Binaural Rendering of Ambisonic Signals via Magnitude Least Squares

Christian Schörkhuber, Markus Zaunschirm, Robert Höldrich
Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz
schoerkhuber@iem.at

Introduction

Binaural rendering of order-limited Ambisonic signals is an active research area due to widespread adoption of the Ambisonic format for headphone-based reproduction of spatial audio content. When signal independent methods are considered, the problem is to find an optimal filter matrix which maps $J = (N+1)^2$ N-th order Ambisonic signals to the two ear signals. The two challenges in the underlying optimization process are (i) how to define a cost function that encodes perceptual dissimilarity, and (ii) how to find the global minimizer. In recent studies [1, 2, 3] it has been shown, that minimizing the squared error between the order N approximated head-related transfer functions (HRTFs) and a large set of measured/modeled head-related transfer functions (HRTFs) is a poor choice for low orders, as severe direction-dependent timbral artifacts are introduced. The observed rapid roll-off at high frequencies for frontal sources can be reduced by a global correction filter as suggested in [2], however, significant direction-dependent signal colorations remain. In [1] it was shown that signal colorations can be reduced by reducing the set of directions in the cost function. This approach is reminiscent of two-staged binaural rendering methods, where Ambisonic signals are first decoded to a set of virtual loudspeakers and then filtered with HFTFs corresponding to the loudspeaker directions. While these methods can rely on a rich body of research concerning loudspeaker-based reproduction of Ambisonic signals, it is unclear if these methods are optimal for headphone-based reproduction.
A rendering method specific to headphone-based reproduction is proposed in [3], where the colorations of rendered signals are significantly reduced by removing the linear phase from all HRTFs at higher frequencies prior to optimization. This approach is perceptually motivated as it can be assumed that altering the interaural phase difference (IPD) at higher frequencies is perceptually irrelevant [4]. Indeed, listening experiments conducted in [3] revealed, that the high-frequency phase modifications achieve a significant quality improvement over state-of-the-art methods.

The contribution of this work is twofold: Firstly, we evaluate the perceptual impact of high frequency phase modifications for different cut-on frequencies and signal types; and secondly, we show that the phase modification proposed in [3] can be viewed as an approximate solution of a noisy phase retrieval problem [5], and that the rendering quality can be further increased by solving the problem exactly.

Notation and Problem Formulation

In a nutshell, Ambisonic signals can be interpreted as the signals recorded by a set of virtual coincident microphones with directivity patterns that are proportional to spherical harmonics[1] up to some order $N \ll \infty$. That is, the j-th Ambisonic signal due to a plane wave signal $s_\Omega(\omega)$ from direction $\Omega = (\varphi, \theta)$, where φ, θ are the azimuth and elevation angle, respectively, is given by $a_j(\omega) = s_\Omega(\omega) Y_j(\Omega)$, where ω denotes frequency and $Y_j(\Omega) = Y_n^m(\Omega)$ is the spherical harmonic of order n and degree m evaluated at Ω, and $j = o(m,n)$ is a single index that depends on the ordering convention defined by the function $o(\cdot)$. In vector notation this can be written as $\boldsymbol{a}(\omega) = s_\Omega(\omega)\boldsymbol{y}(\Omega)$, where $\boldsymbol{a}(\omega) = [a_j(\omega)]_{j=1}^J$, $\boldsymbol{y}(\Omega) = [Y_j(\Omega)]_{j=1}^J$. With the target ear signals $b_l(\omega) = s_\Omega(\omega) H_l(\omega, \Omega)$ for $l \in \{\mathsf{L}, \mathsf{R}\}$, where $H_l(\omega, \Omega)$ is the measured/modeled HRTF, the goal is to find a rendering filter $\boldsymbol{w}(\omega)$ which yields an output signal $\hat{b}_l(\omega) = \boldsymbol{w}_l^{\mathsf{H}}(\omega)\boldsymbol{a}(\omega)$ that is *perceptually* as close as possible to the target signal $b_l(\omega)$. If the soundfield is modeled as a superposition of unknown plane wave signals from unknown directions, the problem can be written as

