# Spatial Audio Through Headphones Based on HRTFs Approximated by Parametric IIR Filters

Von der Fakultät für Elektrotechnik
der Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg
zur Erlangung des akademischen Grades eines Doktor-Ingenieurs
genehmigte

DISSERTATION
vorgelegt von

**Patrick Nowak**

aus Hamburg, Deutschland

Hamburg 2022

| | |
|---|---|
| Gutachter: | Univ.-Prof. Dr.-Ing. habil. Udo Zölzer<br>Helmut-Schmidt-Universität,<br>Universität der Bundeswehr Hamburg |
| | Univ.-Prof. Dr.-Ing. Delf Sachau<br>Helmut-Schmidt-Universität,<br>Universität der Bundeswehr Hamburg |
| Vorsitzender: | Univ.-Prof. Dr.-Ing. Gerd Scholl<br>Helmut-Schmidt-Universität,<br>Universität der Bundeswehr Hamburg |
| Tag der Disputation: | 06.05.2022 |

# Acknowledgments

This dissertation was written during my time as research assistant in the cooperation project between French-German Research Institute of Saint-Louis (ISL) and Helmut Schmidt University / University of the Federal Armed Forces Hamburg. I would like to take this opportunity to thank the people who have accompanied me during this project.

First of all, I would like to thank my supervisor Prof. Udo Zölzer for giving me the opportunity to do research at the Department of Signal Processing and Communication and guiding me through the project. I would also like to thank the directors of ISL, Thomas Czirwitzky and Christian de Villemagne, for realizing the cooperation project. Furthermore, I would like to thank Bernd Fischer, Pierre Naz, and Véronique Zimpfer for supervising and welcoming me during my visits at ISL.

I would also like to thank my colleagues at the Department of Signal Processing and Communication for a great working atmosphere, interesting discussions, and enjoyable board game evenings. A special thanks goes to my former working colleague Piero Rivera Benois for supervising my master's thesis and supporting me with fruitful discussions, common publications, and friendly advice during my time as a research assistant.

Last but not least, I would like to thank my family without whom this work would not have been possible. I am very much thankful to my parents for making it possible for me to study and supporting me all my life. I express my deep sense of gratitude to my wife Elena for her love and encouragement during the last years. I am thankful to my dog Rocky for being by my side during two years in home office. Furthermore, I would like to thank Britta Ehlers for proofreading this dissertation.

Patrick Nowak

# Abstract

The subject of this dissertation is spatial audio through headphones. In the present work, an offline binaural synthesis implementation is proposed using head-related transfer functions (HRTFs) approximated by cascades of parametric infinite impulse response (IIR) filters, parameter interpolation to calculate HRTFs of intermediate directions for generating static as well as moving virtual sound sources, and simulated room effects in order to increase the perceived externalization.

The first contribution to the research field lies in representing HRTFs as cascades of low-order parametric IIR filters together with a delay representing the interaural time difference (ITD). Usually, HRTFs are represented as finite impulse response (FIR) filters containing the corresponding head-related impulse responses (HRIRs) as filter coefficients. However, by using cascades of low-order parametric IIR filters, like first-order shelving or second-order peak filters, memory requirements of the used hardware can be decreased to three parameters per filter stage (cut-off or center frequency, gain, and Q-factor). For this purpose, a two-step procedure is proposed that approximates the magnitude responses of HRTFs by parametric IIR filter cascades. In a first step, the individual filter stages are consecutively integrated, initialized, and tuned. Afterwards, the interaction between individual filter stages is post-optimized. Alternatively, an approach for HRTF magnitude response approximation based on instantaneous backpropagation is proposed. After approximating the HRTF magnitude responses, also the ITDs have to be extracted from the HRIRs or HRTFs of the two ears.

From this, virtual sound sources are generated by filtering a monaural audio signal with the parametric IIR filter cascades of the desired direction and delaying the filtered audio signal of the contralateral ear by the

extracted ITD. In many practical implementations, only a finite number of measured HRTFs is available, resulting in a limited spatial resolution. For HRTFs represented as FIR filters, bilinear rectangular or triangular interpolation can be used to calculate the filter coefficients of intermediate HRTFs. However, when the HRTFs are represented as IIR filters instead, the interpolation is not as straightforward as for FIR filters due to stability considerations. Therefore, in this work, a parameter interpolation algorithm based on bilinear interpolation of the parameters of the individual filter stages together with an assignment of related peak filters is proposed. This interpolation algorithm guarantees the stability of intermediate filters. When generating moving virtual sound sources, two IIR filter cascades are combined in parallel following the cross-fading input-switching combination approach.

For evaluating the proposed methods, three listening tests assessing different aspects of binaural synthesis using HRTFs approximated by parametric IIR filters are performed. In a first listening test, the validity of the proposed parametric IIR filter cascades is proven for static virtual sound sources by comparing their localization results to localization results achieved using HRIRs represented as FIR filters. Additionally, a second listening test proves that adding simulated room effects via the image source model increases the perceived externalization of static virtual sound sources generated using HRTFs approximated by parametric IIR filter cascades up to externalization levels achieved using measured binaural room impulse responses represented as FIR filters. Finally, the audio quality of moving virtual sound sources generated using minimum-phase approximated HRIRs represented as FIR filters and parametric IIR filter cascades is evaluated in a third listening test. By using two IIR filters in parallel following the cross-fading input-switching combination approach, comparable audio quality ratings are achieved as for FIR filter implementations using minimum-phase approximated HRIRs. Thus, HRTFs approximated by parametric IIR filter cascades can be used to reduce the number of saved coefficients. By using two first-order shelving filters, ten second-order peak filters, a mean HRTF magnitude value, and an extracted ITD, only 36 parameters have to be saved per HRTF instead of 200 coefficients as in FIR filter implementations using conventional HRIRs.

# Kurzfassung

Das Thema dieser Dissertation ist räumliches Audio über Kopfhörer. In der vorliegenden Arbeit wird eine Offline-Implementierung der Binauralsynthese vorgeschlagen, die Außenohrübertragungsfunktionen (engl. *head-related transfer functions*, HRTFs) verwendet, die durch Kaskaden parametrischer Filter mit unendlicher Impulsantwort (engl. *infinite impulse response*, IIR) approximiert werden. Außerdem werden eine Interpolation der Parameter zur Berechnung der HRTFs von Zwischenrichtungen und simulierte Raumeffekte zur Erhöhung der wahrgenommenen Externalisierung verwendet.

Der erste Beitrag zum Forschungsgebiet liegt in der Darstellung von HRTFs als Kaskade parametrischer IIR-Filter niedriger Ordnung zusammen mit einer Verzögerung, die die interaurale Zeitdifferenz (engl. *interaural time difference*, ITD) darstellt. Normalerweise werden HRTFs als Filter mit endlicher Impulsantwort (engl. *finite impulse response*, FIR) dargestellt, die die entsprechenden Außenohrimpulsantworten (engl. *head-related impulse responses*, HRIRs) als Filterkoeffizienten enthalten. Durch die Verwendung von Kaskaden parametrischer IIR-Filter niedriger Ordnung, wie Shelving-Filter erster Ordnung oder Peak-Filter zweiter Ordnung, kann der Speicherbedarf der verwendeten Hardware auf drei Parameter pro Filterstufe (Grenz- oder Mittenfrequenz, Verstärkungsfaktor und Q-Faktor) reduziert werden. Zu diesem Zweck wird ein zweistufiges Verfahren vorgestellt, das die Betragsfrequenzgänge der HRTFs durch parametrische IIR-Filterkaskaden approximiert. In einem ersten Schritt werden die einzelnen Filterstufen nacheinander integriert, initialisiert und abgestimmt. Anschließend wird das Zusammenspiel der einzelnen Filterstufen optimiert. Alternativ wird ein Ansatz für die Approximation des Betragsfrequenzganges der HRTFs auf Grundlage von *instantaneous backpropagation* vorgeschlagen. Nach der Approximation der Betragsfrequenzgänge werden die ITDs aus den HRIRs

oder HRTFs der beiden Ohren extrahiert.

Virtuelle Schallquellen werden erzeugt, indem ein monaurales Audiosignal mit den parametrischen IIR-Filterkaskaden der gewünschten Richtung gefiltert und das gefilterte Audiosignal des kontralateralen Ohrs um die extrahierte ITD verzögert wird. In vielen praktischen Anwendungen steht nur eine begrenzte Anzahl von gemessenen HRTFs zur Verfügung, was zu einer begrenzten räumlichen Auflösung führt. Für HRTFs, die als FIR-Filter dargestellt werden, kann die bilineare Interpolation verwendet werden, um die Filterkoeffizienten der dazwischenliegenden HRTFs zu berechnen. Wenn die HRTFs jedoch stattdessen als IIR-Filter dargestellt werden, ist die Interpolation aufgrund von Stabilitätsbedingungen nicht so einfach wie bei FIR-Filtern. Daher wird in dieser Arbeit ein Interpolationsalgorithmus vorgeschlagen, der auf der bilinearen Interpolation der Parameter der einzelnen Filterstufen zusammen mit einer Zuordnung zusammengehöriger Peak-Filter basiert. Dieser Interpolationsalgorithmus garantiert die Stabilität der Filter. Bei der Erzeugung bewegter virtueller Schallquellen werden zwei IIR-Filterkaskaden nach dem Ansatz der *cross-fading-input-switching*-Kombination parallelisiert.

Zur Bewertung der vorgeschlagenen Methoden werden drei Hörtests durchgeführt, die verschiedene Aspekte der Binauralsynthese unter Verwendung von HRTFs, die durch parametrische IIR-Filter approximiert werden, bewerten. In einem ersten Hörtest wird die Gültigkeit der vorgeschlagenen parametrischen IIR-Filterkaskaden für statische virtuelle Schallquellen nachgewiesen, indem ihre Lokalisierungsergebnisse mit denen verglichen werden, die mit als FIR-Filter dargestellten HRTFs erzielt werden. Darüber hinaus wertet ein zweiter Hörtest, die wahrgenommene Externalisierung statischer virtueller Schallquellen, die mit HRTFs, die durch parametrische IIR-Filterkaskaden approximiert werden, generiert werden, aus. Durch Hinzufügen simulierter Raumeffekte, die über das Spiegelquellen-Modell erzeugt werden, erhöht sich die wahrgenommene Externalisierung auf das gleiche Level, das mit gemessenen binauralen Raumimpulsantworten erreicht wird. Schließlich wird in einem dritten Hörtest die Audioqualität von sich bewegenden virtuellen Schallquellen bewertet, die mit Hilfe von als FIR-Filter dargestellten minimalphasigen HRIRs und parametrischen IIR-Filterkaskaden erzeugt werden. Durch die parallele Verwendung von zwei IIR-Filtern nach dem *cross-fading-input-switching*-Kombinationsansatz werden vergleichbare Audioqualitätsbewertungen erzielt wie bei FIR-Filterimplementierungen minimalphasiger HRIRs. Somit können HRTFs, die durch parametrische IIR-Filterkaskaden approximiert werden, verwendet werden, um die Anzahl der zu speichernden Koeffizienten zu reduzieren. Durch die Verwendung von zwei Shelving-Filtern, zehn Peak-Filtern, einem HRTF-Durchschnittsbetragswert und einer extrahierten ITD müssen nur 36 Parameter pro HRTF gespeichert werden, anstatt 200 Koeffizienten wie bei FIR-Filterimplementierungen herkömmlicher HRIRs.

# Contents

# Glossary

In this glossary, the used mathematical operators, functions, symbols and abbreviations are introduced. The symbols which are used once and defined directly in their sections are omitted. The elements are arranged in alphabetical order.

## Mathematical Operators and Functions

| | |
|---|---|
| $\lvert \cdot \rvert$ | Absolute value |
| $\arg\max_{(\cdot)}(\cdot)$ | Arguments of the maxima |
| $C_{\mathrm{T}}(\cdot)$ | Centroid |
| $a * b$ | Convolution of $a$ and $b$ |
| $\cos(\cdot)$ | Cosine function |
| $\mathrm{diag}(\cdot)$ | Diagonal matrix with arguments as diagonal elements |
| $\left\lVert \cdot \right\rVert$ | Euclidean norm |
| $e^{(\cdot)}$ | Exponential function |
| $\int (\cdot) dt$ | Integral with variable $t$ |
| $\log_{10}(\cdot)$ | Logarithm with basis 10 |
| $\max(\cdot)$ | Maximum value |
| $\mathrm{mean}(\cdot)$ | Mean value |
| $(\cdot) \bmod (\cdot)$ | Modulo |
| $\ln(\cdot)$ | Natural logarithm |
| $\partial a / \partial b$ | Partial derivative of $a$ with respect to $b$ |
| $\sec(\cdot)$ | Secant function |
| $\mathrm{sgn}(\cdot)$ | Sign |
| $\sin(\cdot)$ | Sine function |
| $\mathrm{sinc}(\cdot)$ | Sinc function |
| $\sqrt{(\cdot)}$ | Square root |
| $\tan(\cdot)$ | Tangent function |

## Signals and Systems

| | |
|---|---|
| $C(n)$ | Cost function in time-domain |
| $E_{\mathrm{dB}}(k)$ | Approximation error across frequency in dB |
| $h(n)$ | Discrete-time impulse response |
| $\bar{h}(n)$ | Interpolated impulse response |
| $h_{\mathrm{brir,d}}(n)$ | Direct part of the measured BRIR |
| $h_{\mathrm{brir,r}}(n)$ | Reverberant part of the measured BRIR |
| $h_{\mathrm{hrir}}(n)$ | Head-related impulse response |
| $h_i(n)$ | Impulse Response of $i^{\mathrm{th}}$ neighboring direction |
| $h_{\mathrm{ism}}(n)$ | RIR simulated via ISM |
| $\tilde{h}_{\mathrm{ism}}(n)$ | Modified simulated RIR |
| $h_{\mathrm{ism,d}}(n)$ | Direct part of the RIR simulated via ISM |
| $h_{\mathrm{ism,r}}(n)$ | Reverberant part of the RIR simulated via ISM |
| $h_{\mathrm{L}}(n)$ | Impulse response of the filter for the left ear |
| $h_{\mathrm{rir}}(n)$ | Room impulse response |
| $h_{\mathrm{R}}(n)$ | Impulse response of the filter for the right ear |
| $h_{\mathrm{sim,L}}(n)$ | Combined impulse response of measured HRIR and simulated RIR for the left ear |
| $\hat{H}(z)$ | Approximated transfer function |
| $\bar{H}(z)$ | Interpolated transfer function |
| $H_{\mathrm{d}}(z)$ | Desired transfer function |
| $H_{\mathrm{eq,L}}(z)$ | Transfer Function of the HpEq for the left ear |
| $H_{\mathrm{eq,R}}(z)$ | Transfer Function of the HpEq for the right ear |
| $H_{\mathrm{hfs}}(z)$ | Transfer function of an HFS |
| $H_{\mathrm{hp,R}}(z)$ | HpTF for the right ear |
| $H_{\mathrm{hrtf,L}}(z)$ | HRTF for the left ear |
| $H_{\mathrm{hrtf,R}}(z)$ | HRTF for the right ear |
| $H_{\mathrm{hte,R}}(z)$ | Transfer function of physical acoustic path between right loudspeaker of the headphone and microphone at the right ear |
| $H_i(z)$ | Transfer function of $i^{\mathrm{th}}$ neighboring direction |
| $\tilde{H}_{\mathrm{ism,L}}(z)$ | Transfer function of the modified simulated RIR for the left ear |
| $H_{\mathrm{L}}(z)$ | Transfer function of the filter for the left ear |
| $H_{\mathrm{lfs}}(z)$ | Transfer function of an LFS |
| $H_{\mathrm{lspk}}(z)$ | Transfer function of the loudspeaker |
| $H_{\mathrm{meas,R}}(z)$ | Measured transfer function from the loudspeaker to the microphone at the right ear |
| $H_{\mathrm{mic}}(z)$ | Transfer function of the microphone |
| $H_{\mathrm{peak}}(z)$ | Transfer function of a peak filter |

| | |
|---|---|
| $H_{\mathrm{ref}}(z)$ | Transfer function from the loudspeaker to a reference microphone positioned at the center of the head without listener |
| $H_{\mathrm{room}}(z)$ | Transfer function of the room |
| $H_{\mathrm{R}}(z)$ | Transfer function of the filter for the right ear |
| $H_{\mathrm{sim,L}}(z)$ | Combined transfer function of measured HRIR and simulated RIR for the left ear |
| $\lvert H(k)\rvert$ | Magnitude response |
| $\lvert H(k)\rvert_{\mathrm{dB}}$ | Magnitude response in dB |
| $\lvert \tilde{H}(k)\rvert_{\mathrm{dB}}$ | Smoothed magnitude response in dB |
| $\lvert \hat{H}(k)\rvert_{\mathrm{dB,approx}}$ | Approximated magnitude response in dB |
| $\lvert \hat{H}(k)\rvert_{\mathrm{dB,opt}}$ | Optimized magnitude response in dB |
| $\lvert H_{\mathrm{d}}(k)\rvert_{\mathrm{dB}}$ | Desired magnitude response in dB |
| $x(n)$ | Discrete-time monaural audio signal |
| $x_{\mathrm{h}}(n)$ | State signal inside all-pass filter |
| $x_{\mathrm{h},m}(n)$ | State signal inside all-pass of $m^{\mathrm{th}}$ filter in the cascade |
| $x_{\mathrm{L}}(n)$ | Discrete-time signal fed to left loudspeaker of headphone |
| $x_m(n)$ | Discrete-time input signal of $m^{\mathrm{th}}$ filter in the cascade |
| $x_{\mathrm{R}}(n)$ | Discrete-time signal fed to the right loudspeaker of the headphone |
| $X(z)$ | Z-transform of the monaural audio signal |
| $y_{\mathrm{ap1}}(n)$ | Discrete-time output signal of a first-order all-pass filter |
| $y_{\mathrm{ap2}}(n)$ | Discrete-time output signal of a second-order all-pass filter |
| $y_{\mathrm{d}}(n)$ | Desired discrete-time output signal |
| $y_m(n)$ | Discrete-time output signal of $m^{\mathrm{th}}$ filter in the cascade |

## Variables and Parameters

| | |
|---|---|
| $a$ | Coefficient inside shelving and peak filters |
| $a_{\mathrm{B}}$ | Boost coefficient inside parametric IIR filters |
| $a_{\mathrm{B},m}$ | Boost coefficient inside $m^{\mathrm{th}}$ filter in the cascade |
| $a_{\mathrm{C}}$ | Cut coefficient inside parametric IIR filters |
| $a_{\mathrm{C},m}$ | Cut coefficient inside $m^{\mathrm{th}}$ filter in the cascade |
| $a_m$ | Coefficient inside the $m^{\mathrm{th}}$ filter in the cascade |
| $c_i$ | Interpolation weight of $i^{\mathrm{th}}$ neighboring direction |
| $c_{i,\mathrm{p}}$ | Interpolation weight of $p^{\mathrm{th}}$ peak filter for $i^{\mathrm{th}}$ neighbor |
| $c_\theta$ | Weight in elevation direction for bilinear rectangular interpolation |
| $c_\varphi$ | Weight in azimuthal direction for bilinear rectangular interpolation |
| $d$ | Coefficient inside peak filters |

| | |
|---|---|
| $\bar{d}_{\text{extern}}$ | Average perceived externalization |
| $d_{\text{head}}$ | Diameter of the virtual head |
| $d_m$ | Coefficient inside $m^{\text{th}}$ peak filter in the cascade |
| $E_{\text{dB}}$ | Approximation error in dB |
| $E_{\text{tol}}$ | Error tolerance in dB |
| $f$ | Frequency |
| $f_b$ | Bandwidth |
| $f_{b,m}$ | Bandwidth of $m^{\text{th}}$ filter in the cascade |
| $f_{b,p}$ | Bandwidth of $p^{\text{th}}$ peak filter |
| $f_c$ | Cut-off or center frequency |
| $f_{c,\text{H}}$ | Cut-off frequency of the HFS |
| $f_{c,\text{L}}$ | Cut-off frequency of the LFS |
| $f_{c,m}$ | Cut-off or center frequency of $m^{\text{th}}$ filter in the cascade |
| $f_{c,p}$ | Center frequency of $p^{\text{th}}$ peak filter |
| $f_{c,p,i}$ | Center frequency of $p^{\text{th}}$ peak filter for $i^{\text{th}}$ neighbor |
| $f_{\text{k}}$ | Corresponding frequency at bin $k$ |
| $f_s$ | Sampling frequency |
| $g_i$ | Gain for $i^{\text{th}}$ filter during cross-fade |
| $g_{\text{r}}$ | Scaling factor for combining HRIRs and simulated RIRs |
| $g_{\text{r,L}}$ | Scaling factor for combining HRIRs and simulated RIRs for the left ear |
| $G$ | Gain in dB |
| $G_{\text{H}}$ | Gain of the HFS in dB |
| $G_{\text{L}}$ | Gain of the LFS in dB |
| $G_m$ | Gain in dB of $m^{\text{th}}$ filter in the cascade |
| $G_p$ | Gain in dB of $p^{\text{th}}$ peak filter |
| $G_{p,i}$ | Gain in dB of $p^{\text{th}}$ peak filter for $i^{\text{th}}$ neighboring direction |
| $H_0$ | Scaling factor inside parametric IIR filters |
| $H_{0,m}$ | Scaling factor inside $m^{\text{th}}$ filter in the cascade |
| $i$ | Index of neighboring direction |
| $i_{\text{ref}}$ | Closest neighboring direction |
| ITD | Interaural time difference |
| $\bar{\text{ITD}}$ | Interpolated ITD |
| $\text{ITD}_i$ | ITD of $i^{\text{th}}$ neighboring direction |
| $j$ | Imaginary number |
| $k$ | Frequency bin index |
| $k_{\text{max}}$ | Frequency bin index of maximum approximation error |
| $L_h$ | Length of an impulse response $h(n)$ |
| $m$ | Index of filter inside the cascade |
| $M$ | Number of cascaded parametric IIR filters |
| $n$ | Sample time index |
| $n_{\text{max}}$ | Sample time index of the maximum amplitude |
| $p$ | Index of peak filter inside cascade |

| | |
|---|---|
| $p_i$ | Index of peak filter for $i^{\text{th}}$ neighboring direction |
| $p_m$ | Parameter of $m^{\text{th}}$ filter in the cascade |
| $Q$ | Q-factor |
| $Q_{p,i}$ | Q-factor of $p^{\text{th}}$ peak filter for $i^{\text{th}}$ neighboring direction |
| $r$ | Distance |
| $t$ | Time |
| $T_{60}$ | Reverberation time |
| $V_0$ | Gain parameter inside parametric IIR filters |
| $V_{0,m}$ | Gain parameter inside $m^{\text{th}}$ filter in the cascade |
| $z$ | Complex variable of the $Z$-domain |
| $\alpha$ | Absorption coefficient |
| $\beta$ | Reflection coefficient |
| $\Delta\theta$ | Elevation resolution |
| $\Delta\varphi$ | Azimuthal resolution |
| $\eta$ | Step-size or learning rate |
| $\theta$ | Elevation |
| $\theta_i$ | Elevation of $i^{\text{th}}$ neighboring direction |
| $\mu_{\text{H,dB}}$ | Mean value of the magnitude response in dB |
| $\mu_{\bar{H},\text{dB}}$ | Interpolated mean value of magnitude response in dB |
| $\mu_{\text{H}_i,\text{dB}}$ | Mean value of the magnitude response in dB for $i^{\text{th}}$ neighboring direction |
| $\rho_{\text{fb}}$ | Front/back confusion rate |
| $\varphi$ | Azimuth |
| $\bar{\varphi}_{\text{error}}$ | Mean angular error |
| $\varphi_i$ | Azimuth of $i^{\text{th}}$ neighboring direction |
| $\varphi_{\text{interp}}$ | Interpolated azimuthal direction |
| $\varphi_{\text{original}}$ | Original azimuthal direction |
| $\varphi_{\text{perceived}}$ | Perceived azimuthal direction |
| $\varphi_{\text{rel}}$ | Relative azimuth |
| $\omega$ | Angular frequency |

## Matrices and Vectors

| | |
|---|---|
| $\mathbf{d}_{\text{extern}}$ | Vector of perceived externalizations |
| $\mathbf{f_b}$ | Vector of bandwidths |
| $\mathbf{f_c}$ | Vector of center frequencies |
| $\mathbf{G}$ | Vector of gains |
| $\mathbf{H}_{\text{mean}}$ | Vector of mean values of magnitude responses in dB |
| $\mathbf{P}_{\text{approx}}$ | Parameter matrix after approximation |
| $\bar{\mathbf{P}}_{\text{approx}}$ | Interpolated parameter matrix after approximation |
| $\mathbf{P}_{\text{approx},i}$ | Parameter matrix after approximation for $i^{\text{th}}$ neighbor |
| $\mathbf{P}_{\text{opt}}$ | Parameter matrix after optimization |

| | |
|---|---|
| $\varphi_{\text{error}}$ | Vector of azimuthal errors |
| $\tilde{\varphi}_{\text{error}}$ | Vector of signed azimuthal errors |
| $\varphi_{\text{original}}$ | Vector of original azimuthal directions |
| $\varphi_{\text{perceived}}$ | Vector of perceived azimuthal directions |

## Abbreviations

| | |
|---|---|
| 2D | Two-dimensional |
| 3D | Three-dimensional |
| BRIR | Binaural room impulse response |
| CIPIC | Center for Image Processing and Integrated Computing |
| DRR | Direct-to-reverberant energy ratio |
| DTF | Directional transfer function |
| EDC | Energy decay curve |
| ESS | Exponential sine sweep |
| FIR | Finite impulse response |
| GUI | Graphical user interface |
| HFS | High-frequency shelving filter |
| HpEq | Headphone equalization |
| HpTF | Headphone-to-ear transfer function |
| HRIR | Head-related impulse response |
| HRTF | Head-related transfer function |
| IACC | Interaural cross-correlation |
| IBPTT | Instantaneous backpropagation through time |
| IIR | Infinite impulse response |
| ILD | Interaural level difference |
| IPD | Interaural Phase Difference |
| IPTF | Inter-positional transfer function |
| IQR | Interquartile range |
| ISM | Image source model |
| ITD | Interaural time difference |
| JND | Just noticeable difference |
| KEMAR | Knowles Electronics Manikin for Acoustic Research |
| LFS | Low-frequency shelving filter |
| LSD | Log-spectral distance |
| MAA | Minimum audible angle |
| MATLAB | Matrix Laboratory |
| MESM | Multiple exponential sweep method |
| O/I | Output and input |
| RIR | Room impulse response |

CHAPTER 1

# Introduction

Based on the influence of the human anatomy on the impinging sound field, humans are able to localize sound sources by extracting directional cues from the sound field reaching their eardrums. For horizontal sound source localization, mainly interaural cues are used, whereas vertical sound source localization relies on monaural spectral cues as well as characteristic peaks and notches inside the frequency spectrum. Furthermore, cues for the distance perception are given by the absolute loudness of the sound combined with the familiarity of the source and the direct-to-reverberant energy ratio (DRR). The cues for all of these dimensions of human sound source localization are summarized in the corresponding head-related impulse responses (HRIRs) and transfer functions (HRTFs). Due to natural differences in the human anatomy, these HRIRs and HRTFs are highly individual.

Spatial audio targets on improving the immersion of virtual environments by including directional information into the audio signals that are played back through headphones. Although stereo signals are able to shift the virtual sound sources along the interaural axis between the two ears, neither vertical sound source localization nor externalization is achieved. Contrarily, binaural rendering restores the three-dimensional (3D) localization by reproducing the signals that would appear at the two ears during natural listening. For this purpose, measured HRIRs can be used to filter the monaural audio signals. Additionally, headphone equalization (HpEq) can be used to get rid of the influence of the headphone during playback.

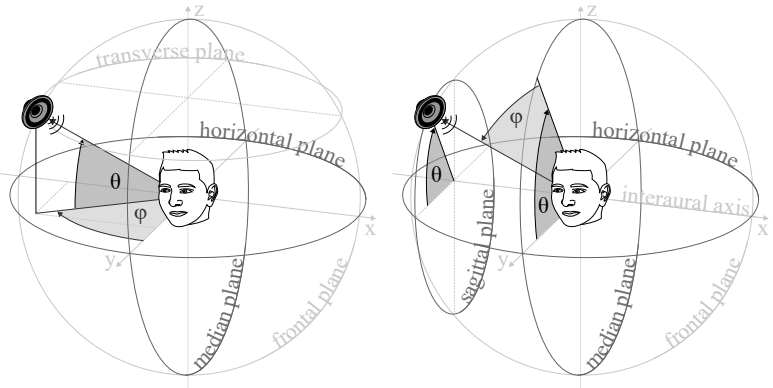
Nowadays, virtual reality gaming and multimedia are the most famous applications for 3D spatial audio. In virtual reality gaming, 3D spatial

audio through headphones is required to improve the immersion of the 3D environment created by the virtual reality glasses. Additionally, for home cinema systems, spatial audio through headphones can be used to replace the loudspeakers inside the room with a single headphone. However, not only the gaming and multimedia industry can be improved by using 3D spatial audio, also teleconferencing and videoconferencing can benefit from combining content and spatial information inside audio signals. By spatially separating the participants similar as for meetings inside real rooms, both the situational awareness and the speech intelligibility can be enhanced. Further applications could be guided tours in museums or sightseeing buses, where the information about an object can be heard from the direction of this object. Also hearing aids can be improved through spatial audio by retaining the directional information during the amplification of impinging sound waves.

With the increasing number of applications for virtual reality also the research activity on 3D spatial audio through headphones has risen. Thus, different approaches for improving the perception of the virtual experience have been developed. This work targets on reducing memory requirements of the used hardware by decreasing the number of saved coefficients per HRTF. For this, the HRTFs are approximated by a cascade of low-order parametric infinite impulse response (IIR) filters combined with a single delay representing the interaural time difference (ITD) between the two ears. In this way, the amount of stored data can be reduced to three parameters (cut-off or center frequency, gain, and Q-factor) per filter stage. Additionally, a method for interpolating the parameters of the individual filter stages is proposed in order to generate HRTFs for intermediate directions. Moreover, room simulation is used to improve the externalization achieved by approximated HRTFs up to the externalization level obtained using long binaural room impulse responses (BRIRs) that contain measured room information.

## Coordinate System

In order to define the position of a sound source relative to the head, commonly two different spherical coordinate systems are used (see Fig. 1.1). These two coordinate systems differ in the definition of azimuth $\varphi$ and elevation $\theta$. Around the human head, a head-centered rectangular coordinate system can be defined, where the $x$-axis points from the right to the left, the $y$-axis from the back to the front, and the $z$-axis from the bottom to the top (see Fig. 1.1). The $x$-axis is also called interaural axis. In this way, three important planes are defined. Firstly, the horizontal plane spanned by the $x$- and $y$-axis divides the environment in an upper and a lower hemisphere. Planes that are parallel to the horizontal plane are called transverse planes (see Fig. 1.1(a)). Secondly, the $xz$-plane, which




***(a)*** *Vertical-polar* ***(b)*** *Interaural-polar*

**Figure 1.1:** Definition of azimuth φ and elevation θ for two different spherical coordinate systems used in spatial audio through headphones: (a) vertical-polar coordinate system and (b) interaural-polar coordinate system.

defines front/back separation, is called the frontal plane. Thirdly, sagittal planes are spanned by the $y$- and $z$-axis (see Fig. 1.1(b)). The sagittal plane that bisects the human head through the center is called median plane or mid-sagittal plane. This median plane separates the environment in a left and a right hemisphere.

The most popular spherical coordinate system is the vertical-polar coordinate system, which is shown in Fig. 1.1(a). Here, azimuth $\varphi$ is defined as angle between $y$-axis and projection of the source vector on the horizontal plane, where the source vector is the vector from the center of the head to the sound source. Then, elevation $\theta$ is defined as angle between this projection and the source vector. Azimuth is defined in the range of $-180^\circ < \varphi \leq 180^\circ$, where positive angles $\varphi > 0^\circ$ define directions on the right and absolute angles $|\varphi| > 90^\circ$ directions in the back. Additionally, elevation is defined in the range of $-90^\circ \leq \theta \leq 90^\circ$ with positive angles $\theta > 0^\circ$ for upper directions. Hereby, cones of constant elevation $\theta$ are formed concentric to the $z$-axis. If the distance $r$ from the sound source to the human head is fixed, constant elevations $\theta$ are found on circular intersection lines between the sphere with radius $r$ and the transverse planes. In contrast to this, constant azimuths $\varphi$ are located on the half-plane spanned by the $z$-axis and the source vector.

In spatial audio through headphones, often the interaural-polar coordinate system is used as an alternative to the vertical-polar coordinate system (see Fig. 1.1(b)). Here, elevation $\theta$ is defined as angle between

$y$-axis and projection of the source vector on the median plane. Afterwards, azimuth $\varphi$ can be measured as angle between this projection and the source vector. In this way, constant elevations $\theta$ are located in a plane spanned by source vector and interaural axis, whereas constant azimuths $\varphi$ are found on cones concentric to the interaural axis. These cones are known as cones of confusion in spatial audio through headphones. For a fixed distance $r$ from the sound source to the human head, these cones of constant azimuth $\varphi$ transform into circular lines of intersection between the sphere with radius $r$ and the sagittal planes. In contrast to the vertical-polar coordinate system, the interaural-polar coordinate system defines azimuth only for angles between $\varphi = -90^\circ$ on the left and $\varphi = 90^\circ$ on the right. Therefore, the distinction between front and back has to be done by elevation $\theta$. This is done by defining elevation in the range of $-90^\circ \leq \theta < 270^\circ$, where $\theta = -90^\circ$ specifies the bottom, $\theta = 0^\circ$ the front, $\theta = 90^\circ$ the top, and $\theta = 180^\circ$ the back.

Although the vertical-polar coordinate system seems more straightforward in the definition of azimuth $\varphi$ and elevation $\theta$, the interaural-polar coordinate system is often preferred in spatial audio through headphones due to the relation between the cones of constant azimuth and the interaural differences used for horizontal sound source localization.

## Structure of the Work

The content of this work can be structured as follows.

In Chapter 2, the theoretical background about human natural hearing and spatial audio through headphones is summarized. Here, firstly, the principles of human sound source localization are explained by specifying the cues that are used for horizontal and vertical sound source localization as well as distance perception, including the definition of HRIRs and HRTFs. Secondly, methods for reproducing these cues during 3D spatial audio through headphones are given.

Afterwards, Chapter 3 describes the proposed cascade of parametric IIR filters used to approximate the magnitude responses of HRTFs with less coefficients. This description contains the evaluation of the minimum number of peak filters required to fulfill a given approximation error tolerance as well as approximating the HRTF magnitude responses with a given number of ten peak filters. Alternatively, also an approach for HRTF magnitude response approximation based on instantaneous backpropagation is proposed.

In order to enhance the spatial resolution of the approximated HRTFs, Chapter 4 proposes a parameter interpolation algorithm that is based on bilinear rectangular interpolation of the parameters of neighboring directions. Here, the simple parameter interpolation is improved by introducing an assignment of peak filters for neighboring directions. Furthermore,

switching between parallel cascades of parametric IIR filters enables the generation of moving virtual sound sources without audible artifacts.

Since the usage of short HRIRs during binaural synthesis results in a poor externalization of the virtual sound sources, Chapter 5 combines the approximated HRTFs with room impulse responses (RIR) simulated via the image source model (ISM). Here, simulated RIRs are scaled according to DRRs of real BRIR measurements.

In Chapter 6, the proposed approaches are evaluated using three listening tests. The first listening test focuses on the localization capabilities using the proposed cascades of parametric IIR filters and conventional FIR filter implementations. The second listening test evaluates the influence of room effects on the perceived externalization of virtual sound sources, including added simulated RIRs via ISM. Contrarily, the third listening test utilizes moving virtual sound sources in order to evaluate the audio quality of the movements generated using parameter interpolation.

Finally, in Chapter 7, conclusions are drawn and suggestions for further research are made.

CHAPTER 2

# Binaural Hearing

Binaural hearing means hearing with two ears. Based on the audio signals reaching these two ears, human beings are able to localize sound sources. In order to understand the principles of human natural hearing, the first section of this chapter summarizes the cues used by human beings to localize sound sources. Afterwards, methods are explained that try to reproduce these cues using headphone playback in order to offer the possibility of localizing virtual sound sources using only the capabilities of human's natural hearing. Finally, the topics of natural hearing and binaural reproduction are summarized.

## 2.1 Natural Hearing

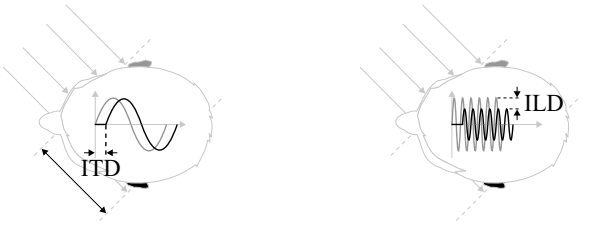
Human beings are able to localize sound sources based on the influence that the human anatomy has over the generated sound field. By evaluating the audio signals reaching the two ears, human beings are able to extract information about azimuth, elevation, and distance. In the following subsections, the cues that are used to extract the different information are explained. Additionally, HRTFs are introduced, which summarize the localization cues.

### 2.1.1 Horizontal Sound Source Localization

In the horizontal plane, primarily interaural cues, which represent differences between the left and the right ear, are used to localize a sound source [Algazi and Duda, 2011]. These differences appear as disparities in

the time of arrival between the two ears, called ITD, as well as interaural level differences (ILD) due to the head shadow effect.

In 1907, Strutt, also known as Lord Rayleigh, published the Duplex theory for lateral sound source localization of pure sine tones [Rayleigh, 1907]. In this theory, a distinction between the main cue for lateral sound source localization is done for low and high frequencies. For low frequencies ($f < 640\,\mathrm{Hz}$), the lateral sound source localization relies on the phase difference between the signals at the two ears, which is related to the ITD in case of sine tones. The threshold is based on the wavelength of the sound at the given frequency, whose half is close to the semi-circumference of the human head. Outgoing this condition would mean that the phase difference exceeded $\pi$, thus it would be unclear which of the two sounds was phase-shifted in comparison to the other. Due to the fact that we can still localize high-frequency sine tones, it was concluded that this localization relies on the ILDs.




***(a)*** *Interaural time difference* ***(b)*** *Interaural level difference*

**Figure 2.1:** For horizontal sound source localization, (a) ITDs are used for low frequencies and (b) ILDs are used for high frequencies.

Figure 2.1 shows the interaural differences for two sine waves with different frequencies arriving from the front-right. In both subfigures, the signals reaching the left and the right ear are plotted in black and gray, respectively. In Fig. 2.1(a), the arrival of a low-frequency sine tone is visualized. By comparing the signals arriving at the two ears, the ITD is clearly visible. This difference in time of arrival is explained by the longer distance that the sound wave arriving from the front-right has to travel to the left ear in comparison to the right one. When increasing the frequency of the sine tone (see Fig. 2.1(b)), the difference in time of arrival is still there but due to the fact that this difference is larger than the period of the sine tone, ambiguities on the correct value of the ITD exist. Therefore, another cue has to be used for high frequencies. The cue that can be used in this case is also visible in Fig. 2.1(b), namely the ILD, which appears as level difference between the sine waves arriving at the two ears. Again

comparing Figs. 2.1(a) and 2.1(b), it can be seen that almost no ILD exists for low frequencies, where the wavelength of the sine wave is long relative to the head diameter. However, for high frequencies the head shadow effect can lead to ILDs up to 20 dB or higher.

In 1936, Stevens and Newman [Stevens and Newman, 1936] showed that lateral sound source localization is accurate for tonal stimuli with low frequencies smaller than 1.5 kHz and high frequencies greater than 4 kHz. Here, the lower frequency ($f < 1.5$ kHz) gives an upper threshold for the usage of ITDs in lateral sound source localization and the higher frequency value ($f > 4$ kHz) gives a lower threshold for the usage of ILDs for lateral sound source localization. Between these two frequencies (1.5 kHz $< f <$ 4 kHz), the sound source localization is worse, which indicates that neither ITD nor ILD are particularly salient in this region. These results confirm the existence of the Duplex theory, where low- and high-frequency sine tones are localized by the ITD and ILD, respectively. However, the cross-over region is shifted to higher frequencies. Moreover, Zwislocki and Feldman [Zwislocki and Feldman, 1956] performed psychophysical studies showing that the sensitivity of human beings on the ITD for sine tones has a minimum near a frequency of $f = 800$ Hz and is limited to frequencies below 1.3 kHz. Well-practiced human listeners have just noticeable difference (JND) of ITDs around 9 to 27 $\mu$s for broadband noise signals, narrow-band noise signals below 600 Hz, and sine tone stimuli between 250 Hz and 1300 Hz [Klumpp and Eady, 1956]. Above this frequency, listeners are unable to detect ITDs for sine tone stimuli. The JND for ILDs is around 1 dB at $f = 1$ kHz and 0.5 dB for higher frequencies [Mills, 1960]. Additionally, Kuhn [Kuhn, 1977] showed that the effective radius of the head is increasing by a factor of 1.5 when decreasing the frequency from 2 kHz to 500 Hz. Thus, the ITD of low frequencies below 500 Hz is 1.5 times larger than for high frequencies above 2 kHz. Furthermore, Henning [Henning, 1974] extended the Duplex theory by showing that human beings can detect interaural timing information from the envelopes of high-frequency sine tones modulated by a low-frequency sine tone below 1 kHz.

Mills [Mills, 1958] studied the minimum audible angle (MAA) between two tonal sound sources that listeners are able to detect. The MAA is similar to the localization blur introduced by Blauert [Blauert, 1997]. In [Mills, 1958], the MAA is evaluated for different frequencies between 250 Hz and 10 kHz as well as different azimuths. The evaluation of the MAA for different frequencies validates the loss of localization accuracy for mid-frequencies (1.4 kHz $< f <$ 3 kHz), which was already seen in [Stevens and Newman, 1936]. Additionally, an increment of the MAA is found for lateral sound sources, which indicates that human beings are more accurate in localizing sound sources in the front than on the sides [Mills, 1958]. For low frequencies ($f < 1$ kHz), MAAs of 1° and roughly 7° are achieved for azimuths of $\varphi = 0°$ and $\varphi = 75°$, respectively.

One similarity of the mentioned studies [Rayleigh, 1907, Stevens and Newman, 1936, Mills, 1958] is that all of them use sine tone stimuli in order to evaluate the influence of ITDs and ILDs on lateral sound source localization. However, in real environments usually sounds with a broader frequency spectrum appear. Therefore, Wightman and Kistler [Wightman and Kistler, 1992] studied the role of ITD and ILD for broadband signals. In their experiment, listeners equipped with headphones were faced with broadband sounds containing conflicting ITD and ILD cues, which carry information for different directions. In this way, Wightman and Kistler showed that lateral localization of broadband sound sources containing low frequencies ($f < 2\,\mathrm{kHz}$) is dominated by the ITD. When removing the low frequencies by high-pass filtering, this dominance disappears and the listener's localization judgments follow the directional information given by the ILD. Similar results were also shown by Macpherson and Middlebrooks [Macpherson and Middlebrooks, 2002] for broadband stimuli ($500\,\mathrm{Hz} < f < 16\,\mathrm{kHz}$), low-pass filtered stimuli ($500\,\mathrm{Hz} < f < 2\,\mathrm{kHz}$), and high-pass filtered stimuli ($4\,\mathrm{kHz} < f < 16\,\mathrm{kHz}$). Additionally, it was shown that changing the monaural spectral cues while letting the interaural spectrum unchanged has almost no influence on lateral sound source localization.

Since ITD and ILD depend on different acoustic paths from the sound source to the two ears, they are mainly determined by the size and shape of the head as well as the position of the ears on the head [Blauert, 1997].

While interaural differences contain information about the lateral direction, they do not discriminate between frontal and rear sound source directions, which can also be seen as part of horizontal sound source localization. However, when using the interaural-polar coordinate system shown in Fig. 1.1(b), ITDs and ILDs can dissolve the whole azimuthal range $-90^\circ < \varphi < 90^\circ$. There, the front/back discrimination is contained in elevation $\theta$, which is evaluated in the vertical sound source localization. Thus, ITD and ILD can be called as cues for horizontal sound source localization.

### 2.1.2 Vertical Sound Source Localization

Although interaural cues enable horizontal sound source localization, they do not provide information about the vertical direction of the sound source due to areas around the interaural axis where the interaural differences have identical values [Blauert, 1997]. In the interaural-polar coordinate system shown in Fig. 1.1(b), directions with the same azimuth $\varphi$ and a different elevation $\theta$ define these so-called cones of confusion. Although Duda et al. [Duda et al., 1999] observed that the ITD can vary as much as 0.12 ms for directions on the same cone of confusion, these variations provide no information for vertical sound source localization. Furthermore, Wightman

and Kistler [Wightman and Kistler, 1993] emphasize that, despite the asymmetry of the head and the directivity of the outer ear, similarities can be seen in the measured iso-ITD and iso-ILD curves as a function of sound direction. This results in only minor changes in the relation of ITDs and ILDs that are not strong enough for a certain localization of the elevation. Thus, the cones of confusion are the basis for front/back ambiguities and elevation errors.

Monaural spectral filtering adds additional information to localize vertical sound sources [Blauert, 1997]. These spectral effects result from diffraction and reflection of the incident sound wave at human's outer ears, head, and torso. Hereby, the physical effects highly depend on the direction of the incoming sound wave. For instance, high frequencies are stronger attenuated for rear sound sources than for frontal ones due to the orientation and the shell-like structure of the pinna. Additionally, peaks and notches inside the spectrum are important features for perceiving elevated sound sources and solving front/back confusions [Middlebrooks, 1992]. These peaks and notches are the result of constructive and destructive interference of differently reflected sound waves at the eardrum. Therefore, different parts of the human body influence different frequency regions dependent on their distance to the eardrum.

The most important spectral features for vertical sound source localization are formed by reflections at the pinna. Due to the given pinna size, these features appear above 3 to 3.5 kHz, where the wavelength of the incoming sound wave is comparable to the size of the pinna [Gardner, 1973, Kuhn, 1987]. In [Algazi et al., 2001a], Algazi et al. confirmed the major effect of the pinna on the spectrum above 3 kHz. Additionally, it was shown that removing the pinna introduces only an average difference of 0.86 dB in the spectrum below 3 kHz, which indicates the negligible effect of the pinna on low frequencies. However, listening tests have shown that human beings are able to localize the vertical direction of a sound source even for low frequencies [Algazi et al., 2001a]. Here, the vertical sound source localization is based on torso reflections and head diffraction, where the influence of head diffraction is stronger at the contralateral ear and the influence of the torso reflection is stronger at the ipsilateral ear. In [Begault, 1994], the spectral influence of the different parts of the human body is separated into directional and non-directional components. Additionally, the influence of the body parts is assigned to frequency regions. The directional components consist of the torso influencing frequencies between 0.1 and 2 kHz, shoulder reflections influencing frequencies between 0.8 and 1.2 kHz, head diffraction and reflection influencing frequencies between 0.5 and 1.6 kHz, and pinna and cavum conchae reflections influencing frequencies between 2 and 14 kHz. In contrast to these directional components, the influences of the cavum conchae dominant resonance at around 3 kHz and the ear canal and eardrum impedance for frequencies between approximately 3

and 18 kHz are independent from the direction.

In [Middlebrooks, 1992], Middlebrooks showed that human's vertical sound source localization of narrow-band sounds relies consistently on the maximum correlation between the spectrum of the sound wave at the eardrum and the directional transfer functions that are associated with the different sound source directions. This indicates that the auditory system of human beings assumes natural sounds having a flat and broad spectrum. When the spectrum of the original sound deviates from this assumption, the reliability of vertical sound source localization is reduced [Wightman and Kistler, 1997]. These effects are also included in the boosted and directional bands formulated by Blauert [Blauert, 1997]. Here, every frequency band is related to one of the vertical directions (front, back, or overhead), such that when facing a listener with a specific narrow-band sound, the vertical sound source localization relies on this frequency band rather than on the true sound location.

These uncertainties when localizing sound sources with an unknown spectrum can be solved by voluntary head movements [Mackensen et al., 1998, Wightman and Kistler, 1999]. Especially front/back confusions can be easily suspended by rotating the head, because with clockwise rotations frontal sound sources will move to the left, whereas rear sound sources will move to the right. These horizontal shifts can be evaluated by corresponding changes in ITD and ILD. Rotating the head counter-clockwise will change the position vice versa. Not only front/back ambiguities can be solved by spontaneous head movements, every vertical sound source localization can benefit from these head movements.

The smallest change of elevation angle that is resolvable by the human auditory system is about 4° in the median plane and increases up to 10° for the sides [Seeber, 2003].

### 2.1.3 Distance Perception

The third dimension of human sound source localization is the distance of the sound source. Three of the main cues for distance perception are the absolute loudness of the sound combined with the familiarity of the source, the DRR, and the low-frequency ILD [Algazi and Duda, 2011]. However, the importance of these cues is related to the environment and the distance. In [Blauert and Braasch, 2008], the distance perception is sorted in three different distance regions.

For nearby sound sources ($r < 25\,\text{cm}$), the interaural and monaural cues used for horizontal and vertical sound source localization are drastically changed [Blauert and Braasch, 2008]. Since these changes are known by the human auditory system, they can be used for distance perception. This effect is called auditory parallax [Blauert, 1997]. Brungart and Rabinowitz have evaluated these changes in dependence of the distance with a sphere

model and dummy-head measurements [Brungart and Rabinowitz, 1999]. Both methods indicate the substantial increase in ILD for lateral sound sources closer than one meter. For distances $r < 0.5\,\mathrm{m}$, this effect is even higher. The reason for the increase of ILD when decreasing the distance of lateral sound sources is the increase of the level at the ipsilateral ear and the decrease of the level at the contralateral ear due to the head shadow effect. Even at low frequencies, where the ILD is very small for medium range sound sources, the ILD can increase up to 20 dB for nearby sound sources at a distance of $r = 12\,\mathrm{cm}$ [Brungart and Rabinowitz, 1999]. Although high-frequency ILDs change in a similar way, having low-frequency ILDs only for nearby sound sources makes it an important cue for distance perception. In contrast to the high dependence of the ILD on the distance, the dependence is considerably weaker for the ITD [Brungart and Rabinowitz, 1999].

For medium range distances ($0.25\,\mathrm{m} < r < 15\,\mathrm{m}$), the sound level decreases with distance. In free-field, the level decreases with $1/r$, which means that doubling the distance leads to a loss of 6 dB in sound level. However, the perceived distance doubles with a loss of 20 dB in sound level [Seeber, 2003]. Hereby, an auditory horizon is formed, which limits the acoustical distance perception of human beings to 15 m [Blauert and Braasch, 2008]. Nevertheless, inside medium range distances, the sound level is the primary cue for distance perception. In order to localize the distance of a sound source accurately, a priori knowledge about the loudness of the sound source is needed. Otherwise, wrong predictions of the original loudness will lead to errors in distance perception. In [Gardner, 1969], Gardner showed that listeners use the vocal effort of whispered and shouted voice for distance perception of sound sources. This additional interpretation led to an overestimated distance of shouted voice and an underestimated distance of whispered voice. Inside rooms, the DRR is used as an objective cue for distance perception [Zahorik et al., 2005]. Although the reverberant energy remains almost constant, the DRR decreases with distance because of the reduction of energy of the direct sound. Since the reverberant energy depends on the original sound level and the room acoustics, usage of the DRR cancels the necessity of knowing the original loudness of the sound source for distance perception. Thus, the DRR is the preferred cue for distance perception inside rooms.

Although higher distances ($r > 15\,\mathrm{m}$) are outside of the auditory horizon, human beings use a priori knowledge and visual cues to localize distant sound sources. Additionally, the frequency-dependent attenuation during sound propagation is a helpful cue for localizing distant sound sources [Blauert and Braasch, 2008]. Since the attenuation increases with frequency, it represents a low-pass filtering. Using this cue and knowledge about the sound source, e.g. a roll of thunder can be estimated in a higher distance than 15 m.

In general, human beings underestimate sound source distances higher

than one meter and overestimate smaller distances [Zahorik et al., 2005]. One reason for this could be a margin of safety for avoiding other objects in the real world. Moreover, the distance of lateral sound sources can be localized relatively accurately, whereas the distance of frontal nearby sound sources is overestimated [Kopčo and Shinn-Cunningham, 2011].

Because of these uncertainties in human perception of absolute distances, human distance perception can be seen as a perception of changes in distance rather than perceiving absolute distances.

### 2.1.4 Head-Related Transfer Functions

In order to summarize all the localization cues that are used for horizontal sound source localization (see Section 2.1.1), vertical sound source localization (see Section 2.1.2), and distance perception (see Section 2.1.3), HRTFs can be used [Wightman and Kistler, 1989]. These HRTFs are specified as transfer functions between a sound source and human eardrums. More specifically, the HRTF is defined as ratio of the Fourier transforms of sound pressure levels at the human's eardrum and sound pressure at the center location of the listener's head without the listener [Algazi and Duda, 2011]. Hence, the HRTF accounts only for the filtering effects introduced by the human outer ear, head, and body. Other names for the HRTF are the free-field-to-eardrum transfer function [Blauert, 1997] or the anatomical transfer function [Hartmann, 1999]. The inverse Fourier transform of the HRTF is the HRIR.

When measuring HRTFs, the human subject is equipped with a probe microphone at the eardrum or a microphone at the blocked ear canal [Møller, 1992]. Additionally, a loudspeaker is placed at the desired direction and distance. In this way, a transfer function

$$H_{\mathrm{meas,R}}(z) = H_{\mathrm{lspk}}(z) \cdot H_{\mathrm{brir,R}}(z) \cdot H_{\mathrm{mic,R}}(z) \tag{2.1}$$

can be measured for the right ear, including the transfer function of the loudspeaker $H_{\mathrm{lspk}}(z)$, the transfer function of the in-ear microphone $H_{\mathrm{mic,R}}(z)$, and the transfer function of the propagation path between loudspeaker and microphone $H_{\mathrm{brir,R}}(z)$ (see Fig. 2.2(a)). As explained in the definition of the HRTF, the HRTF of the right ear is included in the transfer function of the propagation path between the loudspeaker and the microphone located at the human's right ear $H_{\mathrm{brir,R}}(z)$. In order to separate the HRTF, a reference measurement is performed with the listener absent. Here, the microphone is placed at the center of the previous location of the listener's head (see Fig. 2.2(b)). The resulting transfer function

$$H_{\mathrm{ref}}(z) = H_{\mathrm{lspk}}(z) \cdot H_{\mathrm{room}}(z) \cdot H_{\mathrm{mic}}(z) \tag{2.2}$$

contains the same transfer function $H_{\mathrm{lspk}}(z)$, but includes a transfer function

$H_{\mathrm{room}}(z)$ of the propagation path between loudspeaker and microphone without the effects of the human's outer ear, head, and body. As microphone for the reference measurement either the same in-ear microphone as in the binaural measurement ($H_{\mathrm{mic}}(z) = H_{\mathrm{mic,R}}(z)$) or a reference microphone ($H_{\mathrm{mic}}(z) \approx 1$) should be used. Dividing Eqs. (2.1) and (2.2) gives

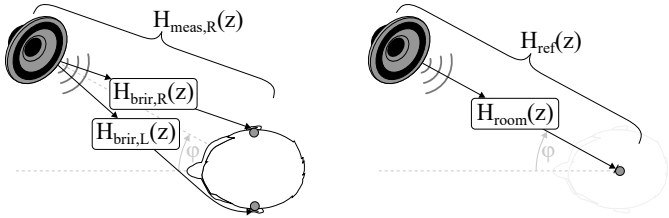
$$\frac{H_{\mathrm{meas,R}}(z)}{H_{\mathrm{ref}}(z)} = \frac{\cancel{H_{\mathrm{lspk}}(z)} \cdot H_{\mathrm{brir,R}}(z) \cdot \cancel{H_{\mathrm{mic,R}}(z)}}{\cancel{H_{\mathrm{lspk}}(z)} \cdot H_{\mathrm{room}}(z) \cdot \cancel{H_{\mathrm{mic}}(z)}}, \tag{2.3}$$

such that the transfer functions of loudspeaker $H_{\mathrm{lspk}}(z)$ and microphone $H_{\mathrm{mic,R}}(z)$ cancel out if the same microphone is used ($H_{\mathrm{mic,R}}(z) = H_{\mathrm{mic}}(z)$). Otherwise, a reference microphone should be used for the reference measurement and the in-ear microphone has to be equalized by it's sensitivity. The remaining ratio

$$H_{\mathrm{hrtf,R}}(z) = \frac{H_{\mathrm{brir,R}}(z)}{H_{\mathrm{room}}(z)} \tag{2.4}$$

contains only the effects of the human's outer ear, head, and body. Since this content matches the definition of the HRTF, $H_{\mathrm{hrtf,R}}(z)$ is called the HRTF of the right ear. Due to differences in size and shape of human's outer ears, heads, and bodies, HRTFs are highly individual [Møller, 1992]. Furthermore, HRTFs depend on azimuth $\varphi$, elevation $\theta$, and distance $r$. These characteristic dependencies are related to the cues explained in Sections 2.1.1 to 2.1.3. However, the distance-dependency is only noticeable for nearby sound sources [Duda and Martens, 1998, Brungart and Rabinowitz, 1999, Shinn-Cunningham et al., 2000] (see Section 2.1.3). Since HRTFs are usually measured in far-field at fixed distances between one and two meters, this distance-dependency can be neglected [Brungart, 2002, Kan et al., 2009]. Therefore, in the following, HRTFs are seen as functions that depend solely on azimuth $\varphi$ and elevation $\theta$.

In Fig. 2.3, exemplary HRIRs and HRTFs of *Subject_003* from the Center for Image Processing and Integrated Computing (CIPIC) database [Algazi et al., 2001b] are shown for a frontal sound source ($\varphi = 0°$, $\theta = 0°$) and a sound source on the left ($\varphi = -80°$, $\theta = 0°$). By comparing HRIRs and HRTFs for the frontal direction in the left column, similar characteristics can be seen for both ears. However, small differences are visible, which can be explained by asymmetries between the two ears. In the magnitude responses of the HRTFs in Fig. 2.3(b), the monaural spectral cues used for vertical sound source localization can be seen. For instance, strong notches are formed by the pinna at $f = 9.5\,\mathrm{kHz}$ for the frontal HRTFs of *Subject_003*. When moving toward the sides, also the interaural differences can be seen from the HRIRs and HRTFs (see right column of Fig. 2.3). The ITD is found as difference in the onsets of the HRIRs for the left and right ear (see Fig. 2.3(a)) and as difference in the slopes of the phase responses



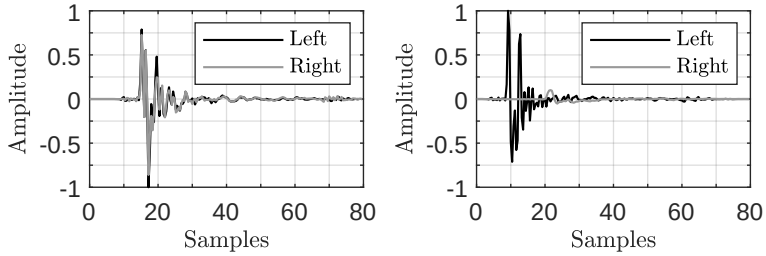

***(a)*** *Binaural measurement* ***(b)*** *Reference measurement*

**Figure 2.2:** In order to measure an HRTF, a two stage measurement procedure is performed, consisting of (a) a binaural measurement $H_{\mathrm{meas,R}}(z)$ and (b) a reference measurement $H_{\mathrm{ref}}(z)$. The HRTF is taken as ratio between these two measurements $H_{\mathrm{hrtf,R}}(z) = H_{\mathrm{meas,R}}(z)/H_{\mathrm{ref}}(z)$.

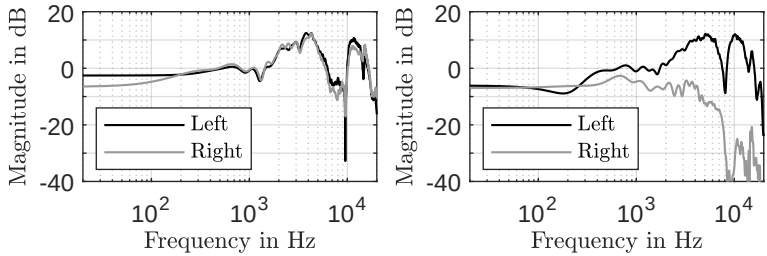
in Fig. 2.3(c). The ILD is observable by comparing the amplitudes of the HRIRs in Fig. 2.3(a) and the magnitudes of the HRTFs in Fig. 2.3(b). In the latter also the frequency-dependency of the ILD is clearly visible by subtracting the magnitude responses in decibels of the two ears. For frequencies below 300 Hz, almost no difference is visible, whereas an ILD of 7.5 dB is achieved for 2 kHz. By further increasing the frequency, the ILD reaches values above 30 dB for frequencies above 10 kHz.

