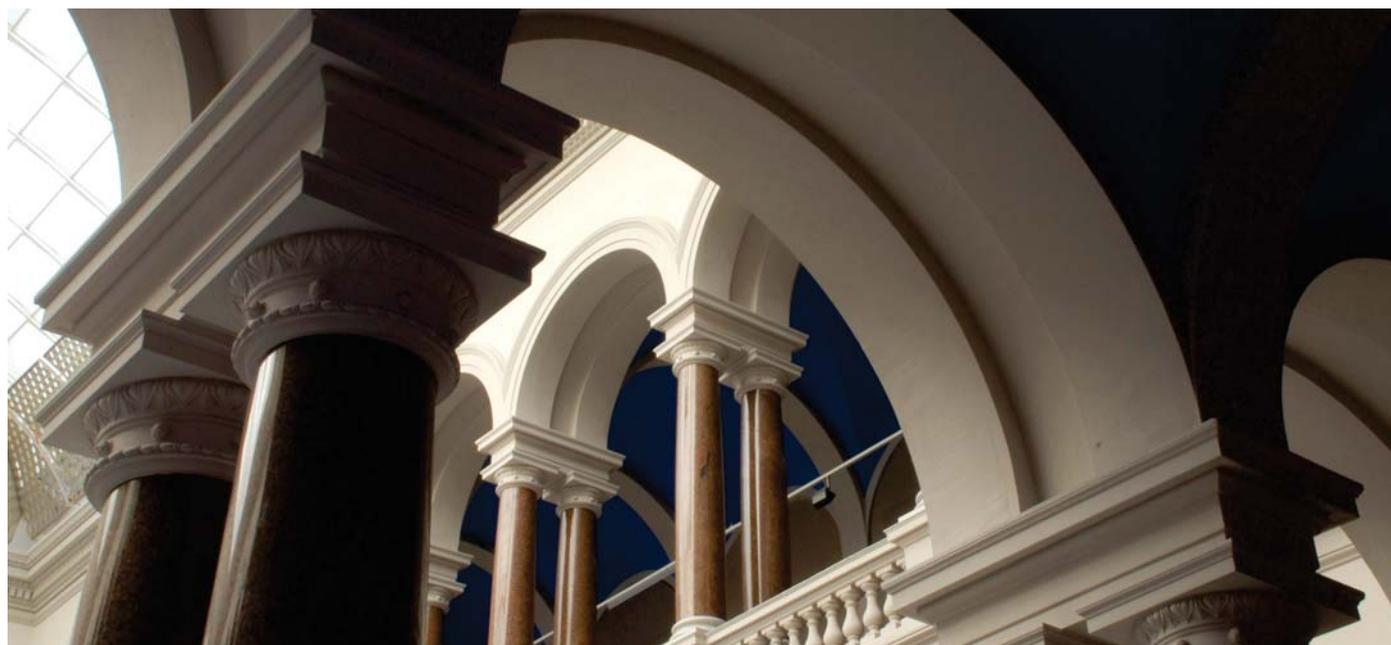


Fabian Brinkmann, Reinhild Roden, Alexander Lindau, Stefan Weinzierl

Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology

Journal article | Accepted manuscript (Postprint)

This version is available at <https://doi.org/10.14279/depositonce-9004>



Brinkmann, F., Roden, R., Lindau, A., & Weinzierl, S. (2015). Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology. *IEEE Journal of Selected Topics in Signal Processing*, 9(5), 931–942. <https://doi.org/10.1109/jstsp.2015.2414905>

Terms of Use

© © 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

WISSEN IM ZENTRUM
UNIVERSITÄTSBIBLIOTHEK

Technische
Universität
Berlin

Audibility and interpolation of head-above-torso orientation in binaural technology

Fabian Brinkmann*, Reinhild Roden, Alexander Lindau, and Stefan Weinzierl

Abstract—Head-related transfer functions (HRTFs) incorporate fundamental cues required for human spatial hearing and are often applied to auralize results obtained from room acoustic simulations. HRTFs are typically available for various directions of sound incidence and a fixed head-above-torso orientation (HATO). If – in interactive auralizations – HRTFs are exchanged according to the head rotations of a listener, the auralization result most often corresponds to a listener turning head *and* torso simultaneously, while – in reality – listeners usually turn their head *independently* above a fixed torso. In the present study, we show that accounting for HATO produces clearly audible differences, thereby suggesting the relevance of correct HATO when aiming at perceptually transparent binaural synthesis. Furthermore, we addressed the efficient representation of variable HATO in interactive acoustic simulations using spatial interpolation. Hereby, we evaluated two different approaches: interpolating between HRTFs with identical torso-to-source but different head-to-source orientations (*head interpolation*) and interpolating between HRTFs with the same head-to-source but different torso-to-source orientations (*torso interpolation*). Torso interpolation turned out to be more robust against increasing interpolation step width. In this case the median threshold of audibility for the head-above-torso resolution was about 25 degrees, whereas with head interpolation the threshold was about 10 degrees. Additionally, we tested a non-interpolation approach (nearest neighbor) as a suitable means for mobile applications with limited computational capacities.

Index Terms—HRTF/HRIR, interpolation, dynamic auralization, psychoacoustics.

I. INTRODUCTION

INTERACTIVE auralization, such as dynamic binaural synthesis, accounts for head rotations of the listener by real-time exchange of corresponding binaural transfer functions. Rendering is often based on sound fields obtained from room acoustic simulations, making it possible to auralize rooms while using arbitrary HRTF sets. Binaural room impulse responses (BRIRs) required for auralization are then obtained by superposition of head-related impulse responses (HRIRs) corresponding to the respective incident angles of direct sound and reflections [1, p. 272]. Interactivity with respect to head rotations fosters a realistic overall impression, helps in resolving front-back confusions [2], and when judging timbre [3]. However, HRTFs usually represent different angles of sound incidence relative to a *fixed* dummy head or human subject. At the reproduction stage, head rotations will thus correspond to a listener moving head *and* torso whereas in a typical situation

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org. The authors are affiliated with the Audio Communication Group, TU Berlin, Einsteinufer 17c, 10587 Berlin, Germany. Correspondence should be addressed to fabian.brinkmann@tu-berlin.de; phon +4930 314 29094.

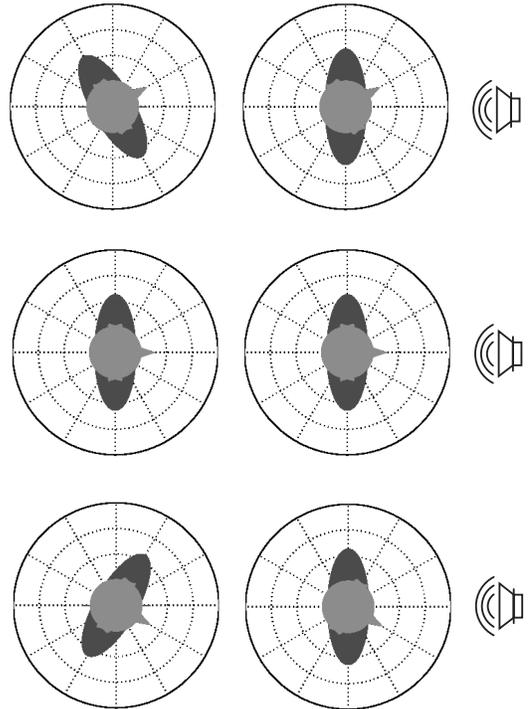


Fig. 1: Illustration of head rotations with constant (left) and variable (right) HATO and head orientations of 30° (top), 0° (middle), and 330° (bottom). HATO is always 0° for the head rotation displayed in left column and otherwise equals the displayed head orientation.

the head is rotated *independently* above a fixed torso (Fig. 1).

The effect of the torso on HRTFs was extensively studied by Algazi et al. [4] for static binaural synthesis and a neutral head-above-torso orientation (HATO). The authors showed that if the torso blocks the direct path from the sound source to the ear, shadowing occurs for frequencies above approximately 100 Hz, causing increasing attenuation of up to 25 dB. For other directions of sound incidence, the torso acts as a reflector causing comb-filters with an amplitude of up to ± 5 dB, whereas the exact positions of peaks and dips of the comb-filter mainly depends on the source elevation. For a source above the listener the first dip occurs already at a frequency as low as 700 Hz. While the torso influence can be shown to extend across the complete audio range, pinnae cues increasingly dominate the spectral shape of the HRTF above 3 kHz (variations up to approx. ± 20 dB) [5], [6].

From an analysis of HRTFs measured for various HATOs,

Guldenschuh et al. [7] found that the most prominent torso reflections occur when ear, shoulder, and source are approximately aligned, and the source elevation is within 20° below the horizontal plane to 40° above. The authors further hypothesized that effects caused by the torso should be audible at least for critical source positions.

Despite the dominating role of head and pinnae effects on the HRTF, Genuit [8] assumed that the torso induces localization cues at frequencies below 3.5 kHz. This was supported with evidence by Algazi et al. [6]. Using 3kHz low-pass-filtered stimuli in localization experiments, the authors could show that torso cues indeed help in detecting the elevation of sound sources outside the median plane.

