalized HPTFs measured with $DH_1$. |
| $DH_2$ | Binaural reproduction using artificial head $DH_2$ and equalized HPTFs measured with $DH_2$. |
| $DH_3$ | Binaural reproduction using artificial head $DH_3$ and equalized HPTFs measured with $DH_3$. |

These five methods (test setting, cf. Figure 7.3) were evaluated by the subjects with reference to the test signal presented via loudspeakers (reference setting, cf. Figure 7.3), where the subjects could switch between the test setting (presented in a hidden, randomized order) and the reference

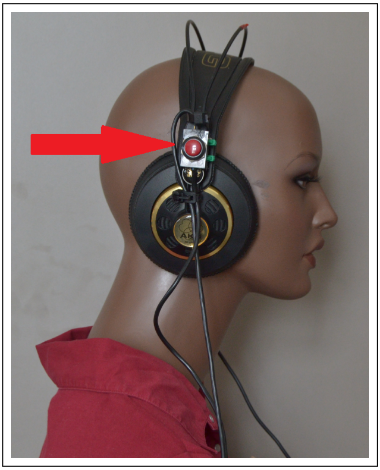


**Figure 7.4:** Used headphones seated on the $DH_3$ with a toggle button (cf. red arrow) that allows to switch between the test and the reference setting.

setting. The subjects could toggle between headphone presentation and loudspeaker presentation by pushing a toggle-button attached to the headphones (cf. arrow in Figure 7.4). Within each condition, the test signal was initially played via headphones (test setting) until the subject pushed the toggle-button. The signals were played in an infinite loop and subjects were able to pause playback using the stop button on the graphical user interface (GUI, cf. Figure 7.5). It was possible to sort the methods according to the entered ratings and hence to further compare the methods among each other by pushing the sort button (cf. Michaud *et al.* (2013)). Moreover, subjects were able to switch between the test and reference setting as often as desired.

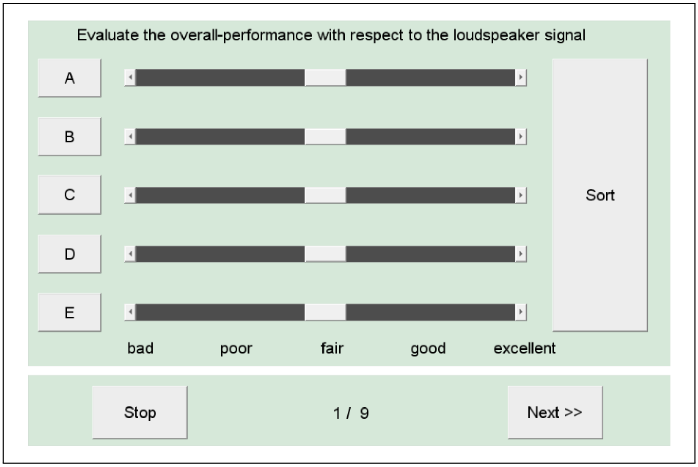

Evaluate the overall-performance with respect to the loudspeaker signal
A
B
C
D
E
Sort
bad
poor
fair
good
excellent
Stop
1 / 9
Next >>

**Figure 7.5:** GUI for evaluating the five methods with regard to loudspeaker presentation for each direction and perceptual quality.

The subjects were instructed to evaluate the perceptual qualities of the test settings regarding localization, spectral coloration and overall performance in three separate (subsequent) experiments with respect to the reference setting (test signal played via loudspeakers) on an English category scale ranging between *bad*, *poor*, *fair*, *good* and *excellent*. Each condition (unique combination of direction and tested method) was evaluated three times in a randomized order. One implementation of a test lasted approximately 30 minutes and subjects had time for familiarization before the test started. All subjects were encouraged to imagine the position of the sound source outside of the head (also when listening to the test signals over headphones during the familiarization phase), which was assumed to enhance the immersive reconstruction of a spatial scenario.

## 7.3 Results

The results[1] of the performed experiments are presented as boxplots (cf. Figure 7.6-7.11). The following description is divided into the results for explicitly considered directions (cf. section 7.3.1) and intermediate directions (cf. section 7.3.2) by the VAH.

### 7.3.1 Explicitly considered directions

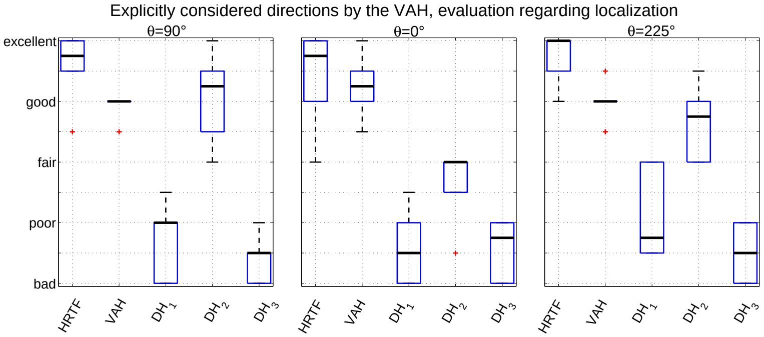


**Figure 7.6:** Aggregated ratings (y-axis) for all six subjects regarding the perception of localization for the five tested methods on the x-axis and three explicitly considered directions (left, middle and right panel).

The aggregated results (over all subjects) with regard to the perceived localization are illustrated in Figure 7.6 with the tested methods on the x-axis and the (explicitly considered) directions separated in three boxes. Consistent with the expectations, the best ratings are obtained for the HRTF-method followed by the individualized synthesis using the VAH. The median of the ratings of $DH_2$ is slightly better then the median of the ratings of the VAH-method at $\theta = 90°$. However, please note that the variance of the ratings across subjects is generally lower for the individualized methods (HRTF and VAH, except for $\theta = 0°$), while the ratings vary more clearly across the subjects for the traditional artificial heads ($DH_1$-$DH_3$). On average, the ratings for the traditional artificial heads are at or below

---

[1] After the implementation of the evaluation experiment, it was noticed that the (left ear) microphone membrane of $DH_1$ was perforated. This may be a possible reason for the relative bad ratings associated with $DH_1$ and should be considered in the discussion. For this reason, the results obtained for $DH_1$ were removed from the analysis in Rasumow *et al.* (2014e).

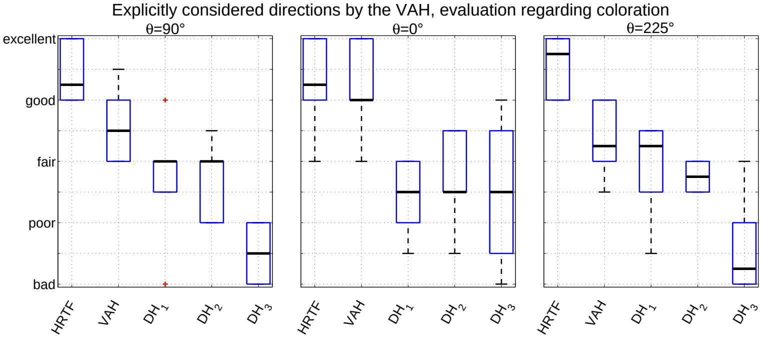


**Figure 7.7:** Aggregated results for all six subjects regarding the perception of spectral coloration for the five tested methods and three explicitly considered directions (left, middle and right panel).

the HRTF- and VAH-ratings, with the best ratings for the HRTF-method and the worst ratings for $DH_3$.

The aggregated ratings regarding spectral coloration are illustrated in Figure 7.7. Again, for the frontal direction the best ratings are obtained for the HRTF- and the VAH-method, and the rating of the VAH-synthesis is approximately equivalent to the HRTF-method. For the lateral direction $\theta = 90°$, the ratings for the VAH-method vary between *fair* and *excellent*, being slightly above the ratings for $DH_1$ and $DH_2$ and clearly above $DH_3$. The median of the ratings associated with the VAH-method at $\theta = 225°$ (approximately varying between *fair* and *good*) is similar to the median of the ratings of $DH_1$ (approximately varying between *poor* and *good*) and slightly above $DH_2$ and clearly above $DH_3$. In general, the ratings with regard to spectral coloration exhibit a more pronounced variability across subjects compared to the ratings with regard to localization.

The aggregated ratings with regard to the overall performance are illustrated in Figure 7.8. More clearly than for the ratings regarding localization and spectral coloration, the HRTF- and VAH-method are evaluated considerably better compared to the traditional artificial heads. For all three explicitly considered directions by the VAH, the median ratings of the VAH-method are *good* or better, whereas the median overall performance ratings of the traditional artificial heads are *fair* or worse. Again, the overall performance ratings of the binaural reproductions using traditional artificial heads (especially $DH_1$) are associated with a rather large inter-subject variance (including the whiskers of the boxplots), which indicates that the suitability of the traditional artificial heads varied notably

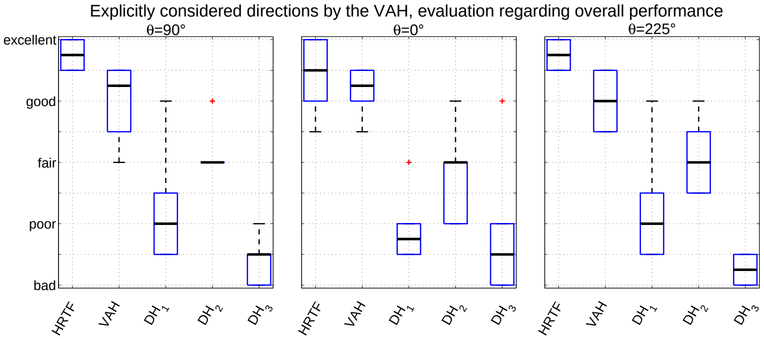


**Figure 7.8:** Aggregated results for all six subjects regarding the overall performance for the five tested methods and three explicitly considered directions (left, middle and right panel).

for the participating subjects.

### 7.3.2 Intermediate directions

The aggregated ratings with regard to localization for the three intermediate directions are illustrated in Figure 7.9. The HRTF-method results in the best (median) ratings, respectively followed by $\mathrm{DH}_2$ ($\theta = 97.5°$ and $\theta = 232.5°$) and the VAH-method ($\theta = 7.5°$). In contrast to the ratings for explicitly considered directions, the (median) localization ratings for in-

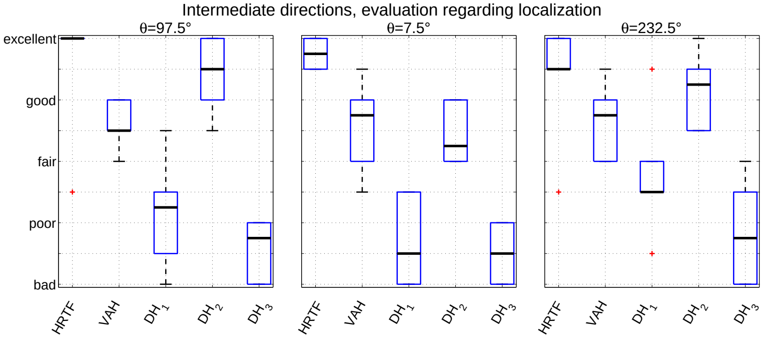


**Figure 7.9:** Aggregated ratings (y-axis) for all six subjects regarding the perception of localization for the five tested methods on the x-axis and three intermediate directions ($\theta = 7.5°$, $97.5°$ & $232.5°$).

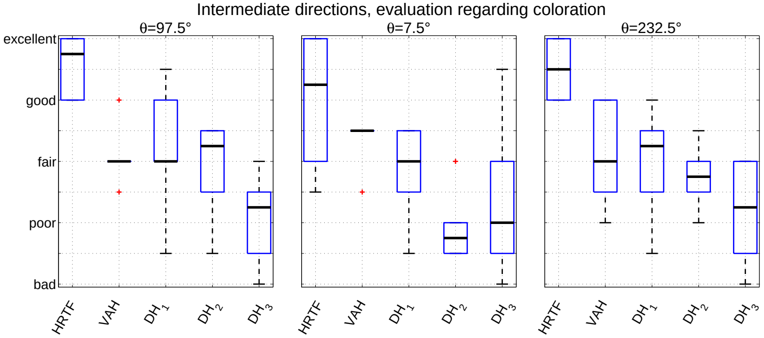


**Figure 7.10:** Aggregated results for all six subjects regarding the perception of spectral coloration for the five tested methods and three intermediate directions.

termediate directions for the VAH-method slightly drop between *fair* and *good* and exhibit a slightly larger variance across the subjects. The (median) localization ratings associated with $\mathrm{DH}_2$, however, unalteredly vary between *fair* and *excellent* (cf. Figure 7.6). Also the localization ratings associated with $\mathrm{DH}_1$ and $\mathrm{DH}_3$ are similar to those for explicitly considered directions, with the median ratings ranging approximately between *bad* and *fair*.