$$\boldsymbol{w}^*(\omega) = \arg\min_{\boldsymbol{w}} \int_{\Omega \in \mathcal{S}^2} D\left(\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega),\ H(\omega, \Omega)\right) \mathrm{d}\Omega, \quad (1)$$

where $D(\cdot,\cdot)$ is a distance function which models the perceived dissimilarity. In the remainder we omit the dependency on ω for brevity, and we drop the subscript $l \in \{\mathsf{L}, \mathsf{R}\}$ as the HRTF set is assumed to be symmetric about the sagittal plane.

Least-Squares Methods

From a perceptual viewpoint it seems reasonable to define the distance function in (1) in terms of important binaural and monaural cues. However, defining and minimizing a perceptually motivated cost function is not trivial. Therefore, all methods proposed in [1, 2, 3] use some variation of a least-squares (LS) formulation, i.e.

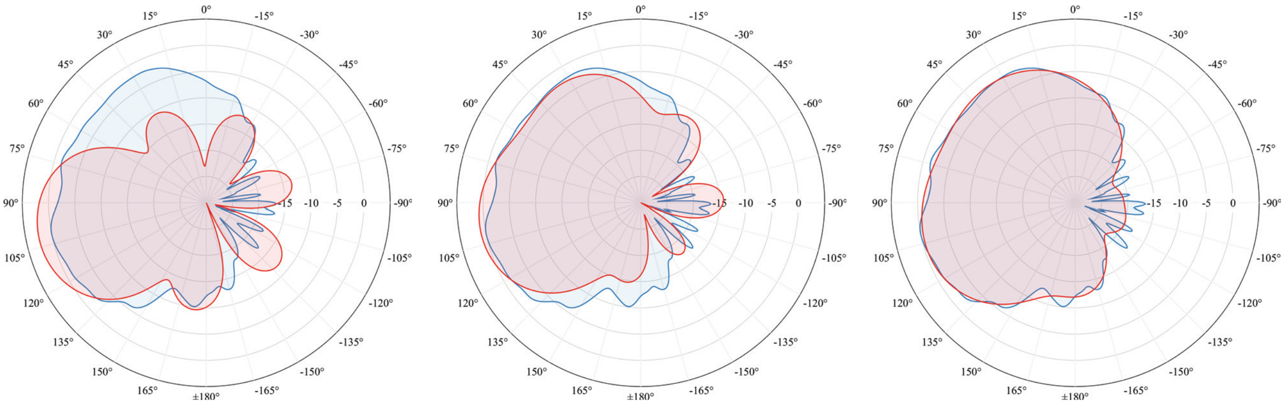
$$\min_{\boldsymbol{w} \in \mathcal{K}} \int_{\Omega \in \mathcal{S}^2} \left|\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega) - H(\Omega)\right|^2 \mathrm{d}\Omega, \quad (2)$$

or its discrete approximation

$$\min_{\boldsymbol{w} \in \mathcal{K}} \sum_{\Omega \in \mathcal{M}} \left|\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega) - H(\Omega)\right|^2 \quad (3)$$

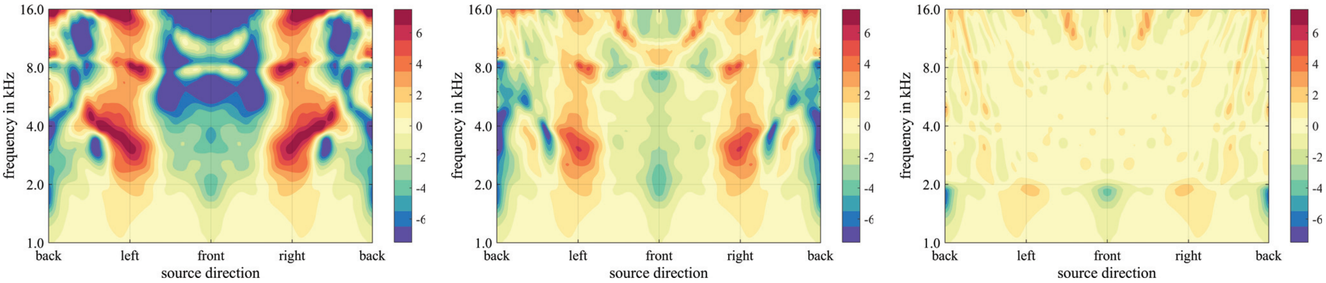
$$\equiv \min_{\boldsymbol{w} \in \mathcal{K}} \left\|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{h}_{\mathcal{M}}\right\|_2^2, \quad (4)$$