The studies discussed above support the hypothesis that accounting for correct HATO will be necessary for a perceptually transparent binaural synthesis. Yet, measuring HRTFs with high angular resolution and a large number of HATOs is time consuming making efficient methods for interpolation between different HATOs desirable. Various interpolation approaches were described for HRTFs obtained for different directions of sound incidence but constant HATO [9], [10, pp. 43].

Hartung et al. [11] applied inverse distance weighting and spherical spline interpolation on HRIRs (time domain), and HRTFs log magnitude and phase spectra (frequency domain). Before applying interpolation in the time domain, HRIRs were time aligned on sample basis according to their maximum values (sampling rate 44.1 kHz). Inverse distance weighting is essentially a linear interpolation using a weighted average according to the great circle distance between the desired and actual source position, thus accounting for the spherical nature of HRTF data sets. When using spherical splines, interpolation is obtained by fitting polynomial functions to the data and evaluating them at the desired position given by azimuth and elevation. Smaller errors between interpolated and measured HRTFs were found for the frequency domain based methods with spherical spline interpolation tending to be superior to inverse distance weighting.

Using inverse distance weighting and minimum phase HRTFs, Minnaar et al. [12] investigated the minimum angular resolution needed for interpolating HRTFs without introducing audible artifacts. Physical evaluation revealed increasing interpolation errors for frequencies above 1 kHz. The largest errors were found at the contralateral ear, and at elevations below the horizontal plane, which is in good agreement with results of Hartung et al. [11]. Audibility of interpolation errors was assessed in a 3AFC listening test using a pink noise stimulus, and covering directions of sound incidence from the horizontal, median and frontal plane. For most source positions, subjects failed to discriminate between measured HRTFs and HRTFs that were interpolated from a 4° grid. Occasionally differences remained detectable for lateral directions and below the horizontal plane.

Moreover, several studies transformed HRTF data sets into the spherical harmonic domain, where interpolation can be achieved by evaluating the spherical harmonic functions at the desired position given by azimuth and elevation [13]–[15].

TABLE I: Measured source positions.

Source	1	2	3	4	5	6
Azim. φ_s [$^\circ$]	0	315	0	45	90	315
Elev. ϑ_s [$^\circ$]	90	30	0	0	0	-30
Distance [m ²]	2.2	2.5	2.1	2.2	2.1	2.6

In the present study, we physically and perceptually examined differences between dynamic auralizations of (a) HRTFs with constant and variable HATOs, as well as (b) measured and interpolated HRTFs. In the latter case, we specifically investigated the minimal resolution of HATOs required for interpolation artifacts to stay below the threshold of perception. We inferred that the torso effects should be most audible for head rotation to the left and right (termed *horizontal head rotations*), because in this case the largest changes of the ears' position relative to the torso occur. We hence limited our investigations accordingly.

II. HEAD-RELATED TRANSFER FUNCTIONS MEASUREMENT

Before being able to assess the effect of HATO, an appropriate HRTF data set was measured with the head and torso simulator FABIAN, which is equipped with a software-controlled neck joint, allowing for a precise control of the HATO in multiple degrees of freedom [16]. FABIAN's head and pinnae are casts of a human subject. The torso and the position of head and pinnae relative to the torso were designed according to anthropometric measures averaged across age and gender [8], [17]–[19]. Accordingly, FABIAN's ear canal entrances are located 17.5 cm above and 1.5 cm in front of the acromion which is the highest point of the shoulder blade.

HRTFs were measured for six source positions given in Tab. I. Thereby, azimuth angles $\varphi_s = \{0^\circ, 180^\circ, 90^\circ, 270^\circ\}$ denote sources in front and back, and to the left, and right of a listener's torso. Positive elevations ϑ_s denote sources above the horizontal plane. Accordingly, HATOs $\varphi_{HATO} = \{45^\circ, 315^\circ\}$ refer to a head rotation above the torso of 45° to the left, and right, respectively. Source position and HATO are independent, i.e. the source positions stays constant if the HATO changes and vice versa. Thus, torso-to-source azimuth φ_{t2s} is given by $360 - \varphi_s$; the head-to-source azimuth φ_{h2s} by $(\varphi_{HATO} - \varphi_s) \bmod 360$.

Source positions were chosen to be typical (e.g. on the horizontal plane) and particularly critical/non-critical with respect to a strong shoulder/torso effect and interpolation artifacts. Generally, source positions are critical for head orientations where ear, shoulder, and source are aligned (sources 2 to 5), this way giving rise to pronounced comb filters, or when the head and torso act as an obstacle for the sound field at the ears (sources 3 to 6), which results in strong shadowing at the contralateral ear, respectively. Source positions are less critical for sources well above the horizontal plane (source 1). Source distances between 2.1 m and 2.6 m were chosen to avoid proximity effects [20], [21] and to ensure that reflections from the speakers could be removed by windowing.

The data set allowed the auralization of horizontal head rotations with constant and variable HATO within the physiological maximum range of motion $\varphi_{HATO, \max} = \pm 82^\circ$ [19],



Fig. 2: Photo of the HRTF measurement setup taken while adjusting the source position with the help of a laser mounted below FABIAN’s left ear.

and a resolution of $\Delta\varphi_{\text{HATO,ref}} = 0.5^\circ$. This spatial resolution is smaller than the worst-case localization blur of 0.75° reported by Blauert [3, p. 39], and is termed *reference* in the following. Accordingly, 329 HRTFs for head rotations with constant, and 329 HRTFs for head rotations with variable HATO were measured for each source position. Moreover, additional HRTFs were measured to account for the different interpolation approaches. This will be described in more detail in Sec. IV-C after introducing head and torso interpolation.

Measurements were conducted in the fully anechoic chamber of the TU Berlin ($V = 1850 \text{ m}^3$, $f_c = 63 \text{ Hz}$) using sine sweeps between 50 Hz and 21 kHz with an FFT order of 16 while achieving a peak-to-tail SNR of about 90 dB. FABIAN was mounted onto the turntable of a *VariSphear* microphone array (with the microphone removed) which gave high precision control of the torso-to-source orientation [22]. As sound sources we used *Genelec 8030a* active studio speakers with the tweeters aiming at FABIAN’s interaural center (cross-over at 3 kHz, centers of tweeter and woofer 11 cm apart). Directivity measurements [23] showed that the major part of the torso laid within the speaker’s main lobe for all source positions and frequencies ensuring that the effect of the torso is well represented in the measured HRTFs (cf. Fig 3). The time variability of the loudspeakers’ frequency response could be reduced to $\pm 0.2 \text{ dB}$ by means of an one-hour warming up procedure. The measurement setup is shown in Fig. 2.

Subsequent to the HRTF measurements, FABIAN was removed and its *DPA 4060* miniature electret condenser microphones were detached for conducting reference measurements. The positions of the microphones were adjusted to be identical to FABIAN’s interaural center. Finally, HRTFs were calculated by spectral division of the measured HRTF and the reference spectrum, simultaneously compensating for transfer functions of loudspeakers and microphones. Further processing of HRTFs included high-pass filtering for rejection of low frequency noise, and shortening of the HRIRs to a length of 425 samples. Finally, HRIRs were saved as original phase, and minimum phase plus time of arrival (TOA) filters. Arrival times were estimated using onset detection on the ten times up-sampled HRIRs. Onsets were defined separately for the left and right channel by the first sample exceeding

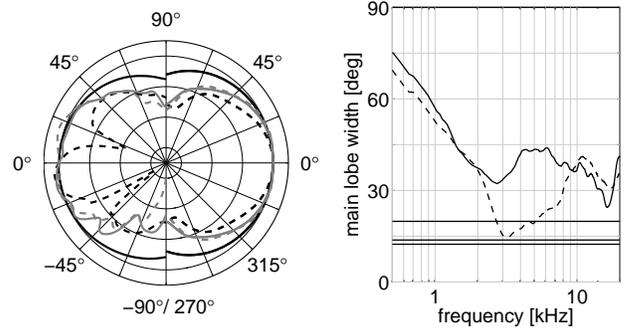


Fig. 3: Left: Vertical and horizontal directivity (left and right semicircle) of a Genelec 8030a normalized at 0° (0.5 kHz, 3 kHz given by solid/dashed black lines; 10 kHz, 20 kHz by solid/dashed gray lines. Grid spacing equals 10 dB). Right: Vertical (dashed) and horizontal (solid) main lobe width given by the angular distance between 0° and the -3 dB point. Horizontal lines mark FABIAN’s shoulder-to-shoulder, ear-to-elbow, and ear-to-hip distance (46 cm; 51 cm; 76 cm) translated to an angular distance for a source at 2.1 m distance.

$$\max(|HRIR_{l,r}|) - 6 \text{ dB}.$$

III. EFFECTS OF HEAD-ABOVE-TORSO ORIENTATION

A. Physical evaluation

This section presents a physical evaluation of the torso’s influence on HRTFs as a function of HATO and source position. Observed differences between head rotations with constant and variable HATO are discussed and a subset of source positions is selected for perceptual evaluation in a subsequent listening test.

1) *Method*: Differences in HRTFs were examined with respect to interaural time and level differences (ITD, ILD), as well as spectral fine structure. Therefore, ILDs were estimated as RMS level differences between left and right ear, whereas ITDs were calculated as differences in TOAs taken from the original phase HRIRs.