In Fig. 2.4, magnitude responses of the right ear are shown in a three-dimensional representation for three different subjects from the CIPIC database [Algazi et al., 2001b]. In the left column, magnitude responses are shown for different azimuths $\varphi$ in the horizontal plane ($\theta = 0°$ (front) and $\theta = 180°$ (back)). Likewise, the right column represents the alteration of magnitude responses with elevation $\theta$ for a fixed azimuth $\varphi = 0°$ (median plane). In all of the subplots, the circular angle gives the direction of the sound source. The head-orientation is displayed in the center of the circle. Although the CIPIC database is based on the interaural-polar coordinate system, the azimuths in the left column are defined in the range of $-180° < \varphi \leq 180°$ for an easier representation. The radius of the circle contains information about the frequency, starting with $f = 200$ Hz at the inner-most position and ending at $f = 20$ kHz for the outer-most points of the circle. The information about magnitude is included in the grayscale of the diagrams, which ranges from $-50$ dB to 10 dB. Values outside of this range are clipped to the minimum or maximum value, respectively. Additionally, the color representation interpolates points in between of the angles defined in the database [Algazi et al., 2001b].

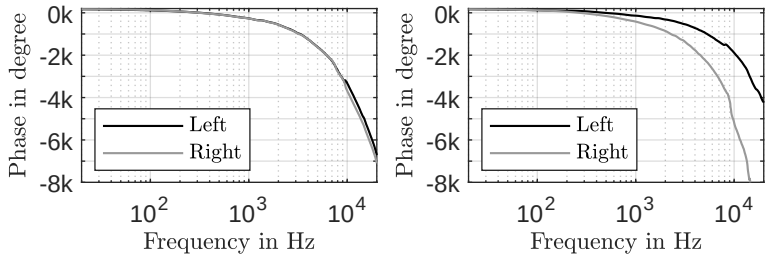
When looking at the magnitude responses in the horizontal plane (left column of Fig. 2.4), dark regions at high frequencies in the left half-plane ($\varphi < 0°$) clearly indicate the head shadow effect for contralateral sound




***(a)*** *Head-related impulse responses*




***(b)*** *Magnitudes of head-related transfer functions*




***(c)*** *Phases of head-related transfer functions*

**Figure 2.3:** Exemplary (a) HRIRs, (b) HRTF magnitudes, and (c) HRTF phases of *Subject_003* from the CIPIC database for azimuths of $\varphi = 0°$ (l.) and $\varphi = -80°$ (r.) in the front ($\theta = 0°$).

sources. Additionally, the pinna-related notch is visible for ipsilateral sound sources at frequencies between 7 and 10 kHz. However, for *Subject_065* in Fig. 2.4(c) this notch is not as pronounced as for the other two subjects. Moreover, the bright regions at frontal ipsilateral sound source positions represent the amplification of frequencies between 2 and 6 kHz, which was

already visible in Fig. 2.3(b).

In the right column of Fig. 2.4, the monaural spectral cues needed for vertical sound source localization can be found in the magnitude responses. The attenuation of high frequencies can be seen as an indicator for rear sound sources. Especially *Subject_003* and *Subject_065* in Figs. 2.4(a) and 2.4(c) show a distinct attenuation of high frequencies. Additionally, the amplification of frequencies between 3 and 6 kHz indicates elevated frontal sound sources. When comparing the different subjects in Figs. 2.4(a) to 2.4(c), some common characteristics that are based on human anatomy, e.g. shell-like structure of the pinna, can be found as described before. However, due to differences between individuals in the size of ear, head, and body, also a lot of differences can be found. These differences are mainly found in individual notch frequencies at high frequencies.

### Extraction of the interaural cues from HRTFs and HRIRs

In [Blauert and Braasch, 2008], Blauert introduced the interaural outer ear frequency response

$$H_{\mathrm{i}}(\omega, \varphi, \theta) = \frac{H_{\mathrm{hrtf,R}}(\omega, \varphi, \theta)}{H_{\mathrm{hrtf,L}}(\omega, \varphi, \theta)} \tag{2.5}$$

as ratio of right and left ear frequency responses, where $\omega = 2\pi f$ is the angular frequency. Note that distance-dependency of HRTFs is neglected here due to the assumption of medium range HRTFs (see Section 2.1.3). The relation between the previously used transfer functions in $Z$-domain and frequency responses, which are used here, is $z = e^{j\omega/f_s}$, where $f_s$ is the sampling frequency and $j = \sqrt{-1}$ is the imaginary number. Here, the ILD is defined as magnitude of the interaural outer ear frequency response

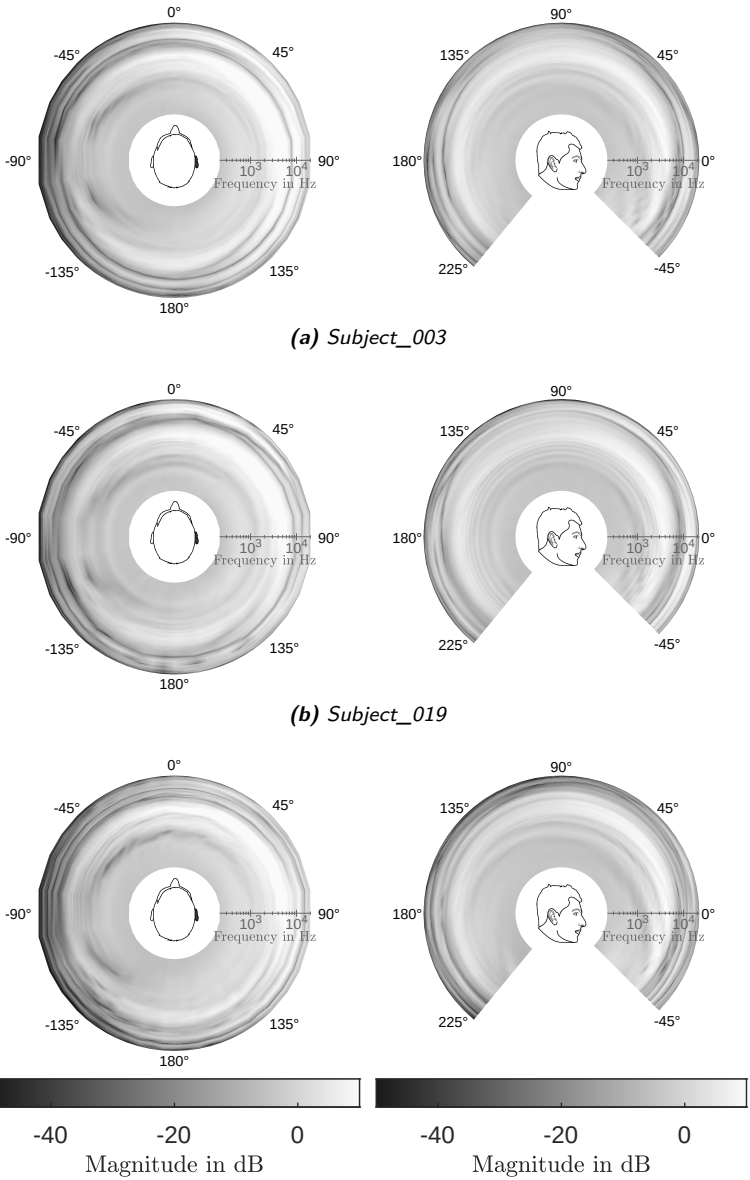
$$\mathrm{ILD}_{\mathrm{dB}}(\omega, \varphi, \theta) = 20 \log_{10} \left| H_{\mathrm{i}}(\omega, \varphi, \theta) \right|. \tag{2.6}$$

When having the magnitude responses for the two ears in decibels as seen in Fig. 2.3(b), the ILD can also be calculated as

$$\mathrm{ILD}_{\mathrm{dB}}(\omega, \varphi, \theta) = 20 \log_{10} |H_{\mathrm{hrtf,R}}(\omega, \varphi, \theta)| - 20 \log_{10} |H_{\mathrm{hrtf,L}}(\omega, \varphi, \theta)|. \tag{2.7}$$

According to [Katz and Noisternig, 2014], the extraction of the ITD from HRIRs or HRTFs can be divided into three categories, namely estimating the onset of the HRIRs based on threshold detection, computing the cross correlation between the HRIRs, and estimating the overall group delay from the phases of the HRTFs.

Threshold detection is a method to estimate the onset of HRIRs based on the first time exceeding a given threshold [Katz and Noisternig, 2014]. A low threshold will be more sensitive to fluctuations, whereas a higher




***(a)*** *Subject_003*

***(b)*** *Subject_019*

***(c)*** *Subject_065*

**Figure 2.4:** HRTF magnitudes of the right ear for three different subjects from the CIPIC database in the horizontal plane (l.) and the median plane (r.) in a three-dimensional representation.

threshold will be more robust in identifying a major peak. Since the amplitude of HRIRs varies strongly with azimuth, a relative threshold based on the maximum value can be used. Finally, the ITD is calculated as difference

$$\mathrm{ITD}(\varphi,\theta) = t_{0,\mathrm{L}}(\varphi,\theta) - t_{0,\mathrm{R}}(\varphi,\theta) \tag{2.8}$$

between the onsets $t_{0,\mathrm{L}}(r,\varphi,\theta)$ and $t_{0,\mathrm{R}}(r,\varphi,\theta)$ of the HRIRs for the left and right ear, respectively. Multiple peaks inside HRIRs arising from multi-path propagation of the sound around the head [Algazi et al., 2002] can disturb the onset detection. Especially for contralateral directions just behind the interaural axis, the shell-like structure of the pinna can lead to a lower amplitude of the main peak resulting from the path around the back of the head in comparison to the one traveling along the front of the head. A medium relative threshold can help to avoid errors in the ITD due to the presence of these multiple peaks. In [Katz and Noisternig, 2014], thresholds of $-10$ and $-20\,\mathrm{dB}$ relative to the maximum value have shown less fluctuations and errors in the calculated ITD than thresholds of 0 and $-30\,\mathrm{dB}$.

The second category is the extraction of the ITD based on interaural cross-correlation (IACC), which is defined as

$$\mathrm{IACC}(\tau,\varphi,\theta) = \frac{\int_{t_1}^{t_2} h_{\mathrm{hrir,L}}(t,\varphi,\theta) h_{\mathrm{hrir,R}}(t+\tau,\varphi,\theta)\,dt}{\sqrt{\int_{t_1}^{t_2} h_{\mathrm{hrir,L}}^2(t,\varphi,\theta)\,dt \int_{t_1}^{t_2} h_{\mathrm{hrir,R}}^2(t,\varphi,\theta)\,dt}}. \tag{2.9}$$

Here, $h_{\mathrm{hrir,L}}(t,\varphi,\theta)$ and $h_{\mathrm{hrir,R}}(t,\varphi,\theta)$ define the continuous time HRIRs for the left and right ear, respectively. Having the IACC, the ITD can be calculated either from the lag of the maximum value

$$\mathrm{ITD}(\varphi,\theta) = \arg\max_{\tau} \mathrm{IACC}(\tau,\varphi,\theta) \tag{2.10}$$

or as centroid

$$\mathrm{ITD}(\varphi,\theta) = C_\tau \left[\mathrm{IACC}(\tau,\varphi,\theta)\right]. \tag{2.11}$$

Additionally, envelopes of HRIRs can be used instead of raw impulse responses. According to [Katz and Noisternig, 2014], using the centroid instead of the maximum value has reduced the sharpness of the discontinuity in the calculated ITD.

Instead of using HRIRs for the extraction of the ITD, also the phases of the corresponding HRTFs can be used. In order to do so, the HRTF can be separated into

$$H_{\mathrm{hrtf}}(\omega,\varphi,\theta) = |H_{\mathrm{hrtf}}(\omega,\varphi,\theta)| \cdot e^{j\phi_{\mathrm{min}}(\omega,\varphi,\theta)} \cdot e^{j\phi_{\mathrm{ex}}(\omega,\varphi,\theta)}, \tag{2.12}$$

where $|H_{\mathrm{hrtf}}(\omega, \varphi, \theta)|$ defines the magnitude response, $\phi_{\mathrm{min}}(\omega, \varphi, \theta)$ the minimum-phase response, and $\phi_{\mathrm{ex}}(\omega, \varphi, \theta)$ the excess phase response [Nam et al., 2008]. The excess phase response can be further separated into a linear phase component corresponding to a pure delay and a nonlinear all-pass phase component. Note that the all-pass component can be neglected in binaural synthesis due to it's inaudibility for most of the directions [Minnaar et al., 1999]. In this way, linear curve fitting can be used to find the time delay $t_{\mathrm{del}}$ that best matches the group delay

$$\tau_{\mathrm{gr}}(\omega, \varphi, \theta) = -\frac{d\phi_{\mathrm{ex}}(\omega, \varphi, \theta)}{d\omega} \tag{2.13}$$

of the excess phase response for the entire frequency range. For the selection of the entire frequency range, various ranges can be found in literature: 0 to 1.5 kHz, 0.5 to 2 kHz, and 1 to 5 kHz [Katz and Noisternig, 2014]. By calculating the difference

$$\mathrm{ITD}(\varphi, \theta) = t_{\mathrm{del,L}}(\varphi, \theta) - t_{\mathrm{del,R}}(\varphi, \theta) \tag{2.14}$$

between the matched delays $t_{\mathrm{del,L}}(\varphi, \theta)$ and $t_{\mathrm{del,R}}(\varphi, \theta)$ of the left and right ear's excess phase responses, respectively, the ITD can be estimated. Additionally, this method can be combined with cross-correlation as used in the IACC method [Katz and Noisternig, 2014]. Hereby, the delays for the two ears are calculated based on the cross-correlation between the original HRIRs and the minimum-phase approximations of the HRIRs

$$h_{\mathrm{hrir,min}}(t, \varphi, \theta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H_{\mathrm{hrtf}}(\omega, \varphi, \theta)| e^{j\phi_{\mathrm{min}}(\omega, \varphi, \theta)} e^{j\omega t} d\omega. \tag{2.15}$$

Although all methods deliver estimations of the ITD, an analytical comparison [Katz and Noisternig, 2014] between the different methods has shown that the differences in the estimated ITD exceed the JND for ITDs.

#### Individualization of HRTFs

Since measuring individual HRTFs needs a lot of requirements that have to be fulfilled, e.g. an anechoic measurement room and hundreds of different loudspeaker positions, the process is time-consuming and financially expensive [Xu et al., 2007]. Thus, new methods for the individualization of HRTFs are in the main focus of 3D audio reconstruction via headphones. In [Xu et al., 2007], seven methods are summarized that were published before 2007. These methods include the individualization by direct HRTF measurements, averaging or using typical HRTFs, subjective selection, scaling or grouping non-individual HRTFs, theoretical computation, physical features, and tuning. Based on these earlier publications enhanced and new developments were made during the last years. A review of the last

ten years is given in [Nowak et al., 2018a], including faster individual measurements [Richter et al., 2016, Li and Peissig, 2017], anthropometric matching or interpolation with databases [Zeng et al., 2010, Geronazzo et al., 2014, Parviainen and Pertilä, 2017, Bilinski et al., 2014, Tashev, 2014, Zhu et al., 2017, Bomhardt et al., 2016, Reddy and Hegde, 2015, Sridhar and Choueiri, 2017, Haraszy et al., 2010, Chun et al., 2017], finite element simulations of the head [Huttunen et al., 2014, Schmidt and Hudde, 2009, Hiipakka, 2012], and perceptually based selection [Katz and Parseihian, 2012].

Faster individual HRTF measurements target on shorten the measurement procedure in time, in order to increase the comfort for human subjects. Here, either the combination of interleaving and overlapping exponential sine sweeps (ESS) [Majdak et al., 2007] and a rotating arc with numerous loudspeakers [Richter et al., 2016], or an adaptive measurement procedure based on arbitrary head movements of a human subject sitting in front of a loudspeaker [Li and Peissig, 2017] are used. Additionally, also simulations based on boundary element methods have been proposed to estimate the influence of the head [Huttunen et al., 2014] or the ear canal [Schmidt and Hudde, 2009, Hiipakka, 2012]. Here, models created from the anthropometric data are used to simulate the acoustic sound field.

In addition to measuring or simulating individual HRTFs, matching new subjects to subjects already existing in HRTF databases is another approach for individualizing HRTFs [Zotkin et al., 2003]. In this approach, the best match is often found by comparing the anthropometric data of the new subject with the anthropometric data of subjects inside the given database. However, also measured ILDs [Parviainen and Pertilä, 2017] or calculated notch frequencies [Geronazzo et al., 2014] can be used as base for the matching procedure. Finally, the HRTF of the matched subject is used as individualized HRTF. Instead of simply matching the anthropometric data, also interpolation between a small group of subjects inside the database can be used to find individualized HRTF magnitudes [Bilinski et al., 2014] and phases [Tashev, 2014]. In [Zhu et al., 2017], the interpolation process is extended with a different weighting for the various anthropometric features. Furthermore, the dimensions of the data can be reduced by using coordinate transforms like the principal component analysis [Bomhardt et al., 2016, Reddy and Hegde, 2015] or spherical harmonics [Sridhar and Choueiri, 2017]. Additionally, artificial neural networks [Haraszy et al., 2010] or deep neural networks [Chun et al., 2017] can be used to perform the matching process, too.

## 2.2 Spatial Audio Through Headphones

Spatial audio through headphones aims to reproduce the acoustic signals reaching the listener's eardrum or entrance of the ear canal while natural

listening. Since the two signals reaching the ears are the only information used by the auditory system to localize sound sources, reproducing the signals should be sufficient to replicate the sound scene [Nicol, 2010]. Therefore, the cues used for sound source localization, e.g. ITD, ILD, and monaural spectral cues, have to be restored by the binaural reproduction method. When comparing binaural reproduction through headphones with binaural reproduction through loudspeakers, these cues are not part of the physical acoustic path from the sound source used for the reproduction to the human ear. Thus, the localization cues have to be included in the signals fed to the internal loudspeakers of the headphone [Katz and Nicol, 2018].

Generally, binaural reproduction through headphones can be split into two categories, namely binaural recordings and binaural synthesis. These two methods will be explained in the following subsections. Additionally, HpEq can be used in order to get rid of the spectral influence of the headphone reproduction. Finally, major challenges of spatial audio through headphones are summarized.

### 2.2.1 Binaural Recordings

Binaural recordings are one of the two methods used for binaural sound reproduction via headphones. The method is easy to implement but fixed to the reproduction of the real sound scene happening during the recording.

The principle of the method is separated into two phases, namely the recording and the playback phase (see Fig. 2.5). In the recording phase, either a human head equipped with microphones at the eardrums or the entrances of the ear canals, or a dummy-head are used to record a sound scene [Møller, 1992]. The recorded signals $y_{\mathrm{L}}(n)$ and $y_{\mathrm{R}}(n)$ measured at the left and right ear, respectively, are saved for usage in the second phase. This second phase is the playback phase, which consists of feeding the recorded signals to the corresponding loudspeakers. When using headphones for the reproduction, the two ears are inevitable separated from each other, thus the signals $y_{\mathrm{L}}(n)$ and $y_{\mathrm{R}}(n)$ can be directly fed to the left and right loudspeaker of the headphone, respectively. In this way, the localization cues captured during recording can immediately be used for localization of the reproduced sound scene. However, although the two channels are easily separated while headphone playback, the headphone playback adds further spectral filtering, which has to be equalized for a proper sound scene reproduction [Møller, 1992]. Thus, HpEq can be added in order to get rid of the influence of the headphone while playback of $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$. More details on HpEq will be explained in Section 2.2.3.

When the same subject is used for recording and playback, the method is referred as individual binaural recording. Contrarily, if another subject or a dummy-head are used for recording, the method is called non-individual

binaural recording.

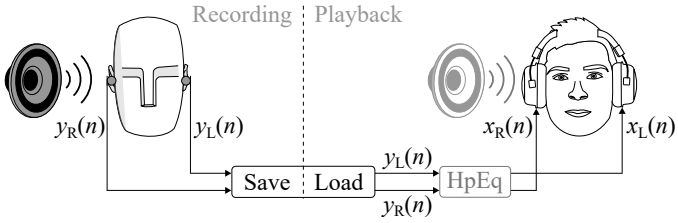



**Figure 2.5:** The principle of binaural reproduction via binaural recordings consists of two phases. In the recording phase, the sound scene is recorded by measuring the signals $y_L(n)$ and $y_R(n)$ at the two ears. In the playback phase, these signals are directly fed to the headphone as signals $x_L(n)$ and $x_R(n)$. Additionally, HpEq can be added in order to get rid of the influence of the headphone while playback.

In 1881, Ader performed the first wire transmission of paired microphone signals at Paris Electrical Exhibition [Hospitalier, 1881]. In order to record singing and orchestra music at the Grand Opera in Paris, microphones were placed pairwise on stage. These signals were then fed to monaural headphones (one for each ear) placed in the demonstration rooms of the Palais de l'Industrie. Later, this invention was called théâtrophone or electrophone [Paul, 2009]. Moreover, in 1927, Fletcher and Sivian [Fletcher and Sivian, 1927] patented a telephone system, where recording was performed by two microphones placed on opposite sides of a balloon. Here, the balloon was made of leather, cloth, or another suitable material. Additionally, the balloon was filled with sponge rubber, packed wool, or cotton. Due to this balloon important effects of the human head, like shadowing and diffraction, were modeled during the recordings. This approach is called stereophony with an acoustic septum rather than binaural recording, because of the missing pinnae [Paul, 2009].

In the 1920s, also the first ideas for constructing dummy-heads were formulated [Paul, 2009]. One of these early dummy-heads is *Oscar*, which was developed at Bell Labs around 1930. *Oscar* was a wax figure equipped with microphones mounted on the cheeks directly in front of the ears. This placement was chosen due to the huge size of the microphones. Also in the 1930s, De Boer and Vermeulen, from the company Philips, constructed the first dummy-head with a microphone placed in the pinna [Paul, 2009]. Additionally, this was the first dummy-head imitating a woman. A more detailed review on the history of dummy-heads is given in [Paul, 2009]. The first reference dummy-head was Knowles Electronics Manikin for Acoustic Research (KEMAR) [Burkhard and Sachs, 1975], which was developed for evaluating hearing aids under in-situ conditions. The anthropometric

data of the dummy-head were chosen to represent an average human adult. Additionally, an ear canal and eardrum simulator replicating the impedance based on an acoustic coupler were added. The HRTFs of KEMAR [Gardner and Martin, 1994] are still used as reference in spatial audio.

In 1978, Usami and Kato [Usami and Kato, 1978] patented a combination of headphone and microphones. In this way, a single device was able to perform binaural recording and binaural reproduction. However, commercialized systems were discontinued after a few years [Paul, 2009]. Nowadays, commercial products that combine headphone and microphones have returned to the market. Additionally, several types of dummy-heads are available by different manufacturers.

From the effort of constructing dummy-heads with replicated ear canal and eardrum characteristics, one could imagine that binaural recordings have to be performed with microphones placed at the eardrum of the listener. However, Middlebrooks et al. [Middlebrooks et al., 1989] showed that although the microphone position inside the ear canal influences the measured amplitude spectra, the directional sensitivity of broadband stimuli is independent of the position. Moreover, Hammershøi and Møller [Hammershøi and Møller, 1996] showed that comparing binaural signals recorded at the open and the blocked ear canal entrance suggests that the full spatial information is contained in recordings at the blocked ear canal entrance. Thus, the blocked ear canal entrance is the most suitable position for performing binaural recordings.

### 2.2.2 Binaural Synthesis

In binaural synthesis through headphones, level and time differences can be introduced respectively to the left and the right channel in order to offer the possibility of shifting the perceived position of the virtual sound source to one of the sides. Additionally, HRIRs can be used to filter the two channels in order to add information about elevation and to create a complete 3D space for virtual sound sources.

One advantage of binaural synthesis in comparison to binaural recordings is the flexibility of creating new sound scenes rather than simply reproducing a real sound scene. On the other hand, the signal processing effort is increased.

In 1989, Wightman and Kistler [Wightman and Kistler, 1989] have explained the principle of binaural synthesis using headphones. Here, the headphones have to reproduce the same signal at the eardrum as a natural sound from the same direction. Therefore, measured HRTFs can be used to filter the original sound source in a similar way than the human body would do during natural listening.

The principle of binaural synthesis is shown in Fig. 2.6. An overview of binaural synthesis is also given in [Begault, 1994]. The process of binaural

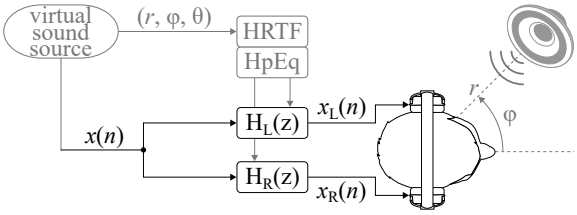



**Figure 2.6:** The principle of binaural synthesis is based on a set of measured HRTFs. In order to create a virtual sound source at distance $r$, azimuth $\varphi$, and elevation $\theta$, the corresponding HRTFs are chosen from a set of HRTFs and used as filters $H_\mathrm{L}(z)$ and $H_\mathrm{R}(z)$. Then, a monaural signal $x(n)$ is filtered by these filters to yield the binaural signals $x_\mathrm{L}(n)$ and $x_\mathrm{R}(n)$ for the left and right ear, respectively. Additionally, HpEq can be included in the filters $H_\mathrm{L}(z)$ and $H_\mathrm{R}(z)$ in order to get rid of the influence of the headphone while playback.

synthesis begins with measuring the HRTFs for the desired spatial directions (distance $r$, azimuth $\varphi$, and elevation $\theta$). Afterwards, these HRTFs are saved in a set of HRTFs for the given subject. In order to create a virtual sound source at distance $r$, azimuth $\varphi$, and elevation $\theta$, the corresponding HRTFs are chosen from the set of HRTFs and used as filters $H_\mathrm{L}(z)$ and $H_\mathrm{R}(z)$. Then, a monaural signal $x(n)$ is filtered by these filters to yield the binaural signals

$$x_\mathrm{L}(n) = h_\mathrm{L}(n) * x(n), \tag{2.16}$$

$$x_\mathrm{R}(n) = h_\mathrm{R}(n) * x(n), \tag{2.17}$$

for the left and right ear, respectively. Here, $h_\mathrm{L}(n)$ and $h_\mathrm{R}(n)$ are the impulse responses of the filters for the corresponding directions. If no HpEq is performed, these impulse responses match the HRIRs for the given direction. When performing the filtering in frequency-domain, the convolution is transformed into a multiplication

$$X_\mathrm{L}(z) = H_\mathrm{L}(z) \cdot X(z), \tag{2.18}$$

$$X_\mathrm{R}(z) = H_\mathrm{R}(z) \cdot X(z), \tag{2.19}$$

where $X(z)$, $X_\mathrm{L}(z)$, and $X_\mathrm{R}(z)$ define the $Z$-transforms of the discrete audio signals $x(n)$, $x_\mathrm{L}(n)$, and $x_\mathrm{R}(n)$, respectively.

Similar as in binaural reproduction via binaural recordings, HpEq can be added in order to get rid of the influence of the headphone. Here, HRTFs and HpEq can be combined before filtering the monaural signal $x(n)$ with

$H_{\mathrm{L}}(z)$ and $H_{\mathrm{R}}(z)$. In this way, the used filters are defined as

$$H_{\mathrm{L}}(z) = H_{\mathrm{hrtf,L}}(z) \cdot H_{\mathrm{eq,L}}(z), \tag{2.20}$$

$$H_{\mathrm{R}}(z) = H_{\mathrm{hrtf,R}}(z) \cdot H_{\mathrm{eq,R}}(z), \tag{2.21}$$

where $H_{\mathrm{eq,L}}(z)$ and $H_{\mathrm{eq,R}}(z)$ define the transfer functions for equalizing the influence of the headphone. More details on HpEq will be provided in Section 2.2.3.

In order to further improve the method of binaural synthesis, a dynamic binaural synthesis can be implemented [Xie, 2020, Wenzel, 1996]. For this, a head tracker is used to capture the head motion of the subject. Based on the orientation of the subject's head relative to the virtual sound source position, the HRTF selection is updated in real-time. In this way, the virtual sound source remains at the same position while rotating the head during dynamic binaural synthesis. This fixing of the virtual sound source's position improves the natural and immersive sense of binaural reproduction.

### 2.2.3 Headphone Equalization

Although headphones easily separate the desired signals at the two ears, they introduce additional spectral coloration (see Fig. 2.7), which degrades externalization and determination of the elevation [Griesinger, 2016]. Nevertheless, if the sound pressure at the eardrum of a listener can be precisely duplicated during headphone playback, a 3D sound experience can be recreated [Schroeder et al., 1974]. Therefore, in order to have a good 3D audio reproduction through headphone playback, HpEq is needed in addition to the usage of HRTFs to get rid of the influence of the headphone-to-ear transfer functions (HpTFs).

In [Møller, 1992], the HpTF is defined as ratio between pressure at a specified position inside the human ear canal and voltage at the headphone terminals.

Figure 2.7 shows an exemplary HpTF, which was measured with a Beyerdynamic DT770 Pro 250 Ohm headphone at the left ear of the Neumann KU100 dummy-head. The measured HpTF is shown in a frequency range between 100 Hz and 16 kHz. The shown HpTF was pre-processed to have zero mean in the given frequency range. It can be seen that frequencies below 2.5 kHz are amplified with gains up to 10.5 dB. Furthermore, two additional frequency regions (6.9 - 8 kHz and 8.9 - 13.2 kHz) are amplified, too. In the other frequency regions the attenuation creates magnitudes down to $-9$ dB. Thus, the headphone playback introduces spectral differences of up to 19.5 dB in magnitude. In order to perfectly cancel the influence of the headphone and to achieve a flat transfer function, the HpEq has to be the inverse of the HpTF (see Fig. 2.7). However, calculating the perfect inverse of a transfer function can be problematic or even not

realizable [Schärer and Lindau, 2009]. Firstly, inverting a transfer function with more poles than zeros would result in an inverse with more zeros than poles, which is non-causal and not realizable. Secondly, zeros in the transfer function would lead to a gain of infinity in the inverse at that frequency. Not only infinite gains are problematically, already strong peaks in the HpEq resulting from deep notches in the HpTF are far more audible than similar notches [Bücklein, 1981]. In order to overcome these problems, the calculated inverse is often approximated in practical applications rather than perfect [Schärer and Lindau, 2009].

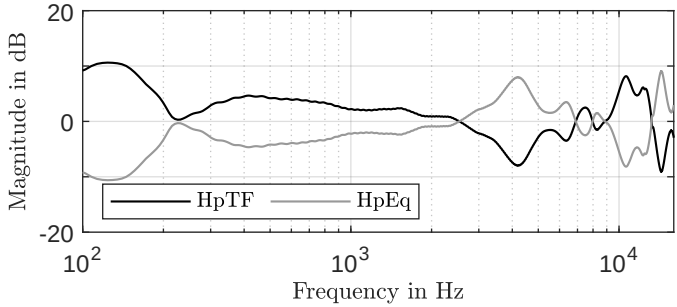



**Figure 2.7:** Measured HpTF for the Beyerdynamic DT770 Pro 250 Ohm headphone at the left ear of the Neumann KU100 dummy-head in the frequency range from 100 Hz to 16 kHz. Additionally, the ideal HpEq, which perfectly inverts the HpTF, is shown.

In [Wightman and Kistler, 1989], Wightman and Kistler explained the needed HpEq for the purpose of binaural synthesis. The principle of HpEq is shown in Fig. 2.8. During playback of a sound through headphones, the audio signal $x_{\mathrm{R}}(n)$ that is fed to the internal loudspeaker is changed by the transfer function $H_{\mathrm{hp,R}}(z)$, resulting in

$$Y_{\mathrm{R}}(z) = H_{\mathrm{hp,R}}(z) \cdot H_{\mathrm{mic,R}}(z) \cdot X_{\mathrm{R}}(z) \tag{2.22}$$

at the microphone position, with $X_{\mathrm{R}}(z)$ and $Y_{\mathrm{R}}(z)$ being the $Z$-transforms of the signals $x_{\mathrm{R}}(n)$ and $y_{\mathrm{R}}(n)$, respectively. Additionally, $H_{\mathrm{mic,R}}(z)$ defines the transfer function of the in-ear microphone and $H_{\mathrm{hp,R}}(z)$ can be separated into

$$H_{\mathrm{hp,R}}(z) = H_{\mathrm{lspk,R}}(z) \cdot H_{\mathrm{hte,R}}(z), \tag{2.23}$$

with $H_{\mathrm{lspk,R}}(z)$ and $H_{\mathrm{hte,R}}(z)$ being the transfer functions of the loudspeaker and the physical acoustic path between the loudspeaker and the microphone, respectively.

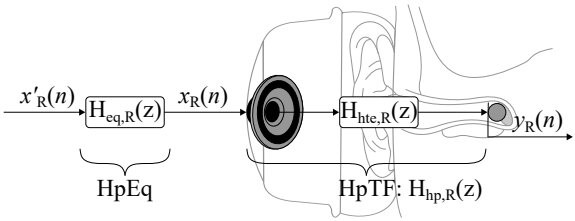



**Figure 2.8:** The principle of HpEq is based on filtering the signal $x'_{\mathrm{R}}(n)$ by $H_{\mathrm{eq,R}}(z)$ before feeding it to the loudspeaker in order to get rid of the influence of the HpTF during headphone playback. Here, $H_{\mathrm{eq,R}}(z)$ is defined as the inverse of $H_{\mathrm{hp,R}}(z)$.

In order to get rid of the influence of the HpTF during headphone playback, the signal $x'_{\mathrm{R}}(n)$ has to be filtered by $H_{\mathrm{eq,R}}(z)$ before it is fed to the loudspeaker. In this way, the $Z$-transform of the recorded signal at the microphone $y_{\mathrm{R}}(n)$ is given as

$$Y_{\mathrm{R}}(z) = H_{\mathrm{eq,R}}(z) \cdot H_{\mathrm{hp,R}}(z) \cdot H_{\mathrm{mic,R}}(z) \cdot X'_{\mathrm{R}}(z). \tag{2.24}$$

Here, $H_{\mathrm{mic,R}}(z)$ only exists during measurement of HpTF and has therefore to be equalized during measurement of HpTF through the microphone sensitivity. When the microphone is absent during normal headphone playback, the multiplication of the remaining transfer functions has to yield

$$H_{\mathrm{eq,R}}(z) \cdot H_{\mathrm{hp,R}}(z) = 1 \tag{2.25}$$

in order to achieve the desired signal $x'_{\mathrm{R}}(n)$ at the ear of the listener, $Y_{\mathrm{R}}(z) = X'_{\mathrm{R}}(z)$. Thus, the ideal HpEq is given as

$$H_{\mathrm{eq,R}}(z) = \frac{1}{H_{\mathrm{hp,R}}(z)}. \tag{2.26}$$

The most direct way to achieve individual HpEq is to measure HpTFs with probe microphones located at the eardrum [Møller, 1992]. However, due to the invasive nature of the microphone placement, the measurement is dangerous and impractical for widespread use. Thus, Hiipakka [Hiipakka, 2012] proposed a method for estimating the sound pressure at the eardrum from measurements taken at the blocked ear canal. Although this method gives the possibility for estimating the sound pressure level at the eardrum, Hammershøi and Møller [Hammershøi and Møller, 1996] showed that measuring binaural signals at the blocked ear canal contains full spatial information. Moreover, in [Oberem et al., 2016], Oberem et al. showed that when measuring both HRTFs and HpTFs with the same microphone setup

(open or close) at the entrance of the ear canal, no differences are found in the results of the performed listening test for binaural reproduction. Additionally, according to [Lindau and Brinkmann, 2012], HpTFs should be measured for the same subject as HRTFs or binaural recording. Thus, for non-individual binaural recordings, also HpEq should be non-individual matching the subject of binaural recordings. In order to avoid the necessity of microphones during measurement of individual HpTFs, Griesinger proposed an application for individual headphone equalization based on equal loudness matching of sine tones at different frequencies [Griesinger, 2016].

Another problem of usage of HpEq is the alteration of HpTFs with headphone repositioning. When perfectly equalizing the measured HpTF for one position of the headphone as shown in Fig. 2.7, peaks inside the transfer function of HpEq will lead to amplifications of this frequency region if the corresponding notch of the HpTF does not occur during playback due to headphone repositioning. For addressing this matter, Masiero and Fels [Masiero and Fels, 2011] proposed a method for robust HpEq based on calculation of the upper limit of several measured HpTFs. In this way, the transfer function of the resulting HpEq is devoid of strong peaks, which may have led to high amplifications. Another approach for considering headphone repositioning is the online measurement of HpEq [Ranjan and Gan, 2015]. Hereby, a filtered-x least mean squares algorithm can be used to adapt the filter coefficients of HpEq. In order to yield the adaptation for a virtual microphone at the eardrum, a microphone-to-eardrum-reference-point response calculated from an ear canal model can be added to the algorithm [Liski et al., 2017]. A summary of different approaches developed during the last ten years is given in [Nowak et al., 2018a].

### 2.2.4 Major Challenges

Although binaural synthesis has a long history, there are still major challenges unsolved [Xie, 2020]. These challenges include vertical sound source localization, front/back discrimination, and externalization of virtual sound sources. In the following subsections, these three challenges are described in more detail. The reason for most of the errors in sound source localization are spatial and timbral distortions due to anthropometric differences between human beings when using non-individualized HRTFs during binaural synthesis through headphones [Wenzel et al., 1993, Møller et al., 1996]. Additional attributes for differentiating between HRTF sets are given in [Simon et al., 2016]. However, these attributes will not be evaluated here.

#### Vertical sound source localization

The spatial and timbral distortions lead to a higher rate of angular errors in sound source localization of synthesized sound sources in comparison

to real sound sources [Wenzel et al., 1993]. Since vertical sound source localization is based on monaural spectral cues, spatial distortions by using non-individualized HRTFs considerably affect the localization process, leading to angular errors in elevation perception. In addition to non-individualized HRTFs, also using inappropriate or even no HpEq during binaural reproduction through headphones impairs the vertical sound source localization, too [Xie, 2020]. In contrast to vertical sound source localization, horizontal sound source localization is based on interaural rather than monaural cues, thus horizontal sound source localization is robust against spectral coloration [Wenzel et al., 1993]. Nevertheless, a difference in the head size between the listener and the subject during HRTF measurement also leads to differences in interaural cues. These non-individual interaural cues result in angular errors in azimuth perception by shifting the perceived sound source to the sides for smaller heads and to the center for broader heads than used for measuring the HRTFs.

These errors in angular sound source localization can be reduced by using individual HRTFs or matched HRTFs that are close to the listener's ones [Wenzel et al., 1993]. Additionally, dynamic cues by voluntary head movements help to improve vertical sound source localization [Xie, 2020].

#### Front/back discrimination

A special case of vertical sound source localization is the discrimination between frontal and rear sound sources, in the following referred as front/back discrimination. The basis for confusions in the front/back discrimination are inappropriate spectral cues inside the used HRTFs and missing visual cues.

Similar as in vertical sound source localization, head movements during dynamic binaural synthesis can help to reduce the front/back confusion rate [Wightman and Kistler, 1999]. These spontaneous and small motions are experienced during natural hearing and lead to monaural and interaural cue changes that can be evaluated by the human brain [Theile, 2016]. When listening to a rear sound source, rotating the head a bit to the left moves the sound source to the left, too, whereas the same rotation would result in a movement of the sound source to the right in case of a frontal sound source. Moreover, Wightman and Kistler [Wightman and Kistler, 1999] have shown that not only movements of the human head can help to reduce front/back confusion rate but also movements of the sound source that are controlled by the listener reduce the number of these confusions. Contrarily, unknown movements of the sound source are not sufficient to reduce the front/back confusion rate.

Furthermore, an additional high-frequency shelving filter can be used to increase the natural effect of the attenuation of high frequencies in monaural spectral cues for rear sound sources [Frank and Zotter, 2018]. For

this purpose, the additional filter has to amplify high frequencies for frontal directions and attenuate them for rear directions. This unnatural increase in difference of spectral cues between frontal and rear sources simplifies perception of differences in high-frequency spectral cues for the listener.

#### Externalization

Although listening to stereo audio signals through headphones enables the possibility to move the position of the sound source between the two ears, the perceived sound source stays inside the head. Contrarily, externalization is the perception of virtual sound sources outside the head, which has proven to be a difficult challenge in binaural reproduction through headphones, especially for sources directly in front or behind the listener [Algazi and Duda, 2011]. The success of externalizing a sound source during binaural synthesis depends on many factors, e.g. room effects, spectral cues, and head movements.

Since externalization is linked to distance perception, the cues for distance perception like room effects, especially the DRR, can be seen as factors for externalized virtual sound sources, too. In [Völk, 2009], Völk has shown that extending the length of a measured impulse response up to a duration of 100 ms increases the perceived externalization. This extension of the impulse response length inserts room effects, like early reflections and reverberation, into the measured impulse response. This combination of HRIR and RIR is called BRIR. Furthermore, Li et al. [Li et al., 2018] have deepened research on the influence of reverberation on the externalization of virtual sound sources by showing that reverberation at the contralateral ear has a higher influence on the perceived externalization than reverberation at the ipsilateral ear.

Not only room effects inside measured impulse responses have proven to affect externalization, also spatial and timbral distortions introduced by using non-individual HRIRs can reduce externalization [Begault et al., 2001, Mróz et al., 2018]. Additionally, Li et al. [Li et al., 2019] have shown that dynamic binaural synthesis helps to increase externalization when using short BRIRs. Contrarily, included head movements have no influence when using long BRIRs.

## 2.3 Summary

Based on the influence of human anatomy on the generated sound field, human beings are able to localize sound sources. Firstly, in the horizontal plane, mainly ITDs and ILDs are used to localize a sound source. Secondly, monaural spectral cues as well as characteristic peaks and notches inside the frequency spectrum give information for vertical sound source localization. Thirdly, cues for distance perception are given by the absolute loudness

of the sound combined with the familiarity of the source, the DRR, and the low-frequency ILD for nearby sound sources. Due to required a priori knowledge about the sound source when evaluating the distance from the absolute loudness, the auditory system is more accurate for differential localization of distance rather than absolute distance. The cues for all of these dimensions of human sound source localization are summarized in the corresponding HRIRs and HRTFs.

In spatial audio through headphones, a sound scene is either reproduced by playing back recorded signals (binaural recording) or generated by synthesizing signals that would appear at the two ears during natural listening (binaural synthesis). For this purpose, binaural synthesis uses measured HRIRs to filter a monaural signal. Although binaural recordings are easier to implement, only real acoustic scenes that were present during recording can be reproduced, whereas binaural synthesis enables the possibility of creating totally new acoustic scenes. In both methods, HpEq can be used in order to get rid of the influence of the headphone during playback.

Major challenges that still remain unsolved during static binaural reproduction are angular errors in vertical sound source localization, front/back confusions, and missing externalization. The reasons for these problems are spectral distortions due to usage of non-individual HRTFs, inappropriate HpEq, and missing room effects when using short BRIRs.

CHAPTER 3

# HRTF Magnitude Approximation with Parametric IIR Filters

HRIRs are used in the application of binaural synthesis through headphones to create a virtual sound source at a given position. In order to achieve a good spatial resolution during synthesis, HRIRs have to be saved for a high number of directions. In most of the cases, these HRIRs are implemented as finite impulse response (FIR) filters. However, IIR filters can approximate the magnitude of the given HRTF with a much lower order compared to the equivalent FIR filter. As a result, IIR filters have a considerably lesser number of coefficients to calculate and adapt. In this way, also memory requirements of saving the HRIRs are reduced when using IIR filters. Additionally, using parametric IIR filters provides the option to tune each parameter independently rather than recomputing the entire set of filter coefficients.

In the following, firstly, earlier research on the topic of IIR filter approximation of HRTFs is summarized. Afterwards, parametric IIR filters are introduced and formulas that define shelving and peak filters are given. Based on these two filter types, an approach for HRTF magnitude approximation is proposed. Additionally, a method of updating parametric IIR filters with backpropagation algorithm is evaluated. Then, the previously explained magnitude approximation is applied to HpEq, too. Finally, the method of approximating HRTF magnitudes with parametric IIR filters is summarized.

## 3.1 Research on IIR Filter Approximation of HRTFs

As described in Section 2.1.4, the phase of HRTFs can be separated into a minimum-phase and an excess phase component, where the excess phase is nearly linear and can therefore be approximated by a pure delay. Thus, minimum-phase HRTFs in combination with a delay representing the ITD can be used to implement binaural synthesis. Listening tests proved that using this combination of minimum-phase HRTFs and a delay representing the ITD shows similar localization results than using measured HRTFs [Kistler and Wightman, 1992, Kulkarni et al., 1995]. In [Huopaniemi et al., 1998], different methods for approximating HRTFs with lower order are summarized, including shorter FIR filters, pole-zero models, approximations with IIR filters, and balanced model truncation. Furthermore, Begault [Begault, 1994] separated the data reduction of saved HRTFs into three categories, namely downsampling of HRTFs, windowing of long HRTFs, and using synthetic HRTFs. Here, synthetic HRTFs are defined as shorter impulse response length filters that approximate the given HRTF by fulfilling perceptual and engineering criteria.

In 2000, Hasegawa et al. [Hasegawa et al., 2000] used a cascade of four to seven second-order IIR filters to approximate dummy-head HRTFs. A localization test in the horizontal plane showed that using these four to seven second-order IIR filters for the approximation of measured HRTFs can lead to similar localization results than using the original FIR filter implementation. Furthermore, Kulkarni and Colburn [Kulkarni and Colburn, 2004] approximated HRTFs by using a common mean HRTF across all directions and modeled directional transfer functions (DTFs) for every individual direction. A DTF is given by subtraction of the HRTF for a given direction and the mean HRTF across all directions, thus the DTF contains all direction-dependent features from the given HRTF. In [Kulkarni and Colburn, 2004], six poles and six zeros were sufficient to model the given DTF with inaudible differences for most of the directions. Additionally, Wang and Chan [Wang and Chan, 2015] used common factor decomposition analysis in order to model HRTFs by IIR filters, where HRTFs from the same azimuth share common poles and HRTFs from the same elevation share common zeros. In order to concurrently estimate the required IIR filter order and globally optimal parameters, Botts et al. used the Bayesian approach to approximate HRTFs with IIR filters [Botts et al., 2013].

In [Breebaart et al., 2009], a pair of HRTFs is parameterized by one gain vector per ear and a delay value representing the ITD. The gain vectors contain gain values for 20 to 40 non-linearly spaced frequency bands. In this way, the pair of HRTFs for a given direction is modeled with 41 to 81 real-valued parameters. Additionally, Ramos and Cobos [Ramos and Cobos, 2013] have proposed to use a cascade of parametric filters in order to approximate HRTF magnitudes. These cascades consist of a single

second-order low-frequency shelving filter (LFS) and multiple peak filters. The principle of the HRTF approximation is based on the loudspeaker equalization method published by the same authors in [Ramos and López, 2006]. During approximation procedure, a number of peak filters is consecutively added, initialized and optimized. The initial parameters are determined by the biggest error area between the current approximation and the target. Here, the center frequency is initialized by the mean logarithmic frequency of the error area, the gain is initialized as average logarithmic magnitude response inside the error area, and the Q-factor is initialized to 1. After the initialization of a new peak filter, a random search based optimization is performed in order to find an optimized parameter set close to the initial one that minimizes the error. When all peak filters are initialized individually, the random search based optimization is used for post-processing triples of neighboring filters in order to improve their interaction. In order to compute the binaural synthesis on parallel processors, the cascade of second-order shelving and peak filters is transformed into a parallel structure of a low-pass and multiple band-pass filters in [Ramos et al., 2017, Belloch et al., 2020].

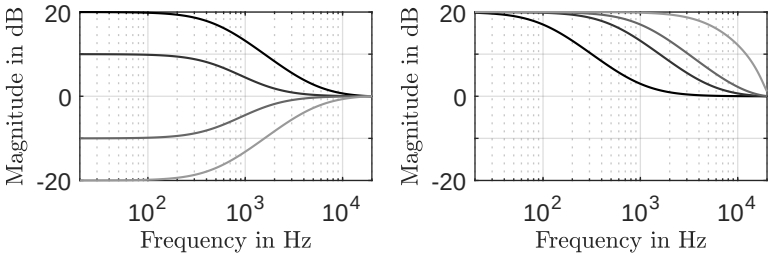
In addition to the usage of IIR filters for approximating HRTFs, parametric IIR filters are also used to tune given HRTFs in order to improve their performance. In [Yao and Chen, 2013], localization capabilities of non-individual HRTFs are improved by adjusting peak filters within a user interface. Additionally, a high-frequency shelving filter (HFS) can be used to increase the natural effect of attenuation of high frequencies in monaural spectral cues for rear sound sources [Frank and Zotter, 2018].

## 3.2 Parametric IIR Filters

The parameterization of IIR filters provides the opportunity of tuning each parameter of the filter independently rather than recomputing the whole set of filter coefficients. Shelving and peak filters are such parametric IIR filters, that amplify or attenuate frequencies in a certain band and let frequencies outside of this band pass. Hereby, shelving filters act in low- or high-frequency region, whereas peak filters act in a specified frequency band. In order to control the entire frequency range, shelving and peak filters can be cascaded.

### 3.2.1 Shelving Filters

A Shelving filter is a parametric IIR filter controlling the low- or high-frequency region and passing other frequencies [Zölzer, 2008]. Depending on the frequency region of control, LFS and HFS are differentiated. Shelving filters are controlled by their cut-off frequency $f_c$ and gain $G$ in decibels. When the gain is positive, the shelving filter is in the boost case and




***(a)*** $f_c = 500\ Hz$ ***(b)*** $G = 20\ dB$

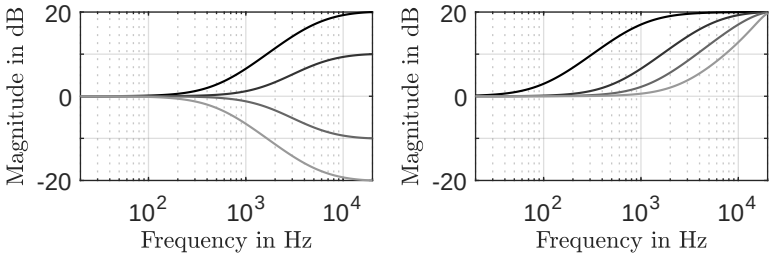
**Figure 3.1:** Exemplary magnitude responses of first-order LFS for (a) a fixed cut-off frequency $f_c = 500\,\text{Hz}$ and variable gains $G = \{20, 10, -10, -20\}\,\text{dB}$, and (b) a fixed gain $G = 20\,\text{dB}$ and variable cut-off frequencies $f_c = \{100, 500, 1000, 5000\}\,\text{Hz}$.

frequencies inside the band of control are amplified. Contrarily, negative gains attenuate those frequencies (cut case). In Fig. 3.1, exemplary magnitude responses of a first-order LFS are shown for a constant cut-off frequency $f_c = 500\,\text{Hz}$ and four different gain values (see Fig. 3.1(a)), and for a constant gain $G = 20\,\text{dB}$ and four different cut-off frequencies (see Fig. 3.1(b)). In the same way, exemplary HFS magnitude responses are shown in Fig. 3.2 for variable gains with a fixed cut-off frequency $f_c = 5\,\text{kHz}$ (see Fig. 3.2(a)) and variable cut-off frequencies with a fixed gain $G = 20\,\text{dB}$ (see Fig. 3.2(b)).

According to [Zölzer, 2008], first-order shelving filters can be implemented using a first-order all-pass filter

$$H_{\text{ap1}}(z) = \frac{a + z^{-1}}{1 + az^{-1}} \tag{3.1}$$

and two direct paths. The block diagram of this implementation is shown in Fig. 3.3(a). By using this implementation, the implementations of LFS and HFS differ only in a single sign. Adding the first-order all-pass filter output $y_{\text{ap1}}(n)$ and the unfiltered input $x(n)$ yields a low-pass filtered version of the input signal, whereas subtracting the all-pass filter output $y_{\text{ap1}}(n)$ from the input signal $x(n)$ introduces a high-pass filtering on the input signal. Additionally, a factor of 0.5 is used after summation to achieve a gain of 0 dB in the passband of the low- or high-pass filter. In this way, the transfer




***(a)*** $f_c = 5\,kHz$ ***(b)*** $G = 20\,dB$

**Figure 3.2:** Exemplary magnitude responses of first-order HFS for (a) a fixed cut-off frequency $f_c = 5\,\text{kHz}$ and variable gains $G = \{20, 10, -10, -20\}\,\text{dB}$, and (b) a fixed gain $G = 20\,\text{dB}$ and variable cut-off frequencies $f_c = \{1, 5, 10, 15\}\,\text{kHz}$.

functions of first-order low- and high-pass filters are given as

$$H_{\text{lp}}(z) = 0.5 \cdot \left[1 + H_{\text{ap1}}(z)\right], \tag{3.2}$$

$$H_{\text{hp}}(z) = 0.5 \cdot \left[1 - H_{\text{ap1}}(z)\right], \tag{3.3}$$

respectively. Scaling the transfer functions of low- and high-pass filter given in Eqs. (3.2) and (3.3), respectively, by $H_0$ and adding a second direct path gives the desired shelving filter effect. Thus, the transfer function

$$H_{\text{lfs}}(z) = 1 + \frac{H_0}{2}[1 + H_{\text{ap1}}(z)] \tag{3.4}$$

defines an LFS and

$$H_{\text{hfs}}(z) = 1 + \frac{H_0}{2}[1 - H_{\text{ap1}}(z)] \tag{3.5}$$

defines an HFS. The relation between the scaling factor $H_0$ and the gain $G$ in decibels is given by

$$H_0 = V_0 - 1 = 10^{\frac{G}{20}} - 1. \tag{3.6}$$

Additionally, the coefficient $a$ of the first-order all-pass filter given in Eq. (3.1) controls the cut-off frequency of the shelving filter. In order to achieve symmetric boost and cut cases for the shelving filters as shown in Figs. 3.1(a) and 3.2(a) [Zölzer, 2008], the calculation of the coefficient $a$ has to be done differently for the two cases. Hereby, the coefficient $a$ is

given by

$$a_{\mathrm{B}} = \frac{\tan(\pi f_c/f_s) - 1}{\tan(\pi f_c/f_s) + 1} \tag{3.7}$$

for the boost case ($G \geq 0\,\mathrm{dB}$) of both shelving filter types, where $f_c$ denotes the cut-off frequency of the shelving filter, $f_s$ is the sampling frequency, and $V_0$ is given in Eq. (3.6). For the cut case ($G < 0\,\mathrm{dB}$), a further differentiation between LFS,

$$a_{\mathrm{C}} = \frac{\tan(\pi f_c/f_s) - V_0}{\tan(\pi f_c/f_s) + V_0}, \tag{3.8}$$

and HFS,

$$a_{\mathrm{C}} = \frac{V_0 \cdot \tan(\pi f_c/f_s) - 1}{V_0 \cdot \tan(\pi f_c/f_s) + 1}, \tag{3.9}$$

has to be done in the calculation of the coefficient $a$. The difference equation that is used to implement the first-order all-pass filter in direct-form-II (see Fig. 3.3(a)) is given by

$$y_{\mathrm{ap1}}(n) = a x_{\mathrm{h}}(n) + x_{\mathrm{h}}(n-1), \tag{3.10}$$

with

$$x_{\mathrm{h}}(n) = x(n) - a x_{\mathrm{h}}(n-1). \tag{3.11}$$

From this, the difference equation of a shelving filter can be calculated to

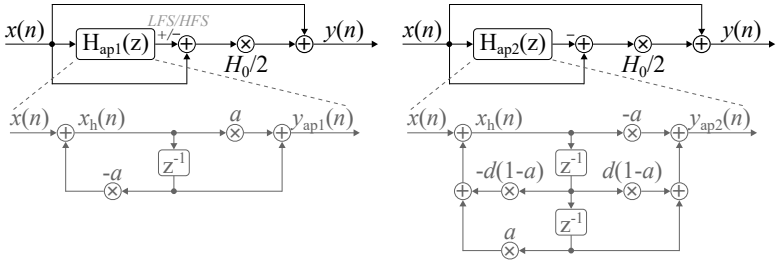
$$y(n) = x(n) + \frac{H_0}{2}\left[x(n) \pm y_{\mathrm{ap1}}(n)\right], \tag{3.12}$$

where the plus sign refers to an LFS and the minus sign to an HFS.

### 3.2.2 Peak Filters

Similar to shelving filters, peak filters amplify or attenuate a certain frequency band, which is specified by a center frequency $f_c$ and a bandwidth $f_b$, and pass frequencies outside of this band [Zölzer, 2008]. The gain parameter $G$ controls the magnitude of the amplification (boost case) or attenuation (cut case) inside the band of control. In Fig. 3.4, exemplary magnitude responses of second-order peak filters are shown for different settings in which two of the parameters are fixed and the third parameter is varied. The Q-factor of a peak filter is defined as

$$Q = \frac{f_c}{f_b} \quad \Leftrightarrow \quad f_b = \frac{f_c}{Q}. \tag{3.13}$$




**(a)** *Shelving filter* **(b)** *Peak filter*

**Figure 3.3:** Block diagrams of (a) shelving and (b) peak filters. In the shelving filter implementation, a first-order all-pass $H_{\text{ap1}}(z)$ is used, whereas the peak filter implementation uses a second-order all-pass $H_{\text{ap2}}(z)$. The $+/-$ in the block diagram of the shelving filter differentiates between LFS and HFS.

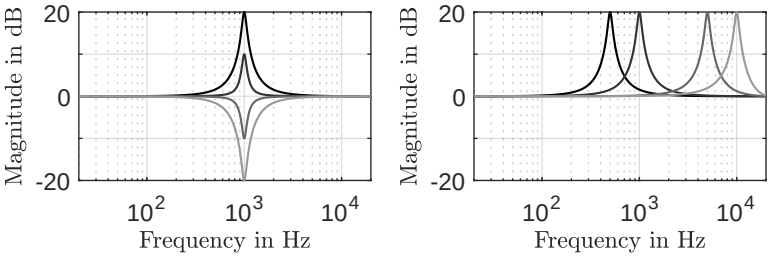
In Fig. 3.4(a), four different gains are shown for a fixed center frequency $f_c = 1\,\text{kHz}$ and a Q-factor of $Q = 10$. It can be seen that given a logarithmic frequency axis peak filters are symmetric around the center frequency $f_c$. Then, in Fig. 3.4(b), gain and Q-factor are fixed to $G = 20\,\text{dB}$ and $Q = 10$, respectively, and the center frequency moves along frequency $f_c = \{0.5, 1, 5, 10\}\,\text{kHz}$. Since the Q-factor is the ratio between center frequency $f_c$ and bandwidth $f_b$ (see Eq. (3.13)), a constant Q-factor means that increasing the center frequency $f_c$ also increases the bandwidth $f_b$ proportionally. Thus, peak filters with the same Q-factor and gain $G$ have the same width when plotted on a logarithmic frequency axis. In Fig. 3.4(c), four different Q-factors are shown for a fixed center frequency ($f_c = 1\,\text{kHz}$) and gain ($G = 20\,\text{dB}$). It can be seen that increasing only the Q-factor narrows the peak, because increasing the Q-Factor while keeping center frequency $f_c$ constant means decreasing bandwidth $f_b$ (see Eq. (3.13)).