[1] For ease of presentation, all numerical results in this contribution are given for the 2-dimensional case, i.e. we use circular harmonics rather than spherical harmonics.



(a) LS solution (equalized) [2] (b) LS solution (with time alignment) [3] (c) MLS solution (proposed)

Figure 1: Desired (blue) and approximated (red) HRTF magnitudes for $f = 6$ kHz and $N = 3$.



(a) LS solution (equalised) [2] (b) LS solution (with time alignment) [3] (c) MLS solution (proposed)

Figure 2: CLL error in dB for $N = 3$.

where $\mathcal{M}$ is a dense set of directions[2] such that $\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{Y}_{\mathcal{M}} = \boldsymbol{I}$, $\boldsymbol{Y}_{\mathcal{M}} = \left[\boldsymbol{y}^{\mathsf{H}}(\Omega)\right]_{\Omega\in\mathcal{M}} \in \mathbb{R}^{|\mathcal{M}|\times J}$, $\boldsymbol{h}_{\mathcal{M}} = \left[H(\Omega)\right]_{\Omega\in\mathcal{M}} \in \mathbb{C}^{|\mathcal{M}|}$, and $\mathcal{K}$ is the domain over which $\boldsymbol{w}$ is optimized. When the LS problem in (4) is unconstrained (i.e. $\mathcal{K} = \mathbb{C}^{J}$), the solution is given by

$$\boldsymbol{w}_{\mathsf{LS}} = \boldsymbol{Y}_{\mathcal{M}}^{\dagger}\boldsymbol{h}_{\mathcal{M}} = \boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{h}_{\mathcal{M}} = \mathcal{SHT}_{N}^{\mathcal{M}}\left(\boldsymbol{h}_{\mathcal{M}}\right), \qquad (5)$$

where $(\cdot)^{\dagger}$ is a pseudo inverse, and $\mathcal{SHT}_{N}^{\mathcal{M}}(\cdot)$ denotes the order-N truncated discrete spherical harmonic transform (SHT). That is, the LS solution is equal to the SHT coefficients up to order N; alas, since the spatial complexity of HRTFs increases with frequency, a significant amount of energy is contained in modes up to order $N = 35$. Hence, the LS solution in (5) yields a severe spectral roll-off towards higher frequencies for low Ambisonic orders. To remedy this timbral artifact, the authors in [2] propose to apply a global diffuse-field equalization filter to the LS solution, which is equivalent to restricting the optimization domain in (4) to $\mathcal{K} = \left\{\boldsymbol{w} \in \mathbb{C}^{J} : \boldsymbol{w}^{\mathsf{H}}\boldsymbol{w} = \int_{\Omega} |H(\Omega)|^{2} = \|\boldsymbol{h}_{\mathcal{M}}\|_{2}^{2}\right\}$ yielding the scaled (equalized) LS solution

$$\boldsymbol{w}_{\mathsf{LSeq}} = \frac{\|\boldsymbol{h}_{\mathcal{M}}\|_{2}}{\left\|\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{h}_{\mathcal{M}}\right\|_{2}}\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{h}_{\mathcal{M}}. \qquad (6)$$

The global equalization term in (6) reduces the overall spectral roll-off, but as the *local* modal order of HRTFs is direction-dependent - with higher orders for frontal directions due to rapid phase changes - direction-dependent colorations remain for lower Ambisonic orders.