In order to obtain an impression of the spectral differences, the log-ratio of the magnitude responses between the HRTFs for constant and variable head-above-torso conditions was calculated (in dB) as

$$\Delta\text{HRTF}(f) = 20\log\frac{|\text{HRTF}_{\text{const}}(f)|}{|\text{HRTF}_{\text{var}}(f)|}, \quad (1)$$

where f is the frequency in Hz. For convenience, the dependency of the HRTF on head orientation, source position, and left and right ear was omitted in (1)-(3).

For a better comparability across source positions, a single value measure was calculated based on Minnaar *et al.* [12], who described the error between a reference and an interpolated HRTF by averaging absolute magnitude differences at 94 logarithmically spaced frequencies, and adding results for left and right ear. This was found to be a good predictor for the listening test results in [12], where subjects had to detect differences between original and interpolated HRTFs. However, instead of calculating the error for discrete frequencies

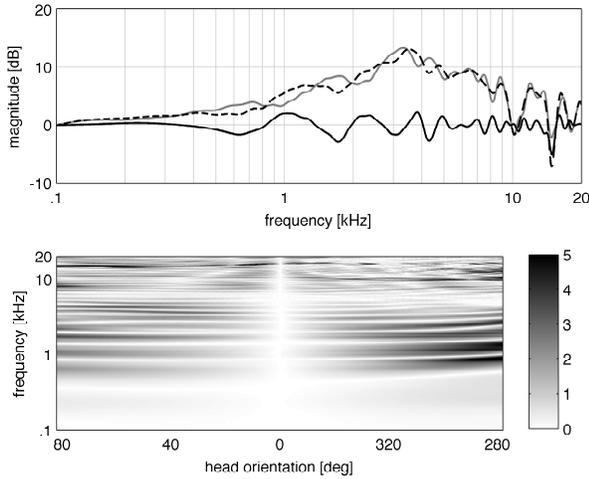


Fig. 4: Right ear HRTFs of the source at $(315^\circ; 30^\circ)$. Top: HRTF for a head orientation of 60° with constant (dashed) and variable (gray) HATO, and difference between them (black). Bottom: Difference between HRTFs with constant and variable HATO for all head orientations. Gray scale indicates magnitude in dB. Differences were calculated according to (1).

we used a Gammatone filter bank, as suggested by Schärer and Lindau [24]. The error level (in dB) in one filter band is given by

$$\Delta\text{HRTF}(f_c) = 20\log \frac{\int C(f, f_c) |\text{HRTF}_{const}(f)| df}{\int C(f, f_c) |\text{HRTF}_{var}(f)| df}, \quad (2)$$

where C is a Gammatone filter with center frequency f_c in Hz as implemented in the Auditory Toolbox [25]. The error level $\Delta\text{HRTF}(f_c)$ was calculated for $N = 39$ auditory filters between 70 Hz and 20 kHz. Then, the results for the left and right ear were added and averaged across f_c resulting in a single value error measure ΔG_μ (in dB) for each pair of HRTFs

$$\Delta G_\mu = \frac{1}{N} \sum_{f_c} (|\text{HRTF}_l(f_c)| + |\text{HRTF}_r(f_c)|). \quad (3)$$

2) *Results:* On average, ITD and ILD differences between head rotations with constant and variable HATO were found to be $2.6 \mu\text{s}$, and 0.24 dB , and hence well below known difference thresholds ($10 \mu\text{s}$ and 0.6 dB) [3, pp. 153]. Maximum deviations of $11.4 \mu\text{s}$ and 0.95 dB exceeded assumed threshold levels only slightly.

HRTFs for head rotations with constant and variable HATO are depicted in Fig. 4. In both cases, a comb-filter caused by the shoulder reflection is visible for frequencies above approx. 400 Hz. Above 3 kHz, it is partly masked by strong peak and notch patterns caused by pinnae resonances. However, when calculating the spectral difference according to (1) high frequency pinna cues cancel out due to identical head-to-source orientations. Expectedly, differences are nearly negligible for head orientations in the vicinity of 0° , as in this case head rotations with constant and variable HATO are very similar. For other head orientations comb-filter-like structures

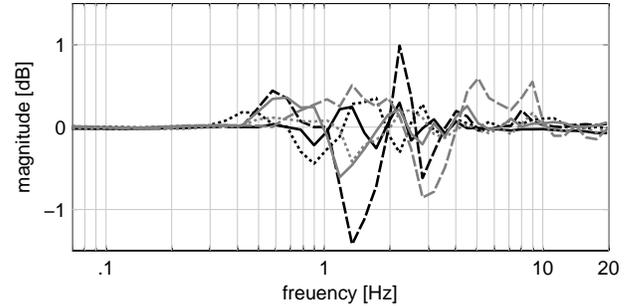


Fig. 5: Differences between HRTFs with constant and variable HATO averaged across head orientations and left and right ear according to (2). Sources 1-3 are given by solid, dotted and dashed black lines; sources 4-6 by solid, dotted and dashed gray lines.

are visible from 0.4 to 20 kHz. In the cases of either constant or variable HATOs distances between ear and shoulder vary, resulting in 'detuned' comb filters whose differences can be seen in Fig. 4. As a general trend, larger deviations occurred at the contralateral ear. Below 700 Hz slight deviations can be seen which are probably due to shadowing effects of the torso. This finding is in good accordance with Algazi et al. [4], where strong shadowing was found for sound sources below -40° elevation and the contralateral ear when using a KEMAR mannequin.

Spectral difference pattern according to (2) were comparable across sources and all exhibited comb-filter like structure (cf. Fig. 5). It was thus assumed that the frequency independent measure according to (3) would give a fair impression of average differences for all source positions and head orientations (cf. Fig. 6). Again, it can be seen that deviations are small in the vicinity of 0° whereas otherwise they reach a maximum of up to 2.4 dB . Moreover, a tendency for the error to increase with decreasing source elevation can be observed. The smallest error of $\leq 1 \text{ dB}$ is found for the source at $(0^\circ; 90^\circ)$. In this case, the shoulder reflection is weak for both constant and variable HATOs as most energy is reflected away from the ear. Intermediate differences of up to 1.4 dB occur for the sources on the horizontal plane and 30° elevation, most likely caused by strong shoulder reflections. The largest error of 2.4 dB is found for the source at -30° elevation and for head-to-source orientations larger than 45° azimuth, because the ear is partly shadowed by the torso in the case of constant HATO.

B. Perceptual evaluation

To test whether or not differences between head rotations with either constant or variable HATO are audible, an ABX listening test was conducted. The setup allowed for instantaneous and repeated comparison between HRTF sets using a dynamic binaural auralization accounting for horizontal head rotations of the listeners.

1) *Method:* Three women and eight men with a median age of 31 years took part in the listening test. All subjects had a musical background; ten subjects had participated in listening tests before; none reported known hearing impairments.

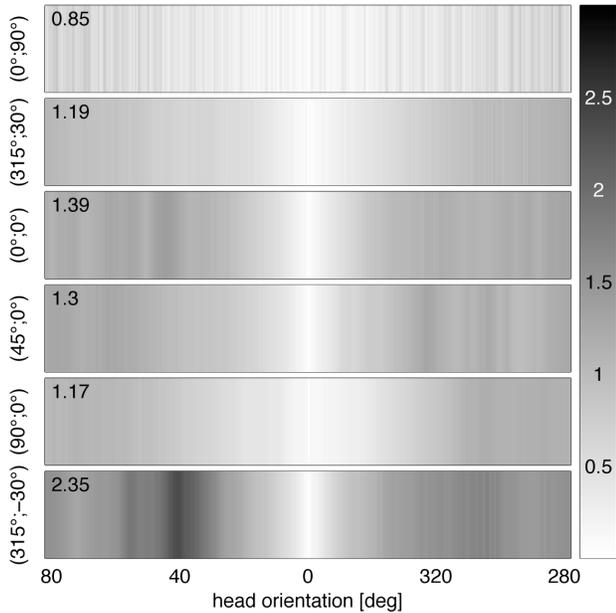


Fig. 6: Differences between HRTFs with constant and variable HATO calculated according to (3). Values inside the plot indicate the maximum error per source. Gray scale indicates magnitude in dB.

Following the ABX paradigm, three stimuli (A , B , and X) were presented to the subjects, whose task was to identify whether A or B equaled X . Conditions representing either head rotations with constant or variable HATO were randomly assigned to A , B , and X . Subjects were instructed and trained to listen to the stimuli in any order they felt to be helpful, to move/hold their heads to/at various positions during listening, to take their time at will before giving an answer, and to switch as fast or slow between stimuli as they wanted.

In order to limit the duration of the experiment, a subset of three sound sources was selected for perceptual evaluation. By drawing on the results of the physical evaluation, particularly critical and non-critical source positions at $(0^\circ; 90^\circ)$, $(90^\circ; 0^\circ)$, and $(315^\circ; -30^\circ)$ were selected. Two different audio stimuli were used: a frozen pink noise with a duration of 5 s (512 samples fade in/out) was chosen in order to reveal spectral differences, and an excerpt of German anechoic male speech with a duration of 5 s was used as a familiar and typical real-life sound. The experiment was split in two blocks whose sequence was balanced across subjects. Within a block, the source position was randomized while the audio content was held constant.

