OPEN ACCESS Freely available online

J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions"
J. Audio Eng. Soc., vol. 69, no. 1/2, pp. 104–117, (2021 January/February).
DOI: https://doi.org/10.17743/jaes.2020.0070

Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions

JOHANNES M. AREND,[1, 2] *AES Student Member*, **FABIAN BRINKMANN,**[2] *AES Associate Member* **AND**
(johannes.arend@th-koeln.de) (fabian.brinkmann@tu-berlin.de)

CHRISTOPH PÖRSCHMANN,[1] *AES Associate Member*
(christoph.poerschmann@th-koeln.de)

[1]*TH Köln – University of Applied Sciences, Cologne, Germany*
[2]*Technical University of Berlin, Berlin, Germany*

High-quality spatial audio reproduction over headphones requires head-related transfer functions (HRTFs) with high spatial resolution. However, acquiring datasets with a large number of (individual) HRTFs is not always possible, and using large datasets can be problematic for real-time applications with limited resources. Consequently, interpolation methods for sparsely sampled HRTFs are of great interest, with spherical harmonics (SH) interpolation becoming increasingly popular. However, the SH representation of sparse HRTFs suffers from spatial aliasing and order truncation errors. To mitigate this, preprocessing methods have been introduced that time-align the sparse HRTFs before SH interpolation. This reduces the effective SH order and thus the number of HRTFs required for SH interpolation. In this paper, we present a physical evaluation of four state-of-the-art preprocessing methods, which showed very similar performance of the methods with notable differences only at low SH orders and contralateral HRTFs. We also performed a listening experiment with one selected method to determine the minimum required SH order required for perceptually transparent interpolation. For the selected method, a sparse HRTF set of order $N \approx 7$ is sufficient for interpolating a frontal source presenting speech or percussion. Higher orders are, however, required for a lateral source and noise.

0 INTRODUCTION

Head-related transfer functions (HRTFs) are one key component for headphone-based spatial audio rendering, as often used in virtual reality (VR) or augmented reality (AR) applications [1, 2]. HRTFs describe the sound incidence from a source to the left and right ear and the associated directional filtering of incoming sound by the pinna, head, and torso. As such, HRTFs include binaural cues (i.e., interaural level differences (ILDs) and interaural time differences (ITDs) primarily used for sound source localization in the horizontal plane) as well as monaural spectral cues primarily used for sound source localization in the median plane [3].

For high-quality spatial audio over headphones, HRTFs with high spatial resolution are essential. Usually, such data are measured on dense spherical sampling grids, which can be achieved by sequential measurements to obtain dummy head HRTFs [4–6], but require procedures and equipment optimized for speed if measuring human subjects [7–9]. For this purpose, measurement systems consisting of (semi)circular loudspeaker arcs are used with signal acquisition techniques that allow for a continuous rotation of the subject or arc. Given this, it is of great interest to measure fewer HRTFs on a sparse spatial sampling grid and generate dense HRTF sets by means of interpolation (also referred to as spatial upsampling). This would decrease the cost and complexity of HRTF measurement systems and allow for faster rotations depending on the acquisition method. Furthermore, interpolation of sparse HRTF sets may reduce the memory and computational load for real-time applications with limited resources (e.g., mobile applications).

Currently very popular is the description and interpolation of HRTFs in the spatially continuous spherical harmonics (SH) domain (see Sec. 1). However, the required number of spatial samples (i.e., measurement directions) increases with frequency, and an SH order (also called spatial order) of $N_{\mathrm{max}} \approx 40$ is needed for a physically correct interpolation up to 20 kHz, resulting in at least $(N + 1)^2 = 1{,}681$ measurement directions [10]. Obviously, sparse HRTF sets

do not meet this requirement, and their SH representation thus suffers from so-called sparsity errors, which is a combination of spatial aliasing and order truncation errors [11]. Because sparse sampling grids only allow SH processing up to $N_{\text{sparse}} < N_{\text{max}}$, energy above N_{sparse} is irreversibly aliased to lower orders, causing spatial ambiguities that result in a high-shelf-like energy increase in SH interpolated HRTFs [12, 13, 11]. The predominant truncation error leads to reduced spatial detail showing up as a severe high frequency roll-off [14, 15, 11], caused by discarding energy above N_{sparse}. In combination, these effects also result in ILD errors and degraded loudness stability in dynamic scenes [16, 11].

To enable accurate SH interpolation of sparse HRTF sets, several preprocessing techniques have been introduced. In the present study, we focus on methods that align the head-related impulse responses (HRIR, time-domain equivalent of the HRTF) in the time [17, 18] or frequency domain [15, 19, 16, 20] prior to the SH interpolation and reverse the alignment afterwards. Since most higher-order HRTF energy stems from rapid spatial phase changes, aligning the HRIRs and thus also the phase components significantly decreases the high-order energy and related sparsity errors [18, 21, 20]. Because the phase changes are caused by the distance of the ears to the coordinate origin—the center of the head, in this case—the alignment can also be interpreted as centering the ears in the origin. For a more comprehensive overview of preprocessing methods, please refer to [19] and Chapter 4.11 of [22].

The studies on preprocessing introduced in the previous paragraph all showed that time-alignment decreases spectral and temporal errors and thus increases the quality of interpolated HRTFs, especially for low-order SH interpolation. However, listening experiments assessing the perceptual performance of SH-based HRTF interpolation either with or without preprocessing are rare. The only study directly related to the topic was presented by Pike and Tew (see Chapter A.8 of [18]). They conducted a Multi-Stimulus Test with Hidden Reference and Anchor (MUSHRA), comparing perceivable differences between a measured reference and SH interpolated HRTFs with and without subsample precise onset-based time-alignment. While interpolated HRTFs were indistinguishable from the reference at $N = 35$ in both cases, at $N = 5$, the time-alignment significantly reduced perceptual differences at least for frontal source positions, whereas for a lateral source position perceptual differences were still clear. Besides that, there are a few studies on the impact of low-order SH representation of HRTFs on localization accuracy [23], perceived loudness stability [11], or speech intelligibility in noise [24].

To the best of our knowledge, a systematic comparison of the different alignment approaches and listening experiments to find the minimum order N that is required for a perceptually transparent SH interpolation is missing so far. Because the methods differ in their computational complexity, a detailed comparison might help to choose the method that is most appropriate for a specific application, whereas the minimum required SH order is of importance for high-quality applications and can provide a starting point for further perceptual studies for applications that allow for a certain quality degradation. To close this gap, we present a physical evaluation of all suggested methods showing that they perform comparably. In addition, we conducted an adaptive forced choice listening experiment with one selected alignment approach to examine the minimum SH order required for interpolated HRTFs to be indistinguishable from a measured reference.

The remainder is structured as follows. Sec. 1 briefly reviews the fundamentals of HRTF representation and interpolation in the SH domain, and Sec. 2 describes the different preprocessing methods in detail. Sec. 3 provides a physical evaluation of the discussed methods by means of spectral and temporal error measures. Sec. 4 describes the listening experiment and results, followed by a discussion and conclusion in Sec. 5 and Sec. 6.

1 SPHERICAL HARMONICS REPRESENTATION OF HRTFS

The HRTF $H^{\text{l, r}}(\omega, \Omega)$ for the left and right ear can be represented in the SH domain by a set of SH coefficients $h_{nm}^{\text{l,r}}(\omega)$, which can be obtained by the spherical Fourier transform (SFT) (see Chapter 1 of [25]) (indices for the left and right ear are omitted in the following whenever the processing is identical for both ears):

$$h_{nm}(\omega) = \int_0^{2\pi} \int_0^{\pi} H(\omega, \Omega) Y_n^m(\Omega)^* \cos\theta \mathrm{d}\theta \mathrm{d}\phi. \tag{1}$$

The angular frequency is given by $\omega = 2\pi f$, with f being the temporal frequency. The direction $\Omega = (\phi, \theta)$ is defined by the azimuth $\phi = [0^\circ, 360^\circ]$ and the elevation $\theta = [-90^\circ, 90^\circ]$, whereby ϕ is measured counterclockwise in the xy-plane, starting at positive x, and θ is 90° at positive z. The notation $(\cdot)^*$ denotes the complex conjugate and Y_n^m the complex SH basis functions of order n and degree m defined as

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\sin\theta) \mathrm{e}^{\mathrm{i}m\phi}, \tag{2}$$

with the associated Legendre functions P_n^m and the imaginary unit $\mathrm{i} = \sqrt{-1}$.