The authors in [1] propose a different modification of the LS problem: rather than compensating for the loss of signal energy with a global filter, $\mathcal{M}$ in (4) is chosen to be sufficiently sparse such that higher-order modes are aliased down to lower-order modes. It has been shown, that this method effectively mitigates timbral artifacts, however, finding the optimal sparse set $\mathcal{M}$ (number of sampling points, sampling scheme, rotation) is not straight-forward as it influences both monaural and binaural cues.

Recently, in [3] another variation of the LS problem has been proposed that yields a significant improvement of the perceived quality. The basic notion is to modify the target HRTF set $H(\Omega)$ prior to optimization such that the energy in higher orders is reduced with minimal perceptual ramifications. The proposed HRTF modifications are based on two observations:

- Most of the energy in higher order modes is caused by rapid phase changes towards higher frequencies due to the off center location of the ears.
- With increasing frequency, the perceptual importance of interaural time differences (ITDs) decreases, while the relative importance of interaural level differences (ILDs) increases.

Consequently, a two-band HRTF modification scheme, referred to as time alignment (TA), has been proposed in

[2] We assume that the set $\mathcal{M}$ is chosen such that the integral is trivially approximated, e.g. by using a spherical t-design with high order. In practice, however, quadrature weights need to be used and the accuracy of the approximation depends on the number of sampling points relative to the modal order of the integrand.

[3], where the modified HRTFs are defined as

$$\tilde{H}(\omega,\Omega) = \begin{cases} H(\omega,\Omega) & \text{if } \omega \le \omega_c \\ H(\omega,\Omega)e^{-i\phi_l(\omega,\Omega)} & \text{if } \omega > \omega_c, \end{cases} \tag{7}$$

where $\phi_l(\omega,\Omega)$ is the linear phase due to the ITD corresponding to direction Ω, and ω_c is the cut-on frequency above which phase modification is applied. By subtracting the linear phase part for high frequencies only, the important ITD cues are preserved at low frequencies while signal energy in higher-order modes is significantly reduced. This in turn reduces the energy loss when the modal order is truncated, thus reducing the spectral roll-off towards higher frequencies.

Proposed Method

We can write the cost function in (4) as

$$\min_{\boldsymbol{w}\in\mathcal{K}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{M}\boldsymbol{p}\|_2^2\,, \tag{8}$$

where $\boldsymbol{M} = \operatorname{diag}\left(|\boldsymbol{h}_{\mathcal{M}}|\right)$, $\boldsymbol{p} = \left(e^{i\gamma(\Omega)}\right)_{\Omega\in\mathcal{M}}$, and usually $\gamma(\Omega) = \angle H(\Omega)$. However, it has been shown in [3], that defining $\gamma(\Omega) = \angle H(\Omega) - \phi_l(\Omega)$ at higher frequencies significantly improves the quality of the solution. While removing the linear phase part from the HRTFs to reduce the modal order is conceptually well motivated, it is not clear which HRTFs phases yield the lowest spatial complexity. Therefore we propose to reformulate the LS problem as a joint minimization over $\boldsymbol{w}$ and $\boldsymbol{p}$, i.e.

$$\min_{\boldsymbol{w},\boldsymbol{p}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{M}\boldsymbol{p}\|_2^2 \tag{9}$$