The combination of three sound sources and two audio contents lead to $2 \times 3 = 6$ conditions which were assessed individually by each subject. For each condition 23 ABX trials were conducted per subject, hence across subjects $23 \cdot 11 = 253$ trials were completed under each of the six conditions. Statistically, the test was designed to test a group averaged detection rate of 65% while guaranteeing cumulated type 1 and type 2 error levels to stay below 0.05 after accounting for repeated testing across conditions by Bonferroni correction [26]. Hence, for one tested condition detectability

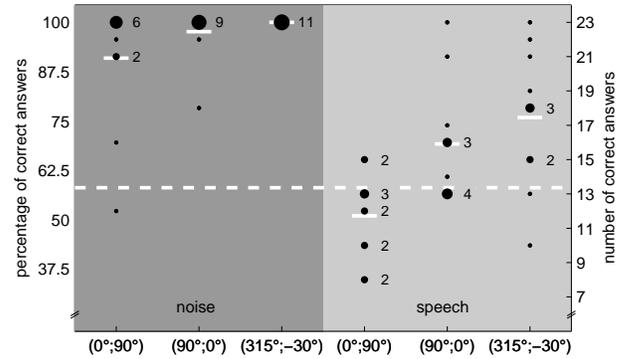


Fig. 7: Listening test results for all subjects and conditions. Dots indicate percentage/number of correct answers; numbers indicate how many subjects had identical results (same number of correct answers). Group mean scores given by solid white lines above the dashed line are significantly above chance.

was significantly above chance when observing 147 or more correct answers.

For reproduction of binaural signals, a thoroughly evaluated dynamic auralization engine and dedicated extraaural headphones were used [27], [28]. The test was conducted in a quiet listening room ($RT_{1 \text{ kHz}} = 0.6 \text{ s}$; $V = 30 \text{ m}^3$; $L_{\text{eq,A}} = 33 \text{ dB SPL}$), where subjects were seated on a revolving chair to comfortably reach and hold arbitrary head orientations. The listening test was administered using the whISPER environment [29], while displaying the user interface on a touchpad. Training prior to the listening test familiarized subjects with the interface and stimuli. Subjects were encouraged to take breaks at will to avoid fatigue, in turn needing maximally 1.5 hours for the test.

2) *Results:* Individual and group-averaged results are shown in Fig. 7 for all tested conditions. Group-averaged results, as given by the white horizontal bars, indicate a clear distinguishability of head rotations with constant and variable HATO: Results were significantly above chance for all tested conditions, except for the non-critical source positions at $(0^\circ; 90^\circ)$ in conjunction with the speech stimulus. Moreover, significantly less correct answers were given for the speech stimulus ($\chi^2 = 44.66$, $p < 0.001$, $df = 1$). When asked for perceived differences between head rotations with constant and variable HATO, the subjects mentioned coloration (11x) and/or localization (3x) in the case of the noise content, and coloration (6x), localization (5x), and/or source width (1x) for the speech sample.

So far, we discussed differences between head rotations with constant and variable HATO. For a number of different conditions we could show that the acoustic deviations between these two situations are audible. We thus conclude that variable HATOs have to be considered when aiming at a perceptually transparent binaural synthesis. In the remainder of this paper, we will discuss interpolation approaches suitable for an efficient representation of HATO in acoustic simulations.

IV. INTERPOLATION OF HEAD-ABOVE-TORSO ORIENTATION

In this section we first introduce and discuss different approaches to spatial interpolation of HRTFs. Second, we show a physical evaluation of in total 17 individual interpolation algorithms. Finally, we present the perceptual evaluation of a selected subset of these algorithms, and extend the results towards all approaches based on a perceptually motivated error measure.

A. Inverse distance weighting and spline interpolation

Interpolation algorithms for spherical data such as HRTFs may be distinguished with respect to whether they operate on neighboring data points only (nearest neighbor, inverse distance weighting, polynomials, splines), or whether they require a full-spherical data set (spherical splines, spherical harmonics). Nevertheless, in principle both families of approaches could be used for the interpolation of HATO. In the latter case however, spherical spline or spherical harmonic coefficients had to be interpolated instead of directly interpolating HRTFs. Consequently, when aiming at finding the difference threshold, full spherical HRTF data sets for HATOs between $\pm 82^\circ$ in the smallest resolution $\Delta\varphi_{\text{HATO, meas}} = 1^\circ$ were needed for calculation of the corresponding coefficients and successive interpolation onto the reference $\Delta\varphi_{\text{HATO, ref}}$. As this would require an unfeasibly large amount of measured data, the current study was restricted to spline – instead of spherical spline – interpolation, and inverse distance weighting. Moreover, interpolation was applied in the time and frequency domain as well as for original and minimum phase HRTFs.

Depending on the head orientation, HRIRs contain different arrival time delays. As a consequence, neighboring HRIRs are temporally misaligned and a direct time domain interpolation would result in double/blurred peak HRIRs. Two alignment strategies were applied to overcome this problem. On the one hand, arrival times were estimated using onset detection as described in Sec. II. On the other hand, we estimated the amount of misalignment from the cross-correlation function between two ten times up-sampled HRIRs ($\arg \max_{\tau} \rho_{xy}(\tau)$). In both cases fractional delays were applied for time alignment [30]. Additionally, TOAs were interpolated based on the the extracted values for both alignment procedures. In the frequency domain, magnitude and unwrapped phase spectra of the original phase HRTFs were interpolated separately, thus again inherently interpolating the TOA and ITD. For minimum phase HRIRs only the magnitude spectrum was interpolated and the result was made minimum phase again using the Hilbert transformation [31, pp. 789]. In this case, the TOA had to be interpolated separately for both time and frequency domain interpolation.

In addition to spline interpolation and inverse distance weighting, the nearest neighbor method was applied. In this case, the HRIR either with the HATO closest to the target orientation (similar to head interpolation) or with the closest torso-to-source azimuth (similar to torso interpolation) was used. This method was included as a possible approach for applications with limited computational resources as, e.g. in

mobile applications. Because the nearest neighbor method yields identical results in the frequency and time domain, as well as for original and minimum phase HRIRs, only one variation had to be tested. In total, 17 interpolation algorithms were investigated as listed in Tab. II, and described in more detail in the following.

With inverse distance weighting, HRTFs for intermediate HATOs φ'_{HATO} , source azimuth φ'_s , and elevation ϑ'_s are obtained as a weighted average of neighboring positions

$$x(\varphi'_{\text{HATO}}, \varphi'_s, \vartheta'_s) = \frac{\sum_{i=1}^2 x(\varphi_{\text{HATO}, i}, \varphi_{s, i}, \vartheta_{s, i}) d_{\varphi, \varphi'}^{-1}}{\sum_{i=1}^2 d_{\varphi, \varphi'}^{-1}} \quad (4)$$

whereby x denotes a sample of the HRIR in the case of time domain interpolation, and a bin of the HRTFs magnitude or phase response in the case of frequency domain interpolation. For head rotations restricted to the horizontal plane, the great circle distance $d_{\varphi, \varphi'}$ reduces to

$$d_{\varphi, \varphi'} = \arccos(\cos(\varphi_{\text{HATO}, i} - \varphi'_{\text{HATO}})). \quad (5)$$

The neighboring HATOs are given by

$$\varphi_{\text{HATO}, i} = \left[\left(\left\lfloor \frac{\varphi'_{\text{HATO}}}{\Delta\varphi_{\text{HATO, meas}}} \right\rfloor + i \right) \Delta\varphi_{\text{HATO, meas}} \right] \bmod 360 \quad (6)$$

where $i \in \{0, 1\}$, $\lfloor \cdot \rfloor$ denotes rounding to the next lower integer, and \bmod is the modulus operator. In contrast, cubic spline interpolation fits a piecewise polynomial through all $x(\varphi_{\text{HATO}, i})$ with a continuous first and second derivative on the entire interval [32], whereby

$$\varphi_{\text{HATO}, i} = (i \Delta\varphi_{\text{HATO, meas}}) \bmod 360, \quad (7)$$

with $-N \leq i \leq N$, $i \in \mathbb{Z}$, and $N = \lceil 82^\circ / \Delta\varphi_{\text{HATO, meas}} \rceil$.

B. Head and torso interpolation

Two different approaches can be considered when interpolating HATO in HRTFs. With head interpolation, intermediate data points are calculated from HRTFs with identical torso-to-source but differing head-to-source orientations (Fig. 8, left). Thus, HRTFs used for interpolation will deviate primarily in the high frequency range, which is dominated by direction-dependent (anti) resonance effects of the pinnae cavities. Hence, this approach is comparable to interpolating HRTFs of different sound source positions and thresholds are expected to be in the order given by Minnaar [12].