In practice, the HRTF is sampled at a finite number of directions, and therefore, the integral in Eq. (1) must be discretized to Q sampling points corresponding to the measurement directions Ω_q. The respective discrete SFT is defined as

$$h_{nm}(\omega) = \sum_{q=1}^{Q} \alpha_q \, H(\omega, \Omega_q) Y_n^m(\Omega_q)^*, \tag{3}$$

where the quadrature weights α_q compensate for an uneven distribution of the sampling points (Chapter 4 of [26]). Alternatively, the discrete SFT can also be formulated in matrix form and then calculated by an inversion of the respective SH transformation matrix (Chapter 3 of [25]), but for the present work, the discrete SFT was always calculated using the closed-form expression according to Eq. (3).

Due to the analytical and spatially continuous basis functions, the SH representation allows for interpolation, that is, HRTFs $\widehat{H}(\omega, \Omega_t)$ for any direction Ω_t can be reconstructed by the discrete inverse spherical Fourier transform (ISFT):

$$\widehat{H}(\omega, \Omega_t) = \sum_{n=0}^{N} \sum_{m=-n}^{n} h_{nm}(\omega) Y_n^m(\Omega_t)\,. \tag{4}$$

However, the discrete sampling in Eq. (3) directly limits the maximum resolvable SH order N,

$$N \leq \lfloor \sqrt{Q/\lambda} - 1 \rfloor, \tag{5}$$

with the efficiency factor $\lambda \geq 1$ that depends on the sampling scheme and the floor operator $\lfloor \cdot \rfloor$. Thus, sparsity errors occur if $N < N_{\mathrm{max}} \approx 40$. As mentioned in the introduction, these errors manifest in spatial ambiguities, reduced spatial resolution, and spectral and temporal distortions in the interpolated HRTFs. SH interpolation of the complex HRTF spectra according to Eqs. (3) and (4) will be referred to as unprocessed (UP) interpolation in the following (i.e., SH interpolation without time-alignment).

2 TIME-ALIGNED SPHERICAL HARMONICS INTERPOLATION

This section introduces the investigated methods in depth. Although the algorithms differ in detail, the underlying idea is the same. All algorithms aim to lower the SH order that is required for high-quality SH interpolation by minimizing the phase changes across space during preprocessing. This is always done separately for the left and right ear and is pursued by aligning the impulse responses by means of time or frequency domain processing. In all cases, this is achieved by a spectral multiplication or division of the HRTF with an alignment function. After the alignment, all algorithms perform the discrete SFT and ISFT according to Eqs. (3) and (4) using the complex HRTF spectra. However, the time of arrival (TOA) (i.e., the time where the onsets occur in the HRIRs) is lost during the alignment. Therefore, it has to be reconstructed after the interpolation, which requires a spatially continuous TOA model in postprocessing. To foster reproducible research, example implementations of the methods under investigation are published as part of the SUpDEq Toolbox for MATLAB[1].

2.1 Onset-Based Time-Alignment

Sample accurate onset-based time-alignment (OBTA) was first proposed by Evans et al. [17] and was refined to subsample accuracy by Pike and Tew [18] as well as by Brinkmann and Weinzierl [19]. In preprocessing, the TOAs of the HRIRs are first detected by threshold-based onset detection and then removed using fractional delays. The time-aligned, complex HRTF spectra and the extracted TOAs are then interpolated separately to any desired (dense) sampling grid using Eqs. (3) and (4). Afterwards, the TOA is reconstructed in postprocessing using fractional delays once again.

We implemented the method as described by Brinkmann and Weinzierl [19] using onset detection with a threshold of −20 dB in relation to the maximum values of the 10 times upsampled and low-pass–filtered HRIRs (8th order Butterworth, $f_c = 3$ kHz, see [27]). The TOAs were removed and inserted in frequency domain using fractional delay filters and circular convolution, which has the advantage that the length of the HRIRs is not changed during the processing. The fractional delays were designed in the time domain using Kaiser windowed sinc filters of order 70 with a side lobe attenuation of 60 dB [28]. The filters exhibit negligible magnitude distortions <0.1 dB and group delay distortions <0.1 samples below 20 kHz.

2.2 Frequency-Dependent Time-Alignment

Zaunschirm et al. [15] presented a frequency-dependent time-alignment (FDTA) as HRTF preprocessing for binaural Ambisonics renderering. FDTA removes the high frequency TOA and thus also the ITD above 1.5 kHz and maintains it at low frequencies. Because the ITDs become less relevant as frequency increases [29], the authors proposed not to resynthesize the high-frequency ITDs for binaural reproduction of the Ambisonics signal. However, the alignment can easily be reversed to reconstruct HRTFs after SH interpolation.

In contrast to the onset-based time-alignment, FDTA does not aim to completely remove the TOAs. Instead, TOA differences between HRIRs are removed and a constant TOA remains. We refer to this as relative TOA alignment in the following.

The relative TOAs $\tau_q^{\mathrm{l,r}}$ are estimated from the time difference by which a plane wave from direction Ω_q arrives at the center of the head and the position of the ear

$$\tau_q^{\mathrm{r}} = \cos\theta_q \sin\phi_q r_0 c^{-1}, \qquad \tau_q^{\mathrm{l}} = -\tau_q^{\mathrm{r}}, \tag{6}$$

with $c = 343$ m/s the speed of sound, r_0 the head radius, and q a spatial sampling point of the HRTF. This inherently assumes that the ears are located at $\phi_{\mathrm{e}} = [90^\circ, 270^\circ]$ and $\theta_{\mathrm{e}} = [0^\circ, 0^\circ]$ and neglects diffraction around the head that might affect the actual TOA. The estimated relative TOAs are the basis for designing an all-pass filter $A_q^{l,r}(\omega)$ for each sampling point q, which is applied by multiplication in the frequency domain to achieve the relative TOA alignment. The filter is defined as

$$A_q^{\mathrm{l,r}}(\omega) = \begin{cases} 1 & \text{for } \omega < \omega_c \\ \mathrm{e}^{-\mathrm{i}(\omega-\omega_c)\tau_q^{\mathrm{l,r}}} & \text{for } \omega \geq \omega_c, \end{cases} \tag{7}$$

where $\omega_c = 2\pi f_c$ with the cut-on frequency $f_c = 1.5$ kHz. Thus, the filter exhibits a group delay of 0 below f_c and $\tau_q^{\mathrm{l,r}}$ above.

After SH interpolation of the time-aligned HRTFs to T desired directions Ω_t, the original ITDs can be reconstructed by reversing the alignment. Thus, all-pass filters for each direction t are calculated according to Eqs. (6) and (7) and applied by division in the frequency domain.

[1] Available: https://github.com/AudioGroupCologne/SUpDEq.

2.3 Spatial Upsampling by Directional Equalization

With Spatial Upsampling by Directional Equalization (SUpDEq), we recently presented a method using a rigid sphere as a simplified head model as the basis for the alignment [16, 30–33]. This has the advantage that scattering effects around the head are approximated. As with FDTA, SUpDEq aims at a relative TOA alignment.

In preprocessing, the sparse HRTF set $H(\omega, \Omega_q)$ with Q measurement directions is equalized by spectral division with rigid sphere transfer functions $H_\mathrm{R}(\omega, \Omega_q)$ described as

$$H_\mathrm{R}(\omega, \Omega_q) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n(kr_0) Y_n^m(\Omega_\mathrm{e}) Y_n^m(\Omega_q)^*, \qquad (8)$$

with Ω_e the left and right ear position. The scattering around the rigid sphere is accounted for by

$$d_n(kr_0) = 4\pi \mathrm{i}^n \left[j_n(kr_0) - \frac{j_n'(kr_0)}{h_n^{(2)'}(kr_0)} h_n^{(2)}(kr_0) \right], \qquad (9)$$

with j_n the spherical Bessel function of the first kind, $h_n^{(2)}$ the spherical Hankel function of the second kind, and j_n' and $h_n^{(2)'}$ their derivatives[2]. The rigid sphere transfer functions are calculated at a high spatial order $N \geq 40$ to avoid sparsity errors.

Because the TOA is contained in the spherical head model, the spectral division of the HRTF by H_R automatically yields the time-alignment and additionally aims at equalizing parts of the magnitude response. H_R may be considered as a simplified HRTF set comprising only basic temporal and spectral features. From an information theory point of view, the result of the equalization can thus be understood as the prediction error between the actual HRTFs and the spherical head model, which has a lower SH order than the original HRTF set.

The equalized HRTFs are interpolated in the SH domain to T desired directions Ω_t using Eqs. (3) and (4). In postprocessing, the interpolated HRTFs are de-equalized by spectral multiplication with rigid sphere transfer functions for the interpolated directions Ω_t to recover previously discarded temporal and spectral components of the HRTF. To maintain valid HRTF data, the equalization and de-equalization were applied in the present study only above the spatial aliasing frequency $f_A = N_s c / 2\pi r_0$, where N_s is the SH order of the sparse sampling grid [35]. This was done by setting $H_\mathrm{R}(\omega, \Omega_q) = 1$ for $0 \leq \omega \leq 2\pi f_A 2^{-1/3}$, where $2^{-1/3}$ represents a third-octave safety margin.