The aggregated ratings with regard to spectral coloration for the intermediate directions are illustrated in Figure 7.10. Again, the best median ratings are obtained for the HRTF-method, however, with a rather large inter-subject variance for $\theta = 7.5°$. This effect is also visible for $\theta = 0°$ (cf. Figure 7.7) and may presumably be explained by the difficulty of some subjects to externalize binaural reproductions for frontal directions (cf. Kim and Choi (2005)). The next-best spectral coloration ratings are obtained for the VAH-method ($\theta = 7.5°$), $\mathrm{DH}_2$ ($\theta = 97.5°$) and $\mathrm{DH}_1$ ($\theta = 232.5°$), with the median ratings at *fair* or slightly above. The worst (median) spectral coloration ratings are obtained for $\mathrm{DH}_3$ ($\theta = 97.5°$ and $\theta = 232.5°$) and $\mathrm{DH}_2$ ($\theta = 7.5°$), ranging approximately around *poor*.

The aggregated ratings with regard to the overall performance for the intermediate directions are illustrated in Figure 7.11. Similarly to the ratings for explicitly considered directions (cf. Figure 7.8), the overall performance ratings obtained for the HRTF-method are again followed by the ratings obtained for the VAH-method, with the median ratings ranging between *fair* and *good*. The next-best median ratings are obtained for $\mathrm{DH}_2$, also ranging between *fair* and *good*, yet constantly below the VAH-method. The worst (median) overall performance ratings are obtained for $\mathrm{DH}_1$ and $\mathrm{DH}_3$ and are constantly below *fair*.

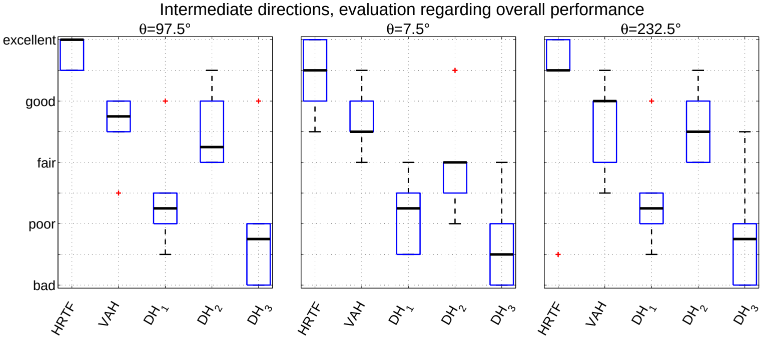


**Figure 7.11:** Aggregated results for all six subjects regarding the overall performance for the five tested methods and three intermediate directions.

Interestingly, the aggregated ratings with regard to spectral coloration are quite similar for $DH_1$ and $DH_3$, while $DH_2$ is generally evaluated better with regard to localization (cf. Figure 7.6 and 7.9) and overall performance (cf. Figure 7.8 and 7.11) compared to $DH_1$ and $DH_3$. This may probably be explained by the greater difficulty of some subjects to externalize the binaural reproductions associated with $DH_1$ and $DH_3$.

## 7.4 Discussion

In line with the expectations, the HRTF-method is constantly evaluated best across the tested methods in each condition, indicating the perceptual benefits of individual binaural reproductions. The next-best ratings vary between the VAH-method and one of the traditional artificial heads. To investigate the specific preferences of the subjects on average, the characteristics of the gathered ratings will be analyzed from a statistical point of view in the next paragraphs.

In general, the experimental data for explicitly considered directions (cf. Figure 7.6 – 7.8) and for intermediate directions (cf. Figure 7.9 – 7.11) show a large variance across the tested methods and across subjects. For the present study it is of primary interest to investigate whether the perceptual ratings vary significantly across the tested methods. A normality test (Lilliefors-test) indicated that the gathered experimental data cannot be assumed to be normally distributed, which may primarily be attributed to the rather low number of subjects. This is why non-parametric statistical tests are used for the further data analysis.

The Friedman-test was used to assess the effect of the tested methods on the ratings in each condition (unique combination of direction and evaluated aspect). In general, a sufficiently small p-value (resulting from the Friedman-test) indicates that the median value of the ratings for at least one method is significantly different from the median values associated with the other methods, which would suggest that there is a main effect due to the tested methods on the ratings. It is common practice to consider a result of the Friedman-test as statistically significant if the associated p-value is smaller than 0.05. However, to control the familywise error rate (probability of at least one type I error), the critical p-value for significance is usually adapted according to the number of the implemented tests (e.g. Bonferroni correction). Since the Friedman-test was implemented for 18 different conditions (3 evaluated aspects × 6 directions), only p-values smaller than $\frac{0.05}{18}$ are considered to indicate a significant effect of the tested methods and are indicated as bold numbers in Table 7.1.

| p-values | $\theta = 0°$ | $\theta = 7.5°$ | $\theta = 90°$ | $\theta = 97.5°$ | $\theta = 225°$ | $\theta = 232.5°$ |
|---|---|---|---|---|---|---|
| Localization | **0.0004** | **0.0002** | **0.0004** | **0.0006** | **0.0002** | **0.0008** |
| Coloration | 0.0189 | 0.0181 | **0.0003** | **0.0021** | **0.0013** | 0.0068 |
| Overall | **0.0020** | **0.0005** | **0.0003** | **0.0019** | **0.0002** | 0.0176 |

**Table 7.1:** p-values resulting from the Friedman test. Values associated with a significant effect of the tested methods on the ratings ($p \leq \frac{0.05}{18}$) are indicated as bold numbers.

The p-values in Table 7.1 indicate that the tested methods seem to have a significant effect on the ratings for a total of 14 conditions, which is in good agreement with the visual impression of the aggregated data in Figures 7.6–7.11. In many conditions, this effect is presumably due to the mainly large difference in ratings between the HRTF-method and $DH_3$, which is, however, not the effect of primary interest in this study. In contrast, one main aspect of interest in this study is to investigate the suitability of the VAH-method and hence to investigate the difference in ratings between the VAH-method and the method associated with the best rating when using a traditional artificial head (best of $DH_1$, $DH_2$ or $DH_3$). To do so, the best performances associated with a traditional artificial head were analyzed manually for the 14 conditions on the basis of the aggregated results from Figure 7.6–7.11. In general, $DH_2$ was determined (by visual inspection of the aggregated data in Figure 7.6–7.11) as the traditional artificial head associated with the best ratings. The Wilcoxon signed-rank test was then used as a post-hoc test within each of the 14 conditions to assess the difference in the mean ranks between the ratings for the VAH-method and $DH_2$. Analogously to the Friedman-test, a sufficiently small p-value resulting from the Wilcoxon signed-rank test indicates that the mean ranks of the two tested methods differ significantly from each other. Note that each of the 14 conditions (bold numbers in Table 7.1) was analyzed separately, where for each condition only one paired difference test was used and hence no correction of the significance level (0.05) was required. The Wilcoxon signed-rank test revealed four conditions indicating significant differences in the mean ranks between the VAH-method and $DH_2$: The results indicate a significant difference in the mean ranks of the ratings with regard to localization between the VAH-method and $DH_2$ for $\theta = 0°$ and $\theta = 97.5°$ ($p = \frac{1}{32} = 0.03125$). Evidently from the aggregated results in Figures 7.6 and 7.9, the ratings for the VAH-method are significantly better compared to the ratings associated with $DH_2$ for $\theta = 0°$ and significantly worse for $\theta = 97.5°$. This fact paradigmatically emphasizes the benefits of the VAH-method for explicitly considered directions but also indicates partial disadvantages of the VAH-method for some intermediate directions compared to traditional artificial heads. Regarding the ratings of the overall performance, the Wilcoxon signed-rank test indicates significant differences in mean ranks for $\theta = 0°$ and $\theta = 225°$ (between the VAH method and $DH_2$, both with $p = \frac{1}{32}$). Here the ratings obtained with the VAH-method are clearly above the ratings obtained with $DH_2$, indicating significantly better ratings of the VAH-method for both conditions. No significant differences in the mean ranks could be found for the ratings regarding spectral coloration between the VAH-method and $DH_2$.

In sum, the ratings emphasize the importance of individualization for binaural reproduction regarding all of the tested aspects (localization, spectral coloration and overall performance). The median ratings obtained with the individualized VAH-synthesis range mainly around *good* for explicitly considered directions and between *fair* and *good* for intermediate directions. In general, the worst results obtained for the VAH-method were gathered for the evaluation regarding spectral coloration, which is also known from previous studies (cf. Rasumow *et al.* (2014a)). On average, the ratings obtained for the VAH-method are better compared to traditional artificial heads, whereas for some conditions (primarily for intermediate directions) the localization is evaluated better for $DH_2$.

Regarding traditional artificial heads, on average $DH_2$ seems to yield the best ratings among the tested traditional artificial heads for the present paradigm. This is mainly assumed to be attributed to its better suitability for headphone equalization (by comparison rather flat HPTFs) compared to the other two artificial heads, which in turn is assumed to results from the very short/non-existing ear canals of $DH_2$.

## 7.5 Conclusion and future work

In conclusion, the gathered ratings emphasize the importance of individualization for binaural reproduction and confirm the validity of the concept to synthesize HRTFs in the horizontal plane using the proposed VAH. The results show that individual HRTFs in conjunction with individually equalized HPTFs result in the best subjective appraisals. The ratings obtained for the VAH-method indicate a high level of acceptance among the subjects. Especially the VAH-synthesis for explicitly considered directions yielded mainly *good* to *excellent* median ratings, which in general are above the ratings associated with traditional artificial heads. The ratings for intermediate directions seem to slightly worsen compared to the explicitly considered directions, however, still performing approximately at the same level (or slightly better regarding the overall performance) as binaural reproductions using traditional artificial heads.

The performance of the VAH for intermediate directions may presumably be enhanced by considering a finer directional grid within the optimization (e.g. by using interpolated steering vectors and HRTFs) and/or when using a higher WNG-constraint. Moreover, it is reasonable to assume that the microphone topology of the used microphone array may have a significant effect on the directivity pattern, also for intermediate directions. Based on this, it seems promising to investigate the performance of the VAH for intermediate directions with various microphone topologies, as for instance was used in Rasumow *et al.* (2013a).

# 8

# General Conclusions

In section 8.1, the main contributions of this thesis are summarized and conclusions are drawn. Moreover, considerations and suggestions for further research are given in section 8.2.

## 8.1 Conclusions

The main objective of this thesis was to develop a novel method to synthesize individual HRTFs in the horizontal plane using a microphone array and to analyze the perceptual importance of individual details of HRTFs that are needed for appropriate binaural reproduction. In this thesis, the general concept of the VAH is based on the optimization of individual filter coefficients by minimizing a least squares cost function using the measured steering vectors and HRTFs for discrete directions, assuming that the intermediate directions will be interpolated implicitly by the virtual artificial head. Several studies have been carried out that provided new insights into the optimization and regularization of multi-directional directivity patterns using microphone arrays and into the perceptual limits when smoothing individual HRTFs.

In **Chapter 2**, the impact of the topology of a two-dimensional microphone array on the accuracy and on the robustness of the synthesis when taking into account small deviations of the assumed steering vectors was investigated. A method to derive planar microphone topologies based on the so-called Golomb-ruler was proposed. The fundamental idea of this Golomb method is to obtain a microphone array topology with as many as possible different inter-microphone distances in all possible directions. Simulations showed that smaller synthesis errors were obtained for the Golomb method in comparison to the Random Sampling method. Moreover, also when taking into account random deviations of the assumed steering vectors the synthesis yielded smaller errors for the Golomb method, indicating the suitability of the proposed microphone topology for synthesizing HRTFs. Hence, the (open) planar microphone arrays used in the subsequent studies were implemented according to the proposed Golomb method. Interestingly, the performed simulations also indicated that the optimum number of microphones (in terms of the synthesis error) decreased for lower frequencies, which are associated with smoother desired directivity patterns. Thus, the results indicate that it may be beneficial to adapt the number of used microphones with respect to the frequency.