In the case of torso interpolation, HRTFs with identical head-to-source but differing torso-to-source orientations are used for the estimation of intermediate points (Fig. 8, right). This approach appears promising because the spectral effect of the torso in HRTFs is less prominent for most directions of sound incidence and the dominating high frequency structure will remain preserved. However, it requires additional HRTFs with source azimuths $\varphi_{s, i}$ for interpolating the desired source azimuth φ'_s

$$\varphi_{s, i} = (\varphi_{\text{HATO}, i} - \varphi'_{\text{HATO}} + \varphi'_s) \bmod 360, \quad (8)$$

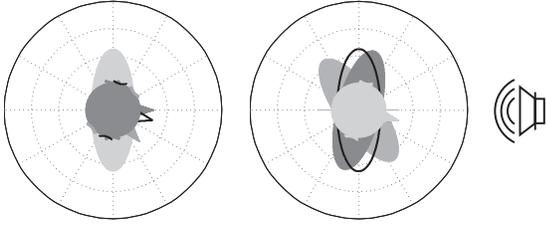


Fig. 8: Illustration of head (left) and torso interpolation (right) for inverse distance weighting ($\Delta\varphi_{\text{HATO, meas}} = 50^\circ$, $\varphi'_{\text{HATO}} = 350^\circ$, $\varphi'_s = 0^\circ$). Positions of measured HRTFs are shown with solid heads and torsi, interpolated HRTFs are indicated by black lines.

where i and $\varphi_{\text{HATO}, i}$ remain as specified for Eq. (6-7). As depicted in Fig. 8, Eq. (8) ensures that the head-to-source azimuth remains constant while the torso is rotated with respect to the source resulting in a change of φ_s . When applying inverse distance weighting to torso interpolation, two additional HRTFs with differing source azimuths are needed for each interpolation, whereas spline interpolation would require a multitude of additional HRTFs. Although this is not a drawback in practice as HRTF data sets usually cover source positions in a high spatial resolution, spline interpolation was excluded from this study in the case of torso interpolation due to the increased measurement effort. Nevertheless, we hypothesized that interpolation artifacts are smaller for torso interpolation compared to head interpolation.

C. Additional head-related transfer function measurements

Head and torso interpolation were investigated for 23 different resolutions of measured HATOs $\Delta\varphi_{\text{HATO, meas}} = \{1, 2, \dots, 10, 12, \dots, 30, 35, \dots, 45^\circ\}$ in the range of $\pm 82^\circ$ given by $\varphi_{\text{HATO, max}}$. Hence, additional HRTFs had to be measured: First, they were needed in cases where $\Delta\varphi_{\text{HATO, meas}}$ was not an integer divisor of $\varphi_{\text{HATO, max}}$. For example, HATOs of 60° and 90° were needed to interpolate to 82° , in the case of $\Delta\varphi_{\text{HATO, meas}} = 30^\circ$. Second, additional HRTFs were needed for testing torso interpolation: Because the torso is rotated during interpolation, two additional source positions had to be measured for each intermediate HATO, i.e. if φ'_{HATO} is not an integer divisor of $\Delta\varphi_{\text{HATO, meas}}$ (cf. Eq. (8), and Fig 8). This led to $\sum_{\forall k} 2 \cdot (329 - (2 \cdot \lfloor \varphi_{\text{HATO, max}} / \Delta\varphi_{\text{HATO, meas}, k} \rfloor + 1))$ additional HRTFs. Calculating the corresponding HATOs and source positions using (6)-(8) and removing duplicates resulted in 6679 additional HRTFs that were measured for each sound source listed in Tab. I.

D. Physical evaluation

A physical evaluation of all 17 algorithms was carried out, calculating differences between the reference and interpolated HRTFs according to (1). For this purpose, HRTFs were interpolated in the range of $-82^\circ \leq \varphi'_{\text{HATO}} \leq 82^\circ$ to $\Delta\varphi_{\text{HATO, ref}} = 0.5^\circ$ for each measured resolution $\Delta\varphi_{\text{HATO, meas}}$ and algorithm (cf. Fig. 9). Whereas in the reference, smallest changes in the high frequency fine structure are

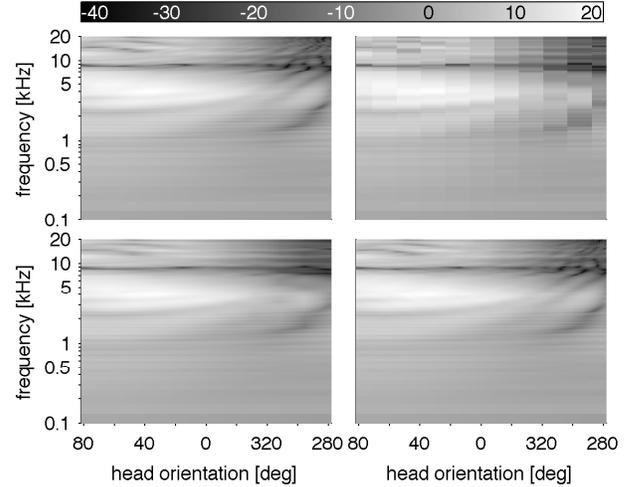


Fig. 9: Magnitude spectra of reference (top left) and interpolated HRTFs (linear interpolation, time domain, $\Delta\varphi_{\text{HATO, meas}} = 16^\circ$, source 3, right ear). Nearest neighbor, via head (top right); head interpolation (bottom left); torso interpolation (bottom right). Gray scale indicates magnitude in dB.

smoothly reproduced, discontinuities are clearly seen for the nearest neighbor algorithm, due to the hard switching between impulse responses for discrete HATOs. When comparing head interpolation to the reference, impairments become visible above approximately 2 kHz. In contrast, with torso interpolation, the spectral fine structure is mostly preserved.

Differences between reference and interpolated HRTFs according to (2) are shown in Fig. 10. They confirm our hypothesis that the errors for torso interpolation are smaller than for head interpolation. In general, and in accordance to Minnaar *et al.* [12], errors increase with frequency which is most likely related to high frequency pinnae cues in the HRTF that underlie a fast spatial fluctuation. Due to the similarity of the error pattern, it was again assumed that (3) still reflects differences between interpolation algorithms and source positions. For an overview of the average performance of algorithms and source positions, median errors averaged across HATO and $\Delta\varphi_{\text{HATO, meas}}$ are given in Tab. II.

Differences between approaches follow the line of argumentation given above. When looking at results for head interpolation, a slight superiority of spline compared over linear interpolation can be seen (0.18 vs. 0.24 dB on average). If excluding the nearest neighbor approach, results for time and frequency domain interpolation as well as for time alignment by cross correlation and onset detection are comparable. In tendency however, smaller errors occur in the frequency domain (0.16 vs. 0.18 dB) and when using cross correlation (0.18 vs. 0.19 dB). Moreover, average performance for original and minimum phase processing (0.17 dB), as well as for the best head interpolation compared to torso interpolation when using the nearest neighbor approach (0.16 dB) were identical.

Results for the source positions depend on the interpolation approach. For head interpolation, errors are largest for sources on the horizontal plane, slightly smaller for sources at

TABLE II: Median error between reference and interpolated HRTFs according to (3) for all interpolation algorithms and source positions (averaged across head orientations and $\Delta\varphi_{\text{HATO,meas}}$). Means across sources are given for ease of interpretation. Errors of 0 dB that occur when the head orientation is a multiple of $\Delta\varphi_{\text{HATO,meas}}$ were excluded from analysis.

#	Approach	Interp.	Domain	Phase	Alignm.	(0°;90°)	(315°;30°)	(0°;0°)	(45°;0°)	(90°;0°)	(315°;-30°)	mean			
1	head interpolation	nearest	time	org.	–	0.14	0.63	0.77	0.83	0.83	0.67	0.64			
2					cross cor.	0.11	0.21	0.29	0.33	0.35	0.22	0.25			
3		linear	time	org.	ons.	0.11	0.22	0.30	0.35	0.37	0.24	0.26			
4					min.	–	0.11	0.21	0.28	0.31	0.34	0.22	0.25		
5		linear	freq.	org.	–	0.11	0.18	0.24	0.27	0.30	0.19	0.22			
6					min.	–	0.11	0.19	0.24	0.27	0.30	0.19	0.22		
7		linear	time	org.	cross cor.	0.12	0.14	0.20	0.23	0.25	0.15	0.18			
8					ons.	0.12	0.14	0.21	0.24	0.25	0.16	0.19			
9		spline	time	org.	min.	–	0.12	0.15	0.20	0.23	0.25	0.16	0.19		
10					spline	freq.	org.	–	0.11	0.12	0.17	0.21	0.22	0.13	0.16
11								min.	–	0.12	0.12	0.17	0.20	0.22	0.13
12	torso interpolation	nearest	time	org.	–	0.14	0.11	0.16	0.18	0.10	0.26	0.16			
13					cross cor.	0.11	0.07	0.09	0.11	0.07	0.16	0.10			
14		linear	time	org.	ons.	0.11	0.07	0.10	0.11	0.07	0.17	0.10			
15					min.	–	0.12	0.07	0.10	0.11	0.07	0.17	0.11		
16		linear	freq.	org.	–	0.11	0.07	0.08	0.10	0.06	0.14	0.09			
17					min.	–	0.11	0.07	0.08	0.10	0.06	0.14	0.09		

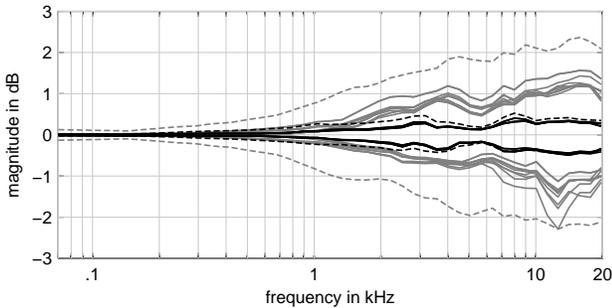


Fig. 10: 5-95% percentile range of the error between reference and interpolated HRTFs according to (2) for all interpolation algorithms (averaged across sources, head orientations, and $\Delta\varphi_{\text{HATO,meas}}$). Gray lines show head, black lines torso interpolation; dashed lines refer to the nearest neighbor approach.