2.4 Phase-Correction

Ben-Hur et al. [20] presented a pre- and postprocessing technique called phase-correction (PC) that is conceptually similar to SUpDEq. In preprocessing, the HRTF set $H(\omega, \Omega_q)$ measured for Q sampling points is equalized by spectral division with open sphere transfer functions $H_\mathrm{O}(\omega, \Omega_q)$ for the corresponding directions Ω_q, given by

$$H_\mathrm{O}(\omega, \Omega_q) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n(kr_0) Y_n^m(\Omega_\mathrm{e}) Y_n^m(\Omega_q)^*, \qquad (10)$$

with

$$d_n(kr_0) = 4\pi \mathrm{i}^n j_n(kr_0) \qquad (11)$$

Compared to Eq. (9), the open sphere transfer function in Eq. (11) does not contain a scattering term and thus results in a frequency-independent time-alignment not accounting for magnitude effects. Therefore, the equalization can also be described as a frequency domain multiplication of the HRTF set $H(\omega, \Omega_q)$ with a phase-correction term (all-pass), which the authors defined as

$$C_q^{\mathrm{l,r}}(\omega) = \mathrm{e}^{-\mathrm{i}kr_0 \cos \Theta_q^{\mathrm{l,r}}}, \qquad (12)$$

where $\Theta_q^{\mathrm{l,r}}$ is the angle between the measured direction Ω_q and the left and right ear position Ω_e and $\cos \Theta_q^{\mathrm{l,r}} = \cos\theta_q \cos\theta_\mathrm{e} + \cos(\phi_q - \phi_\mathrm{e}) \sin\theta_q \sin\theta_\mathrm{e}$.

Applying the phase-correction to the HRTFs in preprocessing results in a time-aligned HRTF set with lower SH order (also referred to as ear-alignment in [20]). After SH interpolation of the phase-corrected HRTFs to T desired directions Ω_t using Eqs. (3) and (4), HRTFs can be reconstructed by applying the inverse phase correction. Thus, phase-correction terms for each direction t are calculated according to Eq. (12) and applied to the interpolated HRTFs by spectral division in the frequency domain.

3 PHYSICAL EVALUATION

The physical evaluation focuses on two aspects: The alignment and restoration of the TOAs as the main methodological difference between the algorithms, and the spectral distortion identified in a previous study as the most problematic artifact [19]. Both aspects are also highly relevant from a perceptual point of view: the TOA is directly related to the ITD, which is the main cue for left/right localization [29], while the perceived coloration and up/down localization errors are attributable to spectral distortions [36].

3.1 HRTFs

HRTFs from a Neumann KU100 measured on a Lebedev grid with 2,702 sampling points [5] were used as the reference allowing for SH interpolation of order $N = 44$ without any sparsity errors. Sparse HRTF sets were then generated by spatially subsampling the reference in the SH domain to Lebedev grids of order $1 \leq N \leq 15$ according to Eqs. (3) and (4). In the last step, the sparse sets were subjected to the processing methods introduced above. Throughout this study, a head radius of $r_0 = 9.19$ cm was used, calculated according to Algazi et al. [37], and the left and right ear position Ω_e required for SUpDEq and PC was defined with $\phi_\mathrm{e} = [90^\circ, 270^\circ]$ and $\theta_\mathrm{e} = [0^\circ, 0^\circ]$.

[2]Please note the dependency of Eq. (9) on the Fourier transform kernel [34, Table I]. We used $p(\omega) = \int_{-\infty}^{\infty} p(t) e^{-\mathrm{i}\omega t} \mathrm{d}t$ as the Fourier transform of the pressure signal $p(t)$.

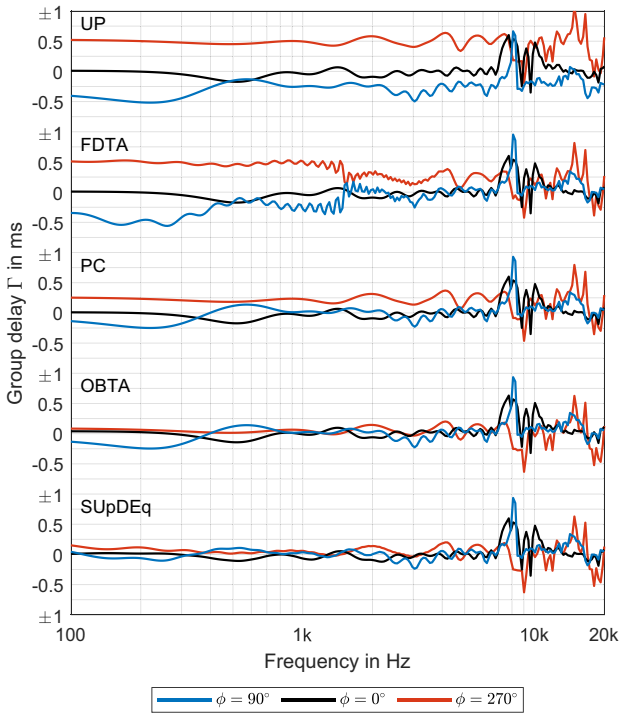

Fig. 1. Group delay of time-aligned HRTFs compared to unprocessed (UP) HRTFs for the left ear and three selected source positions in the horizontal plane ($\theta = 0°$).

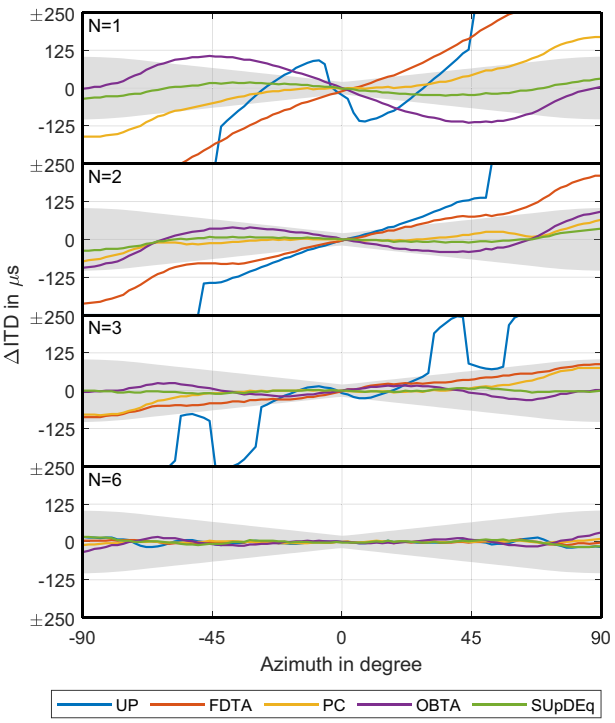

Fig. 2. Difference in ITD relative to the reference for the frontal region of the horizontal plane and selected SH orders N. The shaded area denotes the JND as a function of the reference ITD.

3.2 TOA Alignment

Perfectly aligned HRTFs would show a constant group delay independent of frequency and source position. To assess the performance of the alignment approaches, we calculated the HRTF group delay

$$\Gamma(\omega, \Omega_q) = -\frac{\mathrm{d}\angle H(\omega, \Omega_q)}{\mathrm{d}\omega} \tag{13}$$

for all Q measurement directions, where $\angle H(\cdot)$ is the unwrapped phase response. Because most methods perform a relative alignment, the group delay was centered around 0 ms by subtracting the overall mean separately for each method. Fig. 1 shows group delays of three selected HRTFs in the horizontal plane before the alignment (UP) and after the respective alignment. The unprocessed HRTFs show group delay differences of approximately 1 ms at low frequencies and 0.75 ms at high frequencies. Narrow group delay peaks occur for frequencies above 7 kHz caused by rapid phase changes due to HRTF notches (see also Fig. 3).

As expected, FDTA maintains the group delays below 1.5 kHz and aligns the data for higher frequencies. However, the preprocessing leads to ripples around 1.5 kHz, probably caused by the discontinuity in the alignment function defined in Eq. (7) and the finite HRIR length (Gibbs phenomenon, Chapter 7.5 of [38]). A smooth transition between the two states of the alignment function or windowing the time signal might reduce these ripples. Furthermore, FDTA fails in aligning the contralateral HRTF ($\phi = 270°$) between 1.5 and 8 kHz.

Results for PC are visually very similar to FDTA above 1.5 kHz—apart from the FDTA ripples—which is not surprising, as both methods estimate the TOA based on an open sphere geometry. Below 1.5 kHz, group delay differences of approximately 0.50 ms remain uncompensated because the open sphere TOA alignment does not account for low-frequency phase effects that occur due to the scattering around the head (a visualization of this effect is given in Fig. 2 of [39]).