Since the used regularization by constraining the square norm of the filter coefficients turned out to be suboptimal for the synthesis of HRTFs, more appropriate approaches to enhance the robustness of the VAH-synthesis for small deviations of the steering vectors (e.g. gain, phase and/or positioning errors) were presented and evaluated objectively in **Chapter** 3.

Firstly, an optimization taking into account multiple measured sets of steering vectors was shown to enhance the robustness of the VAH-synthesis compared to a non-regularized optimization. However, an exemplary optimization taking into account only four measured sets of steering vectors resulted in a rather low robustness, especially for lower frequencies.

Secondly, using a mean weighting for the white noise gain ($\mathrm{WNG_m}$) in the regularization constraint was shown to enhance the VAH-synthesis and the associated robustness compared to using a white noise gain for a single direction. One major advantage of the mean weighting for the white noise gain is the equal consideration of the regularization for all directions, which enhances the synthesis considerably compared to a direction-dependent regularization constraint.

Thirdly, a design for incorporating neighbouring frequencies into the optimization and regularization ($\mathrm{WNG_v}$), mimicking the frequency grouping of the human auditory system, was presented. It was shown that the approach incorporating frequencies within ERBs resulted in the best robustness of the synthesis. A psychoacoustically motivated weighting of the white noise gain according to the magnitude of the desired directivity patterns was further shown to enhance the VAH-synthesis only when assuming very small/no deviations of the steering vectors.

In sum, the optimization and regularization for frequencies within relative bandwidths ($\mathrm{WNG_v}$) yielded the best overall results in this study and was consequentially also used in the concluding perceptual study in **Chapter 7**.

In **Chapter 4**, the discriminability when truncating individual hrirs in the time domain was examined. The main motivation of this discrimination experiment was the assumption that a microphone array achieves a more accurate fit to a desired directivity pattern if this pattern is spatially smooth. Since peaky directional characteristics of the HRTFs are also reflected in the frequency response, a discrimination experiment was performed to examine the appropriate length for individual hrirs that led to an unaltered perception compared to a reference setting with a length of 512 samples ($\approx$ 11.6 ms at a sampling frequency of 44100 Hz). It turned out that on average the subjects could discriminate between hrirs with lengths of 512 and $\approx$ 250 samples ($\approx$ 5.7 ms) at a 50% correct score, which was chosen as the threshold of interest. However, truncating the hrirs to psychoacoustically reasonable lengths prior to the VAH-synthesis only slightly enhanced the synthesis. This result, however, emphasized the general validity of smoothing individual HRTFs prior to the synthesis, which was examined in more detail in **Chapter 5**.

In **Chapter 5**, the discriminability when smoothing individual HRTFs in the frequency and spatial domains was examined.

Firstly, the discriminability when simplifying the phase spectrum of measured HRTFs was investigated. It was found that subjects are sensitive to a broadband phase linearization of HRTFs, whereas the original phase can be substituted by a linear phase for frequencies $f > 1000$ Hz without introducing noticeable artifacts. Further, this individual sensitivity to broadband phase linearizations could be predicted by a simple model that is based on interaural phase differences at frequencies $f \leq 1500$ Hz.

Secondly, the imperceptible phase linearization for higher frequencies was used to enable a complex-valued smoothing of HRTFs in relative bandwidths. It was shown that HRTFs may be smoothed into constant relative bandwidths of 1/5 octave without introducing noticeable artifacts, after the original phase was substituted by a linear phase for higher frequencies.

Thirdly, the discriminability when reducing the spatial dynamic range of HRTFs was examined. This aspect is of interest since it is assumed that spatially smooth HRTFs can be synthesized more accurately with the VAH. It turned out that spatial notches in the directivity pattern do not need to be retained in detail if they are less than 29 dB below the maximum value.

These perceptual limits permit to efficiently smooth individual HRTFs in the spectral and spatial domains, which was shown to clearly decrease the synthesis error of the VAH when smoothing the desired HRTFs prior to the synthesis.

In **Chapter 6**, the perceptual impact of the desired $\mathrm{WNG_m}$ on the appraisal of binaural reproductions using the VAH was examined. In general, the robustness of the synthesis increases with larger white noise gains. This, however, comes at the cost of decreasing the synthesis accuracy for the given HRTF set. In order to come up with a compromise between robustness and accuracy, the VAH-synthesis for different $\mathrm{WNG_m}$ was perceptually evaluated with respect to individual binaural reproductions using measured HRTFs. To gain further insight into the dependency of the perceptually preferred $\mathrm{WNG_m}$ for different microphone arrays, the evaluations were performed using two different microphone arrays with different sensor noise. It turned out that the perceived sensor noise was evaluated better with increasing $\mathrm{WNG_m}$, whereas the explicit presence of sensor noise did not seem to consistently deteriorate the ratings of the overall performance. The best compromise with regard to all assessed aspects varied approximately between $\mathrm{WNG_m} = -3$ and $\mathrm{WNG_m} = 0$ dB for the two used microphone arrays.

Interestingly, the ratings with regard to localization, coloration and overall performance decreased considerably for higher $\mathrm{WNG_m}$ when using the microphone array with the higher sensor noise. This phenomenon can be explained by the larger required Lagrange multipliers for the array with the higher sensor noise. In sum, the microphone array with the lower sensor noise yielded the better ratings and was further used in the concluding perceptual study in **Chapter 7**.

In **Chapter 7**, the individualized synthesis using the VAH, individual binaural reproductions using measured HRTFs and binaural reproduction using three traditional artificial heads were evaluated with respect to free field playback. The motivation of this study was to assess the suitability of the proposed principle to only consider measured steering vectors and HRTFs for discrete directions in the optimization of the filter coefficients and to assume the directivity pattern for intermediate directions to be interpolated implicitly. In order to enable the best possible synthesis of individual HRTFs with a VAH, the microphone topology was derived according to the findings in **Chapter 2**, regularization was applied considering the findings in **Chapter 3**, the regularization parameters were derived considering the findings in **Chapter 6** and the desired HRTFs were smoothed according to the findings in **Chapter 4** and **5** prior to the synthesis using the VAH. The binaural reproductions of all tested methods were evaluated for three explicitly considered directions in the horizontal plane and for three intermediate directions.

As a general result, binaural reproductions using individual HRTFs resulted in the best ratings, mostly followed by the VAH-synthesis for explicitly considered directions. The aggregated ratings obtained for the VAH-synthesis ranged between *fair* and *excellent* and were generally better than the ratings for the binaural reproductions using traditional artificial heads. However, the ratings obtained for the VAH-synthesis slightly decreased for intermediate directions, while still performing approximately at the same level (or slightly better regarding the overall performance) compared to binaural reproductions using traditional artificial heads.

In sum, the perceptual evaluations confirm the suitability of the VAH for synthesizing individual HRTFs and reveal its superiority over traditional artificial heads primarily for explicitly considered directions. However, there is still room for improvement, mainly with regard to intermediate directions, which may be enhanced by e.g. increasing the number of optimized directions (e.g. by interpolating the steering vectors and HRTFs), a more stringent regularization (i.e. a higher white noise gain) or a better suited microphone topology.

## 8.2 Suggestions for future research

In this section an overview over possible directions for further research are given.

The spatial alignment of the microphone array for the described procedure is inherently incorporated in the relative alignment between the steering vectors and the desired HRTFs within the optimization. Thus, the filter coefficients can also be optimized for various alignments of the listeners head. In other words, is becomes possible to virtually rotate the alignment of the listener just by changing the filter coefficients without changing the physical alignment/position of the microphone array. This feature is a great potential for further development, which could enable to consider head movements in already existing recordings, e.g. with the aid of the head tracking technology (dynamic binaural reproduction). This step is assumed to enhance the perceived quality of the binaural reproduction, e.g. with regard to externalization and localization (cf. Blauert (1997), Begault *et al.* (2000), Hirahara *et al.* (2011)). The general functionality of this feature has already been observed in informal listening tests. However, such a scenario would need further research of the spatial resolution required within a dynamic binaural reproduction, the interpolation of desired HRTFs and steering vectors for intermediate directions and the impact of the delays resulting from the signal processing chain and the head tracking device. General investigations concerning the latter two aspect were presented in Völkering *et al.* (2014), Völkering (2014).

Moreover, the proposed HRTF-synthesis in the horizontal plane can be expanded to also take into account elevated sources in order to enable an omnidirectional spatial perception. In general, this can be incorporated by extending the cost function (e.g. equation 1.8 or 3.13) by summing over elevation angles. However, this would require further investigations regarding the suitability of the developed microphone topologies and cost functions. Furthermore, it is known from literature that elevated directions are mainly associated with spectral cues provided by the pinna (cf. Blauert (1997), Macpherson and Sabin (2007)), which incorporate individually varying characteristic spectral peaks and notches (cf. Wright *et al.* (1974), Raykar *et al.* (2005), Iida (2008)). At the same time, the performed evaluations reveal that the VAH-synthesis is evaluated more critically with regard to spectral coloration (cf. **Chapter 6** and **7**), indicating perceptual shortcomings in the synthesis of spectral details. Hence, further psychoacoustical research regarding the smoothing of HRTFs including elevated directions prior to the synthesis may be required.

Using a narrowband optimization, it would be ineffective to apply a frequency-dependent weighting of the cost function (cf. $F$ in equation 1.8). In **Chapter** 3, a method for incorporating multiple frequencies into the optimization and regularization of filter coefficients was proposed. In this thesis, only the filter coefficients of the center frequency of the associated (optimized) frequency band were considered. However, this method may also be used for optimizing filter coefficients in separate adjacent frequency bands. In this case, a psychoacoustically motivated weighting of the cost function over directions and frequencies may bear a further possibility for improvement, which would, however, need further psychoacoustical investigations.

The simulations in **Chapter 2** indicate that the optimum number of microphones with regard to the synthesized accuracy changes with the frequency. Hence, it may be beneficial to use only a subset of the available microphones at lower frequencies, which not only would smooth the synthesized spatial directivity pattern but also simultaneously may enhance the robustness. Moreover, it may be advantageous to examine the perceptual impact associated with a frequency-dependent regularization constraint. In the previous experiments, the desired (mean) white noise gains were implemented constantly over all frequencies, whereas it may be beneficial to vary the (mean) white noise gains within certain frequency bands or alike.

The performed evaluations in **Chapter 6** and informal listening tests indicate that the inherent sensor noise of the used microphones is an important aspect for the appraisal of the VAH-synthesis. Furthermore, the results from **Chapter 6** showed, that a lower sensor noise of the used microphones is associated with smaller Lagrange multipliers $\mu$. Considering that the synthesis accuracy decreases with larger Lagrange multipliers $\mu$, a lower sensor noise of the microphones may further enhance the VAH-synthesis. Hence, it may be beneficial to further reduce the sensor noise, as for instance was done for array$_2$ (cf. right panel in Figure 6.2). Further, it should be examined whether the sensor noise negatively affects the measurement of the steering vectors. In this case, it may be beneficial to appropriately smooth the steering vectors, however without altering their relevant characteristics.

Moreover, an important further research topic is the VAH-synthesis for intermediate directions. It is worth noting that the mainly *fair* to *good* ratings in **Chapter 7** for the intermediate directions could only be reached in conjunction with a rather high WNG-constraint (desired $\mathrm{WNG_v}$=2 dB). This may be explained by a rather smooth spatial interpolation for higher

WNG-constraints. Such high WNG-constraints may, however, at the same time slightly deteriorate the performance for explicitly considered directions (cf. **Chapter 6**). Hence, an alternative procedure to achieve a more appropriate interpolation of the VAH-synthesis for intermediate directions may allow for lower WNG-constraints and consequentially enhance the overall performance of the VAH. In general, the performance of the VAH-synthesis for intermediate directions may, for instance, be enhanced by explicitly considering more directions in the optimization (i.e. a finer spatial grid within the optimization, e.g. by interpolation of the measured steering vectors and HRTFs) or by a more appropriate microphone topology (e.g. considering recent studies as Tourbabin and Rafaely (2014)).