$$\text{s.t. } |p_j| = 1 \;\; \forall j = 1,\dots,|\mathcal{M}|, \tag{10}$$

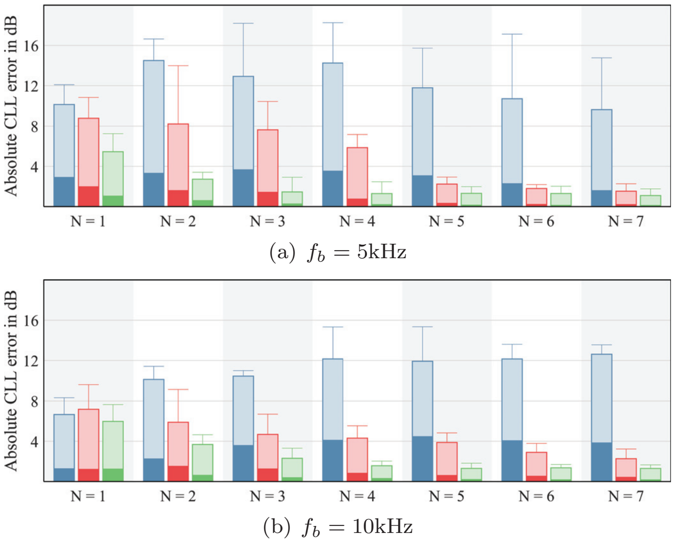
where we simultaneously seek the optimal HRTF phase modification and the corresponding optimal filter coefficients. This quadratically constrained quadratic program (QCQP) might be solved via the popular semidefinite relaxation method [6], however, by observing that this problem is equivalent to the phase-lift formulation of the noisy phase retrieval problem [5], we can rewrite (9)-(10) as

$$\min_{\boldsymbol{w}} \|\, |\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w}| - |\boldsymbol{h}_{\mathcal{M}}|\, \|_2^2\,, \tag{11}$$

where the objective is to approximate only HRTF magnitudes while ignoring the phase error. The proposed rendering filters, referred to as magnitude least squares (MLS) solution, are thus given by

$$\begin{aligned} \boldsymbol{w}_{\mathsf{MLS}}(\omega_k) = \arg\min_{\boldsymbol{w}} \Big[& \lambda(\omega_k) \,\|\, \boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{h}_{\mathcal{M}} \,\|_2^2 \\ & + (1-\lambda(\omega_k)) \,\|\, |\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w}| - |\boldsymbol{h}_{\mathcal{M}}| \,\|_2^2 \Big], \end{aligned} \tag{12}$$

where ω_k is the center frequency of the k-th bin, and $\lambda(\omega) = 1$ if $\omega \le \omega_c$ and 0 otherwise.[3] Finding the global optimizer of (12) for $\lambda(\omega) < 1$ is non-trivial in general, however, we found that any local nonlinear optimization method can be used when the initial estimate for $\boldsymbol{w}_{\mathsf{MLS}}(\omega_k)$ is set to $\boldsymbol{w}_{\mathsf{MLS}}(\omega_{k-1})$. Note that the cost function in (12) is invariant under a global phase change if $\lambda(\omega_k) = 0$, and thus, we apply a phase rotation to the initial solution in order to obtain a smooth phase evolution:



(a) $f_b = 5$kHz

(b) $f_b = 10$kHz

Figure 3: Statistics of the absolute CLL error for all directions and frequencies within octave band around f_b (median, 99th percentile, max). Blue: LS solution (equalized) | Red: LS solution (time aligned) | Green: MLS solution (proposed).

$$\boldsymbol{w}_{\mathsf{MLS}}(\omega_k) \leftarrow \boldsymbol{w}_{\mathsf{MLS}}(\omega_k)e^{i\zeta}, \tag{13}$$

where

$$\zeta = \arg\min_{\tilde{\zeta}} \left\| \boldsymbol{w}_{\mathsf{MLS}}(\omega_k)e^{i\tilde{\zeta}} - \boldsymbol{w}_{\mathsf{MLS}}(\omega_{k-1}) \right\|_2^2 \tag{14}$$

$$= \angle\left(\boldsymbol{w}_{\mathsf{MLS}}(\omega_k)^{\mathsf{H}}\boldsymbol{w}_{\mathsf{MLS}}(\omega_{k-1})\right). \tag{15}$$