Low-frequency group delay differences of approximately 0.25 ms remain for OBTA, which is about half of the differences observed for PC. The remaining differences below 1 kHz are mainly caused by the ipsilateral HRTF ($\phi = 90°$), which shows the strongest fluctuations in this range already for UP. Above 1 kHz, OBTA outperforms FDTA and PC due to a better alignment of the contralateral HRTF.

SUpDEq processing yields the smallest group delay deviations across source positions, reducing low-frequency group delay differences in Fig. 1 to about 0.125 ms. This improvement is clearly related to considering the scattering around the sphere in the alignment process. For frequencies above 1 kHz, SUpDEq and OBTA perform comparably well.

An additional analysis of the group delay standard deviation across all source positions, presented in the supplementary material (Fig. S1 of [40]), confirmed the trends observed for the three selected positions. SUpDEq outperforms all remaining methods up to approximately 1.5 kHz. Above 1.5 kHz, all alignment approaches produce comparable standard deviations that remain below that of the unprocessed HRTFs up to 20 kHz.

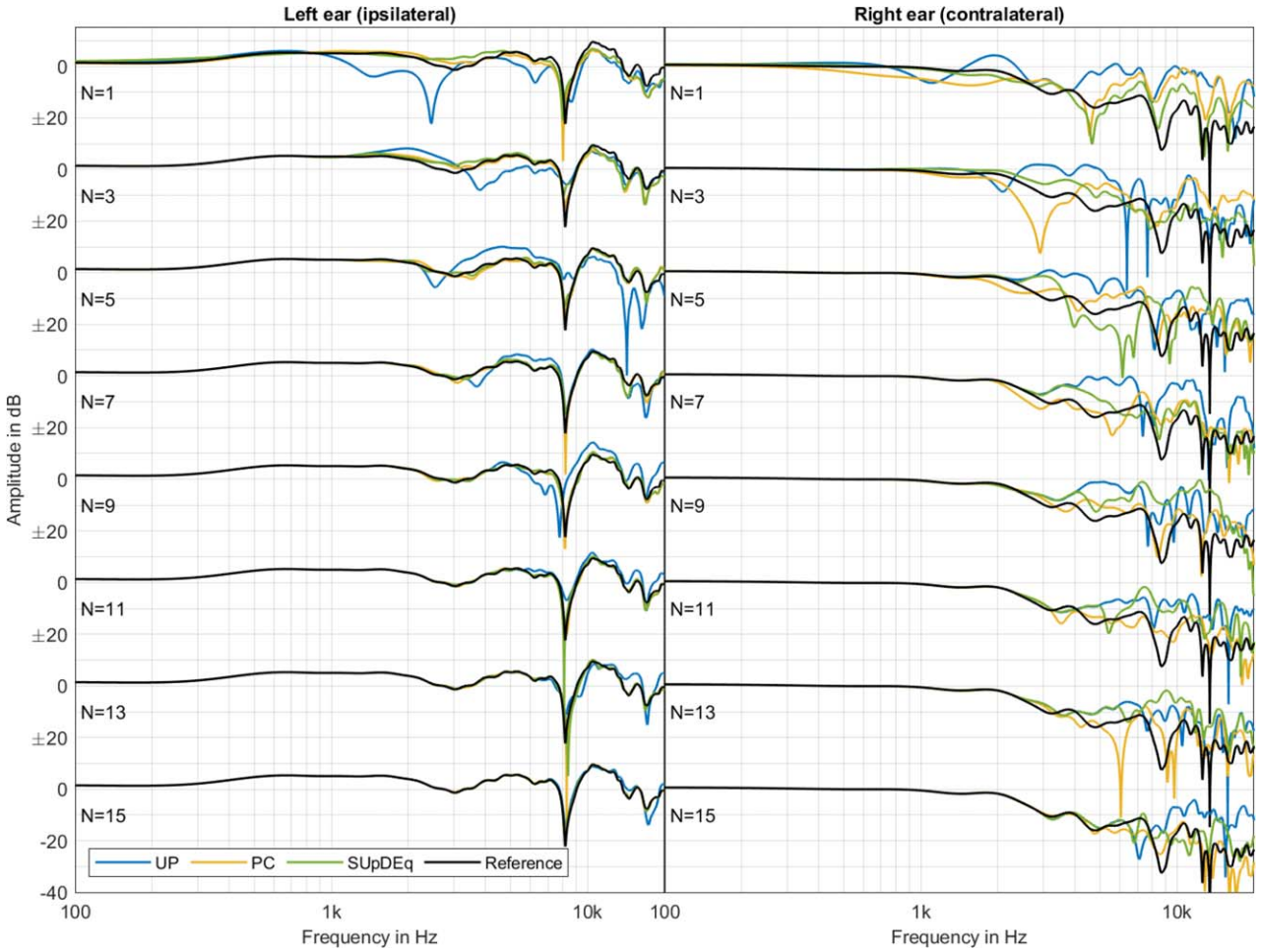

Fig. 3. Reference and interpolated HRTFs for a source at $\Omega = (90^\circ, 0^\circ)$, selected SH orders N and interpolation methods.

3.3 TOA Restoration

To assess the TOA restoration, the horizontal plane ITD was calculated from the difference between the left and right ear TOAs for HRTFs processed with all methods introduced above and SH orders $1 \leq N \leq 15$. The TOAs were estimated from the 10 times upsampled and low-passed HRIRs (8th order Butterworth, $f_c = 3kHz$, see [27]). A threshold of -10 dB was used for TOA detection in all cases. Using a threshold of -30 dB or -20 dB, as recommended by Andreopoulou and Katz [27], would lead to erroneous detections due to preringing in HRIRs processed at low SH orders [16]. Fig. 2 shows the results for selected SH orders by means of differences to the reference ITD for the frontal region of the horizontal plane (results for the rear were almost identical). The gray area denotes the broadband just noticeable difference (JND) as a function of the reference ITD [41]. The JND was linearly interpolated/extrapolated between 20 μs at $\mathrm{ITD_{ref}} = 0$ μs and 100 μs at $\mathrm{ITD_{ref}} = 700$ μs.

For first-order SH interpolation, only SUpDEq manages to keep the ITD errors below the JND, most likely due to the consideration of low-frequency scattering effects described above. While errors only slightly exceed the JND for OBTA and PC in this case, large errors are observed for UP and FDTA. For OBTA and PC, the errors fall below the JND at SH order two, while FDTA requires order three. Thus, starting at SH order three, all alignment methods perform comparably well and yield correct ITDs. However, UP still shows large errors at order three and sudden jumps that are caused by preringing in the HRIRs [16], which can, for example, be reduced by SH tapering [42]. At an SH order of six, the errors finally fall below the JND for all methods.

3.4 Spectral Distortion

To get a first impression of the spectral distortion, Fig. 3 shows HRTFs for two source position at selected SH orders and for selected methods (the supplementary material contains figures for all methods [40, Figs. S2–S4]). For the ipsilateral ear, the errors quickly decrease with increasing SH order and HRTFs are already quite similar to the reference at $N = 3$ for PC and SUpDEq. Results for UP are clearly worse, where high-frequency differences remain up to $N = 15$. Errors are generally larger for the contralateral ear and appear to be less predictable in this case. For example, SUpDEq shows a relatively small error at $N = 3$, where it outperforms UP and PC. At $N = 13$, however, the error for SUpDEq is larger than at $N = 3$ and SUpDEq is outperformed by UP and PC in this case.

For a more systematic analysis, the spectral distortion was calculated as the absolute energetic difference between

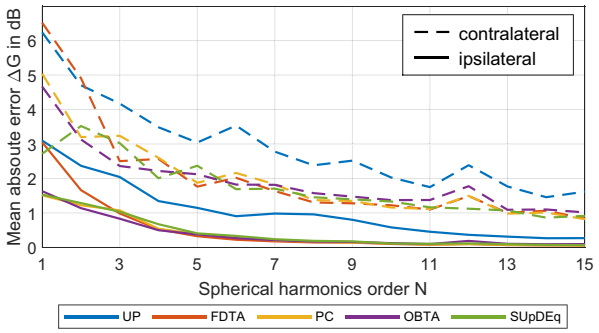

Fig. 4. Left ear energetic error ΔG vs. SH order averaged across frequency and source position in the ipsilateral and contralateral region.

interpolated HRTFs H_i and the reference H_r in 40 auditory filters as implemented in the Auditory Toolbox [43]

$$\Delta G(f_c, \Omega) = \left| 10 \log_{10} \frac{\sum_k C(f_k, f_c)|H_i(f_k, \Omega)|^2}{\sum_k C(f_k, f_c)|H_r(f_k, \Omega)|^2} \right|, \tag{14}$$

with $C(\cdot)$ the auditory filter and f_c the center frequency of the auditory filters and $50\,Hz \le f_k, f_c \le 20\,kHz$. The measure $\Delta G(f_c, \Omega)$ was calculated for 900 source positions on a Fliege sampling grid obtained with the SOFiA Toolbox [44]. In the following, averaged errors are denoted by omitting the corresponding symbol, i.e., $\Delta G(f_c)$ gives the error averaged across source position, $\Delta G(\Omega)$ is the frequency average, and ΔG is averaged across source positions and frequencies. Averaging across source positions was done using the quadrature weights α of the Fliege sampling grid.