Similar to shelving filters, peak filters can be implemented based on an all-pass filter and two direct paths (see Fig. 3.3(b)). However, in order to achieve a second-order peak filter

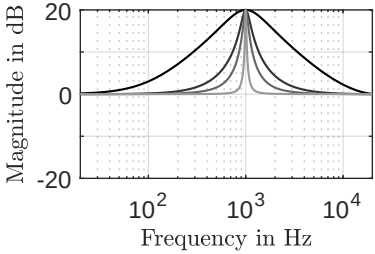
$$H_{\text{peak}}(z) = 1 + \frac{H_0}{2}[1 - H_{\text{ap2}}(z)], \tag{3.14}$$

a second-order all-pass filter given by

$$H_{\text{ap2}}(z) = \frac{-a + d(1-a)z^{-1} + z^{-2}}{1 + d(1-a)z^{-1} - az^{-2}} \tag{3.15}$$




***(a)*** $f_c = 1\,kHz,\ Q = 10$ ***(b)*** $G = 20\,dB,\ Q = 10$




***(c)*** $f_c = 1\,kHz,\ G = 20\,dB$

**Figure 3.4:** Exemplary magnitude responses of second-order peak filters for (a) a fixed center frequency $f_c = 1\,\text{kHz}$, a fixed Q-factor $Q = 10$, and variable gains $G = \{20, 10, -10, -20\}\,\text{dB}$, (b) a fixed gain $G = 20\,\text{dB}$, a fixed Q-factor $Q = 10$, and variable center frequencies $f_c = \{0.5, 1, 5, 10\}\,\text{kHz}$, and (c) a fixed center frequency $f_c = 1\,\text{kHz}$, a fixed gain $G = 20\,\text{dB}$, and variable Q-factors $Q = \{1, 5, 5, 10, 50\}$.

has to be used. Here,

$$d = -\cos(2\pi f_c / f_s) \tag{3.16}$$

controls the center frequency $f_c$, and $a_\text{B}$ and $a_\text{C}$ have the same definitions as the ones from the LFS given in Eqs. (3.7) and (3.8), respectively. However, in this context, the cut-off frequency has to be replaced by the bandwidth $f_b$, resulting in

$$a_\text{B} = \frac{\tan(\pi f_b / f_s) - 1}{\tan(\pi f_b / f_s) + 1} \tag{3.17}$$

for the boost case ($G \geq 0\,\mathrm{dB}$) and

$$a_\mathrm{C} = \frac{\tan(\pi f_b/f_s) - V_0}{\tan(\pi f_b/f_s) + V_0} \tag{3.18}$$

for the cut case ($G < 0\,\mathrm{dB}$). The difference equation that is used to implement the second-order peak filter as shown in Fig. 3.3(b) is given by

$$y(n) = x(n) + \frac{H_0}{2}\left[x(n) - y_\mathrm{ap2}(n)\right], \tag{3.19}$$

with

$$y_\mathrm{ap2}(n) = -a x_\mathrm{h}(n) + d(1-a)x_\mathrm{h}(n-1) + x_\mathrm{h}(n-2), \tag{3.20}$$

$$x_\mathrm{h}(n) = x(n) - d(1-a)x_\mathrm{h}(n-1) + a x_\mathrm{h}(n-2). \tag{3.21}$$

### 3.2.3 Filter Cascade

An important advantage of splitting a higher-order IIR filter into a cascade of lower-order IIR filters is that less filter coefficients have to be changed when modifying the desired filter response. Using parametric IIR filters in the cascaded structure will further reduce the number of changes, because tuning a single parameter does not mean that all filter coefficients have to be recalculated. In order to control the entire frequency range, a cascade of $M$ parametric IIR filters can be used (see Fig. 3.5). This cascade consists of one first-order LFS, $M-2$ second-order peak filters, and one first-order HFS. Additionally, exemplary magnitude responses with varying gain are shown in Fig. 3.5 for the different filter stages. As can be seen, all filter stages can be adapted individually by tuning their parameters $f_c$, $G$ and $Q$.

## 3.3 Approximation Using Parametric IIR Filters

In order to approximate HRTF magnitudes with parametric IIR filters, the cascaded structure shown in Fig. 3.5 is used. Here, the individual filter stages are tuned independently to best approximate the given magnitude response. The flow chart of the procedure for approximating the HRTF magnitudes with parametric IIR filters can be seen in Fig. 3.6. In [Behrends et al., 2011] and [Eichas and Zölzer, 2018], similar methods are used for equalizing loudspeakers and modeling guitar amplifiers, respectively. The following approximations are performed in Matrix Laboratory (MATLAB).

The first step of the approximation procedure seen in Fig. 3.6 is the pre-processing of the original HRIR data $h(n)$. The flow chart of the pre-processing procedure is shown in Fig. 3.7. Firstly, the Fourier transform is used to calculate the HRTFs from the given HRIRs. Note that an

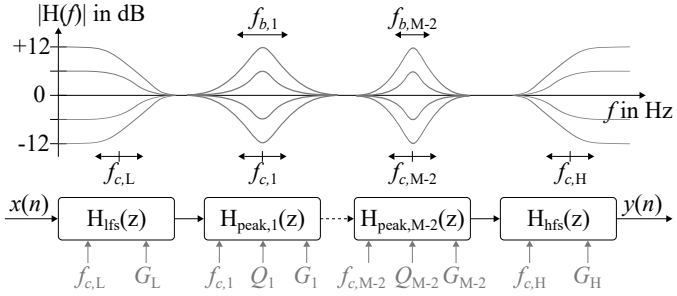



**Figure 3.5:** Block diagram of a parametric IIR filter cascade along with exemplary magnitude responses of variable gain for every filter stage. The cascade consists of one first-order LFS and one first-order HFS controlled by cut-off frequencies ($f_{c,\mathrm{L}}$, $f_{c,\mathrm{H}}$) and gains ($G_\mathrm{L}$, $G_\mathrm{H}$), and $M-2$ second-order peak filters controlled by center frequencies ($f_{c,1}$, ..., $f_{c,\mathrm{M\text{-}2}}$), Q-factors ($Q_1$, ..., $Q_{\mathrm{M\text{-}2}}$), and gains ($G_1$, ..., $G_{\mathrm{M\text{-}2}}$).

exponentially spacing of the frequency bins

$$f_\mathrm{k} = 1000\,\mathrm{Hz} \cdot 2^{\frac{-17K+k}{3K}} \tag{3.22}$$

is used instead of a linear spacing in the evaluation of the HRTFs. Here, $k \in \{0, 1, ..., 30K\}$ represents the number of the frequency bin and $K = 16$ is chosen in order to achieve a $1/48^{\mathrm{th}}$-octave resolution. In this way, the whole audible frequency range between 20 Hz and 20 kHz is evaluated with exponentially spaced frequency bins. Afterwards, the magnitude response in decibels is calculated as

$$|H(k)|_\mathrm{dB} = 20\log_{10}|H(k)|. \tag{3.23}$$

Additionally, the ITD is estimated from the HRIR $h(n)$ or the corresponding HRTF $H(k)$ based on the methods explained in Section 2.1.4. Although the ITD is not part of the following approximation procedure, it has to be saved for usage in the binaural synthesis with approximated HRTF magnitudes. Contrarily, the magnitude response $|H(k)|_\mathrm{dB}$ is further pre-processed. In order to reduce the fluctuations in the magnitude response, the magnitude response is smoothed in the next step using $1/12^{\mathrm{th}}$-octave smoothing

$$|\tilde{H}(k)|_\mathrm{dB} = \sum_{\kappa=0}^{\max(k)} w(k,\kappa) \cdot |H(\kappa)|_\mathrm{dB}, \tag{3.24}$$

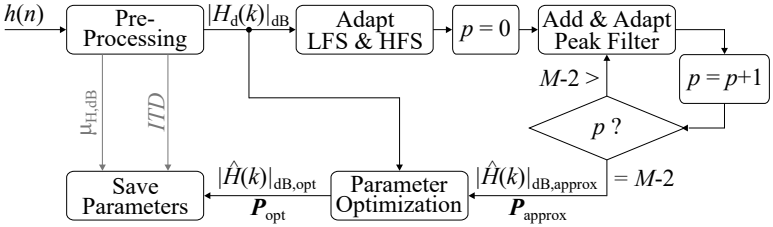



**Figure 3.6:** Flow chart of the approximation procedure using one LFS, one HFS, and $M-2$ peak filters. Here, $h(n)$ defines the original HRIR, $|H_\mathrm{d}(k)|_\mathrm{dB}$ the desired magnitude response in decibels, $|\hat{H}(k)|_\mathrm{dB,approx}$ the approximated magnitude response, $\mathbf{P}_\mathrm{approx}$ the parameter matrix, $|\hat{H}(k)|_\mathrm{dB,opt}$ the optimized magnitude response, and $\mathbf{P}_\mathrm{opt}$ the parameter matrix after optimization. The loop on the right is run through $M-2$ times.

with normalized weighting factor

$$w(k,\kappa) = \frac{e^{-\frac{(f_\kappa - f_\mathrm{k})^2}{2\sigma_\mathrm{k}^2}}}{\sum_{\kappa=0}^{\max(k)} w(k,\kappa)}, \tag{3.25}$$

standard deviation of the Gaussian function around frequency $f_\mathrm{k}$

$$\sigma_\mathrm{k} = \frac{f_\mathrm{k}}{\pi \cdot N_\mathrm{oct}}, \tag{3.26}$$

and order of octave smoothing $N_\mathrm{oct} = 12$. From this smoothed magnitude response $|\tilde{H}(k)|_\mathrm{dB}$, the mean value

$$\mu_\mathrm{H,dB} = \mathrm{mean}|\tilde{H}(k)|_\mathrm{dB} \tag{3.27}$$

of the smoothed magnitude response is calculated. Finally, the desired magnitude response

$$|H_\mathrm{d}(k)|_\mathrm{dB} = |\tilde{H}(k)|_\mathrm{dB} - \mu_\mathrm{H,dB} \tag{3.28}$$

is achieved by subtracting the mean value $\mu_\mathrm{H,dB}$ from the smoothed magnitude response $|\tilde{H}(k)|_\mathrm{dB}$.

Once the pre-processing is done, the shelving filters are adapted to approximate the behavior of the desired magnitude response $|H_\mathrm{d}(k)|_\mathrm{dB}$ in low and high frequencies (see Fig. 3.8). Firstly, the gain of the LFS

$$G_\mathrm{L} = |H_\mathrm{d}(0)|_\mathrm{dB} \tag{3.29}$$

is initialized as magnitude at frequency $f_0$ from Eq. (3.22). Secondly, the

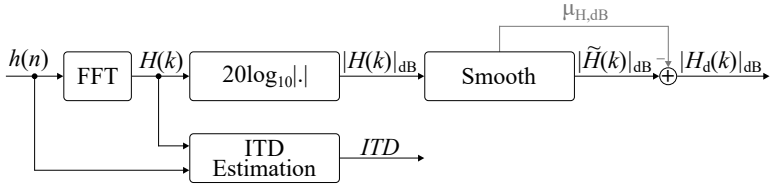



**Figure 3.7:** Flow chart of the pre-processing steps performed on the original HRIR $h(n)$. Firstly, the ITD is estimated and the magnitude response $|H(k)|_{\mathrm{dB}}$ is calculated. Then, the magnitude response is smoothed and the mean is calculated. From this, the desired magnitude response $|H_d(k)|_{\mathrm{dB}}$ is achieved by subtracting the mean $\mu_{\mathrm{H,dB}}$ from the smoothed magnitude response $|\tilde{H}(k)|_{\mathrm{dB}}$.

cut-off frequency $f_{c,\mathrm{L}}$ of the LFS is linearly increased between 20 Hz and 2 kHz for a total number of 500 steps. For every iteration, the approximation error

$$E_{\mathrm{dB}} = \sum_{k=0}^{\max(k)} \left(E_{\mathrm{dB}}(k)\right)^2, \tag{3.30}$$

with

$$E_{\mathrm{dB}}(k) = |H_{\mathrm{d}}(k)|_{\mathrm{dB}} - |\hat{H}(k)|_{\mathrm{dB}}, \tag{3.31}$$

is calculated between desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ and approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB}} = |\hat{H}(k)|_{\mathrm{dB,L}}$ using only the LFS. Finally, the cut-off frequency that minimizes the approximation error $E_{\mathrm{dB}}$ in Eq. (3.30) is taken as initial cut-off frequency $f_{c,\mathrm{L}}$ of the LFS. The remaining approximation error

$$E_{\mathrm{dB,L}}(k) = |H_{\mathrm{d}}(k)|_{\mathrm{dB}} - |\hat{H}(k)|_{\mathrm{dB,L}} \tag{3.32}$$

is then used to initialize the gain of the HFS

$$G_{\mathrm{H}} = E_{\mathrm{dB,L}}(\max(k)) \tag{3.33}$$

as magnitude of the approximation error at the highest evaluated frequency $f_{\max(k)}$. Similar to the adaption of the LFS, the initial cut-off frequency $f_{c,\mathrm{H}}$ of the HFS is found between 5 and 20 kHz by taking the frequency that minimizes the approximation error. As output of the adaptation procedure of the shelving filters, the approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB},0}$ by using only LFS and HFS, and the parameter matrix $\mathbf{P}_0$, including $G_{\mathrm{L}}$, $f_{c,\mathrm{L}}$, $G_{\mathrm{H}}$, and $f_{c,\mathrm{H}}$, are obtained.

After the shelving filters are set, peak filters are added and initialized

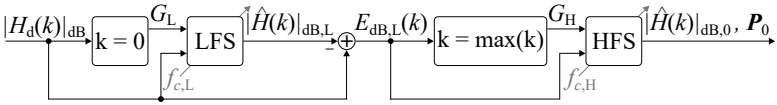



**Figure 3.8:** Flow chart of the adaptation procedure of LFS and HFS, with $|H_\mathrm{d}(k)|_\mathrm{dB}$ defining the desired magnitude response in decibels, $|\hat{H}(k)|_\mathrm{dB,L}$ the approximated magnitude response with only LFS, $E_\mathrm{dB,L}(k)$ the approximation error with only LFS, and $|\hat{H}(k)|_\mathrm{dB,0}$ the approximated magnitude response with both shelving filters. Additionally, $\mathbf{P}_0$ stores the shelving filter parameters including gain $G_\mathrm{L}$ and cut-off frequency $f_{c,\mathrm{L}}$ of LFS, and gain $G_\mathrm{H}$ and cut-off frequency $f_{c,\mathrm{H}}$ of HFS.

consecutively until a maximum number of $p = M-2$ peak filters is reached (see Fig. 3.6). The flow chart of the procedure for adding and adapting the $\{p{+}1\}^{\mathrm{th}}$ peak filter is shown in Fig. 3.9, where $p$ defines the current number of used peak filters. In a first step, the current approximation error $E_\mathrm{dB,p}(k)$ is calculated from the desired magnitude response $|H_\mathrm{d}(k)|_\mathrm{dB}$ and the approximated magnitude response $|\hat{H}(k)|_\mathrm{dB,p}$ by using both shelving filters and the $p$ previously adapted peak filters. Hence, for adding the first peak filter, the approximated magnitude response $|\hat{H}(k)|_\mathrm{dB,0}$ by using only shelving filters is used as current approximation. From this approximation error $E_\mathrm{dB,p}(k)$, center frequency $f_{c,\mathrm{p+1}}$ and gain $G_\mathrm{p+1}$ of the new peak filter are initialized as

$$f_{c,\mathrm{p+1}} = f_{\mathrm{k_{max}}}, \tag{3.34}$$

$$G_\mathrm{p+1} = E_\mathrm{dB,p}(k_\mathrm{max}), \tag{3.35}$$

respectively, where $k_\mathrm{max}$ defines the frequency bin index of the maximum absolute error

$$k_\mathrm{max} = \arg\max_k |E_\mathrm{dB,p}(k)| \tag{3.36}$$

for the current number of used peak filters. Analogous to determine the cut-off frequencies of the shelving filters, the Q-factor of the new peak filter is searched with 500 linear steps between $Q = 1$ and $Q = 100$. Then, the Q-factor that minimizes the approximation error $E_\mathrm{dB,p+1}$ is taken as initial parameter $Q_\mathrm{p+1}$. Finally, the parameters of the new peak filter ($f_{c,\mathrm{p+1}}$, $G_\mathrm{p+1}$, and $Q_\mathrm{p+1}$) are stored together with the previously adapted peak and shelving filters in parameter matrix $\mathbf{P}_\mathrm{p+1}$. Additionally, the approximated magnitude response $|\hat{H}(k)|_\mathrm{dB,p+1}$ using both shelving filters and $p{+}1$ peak filters is calculated.

When having adapted $M-2$ peak filters, the loop of adding and adapting

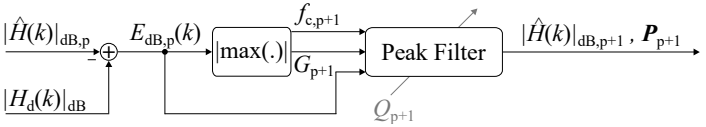



**Figure 3.9:** Flow chart of the procedure for adding and adapting the $\{p+1\}^{\mathrm{th}}$ peak filter. Here, $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ defines the desired magnitude response in decibels, $|\hat{H}(k)|_{\mathrm{dB,p}}$ the approximated magnitude response using both shelving filters and $p$ peak filters, $E_{\mathrm{dB,p}}(k)$ the approximation error using both shelving filters and $p$ peak filters, and $|\hat{H}(k)|_{\mathrm{dB,p+1}}$ the approximated magnitude response using both shelving filters and $p+1$ peak filters. Additionally, $\mathbf{P}_{\mathrm{p+1}}$ stores the shelving filter parameters and the peak filter parameters $G_{\mathrm{p+1}}$, $f_{c,\mathrm{p+1}}$, and $Q_{\mathrm{p+1}}$.

new peak filters is stopped and the approximation result using $M$ parametric IIR filters is achieved. This result includes the parameter matrix

$$\mathbf{P}_{\mathrm{approx}} = \begin{bmatrix} f_{c,\mathrm{L}} & G_{\mathrm{L}} & 1 \\ f_{c,1} & G_1 & Q_1 \\ \vdots & \vdots & \vdots \\ f_{c,\mathrm{M\text{-}2}} & G_{\mathrm{M\text{-}2}} & Q_{\mathrm{M\text{-}2}} \\ f_{c,\mathrm{H}} & G_{\mathrm{H}} & 1 \end{bmatrix} \tag{3.37}$$

and the approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB,approx}}$. However, due to the consecutive addition of new peak filters, previous peak filters in the close environment of the new peak filter can be disturbed by the influence of this new filter, such that the interaction between individual filters has to be optimized in the final stage. Therefore, a parameter optimization has to be performed based on the current parameter matrix $\mathbf{P}_{\mathrm{approx}}$ and the desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$. Hereby, e.g. the Levenberg-Marquardt algorithm [Levenberg, 1944, Marquardt, 1963] can be used. The outputs of this optimization step are the optimized parameter matrix $\mathbf{P}_{\mathrm{opt}}$ and the optimized approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB,opt}}$. Finally, the parameters that are needed for a usage of the IIR filter cascade in binaural synthesis are saved. This includes the parameter matrix $\mathbf{P}_{\mathrm{opt}}$, the mean value of the magnitude response $\mu_{\mathrm{H,dB}}$, and the estimated ITD.

In the following, HRIRs taken from the CIPIC database [Algazi et al., 2001b] are used to evaluate the approximation algorithm. This database contains HRIRs for 45 subjects and a total number of 1250 directions. The directional information of the measured HRIRs is given in the interaural-polar coordinate system and includes 25 azimuths $\varphi$ and 50 elevations $\theta$. Here, the azimuthal resolution is $\Delta\varphi = 5°$ at centered directions ($|\varphi| \leq 45°$).

Additionally, lateral directions are included at $\varphi = \pm 80°$, $\varphi = \pm 65°$, and $\varphi = \pm 55°$. Contrarily, elevations are uniformly distributed with a resolution of $\Delta\theta = 5.625°$ along the whole range between $\theta = -45°$ and $\theta = 225°$. Note that in the following evaluation, approximations are performed for an azimuthal resolution of $\Delta\varphi = 15°$ for center directions inside $|\varphi| \leq 45°$ and all lateral azimuths. Furthermore, elevations are evaluated for a resolution of $\Delta\theta = 45°$, resulting in a total of 91 evaluated directions. Additionally, the results for the two ears are combined by using a relative azimuth $\varphi_{\mathrm{rel}}$, which is positive for ipsilateral directions and negative for contralateral directions. In this way, an augmented data set of 90 magnitude responses per direction is created.

### 3.3.1 Minimum Number of Required Peak Filters

In this section, the minimum number of peak filters needed to approximate HRTF magnitudes within a given error tolerance is evaluated for different subjects and directions [Nowak et al., 2020a]. For this purpose, the previously explained augmented data set is used, including 90 subjects and 91 directions. The approximation procedure in order to find the minimum number of peak filters needed to fulfill a given error tolerance is based on the approximation procedure for a fixed number of peak filters explained in Section 3.3. However, due to the new target of the approximation algorithm, some parts have to be changed. The modified flow chart is shown in Fig. 3.10. Comparing Figs. 3.10 and 3.6 shows two major differences between the two approximation algorithms. Firstly, the condition of the while loop is changed from having a maximum number of $p = M - 2$ peak filters to a limit that the maximum approximation error $\max|E_{\mathrm{dB}}(k)|$ has to fall below. This limit is given by the error tolerance $E_{\mathrm{tol}}$. In this way, peak filters are consecutively added until the condition

$$\max|E_{\mathrm{dB}}(k)| \leq E_{\mathrm{tol}} \tag{3.38}$$

is fulfilled and the maximum approximation error $\max|E_{\mathrm{dB}}(k)|$ is inside the given error tolerance $E_{\mathrm{tol}}$. As a second termination condition of the while loop, a maximum number of 30 peak filters is chosen. Secondly, the parameter optimization is moved into the while loop in order to immediately post-optimize the whole cascade including the latest added peak filter. This optimization of interaction between individual filters yields the best approximation for the current number of used peak filters. Thus, the maximum approximation error is calculated for the optimized filter cascade instead of the simply extended cascade. Note that the target of the post-optimization algorithm is to reduce the overall average error rather than reducing the maximum error. Except for these two changes in the general principle of the approximation algorithm, the working principle

of the individual blocks, namely pre-processing, adaptation of shelving filters, addition and adaptation of peak filters, parameter optimization, and saving of the parameters, is unchanged. Here, the dimension of the saved parameter matrix $\mathbf{P}_M$ varies between subjects and directions due to different numbers of required peak filters $M-2$.

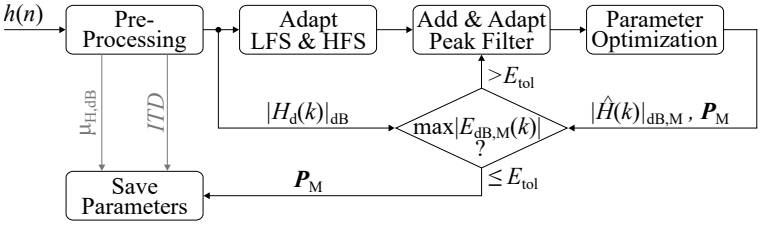



**Figure 3.10:** Flow chart of the approximation procedure for finding the minimum number of peak filters needed to approximate an HRTF magnitude within a given error tolerance $E_{\text{tol}}$. Here, $h(n)$ defines the original HRIR, $|H_{\text{d}}(k)|_{\text{dB}}$ the desired magnitude response in decibels, $|\hat{H}(k)|_{\text{dB,M}}$ the approximated magnitude response using $M-2$ peak filters, $\mathbf{P}_M$ the parameter matrix including $M-2$ peak filters, $E(k)_{\text{dB}}$ the approximation error in decibels. The iterative procedure stops, when the maximum absolute approximation error $\max|E(k)_{\text{dB}}|$ is inside the given error tolerance $E_{\text{tol}}$.

Blommer and Wakefield [Blommer and Wakefield, 1997] stated that for frequencies above 200 Hz differences of 2.4 to 2.7 dB between measured and approximated HRTF magnitudes are inside the discrimination threshold of human subjects. Thus, in [Nowak et al., 2020a] an error tolerance of $E_{\text{tol}} = 2\,\text{dB}$ was chosen for the evaluation of the minimum number of peak filters needed to approximate the HRTF magnitudes of the given data set. In the following, the results from [Nowak et al., 2020a] are summarized and extended.

In Fig. 3.11, an exemplary approximation result is shown, where the left ear's magnitude response of *Subject_008* for the frontal direction ($\varphi = 0°$, $\theta = 0°$) is taken as desired magnitude response $|H_{\text{d}}(k)|_{\text{dB}}$. The approximated magnitude response $|\hat{H}(k)|_{\text{dB}}$ clearly indicates that using an approximation with eight peak filters achieves a magnitude response that is close to the desired one. As can be seen from the approximation error $E(k)_{\text{dB}}$, these eight peak filters are already sufficient to fulfill the given error tolerance $E_{\text{tol}} = 2\,\text{dB}$.

Looking at Fig. 3.12, more detailed information can be gathered about the used shelving and peak filters. It shows that the LFS has a gain close to zero, whereas the HFS has a gain of $G_{\text{H}} = -45\,\text{dB}$ and a cut-off frequency of $f_{c,\text{H}} = 22\,\text{kHz}$. This high gain originates from the uncon-

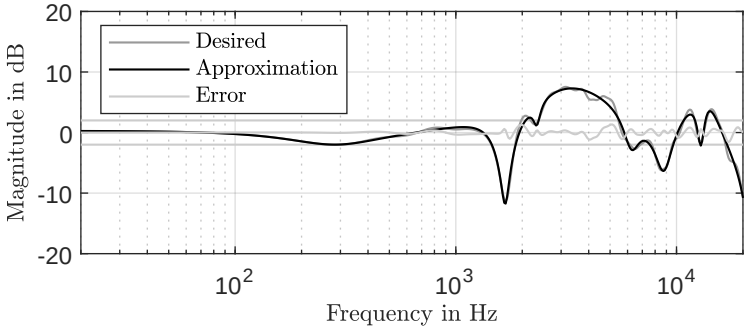



**Figure 3.11:** Approximation of the left ear's HRTF magnitude of *Subject_008* for the frontal direction ($\varphi = 0°$, $\theta = 0°$), with the desired magnitude response $|H_\mathrm{d}(k)|_\mathrm{dB}$ and the approximation $|\hat{H}(k)|_\mathrm{dB}$ using eight peak filters. Additionally, the error tolerance $E_\mathrm{tol} = 2\,\mathrm{dB}$ and the approximation error $E(k)_\mathrm{dB}$ are plotted.

strained Levenberg-Marquardt algorithm that is used as post-optimization method. For the implementation of the Levenberg-Marquardt algorithm, MATLAB's nonlinear curve-fitting function `lsqcurvefit` is used. In order to approximate the entire frequency range, the center frequencies of the eight peak filters are spread from 287 Hz to 13 kHz. Additionally, the gains of the peak filters are between −16 and 10.5 dB, and the Q-factors between 0.9 for the peak filter at $f_c = 287\,\mathrm{Hz}$ and 15.9 for the peak filter with negative gain at $f_c = 12.9\,\mathrm{kHz}$.

Although eight peak filters are sufficient to approximate the given magnitude response (*Subject_008*, left ear, $\varphi = 0°$, and $\theta = 0°$) within an error tolerance of 2 dB, this number of peak filters is not enough for an adequate approximation for all directions and subjects. Thus, the minimum number of required peak filters to fulfill the given error tolerance of $E_\mathrm{tol} = 2\,\mathrm{dB}$ is evaluated for all 45 subjects from the CIPIC database [Algazi et al., 2001b]. For all subjects, 91 directions are evaluated per ear, such that a total number of 8190 approximated magnitude responses arises. Figure 3.13 shows the evaluation of the number of required peak filters for all of these approximations in a histogram. Here, the height of the bar gives the relative frequency of a specific number of required peak filters. The highest relative frequencies are achieved by required numbers of seven (19.07 %) and eight (18.55 %) peak filters. These relative frequencies correspond to absolute frequencies of 1562 and 1519, respectively. The lowest number of required peak filters is three, which is sufficient for nine approximations (0.11 %). Contrarily, a number of 16 to 20 peak filters is

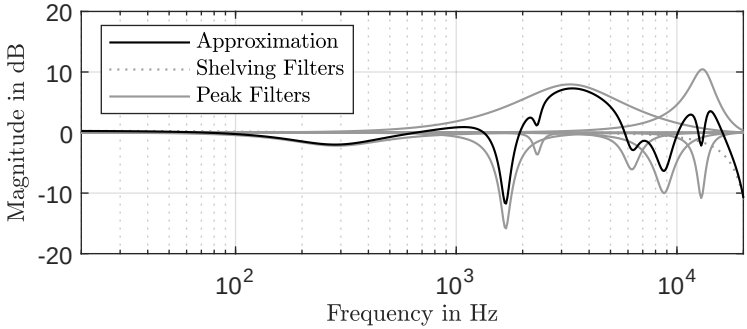



**Figure 3.12:** Approximation of the left ear's HRTF magnitude of *Subject_008* for the frontal direction ($\varphi = 0°$, $\theta = 0°$), with the approximated magnitude response $|\hat{H}(k)|_{\text{dB}}$ from Fig. 3.11 and the magnitude responses of the ten individual filter stages.

needed by 28 approximations (0.34 %). Additionally, seven approximations (0.09 %) fail to estimate the desired magnitude response and are terminated by the maximum number of 30 peak filters. The reason for the failure is the post-optimization by the Levenberg-Marquardt, which moves the center frequencies of the peak filters to the same frequency outside of the evaluated frequency range (20 Hz - 20 kHz). These peak filters have center frequencies higher than $f_c > 20\,\text{kHz}$ and Q-factors greater than $Q > 50$. Since the movement starts already at a small number of used peak filters, these approximations lack of an accurate estimation of the desired magnitude response and result in high approximation errors. One commonality of these seven approximations is that the sound source is located below the head of the subject either in the front ($\theta = -45°$) or the back ($\theta = 225°$). Additionally, when using a much longer unit impulse duration of 2 s for calculating the impulse responses of the parametric IIR filters, 21 (0.26 %) of the 8190 approximations are unstable. The reason for these instabilities are unconstrained parameter optimizations due to the Levenberg-Marquardt algorithm that result primarily in cut-off frequencies of the HFS slightly higher than $f_s/2$. However, also a few small negative cut-off frequencies of the LFS and peak filter bandwidths equal to $f_s/2$ or negative multiples of $f_s/2$ cause unstable impulse responses. Nevertheless, these issue can be solved by post-processing the parameters or constraining the post-optimization algorithm.

In addition to the relative frequency of the number of required peak filters, also the cumulative frequency is given in Fig. 3.13. This cumulative

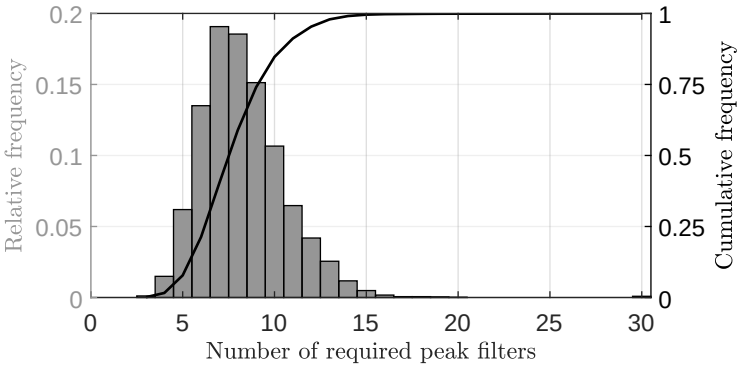
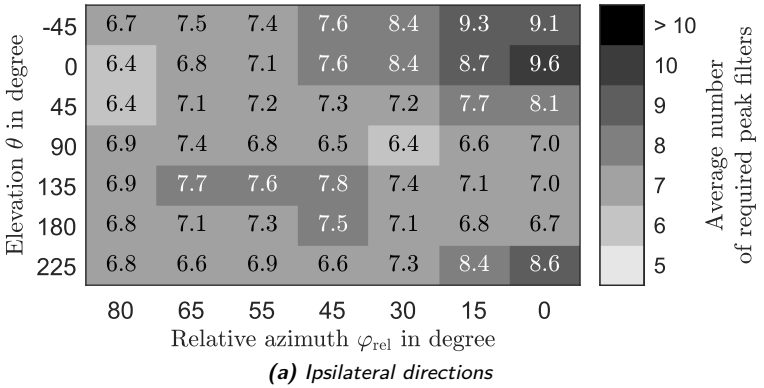



**Figure 3.13:** Histogram of relative frequencies of the required number of peak filters over all of the 8190 approximations. Additionally, the cumulative frequency is plotted.

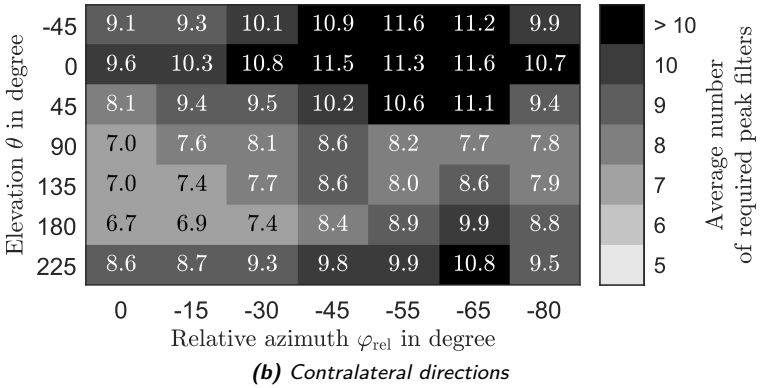
frequency gives the incidence of approximations that require a given number of peak filters or less. Having a fixed number of used peak filters, the cumulative frequency determines the percentage of approximations that still fulfill the given error tolerance $E_{\mathrm{tol}}$. In order to have at least 50 % of the approximations fulfilling the given error tolerance $E_{\mathrm{tol}} = 2\,\mathrm{dB}$, eight peak filters are needed. When using ten peak filters, 84.7 % of the approximations are within the given error tolerance. Further increasing the number of used peak filters up to 12 or 14 gives cumulative frequencies of 95.3 % or 99.1 %, respectively.

Because of the wide spreading of number of required peak filters between three and twenty, a more detailed evaluation is needed. Therefore, Fig. 3.14 separates the evaluation of the required number of peak filters into the 91 different directions. In Fig. 3.14, a heatmap gives the average number of required peak filters for every direction. As described before, the relative azimuth $\varphi_{\mathrm{rel}}$ is used for the horizontal position in order to combine left and right ear data in the evaluation. Here, the relative azimuth is defined as $\varphi_{\mathrm{rel}} = \varphi$ for the right ear and $\varphi_{\mathrm{rel}} = -\varphi$ for the left ear. In this way, positive and negative relative azimuths give ipsilateral and contralateral directions, respectively. For a better representation of the results, Fig. 3.14 is split into two subfigures, having the results for ipsilateral directions in Fig. 3.14(a) and contralateral directions in Fig. 3.14(b). The column for the center directions ($\varphi_{\mathrm{rel}} = 0°$) is included in both of them.

For ipsilateral directions ($\varphi_{\mathrm{rel}} \geq 45°$) in Fig. 3.14(a), the average numbers of required peak filters range from 6.4 to 7.8. In contrast to this, con-




***(a)** Ipsilateral directions*



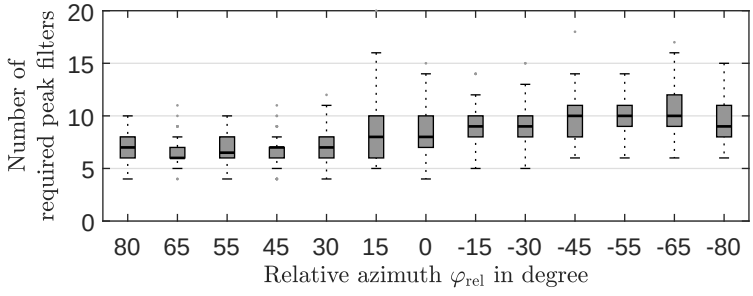

***(b)** Contralateral directions*

**Figure 3.14:** Heatmaps of average number of needed peak filters per direction for 90 subjects separated into (a) ipsilateral and (b) contralateral directions.

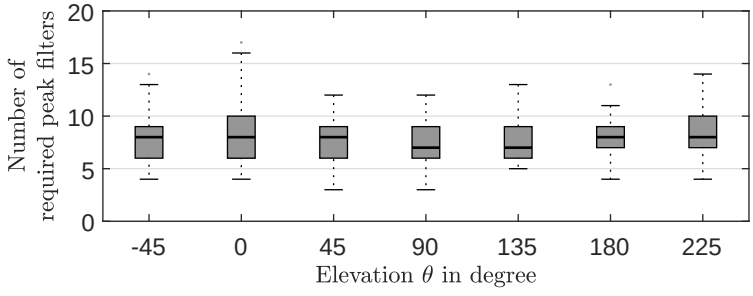
tralateral directions ($\varphi_{\mathrm{rel}} \leq -45°$) require average numbers of 7.8 to 11.6 peak filters (see Fig. 3.14(b)). Additionally, center directions ($|\varphi_{\mathrm{rel}}| < 45°$) need on average 6.4 to 10.8 peak filters in order to achieve approximations within an error tolerance of $E_{\mathrm{tol}} = 2\,\mathrm{dB}$. When moving from ipsilateral to contralateral directions, a rise in the average number of needed peak filters can be seen. For ipsilateral directions only small differences are seen with changes in elevation, whereas contralateral directions tend to need less peak filters for rear directions. The highest number of peak filters is required for magnitude responses that originate from frontal contralateral directions. In this region, on average more than ten peak filters are needed.

Since outliers can strongly affect the average value, also other statistical values should be evaluated. Thus, in Figs. 3.15 and 3.16, box-and-whisker

plots are shown for different sets of directions. Each of the box-and-whisker plots contains the minimum value, the first quartile, the median, the third quartile, and the maximum value. The distance between the third and first quartile is called interquartile range (IQR). From this, the minimum and maximum values are defined as lowest or highest data points, respectively, which fall into a distance of 1.5 times the IQR from the corresponding quartile. Data points outside of this region are stated as outliers and represented as dots.




**(a)** *Azimuth*




**(b)** *Elevation*

**Figure 3.15:** Box-and-whisker plots of the required number of peak filters per (a) relative azimuth $\varphi_{\mathrm{rel}}$ and (b) elevation $\theta$. Each box-and-whisker plot contains information about minimum value, first quartile, median, third quartile, maximum value, and outliers.

In Fig. 3.15, the number of required peak filters is analyzed for azimuth or elevation. In Fig. 3.15(a), one box-and-whisker plot represents a single relative azimuth for all elevations and all subjects, resulting in a total number of 630 data points per box-and-whisker plot. The general trend of requiring more peak filters when moving from ipsilateral to contralateral directions

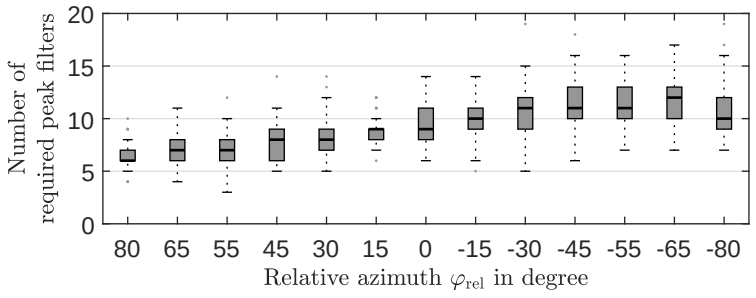
that has already been seen in Fig. 3.14 is confirmed by the box-and-whisker plots in Fig. 3.15(a). Additionally, Fig. 3.15(a) gives information about the spread of the number of required peak filters for different azimuths. For ipsilateral directions ($\varphi_{\mathrm{rel}} \geq 45°$) only a few outliers need eleven peak filters. The third quartile, meaning 75 % of the approximations, not even exceeds eight peak filters. Contrarily, for contralateral directions ($\varphi_{\mathrm{rel}} \leq -45°$), the median reaches values up to ten peak filters and the third quartile values up to twelve peak filters. Additionally, the IQR increases from 1 or 2 peak filters for ipsilateral directions to 2 or 3 peak filters for contralateral directions. By grouping the approximation results in different elevations, 1170 data points are merged in a single box-and-whisker plot in Fig. 3.15(b). All elevations show median values of seven or eight required peak filters. Also values for the first and third quartile vary only by a single peak filter between different elevations, such that only small differences in the required number of peak filters are visible between different elevations.

Although Fig. 3.15(b) has not shown a clear difference in the required number of peak filters between elevations, Fig. 3.14(b) has indicated a strong effect of the elevation on the required number of peak filters for contralateral directions. Therefore, Fig. 3.16 shows the changes in the number of required peak filters with elevation separated into frontal ($\theta = 0°$) and rear directions ($\theta = 180°$). By comparing Figs. 3.16(a) and 3.16(b), the influence of the elevation seen already in Fig. 3.14(b) is clearly visible in the box-and-whisker plots. Starting at frontal ipsilateral directions $\varphi_{\mathrm{rel}} = 15°$ and moving to contralateral directions, the boxes between first and third quartile are lifted to higher numbers of required peak filters for frontal directions in comparison to rear directions. These boxes do not even overlap for most of the relative azimuths, indicating that magnitude responses for contralateral directions in the front need consistently more peak filters in order to be approximated with a given error tolerance than magnitude responses of contralateral directions in the back.

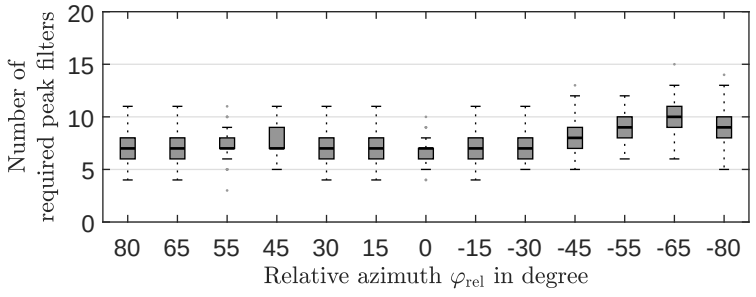
With the given approximation algorithm, the maximum approximation error $\max|E_{\mathrm{dB}}(k)|$ falls in 99.91 % of the magnitude responses within the given error tolerance $E_{\mathrm{tol}} = 2\,\mathrm{dB}$. However, not only the maximum error of the achieved approximations is of interest, also the average magnitude error across frequency is important. For this, the log-spectral distance (LSD) can be used as metric. The LSD in decibels is calculated from the spectral difference between desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ and approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB}}$ as

$$\mathrm{LSD} = \sqrt{\frac{1}{\max(k)+1} \sum_{k=0}^{\max(k)} \left(|H_{\mathrm{d}}(k)|_{\mathrm{dB}} - |\hat{H}(k)|_{\mathrm{dB}}\right)^2}. \tag{3.39}$$

As can be seen, the sum in Eq. (3.39) equates to the definition of the




**(a)** *Front* ($\theta = 0°$)




**(b)** *Back* ($\theta = 180°$)

**Figure 3.16:** Box-and-whisker plots of the required number of peak filters per relative azimuth $\varphi_{\mathrm{rel}}$ for (a) the front ($\theta = 0°$) and (b) the back $\theta = 180°$. Each box-and-whisker plot contains information about minimum value, first quartile, median, third quartile, maximum value, and outliers.

approximation error $E_{\mathrm{dB}}$ in Eq. (3.30). In order to evaluate the performance of the approximation algorithm, the calculated LSD values of the individual approximations are averaged per direction. These average LSD values are shown in Fig. 3.17. The average LSD values per direction range from 0.45 to 0.66 dB with an overall mean of 0.52 dB. The maximum value of 0.66 dB is reached for a relative azimuth of $\varphi_{\mathrm{rel}} = 15°$ and an elevation of $\theta = -45°$, because two of the seven failed approximations fall into this direction. With LSD values of 3.7 dB and 4.0 dB, these approximations strongly increase the average LSD. Except for this outlier, the other average LSD values do not exceed 0.59 dB, which gives a difference in average LSD values of only 0.11 dB between different directions.

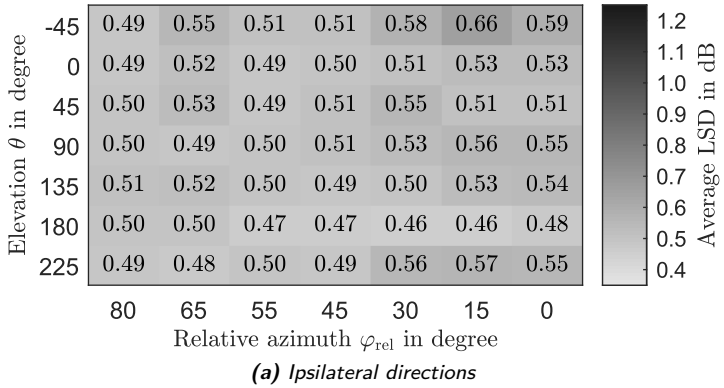


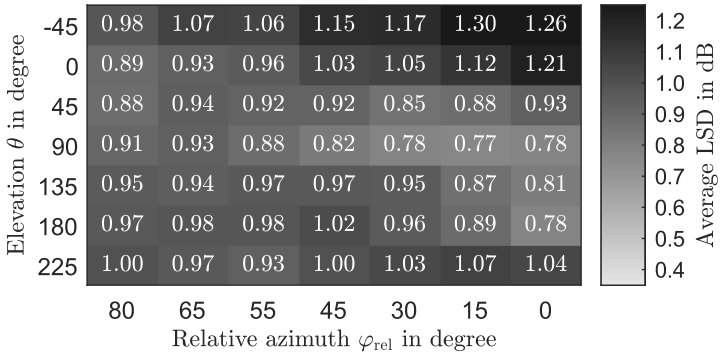

**Figure 3.17:** Heatmaps of average LSD using three to thirty peak filters per direction for 90 subjects separated into (a) ipsilateral and (b) contralateral directions. The overall mean is 0.52 dB.

### 3.3.2 Approximation Using Ten Peak Filters

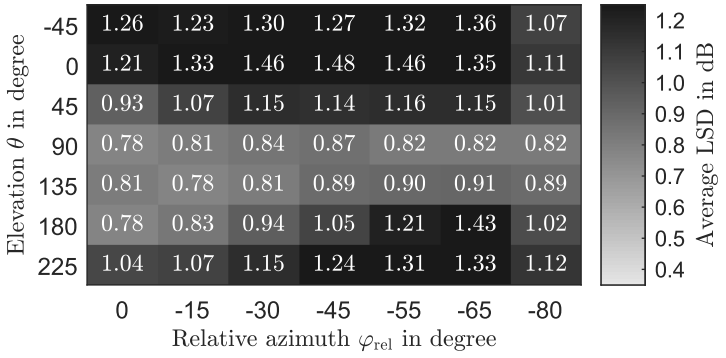
According to Section 3.3.1, ten peak filters are sufficient to approximate 84.7 % of the magnitude responses from the given data set within an error tolerance of $E_{\mathrm{tol}} = 2\,\mathrm{dB}$. Since the implementation of binaural synthesis through headphones requires a fixed number of filters, the HRTF magnitude approximation using ten peak filters for every direction is evaluated in this subsection. For this evaluation, the same subset from the CIPIC database [Algazi et al., 2001b] is used as described in Section 3.3.

The approximation procedure follows the flow chart given in Fig. 3.6 with a total number of $M = 12$ filter stages, including one LFS, one HFS, and ten

peak filters. Firstly, the approximated magnitude responses $|\hat{H}(k)|_{\mathrm{dB,approx}}$ before post-optimization are evaluated. Afterwards, the post-optimization is included in order to show the improvements due to optimized interactions between individual filter stages.




***(a)** Ipsilateral directions*

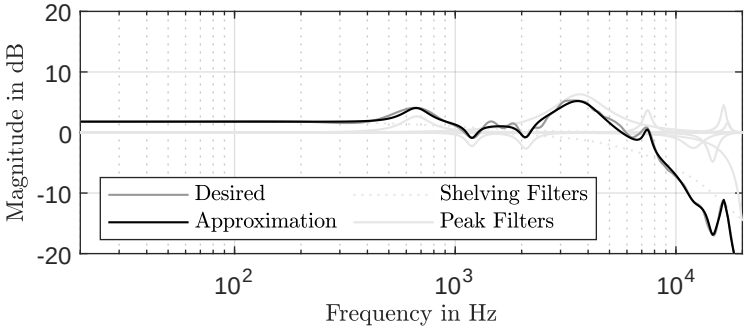



***(b)** Contralateral directions*

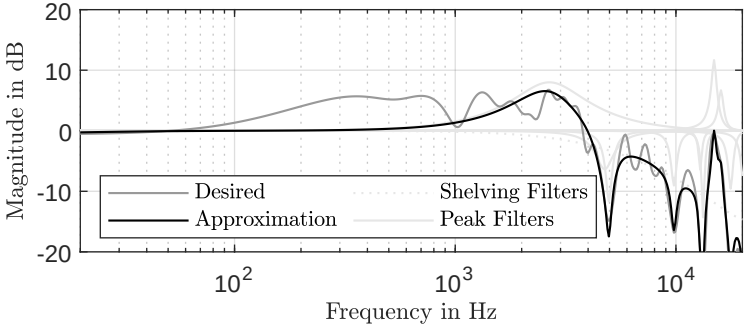
**Figure 3.18:** Heatmaps of average LSD using ten peak filters per direction without post-optimization for 90 subjects separated into (a) ipsilateral and (b) contralateral directions. The overall mean is 1.03 dB.

When using the approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB,approx}}$, average LSD values of 0.77 to 1.48 dB are achieved between approximation and desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ for the different directions. The mean LSD across all directions and subjects is 1.03 dB. The average LSD values per direction are given in the heatmaps shown in Fig. 3.18. The ipsilateral directions ($\varphi_{\mathrm{rel}} \geq 45°$) in Fig. 3.18(a) show an average LSD of 0.96 dB, whereas the contralateral directions ($\varphi_{\mathrm{rel}} \leq 45°$) in Fig. 3.18(b)

show an average LSD of 1.13 dB. For center directions ($|\varphi_{\mathrm{rel}}| < 45°$), the approximations achieve a mean LSD of 1.01 dB. When comparing the heatmaps in Fig. 3.18 with the ones in Fig. 3.14, similarities can be seen in the position of the bright and dark regions. The dark regions in Fig. 3.18(b) with an average LSD higher than 1.2 dB are the result of a non-sufficient number of used peak filters for approximation. Since in Fig. 3.14(b) these directions have shown average numbers of required peak filters higher than ten, the ten peak filters that are used here lead to an increased LSD for these directions in comparison to the others.




***(a)** Lowest LSD of 0.33 dB*




***(b)** Highest LSD of 2.77 dB*

**Figure 3.19:** Approximations with (a) the lowest LSD of 0.33 dB (*Subject_162*, right ear, $\varphi_{\mathrm{rel}} = -65°$, $\theta = 90°$) and (b) the highest LSD of 2.77 dB (*Subject_044*, right ear, $\varphi_{\mathrm{rel}} = -55°$, $\theta = -45°$) between desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ and approximation $|\hat{H}(k)|_{\mathrm{dB,approx}}$. The magnitude responses of the individual filter stages are shown in the background.

The LSD values of the individual approximations range from 0.33 dB for the right ear of *Subject_162* for a direction of $\varphi_{\mathrm{rel}} = -65^\circ$ and $\theta = 90^\circ$ to 2.77 dB for the right ear of *Subject_044* for a direction of $\varphi_{\mathrm{rel}} = -55^\circ$ and $\theta = -45^\circ$. In Fig. 3.19, the approximated magnitude responses $|\hat{H}(k)|_{\mathrm{dB,approx}}$ of these extreme cases are shown together with the desired magnitude responses $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ and the magnitude responses of the individual filter stages. From the two subplots, the differences in the accuracy of the approximations is clearly visible. In Fig. 3.19(a), where an LSD of 0.33 dB is achieved, only small deviations in some frequency regions between desired and approximated magnitude response are visible. Contrarily, in Fig. 3.19(b), very strong variations between the two magnitude responses can be seen. Especially for frequencies between 70 Hz and roughly 1 kHz, errors of up to 5.5 dB occur. Not only these low frequencies lack of a proper approximation, also other frequency bands show deviations of up to 4.5 dB. All these errors sum up to the given LSD of 2.77 dB.

In order to reduce the average LSD by improving the interaction between the individual filter stages, MATLAB's nonlinear curve-fitting function `lsqcurvefit` is used to implement the Levenberg-Marquardt algorithm as post-optimization algorithm.

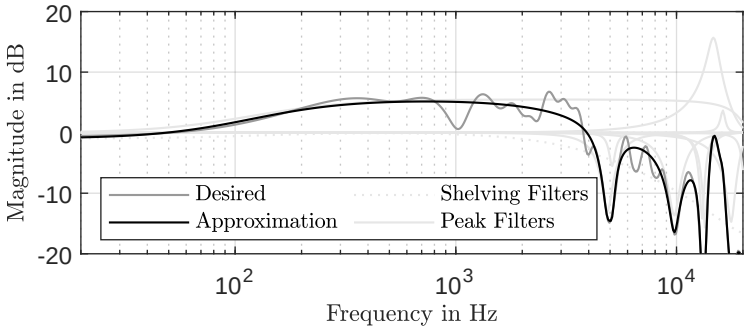
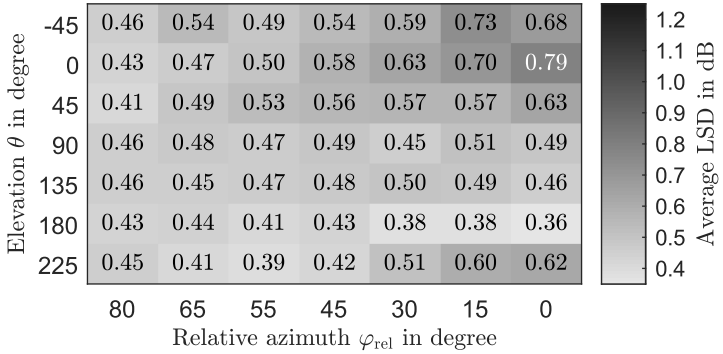


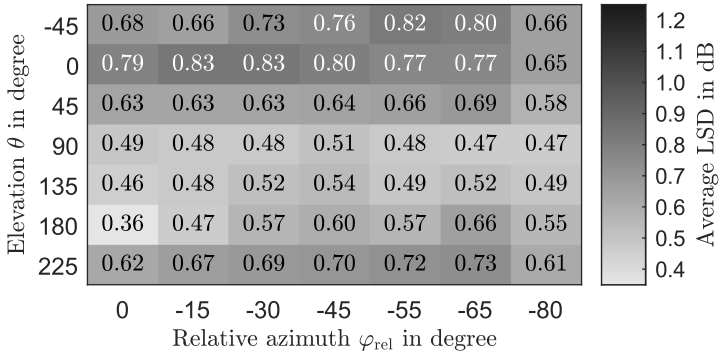

**Figure 3.20:** Post-optimization of the approximation shown in Fig. 3.19(b). The post-optimized approximated magnitude response $|\hat{H}(k)|_{\mathrm{dB,opt}}$ achieves an LSD of only 1.21 dB. The magnitude responses of the individual filter stages are shown in the background.

In Fig. 3.20, the post-optimized magnitude response $|\hat{H}(k)|_{\mathrm{dB,opt}}$ of the approximation from Fig. 3.19(b) is shown. It is clearly visible that the approximation has improved especially in the low frequency region below 800 Hz, where previously errors of up to 5.5 dB occurred. The reason for this improvement is the widening of the peak filter at approximately 2.7 kHz during post-optimization procedure. Here, the Q-Factor of the

peak filter is reduced from $Q = 1.5$ to $Q = 0.025$. In this way, the LSD of the approximation decreases from 2.77 to 1.21 dB.



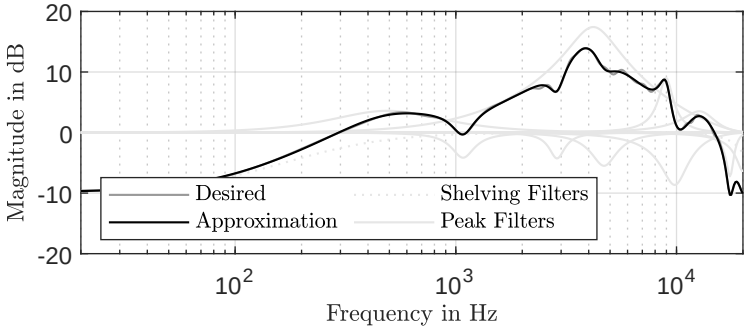

***(a)** Ipsilateral directions*



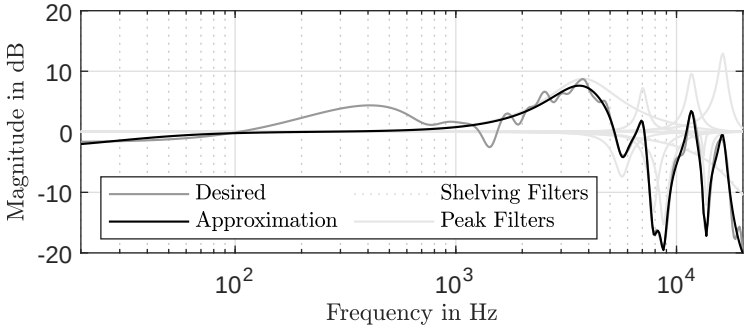

***(b)** Contralateral directions*

**Figure 3.21:** Heatmaps of average LSD using ten peak filters per direction with post-optimization for 90 subjects separated into (a) ipsilateral and (b) contralateral directions. The overall mean is 0.56 dB.

The average LSD values across directions achieved by the post-optimized approximated magnitude responses $|\hat{H}(k)|_{\mathrm{dB,opt}}$ are shown in Fig. 3.21. Here, average LSD values between 0.36 dB for rear directions ($\varphi_{\mathrm{rel}} = 0°$, $\theta = 180°$) and 0.83 dB for slightly contralateral directions in the front ($\varphi_{\mathrm{rel}} = -30°$, $\theta = 0°$) can be seen. By comparing Figs. 3.21 and 3.18, the improvements due to post-optimization are clearly visible. However, the lack of enough peak filters for contralateral directions below ($\theta = -45°$, $\theta = 225°$) and in front ($\theta = 0°$) of the subject is still visible in the heatmap in Fig. 3.21(b).




***(a)** Lowest LSD of 0.14 dB*




***(b)** Highest LSD of 1.73 dB*

**Figure 3.22:** Post-optimized approximations with (a) the lowest LSD of 0.14 dB (*Subject_009*, right ear, $\varphi_{\text{rel}} = 45°$, $\theta = 135°$) and (b) the highest LSD of 1.73 dB (*Subject_012*, left ear, $\varphi_{\text{rel}} = -55°$, $\theta = -45°$) between desired magnitude response $|H_{\text{d}}(k)|_{\text{dB}}$ and post-optimization $|\hat{H}(k)|_{\text{dB,opt}}$. The magnitude responses of the individual filter stages are shown in the background.

Similar to Fig. 3.19, Fig. 3.22 shows the post-optimized approximations with the lowest and highest LSD. In Fig. 3.22(a), only small differences that do not exceed 0.5 dB can be seen between post-optimized magnitude response $|\hat{H}(k)|_{\text{dB,opt}}$ and desired magnitude response $|H_{\text{d}}(k)|_{\text{dB}}$. The approximation error of the worst case in Fig. 3.22(b) shows similarities to the one in Fig. 3.19(b), but this time also the post-optimization does not reproduce the amplification between 100 Hz and 1 kHz. Additionally, the small ripples in the desired magnitude response between 1 kHz and 4 kHz are not approximated due to the restriction to ten peak filters.

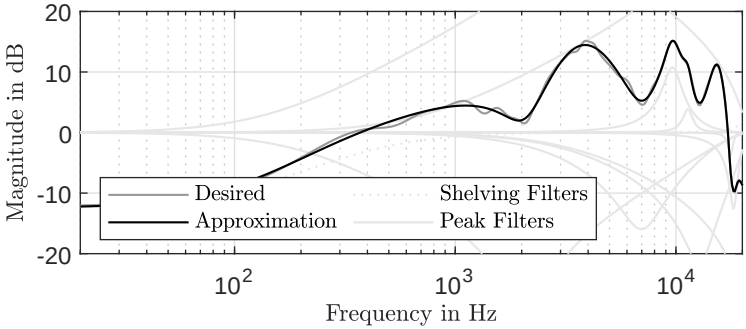



**Figure 3.23:** Post-optimized approximation for the left ear of *Subject_012* for a direction of $\varphi_{\text{rel}} = 55^\circ$ and $\theta = -45^\circ$. The post-optimization yields an LSD of 0.40 dB, but the magnitude responses of the individual filter stages show high positive and negative gains.

Figure 3.23 shows a further post-optimization result, where the left ear's magnitude response of *Subject_012* for a direction of $\varphi_{\text{rel}} = 55^\circ$ and $\theta = -45^\circ$ is approximated. Although the desired magnitude response $|H_{\text{d}}(k)|_{\text{dB}}$ is accurately approximated by the post-optimized magnitude response $|\hat{H}(k)|_{\text{dB,opt}}$, high gains in the individual peak filters that should be avoided are visible. The individual peak filters reach positive gains up to 38 dB and negative gains down to $-29$ dB that interact destructively.

## 3.4 Approximation via Backpropagation Algorithm

As described in [Bhattacharya et al., 2020], also the backpropagation algorithm can be used to update the parameters of a parametric IIR filter cascade in order to approximate a given HRTF magnitude. This section summarizes the results achieved in [Bhattacharya et al., 2020] and deepens the analysis of these results. Additionally, the backpropagation algorithm is used as post-optimization method for the approximations given in Section 3.3.2. In order to understand the underlying principle, firstly the backpropagation algorithm used in [Bhattacharya et al., 2020] is explained and the derived local gradients with respect to the control parameters are given.