# A

# Estimation of the delay and the resulting linear phase $\phi_{lin}(f)$

The main idea of this method is that the estimated linear phase should maintain the delay of the largest magnitude (loudest part) of a *hrir* (cf. Figure 5.10). Consequently, we estimated the delay of the largest magnitude associated with the largest magnitude of the Hilbert envelope of a *hrir*. The Hilbert envelope of a *hrir* is defined as the magnitude of the analytic signal, which was calculated using the function **hilbert** in MATLAB (with the *hrir*-length of $n = 512$ samples and $f_s = 44100$ Hz). The analytic signal is calculated in a three-step algorithm: Firstly, the $n$-point fast Fourier transformation (FFT) of the *hrir* is calculated. Secondly, this FFT-sequence is multiplied with the factor of 2 for frequency bins $i$ ranging from $2 \leq i \leq (n/2)$ and with the factor of 0 for frequency bins ranging from $(n/2) + 2 \leq i \leq n$. Thirdly, the inverse fast Fourier transformation (IFFT) of the manipulated spectrum is calculated, where the first $n$ elements represent the analytic signal.

An exemplary *hrir* and its corresponding Hilbert envelope are depicted in Figure A.1 as a function of time in samples. The maximum of the Hilbert envelope was used to define the estimated delay $\tau$ of the *hrir*. Based on this approach, the estimated linear phase was given by

$$\phi_{lin}(f) = e^{-i \cdot 2\pi \cdot \frac{f}{f_s} \cdot \tau}, \tag{A.1}$$

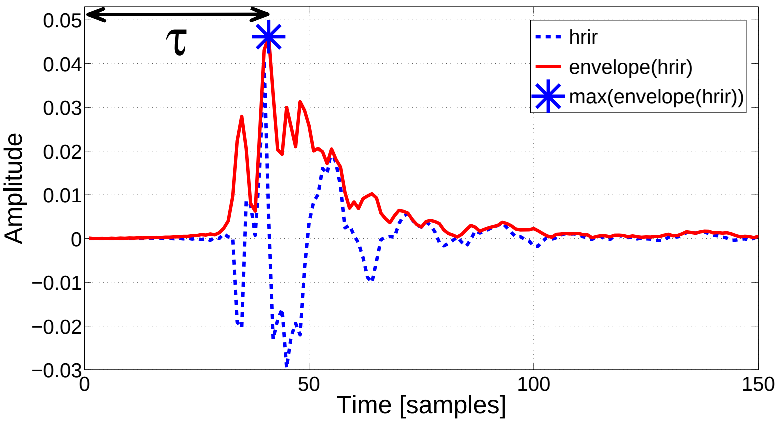


**Figure A.1:** An exemplary *hrir* is depicted (dashed blue line) as a function of time in samples on the x-axis. The red solid line depicts the corresponding Hilbert envelope of this *hrir*. The blue asterisk is the maximum of the Hilbert envelope, characterizing the estimated delay $\tau$ in samples.

with $f$ being the discrete frequency vector (with the length of $n = 512$ bins) equidistantly ranging from zero to $f_s$.

# B

# Applied method for the spectral smoothing of HRTFs

Initially the delay of the measured *hrir* is estimated (step I in Figure B.1) according to the method as described in Appendix A. This step yields a linear phase $\phi_{lin}$, which is used to continue the phase of the HRTF for frequencies $f \geq 5$ kHz (step II). Generally, a smoothing of HRTFs into constant relative bandwidths in the frequency domain is analogous to a truncation of the *hrirs* in the time domain using frequency-dependent truncation lengths. Hence, in order to ensure that the smoothing algorithm processes the main parts of the *hrirs*, it is important to remove the overall delay $\tau$ before smoothing. Analogous to equation. A.1, the removal of the delay $\tau$ is done in step III by multiplying the HRTF with the phase term $e^{i \cdot 2\pi \cdot \frac{f}{f_s} \cdot \tau}$. It is worth noting that due to the preprocessing in step I-III the phase of the HRTF is set to zero for frequencies $f \geq 5$ kHz (cf. phase in step III in Figure B.1). By this means the phase manipulations in step I-III reduce a complex-valued HRTF to a real-valued HRTF for higher frequencies, while the phase characteristics for the lower frequencies (except for the delay $\tau$) are maintained. In step IV the so-preconditioned HRTF is complexly smoothed according to the algorithm proposed in Hatziantoniou and Mourjopoulos (2000) with the bandwidth $\mathrm{B_W}$. In the final step (V) the original delay $\tau$ of the measured HRTF is reconstructed. Analogous to the previous procedure, the delay $\tau$ is reconstructed by multiplying the smoothed HRTF with the phase term $e^{-i \cdot 2\pi \cdot \frac{f}{f_s} \cdot \tau}$.

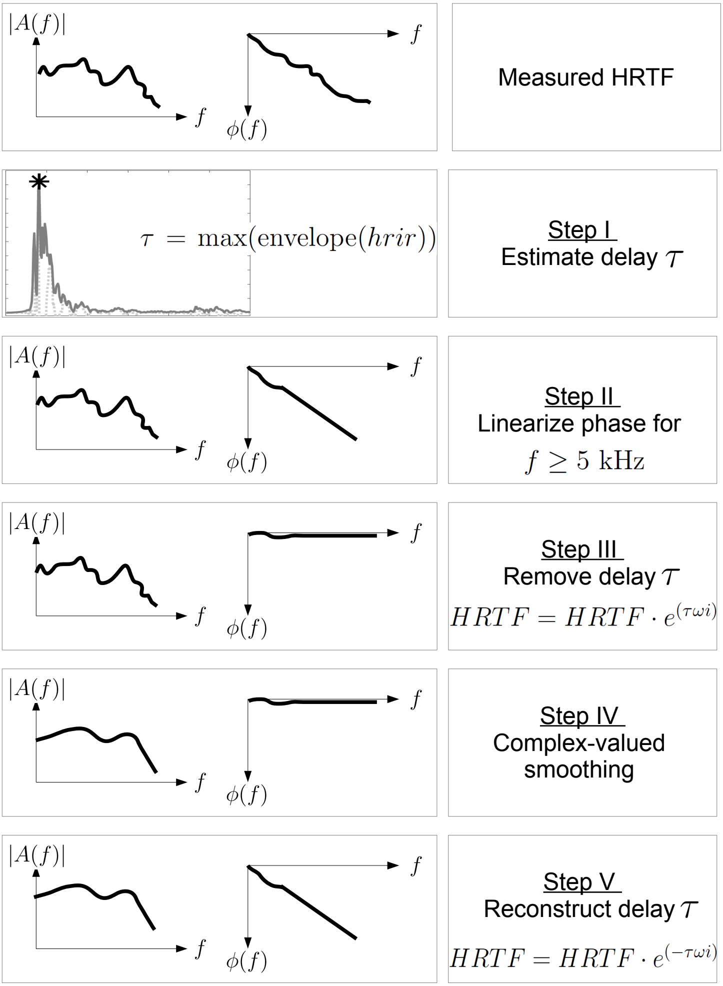


**Figure B.1:** Block diagram characterizing the procedure for the applied complex-valued smoothing of HRTFs. The process can be divided into five separate steps: In step I the delay $\tau$ is estimated as described in Appendix A. The resulting linear phase is applied to the HRTF for frequencies $f \geq 5$ kHz in step II. In step III the delay $\tau$ is removed from the measured HRTF. In step IV the preconditioned HRTF is smoothed complexly according to the algorithm proposed in Hatziantoniou and Mourjopoulos (2000). In step V the initial delay $\tau$ is reconstructed for the smoothed HRTF.

# C

# Measuring apparatus

| Apparatus used for measuring | |
|---|---|
| Device | Description |
| Artificial ear | G.R.A.S. Type 43AA Artificial Ear |
| Calibrator | Norsonic Type 1251 114 dB |
| Converter | AD/DA converter ADI-8 DS RME Audio |
| Free field microphone | G.R.A.S. Microphone Type 40AF |
| Headphones | AKG K-240 Studio |
| Loudspeakers | Lasmex S-01 |
| | Custom-made one-way loudspeakers with $\approx$ 7 cm diameter (cf. Figure C.1) |
| PC Interface | ADI-648 Multichannel Audio Digital Interface RME Audio |
| Software | MATLAB R2012b |
| | Psylab (`http://tgm.jade-hs.de/web/file/psylab.php`) |
| VAH-microphones | Knowles FG-23329 Miniature Electret Microphones (cf. left panel in Figure 6.2) |
| | Analog Devices ADMP 504 Ultralow Noise Microphone (cf. right panel in Figure 6.2) |

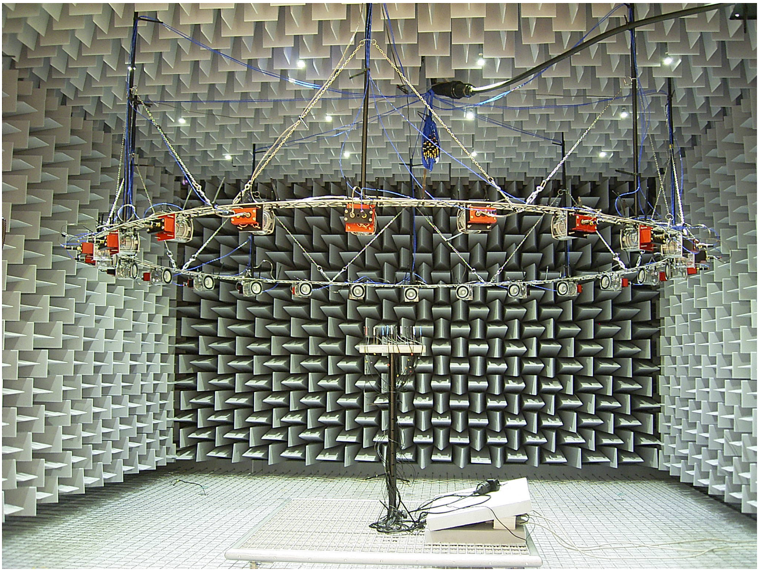


**Figure C.1:** Circular loudspeaker array in the anechoic room at the Institut für Hörtechnik und Audiologie in Oldenburg

# D

# Causality of optimized filter coefficients

A straightforward optimization of the filter coefficients as, for instance, described in section 1.5 may result in noncausal filter coefficients in the time domain, if the used steering vectors and the desired hrirs exhibit similar delays. This aspect is paradigmatically depicted in Figure D.1, with the steering vectors for N=4 microphones and direction $\theta = 0^\circ$ (top), the desired hrir for $\theta = 0^\circ$ (center) and the resulting filter coefficients in the time domain (bottom). Such filter coefficients would, however, need further treatment to result in causal filter coefficients and may bear a possible source of error. A possible way to avoid noncausal filter coefficients is to delay the desired hrirs, as depicted in Figure D.2.