Fig. 4 shows the left ear errors for ipsilateral and contralateral source regions for all methods and SH orders up to $N = 15$. The errors were obtained by averaging across source positions within 25° great circle distance from $\Omega_{\text{ipsi}} = (90°, 0°)$ and $\Omega_{\text{contra}} = (270°, 0°)$. The supplementary material contains another figure showing errors averaged across all source positions [40, Fig. S5]. Fig. 4 confirms the trends found above. Errors for the ipsilateral region are about 3 dB smaller than errors for the contralateral region at $N = 1$, and differences between the two regions slowly decrease to approximately 1 dB at $N = 15$. Moreover, the errors for the ipsilateral region decrease almost monotonically, which is not the case for the contralateral region. While UP clearly performs worst, results for the alignment methods are comparable, except that FDTA produces larger errors for $N \le 2$, and SUpDEq yields the lowest errors at $N = 1$, especially for the contralateral case. For $N \ge 3$, the differences between the methods diminish, and their performances become more and more similar.

To get a better impression of the spatial dependency of the spectral distortion, Fig. 5 shows $\Delta G(\Omega)$ for selected SH orders and SUpDEq. This shows that the region of large errors is generally small and quickly decreases with increasing SH order. For $N = 3$, frequency averaged errors above 3 dB are approximately found within a 45° radius cone around $\Omega = (270°, 0°)$, whereas the cone's radius

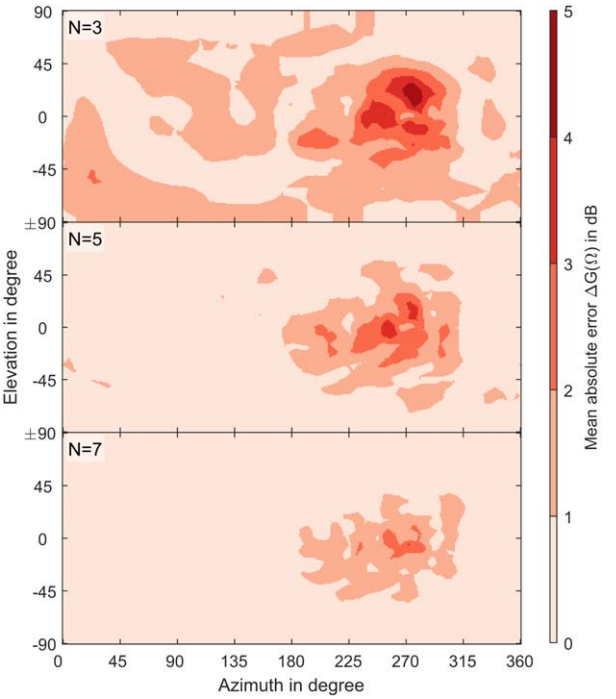

Fig. 5. Left ear energetic error $\Delta G(\Omega)$ for SUpDEq and selected SH orders N.

decreases to about 10° at $N = 7$. For comparison, the supplementary material provides similar plots for SH orders up to $N = 15$ and all methods [40, Figs. S6–S10], indicating similar behavior across the alignment methods.

4 PERCEPTUAL EVALUATION

The aim of the listening experiment was to determine the minimum required SH order N and thus the minimum required number of sampling points of a sparse HRTF set for which interpolated HRTFs are indistinguishable from the reference. To determine this so-called point of subjective equality (PSE), we implemented an adaptive ABX test, in which the SH order N of the sparse HRTF set is adapted according to the response of the subject. This was done for UP and SUpDEq as an example of the approaches discussed in Sec. 3 for three different test signals (noise, speech, and percussion) and two sound source positions (off-center frontal and lateral). We decided to test different source positions and audio content rather than different alignment methods because (first) the physical evaluation revealed that for $N \ge 3$ all methods perform very similarly in terms of TOA restoration and spectral distortion and (second) to limit the duration of the cognitively demanding ABX listening test. We hypothesized that SUpDEq processing generally leads to lower PSEs, that the test signal has a significant influence on the PSEs, and that the lateral sound source position leads to higher PSEs than the frontal position.

4.1 Participants

A total of 32 participants between 21 and 49 years of age ($M = 27.31$ years, $Mdn = 26$ years, $SD = 5.69$) took part

in the experiment for monetary remuneration of €15 per hour. Most of them were students in media technology or electrical engineering. Of those, 23 participants (72%) had already taken part in previous listening experiments and were thus familiar with the dynamic binaural reproduction system and the test environment. All participants had self-reported normal hearing.

4.2 Setup

The experiment was conducted in the sound insulated anechoic chamber of TH Köln, Köln, Germany. Participants were seated on an office chair with a mount holding a tablet computer at eye level about 0.50 m away in front of them, with which the responses were given. We used the MATLAB-based software Scale [45] to implement, control, and execute the experiment. For dynamic binaural rendering, we employed a customized version of the SoundScape Renderer [46], which is capable of loading Spatially Oriented Format for Acoustics (SOFA) files [47] with an arbitrary sampling grid. For three-degrees-of-freedom head tracking (yaw, pitch, and roll), we used a Polhemus Fastrack with 120 Hz update rate.

As digital-to-analog converter and headphone amplifier, we employed an RME Fireface UFX audio interface, and for playback, we used Sennheiser HD600 headphones. To minimize the influence of the headphones, we applied a generic headphone compensation filter, which was designed as a minimum phase finite impulse response filter with 2,048 taps using regularized inversion [48]. The playback level was adjusted to $L_{eq} = 65$ dB(A). The audio interface was set to a buffer size of 256 samples at a sampling rate of 48 kHz. With these settings, the measured overall latency of the system is about 37 ms [49], which is well below assessed thresholds of just detectable system latency of about 60–70 ms [50].

4.3 Stimuli

To obtain HRTFs for the listening test, the reference HRTFs (see Sec. 3.1) were subsampled to Gaussian grids of SH order $1 \leq N \leq 44$ using Eqs. (3) and (4). We chose the Gaussian grid because the order can be increased linearly. In a second step, UP and SUpDEq were used to interpolate HRTFs to a full-spherical spatial sampling grid with a resolution of 1° in horizontal direction and 5° in vertical direction. As with the physical evaluation in Sec. 3, we used an optimal radius of $r_0 = 9.19$ cm and the left and right ear position $\phi_e = [90^\circ, 270^\circ]$ and $\theta_e = [0^\circ, 0^\circ]$ for the spherical head model applied in SUpDEq.

We chose $\Omega = (330^\circ, 0^\circ)$ and $\Omega = (90^\circ, 0^\circ)$ as nominal sound source positions, first to examine PSEs for a frontal position that still contains at least small binaural cues and second to investigate the more critical lateral source position, which was shown to lead to significant artifacts at the contralateral ear, even with preprocessing (see also Sec. 3). As anechoic test signals, we employed a pink noise burst with a length of 0.75 s (including 10-ms cosine-squared onset/offset ramps), a male speech sample of a German sentence with a length of 1.5 s, and a castanet percussion sequence of 1.5 s length. The noise burst represents the most critical test signal with respect to coloration and localization, while speech and castanets are less critical due to the fluctuating spectral content and in the case of speech also due to the natural band limitation. However, percussion and speech signals are more relevant for real-life applications than noise[3].

4.4 Procedure

The experiment was based on an ABX test, that is, a three-interval/two-alternative forced choice (3I/2AFC) paradigm, combined with an adaptive one-up one-down staircase procedure (Chapter 3 and Chapter 5 of [51]). This simple and robust method [52] is free of restrictive assumptions, widely used in psychophysics, and was found to be a good choice to obtain the PSE [53] (i.e., the 50% point on the psychometric function also referred to as the threshold of recognition). Since perceptual differences between HRTFs interpolated from different SH orders are certainly not interval-scaled, more efficient maximum-likelihood procedures such as QUEST [54] could not be used.

According to the $3 \times 2 \times 2$ within-subjects factorial design with the factors *test signal* (noise, speech, and percussion), *method* (UP and SUpDEq), and *sound source position* ($\Omega = (330^\circ, 0^\circ)$ and $\Omega = (90^\circ, 0^\circ)$), each participant had to perform 12 runs. Following the ABX paradigm, a sequence of three intervals was presented at each trial, with X always being played second to ensure direct comparability between the stimuli (the actual playback order was therefore AXB). The middle interval (X) was randomly assigned to the reference HRTF set (A) or the sparse HRTF set (B), resulting in the four possible sequences AAB, BAA, ABB, or BBA. After the sequence was presented, participants had to report whether the first (A) or the third (B) interval was equal to the second (X) interval by pressing the corresponding button on the graphical user interface displayed on the tablet. The three buttons labeled A, X, and B were arranged on a horizontal line and flashed green when the corresponding interval was played. However, the X button was deactivated to prevent wrong entries. Participants could neither repeat a trial nor continue without giving an answer.