The principle of the underlying prediction method is given in Fig. 3.24. Here, the desired HRTF $H_{\text{d}}(z)$ is used as the system under test and the parametric IIR filter cascade from Fig. 3.5 defines the predicted transfer function $\hat{H}(z)$. Additionally, a unit impulse $\delta(n)$ is used as input signal $x(n)$, such that the output signals $y_{\text{d}}(n)$ and $y(n)$ meet the desired and

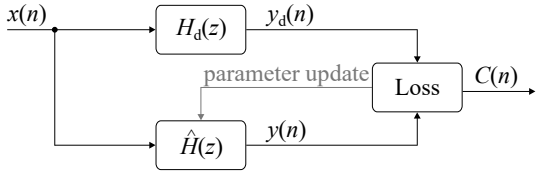



**Figure 3.24:** Block diagram of the model for HRTF magnitude approximation with desired filter output $y_\mathrm{d}(n)$, predicted output $y(n)$, and cost function $C(n)$.

predicted impulse response, respectively. Since the parametric IIR filter cascade represents a minimum-phase system without compensation of the FIR filter's delay, the loss function is defined in the frequency-domain as

$$C(Y_\mathrm{d}(k), Y(k)) = \log_{10}|Y(k)|^2 - \log_{10}|Y_\mathrm{d}(k)|^2 \tag{3.40}$$

in order to match the goal of magnitude response approximation. Here, $|Y_\mathrm{d}(k)|$ and $|Y(k)|$ define the frequency-domain magnitudes of the output signals $y_\mathrm{d}(n)$ and $y(n)$, respectively. In case of $x(n) = \delta(n)$, $|Y_\mathrm{d}(k)|$ and $|Y(k)|$ equal the magnitude responses of the underlying filters ($|H_\mathrm{d}(k)|$ and $|H(k)|$).

When using the gradient descent algorithm for cost function minimization by parameter optimization, the gradient $\frac{\partial C(n)}{\partial p_m}$ is used to update the corresponding parameter $p_m$ according to

$$p_m \leftarrow p_m - \eta \frac{\partial C(n)}{\partial p_m}, \tag{3.41}$$

where $\eta$ defines the step-size or learning rate. The calculation of partial derivatives is not straightforward and has to be done as a product of multiple local derivatives according to the chain rule, resulting in the backpropagation algorithm. The flow of the instantaneous local gradients for the $\{M-1\}^\mathrm{th}$ filter is shown in Fig. 3.25. Here, the local gradient of the cost function with respect to the parameter of the $\{M-1\}^\mathrm{th}$ filter, $\frac{\partial C(n)}{\partial p_{\mathrm{M\text{-}1}}}$, can be factorized into the derivation of the cost function with respect to the output of the whole cascade $\left(\frac{\partial C(n)}{\partial y(n)}\right)$, the derivation of the output of the whole cascade with respect to the input of the last filter $\left(\frac{\partial y(n)}{\partial x_\mathrm{M}(n)}\right)$, and the derivation of the output of the $\{M-1\}^\mathrm{th}$ filter with respect to the parameter of the $\{M-1\}^\mathrm{th}$ filter $\left(\frac{\partial y_{\mathrm{M\text{-}1}}(n)}{\partial p_{\mathrm{M\text{-}1}}}\right)$. Similarly, the gradient of the cost function with respect to the parameter of the $m^\mathrm{th}$ filter in the

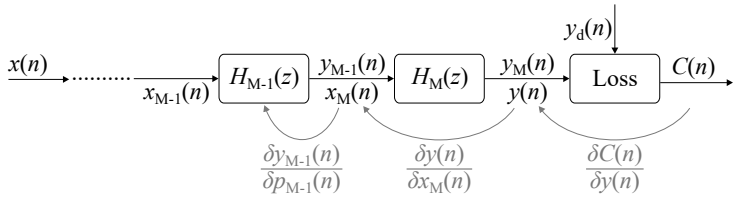



**Figure 3.25:** Flow of the instantaneous local gradients during backpropagation for the $\{M-1\}^{\text{th}}$ filter, with $C(n)$ being the cost function and $y_d(n)$ the desired output. Additionally, $x_m(n)$, $y_m(n)$, and $p_m$ define input, output, and parameter of the $m^{\text{th}}$ filter, respectively.

cascade can be calculated as

$$\frac{\partial C(n)}{\partial p_m} = \begin{cases} \dfrac{\partial C(n)}{\partial y(n)} \cdot \dfrac{\partial y_m(n)}{\partial p_m} \cdot \displaystyle\prod_{q=m+1}^{M} \dfrac{\partial y_q(n)}{\partial x_q(n)} & \text{for} \quad 1 \leq m < M \\ \dfrac{\partial C(n)}{\partial y(n)} \cdot \dfrac{\partial y(n)}{\partial p_M} & \text{for} \quad m = M \end{cases} . \tag{3.42}$$

By using the frequency-domain loss function given in Eq. (3.40), the derivative $\frac{\partial C(n)}{\partial y(n)}$ includes the inverse Fourier transform, which can be solved by using the Wirtinger calculus [Caracalla and Roebel, 2017, Wang et al., 2020]. According to Fig. 3.5, the first filter in the cascade represents a first-order LFS and the $M^{th}$ filter represents a first-order HFS. The other $M-2$ filters are defined as second-order peak filters. In the following two subsections, the local gradients with respect to different filter parameters are listed for shelving and peak filters.

### 3.4.1 Shelving Filters

The formulas that define shelving filters can be found in Eqs. (3.1) to (3.12). Based on the difference equation in Eq. (3.12), the partial derivative of a shelving filter's output $y_m(n)$ with respect to the input $x_m(n)$ is calculated as

$$\frac{\partial y_m(n)}{\partial x_m(n)} = \frac{H_{0,m}}{2} \left[1 \pm a_m\right] + 1, \tag{3.43}$$

where the plus sign is used for LFS and the minus for HFS. The parameter $a_m$ depends on the used filter type (see Eqs. (3.7) - (3.9)). In order to calculate the partial derivatives with respect to the parameters of the shelving filters, the calculation has to be split into boost and cut case due to the different parameters $a_m$ that are defined in Eqs. (3.7) to (3.9). For

the boost case ($G_m \geq 0\,\mathrm{dB}$), the partial derivative of a shelving filter's output $y_m(n)$ with respect to the gain $G_m$ is calculated as

$$\frac{\partial y_m(n)}{\partial G_m} = \frac{\left[x_m(n) \pm y_{\mathrm{ap1},m}(n)\right]}{40} 10^{\frac{G_m}{20}} \ln(10). \tag{3.44}$$

Since the cut case parameter $a_{\mathrm{C},m}$ depends on the gain $G_m$, the derivative of the output of the first-order all-pass $y_{\mathrm{ap1},m}$ with respect to the gain $G_m$

$$\frac{\partial y_{\mathrm{ap1},m}(n)}{\partial G_m} = \frac{\partial a_{\mathrm{C},m}}{\partial G_m} x_{\mathrm{h},m}(n) + a_{\mathrm{C},m} \frac{\partial x_{\mathrm{h},m}(n)}{\partial G_m} + \frac{\partial x_{\mathrm{h},m}(n-1)}{\partial G_m} \tag{3.45}$$

has to be included in the derivation, where

$$\frac{\partial a_{\mathrm{C},m}}{\partial G_m} = \begin{cases} \dfrac{-\ln(10) V_{0,m} \tan\left(\pi \frac{f_{c,m}}{f_s}\right)}{10\left[\tan\left(\pi \frac{f_{c,m}}{f_s}\right) + V_{0,m}\right]^2} & \text{for LFS} \\ \dfrac{\ln(10) V_{0,m} \tan\left(\pi \frac{f_{c,m}}{f_s}\right)}{10\left[V_{0,m} \tan\left(\pi \frac{f_{c,m}}{f_s}\right) + 1\right]^2} & \text{for HFS} \end{cases}, \tag{3.46}$$

$$\frac{\partial x_{\mathrm{h},m}(n)}{\partial G_m} = -\frac{\partial a_{\mathrm{C},m}}{\partial G_m} x_{\mathrm{h},m}(n-1) - a_{\mathrm{C},m} \frac{\partial x_{\mathrm{h},m}(n-1)}{\partial G_m}, \tag{3.47}$$

and $\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial G_m}$ can be calculated via instantaneous backpropagation through time (IBPTT) with the initialization $\frac{\partial x_{\mathrm{h},m}(u)}{\partial G_m}|_{u=0} = 0$ [Back and Tsoi, 1991]. So the partial derivative of a shelving filter's output $y_m(n)$ with respect to the gain $G_m$ for the cut case ($G_m < 0\,\mathrm{dB}$) is calculated as

$$\frac{\partial y_m(n)}{\partial G_m} = \frac{\left[x_m(n) \pm y_{\mathrm{ap1},m}(n)\right]}{40} 10^{\frac{G_m}{20}} \ln(10) \pm \frac{H_{0,m}}{2} \frac{\partial y_{\mathrm{ap1},m}(n)}{\partial G_m}. \tag{3.48}$$

The partial derivative of a shelving filter's output $y_m(n)$ with respect to the cut-off frequency $f_{c,m}$ is calculated as

$$\frac{\partial y_m(n)}{\partial f_{c,m}} = \pm \frac{H_{0,m}}{2} \frac{\partial y_{\mathrm{ap1},m}(n)}{\partial f_{c,m}}, \tag{3.49}$$

where the derivative of the output of the all-pass filter $y_{\mathrm{ap1},m}(n)$ with respect to the cut-off frequency $f_{c,m}$ is

$$\frac{\partial y_{\mathrm{ap1},m}(n)}{\partial f_{c,m}} = \frac{\partial a_m}{\partial f_{c,m}} x_{\mathrm{h},m}(n) + a_m \frac{\partial x_{\mathrm{h},m}(n)}{\partial f_{c,m}} + \frac{\partial x_{\mathrm{h},m}(n-1)}{\partial f_{c,m}}. \tag{3.50}$$

The derivative of the boost case parameter $a_{\mathrm{B},m}$ with respect to the cut-off frequency $f_{c,m}$ is

$$\frac{\partial a_{\mathrm{B},m}}{\partial f_{c,m}} = \frac{2\pi \sec\left(2\pi \frac{f_{c,m}}{f_s}\right)\left[\sec\left(2\pi \frac{f_{c,m}}{f_s}\right) - \tan\left(2\pi \frac{f_{c,m}}{f_s}\right)\right]}{f_s} \tag{3.51}$$

and the derivative of the cut case parameter $a_{\mathrm{C},m}$ with respect to the cut-off frequency $f_{c,m}$ is

$$\frac{\partial a_{\mathrm{C},m}}{\partial f_{c,m}} = \begin{cases} \dfrac{2\pi V_{0,m} \sec\left(\pi \frac{f_{c,m}}{f_s}\right)^2}{f_s\left[\tan\left(\pi \frac{f_{c,m}}{f_s}\right) + V_{0,m}\right]^2} & \text{for LFS} \\ \dfrac{2\pi V_{0,m} \sec\left(\pi \frac{f_{c,m}}{f_s}\right)^2}{f_s\left[V_{0,m}\tan\left(\pi \frac{f_{c,m}}{f_s}\right) + 1\right]^2} & \text{for HFS} \end{cases}. \tag{3.52}$$

The derivative of $x_{\mathrm{h},m}(n)$ with respect to the cut-off frequency $f_{c,m}$ is

$$\frac{\partial x_{\mathrm{h},m}(n)}{\partial f_{c,m}} = -\frac{\partial a_m}{\partial f_{c,m}} x_{\mathrm{h},m}(n-1) - a_m \frac{\partial x_{\mathrm{h},m}(n-1)}{\partial f_{c,m}}. \tag{3.53}$$

### 3.4.2 Peak Filters

Based on Eqs. (3.13) to (3.21), the local derivatives needed to use backpropagation for peak filters can be calculated. At first, the partial derivative of a peak filter's output $y_m(n)$ with respect to it's input $x_m(n)$ is calculated as

$$\frac{\partial y_m(n)}{\partial x_m(n)} = \frac{H_{0,m}}{2}\left[1 + a_m\right] + 1. \tag{3.54}$$

Then, the partial derivatives of a peak filter's output $y_m(n)$ with respect to it's parameters have to be calculated. Firstly, the partial derivative of a peak filter's output $y_m(n)$ with respect to it's gain $G_m$ is calculated as

$$\frac{\partial y_m(n)}{\partial G_m} = \begin{cases} \dfrac{\left[x_m(n) - y_{\mathrm{ap2},m}(n)\right]}{40} 10^{\frac{G_m}{20}} \ln(10) & \text{for } G_m \geq 0\,\mathrm{dB} \\ \dfrac{\left[x_m(n) - y_{\mathrm{ap2},m}(n)\right]}{40} 10^{\frac{G_m}{20}} \ln(10) - \dfrac{H_{0,m}}{2} \dfrac{\partial y_{\mathrm{ap2},m}(n)}{\partial G_m} & \text{for } G_m < 0\,\mathrm{dB} \end{cases}, \tag{3.55}$$

where the partial derivative of the output of the second-order all-pass filter $y_{\mathrm{ap2},m}(n)$ with respect to the gain $G_m$ for the cut case ($G_m < 0\,\mathrm{dB}$) is given as

$$\begin{aligned}\frac{\partial y_{\mathrm{ap2},m}}{\partial G_m} =& -\frac{\partial a_{\mathrm{C},m}}{\partial G_m} x_{\mathrm{h},m}(n) - a_{\mathrm{C},m}\frac{\partial x_{\mathrm{h},m}(n)}{\partial G_m} - d_m\frac{\partial a_{\mathrm{C},m}}{\partial G_m} x_{\mathrm{h},m}(n-1) \\ &+ d_m(1-a_{\mathrm{C},m})\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial G_m} + \frac{\partial x_{\mathrm{h},m}(n-2)}{\partial G_m}, \end{aligned} \tag{3.56}$$

with

$$\frac{\partial a_{\mathrm{C},m}}{\partial G_m} = \frac{-\ln(10)V_{0,m}\tan\left(\pi\frac{f_{b,m}}{f_s}\right)}{10\left[\tan\left(\pi\frac{f_{b,m}}{f_s}\right)+V_{0,m}\right]^2}, \tag{3.57}$$

$$\begin{aligned}\frac{\partial x_{\mathrm{h},m}}{\partial G_m} =& d\frac{\partial a_{\mathrm{C},m}}{\partial G_m} x_{\mathrm{h},m}(n-1) + \frac{\partial a_{\mathrm{C},m}}{\partial G_m} x_{\mathrm{h},m}(n-2) \\ &- d_m(1-a_{\mathrm{C},m})\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial G_m} + a_{\mathrm{C},m}\frac{\partial x_{\mathrm{h},m}(n-2)}{\partial G_m}. \end{aligned} \tag{3.58}$$

Here, the derivatives $\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial G_m}$ and $\frac{\partial x_{\mathrm{h},m}(n-2)}{\partial G_m}$ can be calculated using the initialization $\frac{\partial x_{\mathrm{h},m}(u)}{\partial G_m}|_{u=0} = 0$ and $\frac{\partial x_{\mathrm{h},m}(u)}{\partial G_m}|_{u=-1} = 0$, and IBPTT [Back and Tsoi, 1991]. Secondly, the partial derivative of a peak filter's output $y_m(n)$ with respect to it's center frequency $f_{c,m}$ is calculated as

$$\frac{\partial y_m(n)}{\partial f_{c,m}} = -\frac{H_{0,m}}{2}\frac{\partial y_{\mathrm{ap2},m}(n)}{\partial f_{c,m}}, \tag{3.59}$$

with

$$\begin{aligned}\frac{\partial y_{\mathrm{ap2},m}(n)}{\partial f_{c,m}} =& -a_m\frac{\partial x_{\mathrm{h},m}(n)}{\partial f_{c,m}} + \frac{\partial d_m}{\partial f_{c,m}}(1-a_m)x_{\mathrm{h},m}(n-1) \\ &+ d_m(1-a_m)\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial f_{c,m}} + \frac{\partial x_{\mathrm{h},m}(n-2)}{\partial f_{c,m}}, \end{aligned} \tag{3.60}$$

$$\begin{aligned}\frac{\partial x_{\mathrm{h},m}(n)}{\partial f_{c,m}} =& -\frac{\partial d_m}{\partial f_{c,m}}(1-a_m)x_{\mathrm{h},m}(n-1) - d_m(1-a_m)\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial f_{c,m}} \\ &+ a_m\frac{\partial x_{\mathrm{h},m}(n-2)}{\partial f_{c,m}}. \end{aligned} \tag{3.61}$$

Thirdly, the partial derivative of a peak filter's output $y_m(n)$ with respect to the bandwidth $f_{b,m}$ is calculated as

$$\frac{\partial y_m(n)}{\partial f_{b,m}} = -\frac{H_{0,m}}{2}\frac{\partial y_{\mathrm{ap2},m}(n)}{\partial f_{b,m}}, \tag{3.62}$$

where the partial derivative of the output of the second-order all-pass filter $y_{\mathrm{ap2},m}(n)$ with respect to the bandwidth $f_{b,m}$ is given as

$$\begin{aligned}\frac{\partial y_{\mathrm{ap2},m}(n)}{\partial f_{b,m}} =& - a_m \frac{\partial x_{\mathrm{h},m}(n)}{\partial f_{b,m}} - \frac{\partial a_m}{\partial f_{b,m}} x_{\mathrm{h},m}(n) - d_m \frac{\partial a_m}{\partial f_{b,m}} x_{\mathrm{h},m}(n-1) \\ &+ \frac{\partial x_{\mathrm{h},m}(n-2)}{\partial f_{b,m}} + d_m(1-a_m)\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial f_{b,m}},\end{aligned} \tag{3.63}$$

with

$$\begin{aligned}\frac{\partial x_{\mathrm{h},m}(n)}{\partial f_{b,m}} =& \, d_m \frac{\partial a_m}{\partial f_{b,m}} x_{\mathrm{h},m}(n-1) + \frac{\partial a_m}{\partial f_{b,m}} x_{\mathrm{h},m}(n-2) \\ &- d_m(1-a_m)\frac{\partial x_{\mathrm{h},m}(n-1)}{\partial f_{b,m}} + a_m \frac{\partial x_{\mathrm{h},m}(n-2)}{\partial f_{b,m}}.\end{aligned} \tag{3.64}$$

Additionally, the calculation of the derivative of the all-pass filter coefficient $a_m$ with respect to the bandwidth $f_{b,m}$, $\frac{\partial a_m}{\partial f_{b,m}}$, has to be separated for boost and cut case as

$$\frac{\partial a_m}{\partial f_{b,m}} = \begin{cases} \dfrac{2\pi \sec\left(\pi \frac{f_{b,m}}{f_s}\right)\left[\sec\left(\pi \frac{f_{b,m}}{f_s}\right) - \tan\left(\pi \frac{f_{b,m}}{f_s}\right)\right]}{f_s} & \text{for } G \geq 0\,\mathrm{dB} \\ \dfrac{2\pi V_{0,m} \sec\left(\pi \frac{f_{b,m}}{f_s}\right)^2}{f_s\left[\tan\left(\pi \frac{f_{b,m}}{f_s}\right) + V_{0,m}\right]^2} & \text{for } G < 0\,\mathrm{dB} \end{cases}. \tag{3.65}$$

### 3.4.3 Approximation of HRTF Magnitudes

As a random initialization of shelving and peak filter parameters will yield a random magnitude response $|\hat{H}(k)|_{\mathrm{dB}}$ for the parametric IIR filter cascade having a poor correlation with the desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$, a structured initialization is used in [Bhattacharya et al., 2020]. The initialization procedure for the peak filter parameters is given in Fig. 3.26. The calculation of the desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ is similar to the pre-processing done in Fig. 3.7, but a uniform spacing is used between the frequency bins $k$ instead of the exponentially spacing used in Section 3.3. When the desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$ is derived, the positions of peaks and notches inside the desired magnitude response are calculated with the help of MATLAB's function `findpeaks`. In addition to the positions of the peaks, the proximity, the prominence, and the magnitude difference to neighboring peaks are calculated. The peak

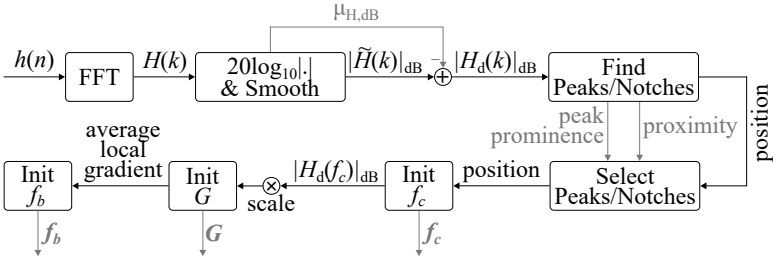



**Figure 3.26:** Initialization procedure for the peak filter parameter vectors, namely center frequency $\mathbf{f_c}$, gain $\mathbf{G}$, and bandwidth $\mathbf{f_b}$. The initialization is based on finding the peaks and notches inside the desired magnitude response in decibels $|H_\mathrm{d}(k)|_\mathrm{dB}$.

prominence is a measure in decibels of how much a peak stands out relative to other peaks, thus an isolated peak with a low magnitude can have a higher prominence than a peak with a higher magnitude that is located in a plateau of high magnitudes. Afterwards, peaks and notches falling below the threshold of one of the three additional metrics are deleted. In the following analysis, these thresholds are given as 0.005 dB for the minimum peak prominence, 300 Hz for the proximity, and 3 dB for the magnitude difference to the neighboring peaks. The positions of the notches are found by using the `findpeaks`-function for the inverted magnitude response. The positions of the remaining peaks and notches are used as initialized center frequencies $f_c$ for the peak filters stored in the vector of center frequencies $\mathbf{f_c}$. Afterwards, the desired magnitude response $|H_\mathrm{d}(f_c)|_\mathrm{dB}$ is evaluated at the center frequencies. Then, a scaled version of these magnitudes is used as initialization of the peak filter gains $\mathbf{G}$. The scaling factor was integrated to the initialization of the gains in order to reduce the summation effect of neighboring peak filters in the initialized magnitude response. Moreover, notches that are located in the positive half-plane will have positive magnitudes, such that the gains for filters at those positions are converted into small negative random values in order to maintain the information of being a notch. Similarly, the gains of peaks located in the negative half-plane are initialized with small positive random values. Finally, the average local gradients of the magnitude response around the positions of the peaks are used to initialize the bandwidths $\mathbf{f_b}$ of the peak filters. In contrast to the elaborated initialization of the peak filters, the shelving filter initialization follows a simple procedure. Here, the gains are initialized by the magnitudes of the desired response at $f_\mathrm{L} = 0\,\mathrm{Hz}$ and $f_\mathrm{H} = 20\,\mathrm{kHz}$, and the cut-off frequencies are initialized depending on the position of the first and last peak filter, respectively.

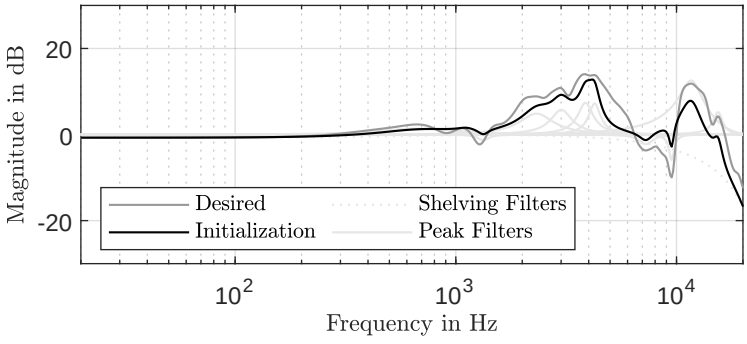
In Fig. 3.27(a), an exemplary initialized magnitude response is shown for the left ear of *Subject_003* and a frontal sound source ($\varphi_{\mathrm{rel}} = 0°$, $\theta = 0°$). In the given example, 13 peaks and notches are found during the initialization procedure. As can be seen, the peak filter positions in the initialized magnitude response match the positions of the peaks and notches in the desired magnitude response. However, due to the scaling factor in the initialization procedure of the peak filters, the gains of these filters are lower than the magnitudes of the desired response at the positions of the center frequencies. Additionally, it can be seen that the gain of the HFS matches the magnitude of the desired magnitude response at $f = 20\,\mathrm{kHz}$. Due to the chosen example, the initialized LFS has only a small negative gain that is barely visible in the given plot.

After initializing the parametric IIR filter cascade, the different parameters are updated based on the backpropagation algorithm using the adaptive moment estimation method [Kingma and Ba, 2015] and the derivatives calculated in Sections 3.4.1 and 3.4.2. Here, 100 epochs consisting of 1024 iterations each and a learning rate of $\eta = 10^{-1}$ are used. Additionally, the learning rate is equipped with a drop factor of 0.99 that is activated when the error of an epoch has increased in comparison to the previous one.

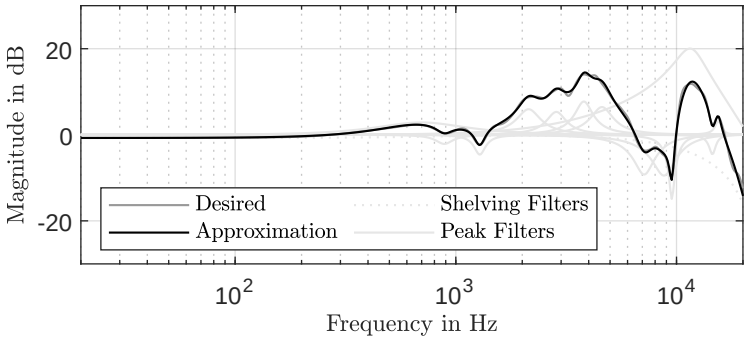
In Fig. 3.27(b), the approximation result is shown for the initialized and desired magnitude responses of Fig. 3.27(a). As can be seen, the backpropagation algorithm has improved the approximation and led to a magnitude response $|\hat{H}(k)|_{\mathrm{dB}}$ close to the desired magnitude response $|H_{\mathrm{d}}(k)|_{\mathrm{dB}}$. Furthermore, the LSD calculated as given in Eq. (3.39) with an exponentially frequency spacing between 20 Hz and 20 kHz (see Eq. (3.22)) has reduced from 2.44 dB for the initialized magnitude response to 0.37 dB for the approximation.

In order to evaluate the given approximation procedure based on the backpropagation algorithm, a subset of HRIRs from the CIPIC database [Algazi et al., 2001b] consisting of the first ten subjects, seven azimuths, and seven elevations is used. Here, the elevation is evaluated between $\theta = -45°$ and $\theta = 225°$ with an angular resolution of $\Delta\theta = 45°$, and the azimuth is evaluated at $\varphi = \{-80°, -55°, -20°, 0°, 20°, 55°, 80°\}$. Left and right ear information are combined by using the relative azimuth $\varphi_{\mathrm{ref}}$ defined in Section 3.3, which is positive for ipsilateral directions and negative for contralateral directions. In this way, an augmented data set of 20 subjects and 49 directions is created, resulting in a total number of 980 approximated magnitude responses.

As described above, the number of peak filters used for approximating the desired magnitude response is determined by the initialization procedure, so that for every approximation a different number of peak filters is used. In average, 17.8 peak filters were used for the approximation. By taking a detailed look into the individual peak filters that are used during the approximation, some of the peak filters have shown Q-factors considerably




***(a)** Initialization*




***(b)** Approximation*

**Figure 3.27:** Exemplary (a) initialization and (b) approximation through backpropagation algorithm of the left ear's magnitude response of *Subject_003* for $\varphi_{\text{rel}} = 0^\circ$ and $\theta = 0^\circ$. The magnitude responses of the individual filter stages are shown in the background. Here, 13 peak filters are used.

higher than 100, which indicate very narrow peaks that have no influence in approximating the smoothed desired magnitude response. In order to reduce the number of used filters, peak filters with a Q-factor higher than 100 are deleted in a post-processing step. Deleting the 525 peak filters in question resulted in an average number of used peak filters of 17.2. In Fig. 3.28, the average number of used peak filters per direction is illustrated. When comparing Fig. 3.28 with Fig. 3.14, similarities can be found in the color distribution of the heatmap. Firstly, ipsilateral directions ($\varphi_{\text{rel}} > 20^\circ$) tend to need less peak filters than the other ones. Especially contralateral directions at frontal elevations show high average numbers of used peak

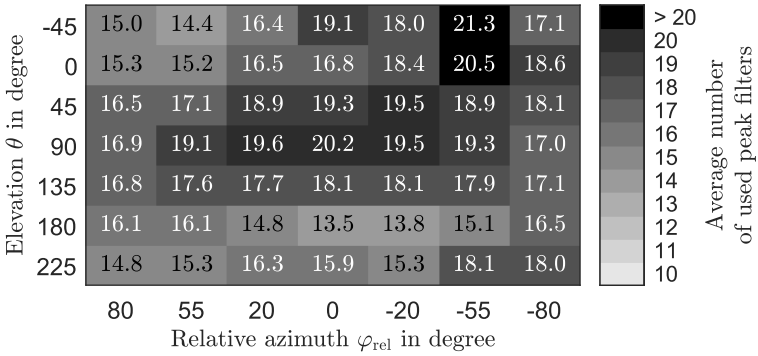



**Figure 3.28:** Heatmap of average number of used peak filters per direction across 20 subjects for 49 different directions.

filters. Here, the highest average number of used peak filters is 21.3 at a relative azimuth of $\varphi_{\text{rel}} = -55°$ and an elevation of $\theta = -45°$. In contrast to this, the rear direction ($\varphi_{\text{rel}} = 0°$, $\theta = 180°$) needs on average only 13.5 peak filters. A difference in the color distributions of the two figures is given by the directions directly above the listener ($|\varphi_{\text{rel}}| \leq 20°$, $\theta = 90°$), where Fig. 3.28 shows a much darker color meaning a comparatively higher average number of used peak filters. Additionally, the average numbers given in Fig. 3.28 are much higher than the ones shown in Fig. 3.14. The reason for this is that in case of the results shown in Fig. 3.14 the peak filters are added consecutively until a given target is fulfilled, whereas here, the initialization procedure determines a number of peak filters without testing whether a lower number of peak filters can achieve acceptable results, too.

By comparing the results for the different desired magnitude responses, big differences are found in the accuracy of the approximation results. In 8.6 % of the approximations, even instabilities can be seen when using a very long unit impulse with a duration of 2 s for calculating the impulse response of the IIR filter cascade. The reason for these instabilities are mainly negative Q-factors of the peak filters. In a few cases also negative center frequencies of the peak filters or cut-off frequencies of the HFS slightly higher than $f_s/2$ occur, which also result in unstable IIR filters. Thus, the occurrence of these instabilities can be fixed by either post-processing of filter parameters or constraining the filter parameters during the approximation procedure. However, constraints in the filter parameters, e.g. giving an upper limit of $f_s/2$ for the cut-off frequency of the HFS, have to be considered in the backpropagation algorithm, too. Thus, derivatives for the used constraints have to be calculated. For the results that are shown here, the instability issues are solved by post-processing. Negative Q-factors

and negative center frequencies of the peak filters are solved by taking the absolute value of the corresponding parameters. Additionally, cut-off frequencies of the HFS are limited to $f_s/2$. In this way, all instabilities were stabilized without a considerable influence on the approximated magnitude responses while using a shorter unit impulse length.

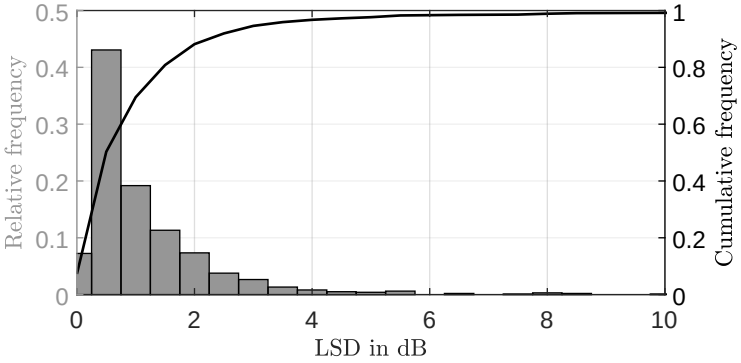
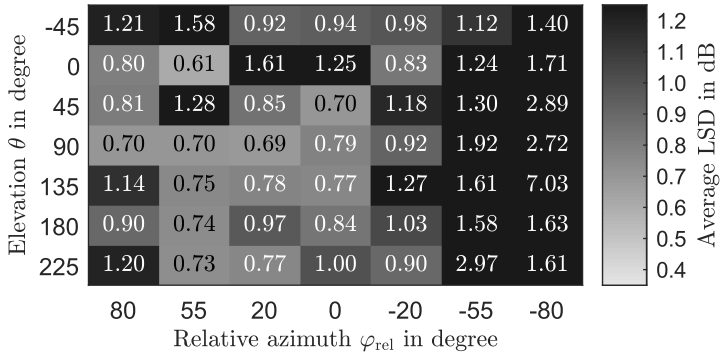


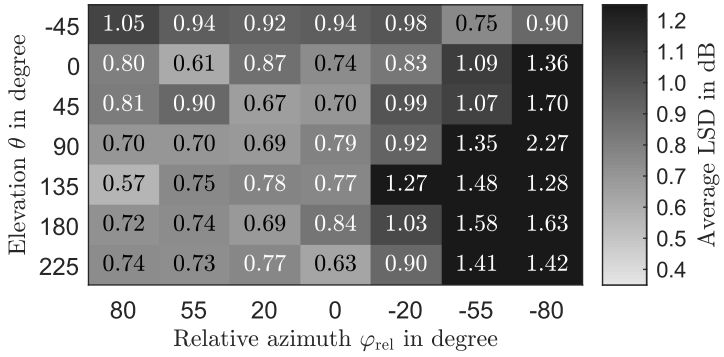

**Figure 3.29:** Histogram of relative frequencies of the achieved LSD over all of the 980 approximations. Here, the LSD values are rounded to an accuracy of 0.5 dB. Additionally, the cumulative frequency is plotted.

The histogram in Fig. 3.29 gives the relative frequency of achieved LSD values across all 980 approximations rounded to an accuracy of 0.5 dB. Most of the approximations (43.1 %) achieve LSD values between 0.25 dB and 0.75 dB. Additionally, 7.2 % of the approximations show even lower LSD values than 0.25 dB. However, 0.8 % of the approximations totally fail to approximate the desired magnitude response and are even outside of the given visualized LSD range of up to 10 dB. Nevertheless, these high LSD values result from very small magnitude responses tending to $-\infty$ dB rather than having instabilities in the approximated magnitude response. In addition to the relative frequency of rounded LSD values, also the cumulative frequency of these values are given in Fig. 3.29. The cumulative frequency indicates that 96.7 % of the approximations have LSD values lower than 4.25 dB and 88.2 % have LSD values smaller than 2.25 dB, showing the effectiveness of the given approximation algorithm for most of the cases. The average LSD across all approximations is 1.30 dB. When considering only the approximations with an LSD lower than 4 dB as successful approximation, the average LSD of these directions is 0.97 dB.

In Fig. 3.30, the average LSD is evaluated per direction. Here, Fig. 3.30(a) contains the average LSD across all approximations, whereas Fig. 3.30(b) contains only the average LSD values for successful approximations with




***(a)** All approximations*

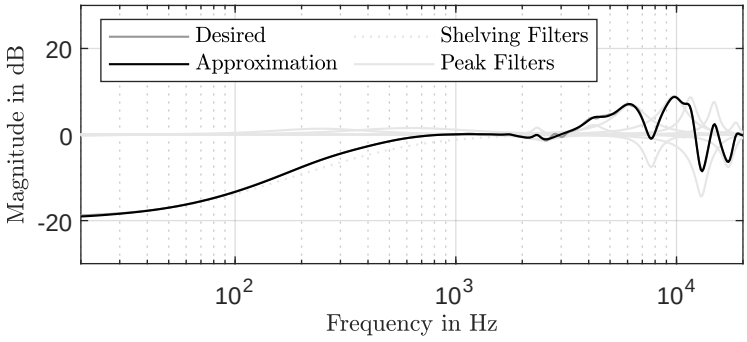



***(b)** Approximations with LSD ≤ 4 dB*

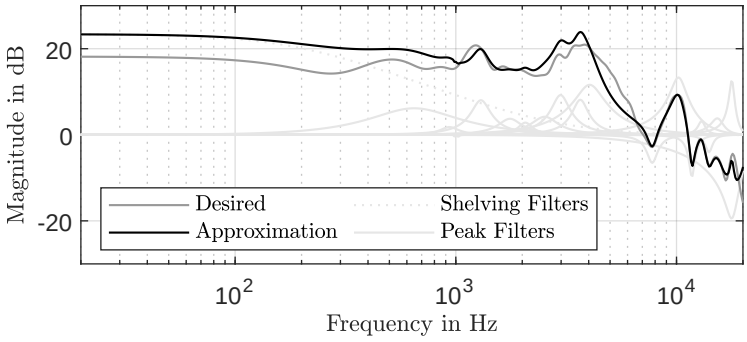
**Figure 3.30:** Heatmaps of average LSD per direction across 20 subjects for (a) all approximations and (b) approximations with an LSD lower than 4 dB. The overall means are 1.30 dB and 0.97 dB, respectively.

an LSD lower than 4 dB. By comparing the two subplots, it can be seen that the average LSD alters for almost half of the directions, meaning that the 32 failed approximations with an LSD higher than 4 dB are spread across 24 different directions. Although using 15.1 to 19.3 peak filters on average, the contralateral rear directions show the highest average LSD values that are 1.27 dB or higher (see Fig. 3.30(b)). In contrast to this, the average LSD for ipsilateral directions exceeds 1 dB only for a single direction ($\varphi_{\mathrm{rel}} = 80°$, $\theta = -45°$, LSD = 1.05 dB).

In addition to the average LSD values per direction, Fig. 3.31 shows two exemplary approximation results. In Fig. 3.31(a), the best approximation result is shown, which achieves the lowest LSD of 0.13 dB between approx-




***(a)** Lowest LSD of 0.13 dB*




***(b)** LSD of 3.95 dB*

**Figure 3.31:** Approximations with (a) the lowest LSD of 0.13 dB using 17 peak filters (*Subject_011*, right ear, $\varphi_{\text{rel}} = 55°$, $\theta = 225°$) and (b) an LSD of 3.95 dB using 19 peak filters (*Subject_018*, left ear, $\varphi_{\text{rel}} = -20°$, $\theta = 225°$) between desired magnitude response $|H_{\text{d}}(k)|_{\text{dB}}$ and approximation through backpropagation $|\hat{H}(k)|_{\text{dB}}$. The magnitude responses of the individual filter stages are shown in the background.

imated and desired magnitude response. As can be seen, only a single small peak and small notch around 3 kHz are not approximated. This small deviation will not be audible. Contrarily, Fig. 3.31(b) shows an approximation with an LSD of 3.95 dB. When comparing approximated and desired magnitude response, the main deviations are seen in low ($f < 1$ kHz) and mid-frequency region (2 kHz $< f <$ 6.5 kHz). One similarity between all these frequencies is that they lie in the passband and the transition band of the LFS, such that the wrong gain of the LFS is the main reason for

the deviation. Especially for the low frequencies ($f < 1\,\text{kHz}$), it is clearly visible that reducing the gain of the LFS by approximately 5 dB would easily reduce the approximation error. Additionally, for the frequency region between 4 kHz and 6.5 kHz, the current approximation is not able to follow the decrease of the magnitude response correctly. The problems inside both of these frequency regions are common for approximations that show LSD values of around 4 dB.

### 3.4.4 Post-Optimization of Approximated HRTF Magnitudes

Since the initialization procedure given in Section 3.4.3 delivers more than an optimal number of filters, this section uses the approximation results achieved in Section 3.3.2 as initialization for the backpropagation algorithm. In this way, the number of peak filters is fixed to ten. Note that the smaller subset of HRIR directions and subjects defined in Section 3.4.3 is used here instead of the one used in Section 3.3.2. Thus, the approximations have to be caught up on azimuth angles $\varphi = 20°$ and $\varphi = -20°$ for the first ten subjects of the database before using them as initialization.

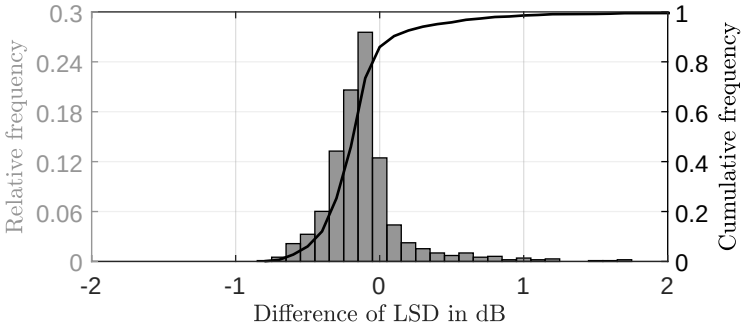


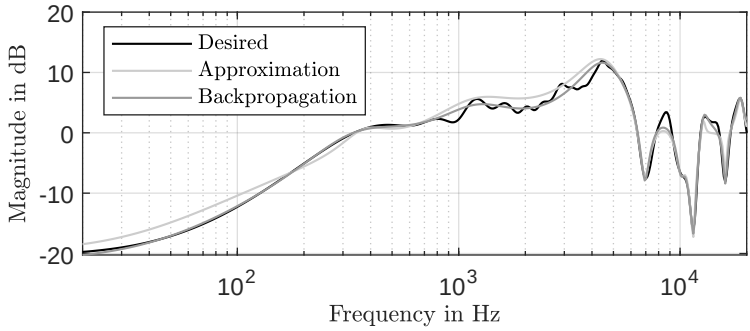

**Figure 3.32:** Histogram of relative frequencies of the achieved LSD improvement or degradation through optimization with the backpropagation algorithm. Here, the LSD values are rounded to an accuracy of 0.1 dB. Additionally, the cumulative frequency is plotted.

In order to use the bandwidth $f_b$ of the peak filters in the backpropagation algorithm as described in Section 3.4.2, the Q-factors from the approximation results are converted into bandwidths according to Eq. (3.13). Due to the already successful approximation results achieved in Section 3.3.2, the task of the backpropagation algorithm is the post-optimization of these results rather than simply approximating the HRTF magnitudes. Therefore, the histogram in Fig. 3.32 visualizes relative and cumulative frequencies

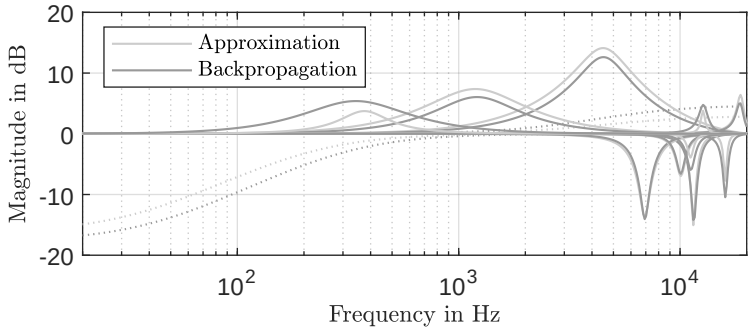
for the differences of LSD between post-optimized results through backpropagation and approximation results from Section 3.3.2 that are taken as initialization. Here, negative differences indicate improvements in the LSD due to the post-optimization procedure. For representation purposes, the differences of LSD are rounded to an accuracy of 0.1 dB, meaning that the bar at a difference in LSD of 0 dB includes slight improvements and degradations up to 0.05 dB. As can be seen, most of the optimizations achieve LSD improvements around 0.1 dB (27.6 %) and 0.2 dB (20.6 %). The highest LSD improvement of 0.84 dB is achieved for the right ear HRTF magnitude of *Subject_010* for a relative azimuth of $\varphi_{\mathrm{rel}} = 20°$ and an elevation of $\theta = -45°$. In Fig. 3.33, the magnitude responses of the approximation used as initialization and the post-optimization through backpropagation algorithm are visualized for this case. As can be seen in Fig. 3.33(a), the backpropagation algorithm fulfills the task of optimizing the interaction between the individual filters. Especially at frequencies below 200 Hz and in the frequency region between 800 Hz and 5 kHz, the backpropagation algorithm reduces the difference between approximation and desired magnitude response. The reason for this reduction in LSD is the adjustment of the gains of the individual filters shown in Fig. 3.33(b).

In addition to the relative frequencies of various levels of difference in LSD shown in Fig. 3.32, the cumulative frequency contains the information that 85.9 % of the post-optimizations achieve improvements in the LSD or at most a degradation of 0.05 dB. For 73.5 % of the post-optimizations, the LSD improvement is even higher than 0.05 dB. Although most of the optimizations achieve an improvement of the LSD, some of them fail and even considerably increase the LSD. Five of the 980 optimizations (0.51 %) even show an increase in the LSD above 2 dB. The highest LSD increase is achieved by the optimization for the left ear HRTF magnitude of *Subject_009* for a relative azimuth of $\varphi_{\mathrm{rel}} = -20°$ and an elevation of $\theta = 180°$. Here, the LSD is increased by 7.59 dB from 1.21 to 8.80 dB. As can be seen in Fig. 3.34, this degradation of the approximation accuracy is caused by the strong increase of the negative gain of the HFS that attenuates all frequencies above 2 kHz. Additionally, also the increase of the negative gain of the LFS diminishes the approximation accuracy in the low frequencies. These problems in the adjustment of the gains for LFS and HFS are also visible for the other optimizations that increase the LSD in comparison to the approximation used as initialization. In most of the cases, the LFS is the main reason for improper optimization increasing the LSD. Similar characteristics have already be seen for the failed approximations through backpropagation in Section 3.4.3.

Nevertheless, Figs. 3.32 and 3.33 indicate that the backpropagation algorithm is able to fine-tune the gains of the individual filters in the parametric IIR filter cascade. Finally, the optimization results of the backpropagation algorithm should be compared to the results achieved by the Levenberg-



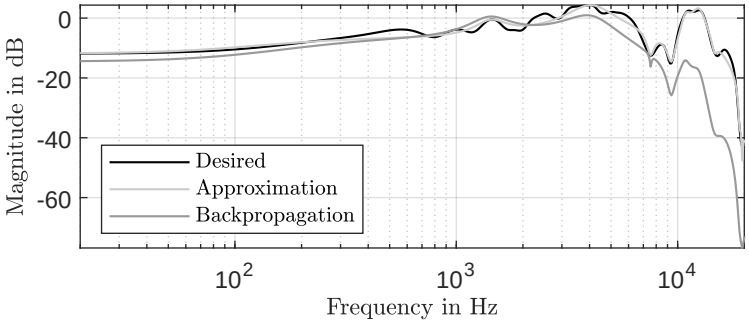

***(a)** Filter cascade*



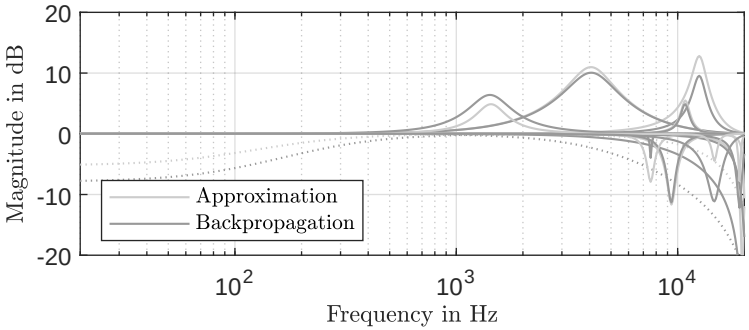

***(b)** Individual filter stages*

**Figure 3.33:** Magnitude responses of (a) the whole filter cascade and (b) the individual filter stages for the optimization through backpropagation with the highest LSD improvement (*Subject_010*, right ear, $\varphi_{\text{rel}} = 20°$, $\theta = -45°$). Here, the LSD of the approximation is reduced from 1.49 to 0.65 dB.

Marquardt algorithm in Section 3.3.2. Therefore, Fig. 3.35 shows the magnitude responses of the approximation results for the case (*Subject_012*, right ear, $\varphi_{\text{rel}} = -55°$, $\theta = 135°$) in which the Levenberg-Marquardt algorithm achieves the highest reduction of the LSD. Here, the Levenberg-Marquardt algorithm reduces the LSD by 1.51 dB, whereas the backpropagation algorithm slightly increases the LSD by 0.07 dB. Similar to Fig. 3.20, the high reduction in the LSD results from the proper approximation of the frequency region between 100 Hz and 1.5 kHz, where both the initial approximation and the optimization through backpropagation algorithm fail to approach the desired magnitude response. The reason for the success of the



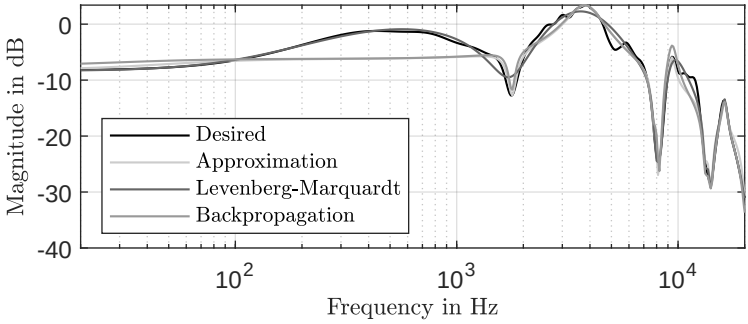

***(a)** Filter cascade*



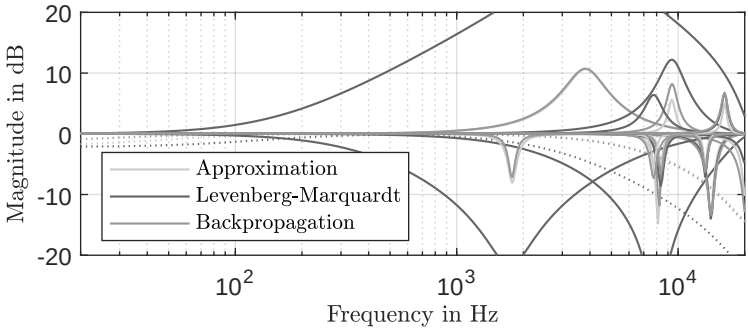

***(b)** Individual filter stages*

**Figure 3.34:** Magnitude responses of (a) the whole filter cascade and (b) the individual filter stages for the optimization through backpropagation with the highest LSD degradation (*Subject_009*, left ear, $\varphi_{\text{rel}} = -20°$, $\theta = 180°$). Here, the LSD of the approximation is increased from 1.21 to 8.80 dB.

Levenberg-Marquardt algorithm is the extreme increase of the gain of the peak filter at 3.8 kHz together with the increase of the negative gain of the peak filter at 1.8 kHz. Additionally, the positive and negative gains of the other peak filters and the HFS are also increased to compensate the high amplification introduced by the peak filter at 3.8 kHz. In Fig. 3.20, also an extreme reduction of the Q-factor of a single peak filter was the reason for the high LSD improvement in the optimization result. In Fig. 3.36, the LSD improvements of the Levenberg-Marquardt algorithm are summarized in a histogram with relative and cumulative frequency. As can be seen, the Levenberg-Marquardt algorithm improves every approximation by at least




***(a)** Filter cascade*




***(b)** Individual filter stages*

**Figure 3.35:** Magnitude responses of (a) the whole filter cascade and (b) the individual filter stages for the optimization through backpropagation and Levenberg-Marquardt algorithm for the right ear of *Subject_012* and a direction of $\varphi_{\text{rel}} = -55^\circ$ and $\theta = 135^\circ$. Here, the Levenberg-Marquardt algorithm reduces the LSD of the approximation from 2.32 to 0.81 dB, whereas the backpropagation algorithm increases the LSD slightly to 2.39 dB.

0.05 dB.

The high variations of the filter parameters that help the Levenberg-Marquardt algorithm to improve the approximation result are not visible in the optimization results of the backpropagation algorithm, so that the optimization through backpropagation algorithm is more restricted, which leads to less improvements. Contrarily, the high gains in the peak filter implementation for the optimized parameters through Levenberg-Marquardt algorithm are generally not wanted and should be avoided.

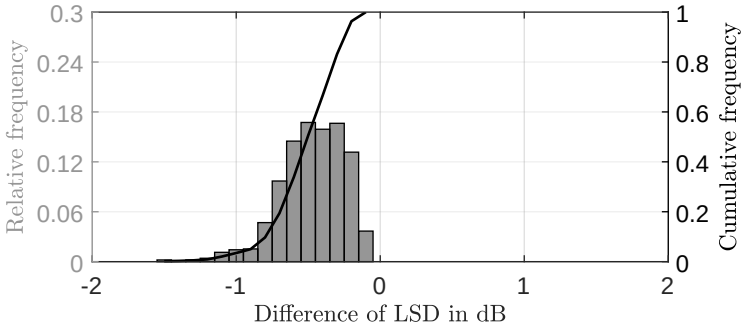
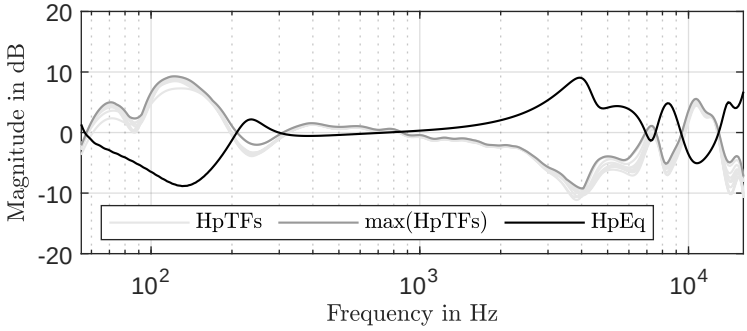



**Figure 3.36:** Histogram of relative frequencies of the achieved LSD improvement through optimization with the Levenberg-Marquardt algorithm. Here, the LSD values are rounded to an accuracy of 0.1 dB. Additionally, the cumulative frequency is plotted.

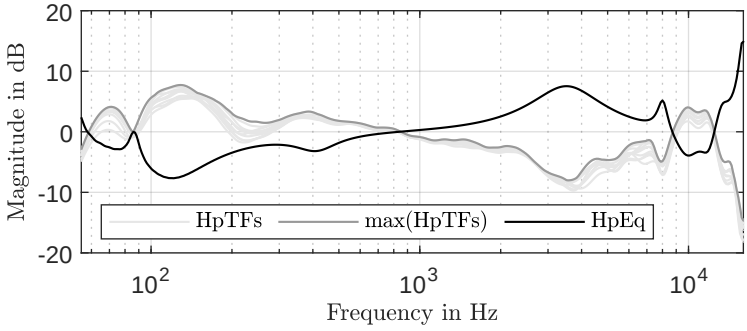
## 3.5 Approximation of Headphone Equalization

In addition to the usage of HRTFs, HpEq is needed to get rid of the additional spectral coloration that is introduced by the HpTF during headphone playback (see Section 2.2.3). In this section, the magnitude responses of the HpEq are approximated with the procedure explained in Section 3.3.

In [Masiero and Fels, 2011], Masiero and Fels proposed a method for robust HpEq. Here, the term robustness means that high peaks inside the transfer function of the HpEq, which would lead to high amplifications of narrow frequency bands, should be avoided. For this, several measurements of the HpTF are taken including a replacement of the headphone in between of the individual measurements. Replacing the headphone leads to small differences in the position of the headphone that introduce differences in the acoustical path, and thus differences in the HpTF (see Fig. 3.37). For some positions, even deeper notches appear that would result in peaks inside the HpEq when inverting the HpTF. If during playback the HpTF contains the same notch, the peak in the HpEq will cancel the notch in the HpTF as desired. However, if the HpTF differs from the one during equalization process, the notch may not appear in the HpTF, and therefore the peak inside the HpEq will introduce undesired effects by amplifying this frequency region. Therefore, in [Masiero and Fels, 2011], several measurements are compared and the maximum magnitude value between all of them is taken for every frequency bin, so that strong notches that appear in a single measurement are ignored. In this way, strong notches




***(a)** Left ear*




***(b)** Right ear*

**Figure 3.37:** Magnitude responses for ten measured HpTFs per ear and the maximum magnitude per frequency bin for the Beyerdynamic DT770 Pro 250 Ohm headphone at (a) the left ear and (b) the right ear of the Neumann KU100 dummy-head. Additionally, the magnitude response of the approximated HpEq by using ten peak filters in the frequency range between 55 Hz and 16 kHz is shown.

are deleted before their inversion leads to strong peaks. In Fig. 3.37, the maximum magnitude value per frequency bin is evaluated for ten measured HpTFs in the frequency range between 55 Hz and 16 kHz.

Since inverting the maximum magnitude across HpTFs requires added information about the phase, the approximation procedure using parametric IIR filters described in Section 3.3 can be used to deliver an equalized magnitude response with a minimum-phase characteristic. In order to allow the described procedure to equalize a given magnitude response rather than

approximating it, the calculation of the approximation error per frequency bin in Eq. (3.31) has to be modified to

$$E_{\mathrm{dB,eq}}(k) = |H_{\mathrm{hp}}(k)|_{\mathrm{dB}} + |H_{\mathrm{eq}}(k)|_{\mathrm{dB}}, \tag{3.66}$$

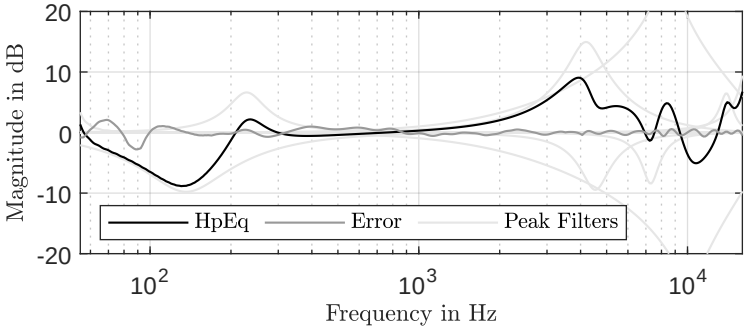
where $|H_{\mathrm{hp}}(k)|_{\mathrm{dB}}$ defines the magnitude response of HpTF and $|H_{\mathrm{eq}}(k)|_{\mathrm{dB}}$ defines the magnitude response of the HpEq. In this way, the equalization error $E_{\mathrm{dB,eq}}(k)$ indicates the frequency bins that are not accurately equalized. Thus, the equalization error achieved by a given number of filters can be used to initialize gain $G$ and center frequency $f_c$ of the next peak filter in the cascade. In addition to the adjustment of the error calculation, also the frequency region of interest is changed to frequencies in between of 55 Hz and 16 kHz in order to avoid high amplifications of frequencies above 16 kHz. Outside the given band, the desired magnitude is faded to 0 dB. Due to the desired magnitude of 0 dB at low and high frequencies, the equalization is performed without the usage of LFS and HFS.

In Fig. 3.37, the magnitude response of the approximated HpEq by using ten peak filters is shown for the left and the right ear. As can be seen, the HpEq inverts the general characteristic of the maximum magnitude values among the magnitude responses of the individual HpTF measurements. Additionally, Fig. 3.38 shows the error of the equalization process and the magnitude responses of the individual peak filters. The magnitude responses of the individual peak filters for the left ear in Fig. 3.38(a) indicate the unconstrained gain in the post-optimization process. Nevertheless, the low error magnitude shows the success of the equalization process. Here, the maximum absolute errors for the left and right ear are 2.77 dB and 1.52 dB, respectively. Except for these maximum error values at frequencies below 100 Hz, the equalization error does not exceed 1 dB for higher frequencies.

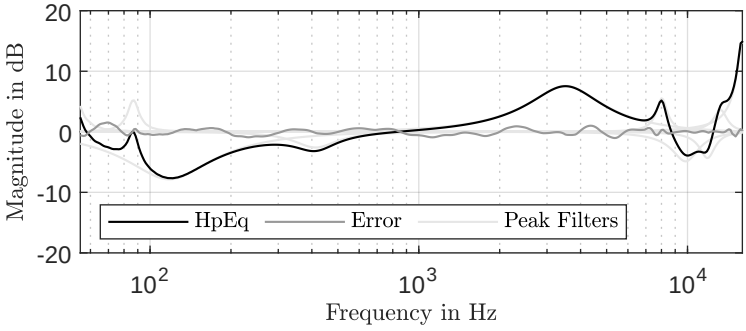
## 3.6 Summary

In binaural synthesis through headphones, HRIRs are used to create a virtual sound source at a given position in 3D space. In most of the cases, these HRIRs are implemented as FIR filters. However, a cascade of low-order IIR filters can approximate the magnitude response of the given HRTFs with a much lower number of coefficients. By using parametric IIR filters like shelving and peak filters, the individual filter stages can be represented by using only three parameters (cut-off or center frequency, gain, and Q-factor). In this way, the memory requirements for saving the HRIRs or HRTFs are reduced.

The approximation of the HRTF magnitude responses is based on a two-step procedure. In a first step, the individual filter stages are consecutively added, initialized and tuned. Here, the remaining approximation error is used as basis for the initialization and the update of the next filter




***(a)** Left ear*




***(b)** Right ear*

**Figure 3.38:** Magnitude responses of the HpEqs shown in Fig. 3.37 together with the magnitude responses of the individual peak filters and the error per frequency bin for (a) the left ear and (b) the right ear.

stage. In a second step, the interaction between the filter stages in the cascade is post-optimized based on the Levenberg-Marquardt algorithm. Using a cascade of one LFS, one HFS, and ten peak filters has shown an accurate approximation of the given HRTF magnitude responses. For 84.7 % of the HRTFs, this number of filter stages is sufficient to produce an approximation error that falls within a 2 dB tolerance.