$\pm 30^\circ$ elevation, and smallest for the source at 90° elevation. Interestingly, this rank order is exactly reflected in the median ILD per source and across head orientations (not shown here). This is in agreement with Hartung and Minnaar [11], [12] who reported interpolation errors to increase with increasing source-to-head azimuth (i.e. with increasing ILD) due to a lower SNR at the contralateral ear. For torso interpolation, in general, differences between sources are smaller. Largest errors occurred for source 6 and smallest for source 5. If averaged across all algorithms, errors for source 1 are smallest and almost identical, indicating its non-critical nature towards interpolation artifacts. Moreover, errors for source 6 are comparable for torso interpolation (0.16 dB @ linear interp.) and head interpolation (0.15 dB @ spline interp.). However, a more detailed analysis revealed that this only holds for $\Delta\varphi_{\text{HATO,meas}} \lesssim 10^\circ$, otherwise, torso interpolation exhibits smaller errors (0.31 dB vs. 0.37 dB).

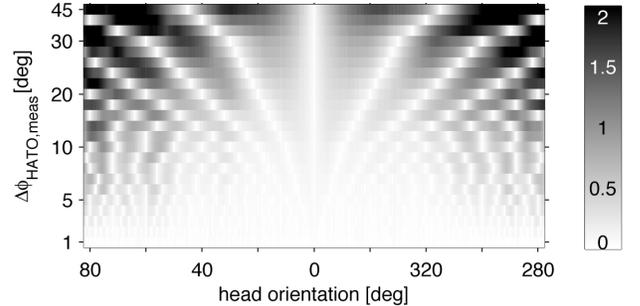


Fig. 11: Example of errors between reference and interpolated HRTFs according to (3): Original phase, time domain, head interpolation with cross correlation for HRIR alignment.

As an example for one interpolation approach, results according to (3) are shown in Fig. 11 for all head orientations and $\Delta\varphi_{\text{HATO,meas}}$. As expected, errors were zero at multiples of $\Delta\varphi_{\text{HATO,meas}}$, largest in between, and increased with increasing measurement grid width.

E. Perceptual evaluation

1) *Method*: Difference thresholds – defined as the inflection point of the sigmoid psychometric function – between reference and interpolated HRTFs were determined using a parametric, adaptive three alternative forced choice test utilizing the ZEST adaptive procedure. ZEST provides a fast and unbiased threshold estimation, which is robust against uncertainties with respect to its proper parameterization [33], [34]. The test was parameterized with a logistic psychometric function (slope parameter $\beta = 1$), a Gaussian a priori probability density function (mean set according to informal listening tests, standard deviation set to $\sigma = 25$), and a lapsing rate of

3%. Again, the whisPER listening test environment was used for conducting the experiment.

Following a 3AFC paradigm, the subjects' task was to detect the interpolated HRTFs by finding the oddball in three presented stimuli. Subjects were carefully instructed and trained to listen to the stimuli in any order they felt to be helpful, to move/hold their heads to/at various positions during listening, to take their time at will before giving an answer, and to switch as fast or slow between stimuli as they wanted. This was important, because the spatial regions of largest interpolation errors strongly depend on the interpolation interval, which changed continuously during the adaptive test procedure. Dynamic auralization for HRTFs with HATOs between $\pm 82^\circ$ was realized as described in Sec. III-B.

Two types of audio stimuli (continuous pink noise; anechoic male speech) and three algorithms ([A] nearest neighbor, via head; [B] head interpolation: time domain, spline interpolation; [C] torso interpolation: frequency domain, linear interpolation) were tested in a two-way factorial, fully repeated measures design ($2 \times 3 = 6$ conditions per subject). The experiment was conducted in an acoustically dry recording studio environment ($RT_{1\text{ kHz}} = 0.5\text{ s}$; $V = 145\text{ m}^3$; $L_{\text{eq,A}} = 23\text{ dB SPL}$).

In order to limit the listening test duration, only the source position most critical towards torso interpolation (315° ; -30°) was tested using minimum-phase HRTFs. To avoid listening fatigue, the test was split in two blocks, each starting with a training followed by three threshold estimates (20 trials each) and an intermediate break of 30 minutes or more. The presentation order of audio stimuli and algorithms was balanced across subjects, while the stimulus was held constant within blocks.

2) *Results*: Thresholds for 25 subjects (6 women, 19 men, median age 27, 24 subjects had musical background, 22 participated in listening tests before) are shown in Fig. 12. Two subjects were discarded from statistical analysis because they were short on time and hurried to finish the test. In turn, both subjects rated noticeably faster than others while showing considerably worse results.

Statistical analysis by means of ANOVA requires normally distributed samples. Because this criterion was violated under some conditions (Lilliefors test), non-parametrical tests were used for analyzing the results. Friedman's test showed highly significant differences between conditions ($\chi^2 = 112.4$, $p < .001$) Hence, as hypothesized, detectability thresholds increase from the nearest neighbor approach to torso interpolation when pooled across stimuli. Additionally, thresholds were higher for the speech as compared to the pink noise stimulus when pooled across algorithms. Post-hoc pairwise comparisons proved all observed differences to be highly significant (Wilcoxon signed rank tests, $p < 0.001$ after accounting for multiple testing by means of Bonferroni correction).

From inspection of the subjects' answers, we assumed that some did not hear differences between reference and interpolated HRTFs regardless of the interpolation interval when being presented with the speech stimulus. To support this assumption, Bernoulli tests [26] were carried out based on answers obtained for the largest measurement grid of $\Delta\varphi_{\text{HATO,meas}} = 45^\circ$. They revealed that one (head inter-

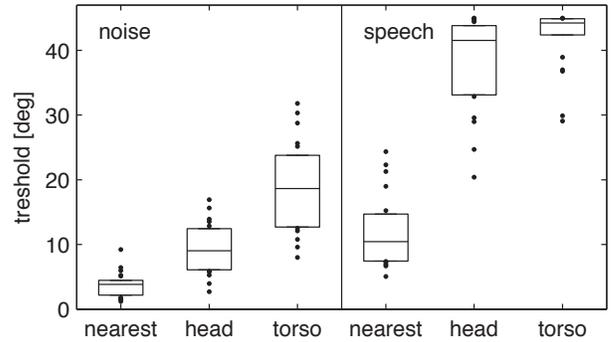


Fig. 12: Distribution of difference thresholds (in degree) for all conditions and subjects. Boxes show median and interquartile range (IQR). Values outside the IQR are marked by dots.

polation), and eight (torso interpolation) subjects failed to significantly discriminate between reference and interpolated HRTFs (type 1 and 2 error 0.025, Bonferroni corrected for multiple testing; testable effect $p = 0.9$). Keeping in mind that the presentation order was balanced and that all subjects detected differences for the noise stimulus, this was believed to be solely related to the speech signal. Its non-stationary and band limited nature made it harder to detect differences, which apparently were below individual thresholds of these subjects for all grid widths. Consequently, we assumed the measured threshold to be underestimated in this case because of this described ceiling effect.

In order to recommend the required measurement grid size that is needed to achieve or fall below a given group-averaged detectability, cumulated probability density functions were estimated from subjects' thresholds using a non-parametric modeling algorithm [35]. Grid width $\Delta\varphi_{\text{HATO}}$ for selected percentiles of average detectability are listed in Tab. III, and will be referred to as *threshold percentiles* in the following. Thereby, for example, the 5% threshold percentile denotes the grid width that is below the threshold of perception for 95% of the population underlying the subjects that participated in the listening test. For the noise stimulus, threshold percentiles increase by a factor of approximately two across algorithms, suggesting that differences between them are perceptually relevant. For the speech stimulus, this factor is even larger from nearest neighbor to head interpolation, but due to the ceiling effect small between head and torso interpolation.

When asked for perceived differences, subjects mentioned coloration (23x), and localization (13x) in the case of the noise stimulus, and localization (20x), and coloration (16x) for the speech sample.

F. Threshold prediction

To extend the results obtained in the perceptual evaluation towards interpolation algorithms that were not included in the listening test, thresholds for all algorithms and source positions were predicted based on an investigation of the interpolation error in dependency of the grid width $\Delta\varphi_{\text{HATO,meas}}$. According to (2), we obtained one error measure per HATO and auditory filter, resulting

TABLE III: Threshold percentiles $\Delta\varphi_{\text{HATO}}$, and corresponding error values ΔG_{95} (cf. Sec. IV-F) for all tested conditions.