If the response was correct, the SH order of the sparse HRTFs was increased by one in the next trial and decreased by one otherwise. Each run started at $N = 1$ to provide clear perceptual differences to the participants. A run was terminated when 16 reversals occurred, where a reversal is defined as a point where a series of steps changes from increasing to decreasing the SH order or vice versa.

Before starting the experiment, participants were briefly introduced to dynamic binaural synthesis and were given instructions about the experimental procedure. They were encouraged to perform small head movements when they felt that this made them more sensitive to differences. To maintain differences between the two nominal source positions, they were additionally instructed to keep their main

[3] Static binaural renderings of the stimuli are part of the supplementary material [40].

Table 1. Mean PSEs across subjects and 95 % between-subjects confidence intervals (CIs) of the means for all tested conditions.

	$\Omega = (330°, 0°)$			$\Omega = (90°, 0°)$		
	Noise	Speech	Perc	Noise	Speech	Perc
Mean PSEs						
Unprocessed	18.27	12.49	13.28	24.29	19.37	18.97
SUpDEq	10.27	6.05	6.92	21.92	17.79	16.55
95% CIs						
Unprocessed	±1.73	±1.61	±1.18	±2.03	±1.42	±1.15
SUpDEq	±1.55	±1.12	±1.12	±1.98	±1.64	±1.90

Table 2. Results of the three-way repeated measures ANOVA with the within-subjects factor test signal (S), method (M), and sound source position (P).

Source	*df*	*F*	*MSE*	ε	η_p^2	*p*
S	2 , 62	48.59	19.90	1	.61	<.001*
M	1 , 31	101.48	19.37	1	.77	<.001*
P	1 , 31	272.92	26.02	1	.90	<.001*
S × M	2 , 62	.89	13.02	1	.03	.416
S × P	2 , 62	1.99	11.64	.97	.06	.147
M × P	1 , 31	58.45	9.49	1	.65	<.001*
S × M × P	2 , 62	.54	10.68	.92	.02	.573

ε, Greenhouse–Geisser (GG) epsilon; *p*, GG–corrected *p*-values. Note that GG correction is appropriate only for within-subject tests with more than one degree of freedom in the numerator.

line of vision straight ahead and were not allowed to rotate their body. The experimenter visually monitored the participants with a camera to ensure that they did not disregard the instructions. In order to get familiar with the setup and the test procedure, the participants had to do a short training session before the actual experiment, which consisted of two runs terminated when eight reversals occurred, one with the noise and one with the speech signal. In total, each session lasted for about 45 to 60 min, including the verbal instruction, the training session, and a break after half of the runs.

4.5 Data Analysis

To calculate the PSEs, the first reversal was omitted (Chapter 7 of [55]), and thus, the PSE estimate was calculated as the averaged N across the last 15 reversals. Visual inspection of the data and Shapiro–Wilk tests for normality, corrected for multiple hypothesis testing according to Hochberg [56], showed no considerable violations of normality (see also [40, Fig. S11]). We thus analyzed the determined PSEs using a three-way repeated measures ANOVA with Greenhouse–Geisser (GG) correction [57] and the within-subjects factors test signal, method, and sound source position. For a more detailed analysis, we conducted a nested GG-corrected repeated measures ANOVA as well as Hochberg-corrected paired t tests (two-tailed) at a 0.05 significance level.

4.6 Results

Table 1 lists the mean PSEs across subjects as well as the 95% between-subjects confidence intervals of the means for all tested conditions. The graphical overview of the data in Fig. 6 shows the interindividual variation in the determined PSEs (left panel) and the mean PSEs across subjects (right panel). The plots clearly support our three initial hypotheses, which are statistically confirmed by the ANOVA summarized in Table 2.

PSEs for SUpDEq are significantly lower than for UP resulting in a drastic decrease of the minimum number of measurement directions required to obtain SH interpolated HRTFs that are indistinguishable from the reference. The strong main effect of method revealed by the ANOVA statistically confirms this finding (Table 2, row M).

The sound source position has a strong influence on the PSEs. With both methods, the minimum required SH order increases significantly for the lateral source (90°, 0°) compared to the frontal (330°, 0°). The ANOVA yielded a strong main effect of source position (effect size $\eta_p^2 = 0.90$, see Table 2, row P) and thus statistically confirms the high perceptual relevance of the sound source position. Furthermore, the benefit of SUpDEq is smaller for the lateral position than for the frontal position, which is confirmed by the significant interaction effect between method and sound source position (Table 2, row M × P). Nevertheless, a nested ANOVA for the six lateral conditions showed a significant main effect of method suggesting that SUpDEq still provides improvements for the lateral position ($F(1,31) = 11.44$, $p = .002$, $\eta_p^2 = .27$, $\varepsilon = 1$).

Regardless of the method, the test signal has a strong influence on the PSEs, which is clearly demonstrated by the significant main effect of test signal revealed by the ANOVA (Table 2, row S). The speech and castanet signals require lower SH orders than the more critical noise signal. Paired t tests at each factor level of method and sound source position (e.g., Noise/UP/(330°, 0°) vs. Speech/UP/(330°, 0°)) confirmed that the PSEs for speech and castanets are always significantly lower than for noise (all $p < .001$). However, similar comparisons between speech and castanets showed no significant differences (all $p > .27$), indicating that both test signals are similarly critical.

5 DISCUSSION

5.1 Comparison Between Algorithms

The physical evaluation in Sec. 3 showed that HRTFs interpolated with the four investigated time-alignment methods are comparable in most cases. However, considerable differences were found in two cases. First, there are differences in the alignment and TOA restoration at low SH orders of $N \leq 2$. SUpDEq performs best in this case, presumably because it correctly models low-frequency phase effects involved in the diffraction around the sphere/head. Second, spectral differences at contralateral source positions remain up to SH orders of $N > 15$. In this region, the HRTF spectra exhibit fast changes across space, which requires higher SH orders for a physically correct interpolation. Caused by the insufficient SH order, aliasing errors

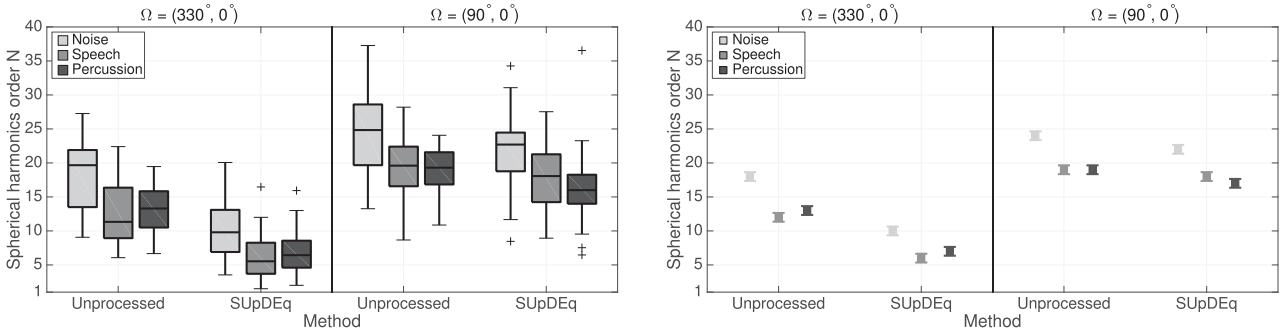

Fig. 6. Interindividual variation in the determined PSEs (left panel) and the mean PSEs across subjects (right panel) as a function of the method (abscissa), the test signal (shades of gray), and the sound source position (left or right half of each panel). The box plots (left panel) show the median and the (across participants) interquartile range (IQR) per condition; whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers are indicated by plus signs. The error bars in the mean plots (right panel) display 95 % within-subjects confidence intervals [58, 59], based on the error term of the respective main effect of method.

occur that drastically differ between algorithms due to differences in the aligned magnitude *and* phase spectra.

5.2 Required SH Order for SUpDEq

The results of the listening experiment in Sec. 4 clearly show the advantages of time-alignment—SUpDEq, in this particular case—compared to SH interpolation of unprocessed HRTFs. Using SUpDEq, a sparse HRTF set with an SH order of $N \approx 7$ was sufficient for speech and castanet content presented from the front $\Omega = (330°, 0°)$ to achieve a binaural rendering that is indistinguishable from the reference. Without preprocessing, this requires an order of $N \approx 13$ and thus about three times more HRTFs ($14^2/8^2$). The pink noise signal presented from the front resulted in mean PSEs of $N \approx 10$ using SUpDEq and $N \approx 18$ without processing. In informal discussions after the experiment, the participants named high-frequency spectral differences as being the dominant cue for distinguishing the stimuli in the direct comparison. A broadband noise signal thus causes stronger perceptual differences and higher PSEs than band-limited speech and spectrally fluctuating castanets.