In addition to the classic approximation procedure, also a novel approach for HRTF magnitude response approximation based on the instantaneous backpropagation algorithm is presented. This algorithm uses the gradient flow through the cascade in order to update control parameters of the individual filter stages. Therefore, the local gradients of the filter outputs

with respect to the control parameters are calculated for shelving and peak filters.

Finally, the parametric IIR filter cascade is used for HpEq, too. Here, the maximum magnitude value among a given set of HpTF measurements is equalized in the frequency range of 55 Hz to 16 kHz using a cascade of ten peak filters.

CHAPTER 4

# Spatial Interpolation of HRTFs

In practical implementations of binaural synthesis through headphones often only a finite number of HRTFs is used. Usually these HRTFs are measured for a given set of azimuths and elevations at a fixed distance. Thus, often a low angular resolution is obtained. Here, the usage of spatial interpolation of the HRTFs has two important reasons. Firstly, the spatial resolution can be increased by interpolating the HRTF of a static direction between two ore more measured directions. Secondly, the realization of moving virtual sound sources requires smooth transitions between the used HRTFs in order to avoid audible discontinuities.

In [Jot et al., 1995], these two cases are separated by using different denominations. Here, the term interpolation is only used for generating a static virtual sound source at a direction not included in the measured HRTF data set. Thus, interpolation is defined as the process of calculating an intermediate HRTF from a finite number of measured HRTFs. Contrarily, updating the HRTF filters while synthesizing a moving virtual sound source in order to achieve smooth transitions without audible artifacts is called commutation [Jot et al., 1995]. Especially for IIR filter implementations, this update causes a mismatch between the internal states of the recursive part of the filter and the new coefficients, which results in audible clicks. Although commutation is usually implemented in the same way as interpolation, Jot et al. [Jot et al., 1995] decided to separate the two cases based on the fact that intermediate filters used for moving virtual sound sources do not have to achieve valid static directional filters. In this work, both cases are summarized under the term interpolation in order to emphasize the equality in calculation of the intermediate filters.

In the following, firstly, previous research on interpolation of HRTFs and HRIRs is summarized. Afterwards, an algorithm for interpolating the parametric IIR filters from Chapter 3 is proposed. Then, the same interpolation algorithm is used to generate moving virtual sound sources with smooth transitions. Finally, spatial interpolation using parametric IIR filters is concluded.

## 4.1 Research on Spatial Interpolation

In order to generate an ideal binaural reproduction, the spatial resolution of measured HRTFs should be equivalent to the MAA that humans are able to detect [Mills, 1958, Perrott, 1984]. However, due to memory limitations and the costly HRTF measurement procedure, it is not practical to measure HRTFs for a resolution similar to the MAA (1°-7°). Thus, interpolation is used to increase the spatial resolution by calculating intermediate HRTFs between measured ones. This procedure is based on the assumption that intermediate directions contain spectral features between the ones of the neighboring measured directions [Begault, 1994]. Although previous investigations on the movement of spectral notches with direction confirm this assumption, interpolation does not always achieve satisfactory results.

Interpolation can be done in time-domain as well as in frequency-domain [Jot et al., 1995]. Note that a compensation of the time of arrival inside the individual HRIRs is needed in both of the domains in order to improve the interpolation result [Matsumoto et al., 2004, Ajdler et al., 2005]. The reason for that can be best explained in the time-domain, where a simple interpolation of two HRIRs with different times of arrival, e.g. two HRIRs in the horizontal plane, would yield an HRIR with two main peaks of half of the amplitude rather than a singular peak in between of the two times of arrival [Begault, 1994]. Additionally, this interpolation will result in comb filtering effects [Lindau et al., 2010]. In order to avoid these effects, magnitude and phase information of the HRTFs should be separated before interpolating them.

In the following, different interpolation approaches for FIR and IIR representations of the corresponding HRTFs are explained. Additionally, the angular resolution of measured HRTFs required for successful interpolation is discussed and methods for dynamic HRTF interpolation are mentioned.

### 4.1.1 FIR Filter Interpolation

As already explained in the previous introduction, the interpolation of HRIRs and HRTFs can be done either in time- or frequency-domain [Jot et al., 1995]. Due to the linearity of the Fourier transform, both methods yield the same result. However, if the delays of the HRIRs and accordingly the phases of the HRTFs vary considerably, disturbing comb filtering effects

are produced in the interpolated HRTFs [Begault, 1994, Huopaniemi, 1999]. Since HRIRs can be represented as minimum-phase approximations in conjunction with a delay representing the ITD (see Section 2.1.4) [Kistler and Wightman, 1992], interpolating this minimum-phase FIR filters and the delays separately can be used to solve this issue [Huopaniemi, 1999].

Bilinear interpolation can be used to calculate an intermediate HRTF or HRIR from a set of neighboring minimum-phase HRIRs or HRTFs in azimuth as well as elevation [Begault, 1994, Savioja et al., 1999]. For this purpose, a grid has to be defined for the entire measurement space. Here, the measurement space is either divided into rectangular or triangular areas [Freeland et al., 2002]. Afterwards, intermediate directions inside a given area are interpolated by using the information of the directions forming this area. Bilinear rectangular interpolation uses four neighboring directions (see Fig. 4.1), whereas for bilinear triangular interpolation only three adjacent directions are interpolated. In the following, the interpolation is described in time-domain.

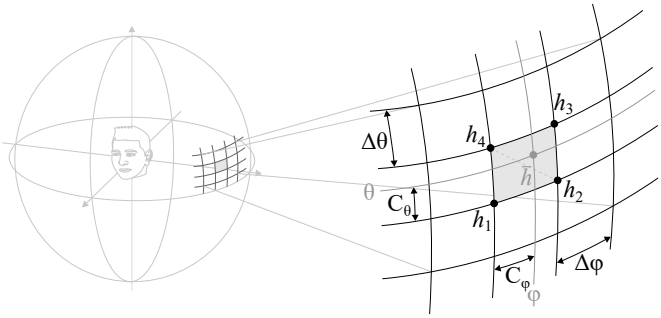



**Figure 4.1:** Principle of bilinear rectangular interpolation in order to calculate an intermediate HRIR $\bar{h}(n)$ at the relative position $(C_\varphi, C_\theta)$ inside a given area defined by measured HRIRs ($h_1(n)$, $h_2(n)$, $h_3(n)$, and $h_4(n)$) in the vertical-polar coordinate system. The azimuthal and elevation resolution of the measurement grid are specified by $\Delta\varphi$ and $\Delta\theta$, respectively. Additionally, bilinear triangular interpolation is indicated.

In Fig. 4.1, the division of the 3D measurement space into a grid of rectangular areas with a horizontal resolution of $\Delta\varphi$ and a vertical resolution of $\Delta\theta$ is illustrated for the vertical-polar coordinate system from Fig. 1.1(a). Here, the four HRIRs at the corners of the rectangle used to calculate the intermediate HRIR $\bar{h}(n)$ are labeled as $h_1(n)$, $h_2(n)$, $h_3(n)$, and $h_4(n)$. Note that always two HRIRs share a common azimuth or elevation. The position of the intermediate direction inside the given rectangle is defined by the relative angular positions $C_\varphi$ and $C_\theta$. Although Fig. 4.1 illustrates the

principle of bilinear interpolation for the vertical-polar coordinate system from Fig. 1.1(a), the same principle can be used for bilinear interpolation in the interaural-polar coordinate system from Fig. 1.1(b).

By calculating the weighted mean across the four neighboring HRIRs, the intermediate HRIR $\bar{h}(n)$ can be obtained as

$$\begin{aligned}\bar{h}(n) =&(1-c_\varphi)(1-c_\theta)h_1(n) + c_\varphi(1-c_\theta)h_2(n) \\ &+ c_\varphi c_\theta h_3(n) + (1-c_\varphi)c_\theta h_4(n),\end{aligned} \tag{4.1}$$

where the weights $c_\varphi$ and $c_\theta$ can be calculated from the relative angular positions, $C_\varphi$ and $C_\theta$, and the spatial resolutions, $\Delta\varphi$ and $\Delta\theta$, as

$$c_\varphi = \frac{C_\varphi}{\Delta\varphi} = \frac{\varphi \bmod \Delta\varphi}{\Delta\varphi}, \tag{4.2}$$

$$c_\theta = \frac{C_\theta}{\Delta\theta} = \frac{\theta \bmod \Delta\theta}{\Delta\theta}, \tag{4.3}$$

respectively. When using bilinear triangular interpolation, only three of the four neighboring directions are used to form a triangle that includes the desired intermediate direction [Freeland et al., 2002]. Thus, for the example shown in Fig. 4.1, the intermediate HRIR $\bar{h}(n)$ can be calculated as

$$\bar{h}(n) =(1-c_{\varphi,\mathrm{tri}}-c_{\theta,\mathrm{tri}})h_3(n) + c_{\varphi,\mathrm{tri}}h_4(n) + c_{\theta,\mathrm{tri}}h_2(n). \tag{4.4}$$

Here, the weighting factors $c_{\varphi,\mathrm{tri}}$ and $c_{\theta,\mathrm{tri}}$ are given as

$$c_{\varphi,\mathrm{tri}} = \frac{|\varphi-\varphi_3|}{\Delta\varphi}, \tag{4.5}$$

$$c_{\theta,\mathrm{tri}} = \frac{|\theta-\theta_3|}{\Delta\theta}, \tag{4.6}$$

with $\varphi_3$ and $\theta_3$ being the azimuth and elevation of $h_3(n)$. Note that for directions inside the other triangle depicted in Fig. 4.1, the azimuthal weight $c_{\varphi,\mathrm{tri}}$ from Eq. (4.5) equals the definition of the weight $c_\varphi$ from Eq. (4.2). This equality also holds for the elevation weights $c_{\theta,\mathrm{tri}}$ from Eq. (4.6) and $c_\theta$ from Eq. (4.3).

In [Hugeng et al., 2017], Hugeng et al. showed that bilinear rectangular interpolation results in a lower average mean squared error and a lower spectral distortion than bilinear triangular interpolation, which may result from the better balancing between used HRIRs. A further advantage of bilinear rectangular interpolation is the lower number of transitions between areas of HRIR sets [de Sousa and Queiroz, 2009]. On the other side, bilinear triangular interpolation requires roughly 25 % less multiplications and additions than the bilinear rectangular interpolation [de Sousa and Queiroz, 2009]. In case of single directions, the linear interpolation for azimuthal

directions in the horizontal plane performs better than the one for elevation directions in the vertical plane [Hugeng et al., 2015]. Additionally, Reddy and Hedge [Reddy and Hegde, 2016] proposed to linearly interpolate the phases of the measured HRTFs in the horizontal plane, too.

Instead of directly using the measured HRTFs for bilinear triangular interpolation, Freeland et al. proposed to use one reference HRTF and two inter-positional transfer functions (IPTFs), which give the ratio between the HRTF of a neighboring direction to the reference HRTF [Freeland et al., 2002, Freeland et al., 2004]. In this way, an HRTF can be represented by a cascade of the reference HRTF and the corresponding IPTF between the desired HRTF and the reference HRTF. The advantage of using these IPTFs is the possibility of representing them with a reduced order in comparison to the corresponding HRTF. Thus, the computational efficiency of the interpolation algorithm can be increased.

Expanding the two-dimensional (2D) triangular interpolation to 3D tetrahedral interpolation includes the distance information into the interpolation algorithm [Gamper, 2013], thus HRTFs at various distances can be interpolated. However, for this purpose, an HRTF data set containing HRTFs at different distances is needed. Similar to bilinear interpolation, firstly, a tetrahedron has to be found that encloses the intermediate position. In this way, the measurement space is divided into a grid of non-overlapping tetrahedrons. Afterwards, the weights $c_i$ for the different HRTFs $H_i(z)$ can be calculated from the position of the intermediate HRTF inside the tetrahedron and the geometry of the used tetrahedron. Here, the sum of all four weights $c_i$ should be

$$\sum_{i=1}^{4} c_i = 1 \tag{4.7}$$

in order to yield an interpolation without introducing an additional gain. By adding the weights in Eqs. (4.1) and (4.4), it can be seen that this constraint is fulfilled for every choice of $c_\varphi$ and $c_\theta$. Furthermore, the individual weights are restricted to $0 \leq c_i \leq 1$. Finally, the weighted sum of all HRTFs can be used to calculate the intermediate HRTF to

$$\bar{H}(z) = \sum_{i=1}^{4} c_i H_i(z). \tag{4.8}$$

Note that Eq. (4.8) is a general form of 4-point interpolation independent from the chosen measurement grid. In [Gamper, 2013], the interpolation was performed solely on the magnitudes of the measured HRTFs. However, Hugeng et al. [Hugeng et al., 2017] showed that a better interpolation is achieved by using minimum-phase HRIRs rather than HRTF magnitudes.

### 4.1.2 IIR Filter Interpolation

A direct translation of the methods for interpolating FIR filters into the interpolation of IIR filters would be the bilinear interpolation of the filter coefficients [Queiroz and de Sousa, 2010]. However, for IIR filters, the numerator and denominator polynomials of the transfer functions have to be simultaneously interpolated, which differs from the direct interpolation of impulse responses and transfer functions. In this way, the positions of poles and zeros are modified, which may result in unstable configurations of the intermediate IIR filter. However, the linear interpolation of stable second-order IIR filters is guaranteed to be stable, too [Jot et al., 1995].

Other interpolation algorithms for IIR filters are based on pole-zero models of the given IIR filters [Wang et al., 2008, Queiroz and de Sousa, 2010]. In [Wang et al., 2008], an indirect interpolation method for pole-zero models was proposed that firstly transforms the pole-zero models into the corresponding HRIRs which are then interpolated by bilinear interpolation as described in the previous section. Finally, the intermediate HRIR is transformed back to a pole-zero model with the same order as the models used for the interpolation. Contrarily, Queiroz and de Sousa [Queiroz and de Sousa, 2010] performed the interpolation directly on the poles and zeros. Here, the HRTFs are approximated by low-order structured IIR filters which are represented by complex poles and zeros in the $z$-plane. These poles and zeros define resonance and anti-resonance frequencies. Applying the interpolation directly on the poles and zeros enables the possibility of interpolating the frequencies of these resonances and anti-resonances in addition to the magnitudes. In order to linearly interpolate the poles and zeros, the poles and zeros of the different HRTFs have to be associated with each other firstly [Queiroz and de Sousa, 2010]. This association is done by ordering the complex poles and zeros based on their frequency and interpolating the $l^{\text{th}}$ pole or zero of all HRTFs linearly with the appropriate weights. In the same way, real poles and zeros are interpolated after sorting them based on their values. Here, an important advantage of structured IIR filters can be noted, because the association of the poles and zeros is already introduced in the approximation procedure, where only a single pole and zero are used per frequency band [Queiroz and de Sousa, 2010]. Since linear interpolation of the gains of the individual HRTFs in both pole-zero models gives poor results in many cases, the overall gain factor of the model is recomputed before the interpolation process.

Moreover, using HRTFs approximated by parametric IIR filters as described in Chapter 3 gives the opportunity of creating intermediate HRTFs by simply interpolating the parameters of the neighboring IIR filters [Ramos and Cobos, 2013]. In [Ramos and Cobos, 2013], both the movement of center frequencies with azimuth and an exemplary interpolation result are shown, indicating that the simple interpolation of IIR filter parameters can

achieve good approximations of intermediate directions.

In [Runkle et al., 1995], two further methods for interpolating pole-zero models approximating DTFs are listed. Firstly, convex combinations of the poles and zeros from two neighboring DTF approximations can be used to compute the intermediate DTF. Secondly, tracked pole-zero configurations from the gradient search algorithm of the approximation procedure can be used to calculate intermediate DTFs. The second method is motivated by the observation that these pole-zero tracks generated in between of two measured directions during the approximation procedure are near to the correct poles and zeros.

### 4.1.3 Localization Accuracy and Spatial Resolution

In [Wightman et al., 1992], Wightman et al. compared the static localization accuracy of measured and interpolated individual HRTFs. Here, the interpolation was performed directly on the measured HRTFs as well as on minimum-phase approximations of these HRTFs. Additionally, different azimuthal ($\Delta\varphi = \{15°, 30°\}$) and elevation resolutions ($\Delta\theta = \{12°, 24°\}$) are used to evaluate their influence on the localization accuracy. The results indicate that measured HRTFs and interpolations using minimum-phase approximations and the higher angular resolution ($\Delta\varphi = 15°$, $\Delta\theta = 12°$) have a comparable localization accuracy. Decreasing the angular resolution has also slightly decreased the accuracy. Contrarily, interpolating directly the measured HRTFs showed worse localization accuracy even for the higher angular resolution. The decreased accuracy is mainly explained by a higher front/back confusion rate.

In contrast to Wightman et al. [Wightman et al., 1992], Wenzel and Foster [Wenzel and Foster, 1993] studied the influence of interpolation on the localization accuracy for non-individual HRTFs of another human subject. Similar to [Wightman et al., 1992], the interpolation was performed on both the measured HRTFs and minimum-phase approximations for different angular resolutions ($\Delta\varphi = \{30°, 60°\}$, $\Delta\theta = 36°$). In order to investigate the influence of different dimensions, the azimuthal and elevation resolution were varied independently, including also the cases in which only one of the dimensions is interpolated. The results indicate that measured and interpolated HRTFs show a similar localization accuracy even for the lowest angular resolution ($\Delta\varphi = 60°$, $\Delta\theta = 36°$). These results are in contradiction to the results from [Wightman et al., 1992]. However, in [Wightman et al., 1992], individual HRTFs were used, thus the conclusion can be made that the usage of non-individual HRTFs degrades the localization accuracy of static virtual sound sources stronger than the interpolation.

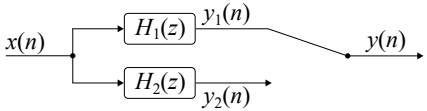
In [Plogsties et al., 2000], both a listening test and LSD were used to identify whether differences between an interpolated HRTF and a measured HRTF are audible for a given resolution of measured HRTFs. The listening

test results indicate that a resolution of 8° is required to achieve inaudible interpolated HRTFs in front of the listener for static virtual sound sources. In order to achieve an LSD of less than 1 dB, a resolution of at least 4° is required. Although the results for dynamic virtual sound sources show similarities, the errors are much less detectable. Furthermore, in [Ajdler et al., 2005], Adjer et al. proposed the azimuthal sampling theorem that relates the maximum azimuthal resolution to the bandwidth of the HRTF. For the entire audible frequency range at a sampling frequency of $f_s = 44.1\,\text{kHz}$, an azimuthal resolution of $\Delta\varphi = 5°$ is required, resulting in 72 different measured azimuths. A lower azimuthal resolution leads to interpolation errors at high frequencies. Thus, HRTF interpolations for lower frequencies need a lower azimuthal resolution, e.g. $\Delta\varphi = 40°$ for frequencies up to 2 kHz [Ajdler et al., 2005].

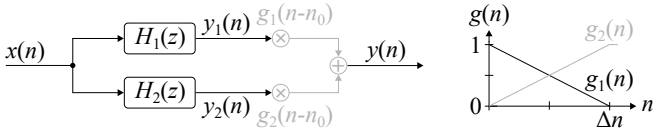
### 4.1.4 Interpolation for Dynamic Virtual Sound Sources

In order to generate dynamic virtual sound sources, interpolation between the measured HRTFs is needed to ensure smooth transitions without audible artifacts even for HRTF databases with high spatial resolution [Duraiswaini et al., 2004]. For this purpose, the approaches described in Sections 4.1.1 and 4.1.2 can be used. Note that fast movements of the virtual sound source demand fast adaptations of the interpolated HRTFs [Jot et al., 1995]. Thus, the interpolation should be done sample-based or block-based for blocks of a few milliseconds [Queiroz and de Sousa, 2010]. Since FIR filter implementations and especially minimum-phase approximations of the HRTFs allow a simple exchange of the coefficients at sample rate, FIR filters are often preferred over IIR filters in dynamic binaural synthesis [Jot et al., 1995]. For IIR filter implementations, updating the filter coefficients causes a mismatch between the new coefficients and the internal states of the recursive part of the filter, which results in audible clicks. In [Välimäki, 1995, Välimäki and Laakso, 1998], different methods for avoiding these artifacts are discussed.

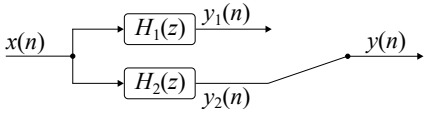
One of these methods that is commonly used in audio applications is the cross-fading method [Välimäki, 1995]. As can be seen in Fig. 4.2, this method uses two or more filters in parallel. While the input signal is connected to all filters, only one filter is used to generate the current output signal. When new coefficients are needed, the connection of the output signal is switched to the corresponding filter. Instead of switching the filters abruptly, the outputs of the two filters are cross-faded by smoothly decreasing the weight of the previous filter from 1 to 0 and simultaneously increasing the weight of the new filter from 0 to 1. The duration of the cross-fading is called transition time $\Delta n$. Using this method for binaural synthesis, would need a cross-fade between three or four filters due to the synthesis in the horizontal as well as the vertical plane [Välimäki, 1995].




***(a)*** *using $H_1(z)$ before transition*




***(b)*** *transition from $H_1(z)$ to $H_2(z)$ starting at $n = n_0$*




***(c)*** *using $H_2(z)$ after transition*

**Figure 4.2:** Principle of cross-fading between two filters $H_1(z)$ and $H_2(z)$. (a) At the beginning, the overall output $y(n) = y_1(n)$ is connected to the output of $H_1(z)$. (b) Then, during the transition from $H_1(z)$ to $H_2(z)$ starting at $n = n_0$, a weighted sum of both filter outputs is used to calculate the overall output $y(n) = g_1(n)y_1(n) + g_2(n)y_2(n)$. The duration of the transition is given by $\Delta n$. (c) Finally, the overall output $y(n) = y_2(n)$ is solely determined by the output of $H_2(z)$.

Furthermore, in [Verhelst and Nilens, 1986], a similar approach for updating IIR filters using a parallel structure of multiple IIR filters is proposed (see Fig. 4.3). In this approach only a single IIR filter is connected to the input, whereas the outputs of all filters are superposed to yield the overall output. After updating another filter with the new coefficients, the connection of the input signal is switched to this filter. Consequently, this approach can be called input-switching method [Välimäki and Laakso, 1998]. In this way, the outputs of the filters that are disconnected from the current input signal can decay. After a given decaying time, the output of the filter becomes negligible, thus the filter can be re-used for being updated and filtering the input signal with the new coefficients [Verhelst and Nilens, 1986]. The number of parallel IIR filters used to implement the interpolation algorithm depends on the ratio between impulse response length of the filters and interpolation period.

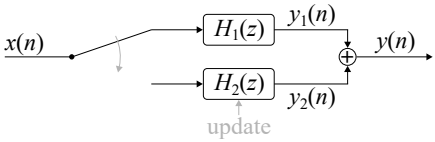



**Figure 4.3:** Input-switching method for updating $H_2(z)$ with the new coefficients while simultaneously filtering the input signal $x(n)$ with the previous coefficients included in $H_1(z)$. After updating $H_2(z)$, the connection of the input signal $x(n)$ is switched to $H_2(z)$ and $H_1(z)$ is available to be updated.

## 4.2 Spatial Interpolation using Parametric IIR Filters

In this section, the spatial interpolation of HRTFs approximated by parametric IIR filters is explained. For this, bilinear interpolation is used to calculate the parameters for an intermediate direction from the parameters of the neighboring directions. At first, this interpolation method is used to create static virtual sound sources between the measured directions. Afterwards, extensions needed to achieve moving virtual sound sources without audible artifacts are described.

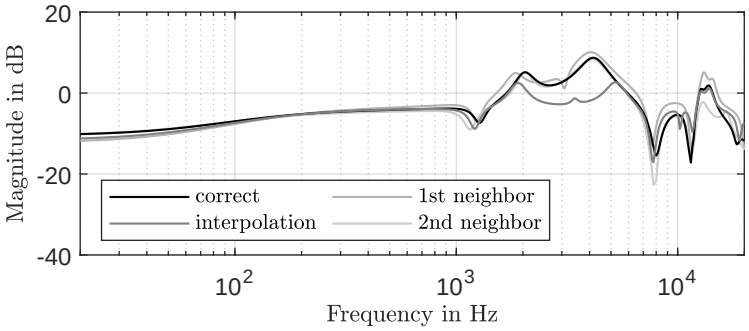
### 4.2.1 Static Virtual Sound Sources

As proposed in [Ramos and Cobos, 2013], HRTFs approximated by parametric IIR filters as described in Chapter 3 can be interpolated by interpolating their parameters. Although an exemplary result for horizontal interpolation between two neighboring directions is shown in [Ramos and Cobos, 2013], no further explanations on the parameter interpolation are given. In the following, the interpolation of the filter parameters is analyzed in details.

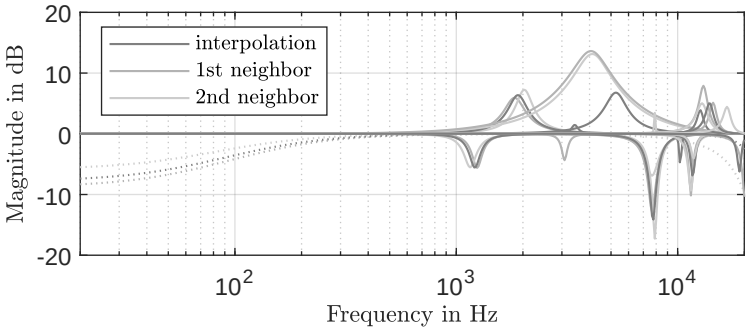
In order to interpolate the filter parameters, bilinear rectangular interpolation can be used. For this, firstly, the peak filters contained in the parameter matrices $\mathbf{P}_{\text{approx}}$ defined in Eq. (3.37) have to be sorted by their center frequencies $f_{\text{c,p}}$ in ascending order. Afterwards, Eq. (4.1) can be modified to

$$\begin{aligned}\bar{\mathbf{P}}_{\text{approx}} =&(1-c_\varphi)(1-c_\theta)\mathbf{P}_{\text{approx,1}} + c_\varphi(1-c_\theta)\mathbf{P}_{\text{approx,2}} \\ &+ c_\varphi c_\theta \mathbf{P}_{\text{approx,3}} + (1-c_\varphi)c_\theta \mathbf{P}_{\text{approx,4}},\end{aligned} \tag{4.9}$$

where $\bar{\mathbf{P}}_{\text{approx}}$ and $\mathbf{P}_{\text{approx,1}}$ to $\mathbf{P}_{\text{approx,4}}$ specify the parameter matrices for interpolated and measured neighboring directions, respectively. Additionally, the coefficients $c_\varphi$ and $c_\theta$ are given by Eqs. (4.2) and (4.3). By interpolating the parameter matrix according to Eq. (4.9), every filter parameter ($f_c$, $G$, and $Q$) is interpolated in the same way. Similarly, bilinear triangular interpolation can be used by modifying Eq. (4.4). However, the




***(a)** Filter cascade*



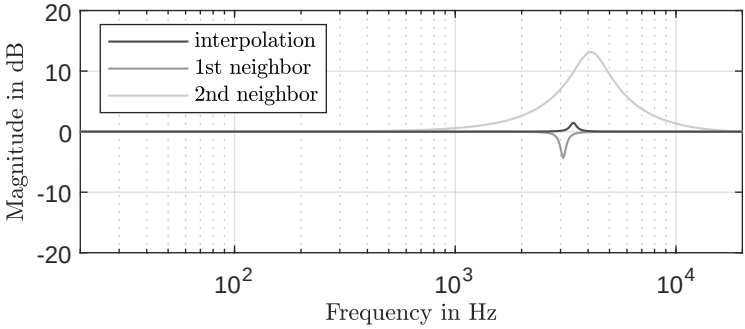

***(b)** Individual filter stages*

**Figure 4.4:** Magnitude responses of (a) the whole filter cascade and (b) all individual filter stages for an intermediate direction (*Subject_065*, left ear, $\varphi = -10°$, $\theta = 0°$) interpolated from two neighboring directions ($\varphi_1 = -15°$, $\theta_1 = 0°$; $\varphi_2 = 0°$, $\theta_2 = 0°$) using simple parameter interpolation. Additionally, the originally approximated magnitude response for the desired direction is drawn as reference.

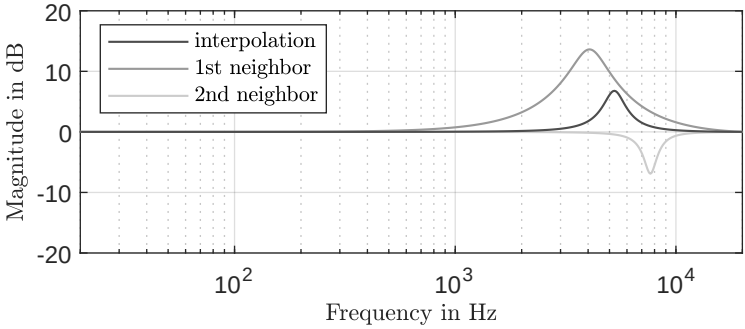
following analysis uses bilinear rectangular interpolation.

Figure 4.4 shows the interpolation result for the left ear of *Subject_065* from the CIPIC database [Algazi et al., 2001b], and a direction of $\varphi = -10°$ and $\theta = 0°$. Here, the resolution of the measurement grid is set to $\Delta\varphi = 15°$ for azimuth and $\Delta\theta = 11.25°$ for elevation. Thus, the neighboring directions are defined as $(\varphi_1 = -15°, \theta_1 = 0°)$, $(\varphi_2 = 0°, \theta_2 = 0°)$, $(\varphi_3 = 0°, \theta_3 = 11.25°)$, and $(\varphi_4 = -15°, \theta_4 = 11.25°)$. According to Eqs. (4.2) and (4.3), the interpolation weights are calculated to $c_\varphi = 1/3$

and $c_\theta = 0$, respectively. Since $c_\theta = 0$, only the first two neighboring directions are used for calculating the intermediate direction. Although the horizontal interpolation shows reasonable results (see Fig. 4.4(a)) in low and high frequencies, the interpolated magnitude response suffers from strong deviations in the frequency range between 2 kHz and 5 kHz. The reason for these deviations can be seen in Fig. 4.4(b), where the magnitude responses of the individual filter stages are plotted. The strong peak at $f_c \approx 4$ kHz that is visible in both neighboring directions is not contained in the interpolation result, which can be explained by an incorrect assignment of the peak filters.




***(a)** Third peak filter*




***(b)** Fourth peak filter*

**Figure 4.5:** Magnitude responses of (a) the third and (b) fourth peak filter of the intermediate and the two neighboring directions from Fig. 4.4.

The assignment of the filter stages is shown in Fig. 4.5 in more detail. In Figs. 4.5(a) and 4.5(b), the magnitude responses of the third and fourth

peak filter are drawn for the two neighboring directions as well as for the intermediate direction. As can be seen, the peak filter at $f_c \approx 4\,\text{kHz}$ is the third peak filter for the second neighboring direction (see Fig. 4.5(a)) and the fourth peak filter for the first neighboring direction (see Fig. 4.5(b)). Thus, the additional peak filter at $f_c \approx 3\,\text{kHz}$ for the first neighboring direction results in the incorrect assignment for this peak filter and subsequent ones that is responsible for the inaccurate interpolation result visible in Fig. 4.4(a).

In order to improve the interpolation accuracy, the assignment of the individual peak filters has to be checked before calculating the interpolation. Therefore, the following algorithm is proposed [Nowak and Zölzer, 2022]:

1. Find the closest neighboring direction

$$i_{\text{ref}} = \arg\max_i c_i \quad \text{for } i \in \{1, 2, 3, 4\} \tag{4.10}$$

   by searching for the maximum weight $c_i$ across directions $(\varphi_i,\, \theta_i)$.

2. The parameter matrix $\mathbf{P}_{\text{approx},i_{\text{ref}}}$ of the closest direction $(\varphi_{i_{\text{ref}}},\, \theta_{i_{\text{ref}}})$ is taken as reference parameter matrix.

3. For every peak filter $p_{i_{\text{ref}}}$ of the reference direction $i_{\text{ref}}$, the other neighboring directions $(i \neq i_{\text{ref}})$ are checked for peak filters $p_i$ that have center frequencies $f_{c,\text{p},i}$ within a threshold of

$$\Delta f_c = \left| 20 \log_{10} \left( \frac{f_{c,\text{p},i_{\text{ref}}}}{f_{c,\text{p},i}} \right) \right| \leq 2\,\text{dB} \tag{4.11}$$

   around the reference center frequency $f_{c,\text{p},i_{\text{ref}}}$.

4a. If a peak filter $p_i$ inside this threshold is found, the sign of the gain $\text{sgn}(G_{\text{p},i})$ is compared to the sign of the gain $\text{sgn}(G_{\text{p},i_{\text{ref}}})$ of the reference peak filter $p_{i_{\text{ref}}}$ in order to interpolate only peak filters that likewise boost or cut the given frequency range.

   If one ore more peak filters $p_i$ are found that fulfill both conditions for a neighboring direction $(i \neq i_{\text{ref}})$, the most similar one in center frequency $\Delta f_c$ and gain $G$ is used for that direction in the interpolation. Additionally, the chosen peak filter $p_i$ can be compared to the next two peak filters of the reference direction before calculating the interpolation in order to check whether a more similar peak filter exists for the reference direction.

   The interpolated parameters of the $p^{\text{th}}_{i_{\text{ref}}}$ peak filter can be calcu-

lated to

$$\left[\bar{f}_{c,\mathrm{p}_{i_{\mathrm{ref}}}} \bar{G}_{\mathrm{p}_{i_{\mathrm{ref}}}} \bar{Q}_{\mathrm{p}_{i_{\mathrm{ref}}}}\right] = \sum_{i=1}^{4} c_i \cdot \left[f_{c,\mathrm{p},i} \;\; G_{\mathrm{p},i} \;\; Q_{\mathrm{p},i}\right]. \tag{4.12}$$

4b. Otherwise, if no peak filter $p_i$ is assigned to the reference peak filter $p_{i_{\mathrm{ref}}}$ for a neighboring direction $i_{\mathrm{ex}}$, this direction is excluded from the interpolation in Eq. (4.12) for that specific peak filter $p_{i_{\mathrm{ref}}}$. In order to still maintain the restriction in Eq. (4.7), the weights of the included neighboring directions ($i \neq i_{\mathrm{ex}}$) have to be modified to

$$c_{i,\mathrm{p}} = \frac{c_i}{1 - c_{i_{\mathrm{ex}}}} \qquad \text{for } i \neq i_{\mathrm{ex}} \tag{4.13}$$

for that specific peak filter $p_{i_{\mathrm{ref}}}$, where $c_{i_{\mathrm{ex}}}$ defines the weight of the excluded direction. In this way, the calculation of the interpolated parameters from Eq. (4.12) is changed to

$$\left[\bar{f}_{c,\mathrm{p}_{i_{\mathrm{ref}}}} \bar{G}_{\mathrm{p}_{i_{\mathrm{ref}}}} \bar{Q}_{\mathrm{p}_{i_{\mathrm{ref}}}}\right] = \sum_{\substack{i=1 \\ i \neq i_{\mathrm{ex}}}}^{4} c_{i,\mathrm{p}} \cdot \left[f_{c,\mathrm{p},i} \;\; G_{\mathrm{p},i} \;\; Q_{\mathrm{p},i}\right]. \tag{4.14}$$

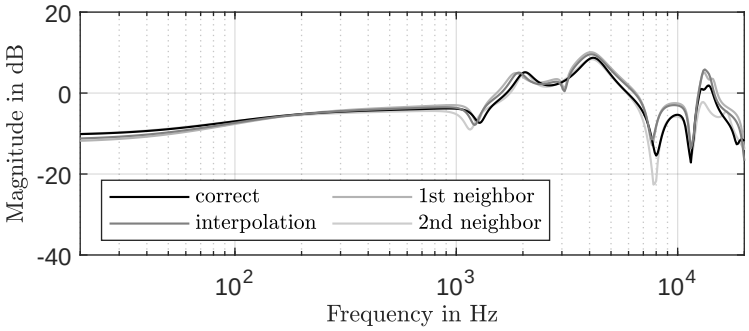
5. Afterwards, the assignment of the peak filters continues with the next peak filter $p_{i_{\mathrm{ref}}} + 1$. Here, the comparison to the neighboring directions starts with the $\{p_i + 1\}^{\mathrm{th}}$ peak filter, where $p_i$ is defined by the peak filter of a neighboring direction $i$ assigned to the previous peak filter $p_{i_{\mathrm{ref}}}$ of the closest direction $i_{\mathrm{ref}}$. This definition of the starting index prevents the double usage of the same peak filter in the interpolation procedure.

6. Finally, the mean magnitude values $\mu_{\mathrm{H}_i,\mathrm{dB}}$ according to Eq. (3.27) and the ITDs of the neighboring directions have to be interpolated as

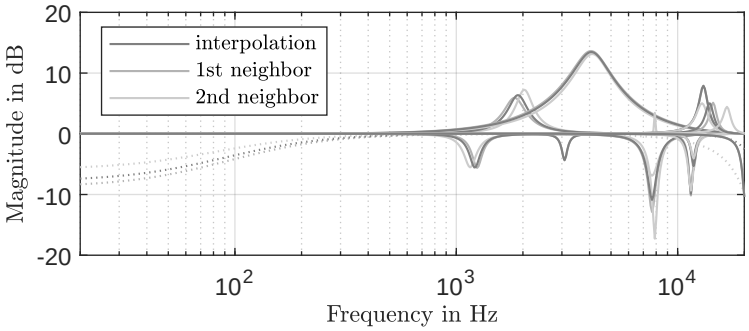
$$\mu_{\bar{\mathrm{H}},\mathrm{dB}} = \sum_{i=1}^{4} c_i \mu_{\mathrm{H}_i,\mathrm{dB}}, \tag{4.15}$$

$$\bar{\mathrm{ITD}} = \sum_{i=1}^{4} c_i \mathrm{ITD}_i. \tag{4.16}$$

In Fig. 4.6, the result of the extended parameter interpolation is plotted for the same example as shown in Fig. 4.4 (*Subject_065*, left ear, $\varphi = -10°$, $\theta = 0°$). By comparing the two interpolated magnitude responses in Figs. 4.6(a) and 4.4(a), the optimized assignment of the peak filters in the extended interpolation algorithm, clearly improves the interpolation result




***(a)** Filter cascade*




***(b)** Individual filter stages*

**Figure 4.6:** Magnitude responses of (a) the whole filter cascade and (b) all individual filter stages of the intermediate direction (*subject_065*, left ear, $\varphi = -10°$, $\vartheta = 0°$) interpolated from two neighboring directions ($\varphi_1 = -15°$, $\vartheta_1 = 0°$; $\varphi_2 = 0°$, $\vartheta_2 = 0°$) using parameter interpolation together with an assignment of the peak filters. Additionally, the originally approximated magnitude response for the desired direction is drawn as reference.

in the mid frequencies. Additionally, Fig. 4.6(b) indicates that all peak filters contained in the closest neighboring direction (first neighbor), are also included in the intermediate direction. Contrarily, the sixth peak filter of the second neighbor at $f_c = 8\,\mathrm{kHz}$ is not considered in the interpolated magnitude response.

By elevating the virtual sound source by $\theta = 5.625°$, the sound source moves from the border of an interpolation rectangle into this area, thus four neighbors are used instead of two. In Fig. 4.7, the magnitude responses

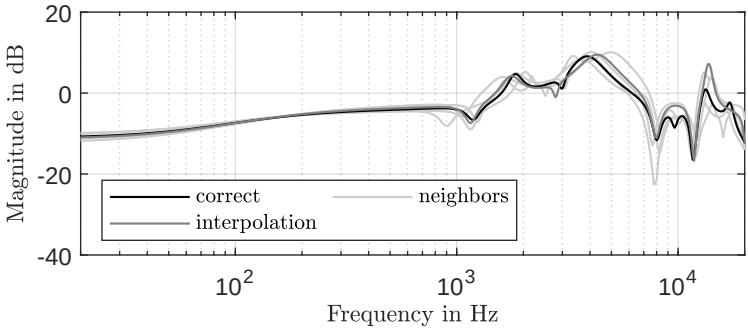
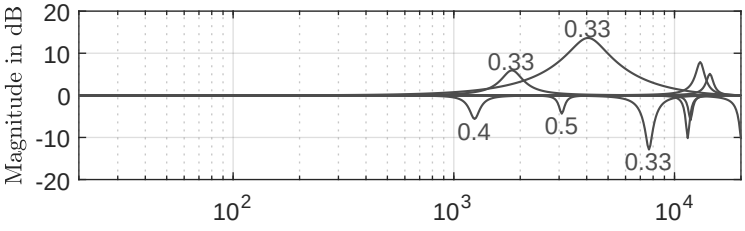


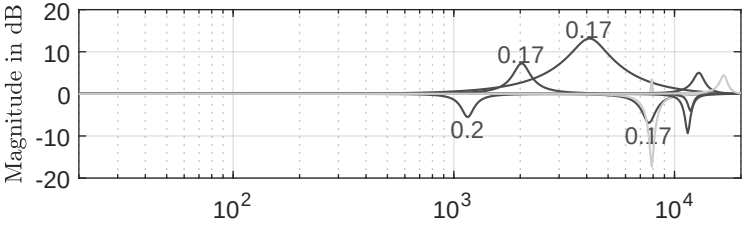

**Figure 4.7:** Magnitude response of the intermediate direction (*Subject_065*, left ear, $\varphi = -10°$, $\vartheta = 5.625°$) interpolated from four neighboring directions ($\varphi_1 = -15°$, $\vartheta_1 = 0°$; $\varphi_2 = 0°$, $\vartheta_2 = 0°$; $\varphi_3 = 0°$, $\vartheta_3 = 11.25°$; $\varphi_4 = -15°$, $\vartheta_4 = 11.25°$) using parameter interpolation together with an assignment of the peak filters. Additionally, the originally approximated magnitude response for the desired direction is drawn as reference.

of the neighboring directions, the interpolated magnitude response for the desired direction ($\varphi = -10°$, $\theta = 5.625°$), and the originally approximated magnitude response for this direction are shown. As can be seen, the interpolated magnitude response lies between the neighboring magnitude responses.

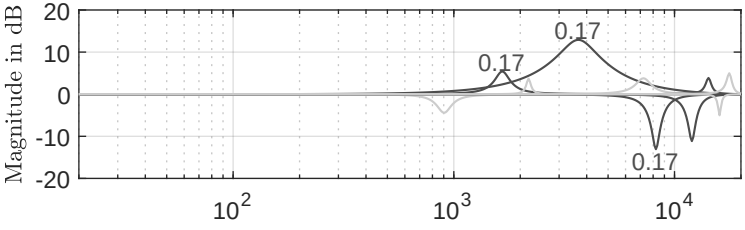
Furthermore, Fig. 4.8 illustrates the assignment of the peak filters of the neighboring directions to the peak filters of the closest direction. In Fig. 4.8(a), the magnitude responses of the individual peak filters for the reference direction ($\varphi_1 = -15°$, $\theta_1 = 0°$) are plotted. Additionally, Figs. 4.8(b) - 4.8(d) show the magnitude responses of the individual peak filters for the other neighboring directions. Here, dark gray peak filters indicate peak filters that are used for the interpolation whereas light gray peak filters are excluded from this calculation, because they are not assigned to the peak filters from the reference direction in Fig. 4.8(a). Besides of the magnitude responses, also the corresponding weights for calculating the first five peak filters of the intermediate direction are given in the figures. As can be seen, the basic weights $c_1 = c_4 = 0.33$ and $c_2 = c_3 = 0.17$ that can be calculated from Eqs. (4.1) - (4.3) are only used for peak filters that are contained in all four neighboring directions. For the other peak filters, e.g. the first peak filter with negative gain at $f_c \approx 1.24\,\mathrm{kHz}$ ($c_{1,1} = c_{4,1} = 0.4$, $c_{2,1} = 0.2$), the weights are changed according to Eq. (4.13). Although the third neighboring direction in Fig. 4.8(c) also contains a similar peak filter with negative gain at $f_c \approx 900\,\mathrm{Hz}$, this peak filter is excluded from



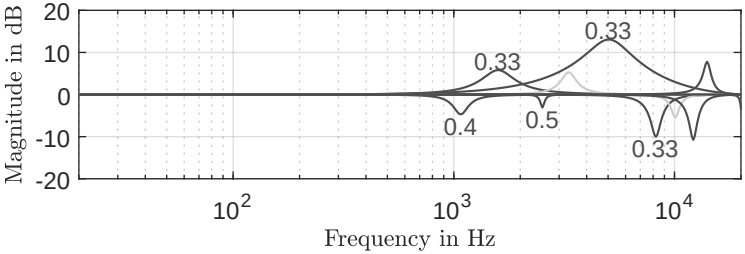

***(a)*** *1st neighbor* ($\varphi_1 = -15°$, $\theta_1 = 0°$, *closest direction*)




***(b)*** *2nd neighbor* ($\varphi_2 = 0°$, $\theta_2 = 0°$)




***(c)*** *3rd neighbor* ($\varphi_3 = 0°$, $\theta_3 = 11.25°$)




***(d)*** *4th neighbor* ($\varphi_4 = -15°$, $\theta_4 = 11.25°$)

**Figure 4.8:** Assignment of the peak filters of the neighboring directions to the ones of the closest direction for the exemplary interpolation ($\varphi = -10°$, $\theta = 5.625°$) from Fig. 4.7. Here, light gray peak filters are unassigned. Additionally, the weights for interpolating the first five peak filters are given.

the interpolation due to it's distance in frequency that is higher than the given threshold $\Delta f_c = 2\,\text{dB}$ from Eq. (4.11). Increasing the threshold would include this peak filter in the interpolation. Nevertheless, with the given threshold, the different neighboring directions contribute to at least five peak filters of the intermediate direction.

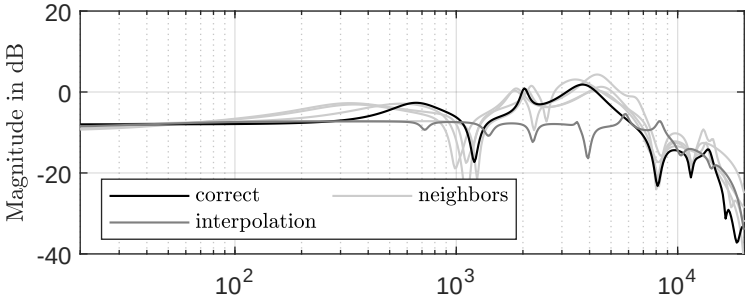
Figures 4.9 and 4.10 show further exemplary interpolated magnitude responses using simple and extended parameter interpolation. Additionally, the parameter interpolation is compared to bilinear rectangular interpolation for minimum-phase approximations of given HRIRs.

In Fig. 4.9, the interpolation is performed for the left ear of *Subject_065* and a contralateral sound incidence ($\varphi = 40°$, $\theta = 5.625°$). The slight elevation of the sound source leads to an interpolation of all four neighboring directions. As can be seen, both the extended parameter interpolation (see Fig. 4.9(b)) and the minimum-phase interpolation (see Fig. 4.9(c)) show reasonable results, whereas the simple parameter interpolation in Fig. 4.9(a) suffers from a wrong allocation of the peak filters. In comparison to the magnitude response by minimum-phase interpolation which stays in between of the magnitude responses of the neighboring directions throughout the entire frequency range, the magnitude response through extended parameter interpolation clearly follows the magnitude response of the closest direction for some frequency ranges, e.g. frequencies below 1 kHz and around 14 kHz. This effect arises from peak filters that are only existent in the closest direction. The absent boost of frequencies below 1 kHz can be explained by the missing peak filter at these frequencies in the approximation for the closest direction. Nevertheless, the extended parameter interpolation in Fig. 4.9(b) shows a proper result that reaches a similar accuracy as the minimum-phase interpolation in Fig. 4.9(c).

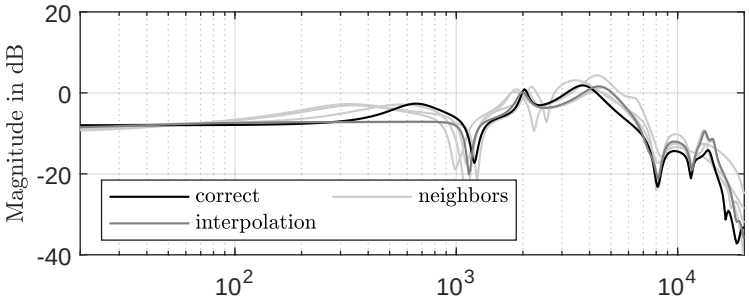
Similar results are visible in Fig. 4.10 for an elevated sound incidence ($\varphi = 5°$, $\theta = 28.125°$) at the left ear of the same subject, so that it can be concluded that the extended parameter interpolation is a proper method for interpolating parametric filter cascades. Note that the correct magnitude response does not always stay in between of the magnitude responses of the neighboring directions, e.g. the notch at 13 kHz in Fig. 4.10(c). In these cases, all interpolation methods are unable to reproduce the correct magnitude response.

### 4.2.2 Moving Virtual Sound Sources

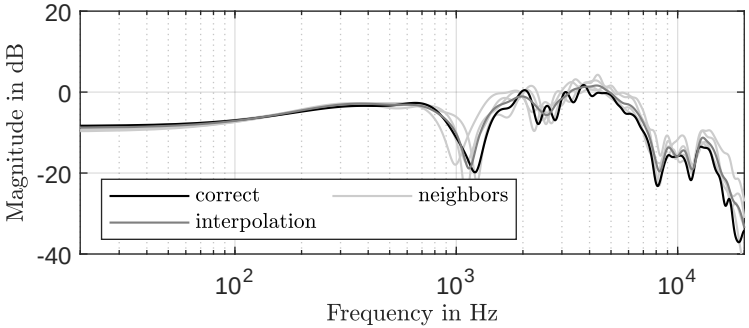
In order to perform dynamic binaural synthesis, the HRTFs used to filter the monaural audio signal have to change with the movement of the virtual sound source. As described in Section 4.1.4, generating these moving virtual sound sources without noticeable artifacts requires smooth transitions between the used directional filters. Thus, even for measurement grids of high spatial resolution, interpolation is used to smoothen the transition




***(a)*** *Simple parameter interpolation*

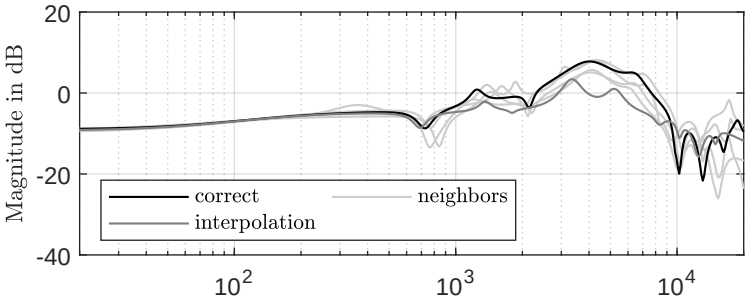



***(b)*** *Extended parameter interpolation*

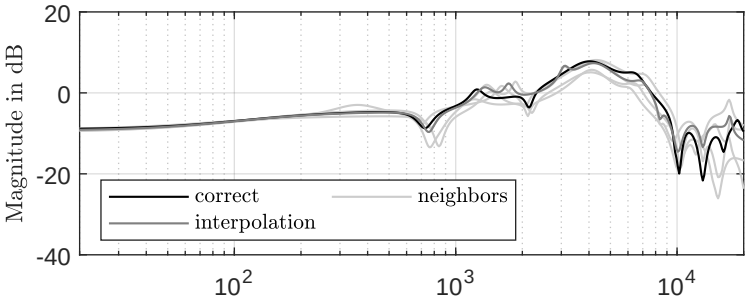



***(c)*** *Minimum-phase interpolation*

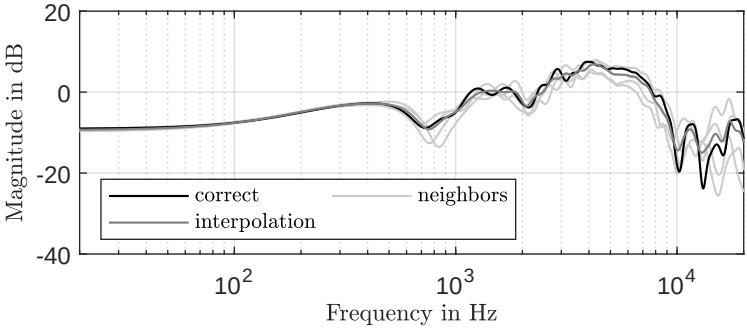
**Figure 4.9:** Comparison of the intermediate magnitude responses calculated through (a) simple parameter interpolation, (b) extended parameter interpolation, and (c) minimum-phase interpolation for an exemplary contralateral direction (*Subject_065*, left ear, $\varphi = 40^\circ$, $\theta = 5.625^\circ$).




***(a)** Simple parameter interpolation*

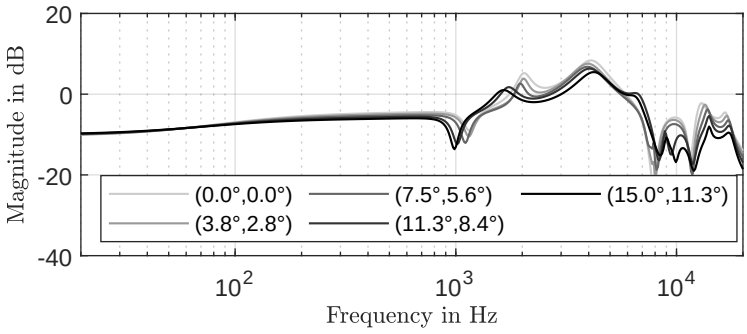



***(b)** Extended parameter interpolation*

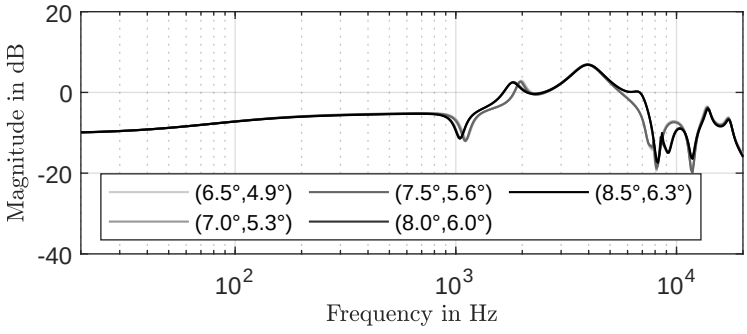



***(c)** Minimum-phase interpolation*

**Figure 4.10:** Comparison of the intermediate magnitude responses calculated through (a) simple parameter interpolation, (b) extended parameter interpolation, and (c) minimum-phase interpolation for an exemplary elevated direction (*Subject_065*, left ear, $\varphi = 5°$, $\vartheta = 28.125°$).




***(a)** Fast movement*




***(b)** Slow movement*

**Figure 4.11:** Evolution of the interpolated magnitude responses for generating moving virtual sound sources across (a) a whole interpolation rectangle inside the measurement grid ($\Delta\varphi = 15^\circ$, $\Delta\theta = 11.25^\circ$) and (b) a lower spatial distance on the same diagonal. In both cases, the movement contains five directions.

between these filters by introducing intermediate directions.

Figure 4.11 illustrates the evolution of the interpolated magnitude response via extended parameter interpolation for two diagonal movements in azimuth and elevation. In Fig. 4.11(a), the virtual sound source moves from the front ($\varphi = 0^\circ$, $\theta = 0^\circ$) to an azimuth of $\varphi = 15^\circ$ and an elevation of $\theta = 11.25^\circ$, meaning that the movement is carried out across the diagonal of a whole rectangle inside the measurement grid ($\Delta\varphi = 15^\circ$, $\Delta\theta = 11.25^\circ$). In this way, the directions at start and end of the movement are originally measured directions that do not have to be interpolated. Additionally,

the interpolated magnitude responses of three equally spaced intermediate directions are tracked during the movement. As can be seen, the magnitude response shows smooth transitions for some of the frequency regions, e.g. for frequencies around 4 kHz or above 12 kHz, whereas other frequency regions, like the region around $f = 10\,\text{kHz}$, suffer from strong deviations in the magnitude response. These differences arise between the third ($\varphi = 7.5°$, $\theta = 5.625°$) and the fourth position ($\varphi = 11.25°$, $\theta = 8.4375°$) of the movement, which can be explained by the change in the closest direction after crossing the center of the rectangle ($\varphi = \Delta\varphi/2 = 7.5°$, $\theta = \Delta\theta/2 = 5.625°$). Peak filters that are not included in the new closest direction will disappear in the magnitude response. In the current implementation, the center position inside a rectangle is assigned to the bottom left corner of this rectangle.

In order to emphasize this problem, Fig. 4.11(b) shows a movement across a lower spatial distance on the same diagonal. As can be seen, the frequency regions mentioned having smooth transitions, show indistinguishable magnitude responses when reducing the spatial distance between the interpolated directions. However, although the directions are closer together, the same deviations in the magnitude responses can be seen for the other frequencies after crossing the center of the rectangle ($\varphi = 7.5°$, $\theta = 5.625°$). Due to the described change in reference direction during the interpolation process, these deviations would stay regardless of how small the angular step size of the movement gets.

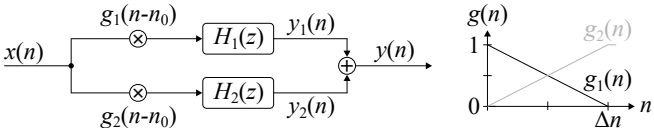



**Figure 4.12:** Improving the input-switching method from Fig. 4.3 by cross-fading the input during the transition from filter $H_1(z)$ to $H_2(z)$.

As described in Section 4.1.4, generating moving virtual sound sources without noticeable artifacts requires smooth transitions between the used directional filters. Especially time-variant IIR filter implementations suffer from audible clicks due to a mismatch between new coefficients and internal states of the recursive parts. Thus, the moving virtual sound sources generated through interpolated intermediate directions seen in Fig. 4.11 need further smoothing. For this, the cross-fading or input-switching method that are shown in Figs. 4.2 and 4.3, respectively, can be used. In this work, a combination of both methods is used that cross-fades the input signal while changing from one filter to another as shown in Fig. 4.12 [Nowak and Zölzer, 2022]. Combining these two methods merges the advantages of both methods by having a smooth switching between filters like in the

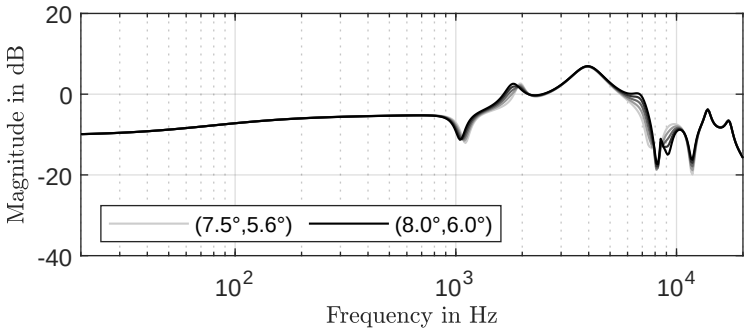



**Figure 4.13:** Using cross-fading of the impulse responses according to Fig. 4.12 in order to smoothen the transition of the interpolated magnitude responses from the third ($\varphi = 7.5°$, $\theta = 5.625°$) to the fourth direction ($\varphi = 8°$, $\theta = 5.9875°$) from Fig. 4.11(b). Here, the cross-fading is performed across $\Delta n = 4$ samples.

cross-fading and including the decay of the previous filter in the overall output like in the input-switching method.

In Fig. 4.13, the transition between the third ($\varphi = 7.5°$, $\theta = 5.625°$) and fourth direction ($\varphi = 8°$, $\theta = 5.9875°$) from Fig. 4.11(b) is smoothened by cross-fading the impulse responses of the two directions. Here, the cross-fading is performed across $\Delta n = 4$ samples, resulting in fade factors of $g_1 = \{(1, 0.75, 0.5, 0.25, 0\}$ for the start direction ($\varphi = 7.5°$, $\theta = 5.625°$) and $g_2 = \{0, 0.25, 0.5, 0.75, 1\}$ for the end direction ($\varphi = 8°$, $\theta = 5.9875°$), with $n \in \{0, 1, \ldots, 4\}$. By cross-fading the impulse responses of the two directions, the instantaneous impulse responses of the parallel structure in Fig. 4.12 are calculated. Evaluating the corresponding magnitude responses in Fig. 4.13 clearly indicates the smooth transition between the two directions. Thus, the cross-faded input-switching method shown in Fig. 4.12 can be used to switch the directional filters while generating moving virtual sound sources.