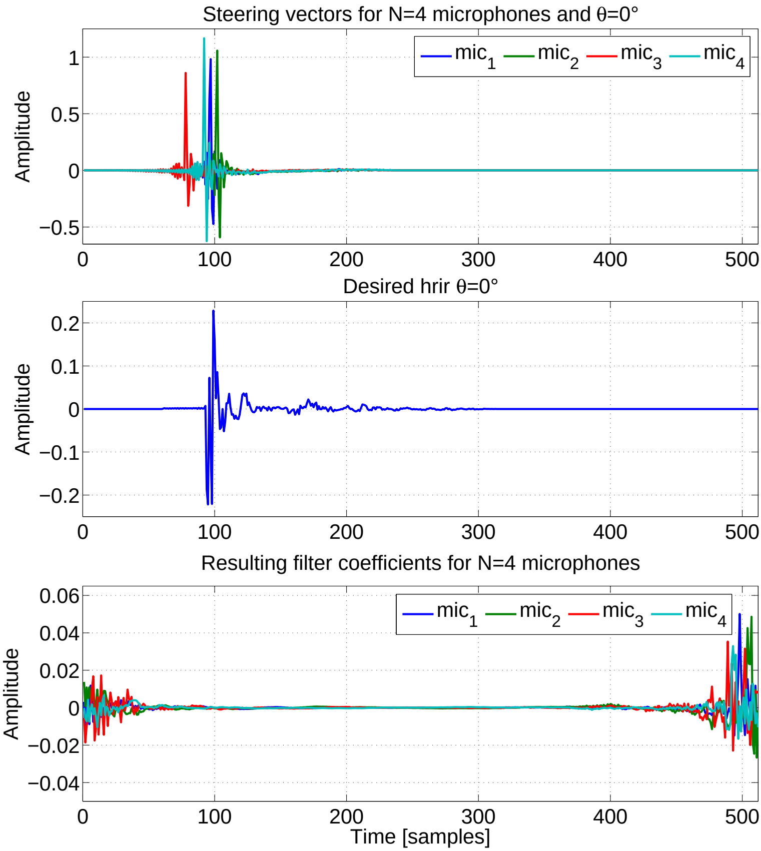


**Figure D.1:** Steering vectors for N=4 microphones (top), the desired hrir for $\theta = 0^\circ$ and the resulting noncausal filter coefficients (bottom) as a function of time in samples ($f_s = 44100$ Hz).

The described optimization procedures in this thesis were based on a NFFT=512. Based on the discrimination study in **Chapter 4**, the individual hrirs may be truncated to 256 samples. Consequently, it is appropriate to truncate the desired hrirs to 256 samples and to delay the desired hrirs by 256 samples (center panel in Figure D.2), which results in causal filter coefficients that do not need further treatment.

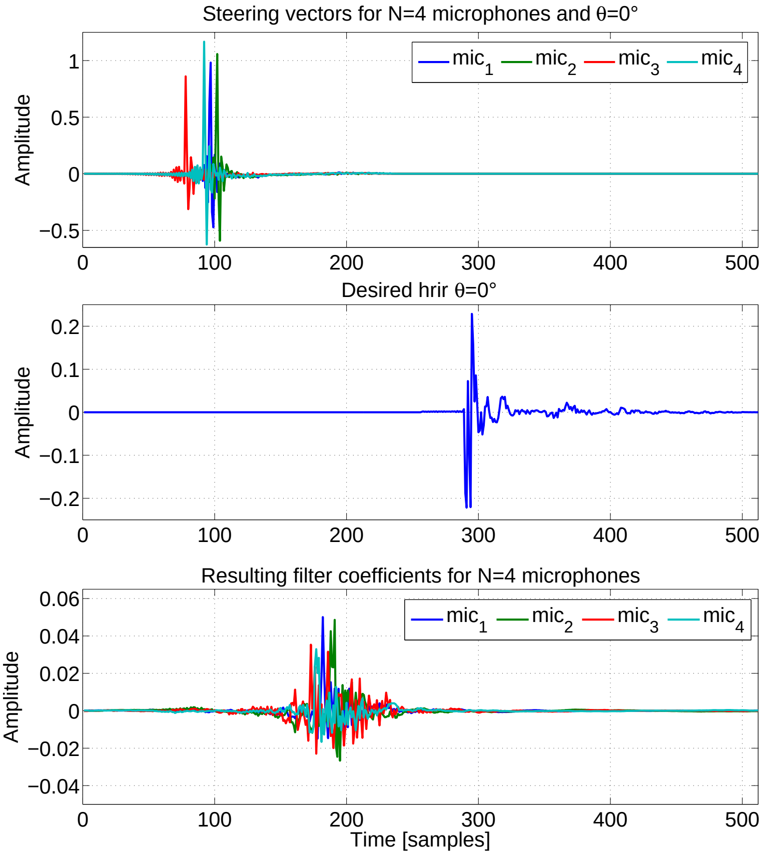


**Figure D.2:** Steering vectors for N=4 microphones (top), the delayed desired hrir for $\theta = 0°$ and the resulting causal filter coefficients (bottom) as a function of time in samples ($f_s = 44100$ Hz).

# E

# Imposing a constraint on the least squares cost function

As described in **Chapter 2-6**, it is important to regularize the filter coefficients in order to yield a robust VAH-synthesis. A common way to increase the robustness is to impose a constraint on the cost function. Based on the least squares cost function in equation 1.8, it is reasonable to impose a constraint[1] on the $\mathrm{WNG_m}$ from equation 3.8, i.e. $\frac{\mathbf{w}^H(f)\mathbf{Q}_{\mathrm{m}}(f)\,\mathbf{w}(f)}{\mathbf{w}^H(f)\mathbf{w}(f)} \geq \beta_{\mathrm{m}}$. This results in the constrained optimization problem

$$\boxed{\min_{\mathbf{w}(f)} \mathrm{J_{LS}}(\mathbf{w}(f)) \quad \text{subject to} \quad \frac{\mathbf{w}^H(f)\mathbf{Q}_{\mathrm{m}}(f)\,\mathbf{w}(f)}{\mathbf{w}^H(f)\mathbf{w}(f)} \geq \beta_{\mathrm{m}}}\,, \tag{E.1}$$

with $\beta_{\mathrm{m}}$ the minimum desired value for $\mathrm{WNG_m}$. The Lagrange function associated with this constrained optimization problem is equal to

$$\begin{aligned}\mathrm{J}_{\mathrm{m}_1}(\mathbf{w}(f)) = &\sum_{i=1}^{\mathrm{P}} F(f,\theta_i)\cdot\left|\mathbf{w}^H(f)\mathbf{d}(f,\theta_i) - \mathrm{D}(f,\theta_i)\right|^2 \\ &+\mu\cdot\left(\frac{\mathbf{w}^H(f)\mathbf{Q}_{\mathrm{m}}(f)\,\mathbf{w}(f)}{\mathbf{w}^H(f)\mathbf{w}(f)} - \beta_{\mathrm{m}}\right),\end{aligned} \tag{E.2}$$

with the Lagrange multiplier $\mu$. This cost function may be solved by setting the gradient $\nabla_{\mathbf{w}(f)}\mathrm{J}_{\mathrm{m}_1}(\mathbf{w}(f))$ to zero. As a disadvantage, $\nabla_{\mathbf{w}(f)}\mathrm{J}_{\mathrm{m}_1}(\mathbf{w}(f))$ cannot easily be derived analogously to the gradient of the initial least squares cost function. However, this may be substantially simplified by reformulating the constraint to $\mathbf{w}^H(f)\mathbf{w}(f) \leq \frac{1}{\beta_{\mathrm{m}}}\,\mathbf{w}^H(f)\mathbf{Q}_{\mathrm{m}}(f)\mathbf{w}(f)$, resulting in the Lagrange function

$$\begin{aligned}\mathrm{J}_{\mathrm{m}_2}(\mathbf{w}(f)) = &\sum_{i=1}^{\mathrm{P}} F(f,\theta_i)\cdot\left|\mathbf{w}^H(f)\mathbf{d}(f,\theta_i) - \mathrm{D}(f,\theta_i)\right|^2 \\ &+\mu\cdot\left(\mathbf{w}^H(f)\mathbf{w}(f) - \frac{1}{\beta_{\mathrm{m}}}\,\mathbf{w}^H(f)\mathbf{Q}_{\mathrm{m}}(f)\mathbf{w}(f)\right).\end{aligned} \tag{E.3}$$

This cost function can equivalently be stated as

$$\begin{aligned}\mathrm{J}_{\mathrm{m}_2}(\mathbf{w}(f)) = &\,\mathbf{w}^H(f)\mathbf{Q}(f)\mathbf{w}(f) - \mathbf{w}^H(f)\mathbf{a}(f) - \mathbf{a}^H(f)\mathbf{w}(f) + d(f) \\ &+\mu\cdot\left(\mathbf{w}^H(f)\mathbf{w}(f) - \frac{1}{\beta_{\mathrm{m}}}\,\mathbf{w}^H(f)\mathbf{Q}_{\mathrm{m}}(f)\mathbf{w}(f)\right),\end{aligned} \tag{E.4}$$

with $\mathbf{Q}(f)$, $\mathbf{a}(f)$ and $d(f)$ given in equation 1.10. The gradient of this cost

---

[1] In this section, an example imposing a constraint on the $\mathrm{WNG_m}$ is described. However, an analogous procedure is also possible for the other described constraints.

function $\nabla_{\mathbf{w}(f)} \mathrm{J}_{\mathrm{m}_2}(\mathbf{w}(f))$ is then given by

$$\nabla_{\mathbf{w}(f)} \mathrm{J}_{\mathrm{m}_2}(\mathbf{w}(f)) = 2\,\mathbf{Q}(f)\mathbf{w}(f) - 2\,\mathbf{a}(f) + \mu \cdot \left(2\,\mathbf{w}(f) - \frac{2}{\beta_{\mathrm{m}}}\,\mathbf{Q}_{\mathrm{m}}(f)\mathbf{w}(f)\right). \tag{E.5}$$

Setting $\nabla_{\mathbf{w}(f)} \mathrm{J}_{\mathrm{m}_2}(\mathbf{w}(f))$ to zero and resolving it into $\mathbf{w}(f)$ leads to the known solution, given in equation 3.11.

# Bibliography

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (**2001**), "The CIPIC HRTF Database," in *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2001)*, New Paltz, NY, USA, pp. 99–102. (Cited on page 86)

Atkins, J. (**2011**a), "Robust beamforming and steering of arbitrary beam patterns using spherical arrays," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 237 – 240. (Cited on pages 23, 34, 47, 74, 84, 88, 120, 121, 136, and 137)

Atkins, J. (**2011**b), "Spatial Acoustic Signal Processing For Immersive Communication," PhD thesis, Johns Hopkins University, Baltimore, MD. (Cited on pages 23, 47, 74, and 137)

Begault, D. (**2000**), "3-D Sound for Virtual Reality and Multimedia," Tech. rep., National Aeronautics and Space Administration, Ames Research Center Moffett Field, California 94035. (Cited on pages xvii, 15, and 16)

Begault, D. R., Lee, A. S., Wenzel, E. M., and Anderson, M. R. (**2000**), "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," in *Audio Engineering Society Convention 108*, URL `http://www.aes.org/e-lib/browse.cfm?elib=9204`. (Cited on page 158)

Belendiuk, K. and Butler, R. A. (**1975**), "Monaural localization of low-pass noise bands in the horizontal plane," *The Journal of the Acoustical Society of America* **58**(3), pp. 701–705, URL `http://scitation.aip.org/content/asa/journal/jasa/58/3/10.1121/1.380717`. (Cited on page 12)

Bendat, J. S. and Piersol, A. G. (**1986**), "Random data: Analysis and measurement procedures," Wiley, chap. 4, pp. 74–109. (Cited on page 63)

Benesty, J., Sondhi, M. M., and Huang, Y. (Eds.) (**2008**), *Springer Handbook of Speech Processing*, Springer, Berlin. (Cited on page 21)

Bitzer, J. and Simmer, K. U. (**2001**), "Superdirective Microphone Arrays," in *Microphone Arrays - Signal Processing Techniques and Applications*, edited by M. S. Brandstein and D. B. Ward, Springer Verlag, pp. 19–38. (Cited on pages 22, 25, 46, 47, 50, 51, and 122)

Blau, M., Sankowsky-Rothe, T., Köhler, S., and Schmidt, J.-H. (**2013**), "Using inter-individual standard deviation of hearing thresholds as a criterion to compare methods aimed at quantifying the acoustic input to the human auditory system in occluded ear scenarios," *Proceedings of Meetings on Acoustics* **19**(1), URL `http://scitation.aip.org/content/asa/journal/poma/19/1/10.1121/1.4801024`. (Cited on page 16)

Blauert, J. (**1997**), *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, URL `http://books.google.de/books?id=wBiEKPhw7r0C`. (Cited on pages 12, 26, and 158)

Brandstein, M. and Griebel, S. (**2001**), *Microphone Arrays*, Digital Signal Processing, Springer Berlin Heidelberg. (Cited on page 21)

Breebaart, J., Nater, F., and Kohlrausch, A. (**2010**), "Spectral and Spatial Parameter Resolution Requirements for Parametric, Filter-Bank-Based HRTF Processing," *Journal of the Audio Engineering Society* **58**(3), pp. 126–140. (Cited on page 86)