The lateral direction $\Omega = (90°, 0°)$ showed to be much more critical than the frontal direction. For speech and castanets, the mean PSEs were in the range $16 \leq N \leq 18$ for SUpDEq and $19 \leq N \leq 20$ for UP. For noise, the mean PSEs further increased to $N \approx 22$ using SUpDEq and $N \approx 24$ for the unprocessed case. The statistical analysis still showed a significant improvement with SUpDEq processing, but the benefits were much smaller than for frontal sound incidence.

Based on the physical evaluation in Sec. 3 and our previous study [16], we expected higher PSEs for lateral sound incidence due to the increased spectral distortions in the contralateral region. The distortion is caused by distinct magnitude interference patterns in the contralateral HRTF that change strongly even for small changes in the source position. This results in high SH orders in the contralateral region that cannot be reduced by means of time-alignment and causes sparsity errors in the interpolated HRTFs. Since the aliasing component of the sparsity error heavily depends on the sampling grid, these errors do not decrease monotonically with order, as can be seen in Figs. 3 and 4. As a result, the interpolated HRTFs show different interference patterns that are clearly distinguishable from the reference in a direct comparison, even more so with small head movements.

It should be kept in mind that the PSE is the most demanding quality criterion and that many applications do not require HRTFs that are indistinguishable from the reference. In the highly critical listening experiment, participants were able to suppress the nearly error-free signals at the ipsilateral "louder" ear and exploit spectral distortions at the contralateral "quieter" ear to distinguish between reference and SH-interpolated HRTFs. However, it is reasonable to assume that the perceived coloration is dominated by the "louder" ipsilateral ear and that spectral distortions at the contralateral ear are often less critical in reference-free listening.

In addition, the largest errors are contained in a narrow cone with a radius of approximately 10° already for an SH order of $N = 7$ (see Fig. 5). Because SUpDEq correctly models the ITD—the main localization cue in the horizontal plane [29]—already at order $N = 1$ (see Fig. 2), the left/right localization should not be problematic, even for lateral source positions. Although up/down localization relies on spectral cues [3], results from listening tests [23] and auditory modeling [19] suggest that an SH order of $N = 4$ maintains enough spectral detail for this task. Accordingly, coloration and localization, which are perhaps the two most important quality aspects besides the PSE, should be sufficiently good even for SH orders that are lower than the values determined with the present listening experiment.

Due to the similarity between the algorithms observed in the physical evaluation, it appears reasonable to assume that results obtained in the perceptual evaluation for SUpDEq also apply (approximately) to the other methods. However, more perceptual studies are required to generalize the results, and different thresholds might be found, especially for lateral sources.

5.3 Comparison to Previous Work

A comparison of our results with other studies is not directly possible because, to the best of our knowledge, there is no other study that has estimated PSEs for SH interpolated HRTFs. Using a 2AFC test, Pike and Tew [18] showed that SH-based HRTF interpolation with and without OBTA is indistinguishable from the reference at $N = 35$. In general, our results support the findings of Pike and Tew, even though one participant in our experiment achieved a PSE of $N \approx 37$ for the condition Noise/UP/(90°, 0°). However, the 95th percentile of this condition is $N \approx 32$, so it can be assumed that using $N = 35$ is sufficient for most listeners.

Using a MUSHRA test, Pike and Tew further showed that time-alignment of a sparse HRTF set with $N = 5$ reduces perceptual differences for a frontal source position, whereas a lateral source at $\Omega = (260°, 0°)$ still produces significant perceptual differences. This agrees with our analysis in Fig. 5, where the lateral source tested by Pike and Tew lies in the region of the largest spectral errors. It is also interesting to note that the frontal sources in the MUSHRA study of Pike and Tew received median quality ratings of about 90% in the case of time-aligned HRTF interpolation and a pink noise test signal. The fact that in the present experiment the median PSE for a similar condition was $N \approx 10$ further supports our assumption that quality-based listening experiments lead to lower minimum required SH orders of sparse HRTF sets.

5.4 Future Work

The physical and perceptual evaluation showed that spectral errors in the contralateral region remain the main challenge for time-alignment–based SH interpolation of HRTFs. Even if the phase components were perfectly eliminated, high SH orders were still necessary to describe the complex interference structure of the HRTF magnitude. To decrease the error in this region, (de-)equalization functions that approximate the HRTF better than the spherical head model used with SUpDEq might help to decrease the error in this region. Furthermore, a qualitative listening test to compare different alignment approaches would be interesting to assess the extent to which the differences discovered in the physical evaluation affect auditory perception.

6 CONCLUSION

In this paper, we performed a physical evaluation of four approaches for SH interpolation of time-aligned HRTFs and a perceptual evaluation of one selected time-alignment approach, namely the SUpDEq method. The systematic comparison showed the similarity of the different pre- and postprocessing techniques. For this reason, it is not surprising that the physical evaluation revealed that all methods perform similarly well in mitigating sparsity and reconstruction errors that occur in SH interpolation of unprocessed HRTFs. However, the analysis also showed that all discussed methods have drawbacks in the region around the contralateral ear.

The listening experiment showed the perceptual benefits of time-alignment on the example of the tested SUpDEq method. In all tested conditions, the minimum SH order required to achieve indistinguishability from a reference was significantly smaller than for SH interpolation without preprocessing. The results suggest that with an SH order of $N \approx 7$ (at least 64 measurement directions), interpolated HRTFs will be indistinguishable or close to indistinguishable from the reference for source positions in the vicinity of the median plane, while perceptual differences will be negligible for most remaining source positions and applications in spatial audio[4]. At order $N = 7$, the physical evaluation showed similar results for all tested methods. Thus, computationally less-demanding methods as PC and FDTA might be preferred in this case. However, differences in low-order processing still exist, and SUpDEq showed the lowest errors when using first-order HRTF sets.

7 ACKNOWLEDGMENT

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the support code 03FH014IX5-NarDasS. We would like to thank Dr. Heinrich R. Liesefeld (LMU Munich, Department of Psychology) for his advice on the experimental design and the statistical analysis. We would also like to thank all participants of the listening experiment.

8 REFERENCES

[1] M. Vorländer, *Auralization* (Springer-Verlag, Berlin, Germany, 2008), http://doi.org/10.1007/978-3-540-48830-9.

[2] A. Roginska and P. Geluso, *Immersive Sound—The Art and Science of Binaural and Multi-Channel Audio* (Routledge, New York, NY, 2018), https://doi.org/10.4324/9781315707525.

[3] J. Blauert, *Spatial Hearing—The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1996).

[4] W. G. Gardner and D. M. Keith, "HRTF Measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908 (1995 Jun.), https://doi.org/10.1121/1.412407.

[5] B. Bernschütz, "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, pp. 592–595 (2013).

[6] F. Brinkmann, A. Lindau, S. Weinzierl, S. Van De Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc.*, vol. 65, no. 10, pp. 841–848 (2017 Oct.), https://doi.org/10.17743/jaes.2017.0033.

[7] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The CIPIC HRTF Database," in *Proceedings of the IEEE Workshop on the Applications of Signal Pro-*

[4]Compare static binaural renderings provided in the supplementary material [40].

cessing to Audio and Acoustics, pp. 99–102 (2001 Oct.), https://doi.org/10.1109/ASPAA.2001.969552.

[8] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718 (2019 Sep.), https://doi.org/10.17743/jaes.2019.0024.

[9] J.-G. Richter, *Fast Measurement of Individual Head-Related Transfer Functions*, doctoral dissertation, RWTH Aachen (2019), https://doi.org/10.30819/4906.

[10] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights Into Head-Related Transfer Function: Spatial Dimensionality and Continuous Representation," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2347–2357 (2010), https://doi.org/10.1121/1.3336399.

[11] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness Stability of Binaural Sound with Spherical Harmonic Representation of Sparse Head-Related Transfer Functions," *EURASIP J. Audio Speech Music Process.*, vol. 2019, no. 5, pp. 1–14 (2019), https://doi.org/10.1186/s13636-019-0148-x.

[12] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143 (2005 Jan.), https://doi.org/10.1109/TSA.2004.839244.

[13] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, doctoral dissertation, TU Berlin (2016), http://dx.doi.org/10.14279/depositonce-5082.

[14] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. M. Arend, "Binaural Reproduction of Plane Waves With Reduced Modal Order," *Acta Acust. united Ac.*, vol. 100, no. 5, pp. 972–983 (2014 Sep./Oct.), https://doi.org/10.3813/AAA.918777.