Although the cross-faded input-switching method solves the problem of audible artifacts during switching between two different IIR filters, peak filters contained only in a single neighboring direction can lead to strong changes in the magnitude response, which results in audible coloration during the movement across the center between two measured directions. In order to tackle this issue, a further smoothing of the transition is achieved by normalizing the interpolation weight $c_{i_{\mathrm{ref}},\mathrm{p}}$ for the gain $G_{\mathrm{p},i_{\mathrm{ref}}}$ of the $p^{\mathrm{th}}$ peak filter of the closest neighboring direction $i_{\mathrm{ref}}$ between 0 and 1, if only this direction contains a peak filter in the given frequency region [Nowak

and Zölzer, 2022]. For an interpolation in the horizontal plane with only two neighboring directions, this normalization can be given as

$$\tilde{c}_{i_{\mathrm{ref}},\mathrm{p}} = \frac{2c_{i_{\mathrm{ref}},\mathrm{p}} - 1}{c_{i_{\mathrm{ref}},\mathrm{p}}}. \tag{4.17}$$

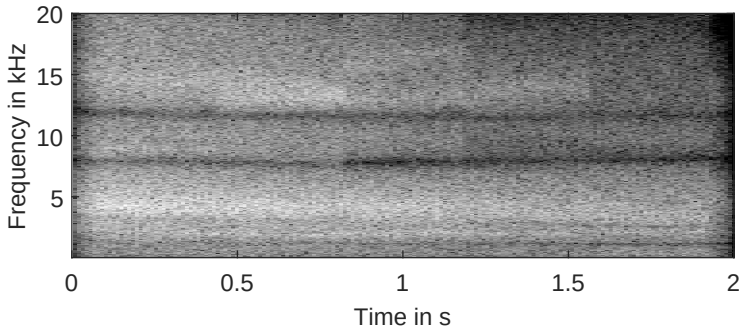
In this way, a peak filter disappears when moving the desired direction to the center between two measured directions where only the reference direction contains this peak filter. Note that only the gain $G_{\mathrm{p},i_{\mathrm{ref}}}$ of the $p^{\mathrm{th}}$ peak filter is interpolated using $\tilde{c}_{i_{\mathrm{ref}},\mathrm{p}}$, whereas the cut-off frequency $f_{c,\mathrm{p},i_{\mathrm{ref}}}$ and the Q-factor $Q_{\mathrm{p},i_{\mathrm{ref}}}$ still use $c_{i_{\mathrm{ref}},\mathrm{p}}$ in order to do not move the peak filter by decreasing these parameters while moving the direction closer to the center. Although the normalization of the gain prevents the magnitude response of the HRTF from changing strongly after passing the center between two measured directions, the extinction of the peak filter can lead to strong attenuation in the magnitude response around the center direction. Thus, this normalization is only recommended for moving virtual sound sources.

Figure 4.14 shows the improvement in smoothness of transition due to the proposed method. Here, a white noise signal and the HRIRs of *Subject_065* from the CIPIC database [Algazi et al., 2001b] are used to generate a moving virtual sound source from $\varphi = -40°$ to $\varphi = 40°$ in front of the subject. The spectrogram of the generated moving virtual sound source via cross-faded input-switching in Fig. 4.14(a) suffers from vertical edges representing strong changes in the magnitude response due to appearing and disappearing peak filters after crossing the center positions of the interpolation areas. When using the proposed smoothed cross-faded input-switching method these edges can be prevented (see Fig. 4.14(b)), indicating the improved transition.

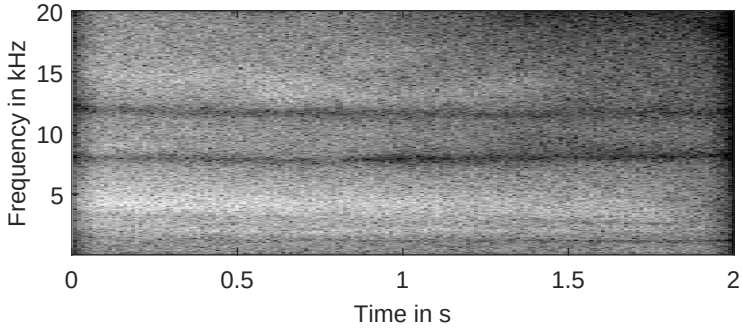
## 4.3 Summary

In binaural synthesis through headphones, the interpolation of measured HRTFs is required for enhancing the spatial resolution of the measurement grid as well as enabling smooth transitions between directional filters used to generate moving virtual sound sources. For FIR filter implementations, bilinear rectangular or triangular interpolation can be used to calculate the HRIRs of intermediate directions from minimum-phase HRIRs of the neighboring directions. Contrarily, interpolating IIR filters needs further restrictions that have to be taken into account, e.g. stability of the interpolated filters.

By using parametric IIR filter approximations of measured HRTFs, an interpolation of the intermediate magnitude responses can be achieved by interpolating the parameters of the neighboring directions. In this




***(a)*** *cross-faded input-switching*




***(b)*** *smoothed cross-faded input-switching*

**Figure 4.14:** Spectrograms of a moving virtual sound source generating a white noise signal while moving in front of *Subject_065* from the CIPIC database ($-40^\circ \rightarrow 40^\circ$) implemented via (a) the cross-faded input-switching method and (b) the smoothed cross-faded input-switching method.

way, the stability of the second-order peak filters is implicitly guaranteed. Although, simple bilinear rectangular interpolation is able to calculate intermediate magnitude responses, an assignment of the peak filters of the neighboring directions is required in order to generate more accurate interpolation results. Therefore, an extended parameter interpolation algorithm is proposed that uses the peak filters of the closest direction as reference for which related peak filters are found in the other neighboring directions. This assignment strongly increases the accuracy of the parameter interpolation.

In order to generate moving virtual sound sources, smooth transitions

between the magnitude responses of the intermediate directions are needed. Since especially time-variant IIR filter implementations suffer from audible clicks due to the mismatch between the updated coefficients and the internal states of the recursive parts, these implementations use parallel structures of IIR filters connected through cross-fading or input-switching. In this work, a combination of these two methods has shown to smooth the transitions between the interpolated magnitude responses, especially for cases in which the change of the reference direction leads to strong deviations in magnitude response. A normalization of the interpolation weight for the gain has shown a further improvement of the transition by reducing the audibility of coloration while moving across the center inside an interpolation area.

CHAPTER 5

# Externalization

Externalization of virtual sound sources is one of the major challenges in 3D spatial audio through headphones (see Section 2.2.4). This chapter deepens the definition of externalization and the influence of room effects inside measured BRIRs on perceived externalization during binaural synthesis through headphones. The room characteristics are given as the most significant factor for externalized virtual sound sources [Begault et al., 2001].

In literature about externalization of virtual sound sources during headphone playback, two different definitions are given for this term. Externalization can be interpreted either as a virtual sound source that is indistinguishable from a real sound source or as a virtual sound source that is perceived outside of the head [Hartmann and Wittenberg, 1996]. In the following, the latter definition is used for the term externalization [Durlach et al., 1992, Völk, 2009]. Here, different levels of externalization can be discriminated as shown in Fig. 5.1. These levels are in-head-, on-head- and out-of-head-localization. Virtual sound sources that are perceived inside the head are also referred to as internalized rather than externalized [Durlach et al., 1992]. The range for this in-head-localization reaches from the midpoint to the boundary surface of the human head [Völk, 2009]. Since perceived externalization of virtual sound sources through headphones can be linked to distance perception of human beings explained in Section 2.1.3, in-head-localization can be seen as a perceived distance of zero. In comparison to internalized virtual sound sources, externalized virtual sound sources can appear on-head, just outside of the head, or far outside of the head [Durlach et al., 1992], where the latter two are referred to as

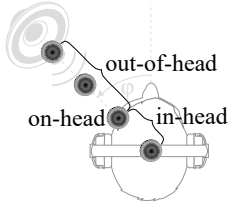



**Figure 5.1:** Levels of perceived externalization during localization of virtual sound sources through headphones. Here, in-head-, on-head-, and out-of-head-localization are discriminated.

out-of-head.

With stereo signals it is possible to move a virtual sound source on the interaural axis between the two ears, but the perceived sound stays inside the head [Durlach et al., 1992]. This process is called lateralization [Plenge, 1974]. In 3D spatial audio, the goal is to have an externalized virtual sound source despite positioning of a virtual sound source in full 3D space. This is called localization rather than lateralization. Although binaural synthesis is able to reproduce a virtual sound source at any position in 3D space, the reproduction often lacks of a successful externalization.

In the following, firstly, the characteristics of room effects and their influence on externalization in binaural synthesis through headphones are described. Afterwards, methods that simulate these room effects are explained, at which the focus is set on ISM. Then, the ISM is implemented using DRRs taken from real BRIR measurements. Finally, the method of adding simulated room effects is summarized.

## 5.1 Room Effects

The basics of room acoustics are given in [Kuttruff, 2000]. When a sound wave impinges at a wall or at the surface of any other object inside a room, a reflection occurs at the given surface. During this reflection, the amplitude and phase of the sound wave are changed by the frequency- and direction-dependent reflection coefficient $\beta$. Here, a part of the energy is lost during the reflection according to the absorption coefficient $\alpha$ of the wall, whose relation to the reflection coefficient $\beta$ is given by

$$\alpha = 1 - |\beta|^2. \tag{5.1}$$

Since the law of reflection that is known from optics is also valid for the reflection of sound waves, the angle of incidence equals the angle of reflection [Kuttruff, 2000]. However, not only reflections occur inside the

room. At the edges of the objects and the walls, also scattering of the impinging sound wave takes place. In this way, the energy of the sound wave will be spread in all directions. This scattering takes place for all objects that are not very small in comparison to the wavelength of the impinging sound.


**Figure 5.2:** Simplified representation of a typical room impulse response in order to show the characteristic components, namely direct signal, early reflections, and subsequent reverberation.

In digital signal processing, the RIR is used to describe the acoustical path between two points inside rooms [Zölzer, 2008]. In Fig. 5.2, the characteristic components of the RIR are shown in a simplified representation. As can be seen, not only the direct signal emitted by the sound source arrives at the receiver but also the reflected signals whose density increases with time. After the early reflections where the individual reflections are still separable, the density increases so much that a separation is no longer possible, resulting in an exponentially decaying tail without information about individual reflections called subsequent reverberation. Generally, an RIR can be split into

$$h_{\mathrm{rir}}(n) = h_{\mathrm{rir,d}}(n) + h_{\mathrm{rir,r}}(n), \tag{5.2}$$

where $h_{\mathrm{rir,d}}(n)$ and $h_{\mathrm{rir,r}}(n)$ describe the impulse responses of direct and reverberant path, respectively. In this way, the DRR can be calculated to

$$\mathrm{DRR} = \frac{\sum_{n=0}^{N-1} |h_{\mathrm{rir,d}}(n)|^2}{\sum_{n=0}^{N-1} |h_{\mathrm{rir,r}}(n)|^2}, \tag{5.3}$$

where $N$ defines the length of the RIR in samples. Usually, the DRR is given in decibels as

$$\mathrm{DRR}_{\mathrm{dB}} = 10 \log_{10} \mathrm{DRR} \tag{5.4}$$

rather than in linear scale. Additionally, one of the most important values for describing the acoustical characteristics of a room is the reverberation time $T_{60}$, which gives the time it takes for the sound pressure level to decrease by 60 dB. According to [Sabine, 1922], the reverberation time can

be approximated by

$$T_{60} = 0.163 \, \frac{\mathrm{s}}{\mathrm{m}} \cdot \frac{V}{\sum_{l=1}^{L} \alpha_l S_l}, \tag{5.5}$$

where $V$ defines the volume of the room in $\mathrm{m}^3$, $S_l$ the area of the walls in $\mathrm{m}^2$, $\alpha_l$ the absorption coefficient of the corresponding area $S_l$, and $L$ the total number of walls. Moreover, the decrease of acoustic energy with time inside a given room can be described by the energy decay curve (EDC), which is defined as

$$\mathrm{EDC}(t) = 10 \log_{10} \frac{\int_t^\infty h_{\mathrm{rir}}^2(\xi) d\xi}{\int_0^\infty h_{\mathrm{rir}}^2(\xi) d\xi}, \tag{5.6}$$

where $h_{\mathrm{rir}}(t)$ is the continuous-time RIR of the room [Lehmann and Johansson, 2008].

In contrast to RIRs, BRIRs are measured by using a dummy-head or a human head equipped with microphones at the ears instead of a single microphone. In this way, every reflection that arrives at the two ears underlies a filtering process with the HRTFs that correspond to the direction of arrival. Thus, every reflection visible in the BRIR contains the directional information. Hence, BRIRs record the full spatial information inside the given room at the position of the two ears.

### 5.1.1 Influence on Externalization

Reflections and reverberation inside a room have two important effects on human sound source localization [Durlach et al., 1992]. Firstly, the reverberant energy degrades the accuracy of the localized sound source direction due to additional sound waves reaching the human ear. However, the human auditory system is able to reduce this degradation by enhancing the perception of the direct sound signal and attenuating the perception of reflected sound signals. This phenomenon is called precedence effect [Durlach et al., 1992]. Secondly, the DRR gives an important cue for distance perception as described in Section 2.1.3. In comparison to using only the loudness of a sound signal arriving at the ear, which depends on distance as well as intensity of the original sound source, having reverberant energy increases the reliability and the objectivity of distance perception.

In [Durlach et al., 1992], the influence of using non-individual BRIRs as well as anechoic BRIRs is investigated. Individual BRIRs with room effects showed the best localization capabilities, including determination of the sound source direction and externalization. Although the usage of non-individual BRIRs strongly increased the front/back confusion rate, the influence on perceived externalization was small. Also BRIRs of the

KEMAR dummy-head with removed pinnae were able to deliver a perceived externalization during headphone playback when reflections and reverberation are included [Durlach et al., 1992]. Contrarily, using anechoic BRIRs substantially reduced the level of perceived externalization. However, adding artificial reverberation to the anechoic BRIRs is able to increase the externalization at the expense of a smeared azimuthal sound source localization [Durlach et al., 1992]. Similar effects of synthetic reverberation on externalization of virtual sound sources filtered with non-individual HRIRs were investigated by Begault in [Begault, 1992]. Although the increase of externalization was consistent across subjects, the absolute perceived distance varied between subjects, confirming the conclusion in Section 2.1.3 that human distance perception is more reliable for differential distances than for absolute ones.

The positive influence of reflections and reverberation inside measured BRIRs on the externalization of virtual sound sources was also shown by Völk et al. [Völk et al., 2008, Völk, 2009]. In [Völk, 2009], Völk showed that extending the length of a measured impulse response up to a duration of 100 ms increases the perceived externalization. Moreover, Völk mentioned in [Völk et al., 2008] that more detailed spectral information at the ears is needed, as contained in BRIRs taken from human heads instead of artificial heads, in order to further increase the externalization. However, it has to be mentioned that in this study, the Neumann KU80 dummy-head was used. In [Hartmann and Wittenberg, 1996], Hartmann and Wittenberg showed that reconstructing the ILD is not sufficient to create an externalized virtual sound source, thus correct spectra have to be reproduced at the two ears instead. Additionally, if the interaural phase difference (IPD) that is related to the ITD is set to zero, then a virtual sound source is perceived inside the head. However, if the optimum constant ITD across frequency and thus the optimum linear IPD is chosen, then the virtual sound source is externalized [Hartmann and Wittenberg, 1996].

In [Werner et al., 2016], a decrease of externalization level is seen when the listening room does not match the recorded or synthesized room, especially for frontal and rear sound sources. This phenomenon is called room divergence effect. Additionally, Leclère et al. [Leclère et al., 2019] criticized that listening tests are often performed with subjects facing a loudspeaker and asking to rate externalization relative to that loudspeaker, which would implicate that the virtual sound source has to be perceived not only outside of the head but also at the correct distance. Therefore, they evaluated the perceived externalization without visual cues and without a real sound source as reference for the perceived distance in [Leclère et al., 2019]. In the absence of this reference distance, non-individual BRIRs showed similar perceived externalization levels than individual BRIRs and HpEq did not improve the perceived externalization. Furthermore, lateral sound sources were more externalized than frontal sound sources. Reverberation

inside used BRIRs has only improved the perceived externalization when interaural differences were included.

Although human beings are able to perceive the distance of a real sound source also in free-field situations without reverberation, for virtual sound sources a lack of this reverberation, e.g. for anechoic BRIR measurements, tends to internalize the localization [Völk et al., 2008]. One reason for this internalization might be that having weak reverberation can also be explained by a close distance to a sound source, which is more probable under realistic conditions than being inside an anechoic room.

## 5.2 Room Simulation

Since stored HRIRs are usually only of short length, no room effects will be saved inside them, thus simulated room effects have to be added afterwards in order to improve the externalization of virtual sound sources. Different approaches exist that try to simulate or synthesize early reflections and/or reverberation inside a room. In the following, an overview of these approaches is given.

Since early reflections are an important characteristic of a room, it is important to reproduce them accurately. Therefore, different approaches for modeling these early reflections exist. Two of the most important approaches are the ray tracing model [Krokstad et al., 1968] and ISM [Allen and Berkley, 1979]. The ray tracing model [Krokstad et al., 1968] assumes a point source with a radial emission of the sound wave. Here, for every direction of emission, the path that is needed to reach the receiver is calculated. From this, the RIR can be determined based on the lengths of the individual paths and the absorption coefficients of the walls that were hit on these paths. Contrarily, the ISM [Allen and Berkley, 1979] begins with the mirroring of the original room at the walls in order to yield a number of image rooms with image sources. Afterwards, the RIR is estimated by the summation of the attenuations of all image sources at the corresponding delays. In Section 5.2.1, ISM is explained in more detail. Since the number of image sources increases strongly with the order of reflection, ISM is practically only used for low-order reflections [Välimäki et al., 2012]. Although the ray tracing model does not suffer from this problem, the modeling of higher-order reflections underlies systematic errors that reduce the accuracy of the modeling [Lehnert, 1993]. In addition to these two approaches, Gerzon [Gerzon, 1992] proposed to use a simple delay line for the first reflections. In this way, the delays of the individual reflections can be controlled together with an attenuation coefficient.

In [Välimäki et al., 2012, Välimäki et al., 2016], the history of artificial reverberation is summarized. The first digital reverberation algorithm was implemented by Schroeder and Logan [Schroeder and Logan, 1961, Schroeder,

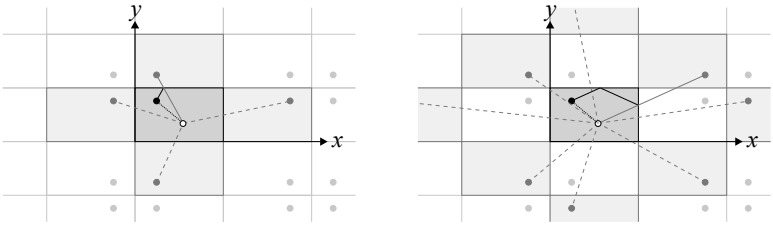
1962] in 1961. The algorithm uses a concatenation of parallel comb filters and cascaded all-pass filters in order to yield artificial reverberation. One advantage of all-pass filters over comb filters is the high echo density without spectral coloration. In contrast to Schroeder, Moorer [Moorer, 1979] proposed to use more parallel comb filters and only a single all-pass filter. Additionally, first-order low-pass filters are introduced in the feedback loop of the comb filters in order to yield a frequency-dependent reverberation time and a more natural sound. Based on these algorithms, further developments were made in order to improve the acoustic quality and the echo density of artificial reverberation [Gerzon, 1971, Stautner and Puckette, 1982, Jot and Chaigne, 1991]. One of these developments is the feedback delay network designed by Jot and Chaigne [Jot and Chaigne, 1991], where parallel comb filters are connected through a feedback matrix in order to increase the echo density. Until now, the feedback delay network is one of the state-of-the-art approaches for artificial reverberation and the choice of the feedback matrix is still under investigation by several researchers, e.g. in [Schlecht, 2020]. Other approaches use FIR filters with pseudo-random coefficients in a feedback loop together with a feedback coefficient or a first-order low-pass filter in order to produce a decay of the amplitude with time [Rubak and Johansen, 1998, Rubak and Johansen, 1999]. Furthermore, Karjalainen and Järveläinen [Karjalainen and Järveläinen, 2007] proposed to use velvet noise instead of pseudo-random coefficients in order to have a sparse sequence with a smoother sound. In [Välimäki et al., 2012], more approaches for acoustic room simulation are summarized including digital waveguide networks, time-varying reverb algorithms, and wave-based methods, which are out of the scope of this section.

Except for calculating the delays and the attenuation levels of reflections and reverberation, all of these models lack of a proper reproduction of interaural and monaural spectral cues at the human ears. In order to solve this issue, Kendall and Martens [Kendall and Martens, 1984] filtered the individual reflections with directional filters. Additionally, in [Jot et al., 1995], an approach for modeling BRIRs that preserves the interaural cues and introduces spectral cues based on average filters is proposed. Due to the existence of interaural and monaural spectral cues inside the room effects, these methods model BRIRs instead of RIRs.

### 5.2.1 Image Source Model

In 1979, Allen and Berkley [Allen and Berkley, 1979] proposed the ISM as a simulation method for early reflections of an RIR between two points inside a rectangular room that is simple, easy to use, and fast. As described before, the basic principle relies on mirroring of the original room at the walls. This can be seen in Fig. 5.3 for the example of a 2D representation and in Fig. 5.4 for the example of a 3D representation. In this way, image

sources are created in image rooms, too. For a better visualization, these image sources are only shown in the 2D representation.




**(a)** First-order **(b)** Second-order

**Figure 5.3:** 2D representation of the image rooms and image sources inside the ISM for (a) first- and (b) second-order reflections.

The order of a reflection is defined as the number of walls crossed by the path from the corresponding image source to the receiver. In Fig. 5.3(a), the image sources leading to first-order reflections are highlighted, whereas Fig. 5.3(b) emphasizes second-order reflections.

Figure 5.4 illustrates the image rooms up to the third order. For a better representation, the different cross sections in the $yz$-plane are stretched along $x$-axis. Here, six first-order reflections and 18 second-order reflections can be counted. For third and fourth order, already 38 and 66 different reflected paths exist in 3D space, respectively. In general, the number $n_\kappa$ of $\kappa^{\text{th}}$-order reflections in 3D space can be calculated to

$$n_\kappa = \begin{cases} 1 & \text{for } \kappa = 0 \\ 2 + \sum_{l=1}^{\kappa} 4 \cdot (2l-1) & \text{for } \kappa > 0 \end{cases}. \tag{5.7}$$

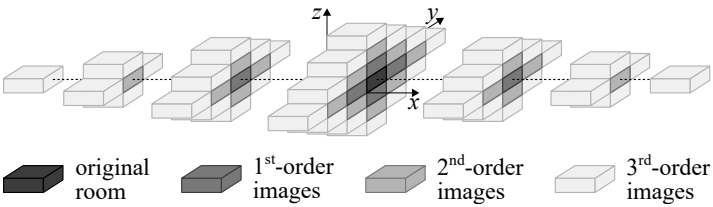



**Figure 5.4:** 3D representation of the image rooms inside the ISM up to third order. For a better representation, the different cross sections in the $yz$-plane are stretched along $x$-axis.

Accumulating the number of reflections per order from Eq. (5.7) up to the $\kappa^{\text{th}}$ order gives the total number of reflections

$$N_\kappa = \sum_{l=0}^{\kappa} n_l \tag{5.8}$$

having $\kappa^{\text{th}}$ order or lower. In Fig. 5.5, the number of reflections $n_\kappa$ per order and the total number of reflections $N_\kappa$ having $\kappa^{\text{th}}$ order or lower are shown for orders up to $\kappa = 10$. As can be seen, the total number of reflections $N_\kappa$ strongly increases with the reflection order $\kappa$. When simulating the early reflections up to the tenth order, 1560 reflected paths have to be taken into account.

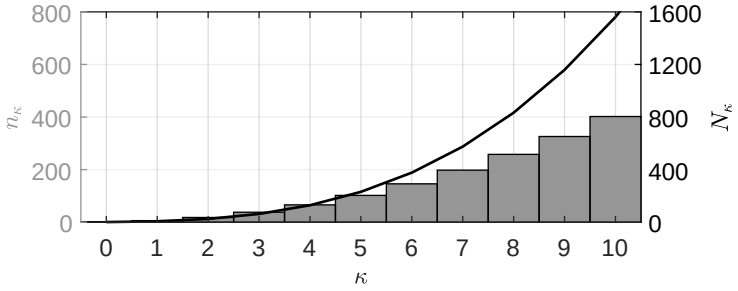



**Figure 5.5:** Histogram of the number of reflections $n_\kappa$ per order $\kappa$ together with the total accumulative number of reflections $N_\kappa$ having $\kappa^{\text{th}}$ order or lower.

Although the assumption that point image sources fulfill the boundary conditions of the walls may not be exact for non-rigid walls, ISM still uses this assumption in order to stay simple [Allen and Berkley, 1979]. Additionally, the reflection coefficient $\beta$ used in Eq. (5.1) is assumed to be frequency- and direction-independent. Using these assumptions, the time of arrival and the amplitude of a reflection can be calculated based on the distance of the corresponding image source to the receiver and the reflection coefficients of the walls crossed by this path. For this, the positions of the receiver (microphone) and the source (loudspeaker) are defined as

$$\mathbf{p}_\text{r} = \begin{bmatrix} x_\text{r} & y_\text{r} & z_\text{r} \end{bmatrix}^\text{T}, \tag{5.9}$$

$$\mathbf{p}_\text{s} = \begin{bmatrix} x_\text{s} & y_\text{s} & z_\text{s} \end{bmatrix}^\text{T}, \tag{5.10}$$

respectively, where $x_\text{r}$, $y_\text{r}$, $z_\text{r}$, $x_\text{s}$, $y_\text{s}$, and $z_\text{s}$ define the coordinates on the corresponding axes as shown in Fig. 5.6 for the $x$-axis. Here, one corner of the room is located in the origin of the coordinate system. The room

dimensions are given as

$$\mathbf{l}_{\text{room}} = \begin{bmatrix} L_x & L_y & L_z \end{bmatrix}^{\text{T}}. \tag{5.11}$$

Moreover, Fig. 5.6 illustrates the mirroring of the original room at the walls parallel to the $yz$-plane creating image sources along the $x$-axis. As can be seen, two parameters $u_x$ and $v_x$ can be used to identify the image source along the $x$-axis. Here, the variable $v_x$ defines the positions of the image walls at multiples of $2L_x$ and the variable $u_x$ defines whether an image source is on the left ($u_x = 1$) or the right ($u_x = 0$) side of these walls. In this way, the position of an image source on the $x$-axis can be calculated to

$$x_{\text{is}}(u_x, v_x) = -(2u_x - 1) \cdot x_{\text{s}} + v_x \cdot 2L_x. \tag{5.12}$$

The original room has the parameters $v_x = 0$ and $u_x = 0$.

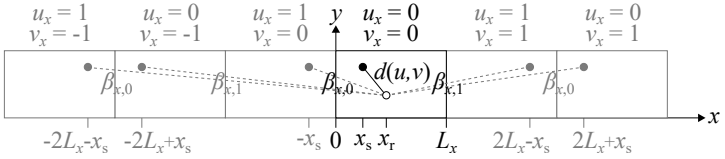



**Figure 5.6:** Parameter explanation for image rooms along the $x$-axis. Here, the variable $v_x$ defines the positions of the image walls at multiples of $2L_x$ and the variable $u_x$ defines the position of the image source relative to the corresponding wall. Additionally, $\beta_{x,0}$ and $\beta_{x,1}$ denote the reflection coefficients of the walls parallel to the $yz$-plane.

Expanding the ISM to all three directions yields image sources at

$$\begin{aligned} \mathbf{p}_{\text{is}}(\mathbf{u}, \mathbf{v}) = &- \operatorname{diag}(2u_x - 1, 2u_y - 1, 2u_z - 1) \cdot \mathbf{p}_{\text{s}} \\ &+ \operatorname{diag}(v_x, v_y, v_z) \cdot 2\mathbf{l}_{\text{room}}, \end{aligned} \tag{5.13}$$

where

$$\mathbf{u} = \begin{bmatrix} u_x & u_y & u_z \end{bmatrix}^{\text{T}} \quad \text{with} \quad u_x, u_y, u_z \in \{0, 1\}, \tag{5.14}$$

$$\mathbf{v} = \begin{bmatrix} v_x & v_y & v_z \end{bmatrix}^{\text{T}} \quad \text{with} \quad v_x, v_y, v_z \in \mathbb{N}, \tag{5.15}$$

and $\operatorname{diag}(\cdot)$ denotes a diagonal matrix with the arguments as diagonal elements [Lehmann and Johansson, 2008]. From this, the distance from an image source to the receiver can be calculated to

$$r(\mathbf{u}, \mathbf{v}) = \left\| \mathbf{p}_{\text{r}} - \mathbf{p}_{\text{is}}(\mathbf{u}, \mathbf{v}) \right\|, \tag{5.16}$$

with $\|\cdot\|$ being the Euclidean norm. This length of the straight path from the

image source to the receiver $r(\mathbf{u}, \mathbf{v})$ equals the length of the corresponding reflected path from the original source to the receiver. Thus, the distance from the image source to the receiver $r(\mathbf{u}, \mathbf{v})$ can be used to calculate the time of arrival

$$\tau(\mathbf{u}, \mathbf{v}) = \frac{r(\mathbf{u}, \mathbf{v})}{c} \tag{5.17}$$

of the reflected signal, where $c$ denotes the speed of sound [Lehmann and Johansson, 2008]. Additionally, the individual paths are attenuated by two factors. Firstly, the propagation loss due to the distance $r(\mathbf{u}, \mathbf{v})$ between an image source and the receiver and secondly, the reflection coefficients $\beta$ of the walls that are hit by the reflected path, resulting in an amplitude of

$$A(\mathbf{u}, \mathbf{v}) = \frac{\beta_{x,0}^{|v_x - u_x|} \beta_{x,1}^{|v_x|} \beta_{y,0}^{|v_y - u_y|} \beta_{y,1}^{|v_y|} \beta_{z,0}^{|v_z - u_z|} \beta_{z,1}^{|v_z|}}{4\pi r(\mathbf{u}, \mathbf{v})} \tag{5.18}$$

for a reflected path [Lehmann and Johansson, 2008], where $\beta_{x,0}$, $\beta_{x,1}$, $\beta_{y,0}$, $\beta_{y,1}$, $\beta_{z,0}$, and $\beta_{z,1}$ denote the reflection coefficients of the different walls inside the room. Here, the indices $x$, $y$, and $z$ define the reflection coefficients of a wall perpendicular to the $x$-, $y$, and $z$-axis, respectively, and the indices 0 and 1 discriminate the location of the two walls referenced by the first index. An index of 0 describes the walls containing the origin of the coordinate system and multiples by twice the corresponding room dimension, whereas an index of 1 describes the opposite walls. By summing up the exponents in the numerator of Eq. (5.18), the order of a reflected path can be calculated to

$$\kappa(\mathbf{u}, \mathbf{v}) = |v_x - u_x| + |v_x| + |v_y - u_y| + |v_y| + |v_z - u_z| + |v_z|. \tag{5.19}$$

The continuous-time RIR from source to receiver can be determined by adding the contributions of all image sources expanding infinitely in all three dimensions, thus the resulting RIR can be calculated as

$$h_{\mathrm{ism}}(t) = \sum_{\mathbf{u}=0}^{1} \sum_{\mathbf{v}=-\infty}^{\infty} A(\mathbf{u}, \mathbf{v}) \cdot \delta(t - \tau(\mathbf{u}, \mathbf{v})), \tag{5.20}$$

where $\tau(\mathbf{u}, \mathbf{v})$ defines the delay of a reflected path according to Eq. (5.17) and $A(\mathbf{u}, \mathbf{v})$ the amplitude factor according to Eq. (5.18). Here, the direct path is given by $\mathbf{u} = (0, 0, 0)^{\mathrm{T}}$ and $\mathbf{v} = (0, 0, 0)^{\mathrm{T}}$.

## 5.3 Implementation

In order to implement the ISM algorithm in the time-domain according to Eq. (5.20), the continuous-time RIR $h_{\mathrm{ism}}(t)$ has to be converted

into the discrete-time RIR $h_{\mathrm{ism}}(n)$ by the sampling operation. However, when performing the sampling operation, the unit impulses $\delta(t-\tau(\mathbf{u},\mathbf{v}))$ shifted to time stamps that differ from multiples of the sampling period $T_s$, $\tau(\mathbf{u},\mathbf{v}) \neq nT_s$, will disappear in the discrete-time RIR $h_{\mathrm{ism}}(n)$ [Lehmann and Johansson, 2008]. In order to tackle this problem, Allen and Berkley [Allen and Berkley, 1979] rounded the delays $\tau(\mathbf{u},\mathbf{v})$ of the individual reflections to the closest multiples of the sampling period ($mT_s$ with $m \in \mathbb{N}^+$). Although in this way the unit impulses of all reflections appear in the discrete-time RIR $h_{\mathrm{ism}}(n)$, the effective room characteristics are modified by rounding the distances of image sources. Therefore, Lehmann and Johansson [Lehmann and Johansson, 2008] proposed the implementation of the ISM approach in frequency-domain. Transforming Eq. (5.20) into frequency-domain by using the Fourier transform yields the frequency response

$$H_{\mathrm{ism}}(\omega) = \sum_{\mathbf{u}=0}^{1} \sum_{\mathbf{v}=-\infty}^{\infty} A(\mathbf{u},\mathbf{v}) \cdot e^{-j\omega\tau(\mathbf{u},\mathbf{v})} \tag{5.21}$$

of the room. The band-limitation to $f_s/2$ which is needed in front of the sampling operation converts the unit impulses $\delta(t-\tau(\mathbf{u},\mathbf{v}))$ from Eq. (5.20) into sinc-like pulses at the given delay, resulting in

$$h_{\mathrm{ism}}(n) = \sum_{\mathbf{u}=0}^{1} \sum_{\mathbf{v}=-\infty}^{\infty} A(\mathbf{u},\mathbf{v}) \cdot \mathrm{sinc}(n-\tau(\mathbf{u},\mathbf{v}) \cdot f_s), \tag{5.22}$$

with

$$\mathrm{sinc}(n) = \frac{\sin(\pi n)}{\pi n}. \tag{5.23}$$

In this way, also reflections at delays that differ from multiples of the sampling period can be taken into account by the sampling operation. Furthermore, when implementing the ISM, negative reflection coefficients

$$\beta = -\sqrt{1-\alpha} \tag{5.24}$$

should be used in order to generate reverberation tails that are similar to the characteristics of real acoustic measurements [António et al., 2002, Lehmann and Johansson, 2008]. Additionally, Lehmann et al. [Lehmann et al., 2007, Lehmann and Johansson, 2008] showed that absorption coefficients $\alpha$ that are calculated based on the reverberation time approximations defined by Sabine [Sabine, 1922] (see Eq. (5.5)), Eyring [Eyring, 1930], or others [Neubauer and Kostek, 2001] do not accurately reproduce the desired reverberation time, especially for non-uniform absorption coefficients $\alpha_l$ between different walls inside the simulated room. Thus, in [Lehmann

et al., 2007, Lehmann and Johansson, 2008], an approach for calculating the absorption coefficients $\alpha_l$ based on a closed-form expression of the EDC shown in Eq. (5.6) is used. Using this prediction, the EDC can be calculated for different sets of absorption coefficients $\alpha_l$ until a coefficient set is found that yields the desired reverberation time. In order to adapt the characteristics of the simulated room, the reverberation time $T_{60}$ and a weighting of the absorption coefficients

$$\mathbf{w}_\alpha = \begin{bmatrix} w_{x,0} & w_{x,1} & w_{y,0} & w_{y,1} & w_{z,0} & w_{z,1} \end{bmatrix} \tag{5.25}$$

have to be set. Here, the assignment of the weighting factors to the walls of the room is done in the same way as for the reflection coefficients $\beta_l$ in Eq. (5.18).

The implementation of the ISM algorithm is based on [Lehmann and Johansson, 2008]. However, some adjustments are made in order to combine the simulated RIRs with measured HRIRs. In the following, firstly, the parameters needed to implement the ISM are given. Then, the BRIR measurements are explained and the results are shown. Finally, these results are used to combine measured HRIRs with simulated RIRs.

### 5.3.1 Image Source Model

In order to implement the ISM based on [Lehmann and Johansson, 2008], firstly, the parameters of the simulated room have to be set. Here, the room dimensions defined in Eq. (5.11) are set to

$$\mathbf{l}_{\text{room}} = \begin{bmatrix} 4.2\,\text{m} & 4.8\,\text{m} & 2.0\,\text{m} \end{bmatrix}^{\text{T}}. \tag{5.26}$$

Additionally, the reverberation time is set to $T_{60} = 0.2\,\text{s}$ and the weights of the absorption coefficients for the individual walls as defined in Eq. (5.25) are given as

$$\mathbf{w}_\alpha = \begin{bmatrix} 0.75 & 0.75 & 0.75 & 0.75 & 1 & 0.75 \end{bmatrix}. \tag{5.27}$$

The decision of having a higher weighting for $w_{z,0}$ in comparison to the other five walls is made in order to simulate a carpeted floor and using the same materials for the other five walls. Having $w_{z,1} = 0.75$ instead of $w_{z,0} = 1$, means that 25 % less of the acoustic energy is absorbed by the ceiling than by the floor.

In addition to the parameterization of the room, also the positions of the loudspeakers and the microphones inside the room have to be defined (see Fig. 5.7). In order to simulate different RIRs for the two ears, two microphones are placed inside the room with a distance of $d_{\text{head}} = 0.2\,\text{m}$. Although this distance introduces a direction-dependent delay between the two microphones, no other effects like head shadowing or HRTF filtering

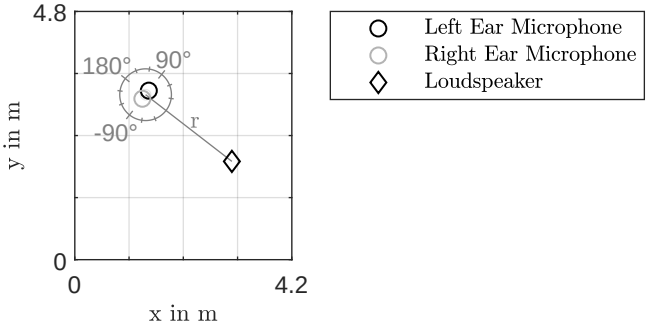


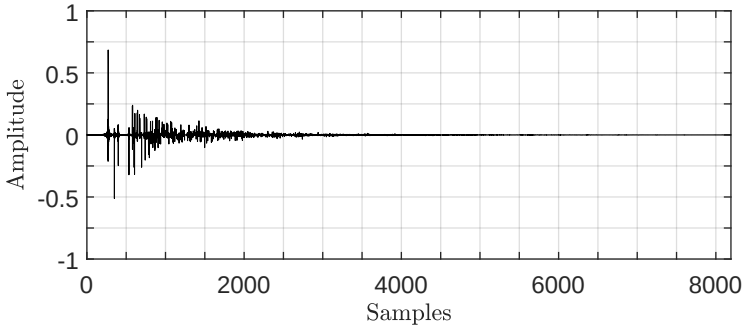

**Figure 5.7:** Positioning of the microphones and the loudspeaker in the $xy$-plane of the simulated room. The two microphones are separated by $d_{\text{head}} = 0.2\,\text{m}$. In order to simulate different azimuthal directions $\varphi$, the virtual head is rotated counter-clockwise in steps of $\Delta\varphi = 30^\circ$. Since all components are positioned at a height of $z = 1.14\,\text{m}$, the simulated elevation is fixed to $\theta = 0^\circ$.

are simulated. While adding these effects would simulate BRIRs which are needed for a complete reproduction of the signals at the human ears during natural hearing inside a room, filtering every impinging reflection with the corresponding HRTF would highly increase the computational complexity. Thus, spatially separating the two microphones is a trade-off between increasing the computational complexity and reproducing the effects of the human head on incoming reflections.

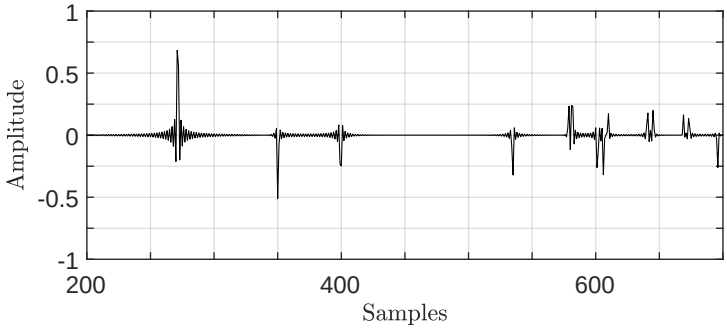
The position of the virtual head formed by the distance of $d_{\text{head}} = 0.2\,\text{m}$ between the two microphones and the loudspeaker can be seen in Fig. 5.7. In order to prevent symmetry of the RIRs between the two ears, the center of the virtual head and the loudspeaker are located on a line that is neither perpendicular to one of the walls nor directly on the diagonal of the room. Additionally, both the loudspeaker and the microphones are positioned at a height of $z = 1.14\,\text{m}$, resulting in an elevation of $\theta = 0^\circ$ for all simulated RIRs.

At first, the virtual head is oriented in the direction of the loudspeaker, thus simulating an azimuth of $\varphi = 0^\circ$. Afterwards, the positions of the microphones are rotated on a circle with a radius of $\frac{d_{\text{head}}}{2} = 0.1\,\text{m}$ around the center of the virtual head to simulate different azimuthal directions $\varphi$. In order to fulfill the clockwise definition of the azimuthal direction $\varphi$ of the loudspeaker relative to the orientation of the virtual head as seen in Fig. 1.1(b), the virtual head has to be rotated counter-clockwise. In this way, RIRs with an azimuthal resolution of $\Delta\varphi = 30^\circ$ are consecutively calculated via ISM as explained in Sections 5.2.1 and 5.3. For every azimuth

$\varphi$, two RIRs are simulated, one for the left ear microphone $h_{\mathrm{ism,L}}(n, \varphi, \theta)$ and one for the right ear microphone $h_{\mathrm{ism,R}}(n, \varphi, \theta)$. All simulations are performed in MATLAB.




***(a)** RIR of length $L_h = 8192$*




***(b)** Direct path and first reflections*

**Figure 5.8:** Simulated RIR at the left ear microphone $h_{\mathrm{ism,L}}(n, \varphi, \theta)$ for an azimuth of $\varphi = 0^\circ$ and an elevation of $\theta = 0^\circ$ (a) with a length of $L_h = 8192$ and (b) for a cutout of direct path and first reflections.

In Fig. 5.8, the simulated RIR for the left ear and an azimuthal direction of $\varphi = 0^\circ$ is shown. Here, Fig. 5.8(a) illustrates the RIR for a length of $L_h = 8192$, whereas Fig. 5.8(b) shows a cutout of the direct path and the first reflections. In Fig. 5.8(a), the characteristic components of a typical room impulse response as defined in Fig. 5.2 are clearly visible. Firstly, a strong impulse can be seen, representing the direct path. Note that the RIRs are normalized to a maximum absolute amplitude of 1 for the whole set of directions. Otherwise, distances of around $r \approx 2.11\,\mathrm{m}$ between

loudspeaker and microphone would lead to amplitudes of approximately 0.038 (see Eq. (5.18)). Secondly, the early reflections show impulses at different delays that are still separable. These individual reflections can be seen better in the cutout in Fig. 5.8(b). Here, the influence of the band-limitation on the shifted unit impulses (see Eq. (5.22)) is clearly visible, leading to sinc-like pulses. Finally, the increase of density of reflections leads to a smearing of the individual reflections, resulting in the subsequent reverberation.

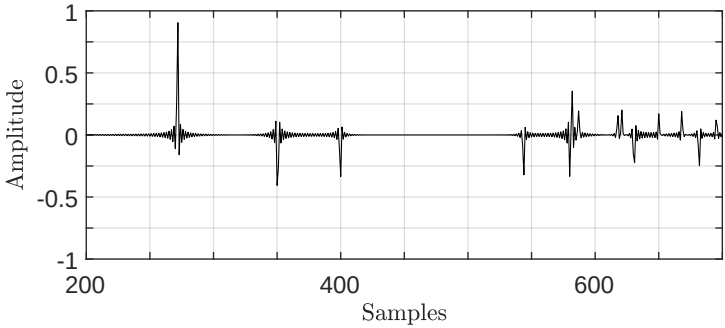
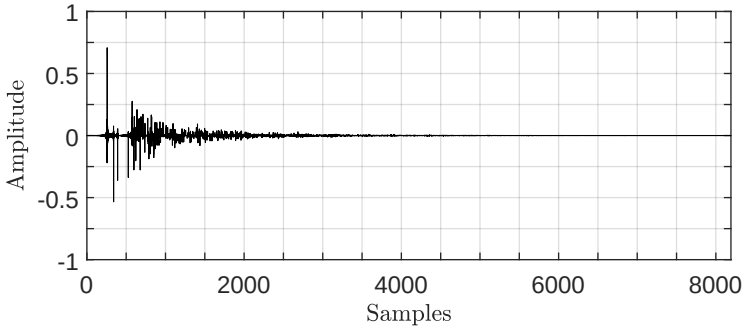


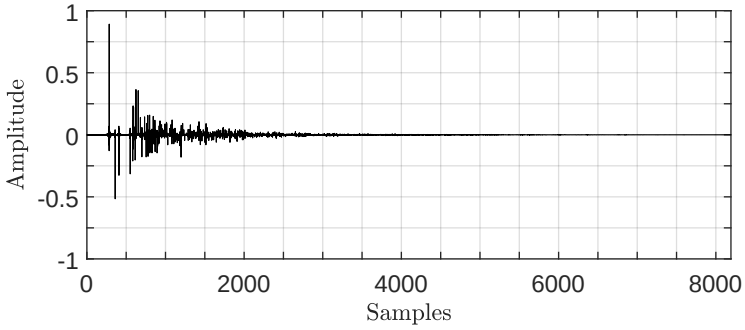

**Figure 5.9:** Cutout of the simulated RIR at the right ear microphone $h_{\mathrm{ism,R}}(n, \varphi, \theta)$ for an azimuth of $\varphi = 0^\circ$ and an elevation of $\theta = 0^\circ$.

The equality of the delay and the amplitude of the direct path at the positions of the left and right ear microphones in Figs. 5.8(b) and 5.9, respectively, is explained by the same distance of the two microphones to the loudspeaker for a frontal sound incident ($\varphi = 0^\circ$). The first two reflections do also have exactly the same delay and amplitude in both figures, confirming that these reflections are induced by the ceiling and the floor, which cause symmetric reflection paths for both ears. Due to the unsymmetrical placing of the microphones and the loudspeaker inside the room, the reflections from the walls introduce different delays and amplitudes for both ears.

Figure 5.10 shows the simulated RIRs for an ipsilateral ($\varphi = -90^\circ$) and a contralateral ($\varphi = 90^\circ$) sound incidence at the position of the left ear. Due to the distance of $d_{\mathrm{head}} = 0.2\,\mathrm{m}$ between the two ears, the time of arrival of the direct signal differs by approximately 26 samples between the ipsilateral and the contralateral sound incidence. Additionally, a difference in the amplitude of the direct signals can be seen. However, due to the lack of the head shadow effect, this difference is really small in comparison to the ILD that can be seen in Fig. 2.3(a). Moreover, Fig. 5.10 shows comparable amplitudes for the reflections and subsequent reverberation.




***(a)*** *ipsilateral* ($\varphi = -90°$)




***(b)*** *contralateral* ($\varphi = 90°$)

**Figure 5.10:** Simulated RIRs at the left ear microphone $h_{\text{ism,L}}(n, \varphi, \theta)$ for an elevation of $\theta = 0°$ and azimuthal directions of (a) $\varphi = -90°$ (ipsilateral sound incidence) and (b) $\varphi = 90°$ (contralateral sound incidence).

### 5.3.2 Measurement of Binaural Room Impulse Responses

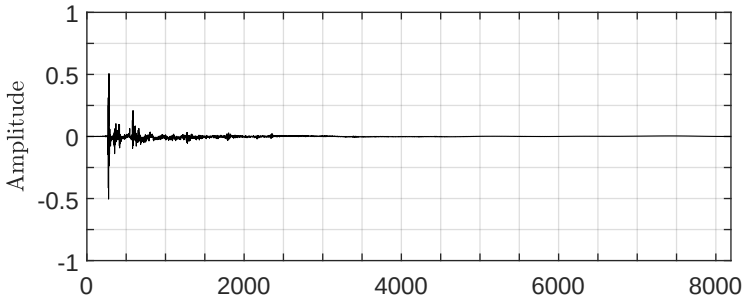
In addition to the simulation of RIRs, also measurements of BRIRs are performed. An overview of the measurement setup is presented in Fig. 5.11. The measurements are performed inside a room for audio listening in the Department of Signal Processing and Communication with the same dimensions ($4.2\,\text{m} \times 4.8\,\text{m} \times 2.0\,\text{m}$) as used for the RIR simulation in Section 5.3.1. Additionally, the positions of the KU100 dummy-head and the Genelec 8030C loudspeaker match the ones shown in Fig. 5.7. In order to generate the excitation signal and to record by means of the dummy-head at a sampling rate of $f_s = 44.1\,\text{kHz}$, an RME Fireface UCX audio interface is connected to the loudspeaker and the dummy-head.

**Figure 5.11:** Setup for measuring BRIRs using a Genelec 8030C loudspeaker and a Neumann KU100 dummy-head connected to an RME Fireface UCX audio interface inside a room for audio listening with dimensions 4.2 m × 4.8 m × 2.0 m.

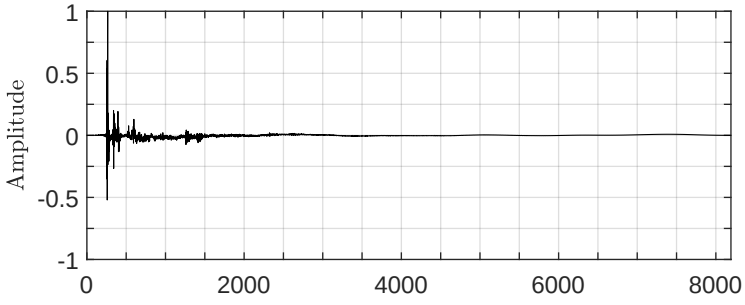
As excitation signal, an ESS in the frequency range between $f_{\text{start}} = 55\,\text{Hz}$ and $f_{\text{end}} = f_s/2$ is used [Farina, 2000, Farina, 2007]. The duration of the ESS is given as $T_{\text{sweep}} = 3\,\text{s}$. Additionally, fade-in and fade-out of the ESS are implemented with a length of $T_{\text{fade}} = 0.1\,\text{s}$ each, ensuring a smooth beginning and ending of the excitation signal. The whole measurement procedure is implemented in MATLAB, containing the generation of the ESS and the convolution of the recorded signal with the inverse ESS. Finally, scaling and delay that are introduced by the convolution of recorded and inverse ESS have to be reversed [Holters et al., 2009]. Both scaling and delay depend on the parameters of the ESS.

Besides the measurement of BRIRs from loudspeaker to dummy-head microphones, also output and input (O/I) of the audio interface are connected in order to identify the influence of the operation system and the audio interface on the measured BRIRs. Using the impulse response of this O/I connection and the least squares minimization method proposed in [Rivera Benois et al., 2016], cleaned BRIRs without an influence of the operation system and the audio interface can be calculated. Otherwise, operation system and audio interface would mainly introduce an additional delay to the measured BRIRs.

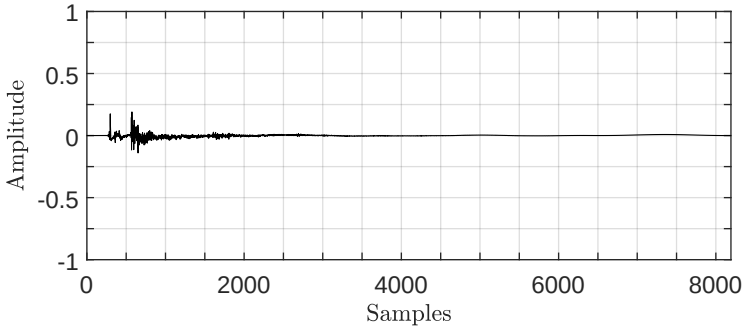
In this way, BRIRs are measured for both ears with a fixed elevation of $\theta = 0°$ and an azimuthal resolution of $\Delta\varphi = 30°$. In Fig. 5.12, measured BRIRs $h_{\text{brir,L}}(n, \varphi, \theta)$ of the left ear are shown for three directions including a frontal ($\varphi = -0°$), an ipsilateral ($\varphi = -90°$), and a contralateral ($\varphi = 90°$) sound incidence. Note that the BRIRs are scaled such that the maximum absolute amplitude between all directions is normalized to 1. As can be seen in Fig. 5.12(b), the highest amplitudes of the direct path are achieved




**(a)** *frontal* ($\varphi = 0°$)




**(b)** *ipsilateral* ($\varphi = -90°$)




**(c)** *contralateral* ($\varphi = 90°$)

**Figure 5.12:** Measured BRIRs at the left ear microphone of the Neumann KU100 dummy-head $h_{\text{brir,L}}(n, \varphi, \theta)$ for an elevation of $\theta = 0^\circ$ and azimuthal directions of (a) $\varphi = 0^\circ$ (frontal), (b) $\varphi = -90^\circ$ (ipsilateral), and (c) $\varphi = 90^\circ$ (contralateral).

for ipsilateral sound incidence. In contrast to this, contralateral directions strongly suffer from the head shadow effect, resulting in amplitudes of the direct path that are comparable to those of the strongest reflections (see Fig. 5.12(c)). For some contralateral directions, the reflections of the windows are even stronger than the direct path.

Additionally, the distance of $r = 2.11\,\mathrm{m}$ between the loudspeaker and the center of the dummy-head leads to delays of around 271 samples in the measured BRIRs. Because of the same setup inside the room, the delays are similar to the ones seen in Figs. 5.8(b) and 5.9. Due to the presence of the dummy-head in the measurements the delays for the contralateral directions are, however, slightly higher than the ones of the simulations.

Nevertheless, the biggest difference between simulated RIRs and measured BRIRs is the fact that in case of BRIR measurements, all reflections are filtered with the corresponding HRTFs of the direction of arrival.

### 5.3.3 Combination of Measured HRIRs and Simulated RIRs

Although the usage of long BRIRs during binaural synthesis ensures externalized virtual sound sources, storing these BRIRs for several directions requires a lot of memory. Thus, usually shorter HRIRs are used during binaural synthesis through headphones. The lack of room effects inside these HRIRs, however, results in a poor externalization. In order to ensure externalized sound sources even for short HRIRs, these HRIRs have to be enriched with simulated room effects. In Fig. 5.13, a method for combining measured HRIRs as described in Section 5.3.2 and simulated RIRs via ISM (see Section 5.3.1) is shown.

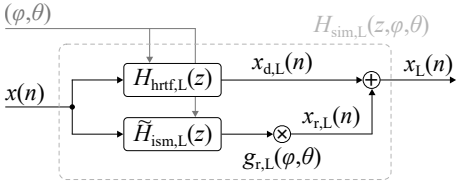
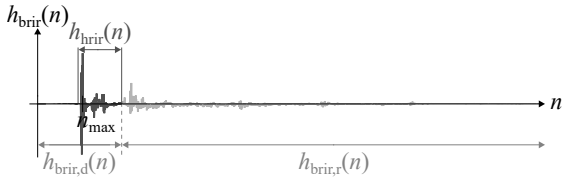


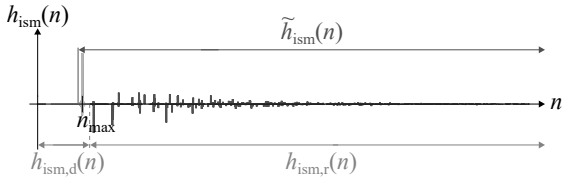

**Figure 5.13:** Combination of measured HRTFs $H_{\mathrm{hrtf,L}}(z)$ and modified simulated RIRs $\tilde{h}_{\mathrm{ism,L}}(n)$ of the left ear in a parallel structure. In order to achieve a desired DRR between direct signal $x_{\mathrm{d,L}}(n)$ and reflected signal $x_{\mathrm{r,L}}(n)$, a direction-dependent scaling factor $g_{\mathrm{r,L}}(\varphi, \theta)$ is added in the reflected path.

Firstly, the shorter HRIRs $h_{\mathrm{hrir}}(n, \varphi, \theta)$ have to be extracted from the BRIRs $h_{\mathrm{brir}}(n, \varphi, \theta)$ by separating the direct path. Here, a length of $L_h = 200$ samples is chosen for the HRIRs. The principle of the extraction procedure is shown in Fig. 5.14(a). In order to prevent a changing of the ITD, the left and right ear BRIRs have to be cut at the same position. Therefore,

the position of the maximum absolute amplitude $n_{\max}$ is evaluated for both impulse responses and the lower one of them is used to define the starting position of the HRIRs. Since a cutting at this position would erase the onset of the impulse response, the starting position is set to 20 samples before the first maximum. The cutting of the impulse response at this position just before $n_{\max}$ is visualized in Fig. 5.14(a), resulting in the extracted HRIR $h_{\mathrm{hrir}}(n, \varphi, \theta)$ of length $L_h = 200$ samples.




***(a)*** *Measured BRIR*




***(b)*** *Simulated RIR via ISM*

**Figure 5.14:** Principle of splitting (a) measured BRIRs $h_{\mathrm{brir}}(n)$ and (b) simulated RIRs $h_{\mathrm{ism}}(n)$ into direct and reverberant parts. Additionally, the extraction of HRIRs $h_{\mathrm{hrir}}(n)$ and modified simulated RIRs $\tilde{h}_{\mathrm{ism}}(n)$ is shown. Here, $n_{\max}$ defines the sample time index of the maximum absolute amplitude.

In Fig. 5.15, the HRIR extracted from the measured BRIR given in Fig. 5.12(a) is shown. A filtering of a monaural audio signal $x(n)$ with the shortened HRTFs of the two ears, $H_{\mathrm{hrtf,L}}(z)$ and $H_{\mathrm{hrtf,R}}(z)$, yields the signals $x_{\mathrm{d,L}}(n)$ and $x_{\mathrm{d,R}}(n)$ that contain the directional information of the virtual sound source.

For the purpose of including room effects in the output signal $x_{\mathrm{L}}(n)$ of Fig. 5.13, a parallel branch is added that filters the monaural audio signal $x(n)$ with the simulated RIRs. However, before using the RIRs from Section 5.3.1 in the parallel branch, these RIRs have to be pre-processed according to Fig. 5.14(b) in order to adapt to the shortened HRIRs. As mentioned in the previous sections, both the BRIRs and the simulated RIRs involve the delay of the direct path resulting from the distance between loudspeaker and dummy-head. When shortening the BRIR, this delay is too long to be considered. Therefore, the delay at the beginning of the RIR has to be trimmed in the same way as in the case of HRIR. This trimming

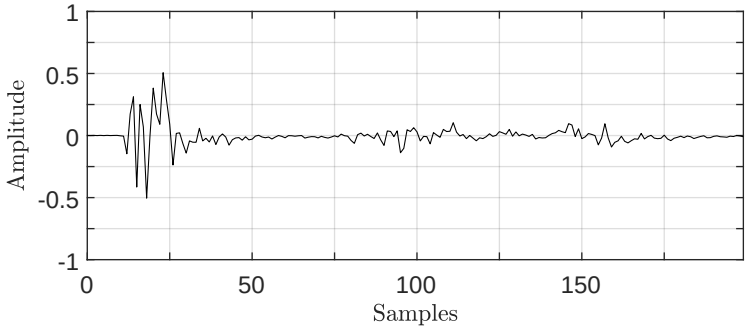



**Figure 5.15:** Measured HRIR at the left ear microphone of the Neumann KU100 dummy-head $h_{\mathrm{hrir,L}}(n, \varphi, \theta)$ for a frontal sound incidence ($\varphi = 0°$, $\theta = 0°$). The given HRIR coincides with the direct path of the BRIR from Fig. 5.12(a).

can be seen in Fig. 5.14(b). Additionally, the direct pulse has to be deleted from the RIR in order to maintain only the reflected paths in the second branch. In this way, the modified simulated RIR $\tilde{h}_{\mathrm{ism}}(n, \varphi, \theta)$ that can be seen in Fig. 5.14(b) is achieved. Due to the deletion of the direct pulse in the modified simulated RIR $\tilde{h}_{\mathrm{ism}}(n, \varphi, \theta)$, the modified simulated RIR equals a time-shifted version of the reverberant part of the simulated RIR $h_{\mathrm{ism,r}}(n, \varphi, \theta)$.