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**), "Binaural processing model based on contralateral inhibition. I. Model structure," *The Journal of the Acoustical Society of America* **110**(2), pp. 1074–1088, URL `http://scitation.aip.org/content/asa/journal/jasa/110/2/10.1121/1.1383297`. (Cited on page 26)

Castaneda, C. D. S., Sakamoto, S., Lopez, J. A. T., Li, J., Yan, Y., and Suzuki, Y. (**2013**), "Accuracy of head-related transfer functions synthesized with spherical microphone arrays," in *Proceedings of the 21st International Congress on Acoustics, Montreal (Canada)*, ASA, vol. 19. (Cited on pages 23, 88, 120, and 137)

Chen, J., Veen, B. D. V., and Hecox, K. E. (**1992**), "External ear transfer function modeling: A beamforming approach," *The Journal of the Acoustical Society of America* **92**(4), pp. 1933–1944. (Cited on pages 47, 74, 84, and 136)

Ćirić, D. G. and Hammershøi, D. (**2006**), "Coupling of earphones to human ears and to standard coupler," *The Journal of the Acoustical Society of America* **120**(4), pp. 2096–2107, URL `http://scitation.aip.org/content/asa/journal/jasa/120/4/10.1121/1.2258929`. (Cited on page 19)

Cox, H., Zeskind, R., and Kooij, T. (**1986**), "Practical supergain," *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**(3), pp. 393–398. (Cited on pages 25, 37, 46, 47, 50, and 51)

Cox, H., Zeskind, R. M., and Owen, M. M. (**1987**), "Robust Adaptive Beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-35**(10), pp. 1365–1376. (Cited on page 38)

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**), "A quantitative model of the effective signal processing in the auditory system. I.Model structure," *The Journal of the Acoustical Society of America* **99**(6), pp. 3615–3622. (Cited on page 26)

DE 10 2010 012 388 A1 (), *Sender- bzw. Sensoranordnung zur Erzielung optimierter Richtcharakteristik für Sende- oder Empfangsvorrichtungen in zwei und drei Dimensionen.* (Cited on pages xviii, 36, and 37)

Dietz, M., Ewert, S. D., and Hohmann, V. (**2011**), "Auditory Model Based Direction Estimation of Concurrent Speakers from Binaural Signals," *Speech Commun.* **53**(5), pp. 592–605, URL `http://dx.doi.org/10.1016/j.specom.2010.05.006`. (Cited on page 26)

Dobrucki, A., Plaskota, P., Pruchnicki, P., Pec, M., Bujacz, M., and Strumillo, P. (**2010**), "Measurement System for Personalized Head-Related Transfer Functions and Its Verification by Virtual Source Localization Trials with Visually Impaired and Sighted Individuals," *Journal of the Audio Engineering Society* **58**(9), pp. 724–738, URL `http://www.aes.org/e-lib/browse.cfm?elib=15518`. (Cited on page 17)

Doclo, S., Gannot, S., Moonen, M., and Spriet, A. (**2010**), "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, edited by S. Haykin and K. J. Ray Liu, Wiley, chap. 9, pp. 269–302. (Cited on page 21)

Doclo, S. and Moonen, M. (**2003**a), "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing* **83**(12), pp. 2641–2673. (Cited on pages xviii, 22, 23, 46, 48, and 74)

Doclo, S. and Moonen, M. (**2003**b), "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Transactions on Signal Processing* **51**(10), pp. 2511–2526. (Cited on pages 24, 25, 31, 37, 39, 47, 50, 121, and 122)

Doclo, S. and Moonen, M. (**2007**), "Superdirective Beamforming Robust Against Microphone Mismatch," *IEEE Transactions on Audio, Speech, and Language Processing* **15**(2), pp. 617–631. (Cited on pages 25, 37, 47, and 50)

Dörbecker, M. (**1998**), "Mehrkanalige Signalverarbeitung zur Verbesserung akustisch gesrörter Sprachsignale am Beispiel elektronischer Hörhilfen," Ph.D. thesis, Verlag Mainz in Aachen. (Cited on page 38)

Duraiswaini, R., Zotkin, D., and Gumerov, N. (**2004**), "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 4, pp. iv–45 – iv–48. (Cited on page 88)

Enzner, G. (**2009**), "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09*, pp. 325–328. (Cited on page 17)

Fletcher, H. (**1940**), "Auditory Patterns," *Rev. Mod. Phys.* **12**, pp. 47–65. (Cited on pages 27, 53, 55, and 86)

Glasberg, B. R. and Moore, B. C. J. (**1990**), "Derivation of auditory filter shapes from notched-noise data," *Hearing Research* **47**, pp. 103–138. (Cited on pages 90, 100, 104, and 107)

Golomb, S. and Taylor, H. (**1982**), "Two-dimensional synchronization patterns for minimum ambiguity," *IEEE Transactions on Information Theory* **28**(4), pp. 600–604. (Cited on page 36)

Hacker, M. J. and Ratcliff, R. (**1979**), "A revised table of d' for M-alternative forced choice," *Perception & Psychophysics* **26**(2), pp. 168–170. (Cited on page 94)

Hammershøi, D. and Hoffmann, P. F. (**2011**), "Control of earphone produced binaural signals," in *Proc. Forum Acusticum 2011, Aalborg (Denmark)*, pp. 2235–2239. (Cited on page 19)

Hammershøi, D. and Møller, H. (**1991**), "Free-Field Sound Transmission to the External Ear: A Model and some Measurements," in *Fortschritte der Akustik, Deutsche Arbeitsgemeinschaft für Akustik. DPG-GmbH*, Dega, Bochum, pp. 473–476. (Cited on pages )

Hammershøi, D. and Møller, H. (**1996**), "Sound transmission to and within the human ear canal," *The Journal of the Acoustical Society of America* **100**(1), pp. 408–427. (Cited on pages 16, 57, 75, 92, and 125)

Hammershøi, D. and Sandvad, J. (**1994**), "Binaural Auralization, Simulating Free Field Conditions by Headphones," in *Audio Engineering*

*Society Convention 96*, URL `http://www.aes.org/e-lib/browse.cfm?elib=6369`. (Cited on pages )

Hansen, M. (**2006**), "Lehre und Ausbildung in Psychoakustik mit psylab: Freie Software für psychoakustische Experimente," in *Fortschritte der Akustik – DAGA '06*, Dega, Braunschweig, pp. 591–592. (Cited on page 77)

Hartmann, W. M. and Wittenberg, A. (**1996**), "On the externalization of sound images," *The Journal of the Acoustical Society of America* **99**(6), pp. 3678–3688, URL `http://scitation.aip.org/content/asa/journal/jasa/99/6/10.1121/1.414965`. (Cited on pages 10, 12, and 46)

Hatziantoniou, D. P. and Mourjopoulos, J. N. (**2000**), "Generalized Fractional-Octave Smoothing of Audio and Acoustic Responses," *Journal of the Audio Engineering Society.* **48**(4), pp. 259–280. (Cited on pages xxiv, 90, 95, 166, and 167)

Haut, C. (**2009**), "Entwicklung einer adaptiven Mikrofonanordnung mit HRTF-Richtcharakteristik," Masterarbeit, Universität Oldenburg. (Cited on page 23)

Hill, P. A., Nelson, P. A., Kirkeby, O., and Hamada, H. (**2000**), "Resolution of front-back confusion in virtual acoustic imaging systems," *The Journal of the Acoustical Society of America* **108**(6), pp. 2901–2910, URL `http://scitation.aip.org/content/asa/journal/jasa/108/6/10.1121/1.1323235`. (Cited on pages 10 and 46)

Hirahara, T., Sawada, Y., and Morikawa, D. (**2011**), "Impact of dynamic binaural signals on three-dimensional sound reproduction," in *Proceedings of Internoise 2011, Osaka (Japan).* (Cited on page 158)

Huopaniemi, J. and Karjalainen, M. (**1996**), "HRTF Filter Design Based On Auditory Criteria," in *Proc. Nordic Acoustical Meeting (NAM'96)*, pp. 323–330. (Cited on page 87)

Huopaniemi, J., Zacharov, N., and Karjalainen, M. (**1999**), "Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design," *Journal of the Audio Engineering Society* **47**(4), pp. 218–239. (Cited on pages 71 and 86)

Iida, K. (**2008**), "Estimation of sound source elevation by extracting the vertical localization cues from binaural signals," *Proceedings of Meetings on Acoustics* **4**(1), URL `http://scitation.aip.org/content/asa/journal/poma/4/1/10.1121/1.2980020`. (Cited on page 158)

Jeffress, L. A. (**1948**), "A place theory of sound localization," *Journal* *of Comparative and Physiological Psychology* **41**(1), pp. 35–39, URL `http://psycnet.apa.org/index.cfm?fa=search.displayRecord&#38;id=F0BE069D-BF68-BB7F-0A03-D83AB1C01E7C&#38;resultID=1&#38;page=1&#38;dbTab=all`. (Cited on page 26)

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**), "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America* **124**(1), pp. 422–438, URL `http://scitation.aip.org/content/asa/journal/jasa/124/1/10.1121/1.2924135`. (Cited on page 26)

Kaernbach, C. (**1991**), *"Simple adaptive testing with the weighted up-down method,"* . (Cited on page 77)

Kahana, Y., Nelson, P. A., Kirkeby, O., and Hamada, H. (**1999**), "A multiple microphone recording technique for the generation of virtual acoustic images," *The Journal of the Acoustical Society of America* **105**(3), pp. 1503–1516. (Cited on pages 23, 47, 74, 84, 120, and 136)

Kajala, M. and Hamaldinen, M. (**1999**), "Broadband Beamforming Optimization for Speech Enhancement in Noisy Environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, USA, pp. 19–22. (Cited on pages 23, 46, and 48)

Karjalainen, M. (**1987**), "Auditory Models for Speech Processing," in *Proc. of Int. Congr. of Phonetic Sciences, Invited paper*, Tallinn, Estonia, U.S.S.R. (Cited on page 26)

Kates, J. M. (**1993**), "Superdirective arrays for hearing aids," *The Journal of the Acoustical Society of America* **94**(4), pp. 1930–1933. (Cited on pages 46, 50, and 51)

Köhler, S., Blau, M., van de Par, S., and Rasumow, E. (**2014**), "Simultane Messung mehrerer HRTFs in nichtreflexionsarmer Umgebung," in *Fortschritte der Akustik - DAGA 2014*, Oldenburg, Germany, pp. 202–203. (Cited on page 17)

Kim, S.-M. and Choi, W. (**2005**), "On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach," *The Journal of the Acoustical Society of America* **117**(6), pp. 3657–3665, URL `http://scitation.aip.org/content/asa/journal/jasa/117/6/10.1121/1.1921548`. (Cited on pages 19, 139, and 148)

Kirkeby, O., Nelson, P., Hamada, H., and Orduna-Bustamante, F. (**1998**), "Fast deconvolution of multichannel systems using regularization," *Speech and Audio Processing, IEEE Transactions on* **6**(2), pp. 189–194. (Cited on page 20)

Kirkeby, O. and Nelson, P. A. (**1999**), "Digital Filter Design for Inversion Problems in Sound Reproduction," *Journal of the Audio Engineering Society* **47**(7/8), pp. 583–595, URL `http://www.aes.org/e-lib/browse.cfm?elib=12098`. (Cited on pages 18, 20, 122, and 141)

Kistler, D. J. and Wightman, F. L. (**1992**), "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *The Journal of the Acoustical Society of America* **91**(3), pp. 1637–1647. (Cited on pages )

Klumpp, R. G. and Eady, H. R. (**1956**), "Some Measurements of Interaural Time Difference Thresholds," *The Journal of the Acoustical Society of America* **28**(5), pp. 859–860. (Cited on pages 100 and 106)

Kohlrausch, A., Braasch, J., Kolossa, D., and Blauert, J. (**2013**), "An Introduction to Binaural processing," in *The Technology of Binaural Listening*, edited by J. Blauert, Springer, pp. 1–32. (Cited on page 10)