[15] M. Zaunschirm, C. Schoerkhuber, and R. Hoeldrich, "Binaural Rendering of Ambisonic Signals by HRIR Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018), https://doi.org/10.1121/1.5040489.

[16] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 1060–1071 (2019 Jun.), https://doi.org/10.1109/TASLP.2019.2908057.

[17] 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 (1998), https://doi.org/10.1121/1.423749.

[18] C. W. Pike, *Evaluating the Perceived Quality of Binaural Technology*, doctoral dissertation, University of York (2019).

[19] F. Brinkmann and S. Weinzierl, "Comparison of Head-Related Transfer Functions Pre-Processing Techniques for Spherical Harmonics Decomposition," presented at the *AES International Conference on Audio for Virtual and Augmented Reality* (2018 Aug.), conference paper P9-3.

[20] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient Representation and Sparse Sampling of Head-Related Transfer Functions Using Phase-Correction Based on Ear Alignment," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2249–2262 (2019 Dec.), https://doi.org/10.1109/TASLP.2019.2945479.

[21] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in Proceedings of the *44th DAGA*, pp. 339–342 (2018).

[22] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording* (Springer, Cham, Switzerland, 2019), https://doi.org/10.1007/978-3-030-17207-7.

[23] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 921–930 (2015 Aug.), https://doi.org/10.1109/JSTSP.2015.2421876.

[24] G. Dagan, N. R. Shabtai, and B. Rafaely, "Spatial Release from Masking for Binaural Reproduction of Speech in Noise with Varying Spherical Harmonics Order," *Appl. Acoust.*, vol. 156, pp. 258–261 (2019 Dec.), https://doi.org/10.1016/j.apacoust.2019.07.015.

[25] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer-Verlag, Berlin, Germany, 2015), https://doi.org/10.1007/978-3-662-45664-4.

[26] J. Ahrens, *Analytic Methods of Sound Field Synthesis* (Springer-Verlag, Berlin, Germany, 2012), https://doi.org/10.1007/978-3-642-25743-8.

[27] A. Andreopoulou and B. F. G. Katz, "Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 588–598 (2017), https://doi.org/10.1121/1.4996457.

[28] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the Unity Delay," *IEEE Sig. Proc. Mag.*, vol. 13, no. 1, pp. 30–60 (1996 Jan.), https://doi.org/10.1109/79.482137.

[29] F. L. Wightman and D. J. Kistler, "The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661 (1992), https://doi.org/10.1121/1.402445.

[30] C. Pörschmann and J. M. Arend, "Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments," presented at the *AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 15.

[31] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Spatial Upsampling of Individual Sparse Head-Related Transfer Function Sets by Directional Equalization," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 4870–4877 (2019), http://doi.org/10.18154/RWTH-CONV-239484.

[32] J. M. Arend and C. Pörschmann, "Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field

Datasets," in *Proceedings of the 45th DAGA*, pp. 1454–1457 (2019).

[33] J. M. Arend and C. Pörschmann, "Spatial Upsampling of Sparse Head-Related Transfer Function sets by Directional Equalization—Influence of the Spherical Sampling Scheme," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 2643–2650 (2019), http://doi.org/10.18154/RWTH-CONV-238939.

[34] V. Tourbabin and B. Rafaely, "On the Consistent Use of Space and Time Conventions in Array Processing," *Acta Acust. united Ac.*, vol. 101, no. 3, pp. 470–473 (2015 May/Jun.), https://doi.org/10.3813/AAA.918843.

[35] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial Aliasing in Spherical Microphone Arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 55, no. 3, pp. 1003–1010 (2007 Mar.), https://doi.org/10.1109/TSP.2006.888896.

[36] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2014), https://doi.org/10.1121/1.4887447.

[37] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a Spherical-Head Model from Anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479 (2001 Jun.).

[38] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. (Pearson Higher Education, Inc., Upper Saddle River, NJ, 2010).

[39] V. Benichoux, M. Rébillat, and R. Brette, "On the Variation of Interaural Time Differences with Frequency," *J. Acoust. Soc. Am.*, vol. 139, no. 4, pp. 1810–1821 (2016), https://doi.org/10.1121/1.4944638.

[40] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions – Supplementary Material," (2021), https://doi.org/10.5281/zenodo.4289971.

[41] J. E. Mossop and J. F. Culling, "Lateralization for Large Interaural Delays," *J. Acoust. Soc. Am.*, vol. 104, no. 3, pp. 1574–1579 (1998), https://doi.org/10.1121/1.424369.

[42] C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265 (2019), https://doi.org/10.1109/ICASSP.2019.8683751.

[43] M. Slaney, "Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work - Technical Report #1998-010" (1998).

[44] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA Sound Field Analysis Toolbox," in *Proceedings of the International Conference on Spatial Audio (ICSA)*, pp. 8–16 (2011).

[45] A. Vazquez Giner, "Scale—Conducting Psychoacoustic Experiments with Dynamic Binaural Synthesis," in *Proceedings of the 41st DAGA*, pp. 1128–1130 (2015).

[46] M. Geier, J. Ahrens, and S. Spors, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," presented at the *124th Convention of the Audio Engineering Society* (2008 May), convention paper 7330.

[47] Audio Engineering Society, "AES69-2015: AES Standard for File Exchange—Spatial Acoustic Data File Format" (2015).

[48] V. Erbes, M. Geier, H. Wierstorf, and S. Spors, "Free Database of Low-Frequency Corrected Head-Related Transfer Functions and Headphone Compensation Filters," presented at the *127th Convention of the Audio Engineering Society* (2017 May), eBrief 325.

[49] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 9.

[50] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," in *Proceedings of the 35th DAGA*, pp. 1063–1066 (2009).

[51] F. A. A. Kingdom and N. Prins, *Psychophysics: A Practical Introduction*, 1st ed. (Academic Press, London, United Kingdom, 2009), https://doi.org/10.1016/C2012-0-01278-1.

[52] H. Levitt, "Transformed Up-Down Methods in Psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 467–477 (1971), https://doi.org/10.1121/1.1912375.

[53] T. S. Meese, "Using the Standard Staircase to Measure the Point of Subjective Equality: A Guide Based on Computer Simulations," *Perc. Psychophys.*, vol. 57, no. 3, pp. 267–281 (1995), https://doi.org/10.3758/bf03213053.

[54] A. B. Watson and D. G. Pelli, "QUEST: A Bayesian Adaptive Psychometric Method," *Perc. Psychophys.*, vol. 33, no. 2, pp. 113–120 (1983), https://doi.org/10.3758/bf03202828.

[55] S. A. Gelfand, *Hearing—An Introduction to Psychological and Physiological Acoustics*, 6th ed. (CRC Press, Boca Raton, FL, 2017).

[56] Y. Hochberg, "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, vol. 75, no. 4, pp. 800–802 (1988 Dec.), https://doi.org/10.1093/biomet/75.4.800.

[57] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," *Rev. Educ. Res.*, vol. 42, no. 3, pp. 237–288 (1972), https://doi.org/10.3102/00346543042003237.

[58] G. R. Loftus and M. E. J. Masson, "Using Confidence Intervals in Within-Subject Designs," *Psychon. Bull. Rev.*, vol. 1, no. 4, pp. 476–490 (1994), https://doi.org/10.3758/bf03210951.

[59] J. Jarmasz and J. G. Hollands, "Confidence Intervals in Repeated-Measures Designs: The Number of Observations Principle." *Can. J. Exp. Psychol.*, vol. 63, no. 2, pp. 124–138 (2009), https://doi.org/10.1037/a0014164.

THE AUTHORS


Johannes M. Arend


Fabian Brinkmann


Christoph Pörschmann

Johannes M. Arend received a B.Eng. degree in media technology from HS Düsseldorf (Germany) in 2011 and an M.Sc. degree in media technology from TH Köln, Köln, Germany, in 2014. Since 2015, he has been a Research Fellow and working toward a Ph.D. at TH Köln, Köln, Germany, and TU Berlin, Berlin, Germany, in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing.

•

Fabian Brinkmann received an M.A. degree in communication sciences and technical acoustics in 2011 and a Dr. rer. nat. degree in 2019 from the Technical University of Berlin, Berlin, Germany. He focuses on the fields of signal processing and evaluation approaches for spatial audio.

•

Christoph Pörschmann studied Electrical Engineering at the Ruhr-Universität Bochum, Bochum, Germany, and Uppsala Universitet, Uppsala, Sweden. In 2001, he obtained his Dr.-Ing. degree from the Electrical Engineering and Information Technology Faculty of the Ruhr-Universität Bochum as a result of his research at the Institute of Communication Acoustics. Since 2004, he has been Professor of Acoustics at TH Köln, Köln, Germany. His research interests are in the field of virtual acoustics, spatial hearing, and the related perceptual processes.