		Noise			Speech		
		Near.	Head	Torso	Near.	Head	Torso
50%	$\Delta\varphi_{\text{HATO}}$	4.4	10.5	20.8	12.7	41.1	43.6
	ΔG_{95}	0.89	1.09	0.82	2.15	3.64	1.28
25%	$\Delta\varphi_{\text{HATO}}$	3.3	7.9	15.5	9.5	35.6	41.4
	ΔG_{95}	0.69	0.71	0.68	1.74	2.81	1.24
5%	$\Delta\varphi_{\text{HATO}}$	1.5	4.6	9.6	5.7	23.7	31.7
	ΔG_{95}	0.41	0.33	0.44	1.12	2.44	1.12

in 329 (HATOs) \times 39 (audit. filter) = 12,831 ($\Delta\text{HRTF}(f_c)$) values for each grid width and source position. By assuming that (a) differences between reference and interpolated HRTFs are audible if any $\Delta\text{HRTF}(f_c)$ exceeds a certain threshold, and (b) that due to the dynamic auralization the highest $\Delta\text{HRTF}(f_c)$ might not always be discovered, we expected the arithmetic mean across the largest five percent of the 12,831 values to be a perceptually suitable and robust error measure. This measure was termed ΔG_{95} and is depicted in Fig. 13. Expectedly, head interpolation exhibits larger errors than torso interpolation, and the nearest neighbor approach represents the upper error bound except for grid widths larger than 40° , where occasionally errors are largest for spline interpolation. In general, the error increases with increasing grid width, but especially for head interpolation local maxima and minima emerge. This indicates that the quality of interpolation is not only a function of grid width.

To establish a link to the results for source $(315^\circ; -30^\circ)$ obtained from perceptual evaluation, ΔG_{95} was calculated at the 5%, 25%, and 50% threshold percentiles given in Tab. III. If $\Delta\varphi_{\text{HATO}}$ was not included in the measured HATO resolution $\Delta\varphi_{\text{HATO,meas}}$, ΔG_{95} was calculated using a weighted average of the two neighboring values. For the noise stimulus the ΔG_{95} values as expected (a) are approximately equal within a given threshold percentile, (b) do not overlap across threshold percentiles, and (c) decrease with decreasing threshold percentile. In this cases this indicates their perceptual relevance, and their suitability to be used for predicting threshold percentiles for sources and interpolation algorithms that were not included in the perceptual evaluation. This is, however, not the case for the speech stimulus which might either be caused by the ceiling effect that biased the threshold percentiles in Tab. III, or it might suggest that the pure spectral error measure ΔG_{95} loses its validity in this case.

Finally, for predicting the threshold percentiles for all interpolation approaches, only the ΔG_{95} values obtained for the noise stimulus were used. For robustness, ΔG_{95} was averaged across the tested interpolation algorithms, which lead to the following values that were used for prediction: $\Delta G_{95,\text{pred.},50\%} = 0.93$ dB, $\Delta G_{95,\text{pred.},25\%} = 0.69$ dB, and $\Delta G_{95,\text{pred.},5\%} = 0.39$ dB. Subsequently, thresholds percentiles for all interpolation algorithms and sources were predicted by finding the first ΔG_{95} value exceeding the corresponding $\Delta G_{95,\text{pred.}}$. To make this prediction more exact, the curves in Fig. 13 were interpolated to a resolution of 0.01°

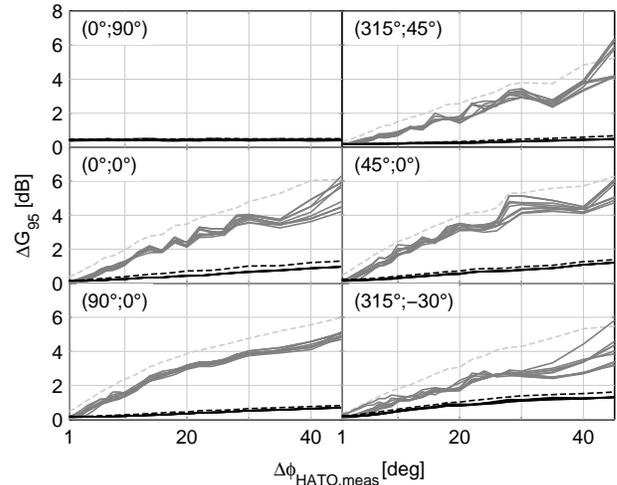


Fig. 13: ΔG_{95} for all source positions and grid widths $\Delta\varphi_{\text{HATO,meas}}$. Gray lines show head, black lines torso interpolation; dashed lines refer to the nearest neighbor approach.

beforehand.

Noteworthy, the average difference between the nine threshold percentiles estimated from the perceptual evaluation and the predicted threshold percentiles was only 0.9° . However, a deviation of 5° between thresholds occurred for the torso interpolation (algorithm #17, source 6, 50% percentile). This was caused by the slow increase of ΔG_{95} for $18^\circ \leq \Delta\varphi_{\text{HATO}} \leq 25^\circ$ (cf. Fig. 13) and was thus considered to be perceptually non-critical. The smallest predicted threshold percentile for each interpolation algorithm across sources is listed in Tab. IV.

V. DISCUSSION

Our evaluation of the effect of HATO in HRTFs supported findings of earlier studies regarding the comb-filter like nature of the shoulder reflection, which was found to be most prominent if sound source, shoulder, and ear are aligned [4], [6], [7]. Because observed deviations in ITDs and ILDs were below the threshold of audibility for the vast majority of HATOs and source positions, we suppose perceived differences to be mostly due to spectral deviations. Perceived differences in localization might also be due to spectral cues, related to mismatched comb-filters in HRTFs exciting different directional bands [3, pp. 93] and thus evoking differences in perceived elevation. This assumption would be in accordance with Algazi et al. [6], who found torso and shoulder related cues to be involved in the perception of elevation for sources outside the median plane.

Best interpolation results were – as presumed a priori and predicted by physical evaluation – achieved for torso interpolation using HRTFs with identical head-to-source but varying torso-to-source orientation. Compared to head interpolation, this provided a better preservation of high frequency pinnae cues when interpolating between HRTFs with identical head-to-source orientation. Remarkably, torso interpolation in conjunction with the nearest neighbor approach outperformed most head interpolation algorithms, thus suggesting that the

TABLE IV: Threshold estimates in degree for all 17 interpolation algorithms. For easy of display, only the smallest (most critical) estimate across the six source positions is shown. The interpolation algorithms are numbered according to Tab. II (1: head interp., near. neighb.; 2-11: head interp., lin./spline; 12 torso interp., near. neighb.; 13-17: torso interp., lin./spline).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
50%	3.1	5.0	4.6	5.6	5.7	5.7	6.5	6.2	6.5	6.2	6.3	17.2	23.6	20.9	23.6	25.1	25.2
25%	2.1	3.9	3.5	4.3	4.5	4.5	5.6	4.2	4.9	5.0	5.0	11.9	14.9	14.4	14.7	15.9	16.0
5%	1.0	2.8	2.1	2.8	3.0	3.0	4.4	2.0	3.5	3.6	3.6	5.8	8.6	7.6	8.3	8.7	8.7

effect of the torso on the HRTF is small compared to that of head and pinnae. In tendency, and according to Hartung [11], the physical evaluation revealed smaller errors for frequency compared to time domain interpolation as well as for spline compared to linear interpolation, whereas original and minimum phase interpolation on average performed identical.

Difference thresholds that represent the minimally needed angular resolution of HATO were (a) estimated from perceptual evaluation, and (b) predicted based on the latter. Both reflect the superiority of torso interpolation: For the vast majority of tested algorithms, thresholds for torso interpolation outperform those of head interpolation by a factor of two to three. As assumed a priori, the median threshold of 10.5° for the noise stimulus and head interpolation (cf. Tab. III) is comparable to results of Minnaar et al. [12]. For a source at $(315^\circ; -30^\circ)$, the authors found a resolution of 8° to be sufficient for interpolation artifacts to be inaudibly small. Note that the criterion of audibility applied by Minnaar et al. is stricter than the threshold criterion applied in our study which might account for the gap between the results. The similarity is due to the fact that in both cases HRTFs with different head-to-source orientations were used for interpolation.

While the perceptual evaluation was carried out using dynamic binaural synthesis allowing for head rotations in the horizontal plane, it can be assumed that different subjects listened to HRTFs for different head orientations during rating. This makes it likely that not all subjects discovered the head orientations where largest differences appeared. However, it seems unlikely that the dynamic auralization biased the results keeping in mind that (a) subjects were carefully instructed and trained to listen for differences at various head orientation, (b) an inspection of the raw data (hit rates in listening test I; thresholds in listening test II) suggest that most subjects actually detected differences, and (c) results are comparable to listening tests carried out using static binaural synthesis [12]. In turn, we suggest that the results are generalizable to a wide range of head orientations *because* different subjects evaluated HRTFs at different head orientations.

The interpolation of HATO requires HRTF data sets with a high resolution and various HATOs. Different proposals were made regarding the required resolution of source positions. Zhang et al. [36] transformed HRTFs into the spherical harmonic domain and found the reconstruction to be *reasonably accurate* if using 2304 HRTFs. Minnaar et al. [12] suggested that interpolation errors will remain inaudible for 1130 HRTFs if using minimum phase interpolation in the time domain. Consequently, a perceptual transparent representation of the HATO in the range of $\pm 75^\circ$ will require about 8,000 to 16,000

HRTFs using $\Delta\varphi_{\text{HATO}} = 25^\circ$ (predicted 50% threshold percentile for torso interpolation; Tab. IV, #16). This appears to be feasible – even for human subjects – when considering fast HRTF measurement and modeling techniques [37], [38].