Kohlrausch, A. and Breebaart, J. (**2001**), "Perceptual (ir)relevance of HRTF magnitude and phase spectra," in *Audio Engineering Society Convention 110*. (Cited on pages 86 and 107)

Kulkarni, A. and Colburn, H. S. (**2000**), "Variability in the characterization of the headphone transfer-function," *The Journal of the Acoustical Society of America* **107**(2), pp. 1071–1074, URL `http://scitation.aip.org/content/asa/journal/jasa/107/2/10.1121/1.428571`. (Cited on page 19)

Kulkarni, A. and Colburn, S. (**1998**), "Role of spectral detail in sound-source localization," *Letters to nature* **396**(24/31), pp. 747–749. (Cited on pages 17, 71, 86, and 89)

Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (**1999**), "Sensitivity of human subjects to head-related transfer-function phase spectra," *The Journal of the Acoustical Society of America* **105**(5), pp. 2821–2840. (Cited on pages 15, 71, 86, 87, 89, 90, 101, 106, 107, and 109)

Le Goff, N., Buchholz, J., and Dau, T. (**2013**), "Modeling horizontal localization of complex sounds in the impaired and aided impaired auditory system," in *The Technology of Binaural Listening*, edited by J. Blauert, Springer, pp. 121–144. (Cited on page 26)

Levin, D., Habets, E., and Gannot, S. (**2013**), "Robust beamforming using sensors with nonidentical directivity patterns," in *IEEE International*

*Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, pp. 91–95. (Cited on pages 25 and 47)

Li, J., Stoica, P., and Wang, Z. (**2003**), "On robust Capon beamforming and diagonal loading," *IEEE Transactions on Signal Processing* **51**(7), pp. 1702–1715. (Cited on page 122)

Licklider, J. C. R. (**1951**), "A Duplex Theory of Pitch Perception," *The Journal of the Acoustical Society of America* **23**(1), pp. 147–147, URL `http://scitation.aip.org/content/asa/journal/jasa/23/1/10.1121/1.1917296`. (Cited on page 12)

Lindau, F., Alexander; Brinkmann (**2012**), "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings," *Journal of the Audio Engineering Society* **60**(1/2), pp. 54–62, URL `http://www.aes.org/e-lib/browse.cfm?elib=16166`. (Cited on page 109)

Mabande, E., Schad, A., and Kellermann, W. (**2009**), "Design of robust superdirective beamformers as a convex optimization problem," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, pp. 77–80. (Cited on page 122)

Macpherson, E. A. and Sabin, A. T. (**2007**), "Binaural weighting of monaural spectral cues for sound localization," *The Journal of the Acoustical Society of America* **121**(6), pp. 3677–3688, URL `http://scitation.aip.org/content/asa/journal/jasa/121/6/10.1121/1.2722048`. (Cited on page 158)

Majdak, P., Balazs, P., and Laback, B. (**2007**), "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions," *Journal of the Audio Engineering Society* **55**(7/8), pp. 623–637, URL `http://www.aes.org/e-lib/browse.cfm?elib=14190`. (Cited on page 17)

Masiero, B. S. (**2012**), "Individualized Binaural Technology. Measurement, Equalization and Subjective Evaluation," Ph.D. thesis, Institute of Technical Acoustics RWTH Aachen University. (Cited on pages 10, 18, 19, 20, 139, and 141)

May, T., van de Par, S., and Kohlrausch, A. (**2011**), "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Transactions on Audio, Speech, and Language Processing* **19**(1), pp. 1–13. (Cited on page 26)

McCowan, I. A. (**2004**), "Microphone Arrays: A Tutorial," URL `https://`

`www.idiap.ch/~mccowan/arrays/tutorial.pdf`, called on 29.09.2014. (Cited on pages 21 and 22)

Mehrgardt, S. and Mellert, V. (**1977**), "Transformation characteristics of the external human ear," *The Journal of the Acoustical Society of America* **61**(6), pp. 1567–1576. (Cited on pages 12 and 87)

Mellert, V. (**1972**), "Construction of a Dummy Head after New Measurements of Threshold of Hearing," *The Journal of the Acoustical Society of America* **51**, pp. 1359. (Cited on page 10)

Mellert, V., Siebrasse, K. F., and Mehrgardt, S. (**1974**), "Determination of the transfer function of the external ear by an impulse response measurement," *The Journal of the Acoustical Society of America* **56**(6), pp. 1913–1915, URL `http://scitation.aip.org/content/asa/journal/jasa/56/6/10.1121/1.1903534`. (Cited on pages 12 and 17)

Mellert, V. and Tohtuyeva, N. (**1997**), "Multimicrophone arrangements as suitable for dummyhead recording technique," in *Proc. 137th ASA Meeting*, p. 3117. (Cited on pages 23, 34, and 120)

Michaud, P.-Y., Meunier, S., Herzog, P., Lavandier, M., and d'Aubigny, G. D. (**2013**), "Perceptual Evaluation of Dissimilarity Between Auditory Stimuli: An Alternative to the Paired Comparison," *Acta Acustica united with Acustica* **99**(5), pp. 806–815, URL `http://www.ingentaconnect.com/content/dav/aaua/2013/00000099/00000005/art00014`. (Cited on page 144)

Mitchell, L. D. (**1982**), "Improved Methods for the Fast Fourier Transform (FFT) Calculation of the Frequency Response Function," *Journal of Mechanical Design* **104**(2), pp. 277–279. (Cited on pages 17, 57, and 92)

Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. (**1995**), "Transfer Characteristics of Headphones Measured on Human Ears," *Journal of the Audio Engineering Society* **43**(4), pp. 203–217, URL `http://www.aes.org/e-lib/browse.cfm?elib=7954`. (Cited on pages 17, 19, 20, and 138)

Moore, B. C. J. (**2003**), *An Introduction to the Psychology of Hearing, Fifth Edition*, Academic Press, Paperback. (Cited on pages 12 and 86)

Moore, B. C. J. and Glasberg, B. R. (**1983**), "Suggested formulae for calculating auditory filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America* **74**(3), pp. 750–753, URL `http://scitation.aip.org/content/asa/journal/jasa/74/3/10.1121/1.389861`. (Cited on pages 27, 53, and 139)

Nordebo, S., Claesson, I., and Nordholm, S. (**1994**), "Weighted Chebyshev Approximation for the Design of Broadband Beamformers Using Quadratic Programming," *IEEE Signal Processing Letters* **1**(7), pp. 103–105. (Cited on pages 23, 46, and 74)

Paquier, M., Koehl, V., and Jantzem, B. (**2011**), "Effects of headphone transfer function scattering on sound perception," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011*, pp. 181–184. (Cited on pages 20 and 141)

Parodi, Y. (**2010**), *A Systematic Study of Binaural Reproduction Systems Through Loudspeakers: A Multiple Stereo-dipole Approach : PhD Thesis*, Section of Acoustics, Aalborg University, URL `http://books.google.de/books?id=SxZRtwAACAAJ`. (Cited on page 18)

Patterson, R. D. and Nimmo-Smith, I. (**1980**), "Off-frequency listening and auditory-filter asymmetry," *The Journal of the Acoustical Society of America* **67**(1), pp. 229–245. (Cited on pages 86 and 107)

Paul, S. (**2009**), "Binaural Recording Technology: A Historical Review and Possible Future Developments," *Acta Acustica united with Acustica* **95**(5), pp. 767–788. (Cited on pages 10, 74, 84, and 136)

Perrott, D. R. and Nelson, M. A. (**1969**), "Limits for the Detection of Binaural Beats," *The Journal of the Acoustical Society of America* **46**(6B), pp. 1477–1481. (Cited on pages 100 and 107)

Pulkki, V., Karjalainen, M., and Huopaniemi, J. (**1998**), "Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model," in *Audio Engineering Society Convention 104*. (Cited on pages )

Rasumow, E., Blau, M., Doclo, S., Hansen, M., Mellert, M., Püschel, D., and van de Par, S. (**2012**a), "Psychoakustisch motivierte Glättung von kopfbezogenen Übertragungsfunktionen: Hörbarkeit der Linearisierung von Phasengängen," in *Fortschritte der Akustik - DAGA 2012*, Darmstadt, Germany, pp. 633–634. (Cited on page 83)

Rasumow, E., Blau, M., Doclo, S., Hansen, M., van de Par, S., Püschel, D., and Mellert, M. (**2014**a), "Individualized binaural reproduction using a virtual artificial head," in *Fortschritte der Akustik - DAGA 2014*, Oldenburg, Germany, pp. 26–27. (Cited on pages 135 and 151)

Rasumow, E., Blau, M., Doclo, S., Hansen, M., van de Par, S., Püschel, D., and Mellert, V. (**2013**a), "Least squares versus non-linear cost functions for a vitual artificial head," in *Proceedings of the 21st International Congress on Acoustics, Montreal (Canada).* (Cited on pages 47, 48, 49, 57, 60, 74, 84, 85, 110, 121, 122, 127, 128, 136, 137, and 152)

Rasumow, E., Blau, M., Doclo, S., Püschel, D., Hansen, M., Mellert, M., van de Par, S., and Püschel, D. (**2011**a), "Zahlentheoretisch motivierte Optimierung von Mikrofonpositionen für einen virtuellen Kunstkopf," in *Fortschritte der Akustik - DAGA 2011*, Düsseldorf, Germany, pp. 637–638. (Cited on page 33)

Rasumow, E., Blau, M., Hansen, M., Doclo, S., van de Par, S., Mellert, V., and Püschel, D. (**2011**b), "Robustness of virtual artificial head topologies with respect to microphone positioning errors," in *Proc. Forum Acusticum, Aalborg*, Aalborg, pp. 2251–2256. (Cited on pages xxii, 25, 33, 47, 57, 74, 81, 82, 84, 85, 110, 121, 122, 123, 124, 136, and 137)

Rasumow, E., Blau, M., Hansen, M., Doclo, S., van de Par, S., Mellert, V., and Püschel, D. (**2014**b), "The Impact of the White Noise Gain (WNG) of a Virtual Artificial Head on the Appraisal of Binaural Sound Reproduction," in *Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014*, pp. 174–180. (Cited on pages 47, 48, 49, 52, 57, 119, 137, and 139)

Rasumow, E., Blau, M., Hansen, M., Doclo, S., van de Par, S., Püschel, D., and Mellert, V. (**2012**b), "Smoothing head-related transfer functions for a virtual artificial head," in *Acoustics 2012*, Nantes, France, pp. 1019–1024. (Cited on pages 73, 95, 107, 125, 137, and 138)

Rasumow, E., Blau, M., Hansen, M., van de Par, S., Doclo, S., Mellert, V., and Püschel, D. (**2014**c), "Smoothing individual head- related transfer functions in the frequency and spatial domains," *Journal of the Acoustical Society of America* **135**(4), pp. 2012–2025. (Cited on pages 27, 57, 58, 83, 125, 137, and 138)

Rasumow, E., Blau, M., van de Par, S., Hansen, M., Doclo, S., Püschel, D., and Mellert, M. (**2013**b), "Subjective importance of individual HRTF phase," in *Proc. Annual Conference on Acoustics (AIA-DAGA)*, Merano, Italy, pp. 604–607. (Cited on page 83)

Rasumow, E., Hansen, M., van de Par, S., Püschel, D., Mellert, V., Doclo, S., and Blau, M. (**2014**d), "Regularization approaches for synthesizing head-related transfer functions with microphone arrays," *Journal paper in preparation for submission to the IEEE Transactions on Audio, Speech and Language Processing* . (Cited on pages 45, 136, 137, and 139)

Rasumow, E., Hansen, M., van de Par, S., Püschel, D., Mellert, V., Doclo, S., and Blau, M. (**2014**e), "Perceptual evaluation of individualized binaural reproduction using a virtual artificial head," *Journal paper in* *preparation* . (Cited on pages 135 and 145)

Raufer, S., Rasumow, E., and Blau, M. (**2013**), “HRTF- Measurements with Earmolds and Conventional Ear Plugs - A Comparison,” in *Proceedings of AIA-DAGA 2013, Merano (Italy)*, pp. 1320–1321. (Cited on pages xvii, 17, and 138)