After pre-processing the RIR, the direction-dependent scaling factor $g_{\mathrm{r}}(\varphi, \theta)$ has to be tuned. This scaling factor targets an adjustment of the DRR for measured BRIRs and combined impulse response of measured HRIR and simulated RIR $h_{\mathrm{sim}}(n, \varphi, \theta)$ from Fig. 5.13. In order to leave the directional cues inside the direct signal unchanged, the scaling factor $g_{\mathrm{r}}(\varphi, \theta)$ is added in the reflected path. Thus, the combined impulse response for the left ear and a given direction ($\varphi$, $\theta$) can be calculated to

$$h_{\mathrm{sim,L}}(n, \varphi, \theta) = h_{\mathrm{hrir,L}}(n, \varphi, \theta) + g_{\mathrm{r,L}}(\varphi, \theta) \cdot \tilde{h}_{\mathrm{ism,L}}(n, \varphi, \theta), \qquad (5.28)$$

where $h_{\mathrm{hrir,L}}(n, \varphi, \theta)$ defines the shortened HRIR, $\tilde{h}_{\mathrm{ism,L}}(n, \varphi, \theta)$ the modified simulated RIR, and $g_{\mathrm{r,L}}(\varphi, \theta)$ the scaling factor for achieving a given DRR at the left ear.

For calculating the scaling factor, firstly, the DRR of the measured BRIRs from Section 5.3.2 is evaluated according to Eq. (5.3). Here, the direct path of the measured BRIR $h_{\mathrm{brir,d}}(n)$ is defined as the shortened HRIR $h_{\mathrm{hrir}}(n)$ including the delay of the direct path as can be seen in Fig. 5.14(a). In this way, the direct path of the measured BRIR $h_{\mathrm{brir,d}}(n)$ and the direct path of the combined impulse response $h_{\mathrm{sim}}(n)$ that is set

to $h_{\mathrm{hrir}}(n)$, coincide except for the delay of the direct path that contains negligible energy. Thus, the scaling factor for the left ear can be calculated based on the energy content of the reflected paths as

$$g_{\mathrm{r,L}}(\varphi,\theta) = \sqrt{\frac{\sum_{n=0}^{N-1} |h_{\mathrm{brir,r,L}}(n,\varphi,\theta)|^2}{\sum_{n=0}^{N-1} |\tilde{h}_{\mathrm{ism,L}}(n,\varphi,\theta)|^2}}, \tag{5.29}$$

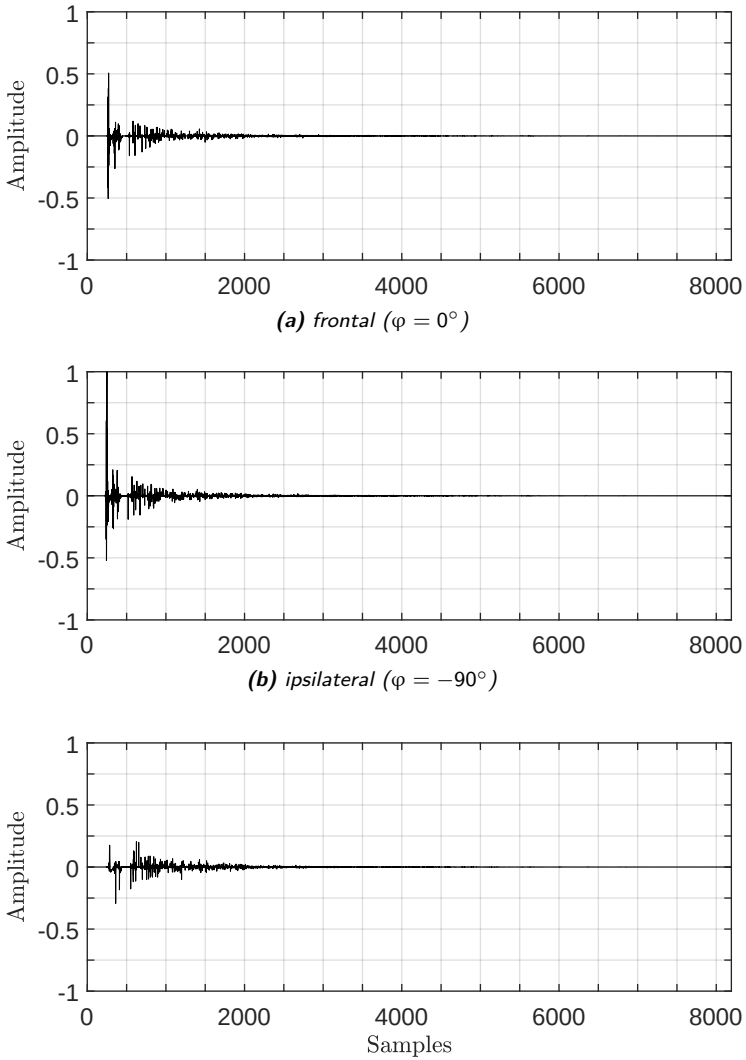
where $h_{\mathrm{brir,r,L}}(n,\varphi,\theta)$ and $\tilde{h}_{\mathrm{ism,L}}(n,\varphi,\theta)$ respectively define the reflected paths of the measured BRIRs and the combined impulse response of the left ear according to Figs. 5.14 and 5.13, and $N$ defines the length of the reflected path.

Using the given impulse responses of the left ear, $h_{\mathrm{hrir,L}}(n,\varphi,\theta)$ and $\tilde{h}_{\mathrm{ism,L}}(n,\varphi,\theta)$, and the scaling factor $g_{\mathrm{r,L}}(\varphi,\theta)$ in the parallel structure from Fig. 5.13 results in the combined impulse responses for the left ear $h_{\mathrm{sim,L}}(n,\varphi,\theta)$ shown in Fig. 5.16. Note that the same cases are plotted like in Fig. 5.12. For a better comparison, the delays of the direct paths inside the BRIRs are also added in front of the combined impulse responses. From Fig. 5.16, the combination of measured HRIRs as direct path and simulated RIRs as reflected path is clearly visible. Nevertheless, the first reflections of the simulated RIRs overlap with the tail of the measured HRIRs. By comparing the early reflections in Fig. 5.16 with the ones in Fig. 5.12, the absent filtering of the impinging reflections with HRIRs of the corresponding directions can be seen. Since this filtering spreads the energy of the reflections along time, the maximum amplitudes of the reflections in Fig. 5.12 are lower than the ones of the sinc-like pulses in Fig. 5.16. In this way, also the amplitude of the reverberation tail is higher than the one of the measured BRIRs. Another explanation of this effect might be that the used parameters for the reverberation time $T_{60}$ and the weighting of the absorption coefficients $\mathbf{w}_\alpha$ in the simulation do not match the real characteristics of the room. Note that the aim of the combined impulse response was to include room effects into measured HRIRs and reproduce the DRR of measured BRIRs rather than completely replicate the original room characteristics.

## 5.4 Summary

Nowadays, the externalization of virtual sound sources is one of the major challenges in 3D spatial audio through headphones. Here, virtual sound sources are specified as being externalized if the virtual sound source is perceived outside of the head. Room effects inside the used BRIRs, e.g. reflections and reverberation, are given as the most significant factors for a successful externalization of virtual sound sources.

Since HRIRs are usually only of short length, no room effects will be




**(a)** *frontal* ($\varphi = 0°$)

**(b)** *ipsilateral* ($\varphi = -90°$)

**(c)** *contralateral* ($\varphi = 90°$)

**Figure 5.16:** Combined impulse responses of the left ear $h_{\mathrm{sim,L}}(n, \varphi, \theta)$ for an elevation of $\theta = 0°$ and azimuthal directions of (a) $\varphi = 0°$ (frontal), (b) $\varphi = -90°$ (ipsilateral), and (c) $\varphi = 90°$ (contralateral). Note that the delays in front of the direct paths are added to the combined impulse responses in order to yield a better comparison to the measured BRIRs in Fig. 5.12.

contained inside of them. Thus, these HRIRs have to be enriched with simulated room effects in order to improve the externalization of the virtual sound sources. In literature, different approaches exist for simulating early reflections and/or reverberation inside a room. In this work, ISM is used to model the room effects.

Although ISM yields simulated reflections with proper delays and amplitudes, these reflections lack of a reproduction of directional information. While filtering the reflections with the corresponding HRTFs would simulate BRIRs which are needed for a complete reproduction of the signals at the human ears during natural hearing inside a room, filtering every impinging reflection with the corresponding HRTF would highly increase the computational complexity. Thus, spatially separating the microphones for the two ears by $d_{\mathrm{head}} = 0.2\,\mathrm{m}$ in the ISM algorithm is a trade-off to achieve different RIRs for them.

Finally, measured HRIRs and RIRs simulated via ISM are combined in a parallel structure. Here, the RIRs are scaled by an additional gain factor in order to yield similar DRRs inside the combined impulse responses than for the measured BRIRs from the same direction. In this way, a combination of directional information inside measured HRIRs and simulated room effects is ensured.

CHAPTER 6

# Evaluation

In the previous chapters, methods are proposed that enable the usage of parametric IIR filter approximations of given HRTFs in 3D spatial audio through headphones. The whole application that combines binaural synthesis, HpEq, and room simulation is shown in Fig. 6.1. The main part of the application is given by the binaural synthesis, in which the binaural signals containing the directional information are generated. Here, the filtering of the monaural audio signal $x(n)$ with the directional filters $H_{\mathrm{L}}(z)$ and $H_{\mathrm{R}}(z)$ is split into two steps. Firstly, the monaural audio signal $x(n)$ is filtered by the parametric IIR filter cascades that approximate the magnitude responses of the HRTFs for the desired direction, resulting in binaural audio signals containing the spectral information as wells as the ILD between the given HRTFs. Secondly, the ITD is introduced by delaying the binaural signal for the contralateral ear according to the extracted ITD.

Although these binaural signals deliver the important cues for virtual sound source localization, HpEq and room simulation can be added afterwards in order to improve the immersion of the virtual environment. HpEq cancels the spectral coloration introduced by headphone playback in order to generate natural sound scenes, whereas room simulation adds room effects needed for improving the externalization of static virtual sound sources.

In order to evaluate the present work, three listening tests are performed, which analyze different components of the application shown in Fig. 6.1. Firstly, the localization ability of parametric IIR filter cascades is compared to FIR filter representations of measured HRIRs both for individual and non-individual measurements. Secondly, the perceived externalization is

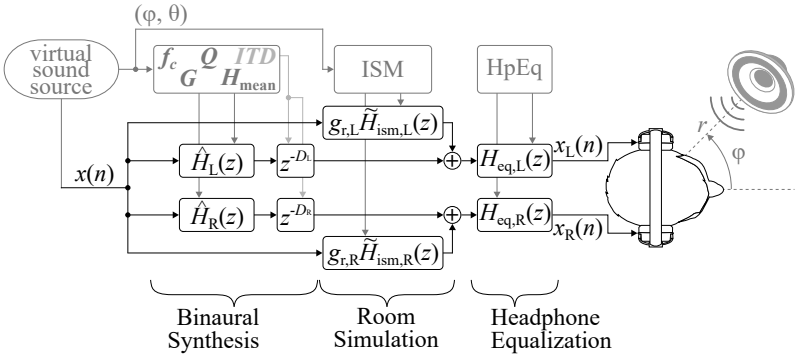



**Figure 6.1:** Binaural synthesis using HRTFs approximated by parametric IIR filters combined with a delay representing the ITD. Additionally, HpEq and room simulation are added afterwards.

evaluated for non-individual parametric IIR filter cascades equipped with simulated room effects and FIR filter representations of measured BRIRs including room effects of the measuring room. Thirdly, the audio quality of moving virtual sound sources generated by parametric IIR filter cascades and FIR filter implementations is rated.

For the evaluation of the first two listening tests, three different metrics are used, namely mean angular error $\bar{\varphi}_{\text{error}}$, front/back confusion rate $\rho_{\text{fb}}$, and average perceived externalization $\bar{d}_{\text{extern}}$. Firstly, the mean angular error is given as

$$\bar{\varphi}_{\text{error}} = \text{mean}(\boldsymbol{\varphi}_{\text{error}}), \tag{6.1}$$

$$\boldsymbol{\varphi}_{\text{error}} = |\boldsymbol{\varphi}_{0,\text{original}} - \boldsymbol{\varphi}_{0,\text{perceived}}|. \tag{6.2}$$

Here, $\boldsymbol{\varphi}_{0,\text{original}}$ and $\boldsymbol{\varphi}_{0,\text{perceived}}$ are respectively calculated from the vectors of original azimuthal direction $\boldsymbol{\varphi}_{\text{original}}$ and perceived azimuthal direction $\boldsymbol{\varphi}_{\text{perceived}}$ by projecting them element by element to the frontal plane according to

$$\varphi_{0,\text{original}}(l) = \begin{cases} \varphi_{\text{original}}(l) & \text{for} \quad |\varphi_{\text{original}}(l)| \leq 90^\circ \\ \text{sgn}(\varphi_{\text{original}}(l)) \cdot 180^\circ - \varphi_{\text{original}}(l) & \text{else} \end{cases}, \tag{6.3}$$

$$\varphi_{0,\text{perceived}}(l) = \begin{cases} \varphi_{\text{perceived}}(l) & \text{for} \quad |\varphi_{\text{perceived}}(l)| \leq 90^\circ \\ \text{sgn}(\varphi_{\text{perceived}}(l)) \cdot 180^\circ - \varphi_{\text{perceived}}(l) & \text{else} \end{cases}, \tag{6.4}$$

where $l \in \{1, ..., L\}$ with $L$ defining the length of the individual vectors. Furthermore, an additional error criterion $\tilde{\boldsymbol{\varphi}}_{\text{error}}$, which takes into account

whether the perceived azimuthal direction $\varphi_{\text{perceived}}(l)$ is more lateral ($\tilde{\varphi}_{\text{error}}(l) > 0°$) or more central ($\tilde{\varphi}_{\text{error}}(l) < 0°$) than the original azimuthal direction $\varphi_{\text{original}}(l)$, can be defined. This error criterion can be calculated element by element from the absolute error in Eq. (6.2) as

$$\tilde{\varphi}_{\text{error}}(l) = \begin{cases} \varphi_{\text{error}}(l) & \text{for} \quad |\varphi_{0,\text{perceived}}(l)| \geq |\varphi_{0,\text{original}}(l)| \\ -\varphi_{\text{error}}(l) & \text{for} \quad |\varphi_{0,\text{perceived}}(l)| < |\varphi_{0,\text{original}}(l)| \end{cases}. \tag{6.5}$$

By using $\boldsymbol{\varphi}_{0,\text{original}}$ and $\boldsymbol{\varphi}_{0,\text{perceived}}$ from Eqs. (6.3) and (6.4), front/back confusions do not influence the angular localization error. Therefore, secondly, the front/back confusion rate $\rho_{\text{fb}}$ is included in the evaluation. Here, front/back confusions are only counted for central original azimuthal directions

$$|\varphi_{\text{original}}(l)| \leq 60° \tag{6.6}$$

and perceived azimuthal directions not rounded to $|\varphi_{\text{perceived}}(l)| = 90°$ for an azimuthal resolution of $\Delta\varphi = 15°$, such that only perceived azimuthal directions that meet

$$|\varphi_{\text{perceived}}(l)| \leq 82.5° \quad \& \quad |\varphi_{\text{perceived}}(l)| \geq 97.5° \tag{6.7}$$

are counted. After excluding pairs of original and perceived azimuthal directions that do not match the conditions given in Eqs. (6.6) and (6.7), the front/back confusion rate can be calculated to

$$\rho_{\text{fb}} = \frac{N_{\text{fb}}}{N_{\text{included}}}, \tag{6.8}$$

where $N_{\text{included}}$ defines the total number of included pairs of original and perceived azimuthal directions and $N_{\text{fb}}$ defines the number of pairs underlying front/back confusions. Thirdly, the externalization is evaluated by averaging the vector of perceived externalization $\boldsymbol{d}_{\text{extern}}$ according to

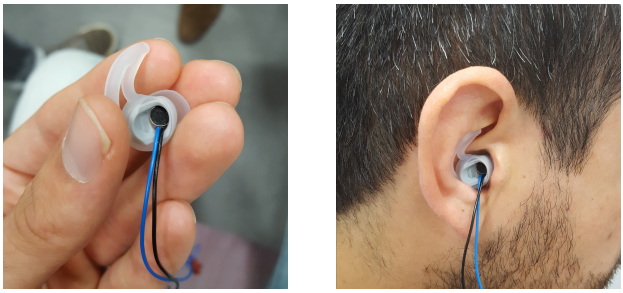
$$\bar{d}_{\text{extern}} = \text{mean}(\boldsymbol{d}_{\text{extern}}). \tag{6.9}$$

In contrast to the first two listening tests, the third one evaluates the audio quality of moving virtual sound sources, which will be explained in the corresponding subsection.

## 6.1 Listening Test I: Localization Accuracy

In the first listening test, the localization capability of parametric IIR filter cascades is compared to the one of FIR filter representations of measured HRIRs using both individual and non-individual measurements. In order to measure individual HRIRs, an own small in-ear microphone prototype

is build as presented in [Nowak et al., 2018b]. This prototype consists of ear inserts like the Bose StayHear and small electret microphones (see Fig. 6.2(a)). The ear inserts are used to place the microphones at the entrance of the ear canal as shown in Fig. 6.2(b). By using a pair of these in-ear microphones, individual binaural signals can be recorded for different subjects. Additionally, individual BRIRs and HRIRs can be measured using ESS as excitation signal and following the procedure explained in Section 5.3.2. For measuring different horizontal directions either a single loudspeaker can be positioned in front of a subject rotating in between of the different measurements as shown in Fig. 5.11 or the subject can be sat in the center of a circle containing several loudspeakers separated by the desired angular resolution.



***(a)*** *prototype* ***(b)*** *positioning*

**Figure 6.2:** The in-ear microphone prototype is build by (a) combining an electret microphone and an ear insert. In order to measure individual HRIRs, (b) the prototype is placed at the entrance of the ear canal of the subject.

In order to accelerate the measurement procedure, the multiple exponential sweep method (MESM) can be used [Majdak et al., 2007]. The principle of MESM consists of using overlapping and interleaving ESS as excitation signals for the different loudspeakers surrounding the subject. By setting the delays of the individual ESS according to the measurement setup and the room characteristics, HRIRs for different directions can be measured simultaneously, resulting in a shortened measurement duration. In this work, the MESM is implemented for six delayed ESS enabling the simultaneous measurement of HRIRs for six different directions. The six corresponding loudspeakers are placed in an angular resolution of $\Delta\varphi = 15^\circ$ around the subject starting with the loudspeaker directly in front of the subject as shown in Fig. 6.3. Additionally, three further loudspeakers marked with an orange stripe are placed at the left, the right, and in the back of the subject. The entire azimuthal space can be covered by successively

rotating the subject to these loudspeakers and repeating the measurement. This results in a total of 24 measured directions in the horizontal plane. In addition to individual HRIRs, also non-individual HRIRs of the Neumann KU100 dummy-head are measured using the internal microphones of the dummy-head. All measurements are performed inside a room for audio listening with dimensions 4.2 m × 4.8 m × 2.0 m and for HRIR lengths of 200 coefficients.

**Figure 6.3:** Measurement setup for using MESM with six loudspeakers at $\varphi_{\text{lspk}} = \{0°, -15°, ..., -75°\}$. For measuring the entire horizontal plane, the subject has to rotate successively into the directions of the marked loudspeakers at $\varphi_{\text{subject}} = \{0°, 90°, 180°, -90°\}$.

After measuring the individual and non-individual HRIRs, the magnitude responses of the corresponding HRTFs are approximated using a cascade of ten peak filters and two shelving filters as explained in Section 3.3.2. Additionally, the ITD between HRIRs of the two ears is calculated based on the methods from Section 2.1.4. Finally, the parameter vectors of the parametric IIR filter cascades for the two ears ($\mathbf{f_c}$, $\mathbf{G}$, $\mathbf{Q}$), the subtracted mean values of the two magnitude responses ($\mathbf{H}_{\text{mean}} = \begin{bmatrix}\mu_{\text{H,dB,L}} & \mu_{\text{H,dB,R}}\end{bmatrix}$), and the ITD are saved for using them inside the binaural synthesis implementation shown in Fig. 6.1.

Furthermore, minimum-phase approximations of measured HRIRs are calculated in order to use them together with the extracted ITDs instead of the measured HRIRs. As explained in Section 4.1.1, these minimum-phase HRIRs are helpful to enable a smooth interpolation especially for moving virtual sound sources. Although the first listening test does not contain moving virtual sound sources, minimum-phase HRIRs are used in order to evaluate the influence of rounding the ITDs to integer valued delays during the extraction on the localization accuracy. In this way, a fair comparison between the proposed parametric IIR filter cascade implementation and the FIR filter implementation of the binaural synthesis is achieved.

During the listening test, five different filter types are evaluated with an azimuthal resolution of $\Delta\varphi = 15°$:

- `dummy HRIR`: non-individual HRIR ($L_h = 200$)
- `ind. HRIR`: individual HRIR ($L_h = 200$)
- `ind. minPh`: individual minimum-phase HRIR ($L_h = 200$)
- `dummy IIR`: non-ind. parametric IIR filter cascade ($M = 12$)
- `ind. IIR`: individual parametric IIR filter cascade ($M = 12$)

In addition to measurements with an azimuthal resolution of $\Delta\varphi = 15°$, also a measurement grid with a resolution of $\Delta\varphi = 30°$ is used to evaluate interpolated directions between two measurements. In this way, the interpolation algorithms can be evaluated for twelve intermediate directions ($\varphi_{\text{interp}} = \{-165°, -135°, ..., 165°\}$) by comparing the localization results to the results achieved by original measurements. The interpolation contains bilinear rectangular interpolation of minimum-phase HRIRs and parameter interpolation according to Section 4.2.1 both for individual as well as non-individual dummy-head measurements. Overall, 120 audio signals using measured directional filters and 48 audio signals using interpolated directional filters are tested, resulting in a total of 168 audio signals using different filter types and directions.

### 6.1.1 Listening Test Procedure

For evaluating the localization accuracy of the defined filter types, the graphical user interface (GUI) shown in Fig. 6.4 is designed for performing the listening test. Due to the high number of different filter types and directions, the listening test uses only one stimulus containing a single snap in order to maintain an acceptable duration. Thus, the listening test consists of 168 different test stimuli. All test stimuli are generated beforehand by filtering the monaural stimulus with the corresponding directional filters and saving the resulting binaural audio signals together with the information about stimulus, filter type, and azimuthal direction. This filtering process contains only the binaural synthesis part of Fig. 6.1 without room simulation or HpEq.

In a first step, the subject has to choose his name from a drop-down list to load the audio files containing his individual measurements. Then, one of the two sets has to be chosen for the listening test. Here, both sets contain the same test stimuli in order to have two results per combination of filter type and direction for every subject. Afterwards, the order of the 168 test stimuli is randomized and the `Play sound` button is enabled. Pressing this button plays the first binaural audio signal and enables the submission of the perceived direction and the `Submit` button. The test stimulus can be played as many times as desired. The submission of the perceived direction $\varphi_{\text{perceived}}$ is performed by moving a point source to the desired direction as shown in Fig. 6.4. The shown grid has an azimuthal

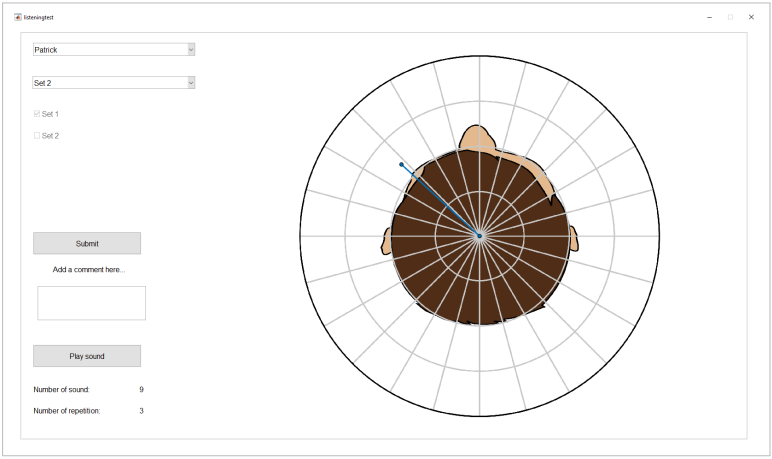



**Figure 6.4:** GUI of the first listening test which evaluates the perceived azimuthal direction $\varphi_{\text{perceived}}$ as well as the perceived externalization $d_{\text{extern}}$. The information is submitted by adjusting the perceived position ($\varphi_{\text{perceived}}$, $d_{\text{extern}}$) of the virtual sound source in the horizontal plane.

resolution of $\Delta\varphi = 15°$, but the subject is able to choose any azimuth in the horizontal plane. In addition to the azimuthal direction, also the perceived externalization $d_{\text{extern}}$ has to be submitted by adjusting the distance to the center of the head. Here, the center of the head has a perceived externalization of 0 and the outer most circle inside the GUI has a value of 1. The threshold between in-head and out-of-head localization is defined by a perceived externalization of 0.5. Furthermore, a text box is provided for adding comments about the current test stimulus. After setting the perceived azimuth and externalization as well as giving an optional comment, the result has to be confirmed by pressing the `Submit` button. This confirmation prepares the GUI for the next test stimulus enabling the `Play sound` button and disabling the submission of the perceived direction and the `Submit` button. This disabling prevents the subject from submitting a result without listening to the corresponding test stimulus. When completing the $168^{\text{th}}$ test stimulus, the listening test stops and a message window pops up saying "Thank you for attending the Listening Test! Ready for next Listening Test.". Finally, the results are saved and the check box of the chosen set is checked.

Overall, eight subjects participated in the listening test. All subjects are male research assistants in the Department of Signal Processing and Communication in the age of 26 to 37 years with an average age of 30.9 years.

Every subject performed the listening test at home using a different over-ear headphone. The listening test took about 30 to 40 minutes per set with a recommendation of taking a pause between the sets or performing them on different days. In the following, firstly, the results of the five measured filter types are evaluated. Afterwards, the results of the interpolated directions are compared to the results of the corresponding measured directions.

### 6.1.2 Results for Measured Directions

In order to evaluate the localization accuracy of parametric IIR filter cascades, their localization results are compared to localization results achieved using FIR filter representations of measured HRIRs. Figures 6.5 and 6.6 show scatter plots summarizing the localization results achieved using FIR filter representations of individual HRIRs (`ind. HRIR`) and individual parametric IIR filter cascades (`ind. IIR`), respectively.

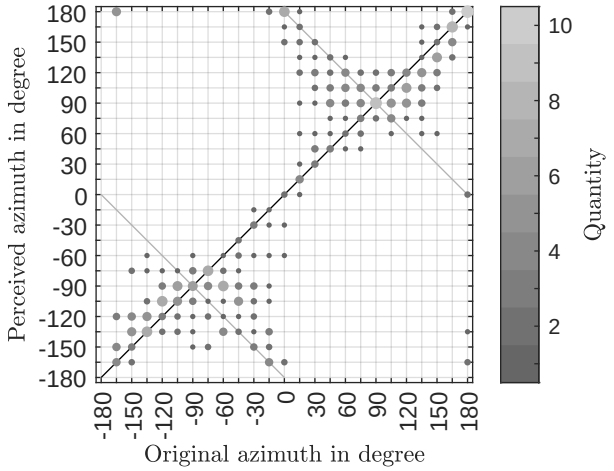



**Figure 6.5:** Scatter plot of the localization results achieved using FIR filter representations of individual HRIRs (`ind. HRIR`). Here, perceived azimuths are rounded to an azimuthal resolution of $\Delta\varphi = 15^\circ$. The black main diagonal indicates correct localization, whereas the gray negative diagonals indicate front/back confusions.

These scatter plots visualize all pairs of original azimuth $\varphi_{\text{original}}$ and perceived azimuth $\varphi_{\text{perceived}}$ submitted during the listening tests by inserting a point at the corresponding coordinates. Here, radius and color of the individual points illustrate the quantities of the corresponding pairs of

original and perceived azimuth. In order to reduce the number of individual points inside the scatter plots, the perceived azimuth is rounded to the same azimuthal resolution as used for the measured HRIRs ($\Delta\varphi = 15^\circ$). Points that lie on the black main diagonal are counted as correct localization. Contrarily, points lying on the gray negative diagonals indicate front/back confusions. The angular error $\varphi_{\text{error}}$ of a localization defined in Eq. (6.2) can be evaluated by calculating the vertical distance to the closest of these diagonals. Localization points having a smaller vertical distance to the gray negative diagonals than the black main diagonal suffer from front/back confusions. Note that in the following, only localization points satisfying the conditions from Eqs. (6.6) and (6.7) account for the calculation of the front/back confusion rate $\rho_{\text{fb}}$.

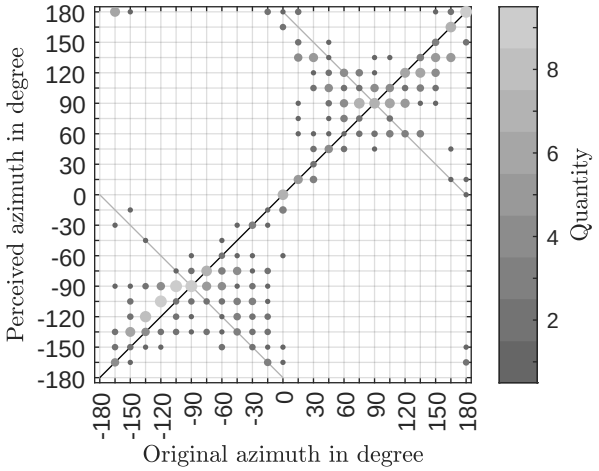



**Figure 6.6:** Scatter plot of the localization results achieved using individual parametric IIR filter cascades (`ind. IIR`). Here, perceived azimuths are rounded to an azimuthal resolution of $\Delta\varphi = 15^\circ$. The black main diagonal indicates correct localization, whereas the gray negative diagonals indicate front/back confusions.

By comparing Figs. 6.5 and 6.6, similar characteristics can be seen inside the scatter plots. Firstly, virtual sound sources are perceived more likely in the back ($|\varphi_{\text{perceived}}| > 90^\circ$) than in the front ($|\varphi_{\text{perceived}}| < 90^\circ$). Additionally, a lot of virtual sound sources are perceived laterally at $|\varphi_{\text{perceived}}| = 90^\circ$, indicating shifts of virtual sound sources towards the side. Furthermore, no localization points are visible in the top left or bottom right quarter, exhibiting the absence of left/right confusions in the localiza-

tion results, which is expected for 3D spatial audio through headphones due to the separation of the two loudspeakers inside the headphone. The presence of left/right confusions would indicate that either the headphone is not working properly or a subject was lacking in concentration.

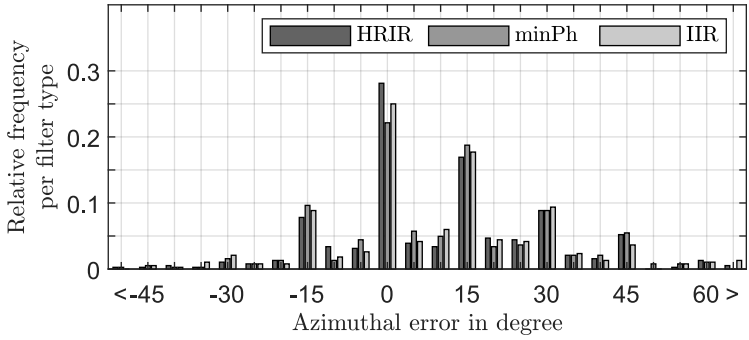
For a more detailed evaluation of the localization results, Table 6.1 gives the mean angular error $\bar{\varphi}_{\text{error}}$ from Eq. (6.1), the front/back confusion rate $\rho_{\text{fb}}$ from Eq. (6.8), and the average perceived externalization $\bar{d}_{\text{extern}}$ from Eq. (6.9) achieved by the different filter types. The similarities of the localization results seen in Figs. 6.5 and 6.6 are also visible in Table 6.1 by comparing the second and the fourth column. Individual parametric IIR filter cascades (`ind. IIR`) show a slightly higher mean angular error $\bar{\varphi}_{\text{error}}$ and a slightly lower front/back confusion rate $\rho_{\text{fb}}$ than FIR filter representations of individual HRIRs (`ind. HRIR`). However, when comparing the results of individual parametric IIR filter cascades (`ind. IIR`) in the fourth column to those of FIR filter representations of individual minimum-phase approximated HRIRs (`ind. minPh`) in the third column equal values are found for the different metrics. By comparing the localization results of individual and non-individual dummy-head filters, a slight degradation of the localization accuracy can be seen for FIR filter representations of HRIRs as well as parametric IIR filter cascades. All average perceived externalization values $\bar{d}_{\text{extern}}$ are close to 0.5, indicating that the virtual sound sources are perceived on-head rather than out-of-head. The evaluation of the average perceived externalization is treated in the second listening test.

**Table 6.1:** Summary of the localization results of the first listening test across all subjects separated for the different filter types.

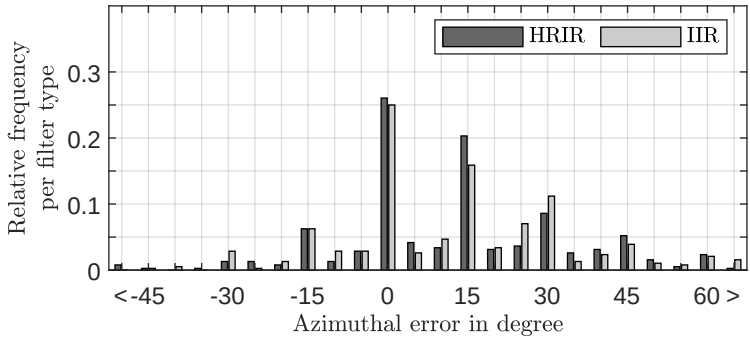
| Filter type | Individual | | | Dummy-head | |
|---|---|---|---|---|---|
| | HRIR | minPh | IIR | HRIR | IIR |
| $\bar{\varphi}_{\text{error}}$ in degree | 15.9 | 16.9 | 16.8 | 17.0 | 17.4 |
| $\rho_{\text{fb}}$ in percent | 33.5 | 29.3 | 30.0 | 33.2 | 40.1 |
| $\bar{d}_{\text{extern}}$ | 0.551 | 0.557 | 0.562 | 0.559 | 0.558 |

In Fig. 6.7, the signed azimuthal error $\tilde{\varphi}_{\text{error}}$ from Eq. (6.5) is plotted for individual and non-individual filters using an azimuthal resolution of $\Delta\varphi = 5°$. Positive azimuthal errors $\tilde{\varphi}_{\text{error}}$ indicate that the perceived azimuthal direction $\varphi_{\text{perceived}}$ is more lateral than the original azimuthal direction $\varphi_{\text{original}}$, whereas negative azimuthal errors $\tilde{\varphi}_{\text{error}}$ indicate central shifts. As can be seen, the perception of most of the virtual sound sources is shifted towards the sides. Additionally, the highest relative frequencies are visible for azimuthal error values that are multiples of 15°, which can

be explained by the measurement grid shown inside the GUI (see Fig. 6.4). Besides these similarities between different filter types, also small deviations can be seen. Firstly, FIR filter representations of measured HRIRs show a slightly higher relative frequency of correct localization than FIR filter representations of minimum-phase approximated HRIRs or parametric IIR filter cascades. Secondly, the localization results of individual HRIRs are slightly better than the ones of non-individual HRIRs.




**(a)** *individual*




**(b)** *dummy-head*

**Figure 6.7:** Relative frequency of the signed azimuthal error $\tilde{\varphi}_{\text{error}}$ from Eq. (6.5) for (a) individual and (b) non-individual filters using an azimuthal resolution of $\Delta\varphi = 5°$. Here, positive errors indicate that the perceived azimuthal direction is more lateral than the original azimuthal direction, whereas negative errors indicate a shift towards central directions.

In contrast to the other figures and tables, Table 6.2 separates the localization results for the different subjects. Here, the mean angular error

$\bar{\varphi}_{\text{error}}$ from Eq. (6.1) and the front/back confusion rate $\rho_{\text{fb}}$ from Eq. (6.8) are given for FIR filter representations of individual HRIRs (`ind. HRIR`) as well as individual parametric IIR filter cascades (`ind. IIR`). Four subjects suffer from front/back confusion rates of almost 50 %, indicating that the subjects are either guessing whether a virtual sound source is located in the front or the back, or that most of the virtual sound sources are perceived in the back, which is often the case for inexperienced subjects due to missing visual cues that would appear for frontal sound sources during natural hearing. Additionally, two subjects reach very high mean angular errors $\bar{\varphi}_{\text{error}}$ around 25°. Contrarily, *Subject 4* achieves mean angular errors $\bar{\varphi}_{\text{error}}$ of only 4.0° and 6.2° for the two different filter types. One similarity between all subjects are the comparable localization results achieved using FIR filter representations of HRIRs (`ind. HRIR`) and parametric IIR filter cascades (`ind. IIR`), confirming the validity of parametric IIR filter cascades in static binaural synthesis through headphones.

**Table 6.2:** Comparison of the localization results achieved by the individual subjects using FIR filter representations of individual HRIRs (`ind. HRIR`) and individual parametric IIR filter cascades (`ind. IIR`).

| Metric | $\bar{\varphi}_{\text{error}}$ in degree | | $\rho_{\text{fb}}$ in percent | |
|---|---|---|---|---|
| Filter type | `ind. HRIR` | `ind. IIR` | `ind. HRIR` | `ind. IIR` |
| *Subject 1* | 15.9 | 16.4 | 48.4 | 32.4 |
| *Subject 2* | 26.2 | 25.6 | 46.4 | 34.5 |
| *Subject 3* | 15.4 | 15.2 | 44.4 | 45.7 |
| *Subject 4* | 4.0 | 6.2 | 19.4 | 25.0 |
| *Subject 5* | 12.0 | 12.4 | 12.5 | 18.2 |
| *Subject 6* | 12.6 | 11.8 | 25.8 | 20.0 |
| *Subject 7* | 25.2 | 28.2 | 33.3 | 30.0 |
| *Subject 8* | 15.7 | 16.6 | 50.0 | 50.0 |

### 6.1.3 Results for Interpolated Directions

In practical implementations, only a finite number of measured HRIRs is available, resulting in a limited spatial resolution. Thus, spatial interpolation is required to increase the spatial resolution. In order to evaluate the parameter interpolation algorithm from Chapter 4, interpo-

lated directions are included in the listening test. In the following, the results presented in [Nowak and Zölzer, 2022] are evaluated in more detail. As described previously, twelve interpolated directions are generated at $\varphi_{\text{interp}} = \{-165^\circ, -135^\circ, ..., 165^\circ\}$ using a measurement grid with an azimuthal resolution of $\Delta\varphi = 30^\circ$. The interpolation contains bilinear rectangular interpolation of minimum-phase HRIRs and parameter interpolation according to Section 4.2.1 both for individual and non-individual dummy-head measurements.

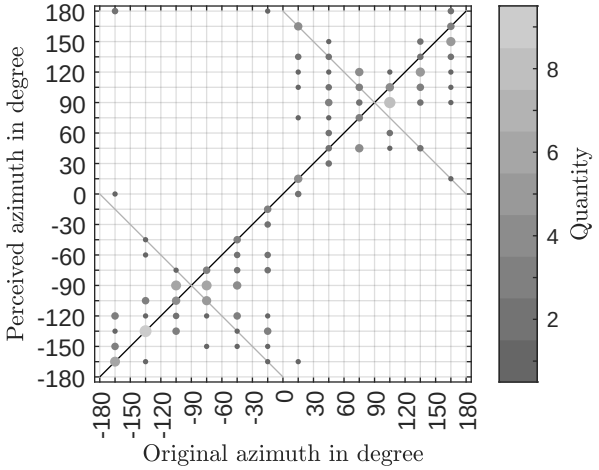



**Figure 6.8:** Scatter plot of the localization results achieved using individual parametric IIR filter cascades and parameter interpolation (`ind. interp. IIR`) for generating the interpolated directions. Perceived azimuths are rounded to an azimuthal resolution of $\Delta\varphi = 15^\circ$. The black main diagonal indicates correct localization, whereas the gray negative diagonals indicate front/back confusions.

Figure 6.8 summarizes the localization results of individual parametric IIR filter cascades and parameter interpolation (`ind. interp. IIR`) for the interpolated directions $\varphi_{\text{interp}}$ in a scatter plot. By comparing the results of the interpolated directions in Fig. 6.8 to the results of the measured directions in Fig. 6.6, similar characteristics can be seen. Firstly, more virtual sound sources are perceived rear than frontal. Secondly, a lot of virtual sound sources are perceived laterally ($|\varphi_{\text{perceived}}| = 90^\circ$). Note that the original azimuth axis in Fig. 6.8 is more sparse due to the azimuthal resolution of $\Delta\varphi = 30^\circ$ for interpolated directions instead of the azimuthal resolution of $\Delta\varphi = 15^\circ$ for measured directions in Fig. 6.6.

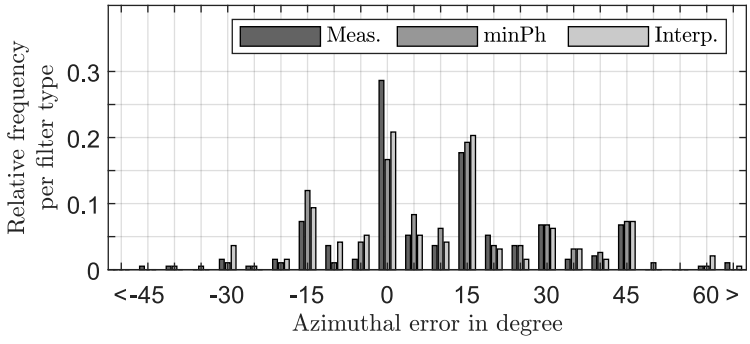
In addition to the scatter plots, Tables 6.3 and 6.4 summarize the mean angular error $\bar{\varphi}_{\text{error}}$ from Eq. (6.1), the front/back confusion rate $\rho_{\text{fb}}$ from Eq. (6.8), and the average perceived externalization $\bar{d}_{\text{extern}}$ from Eq. (6.9) achieved by the different measured and interpolated filter types. Note that for a fair comparison, also for measured filter types only directions that are interpolated ($\varphi_{\text{interp}} = \{-165°, -135°, ..., 165°\}$) are used to calculate the evaluation criteria. Thus, the values in Tables 6.3 and 6.4 differ from the ones in Table 6.1. The increase in the mean angular error $\bar{\varphi}_{\text{error}}$ especially for non-individual filter types might be explained by the missing lateral directions ($|\varphi_{\text{perceived}}| = 90°$), which show a high rate of correct localization in the localization results of measured directions.

**Table 6.3:** Summary of the localization results of the first listening test for interpolated directions $\varphi_{\text{interp}} = \{-165°, -135°, ..., 165°\}$ across all subjects separated for the different individual filter types.

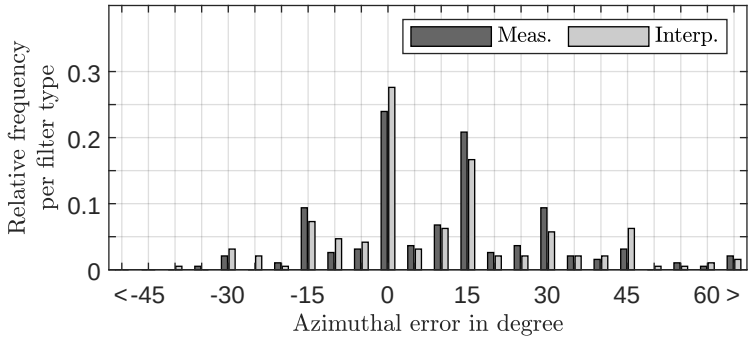
| Filter type | HRIR | | | IIR | |
|---|---|---|---|---|---|
| | Meas. | minPh | Interp. | Meas. | Interp. |
| $\bar{\varphi}_{\text{error}}$ in deg. | 15.8 | 17.0 | 16.8 | 16.3 | 16.2 |
| $\rho_{\text{fb}}$ in pct. | 33.6 | 30.5 | 37.6 | 36.1 | 28.0 |
| $\bar{d}_{\text{extern}}$ | 0.587 | 0.588 | 0.584 | 0.598 | 0.583 |

In Table 6.3, the localization results of the individual filter types are evaluated for FIR filter representations of HRIRs (`ind. HRIR`) as well as parametric IIR filter cascades (`ind. IIR`). The interpolation of FIR filter representations of HRIRs uses the minimum-phase approximations of the measured HRIRs (`ind. minPh`). When comparing the localization results achieved by the FIR filter representations of these minimum-phase HRIRs (`ind. minPh`) with the ones achieved by FIR filter representations of interpolated minimum-phase HRIRs (`ind. interp. minPh`) for the same directions, almost no difference can be seen in the mean angular error $\bar{\varphi}_{\text{error}}$. Similar results are also visible for the individual parametric IIR filter cascades. Contrarily, the evaluation of the front/back confusion rate $\rho_{\text{fb}}$ shows different tendencies for the two interpolation methods.

In Fig. 6.9, the signed azimuthal error $\tilde{\varphi}_{\text{error}}$ from Eq. (6.5) is plotted for FIR filter representations of individual HRIRs and individual parametric IIR filter cascades. Positive azimuthal errors $\tilde{\varphi}_{\text{error}}$ indicate that the perceived azimuthal direction $\varphi_{\text{perceived}}$ is more lateral than the original azimuthal direction $\varphi_{\text{original}}$, whereas negative azimuthal errors $\tilde{\varphi}_{\text{error}}$ indicate central shifts. In Fig. 6.9(a), the different filter types based on FIR filter representations of HRIRs are compared, namely measured HRIRs




*(a) HRIR*




*(b) IIR*

**Figure 6.9:** Relative frequency of the signed azimuthal error $\tilde{\varphi}_{\text{error}}$ from Eq. (6.5) for (a) FIR filter representations of individual HRIRs and (b) individual parametric IIR filter cascades using an azimuthal resolution of $\Delta\varphi = 5^\circ$. Here, positive errors indicate that the perceived azimuthal direction is more lateral than the original azimuthal direction, whereas negative errors indicate a shift towards central directions.

(`ind. HRIR`), minimum-phase approximated HRIRs (`ind. minPh`), and interpolated minimum-phase HRIRs (`ind. interp. minPh`). This comparison shows a higher relative frequency of correct localization for measured HRIRs (`ind. HRIR`) than for minimum-phase approximated HRIRs (`ind. minPh`), which might be explained by using rounded extracted ITDs as delay for the contralateral ear. Also the interpolated minimum-phase HRIRs (`ind. interp. minPh`) show a slightly higher relative frequency of correct localization than the minimum-phase approximated HRIRs (`ind. minPh`).

Similar as in Fig. 6.7, most of the virtual sound sources are shifted towards the sides with the highest relative frequencies appearing at azimuthal errors that are multiples of 15°. In Fig. 6.9(b), a low enhancement of the relative frequency of correct localization is visible for interpolated parametric IIR filter cascades (`ind. interp. IIR`) in comparison to measured ones (`ind. IIR`). This difference is also responsible for the slight decrease in the mean angular error $\bar{\varphi}_{\text{error}}$ seen in Table 6.3.

In Table 6.4, the localization results of the interpolation methods are evaluated for FIR filter representations of non-individual dummy-head HRIRs (`dummy HRIR`) and non-individual dummy-head parametric IIR filter cascades (`dummy IIR`). As can be seen, these non-individual filter types show higher mean angular errors $\bar{\varphi}_{\text{error}}$ and front/back confusion rates $\rho_{\text{fb}}$ than the corresponding individual filter types in Table 6.3. For parametric IIR filter cascades (`dummy IIR`) the same trends can be seen for measured and interpolated filters as in Table 6.3. However, FIR filter representations of interpolated minimum-phase HRIRs (`dummy interp. minPh`) show an obvious decrease in the mean angular error $\bar{\varphi}_{\text{error}}$ in comparison to FIR filter representations of measured HRIRs (`dummy HRIR`).

**Table 6.4:** Summary of the localization results of the first listening test for interpolated directions $\varphi_{\text{interp}} = \{-165°, -135°, ..., 165°\}$ across all subjects separated for the different non-individual dummy-head filter types.

| Filter type | HRIR | | IIR | |
|---|---|---|---|---|
| | Meas. | Interp. | Meas. | Interp. |
| $\bar{\varphi}_{\text{error}}$ in degree | 18.3 | 17.0 | 18.6 | 18.2 |
| $\rho_{\text{fb}}$ in percent | 33.3 | 35.6 | 48.3 | 45.7 |
| $\bar{d}_{\text{extern}}$ | 0.596 | 0.603 | 0.593 | 0.582 |

Overall, the similar localization results between measured and interpolated filter types indicate the validity of the interpolation methods. Thus, also the validity of parameter interpolation is proven for static virtual sound sources.

## 6.2 Listening Test II: Externalization

As explained in Section 5, room effects are an important cue for perceiving static virtual sound sources externalized. Thus, simulated room effects are used to equip short HRIRs with the required room effects. In order to evaluate the effect of room effects, a second listening test is performed that inquires both the perceived externalization as well as the perceived

azimuthal direction [Nowak et al., 2020b]. The procedure of the listening test equals the one from Section 6.1.1, thus also the GUI is the same as shown in Fig. 6.4. However, different filter types are used to generate the test stimuli. In order to include HpEq, the same headphone is used for every subject, namely the Beyerdynamic DT770 Pro 250 Ohm over-ear headphone. Due to this reason, the listening test was performed in the laboratory of the Department of Signal Processing and Communication.

Overall, five different filter types are evaluated during the listening test:

- `BRIR`: non-individual BRIR ($L_h = 8192$)
- `HRIR`: non-individual HRIR ($L_h = 200$)
- `IIR`: non-individual parametric IIR filter cascade ($M = 12$)
- `ISM`: `IIR` with simulated room effects via ISM ($L_h = 8192$)
- `HpEq`: `ISM` with non-individual HpEq ($M = 12$)

All filter types are based on non-individual BRIR measurements of the Neumann KU100 dummy-head taken inside a room for audio listening with dimensions $4.2\,\mathrm{m} \times 4.8\,\mathrm{m} \times 2.0\,\mathrm{m}$ (see Fig. 5.11). The measurements are performed for an azimuthal resolution of $\Delta\varphi = 30°$, resulting in a total of twelve directions. From these measurements, firstly the BRIRs of length $L_h = 8192$ and the shortened HRIRs of length $L_h = 200$ are extracted as explained in Section 5.3.3. The FIR filter representations of these impulse responses are implemented in the filter types `BRIR` and `HRIR`. Secondly, a parametric IIR filter cascade containing an LFS, an HFS, and ten peak filters is used to approximate the magnitude responses of the shortened HRIRs according to Section 3.3. Additionally, the ITDs are extracted as explained in Section 2.1.4. Thus, the filter type `IIR` is implemented as shown in the binaural synthesis part from Fig. 6.1. Afterwards, room simulation is added as explained in Section 5.3.3 and shown in Fig. 6.1 to form the filter type referred as `ISM`. Finally, HpEq is appended as a third step in order to evaluate the whole application shown in Fig. 6.1, resulting in the filter type `HpEq`. As explained in Section 2.2.3, for non-individual HRIRs also HpEq should be performed non-individually using the same subject as for the HRIR measurements. Thus, HpEq is calculated based on HpTFs measured for the Neumann KU100 dummy-head using the Beyerdynamic DT770 Pro 250 Ohm over-ear headphone. As described in Section 3.5, HpEq is implemented using a cascade of parametric IIR filters, too. Similar as for the magnitude responses of HRTFs, the HpEq uses ten peak filters.

During the listening test, two different stimuli are used, namely a speech signal that contains only a single word ("matter") and a snap representing a broadband signal. Having two different stimuli, five different filter types, and twelve different directions results in a total of 120 test stimuli. In order

to enhance the number of results per test stimulus, every subject listens to the set of 120 test stimuli twice.

Overall, eight subjects participated in the listening test. All subjects are male in the age of 23 to 33 years with an average age of 27.4 years. Six of them are research assistants in the Department of Signal Processing and Communication and the other two subjects are students writing their thesis in the Department of Signal Processing and Communication. The listening test took about 25 to 30 minutes per set with a pause between the sets. In the following, the results of the listening test presented in [Nowak et al., 2020b] are evaluated in more detail.

### 6.2.1 Results

In a first step, the reliability of the listening test results is proven by evaluating the localization accuracy (see Fig. 6.10). Here, the listening test results of all subjects are summarized in a single scatter plot. Although Fig. 6.10 shows a lot of front/back confusions, the results can be seen as plausible, because of the absence of left/right confusions. Otherwise, the presence of left/right confusions would indicate that either the headphone is not working properly or a subject was lacking in concentration.

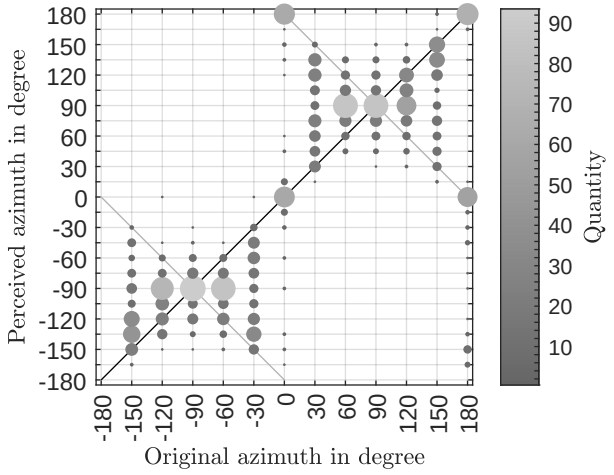



**Figure 6.10:** Scatter plot for evaluating the reliability of the results of the second listening test by monitoring the localization accuracy across all subjects.

Furthermore, Table 6.5 gives the mean angular error $\bar{\varphi}_{\text{error}}$ from Eq. (6.1), the front/back confusion rate $\rho_{\text{fb}}$ from Eq. (6.8), and the average perceived

externalization $\bar{d}_{\text{extern}}$ from Eq. (6.9) achieved by the different filter types. All filter types show mean angular errors of 16.9° to 18.7° and very high front/back confusion rates around 40 %, which confirm the results visible in the scatter plot in Fig. 6.10. Front/back confusion rates of almost 50 % indicate that the subjects are either guessing whether a sound source is located in the front or the back or that most of the sound sources are perceived in the back, which is often the case for inexperienced subjects due to missing visual cues that would appear for frontal sound sources during natural hearing. Nevertheless, the primary evaluation criterion of the underlying listening test is given by the perceived externalization rather than the localization capabilities. For this, the average perceived externalization $\bar{d}_{\text{extern}}$ achieved by using the different filter types is included in Table 6.5, too. As can be seen, reducing the length of the FIR filter from $L_h = 8192$ (`BRIR`) to $L_h = 200$ (`HRIR`) leads to a drastic decrease in the average perceived externalization. This decrease can be explained by the missing room effects inside the shortened impulse responses of `HRIR`. Approximating the magnitude responses of `HRIR` using parametric IIR filter cascades (`IIR`) has almost no influence on the perceived externalization. By adding simulated room impulse responses (`ISM`), the average perceived externalization is raised to levels that are comparable to those of measured BRIRs in `BRIR`. Moreover, appending HpEq in addition to simulated room effects (`HpEq`) further increases the average perceived externalization slightly.

**Table 6.5:** Summary of the results of the second listening test across all subjects separated for the different filter types.

| Filter type | `BRIR` | `HRIR` | `IIR` | `ISM` | `HpEq` |
|---|---|---|---|---|---|
| $\bar{\varphi}_{\text{error}}$ in degree | 16.9 | 17.6 | 18.7 | 17.8 | 18.5 |
| $\rho_{\text{fb}}$ in percent | 41.0 | 38.4 | 42.0 | 42.4 | 44.2 |
| $\bar{d}_{\text{extern}}$ | 0.670 | 0.608 | 0.605 | 0.687 | 0.699 |

In addition to the calculation of the average perceived externalization per filter type in Table 6.5, also box-and-whisker plots are shown in Fig. 6.11, containing the lower whisker $w_{\text{l}}$, the first quartile $q_1$, the median $q_2$, the third quartile $q_3$, the upper whisker $w_{\text{u}}$, and outliers. Here, $q_1$, $q_2$, and $q_3$ define thresholds that are fallen below by 25 %, 50 %, and 75 % of the values inside $\mathbf{d}_{\text{extern}}$, respectively. Additionally, the IQR is defined as

$$\text{IQR} = q_3 - q_1. \tag{6.10}$$

From this, the whiskers $w_{\text{l}}$ and $w_{\text{u}}$ are given by the lowest and highest

values inside $\mathbf{d}_{\text{extern}}$ fulfilling

$$w_{\text{l}} = \min\left(\mathbf{d}_{\text{extern}} \geq q_1 - 1.5 \cdot \text{IQR}\right), \tag{6.11}$$

$$w_{\text{u}} = \max\left(\mathbf{d}_{\text{extern}} \leq q_3 + 1.5 \cdot \text{IQR}\right). \tag{6.12}$$

All values inside $\mathbf{d}_{\text{extern}}$ that are outside of the range spanned by $w_{\text{l}}$ and $w_t extu$ in Eqs. (6.11) and (6.12) are taken as outliers and marked as single dots inside the box-and-whisker plots.

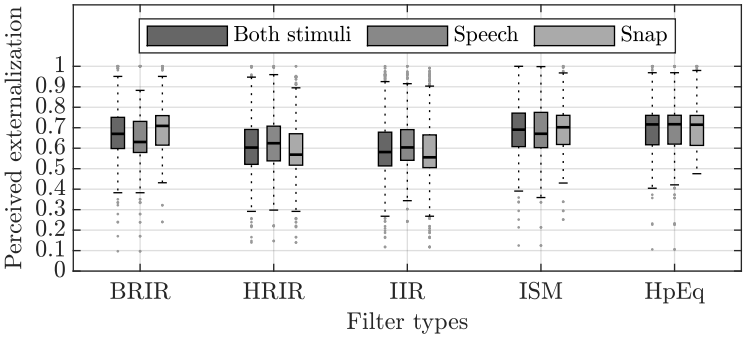



**Figure 6.11:** Box-and-whisker plots for the different filter types containing the lower whisker $w_{\text{l}}$, the first quartile $q_1$, the median $q_2$, the third quartile $q_3$, the upper whisker $w_{\text{u}}$, and outliers. Additionally, the results are separated for the different stimuli.

In Fig. 6.11, these metrics are used to evaluate the perceived externalization achieved by using the different filter types. A comparison of the median $q_2$ for the different filter types confirms the relations deduced from the average perceived externalization $\bar{d}_{\text{extern}}$ in Table 6.5. Additionally, IQRs of roughly 0.15 to 0.2 for all filter types indicate a similar spreading of perceived externalization for all filter types. These variations in perceived externalization per filter type can be explained by different biases in the interpretation of externalization for individual subjects. Since all boxes lie above 0.5, all filter types can be counted as outside of the head. However, the two filter types containing no room effects (`HRIR`, `IIR`) are closer to this threshold than the others. Thus, they can be stated as being on-head rather than out-of-head. Furthermore, results that are separated for the different stimuli indicate that a stronger influence of included room effects on perceived externalization is visible for the broadband snap.

By using the spider plots in Figs. 6.12 and 6.13, the perceived externalization in dependence of the original azimuthal direction is evaluated. Here, a separate average perceived externalization $\bar{d}_{\text{extern},\varphi}$ is calculated

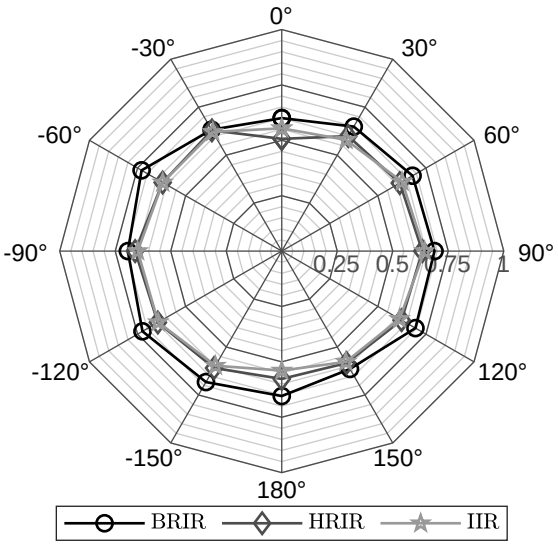



**Figure 6.12:** Spider plot for evaluating the influence of the filter length on the perceived externalization per azimuthal direction.

per azimuthal direction $\varphi_{\text{original}}$ according to Eq. (6.9) using only the values inside $\mathbf{d}_{\text{extern}}$ related to the given original azimuthal direction $\varphi_{\text{original}}$. A commonality between both spider plots is the fact that lateral directions show higher average perceived externalizations than frontal and rear directions, which can be explained by the position of the loudspeakers of the headphone at these directions. In contrast to this, frontal and rear sound sources have to be generated completely virtual. In Figs. 6.12 and 6.13, the influence of room effects on the perceived externalization is illustrated.