VI. CONCLUSION

In this study, we assessed the audibility of differences occurring during head rotations with constant or variable HATO, as well as the suitability of different algorithms for interpolating the HATO in HRTFs. To this end, we examined spectral and temporal deviations, and conducted two listening tests.

Although the effect of the torso on the HRTF is small compared to that of head and pinnae, we showed that differences between head rotations with constant and variable HATO were audible for the vast majority of source positions and audio contents. This suggests the importance of accounting for correct HATO at least if aiming at an authentic auralization, i.e. an auralization that is indistinguishable from a corresponding real sound field. This might, for example, be the case when benchmarking BRIRs obtained from numerical room modeling techniques against measured BRIRs.

Our evaluation of the interpolation of HATO in HRTFs showed that a grid width between 20° and 25° is sufficient when using torso interpolation, even for critical audio content and source positions. In this case, interpolation artifacts were below threshold for 50% of the subjects. A resolution of 8° was needed for artifacts to be subliminal for 95% of the subjects. If feasible, interpolation should be carried out in the frequency domain separately for the magnitude and unwrapped phase response.

This study was restricted to head rotations in the horizontal plane, because they were considered most important and critical. Nevertheless, future studies could also investigate the effect and interpolation of head rotations in elevation and roll. Moreover, it would be interesting to examine in how far interpolation algorithms in general – not only for HATO – can be applied to BRIRs, too, while assuming that reverberant sound fields will pose higher demands on interpolation algorithms. In addition, perceptual consequences of artifacts arising from larger interpolation intervals might be subjected to further qualitative analysis, as this might be interesting for applications not demanding an authentic reproduction.

ACKNOWLEDGMENT

This work was funded by the German Research Foundation (DFG WE 4057/3-1). We thank Joseph G. Tylka from the 3D3A Laboratory at Princeton University for providing directivity data of the Genelec 8030a.

REFERENCES

- [1] M. Vorländer, *Auralization. Fundamentals of acoustics, modeling, simulation, algorithmics and acoustic virtual reality*, 1st ed. Berlin Heidelberg: Springer, 2008.
- [2] P. Minnaar, K. S. Olesen, F. Christensen, and H. Møller, "The importance of head movements for binaural room synthesis," in *International Conference on Auditory Display*, Espoo, Finland, July/August 2001, pp. 21–25.
- [3] J. Blauert, *Spatial Hearing. The psychophysics of human sound localization*, revised ed. Cambridge, Massachusetts: MIT Press, 1997.
- [4] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064, November 2002.
- [5] M. B. Gardner, "Some monaural and binaural facets of median plane localization," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1489–1495, 1973.
- [6] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1110–1122, March 2001.
- [7] M. Guldenschuh, A. Sontacchi, and F. Zotter, "HRTF modelling in due consideration variable torso reflections," in *Acoustics*, Paris, France, June 2008.
- [8] K. Genuit, "Ein Modell zur Beschreibung von Außenohrübertragungseigenschaften," Ph.D. dissertation, Technische Hochschule, Aachen, Germany, 1984.
- [9] S. M. Robeson, "Spherical methods for spatial interpolation: Review and evaluation," *Cartography and Geographic Information Systems*, vol. 24, no. 1, pp. 3–20, 1997.
- [10] R. Nicol, *Binaural Technology*, ser. AES Monograph. New York, USA: Audio Eng. Soc., April 2010.
- [11] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *16th Int. AES Conference*, Rovaniemi, Finland, April 1999, pp. 319–329.
- [12] P. Minnaar, J. Plogsties, and F. Christensen, "Directional resolution of head-related transfer functions required in binaural synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 919–929, October 2005.
- [13] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411, October 1998.
- [14] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, May 2004, pp. IV 45–48.
- [15] M. Follow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig, "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition," *Acta Acust. united Ac.*, vol. 98, pp. 72–82, 2012.
- [16] A. Lindau, T. Hohn, and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *122th AES Convention, Convention Paper 7032*, Vienna, Austria, May 2007.
- [17] DIN 33402-2, *Ergonomics - Human body dimensions - Part 2: Values*. Berlin, Germany: Beuth, 2005.
- [18] DIN IEC/TS 60318-7:2011, *Electroacoustics - Simulators of human head and ear - Part 7: Head and torso simulator for the measurement of hearing aids*. Berlin, Germany: Beuth, 2005.
- [19] C. T. Morgan, A. Chapanis, J. S. Cook, and M. Lund, *Human Engineering Guide to Equipment Design*. New York, USA: McGraw-Hill, 1963.
- [20] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1465–1479, September 1999.
- [21] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *130th AES Convention, Engineering Brief*, 2011.
- [22] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the neumann KU 100," in *AIA-DAGA 2013, International Conference on Acoustics*, Merano, Italy, March 2013, pp. 592–595.
- [23] J. G. Tylka, "On the calculation of full and partial directivity indices," 3D Audio and Applied Acoustics Laboratory, Princeton University, Princeton, New Jersey, USA, Technical Report, November 2014.
- [24] Z. Schärer and A. Lindau, "Evaluation of equalisation methods for binaural signals," in *126th AES Convention, Convention Paper*, Munich, Germany, May 2009.
- [25] M. Slaney, "Auditory toolbox. version 2," Interval Research Corporation, Technical Report, 1998.
- [26] L. Leventhal, "Type 1 and type 2 errors in the statistical analysis of listening tests," *J. Audio Eng. Soc.*, vol. 34, no. 6, pp. 437–453, June 1986.
- [27] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, September/October 2012.
- [28] F. Brinkmann, A. Lindau, M. Vrhovnik, and S. Weinzierl, "Assessing the authenticity of individual dynamic binaural synthesis," in *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany, April 2014, pp. 62–68.
- [29] A. Lindau, "Whisper. a matlab toolbox for performing quantitative and qualitative listening tests," 2014. [Online]. Available: <http://dx.doi.org/10.14279/depositonnce-31>
- [30] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30–60, January 1996.
- [31] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-time signal processing*, 2nd ed. Upper Saddle, USA: Prentice Hall, 1999.
- [32] W. Gautschi, *Numerical analysis*, 2nd ed. Basel, Switzerland: Birkhäuser, 2012.
- [33] B. Treutwein, "Adaptive psychophysical procedures," *Vision Research*, vol. 35, no. 17, pp. 2503–2522, 1995.
- [34] S. Otto and S. Weinzierl, "Comparative simulations of adaptive psychometric procedures," in *NAG/DAGA 2009, International Conference on Acoustics*, Rotterdam, Netherland, 2009, pp. 1276–1279.
- [35] K. Zychaluk and D. H. Foster, "Model-free estimation of the psychometric function," *Attention, Perception, & Psychophysics*, vol. 71, no. 6, pp. 1414–1425, 2009.
- [36] W. Zhang, M. Zhang, R. A. Kennedy, and T. D. Abhayapala, "On high-resolution head-related transfer function measurements: An efficient sampling scheme," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 575–584, February 2012.
- [37] G. Enzner, C. Antweiler, and S. Spors, "Acquisition and representation of head-related transfer functions," in *The technology of binaural listening*, 1st ed., ser. Modern acoustics and signal processing, J. Blauert, Ed. Heidelberg at al.: Springer, 2013, pp. 57–92.
- [38] T. Huttunen, A. Vanne, S. Harder, R. R. Paulsen, S. King, L. Perry-Smith, and L. Kärkkäinen, "Rapid generation of personalized HRTFs," in *55th Int. AES. Conf.: Spatial Audio*, Helsinki, Finland, August 2014.

Fabian Brinkmann received his M.A. degree (magister artium) in communication sciences and technical acoustics from TU Berlin, Germany. Since 2011 he is a research associate in the DFG research consortium *SEACEN* and is currently pursuing the Ph.D. degree in the field of auralization and binaural synthesis. He is reviewer for EAA and Springer Science publishers, and interested in evaluation and signal processing for spatial audio.

Reinhild Roden received a bachelor's degree in musicology and media from HU Berlin and a bachelor's degree in engineering in hearing technology and audiology from Jade University Oldenburg. She is now involved in the research consortium *SEACEN* at TU Berlin while pursuing her master's degree in audio communication and technology in the field of spatial audio and audio signal processing with a focus on auditory perception.

Alexander Lindau obtained an M.A. degree (magister artium) in communication sciences, electrotechnical engineering and technical acoustics, and a doctoral degree (Dr. rer. nat.) from TU Berlin. Dr. Lindau has published about 40 conference papers, journal articles and book chapters and is reviewer for AES, EAA, IEEE, and Springer Science publishers. Currently Mr. Lindau is a research associate at the Audio Communication Group of the TU Berlin and involved the DFG research consortium *SEACEN* where he focuses on new approaches towards the perceptual evaluation of spatial audio technologies.

Stefan Weinzierl is head of the Audio Communication Group at TU Berlin. His activities in research include audio technology, musical acoustics, room acoustics and virtual acoustics. He is coordinating a master program in Audio Communication and Technology at TU Berlin and teaching Tonmeister students at the University of the Arts (UdK) in audio technology and digital signal processing. With a diploma in physics and sound engineering and a two-year study in musicology at UC Berkeley, he received his Ph.D. from TU Berlin. He is currently coordinating a German research consortium on virtual acoustics at TU Berlin.