Raykar, V. C., Duraiswami, R., and Yegnanarayana, B. (**2005**), “Extracting the frequencies of the pinna spectral notches in measured head related impulse responses,” *The Journal of the Acoustical Society of America* **118**(1), pp. 364–374, URL `http://scitation.aip.org/content/asa/journal/jasa/118/1/10.1121/1.1923368`. (Cited on page 158)

Romigh, G. D., Brungart, D., Stern, R. M., and Simpson, B. D. (**2013**), “The role of spatial detail in sound-source localization: Impact on HRTF modeling and personalization.” ASA, vol. 19, p. 050170, URL `http://link.aip.org/link/?PMA/19/050170/1`. (Cited on page 89)

Sakamoto, S., Hongo, S., Kadoi, R., and Suzuki, Y. (**2008**), “SENZI and ASURA: New High-Precision Sound-Space Sensing Systems based on Symmetrically Arranged Numerous Microphones,” in *Universal Communication, 2008. ISUC ’08. Second International Symposium on*, pp. 429–434. (Cited on pages 23, 47, and 74)

Sakamoto, S., Hongo, S., Okamoto, T., Iwaya, Y., and Suzuki, Y. (**2013**), “Improvement of accuracy of three-dimensional sound space synthesized by real-time “SENZI”, a sound space information acquisition system using spherical array with numerous microphones,” in *Proceedings of the 21st International Congress on Acoustics, Montreal (Canada).* (Cited on pages 23, 120, 121, and 137)

Schmidt, J.-H., Mauermann, M., and Blau, M. (**2011**), “Streuung der Hörschwelle von Normalhörenden für Frequenzen oberhalb 1kHz bei Bezug auf den Schalldruck im Ohrsimulator, im Freifeld und am Trommelfell,” in *Fortschritte der Akustik - DAGA 2011*, Düsseldorf, Germany, pp. 353–354. (Cited on page 16)

Schmidt, S. (**2009**), “Finite Element Simulation of External Ear Sound Fields for the Optimization of Eardrum-Related Measurements,” Ph.D. thesis, Institut für Kommunikationsakustik, Ruhr-Universität Bochum. (Cited on page 19)

Schärer, Z. and Lindau, A. (**2009**), “Evaluation of Equalization Methods for Binaural Signals,” in *Audio Engineering Society Convention 126*, URL `http://www.aes.org/e-lib/browse.cfm?elib=14917`. (Cited on page 20)

Searle, C. L., Braida, L. D., Davis, M. F., and Colburn, H. S. (**1976**), "Model for auditory localization," *The Journal of the Acoustical Society of America* **60**(5), pp. 1164–1175, URL `http://scitation.aip.org/content/asa/journal/jasa/60/5/10.1121/1.381219`. (Cited on page 12)

Seeber, B. (**2003**), "Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode," Ph.D. thesis, Lehrstuhl für Mensch-Maschine-Kommunikation der Technischen Universität München. (Cited on pages 11 and 12)

Simmer, K., Bitzer, J., and Marro, C. (**2001**), "Post-Filtering Techniques," in *Microphone Arrays*, edited by M. Brandstein and D. Ward, Springer Berlin Heidelberg, Berlin, Heidelberg, New York, Digital Signal Processing, pp. 39–60. (Cited on page 122)

Søndergaard, P. and Majdak, P. (**2013**), "The auditory modeling toolbox," in *The Technology of Binaural Listening*, edited by J. Blauert, Springer, pp. 33–56. (Cited on page 26)

Tohtuyeva, N. and Mellert, V. (**1999**), "Approximation of dummy-head recording technique by a multimicrophone arrangement," *The Journal of the Acoustical Society of America* **105**(2), pp. 1101–1101. (Cited on pages 23, 47, 74, 84, and 136)

Tourbabin, V. and Rafaely, B. (**2014**), "Theoretical Framework for the Optimization of Microphone Array Configuration for Humanoid Robot Audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(12), pp. 1803–1814. (Cited on page 160)

Van Veen, B. and Buckley, K. (**1988**), "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine* **5**(2), pp. 4–24. (Cited on page 21)

Van Veen, B. and Buckley, K. (**1998**), "Beamforming techniques for Spatial Filtering," in *The Digital Signal Processing Handbook*, edited by V. Madisetti, Crc Pr Inc, pp. 61.1–61.20. (Cited on page 21)

Völk, F. (**2014**), "Inter- and Intra-Individual Variability in the Blocked Auditory Canal Transfer Functions of Three Circum-Aural Headphones," *Journal of the Audio Engineering Society* **62**(5), pp. 315–323, URL `http://www.aes.org/e-lib/browse.cfm?elib=17242`. (Cited on page 141)

Völkering, J.-U. (**2014**), "Räumliche Interpolation von HRTFs für einen virtuellen Kunstkopf," Masterarbeit, Universität Oldenburg. (Cited on page 158)

Völkering, J.-U., Rasumow, E., and Blau, M. (**2014**), “Examination of different HRTF interpolation methods,” in *Fortschritte der Akustik - DAGA 2014*, Oldenburg, Germany, pp. 562–563. (Cited on page 158)

Voigt, C. and Adamy, J. (**2007**), *Formelsammlung der Matrizenrechnung*, Oldenbourg Wissenschaftsverlag, URL `http://books.google.de/books?id=W3dSwLfg64EC`. (Cited on pages )

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (**1993**), “Localization using nonindividualized head-related transfer functions,” *The Journal of the Acoustical Society of America* **94**(1), pp. 111–123, URL `http://scitation.aip.org/content/asa/journal/jasa/94/1/10.1121/1.407089`. (Cited on pages 10 and 46)

Wightman, F. L. and Kistler, D. J. (**1989**), “Headphone simulation of free-field listening. II: Psychophysical validation,” *The Journal of the Acoustical Society of America* **85**(2), pp. 868–878, URL `http://scitation.aip.org/content/asa/journal/jasa/85/2/10.1121/1.397558`. (Cited on page 139)

Wightman, F. L. and Kistler, D. J. (**1992**), “The dominant role of lowfrequency interaural time differences in sound localization,” *The Journal of the Acoustical Society of America* **91**(3), pp. 1648–1661, URL `http://scitation.aip.org/content/asa/journal/jasa/91/3/10.1121/1.402445`. (Cited on page 12)

Wightman, F. L. and Kistler, D. J. (**1997**), “Monaural sound localization revisited,” *The Journal of the Acoustical Society of America* **101**(2), pp. 1050–1063, URL `http://scitation.aip.org/content/asa/journal/jasa/101/2/10.1121/1.418029`. (Cited on page 12)

Wright, D., Hebrank, J. H., and Wilson, B. (**1974**), “Pinna reflections as cues for localization,” *The Journal of the Acoustical Society of America* **56**(3), pp. 957–962, URL `http://scitation.aip.org/content/asa/journal/jasa/56/3/10.1121/1.1903355`. (Cited on page 158)

Xie, B. and Zhang, T. (**2010**), “The Audibility of Spectral Detail of Head-Related Transfer Functions at High Frequency,” *Acta Acustica united with Acustica* **96**(2), pp. 328–339. (Cited on pages 71, 86, and 107)

Yasuno, Y. and Ohga, J. (**2005**), “Temperature characteristics of electret condenser microphones,” in *ISE-12. 2005 12th International Symposium on Electrets, 2005*, pp. 412–415. (Cited on page 59)

Yost, W. A. (**1974**), “Discriminations of interaural phase differences,” *The Journal of the Acoustical Society of America* **55**(6), pp. 1299–1303. (Cited on page 106)

Ziegelwanger, H. and Majdak, P. (**2014**), "Modeling the direction-continuous time-of-arrival in head-related transfer functions," *The Journal of the Acoustical Society of America* **135**(3), pp. 1278–1293, URL `http://scitation.aip.org/content/asa/journal/jasa/135/3/10.1121/1.4863196`. (Cited on page 15)

Zotkin, D. N., Duraiswami, R., and Gumerov, N. A. (**2009**), "Regularized HRTF fitting using spherical harmonics." in *WASPAA*, IEEE, pp. 257–260. (Cited on pages 23, 88, 120, 121, and 137)

# Acknowledgments

First of all, I would like to thank Prof. Dr. Simon Doclo and Prof. Dr. Matthias Blau for accepting me as a Ph.D. candidate. I am honored and very grateful to have had the opportunity to work with them.

I am very grateful to my supervisor, Matthias Blau, who gave me the opportunity to work on a very interesting topic. I would like to thank Matthias Blau for his patience and kindness and for teaching me to do scientific work. It is the great merit of Matthias Blau to have constantly led the numerous discussions to fruitful decisions and to have organized the successful progress of the whole project.

I also wish to thank Simon Doclo for supervising my work. I am very thankful for his support and valuable knowledge that only made this work possible. I would like to thank Simon Doclo for the many helpful suggestions for improvement, which considerably enhanced the performance of the VAH and the resulting publications.

I would like to thank Prof. Dr. Steven van de Par and Prof. Dorte Hammershøi for their kind participation in my thesis committee. Furthermore, I would like to express my gratitude to Steven van de Par for considerably improving my understanding of perception of sound and for his support in enhancing the publications.

My sincere thanks to Prof. Dr. Martin Hansen for his valuable guidance and tireless patience. I wish to thank Martin Hansen for his farsightedness and for his friendly and perpetual support.

I am grateful to Prof. Dr. Volker Mellert and Dr. Dirk Püschel for their support of this work and for the numerous fruitful discussions. Furthermore, I would like to thank Dirk Püschel and Akustik Technologie Göttingen (SoundTec) for the partial economic support of the project.

The economic support by the Bundesministerium für Bildung und Forschung (grant no. 17080X10) is greatly acknowledged.

I owe my deepest gratitude to my mentor Uwe Simmer. I am obliged to thank for his never ending patience, encouraging support and helpful expla-

nations. It was always a great pleasure to discuss with Dr. Uwe.

Thanks to all members of the Institut für Hörtechnik und Audiologie for creating a very nice working environment. Special thanks to my colleagues and friends Tobias Sankowsky-Rothe, Sven Kissner, Matthias Brandt, Andreas Bruns, and Freerk Schmits-Stapelfeld for steadily resetting my focus and for their moral support. Moreover, many thanks to my colleagues Marco Wilmes and Holger Groenewold for their technical support. Special thanks to Holger Groenewold for the great deal of soldering work.

I would like to thank to all subjects who participated in a myriad of experiments and all helpers supporting the project *Virtueller Kunstkopf*: Julia Habicht, Sven Kissner, Stefan Raufer, Christopher Bensch, Simon Köhler and Jan-Uwe Völkering.

A special thank you goes to my parents Leo & Vera: Спасибо Вам огромное за то, что Вы меня всегда и во всём поддерживали и поддерживаети. Спасибо Вам за всё, что Вы для меня сделали.

Finally, I am most indebted to my wife Viviane. Thank you for your moral support, your understanding, your patience and for your ability to ease stressful periods.

# Curriculum Vitae

| **Personal** | |
|---:|---|
| name | Eugen Rasumow |
| date of birth | 5-th of March, 1983 |
| place of birth | Schachtinsk, Kazakhstan |
| nationality | German |
| **Education** | |
| 2003 | **Gymnasium Westerstede**<br>Westerstede, Germany<br>**A-Level** |
| 2003-2007 | **University of Applied Science of Oldenburg**<br>Oldenburg, Germany<br>**Dipl.-Ing. (FH)** in Hearing Technology and Audiology |
| 2007-2009 | **University of Oldenburg**<br>Oldenburg, Germany<br>**Master of Science (M.Sc.)** in Hearing Technology and Audiology |
| 2009-2015 | **University of Applied Science of Oldenburg**<br>Oldenburg, Germany<br>**Scientific staff member** |
| 2010-2015 | **University of Oldenburg**<br>Oldenburg, Germany<br>**Ph.D. candidate** |

# Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig verfasst habe und nur die angegebenen Quellen und Hilfsmittel verwendet habe. Teile der Dissertation wurden bereits veröffentlicht, wie an den entsprechenden Stellen angegeben. Der Anteil der Koautoren an den Veröffentlichungen bestand in der Betreuung der Arbeit und Korrektur der Manuskripte.

Oldenburg, April 13, 2015

.................................................
(Eugen Rasumow)