Firstly, Fig. 6.12 shows the influence of the FIR filter length on the perceived externalization per azimuthal direction. From this, the negative influence of the missing room effects in `HRIR` and `IIR` on the average perceived externalization is clearly visible for all azimuthal directions except $\varphi = -30°$.

Secondly, the influence of adding simulated room effects on the perceived externalization per azimuthal direction is shown in Fig. 6.13 by comparing the average perceived externalization achieved using `BRIR`, `IIR`, and `ISM`. The simulated room effects contained in `ISM` increase the average perceived externalization achieved by using `IIR` for all azimuthal directions $\varphi$. The resulting average perceived externalization reaches levels that are equal or higher than those of real measured room effects contained in `BRIR`.

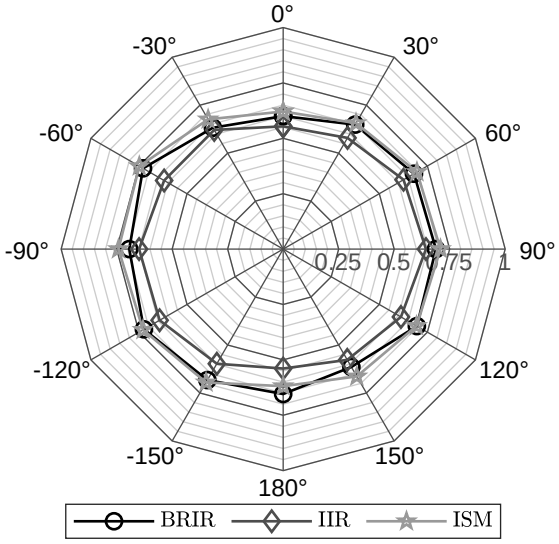



**Figure 6.13:** Spider plot for evaluating the influence of adding simulated room effects on the perceived externalization per azimuthal direction.

Overall, the listening test results confirm that missing room effects inside short HRIRs drastically reduce the perceived externalization of static virtual sound sources. By adding simulated reflections via ISM, this loss in perceived externalization can be compensated. Also for binaural synthesis using parametric IIR filter cascades, appending a room simulator enables externalized static virtual sound sources.

## 6.3 Listening Test III: Moving Virtual Sound Sources

In contrast to the previous listening tests, the third listening test evaluates moving virtual sound sources rather than static ones, thus also the procedure and the utilized GUI have to be changed. The new GUI is shown in Fig. 6.14.

Since the third listening test targets on evaluating the audio quality of generated moving virtual sound sources, the subjects have to rate the individual test stimuli between 0 (bad quality) and 100 (very good quality). Here, the evaluation criteria are given by the plausibility of direction, smoothness of transition, and audibility of artifacts.

Similar as in the other listening test, the subject firstly has to choose his name from the drop-down list in order to save the results afterwards. Since only non-individual HRIRs of *Subject_065* from the CIPIC database [Algazi

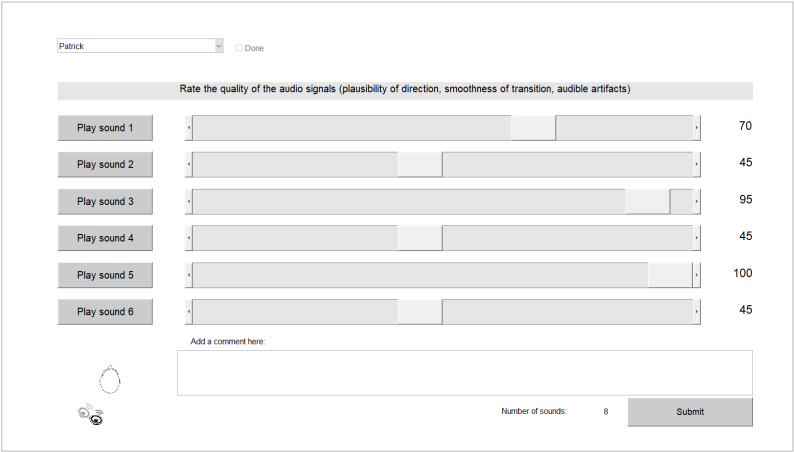



**Figure 6.14:** GUI of the third listening test, which evaluates the quality of audio signals representing moving virtual sound sources. Here, the rating of audio quality contains plausibility of direction, smoothness of transition, and audibility of artifacts.

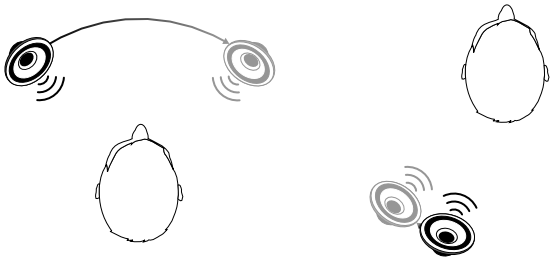
et al., 2001b] are used, the same set of test stimuli is used for every subject. This set contains six different filter types:

- `FIR`: non-individual HRIR ($L_h = 200$)
- `minPh`: non-individual minimum-phase HRIR ($L_h = 200$)
- `IIR`: non-individual parametric IIR filter cascade ($M = 12$)
- `IIR2`: two parallel `IIR` with input-switching
- `IIR2f`: `IIR2` with cross-fading input-switching combination
- `IIR2fs`: `IIR2f` with smoothed parametric IIR filter cascades

The filter type `FIR` simply uses FIR filter representations of HRIRs taken from the database, whereas `minPh` and `IIR` approximate them as explained in Section 6.1. In addition to the usage of a single IIR filter cascade in `IIR`, also three filter types are included that use two parallel IIR filter cascades with different switching methods as explained in Section 4.2.2. Firstly, `IIR2` uses the input-switching method from Fig. 4.3. Secondly, `IIR2f` improves the switching between the two IIR filter cascades by using a cross-fading input-switching combination approach as shown in Fig. 4.12. Thirdly, a further smoothing of the transition is achieved by reducing the interpolation weight $\tilde{c}_{i_{\mathrm{ref}},\mathrm{p}}$ for the gain $G_{\mathrm{p},i_{\mathrm{ref}}}$ according to Eq. (4.17) when moving towards directions that do not contain that peak filter, such that

the peak filter disappears before changing the reference direction. This smoothing is implemented in `IIR2fs`.

Furthermore, three different stimuli and three different scenarios are used. The different stimuli are given by a speech signal, a white noise signal, and a music signal. All signals have a duration of two seconds plus 8192 padded zeros to ensure a decaying of the room effects inside the filtered audio signals. The different scenarios can be separated into two different movements (see Fig. 6.15) generated using an azimuthal measurement grid with $\Delta\varphi = 5^\circ$ and a third one that creates the movement from Fig. 6.15(a) using a coarser measurement grid with an azimuthal resolution of $\Delta\varphi = 15^\circ$.



***(a)** frontal movement* ***(b)** rear movement*

**Figure 6.15:** Two different sound sources moving from the position of the black loudspeaker to the position of the gray loudspeaker in (a) the front ($-40^\circ \rightarrow 40^\circ$) and (b) the back ($-160^\circ \rightarrow -140^\circ$).

Having three different stimuli and three different scenarios leads to a total of nine different comparisons, where every comparison includes six test stimuli filtered by the six different filter types. This filtering is performed beforehand, such that during the listening test, the corresponding test stimuli only have to be loaded. After entering the name, the first comparison is loaded and the six `Play sound` buttons are enabled. The desired movement is shown in the bottom left corner. Furthermore, the loading process contains a randomization of the six test stimuli in order to hide information about the order of the filter types. After listening to one of the test stimuli, the corresponding bar is activated, such that the subject is able to rate the audio quality of the test stimulus. For rating the audio quality of the individual test stimuli, all three criteria have to be summed up to a single audio quality rating. Although plausibility of direction is one of the evaluation criteria, none of the filter types reproduces movements that are not feasible. Thus, smoothness of transition and audibility of artifacts are the most important evaluation criteria. After rating all six test

stimuli, the subject is able to confirm the rating by pressing the `Submit` button. Similar as in the localization test, the subject can listen to every test stimulus as many times as desired. Moreover, the option of giving an additional comment is also included in this listening test. Pressing the `Submit` button guides the listening test to the next comparison by loading the corresponding test stimuli, enabling the `Play sound` buttons, and disabling the bars as well as the `Submit` button. In order to enhance the number of results per comparison, every comparison is performed twice during the listening test, resulting in 18 comparisons that have to be performed by the subject. The order of these comparisons is randomized, too. When completing all comparisons, the listening test stops and a message window pops up saying "Thank you for attending the Listening Test!". Finally, the results are saved and the `Done` check box is checked.

Overall, eight subjects participated in the listening test. All subjects are male research assistants in the Department of Signal Processing and Communication in the age of 26 to 37 years with an average age of 30.9 years. Every subject performed the listening test at home using a different over-ear headphone. The listening test took approximately 30 minutes. In the following, the results of the listening test, which are also presented in [Nowak and Zölzer, 2022], are evaluated.

### 6.3.1 Results

In Table 6.6, the average audio quality ratings per filter type are listed for the eight subjects. Every entry represents the mean across 18 ratings, mixing all three stimuli and all three scenarios. In every row, the maximum value and values reaching at least 95 % of this maximum value are highlighted. As can be seen, the FIR filter representations of minimum-phase approximated HRIRs (`minPh`) are highlighted six times and the smoothed parallel IIR filter cascades (`IIR2fs`) five times. Additionally, the parallel IIR filter cascades using the cross-fading input-switching combination approach (`IIR2f`) are highlighted once. These numbers indicate that both the FIR filter representations of minimum-phase approximated HRIRs (`minPh`) and the smoothed parallel IIR filter cascades (`IIR2fs`) are rated with the highest audio quality for almost every subject. The average audio quality rating across all subjects acknowledges this preference in audio quality. Although the average audio quality ratings of all subjects show similar rankings, the individual ratings differ in range, with *Subject 1* having average ratings in the range of 19.8 (`IIR`) to 58.2 (`IIR2f`) and *Subject 6* having average ratings from 23.1 (`IIR`) to 100.0 (`minPh`).

In addition to the average audio quality ratings per subject in Table 6.6, Fig. 6.16 shows box-and-whisker plots, containing the lower whisker $w_\mathrm{l}$, the first quartile $q_1$, the median $q_2$, the third quartile $q_3$, the upper whisker $w_\mathrm{u}$, and outliers of the audio quality ratings across all subjects. As can be

**Table 6.6:** Summary of the average audio quality ratings of the third listening test separated for the different filter types and subjects. The maximum value per row and values reaching at least 95 % of this maximum value are highlighted.

| Filter type | FIR | minPh | IIR | IIR2 | IIR2f | IIR2fs |
|---|---|---|---|---|---|---|
| *Subject 1* | 30.2 | **55.9** | 19.8 | 25.3 | **58.2** | **56.9** |
| *Subject 2* | 44.1 | 63.6 | 22.9 | 26.0 | 61.9 | **73.4** |
| *Subject 3* | 46.1 | **75.1** | 22.9 | 20.2 | 53.4 | **77.9** |
| *Subject 4* | 47.7 | 87.3 | 23.6 | 33.6 | 88.8 | **93.8** |
| *Subject 5* | 60.0 | **99.7** | 24.4 | 25.8 | 78.9 | 94.7 |
| *Subject 6* | 60.3 | **100.0** | 23.1 | 25.0 | 78.6 | 93.9 |
| *Subject 7* | 49.1 | **75.9** | 22.3 | 25.3 | 52.6 | 67.9 |
| *Subject 8* | 72.3 | **90.3** | 59.9 | 62.4 | 82.8 | **87.4** |
| All Subj. | 51.2 | **81.0** | 27.4 | 30.5 | 69.4 | **80.7** |

seen from the box-and-whisker plots for the all stimuli evaluation, the FIR filter representations of minimum-phase approximated HRIRs (`minPh`) and the smoothed parametric IIR filter cascades (`IIR2fs`) achieve first quartile values above 70, whereas the single parametric IIR filter cascade (`IIR`) and the parallel parametric IIR filter cascade without fading (`IIR2`) show third quartile values below 50. Additionally, FIR filter representations of HRIRs show a very high IQR of 75. This high variation can be explained by the box-and-whisker plots separated for the different stimuli. For speech stimuli, both FIR filter representations (`FIR`, `minPh`) and the two parallel parametric IIR filter cascades with fading (`IIR2f`, `IIR2fs`) show good audio quality ratings. Contrarily, the single IIR filter cascade (`IIR`) and the parallel parametric IIR filter cascade without fading (`IIR2`) show low audio quality ratings even for speech stimuli, which can be explained by the clicks inside the filtered stimuli produced during update of the IIR filter coefficients. The audio quality ratings of FIR filter representations of HRIRs (`FIR`) and parallel parametric IIR filter cascade using the cross-fading input-switching combination approach (`IIR2f`) drastically drop when considering white noise stimuli. Here, `FIR` suffers from comb filtering effects due to interpolation of HRIRs with different delays, whereas `IIR2f` suffers from audible coloration due to missing peak filters after changing the reference direction inside extended parameter interpolation.

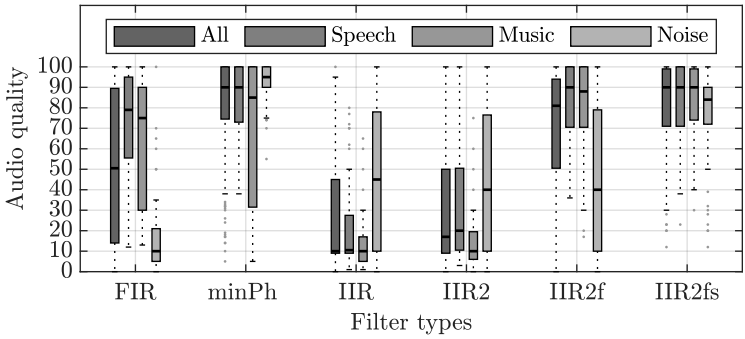



**Figure 6.16:** Box-and-whisker plots representing the audio quality for the different filter types containing the lower whisker $w_l$, the first quartile $q_1$, the median $q_2$, the third quartile $q_3$, the upper whisker $w_u$, and outliers. Additionally, the results are separated for the different stimuli.

In order to reduce the subject-dependent range of audio quality ratings seen in Table 6.6, Fig. 6.17 shows the same results as Fig. 6.16, but uses a scaled audio quality. This scaled audio quality normalizes the audio quality ratings inside every comparison to the entire range of 0 to 100, such that the worst audio quality inside every comparison is scaled to 0 and the best audio quality is scaled to 100. In this way, the ratings of the individual subjects are normalized to the same range. The scaled audio quality in Fig. 6.17 confirms the relations of the audio quality seen in Fig. 6.16, but enhances the conspicuousness of the fact that `IIR` and `IIR2` show the worst audio quality for almost every comparison. Similarly, `minPh` and `IIR2fs` achieve the best audio quality ratings for almost every comparison. Additionally, `IIR2f` shows very good normalized audio quality ratings except for broadband noise stimuli, which can be explained by the coloration effects that are audible when a peak filter disappears while changing the reference direction inside extended parameter interpolation.

In contrast to the other two box-and-whisker plots, Fig. 6.18 separates the listening test results into the different scenarios. While `minPh`, `IIR`, and `IIR2` show constant median values across scenarios, the median values of the other three filter types vary strongly with scenario. For `FIR`, the comb filtering effects that are audible while changing between HRIRs with different delays are the strongest for the frontal movement with an azimuthal resolution of $\Delta\varphi = 5°$. The reason for this is the high number of changes between different HRIRs when moving from $\varphi_1 = -40°$ to $\varphi_2 = 40°$ with an azimuthal resolution of $\Delta\varphi = 5°$. Contrarily, for `IIR2f` and `IIR2fs`, the rear movement suffers from the strongest audible coloration

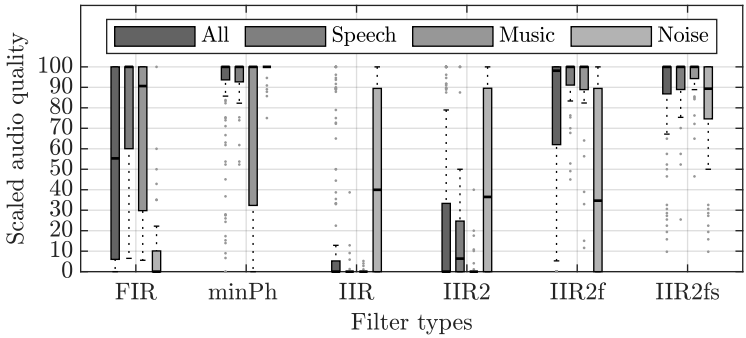



**Figure 6.17:** Box-and-whisker plots representing the scaled audio quality for the different filter types containing the lower whisker $w_l$, the first quartile $q_1$, the median $q_2$, the third quartile $q_3$, the upper whisker $w_u$, and outliers. Additionally, the results are separated for the different stimuli.

due to missing peak filters after changing the reference direction inside parameter interpolation, resulting in a lower audio quality for this scenario. For frontal movements, these two parallel parametric IIR filter cascades with fading (`IIR2f`, `IIR2fs`) show a similar audio quality than FIR filter representations of minimum-phase approximated HRIRs (`minPh`).

Overall, similar audio quality ratings between FIR filter representations of minimum-phase approximated HRIRs (`minPh`) and smoothed parallel parametric IIR filter cascades (`IIR2fs`) indicate the validity of parameter interpolation for moving virtual sound sources. Except for broadband noise stimuli, also parallel parametric IIR filter cascades using the cross-fading input-switching combination approach (`IIR2f`) show very good audio quality ratings.

## 6.4 Summary

In order to evaluate the methods proposed in the previous chapters, three listening tests are presented in this chapter, which assess different aspects of binaural synthesis using HRTFs approximated by parametric IIR filters shown in Fig. 6.1.

In a first listening test, the localization accuracy of the proposed parametric IIR filter cascade is compared to localization results achieved using FIR filter representations of HRIRs for measured as well as interpolated static virtual sound source directions. Similarities in the mean angular errors and the front/back confusion rates confirm the validity of the parametric IIR filter cascades in static binaural synthesis. Additionally, a comparison of

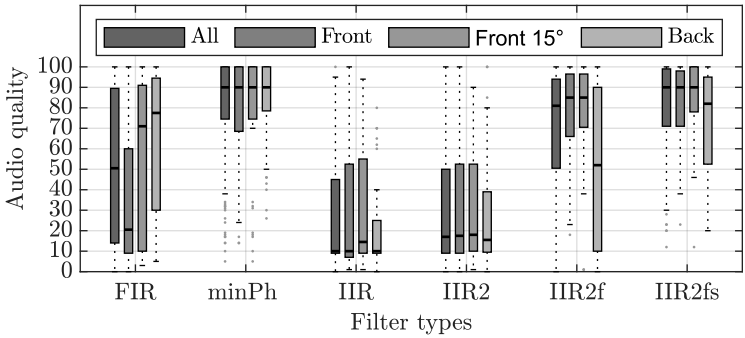



**Figure 6.18:** Box-and-whisker plots representing the audio quality for the different filter types containing the lower whisker $w_l$, the first quartile $q_1$, the median $q_2$, the third quartile $q_3$, the upper whisker $w_u$, and outliers. Additionally, the results are separated for the different scenarios with *Front* and *Back* having azimuthal resolutions of $\Delta\varphi = 5^\circ$, and *Front 15°* having an azimuthal resolution of $\Delta\varphi = 15^\circ$.

virtual sound sources at measured and interpolated directions has proven the validity of parameter interpolation for static virtual sound sources.

In addition to the localization accuracy evaluated in the first listening, a second listening test focuses on the perceived externalization of static virtual sound sources. The results indicate that missing room effects inside short HRIRs drastically reduce the perceived externalization. By adding simulated reflections via ISM, this loss in perceived externalization can be compensated. Thus, also for binaural synthesis using parametric IIR filter cascades, appending a room simulator enables externalized static virtual sound sources.

In contrast to the previous listening tests, a third listening test evaluates moving virtual sound sources rather than static ones. Similar audio quality ratings between FIR filter representations of minimum-phase approximated HRIRs and smoothed parallel parametric IIR filter cascades indicate the validity of the parameter interpolation for moving virtual sound sources. Except for broadband noise stimuli, also parallel parametric IIR filter cascades using the cross-fading input-switching combination approach show very good audio quality ratings.

Overall, the results of all three listening tests evidence the usability of the proposed parametric IIR filter cascades in binaural synthesis through headphones for static as well as moving virtual sound sources.

CHAPTER 7

# Conclusion

In binaural synthesis through headphones, monaural audio signals are filtered by measured HRIRs in order to generate binaural signals that contain localization cues used by human beings during natural hearing. These cues include ITD and ILD for horizontal sound source localization, and monaural spectral cues as well as characteristic peaks and notches inside the frequency spectrum for vertical sound source localization. By using BRIRs, DRRs give cues for distance perception. Additionally, HpEq can be used to compensate for the influence of the headphone during playback.

In this work, a binaural synthesis implementation is presented using HRTFs approximated by cascades of parametric IIR filters combined with delays representing the ITD. By using low-order parametric IIR filters, memory requirements of the used hardware can be decreased to three parameters per filter stage (cut-off or center frequency, gain, and Q-factor). In Chapter 3, a two-step HRTF magnitude response approximation procedure is described. In a first step, the individual filter stages are consecutively integrated, initialized and tuned. Here, the remaining approximation error is used as basis for initializing and updating the next filter stage. In a second step, the interaction between filter stages in the cascade is post-optimized based on the Levenberg-Marquardt algorithm. For 84.7 % of the HRTFs from the CIPIC database, using one LFS, one HFS, and ten peak filters is sufficient to produce an approximation error that falls within a 2 dB tolerance. Especially for ipsilateral directions, ten peak filters are sufficient for accurate approximation results. Thus, this work uses cascades of one LFS, one HFS, and ten peak filters to approximate the HRTF magnitude

responses. A similar parametric IIR filter cascade using ten peak filters is also used to equalize the maximum magnitude value among a given set of HpTF measurements in order to yield HpEq without strong amplifications of narrow frequency bands. In addition to this approximation procedure, also an approach for HRTF magnitude response approximation based on instantaneous backpropagation is presented. This algorithm uses the gradient flow through the cascade in order to update the control parameters of the individual filter stages. For this, local gradients of the filter outputs with respect to the filter parameters are derived in Chapter 3.

Since, in practical implementations, HRTFs are only available for a finite number of directions, spatial interpolation is used to enhance the spatial resolution of the measurement grid and enable smooth transitions between directional filters inside moving virtual sound sources. For FIR filter implementations, bilinear rectangular or triangular interpolation can be used to calculate the HRIRs of intermediate directions from the minimum-phase approximated HRIRs of neighboring directions. Contrarily, interpolating IIR filters is not as straightforward as FIR filter interpolation due to further restrictions that have to be taken into account, e.g. stability of the interpolated filters. In Chapter 4, an algorithm for spatial interpolation of parametric IIR filter cascades is proposed. This algorithm calculates intermediate magnitude responses by interpolating the parameters of neighboring directions guaranteeing the stability of the second-order peak filters implicitly. Although simple bilinear rectangular interpolation of parameters is able to calculate intermediate magnitude responses, an assignment of peak filters of neighboring directions is required in order to generate more accurate interpolation results. Therefore, an extended parameter interpolation algorithm is proposed that uses the peak filters of the closest direction as reference for which related peak filters are found in the other neighboring directions. For moving virtual sound sources, smooth transitions between magnitude responses of intermediate directions are required. Since especially time-variant IIR filter implementations suffer from audible clicks due to a mismatch between updated coefficients and internal states of the recursive parts, two cascades of parametric IIR filters are connected in parallel following a cross-fading input-switching combination approach. A further improvement of transition is achieved by reducing the interpolation weight for the gain of a peak filter when moving towards neighboring directions that do not contain that peak filter, such that the peak filter disappears before changing the reference direction. This extension has shown to reduce the spectral coloration when changing the reference direction.

Similar to short HRIRs, HRTFs approximated by parametric IIR filter cascades contain no room effects, which are given as the most significant factors for successful externalization of virtual sound sources. Thus, in Chapter 5, the parametric IIR filter cascades are enriched with simulated room effects via ISM. Although the ISM yields simulated reflections with

proper delays and amplitudes, these reflections lack of a reproduction of directional information. While filtering the reflections with the corresponding HRTFs would simulate BRIRs needed for a complete reproduction of the signals at the human ears during natural hearing inside a room, filtering every impinging reflection with the corresponding HRTF would highly increase the computational complexity. Thus, spatially separating the microphones for the two ears in the ISM algorithm is a trade-off to achieve different RIRs for the two ears. In order to yield similar DRRs inside the combined impulse responses than for measured BRIRs, an algorithm is proposed, which scales the simulated RIRs before combining them with the parametric IIR filter cascades in a parallel structure.

For evaluating the proposed methods, Chapter 6 presents three listening tests assessing different aspects of binaural synthesis using HRTFs approximated by parametric IIR filters, namely localization accuracy, externalization, and audio quality of moving virtual sound sources. In a first listening test, the localization accuracy of the proposed parametric IIR filter cascades is compared to localization results achieved using HRIRs represented as FIR filters. Similarities in the mean angular errors and front/back confusion rates achieved by the two representations confirm the validity of the parametric IIR filter cascades for measured directions. Additionally, a comparison of virtual sound sources at measured and interpolated directions shows comparable localization results in mean angular error and front/back confusion rate. Thus, also the validity of parameter interpolation for static virtual sound sources is proven. In a second listening test, the perceived externalization of static virtual sound sources is evaluated. Missing room effects inside parametric IIR filter cascades drastically reduce the perceived externalization. By adding simulated reflections via ISM, the perceived externalization is increased up to externalization levels achieved using measured BRIRs represented as FIR filters. In a third listening test, the audio quality of moving virtual sound sources generated using minimum-phase approximated HRIRs represented as FIR filters and parametric IIR filter cascades is evaluated. Using only a single parametric IIR filter cascade or two parallel IIR filter cascades without fading produces audible clicks in the filtered audio signals. However, using two IIR filters in parallel following the cross-fading input-switching combination approach shows comparable audio quality ratings than the FIR implementation using minimum-phase approximated HRIRs. This result indicates the validity of the parameter interpolation also for moving virtual sound sources.

Overall, the listening test results validate the proposed offline binaural synthesis implementation using HRTFs and HpEq represented as parametric IIR filter cascades, parameter interpolation to calculate HRTFs of intermediate directions for generating static as well as moving virtual sound sources, and simulated room effects in order to increase the perceived externalization. Thus, HRTFs approximated by parametric IIR filter

cascades can be used to reduce the number of saved coefficients. By using two first-order shelving filters, ten second-order peak filters, a mean HRTF magnitude value, and an extracted ITD, only 36 parameters have to be saved per HRTF instead of 200 coefficients as in FIR filter implementations using conventional HRIRs.

## 7.1 Further Research

Based on the outcome of this work and observations gathered during its presentation, suggestions for further research are given. Firstly, the post-optimization of the parametric IIR filter approximations can be improved by constraining the parameters inside the Levenberg-Marquardt algorithm. In this way, peak filters with extreme gains and Q-factors can be circumvented. Additionally, the accuracy of the HRTF magnitude response approximation via backpropagation algorithm can be enhanced by improving the update of the shelving filters, which has shown to introduce inaccuracies for some evaluated directions especially for LFS.

Secondly, structured peak filters inside fixed frequency bands can be used in order to simplify the assignment during parameter interpolation. For this purpose, magnitude responses of HRTFs can be separated into fixed frequency bands before approximating them by parametric IIR filters. Afterwards, a single peak filter can be used inside every frequency band to approximate the desired magnitude response.

Thirdly, a real-time application of binaural synthesis through headphones using HRTFs approximated by parametric IIR filter cascades can be implemented to build a demonstrator including a head tracker for capturing head motions of the subject. Hence, efficient implementations of the peak filter assignment inside parameter interpolation as well as the calculation of IIR filter coefficients from the corresponding parameters have to be examined. A solution for implementing the parameter interpolation in real-time can be the saving of previously performed peak filter assignments. While this offline assignment would slightly increase the memory requirements, the number of calculations that have to be done in real-time during parameter interpolation can be decreased. Otherwise, an efficient online peak filter assignment has to be implemented. Since MATLAB is not optimized for real-time audio applications, a real-time application requires an implementation done with another programming language, e.g. C or C++.

Fourthly, although the ISM has shown to generate simulated RIRs that are able to increase the perceived externalization, the high number of calculations inside the ISM constraints its usage in real-time applications. Thus, an alternative room simulator should be found that requires less computations per output sample.

# Bibliography


[Ajdler et al., 2005] Ajdler, T., Faller, C., Sbaiz, L., and Vetterli, M. (2005). Interpolation of head related transfer functions considering acoustics. In *Audio Engineering Society Convention 118*, Barcelona, Spain.

[Algazi and Duda, 2011] Algazi, V. R. and Duda, R. O. (2011). Headphone-based spatial sound. *IEEE Signal Processing Magazine*, 28(1):33–42.

[Algazi et al., 2001a] Algazi, V. R., Avendano, C., and Duda, R. O. (2001a). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122.

[Algazi et al., 2001b] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001b). The CIPIC HRTF database. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA.

[Algazi et al., 2002] Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002). Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5):2053–2064.

[Allen and Berkley, 1979] Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

[António et al., 2002] António, J., Godinho, L., and Tadeu, A. (2002). Reverberation times obtained using a numerical model versus those given by simplified formulas and measurements. *Acta Acustica united with Acustica*, 88(2):252–261.

[Back and Tsoi, 1991] Back, A. D. and Tsoi, A. C. (1991). FIR and IIR synapses, a new neural network architecture for time series modeling. *Neural Computation*, 3(3):375–385.

[Begault, 1992] Begault, D. R. (1992). Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society*, 40(11):895–904.

[Begault, 1994] Begault, D. R. (1994). *3-D Sound for Virtual Reality and Multimedia*. Academic Press Professional, Inc., San Diego, CA, USA.

[Begault et al., 2001] Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916.

[Behrends et al., 2011] Behrends, H., von dem Knesebeck, A., Bradinal, W., Neumann, P., and Zölzer, U. (2011). Automatic equalization using parametric IIR filters. *Journal of the Audio Engineering Society*, 59(3):102–109.

[Belloch et al., 2020] Belloch, J. A., Ramos, G., Badia, J. M., and Cobos, M. (2020). An efficient implementation of parallel parametric HRTF models for binaural sound synthesis in mobile multimedia. *IEEE Access*, 8:49562–49573.

[Bhattacharya et al., 2020] Bhattacharya, P., Nowak, P., and Zölzer, U. (2020). Optimization of cascaded parametric peak and shelving filters with backpropagation algorithm. In *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20)*, Vienna, Austria.

[Bilinski et al., 2014] Bilinski, P., Ahrens, J., Thomas, M. R. P., Tashev, I. J., and Platt, J. C. (2014). HRTF magnitude synthesis via sparse representation of anthropometric features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy.

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

[Blauert and Braasch, 2008] Blauert, J. and Braasch, J. (2008). Räumliches Hören. In Weinzierl, S., editor, *Handbuch der Audiotechnik*, chapter 3, pages 87–121. Springer, Berlin, Heidelberg.

[Blommer and Wakefield, 1997] Blommer, M. A. and Wakefield, G. H. (1997). Pole-zero approximations for head-related transfer functions using a logarithmic error criterion. *IEEE Transactions on Speech and Audio Processing*, 5(3):278–287.

[Bomhardt et al., 2016] Bomhardt, R., Braren, H., and Fels, J. (2016). Individualization of head-related transfer functions using principal component analysis and anthropometric dimensions. *Proceedings of Meetings on Acoustics*, 29(1).

[Botts et al., 2013] Botts, J., Escolano, J., and Xiang, N. (2013). Design of IIR filters with Bayesian model selection and parameter estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):669–674.

[Breebaart et al., 2009] Breebaart, J., Nater, F., and Kohlrausch, A. (2009). Parametric binaural synthesis: Background, applications and standards. In *Proceedings of the International Conference on Acoustics - NAG/DAGA 2009*, Rotterdam, Netherlands.

[Brungart, 2002] Brungart, D. S. (2002). Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environments*, 11(1):93–106.

[Brungart and Rabinowitz, 1999] Brungart, D. S. and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479.

[Bücklein, 1981] Bücklein, R. (1981). The audibility of frequency response irregularities. *Journal of the Audio Engineering Society*, 29(3):126–131.

[Burkhard and Sachs, 1975] Burkhard, M. D. and Sachs, R. M. (1975). Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 58(1):214–222.

[Caracalla and Roebel, 2017] Caracalla, H. and Roebel, A. (2017). Gradient conversion between time and frequency domains using Wirtinger calculus. In *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, Scotland.

[Chun et al., 2017] Chun, C. J., Moon, J. M., Lee, G. W., Kim, N. K., and Kim, H. K. (2017). Deep neural network based HRTF personalization using anthropometric measurements. In *Audio Engineering Society Convention 143*, New York, NY, USA.

[de Sousa and Queiroz, 2009] de Sousa, G. H. M. and Queiroz, M. (2009). Two approaches for HRTF interpolation. In *Proceedings of the 12th Brazilian Symposium on Computer Music*, Recife, Brazil.

[Duda and Martens, 1998] Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058.

[Duda et al., 1999] Duda, R. O., Avendano, C., and Algazi, V. R. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, USA.

[Duraiswaini et al., 2004] Duraiswaini, R., Zotkin, D. N., and Gumerov, N. A. (2004). Interpolation and range extrapolation of HRTFs [head related transfer functions]. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, QC, Canada.

[Durlach et al., 1992] Durlach, N., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., and Wenzel, E. M. (1992). On the externalization of auditory images. *Presence: Teleoperators and Virtual Environments*, 1(2):251–257.

[Eichas and Zölzer, 2018] Eichas, F. and Zölzer, U. (2018). Gray-box modeling of guitar amplifiers. *Journal of the Audio Engineering Society*, 66(12):1006–1015.

[Eyring, 1930] Eyring, C. F. (1930). Reverberation time in "dead" rooms. *The Journal of the Acoustical Society of America*, 1(2A):217–241.

[Farina, 2000] Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*, Paris, France.

[Farina, 2007] Farina, A. (2007). Advancements in impulse response measurements by sine sweeps. In *Audio Engineering Society Convention 122*, Vienna, Austria.

[Fletcher and Sivian, 1927] Fletcher, H. and Sivian, L. J. (1927). Binaural telephone system. United States Patent 1,624,486.

[Frank and Zotter, 2018] Frank, M. and Zotter, F. (2018). Simple reduction of front-back confusion in static binaural rendering. In *Fortschritte der Akustik: DAGA 2018*, Munich, Germany.

[Freeland et al., 2002] Freeland, F. P., Biscainho, L. W. P., and Diniz, P. S. R. (2002). Efficient HRTF interpolation in 3D moving sound. In *22nd AES International Conference on Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland.

[Freeland et al., 2004] Freeland, F. P., Biscainho, L. W. P., and Diniz, P. S. R. (2004). Interpositional transfer function for 3D-sound generation. *Journal of the Audio Engineering Society*, 52(9):915–930.

[Gamper, 2013] Gamper, H. (2013). Head-related transfer function interpolation in azimuth, elevation, and distance. *The Journal of the Acoustical Society of America*, 134(6):EL547–EL553.

[Gardner, 1969] Gardner, M. B. (1969). Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45(1):47–53.

[Gardner, 1973] Gardner, M. B. (1973). Some monaural and binaural facets of median plane localization. *The Journal of the Acoustical Society of America*, 54(6):1489–1495.

[Gardner and Martin, 1994] Gardner, B. and Martin, K. (1994). HRTF measurements of a KEMAR dummy-head microphone. Technical Report 280, MIT Media Lab Perceptual Computing.

[Geronazzo et al., 2014] Geronazzo, M., Spagnol, S., Bedin, A., and Avanzini, F. (2014). Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy.

[Gerzon, 1971] Gerzon, M. A. (1971). Synthetic stereo reverberation: Part one. *Studio Sound*, 13:632–635.

[Gerzon, 1992] Gerzon, M. A. (1992). The design of distance panpots. In *Audio Engineering Society Convention 92*, Vienna, Austria.

[Griesinger, 2016] Griesinger, D. (2016). Playback of non-individual binaural recordings without head tracking, and its potential for archiving and analyzing concert hall acoustics. In *Proceedings of the 22nd International Congress on Acoustics (ICA)*, Buenos Aires, Argentina.

[Hammershøi and Møller, 1996] 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):408–427.

[Haraszy et al., 2010] Haraszy, Z., Cristea, D.-G., Tiponut, V., and Slavici, T. (2010). Improved head related transfer function generation and testing for acoustic virtual reality development. In *Proceedings of the 14th WSEAS International Conference on Systems: Part of the 14th WSEAS CSCC Multiconference - Volume II*, Corfu, Greece.

[Hartmann, 1999] Hartmann, W. M. (1999). How we localize sound. *Physics Today*, 52(11):24–29.

[Hartmann and Wittenberg, 1996] Hartmann, W. M. and Wittenberg, A. (1996). On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–3688.

[Hasegawa et al., 2000] Hasegawa, H., Kasuga, M., Matsumoto, S., and Koike, A. (2000). Simply realization of sound localization using HRTF approximated by IIR filter. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E83-A(6):973–978.

[Henning, 1974] Henning, G. B. (1974). Detectability of interaural delay in high-frequency complex waveforms. *The Journal of the Acoustical Society of America*, 55(1):84–90.

[Hiipakka, 2012] Hiipakka, M. (2012). *Estimating pressure at the eardrum for binaural reproduction*. Phd thesis, Aalto University, Espoo, Finland.

[Holters et al., 2009] Holters, M., Corbach, T., and Zölzer, U. (2009). Impulse response measurement techniques and their applicability in the real world. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy.

[Hospitalier, 1881] Hospitalier, E. (1881). The telephone at the paris opera. *Scientific American*, 45:422–423.

[Hugeng et al., 2015] Hugeng, H., Laya, F., Wahidin, W., and Gunawan, D. (2015). Implementation of HRIR interpolations on DSP Board TMS320C5535 eZdsp$^{TM}$. In *Proceedings of the 14th International Conference on Quality in Research (QiR)*, Lombok, Indonesia.

[Hugeng et al., 2017] Hugeng, H., Anggara, J., and Gunawan, D. (2017). Implementation of 3D HRTF interpolation in synthesizing virtual 3D moving sound. *International Journal of Technology (IJTech)*, 8(1):184–193.

[Huopaniemi, 1999] Huopaniemi, J. (1999). *Virtual Acoustics and 3-D Sound in Multimedia Signal Processing*. PhD thesis, Helsinki University of Technology, Espoo, Finland.

[Huopaniemi et al., 1998] Huopaniemi, J., Zacharov, N., and Karjalainen, M. (1998). Objective and subjective evaluation of head-related-transfer function filter design. In *Audio Engineering Society Convention 105*, San Francisco, CA, USA.

[Huttunen et al., 2014] Huttunen, T., Vanne, A., Harder, S., Paulsen, R. R., King, S., Perry-Smith, L., and Kärkkäinen, L. (2014). Rapid generation of personalized HRTFs. In *55th International Conference on Spatial Audio 2014*, Helsinki, Finland.

[Jot and Chaigne, 1991] Jot, J.-M. and Chaigne, A. (1991). Digital delay networks for designing artificial reverberators. In *Audio Engineering Society Convention 94*, Berlin, Germany.

[Jot et al., 1995] Jot, J.-M., Larcher, V., and Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. In *Audio Engineering Society Convention 98*, Paris, France.

[Kan et al., 2009] Kan, A., Jin, C., and van Schaik, A. (2009). A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. *The Journal of the Acoustical Society of America*, 125(4):2233–2242.

[Karjalainen and Järveläinen, 2007] Karjalainen, M. and Järveläinen, H. (2007). Reverberation modeling using velvet noise. In *30th International Conference on Intelligent Audio Environments*, Saariselkä, Finland.

[Katz and Nicol, 2018] Katz, B. F. G. and Nicol, R. (2018). Binaural spatial reproduction. In Zacharov, N., editor, *Sensory Evaluation of Sound*, chapter 11, pages 349–388. CRC Press, Boca Raton, FL, USA.

[Katz and Noisternig, 2014] Katz, B. F. G. and Noisternig, M. (2014). A comparative study of interaural time delay estimation methods. *The Journal of the Acoustical Society of America*, 135(6):3530–3540.

[Katz and Parseihian, 2012] Katz, B. F. G. and Parseihian, G. (2012). Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, 131(2):EL99–EL105.

[Kendall and Martens, 1984] Kendall, G. S. and Martens, W. L. (1984). Simulating the cues of spatial hearing in natural environments. In *Proceedings of the International Computer Music Conference*, Paris, France.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, L. J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

[Kistler and Wightman, 1992] 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):1637–1647.

[Klumpp and Eady, 1956] 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):859–860.

[Kopčo and Shinn-Cunningham, 2011] Kopčo, N. and Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *The Journal of the Acoustical Society of America*, 130(3):1530–1541.

[Krokstad et al., 1968] Krokstad, A., Strøm, S., and Sørsdal, S. (1968). Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125.

[Kuhn, 1977] Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(1):157–167.

[Kuhn, 1987] Kuhn, G. F. (1987). Physical acoustics and measurements pertaining to directional hearing. In Yost, W. A. and Gourevitch, G., editors, *Directional Hearing*, chapter 1, pages 3–25. Springer US, New York, NY, USA.

[Kulkarni and Colburn, 2004] Kulkarni, A. and Colburn, H. S. (2004). Infinite-impulse-response models of the head-related transfer function. *The Journal of the Acoustical Society of America*, 115(4):1714–1728.

[Kulkarni et al., 1995] Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (1995). On the minimum-phase approximation of head-related transfer functions. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Accoustics*, New Paltz, NY, USA.

[Kuttruff, 2000] Kuttruff, H. (2000). *Room Acoustics*. Spon Press, Taylor & Francis Group, London, UK, 4th edition.

[Leclère et al., 2019] Leclère, T., Lavandier, M., and Perrin, F. (2019). On the externalization of sound sources with headphones without reference to a real source. *The Journal of the Acoustical Society of America*, 146(4):2309–2320.

[Lehmann and Johansson, 2008] Lehmann, E. A. and Johansson, A. M. (2008). Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America*, 124(1):269–277.

[Lehmann et al., 2007] Lehmann, E. A., Johansson, A. M., and Nordholm, S. (2007). Reverberation-time prediction method for room impulse responses simulated with the image-source model. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA.

[Lehnert, 1993] Lehnert, H. (1993). Systematic errors of the ray-tracing algorithm. *Applied Acoustics*, 38(2):207 – 221.

[Levenberg, 1944] Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168.

[Li and Peissig, 2017] Li, S. and Peissig, J. (2017). Fast estimation of 2D individual HRTFs with arbitrary head movements. In *22nd International Conference on Digital Signal Processing (DSP)*, London, UK.

[Li et al., 2018] Li, S., Schlieper, R., and Peissig, J. (2018). The effect of variation of reverberation parameters in contralateral versus ipsilateral ear signals on perceived externalization of a lateral sound source in a listening room. *The Journal of the Acoustical Society of America*, 144(2):966–980.

[Li et al., 2019] Li, S., Schlieper, R., and Peissig, J. (2019). The impact of head movement on perceived externalization of a virtual sound source with different BRIR lengths. In *AES International Conference on Immersive and Interactive Audio*, York, UK.

[Lindau and Brinkmann, 2012] Lindau, A. and Brinkmann, F. (2012). Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. *Journal of the Audio Engineering Society*, 60(1/2):54–62.

[Lindau et al., 2010] Lindau, A., Estrella, J., and Weinzierl, S. (2010). Individualization of dynamic binaural synthesis by real time manipulation of the ITD. In *Audio Engineering Society Convention 128*, London, UK.

[Liski et al., 2017] Liski, J., Välimäki, V., Vesa, S., and Väänänen, R. (2017). Real-time adaptive equalization for headphone listening. In *25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece.

[Mackensen et al., 1998] Mackensen, P., Reichenauer, K., and Theile, G. (1998). Einfluß der spontanen Kopfdrehungen auf die Lokalisation beim binauralen Hören. In *Bericht zur 20.Tonmeistertagung*, Karlsruhe, Germany.

[Macpherson and Middlebrooks, 2002] Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236.

[Majdak et al., 2007] Majdak, P., Balasz, 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):623–637.

[Marquardt, 1963] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441.

[Masiero and Fels, 2011] Masiero, B. and Fels, J. (2011). Perceptually robust headphone equalization for binaural reproduction. In *Audio Engineering Society Convention 130*, London, UK.

[Matsumoto et al., 2004] Matsumoto, M., Yamanaka, S., Toyama, M., and Nomura, H. (2004). Effect of arrival time correction on the accuracy of binaural impulse response interpolation–interpolation methods of binaural response. *Journal of the Audio Engineering Society*, 52(1/2):56–61.

[Middlebrooks, 1992] Middlebrooks, J. C. (1992). Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5):2607–2624.

[Middlebrooks et al., 1989] Middlebrooks, J. C., Makous, J. C., and Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *The Journal of the Acoustical Society of America*, 86(1):89–108.

[Mills, 1958] Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246.

[Mills, 1960] Mills, A. W. (1960). Lateralization of high-frequency tones. *The Journal of the Acoustical Society of America*, 32(1):132–134.

[Minnaar et al., 1999] Minnaar, P., Christensen, F., Møller, H., Olesen, S. K., and Plogsties, J. (1999). Audibility of all-pass components in binaural synthesis. In *Audio Engineering Society Convention 106*, Munich, Germany.

[Møller, 1992] Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics*, 36(3):171 – 218.

[Møller et al., 1996] Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469.

[Moorer, 1979] Moorer, J. A. (1979). About this reverberation business. *Computer Music Journal*, 3(2):13–28.

[Mróz et al., 2018] Mróz, B., Zotter, F., Wendt, F., Zaunschirm, M., and Frank, M. (2018). Experiment on externalization in binaural directional-source auralization. In *Fortschritte der Akustik: DAGA 2018*, Munich, Germany.

[Nam et al., 2008] Nam, J., Kolar, M. A., and Abel, J. S. (2008). On the minimum-phase nature of head-related transfer functions. In *Audio Engineering Society Convention 125*, San Francisco, CA, USA.

[Neubauer and Kostek, 2001] Neubauer, R. and Kostek, B. (2001). Prediction of the reverberation time in rectangular rooms with non-uniformly distributed sound absorption. *Archives of Acoustics*, 26(3):183–202.

[Nicol, 2010] Nicol, R. (2010). *Binaural Technology*. AES Monograph, Audio Engineering Society, New York, NY, USA.

[Nowak and Zölzer, 2022] Nowak, P. and Zölzer, U. (2022). Spatial interpolation of HRTFs approximated by parametric IIR filters. In *Fortschritte der Akustik: DAGA 2022*, Stuttgart, Germany.

[Nowak et al., 2018a] Nowak, P., Zimpfer, V., and Zölzer, U. (2018a). 3D virtual audio with headphones: A literature review of the last ten years. In *Fortschritte der Akustik: DAGA 2018*, Munich, Germany.

[Nowak et al., 2018b] Nowak, P., Zimpfer, V., and Zölzer, U. (2018b). Methods for solving front-back confusions in 3D spatial audio headphones. In *6th Workshop on Battlefield Acoustics*, Saint-Louis, France.

[Nowak et al., 2020a] Nowak, P., Zimpfer, V., and Zölzer, U. (2020a). Automatic approximation of head-related transfer functions using parametric IIR filters. In *Fortschritte der Akustik: DAGA 2020*, Hannover, Germany.

[Nowak et al., 2020b] Nowak, P., Zimpfer, V., and Zölzer, U. (2020b). Spatial audio through headphones based on HRTFs approximated by parametric IIR filters. In *7th Workshop on Battlefield Acoustics*, Saint-Louis, France.

[Oberem et al., 2016] Oberem, J., Masiero, B., and Fels, J. (2016). Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods. *Applied Acoustics*, 114:71–78.

[Parviainen and Pertilä, 2017] Parviainen, M. and Pertilä, P. (2017). Obtaining an optimal set of head-related transfer functions with a small amount of measurements. In *Proceedings of the IEEE International Workshop on Signal Processing Systems (SiPS)*, Lorient, France.

[Paul, 2009] Paul, S. (2009). Binaural recording technology: A historical review and possible future developments. *Acta Acustica united with Acustica*, 95:767–788.

[Perrott, 1984] Perrott, D. R. (1984). Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity. *The Journal of the Acoustical Society of America*, 75(4):1201–1206.

[Plenge, 1974] Plenge, G. (1974). On the differences between localization and lateralization. *The Journal of the Acoustical Society of America*, 56(3):944–951.

[Plogsties et al., 2000] Plogsties, J., Minnaar, P., Christensen, F., Olesen, S. K., and Møller, H. (2000). The directional resolution needed when measuring head-related transfer functions. In *Fortschritte der Akustik: DAGA 2000*, Oldenburg, Germany.

[Queiroz and de Sousa, 2010] Queiroz, M. and de Sousa, G. H. M. (2010). Structured IIR models for HRTF interpolation. In *Proceedings of the 36th International Computer Music Conference (ICMC)*, New York, NY, USA.

[Ramos and Cobos, 2013] Ramos, G. and Cobos, M. (2013). Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications. *The Journal of the Acoustical Society of America*, 134(3):1735–1738.

[Ramos and López, 2006] Ramos, G. and López, J. J. (2006). Filter design method for loudspeaker equalization based on IIR parametric filters. *Journal of the Audio Engineering Society*, 54(12):1162–1178.

[Ramos et al., 2017] Ramos, G., Cobos, M., Bank, B., and Belloch, J. A. (2017). A parallel approach to HRTF approximation and interpolation based on a parametric filter model. *IEEE Signal Processing Letters*, 24(10):1507–1511.

[Ranjan and Gan, 2015] Ranjan, R. and Gan, W.-S. (2015). Natural listening over headphones in augmented reality using adaptive filtering techniques. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1988–2002.

[Rayleigh, 1907] Rayleigh, L. (1907). XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232.

[Reddy and Hegde, 2015] Reddy, C. S. and Hegde, R. M. (2015). A joint sparsity and linear regression based method for customization of median plane HRIR. In *49th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA.

[Reddy and Hegde, 2016] Reddy, C. S. and Hegde, R. M. (2016). Horizontal plane HRTF interpolation using linear phase constraint for rendering spatial audio. In *24th European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary.

[Richter et al., 2016] Richter, J.-G., Behler, G., and Fels, J. (2016). Evaluation of a fast HRTF measurement system. In *Audio Engineering Society Convention 140*, Paris, France.

[Rivera Benois et al., 2016] Rivera Benois, P., Bhattacharya, P., and Zölzer, U. (2016). Derivation technique for headphone transfer functions based on sine sweeps and least squares minimization. In *Proceedings of the 45th International Congress and Exposition on Noise Control Engineering (INTER-NOISE)*, Hamburg, Germany.

[Rubak and Johansen, 1998] Rubak, P. and Johansen, L. G. (1998). Artificial reverberation based on a pseudo-random impulse response. In *Audio Engineering Society Convention 104*, Amsterdam, Netherlands.

[Rubak and Johansen, 1999] Rubak, P. and Johansen, L. G. (1999). Artificial reverberation based on a pseudo-random impulse response II. In *Audio Engineering Society Convention 106*, Munich, Germany.

[Runkle et al., 1995] Runkle, P. R., Blommer, M. A., and Wakefield, G. H. (1995). A comparison of head related transfer function interpolation methods. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Accoustics*, New Paltz, New York, USA.

[Sabine, 1922] Sabine, W. C. (1922). *Collected Papers on Acoustics*. Harvard University Press, Cambridge, MA, USA.

[Savioja et al., 1999] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999). Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705.

[Schärer and Lindau, 2009] Schärer, Z. and Lindau, A. (2009). Evaluation of equalization methods for binaural signals. In *Audio Engineering Society Convention 126*, Munich, Germany.

[Schlecht, 2020] Schlecht, S. J. (2020). FDNTB: The feedback delay network toolbox. In *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20)*, Vienna, Austria.

[Schmidt and Hudde, 2009] Schmidt, S. and Hudde, H. (2009). Accuracy of acoustic ear canal impedances: Finite element simulation of measurement methods using a coupling tube. *The Journal of the Acoustical Society of America*, 125(6):3819–3827.

[Schroeder, 1962] Schroeder, M. R. (1962). Natural sounding artificial reverberation. *Journal of the Audio Engineering Society*, 10(3):219–223.

[Schroeder and Logan, 1961] Schroeder, M. R. and Logan, B. F. (1961). "Colorless" artificial reverberation. *IRE Transactions on Audio*, AU-9(6):209–214.

[Schroeder et al., 1974] Schroeder, M. R., Gottlob, D., and Siebrasse, K. F. (1974). Comparative study of european concert halls: Correlation of subjective preference with geometric and acoustic parameters. *The Journal of the Acoustical Society of America*, 56(4):1195–1201.

[Seeber, 2003] Seeber, B. (2003). *Untersuchung der auditiven Lokalisation mit einer Lichtzeigermethode*. Phd thesis, Technische Universität München, Munich, Germany.

[Shinn-Cunningham et al., 2000] Shinn-Cunningham, B. G., Santarelli, S., and Kopčo, N. (2000). Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107(3):1627–1636.

[Simon et al., 2016] Simon, L. S. R., Zacharov, N., and Katz, B. F. G. (2016). Perceptual attributes for the comparison of head-related transfer functions. *The Journal of the Acoustical Society of America*, 140(5):3623–3632.

[Sridhar and Choueiri, 2017] Sridhar, R. and Choueiri, E. (2017). A method for efficiently calculating head-related transfer functions directly from head scan point clouds. In *Audio Engineering Society Convention 143*, New York, NY, USA.

[Stautner and Puckette, 1982] Stautner, J. and Puckette, M. (1982). Designing multi-channel reverberators. *Computer Music Journal*, 6(1):52–65.

[Stevens and Newman, 1936] Stevens, S. S. and Newman, E. B. (1936). The localization of actual sources of sound. *The American Journal of Psychology*, 48(2):297–306.

[Tashev, 2014] Tashev, I. (2014). HRTF phase synthesis via sparse representation of anthropometric features. In *Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA.

[Theile, 2016] Theile, G. (2016). Equalization of studio monitor headphones. In *AES International Conference on Headphone Technology*, Aalborg, Denmark.

[Usami and Kato, 1978] Usami, N. and Kato, T. (1978). Headphone unit incorporating microphones for binaural recording. United States Patent 4,088,849.

[Välimäki, 1995] Välimäki, V. (1995). *Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters*. Phd thesis, Helsinki University of Technology, Espoo, Finland.

[Välimäki and Laakso, 1998] Välimäki, V. and Laakso, T. I. (1998). Suppression of transients in variable recursive digital filters with a novel and efficient cancellation method. *IEEE Transactions on Signal Processing*, 46(12):3408–3414.

[Välimäki et al., 2012] Välimäki, V., Parker, J. D., Savioja, L., Smith, J. O., and Abel, J. S. (2012). Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448.

[Välimäki et al., 2016] Välimäki, V., Parker, J. D., Savioja, L., Smith, J. O., and Abel, J. S. (2016). More than 50 years of artificial reverberation. In *AES International Conference on Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)*, Leuven, Belgium.

[Verhelst and Nilens, 1986] Verhelst, W. and Nilens, P. (1986). A modified-superposition speech synthesizer and its applications. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tokyo, Japan.

[Völk, 2009] Völk, F. (2009). Externalization in data-based binaural synthesis: Effects of impulse response length. In *Proceedings of the International Conference on Acoustics - NAG/DAGA 2009*, Rotterdam, Netherlands.

[Völk et al., 2008] Völk, F., Heinemann, F., and Fastl, H. (2008). Externalization in binaural synthesis: Effects of recording environment and measurement procedure. In *Proceedings of the Acoustics '08*, Paris, France.

[Wang and Chan, 2015] Wang, Z. and Chan, C.-F. (2015). Continuous function modeling of head-related impulse response. *IEEE Signal Processing Letters*, 22(3):283–287.

[Wang et al., 2008] Wang, L., Yin, F., and Chen, Z. (2008). An indirect interpolation method for head-related transfer function pole-zero models. *Acoustical Science and Technology*, 29(5):329–331.

[Wang et al., 2020] Wang, X., Takaki, S., and Yamagishi, J. (2020). Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415.

[Wenzel, 1996] Wenzel, E. M. (1996). What perception implies about implementation of interactive virtual acoustic environments. In *Audio Engineering Society Convention 101*, Los Angeles, CA, USA.

[Wenzel and Foster, 1993] Wenzel, E. M. and Foster, S. H. (1993). Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA.

[Wenzel et al., 1993] 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):111–123.

[Werner et al., 2016] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K. (2016). A summary on acoustic room divergence and its effect on externalization of auditory events. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal.

[Wightman and Kistler, 1989] Wightman, F. L. and Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867.

[Wightman and Kistler, 1992] Wightman, F. L. and Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661.

[Wightman and Kistler, 1993] Wightman, F. L. and Kistler, D. J. (1993). Sound localization. In Yost, W. A., Popper, A. N., and Fay, R. R., editors, *Human Psychophysics*, chapter 5, pages 155–192. Springer US, New York, NY.

[Wightman and Kistler, 1997] Wightman, F. L. and Kistler, D. J. (1997). Monaural sound localization revisited. *The Journal of the Acoustical Society of America*, 101(2):1050–1063.

[Wightman and Kistler, 1999] Wightman, F. L. and Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853.

[Wightman et al., 1992] Wightman, F. L., Kistler, D. J., and Arruda, M. (1992). Perceptual consequences of engineering compromises in synthesis of virtual auditory objects. *The Journal of the Acoustical Society of America*, 92(4):2332–2332.

[Xie, 2020] Xie, B. (2020). Spatial sound-history, principle, progress and challenge. *Chinese Journal of Electronics*, 29(3):397–416.

[Xu et al., 2007] Xu, S., Li, Z., and Salvendy, G. (2007). Individualization of head-related transfer function for three-dimensional virtual auditory display: A review. In *Virtual Reality*, pages 397–407. Springer, Berlin, Heidelberg.

[Yao and Chen, 2013] Yao, S.-N. and Chen, L. J. (2013). HRTF adjustments with audio quality assessments. *Archives of Acoustics*, 38(1):55–62.

[Zahorik et al., 2005] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420.

[Zeng et al., 2010] Zeng, X.-Y., Wang, S.-G., and Gao, L.-P. (2010). A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures. *Journal of Sound and Vibration*, 329(19):4093 – 4106.

[Zhu et al., 2017] Zhu, M., Shahnawaz, M., Tubaro, S., and Sarti, A. (2017). HRTF personalization based on weighted sparse representation of anthropometric features. In *Proceedings of the International Conference on 3D Immersion (IC3D)*, Brussels, Belgium.

[Zölzer, 2008] Zölzer, U. (2008). *Digital Audio Signal Processing*. John Wiley & Sons Ltd, Chichester, UK, 2nd edition.

[Zotkin et al., 2003] Zotkin, D. Y. N., Hwang, J., Duraiswaini, R., and Davis, L. S. (2003). HRTF personalization using anthropometric measurements. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA.

[Zwislocki and Feldman, 1956] Zwislocki, J. and Feldman, R. S. (1956). Just noticeable differences in dichotic phase. *The Journal of the Acoustical Society of America*, 28(5):860–864.

# Curriculum Vitae

## Personal Details

| | |
|---|---|
| Name | Patrick Nowak |
| Date of Birth | 03/04/1991 |
| Place of Birth | Hamburg, Germany |
| Nationality | German |

## Employment History

| | |
|---|---|
| since 02/2021 | **Research Assistant**<br>Helmut Schmidt University<br>Department of Signal Processing and Communication<br>Hamburg, Germany |
| 12/2016 - 11/2020 | **Research Assistant** in cooperation project between<br>French-German Research Institute of Saint-Louis<br>Acoustic and Soldier Protection Group<br>Saint-Louis, France<br>Helmut Schmidt University<br>Department of Signal Processing and Communication<br>Hamburg, Germany |
| 03/2014 - 08/2014 | **Internship**<br>STILL GmbH<br>Hamburg, Germany |
| 02/2014 | **Internship**<br>Deutsches Elektronen-Synchrotron DESY<br>Hamburg, Germany |
| 01/2014 | **Internship**<br>Pinck Ingenieure Consulting GmbH<br>Hamburg, Germany |

## Academic Career

| | |
|---|---|
| 11/2016 - 05/2022 | **Ph.D. Candidate**<br>Helmut Schmidt University<br>Department of Signal Processing and Communication<br>Hamburg, Germany |
| 10/2014 - 10/2016 | **Master of Science**<br>Electrical Engineering<br>Specialization: Information and Communication Systems<br>Hamburg University of Technology<br>Hamburg, Germany |
| 10/2010 - 06/2014 | **Bachelor of Science**<br>General Engineering Science<br>Specialization: Electrical Engineering<br>Hamburg University of Technology<br>Hamburg, Germany |

## School Education

| | |
|---|---|
| 08/2001 - 06/2010 | **Sankt-Ansgar-Schule**<br>Hamburg, Germany |
| 08/1997 - 06/2001 | **St. Paulus Schule**<br>Hamburg, Germany |