Numerical Evaluation

In Fig. 1 the magnitudes of the approximated HRTFs $\hat{H}(\Omega) = \boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega)$ are compared with the magnitudes of the target HRTFs $H(\Omega)$ for $N = 3$ and $\omega_k = 6$ kHz. It can be observed that subtracting the linear phase from the HRTFs significantly improves the approximation (Fig. 1(b)) compared to the LS solution with diffuse-field equalization in Fig. 1(a), and that the proposed MLS approach (Fig. 1(c)) reduces the magnitude error even further. In order to evaluate timbral artifacts for different source directions and frequencies, we use the composite loudness level (CLL), defined as

$$\mathrm{CLL}\left(H(\omega,\Omega)\right) = 10\log\left(|H(\omega,\Omega)|^2 + |H(\omega,\Omega')|^2\right),$$

where $\Omega' = (-\varphi,\theta)$, which is related to the perceived timbre. In Fig. 2 the CLL error

$$e_{\mathrm{CLL}}(\omega,\Omega) = \mathrm{CLL}\left(\hat{H}(\omega,\Omega)\right) - \mathrm{CLL}\left(H(\omega,\Omega)\right)$$

is depicted for different methods with $N = 3$ and a cut-on frequency of 2 kHz. In Fig. 3 the CLL error statistics for different Ambisonic orders are depicted for octave bands around 5 and 10 kHz, respectively. These results show that the proposed method consistently outperforms the two methods recently proposed in [2] and [3].

[3] To avoid rapid filter changes around ω_c we choose a smooth transistion function in practice.

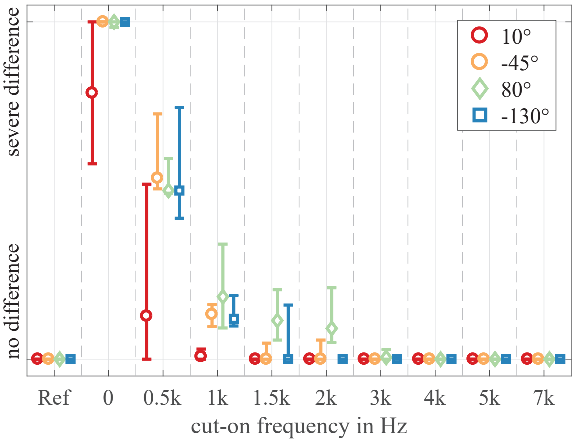


Figure 4: Median an 95% confidence interval of perceived difference ratings for drums as source signal. The cut-on frequency of the modified HRTFs is indicated by the x-ticks.

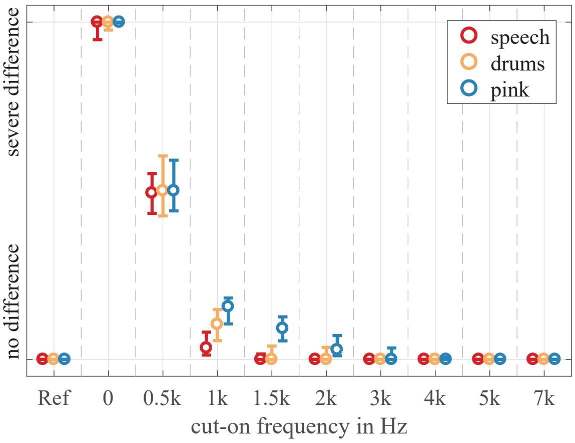


Figure 5: Median an 95% confidence interval of pooled (for all four test directions) ratings per source signal. The cut-on frequency of the modified HRTFs is indicated by the x-ticks.

Optimal Cut-On Frequency

The present method and the method proposed in [3] rely on the assumption, that the relative perceptual importance of IPDs is negligible at high frequencies (compared to interaural level differences and perceived timbre). While this assumption is well motivated, it is not clear how to choose the cut-on frequency ω_c above which IPDs can be disregarded. We therefore conducted a listening experiment using a modified HRTF set as defined in (7). The experiment compared the modified HRTFs with cut-on frequencies $\omega_c = \{0, 0.5, 1, 1.5, 2, 3, 4, 5, 7\}$kHz in a MUSHRA-like procedure against a reference HRTF. Participants were asked to rate the perceived overall difference on a scale from *no audible difference* to *severe difference*. The presented test signals were continuously looped and participants were allowed to seamlessly switch between signals in real-time as often as desired. Overall, three source signals (speech, drum loop, and pulsed pink noise with 150ms hann-windowed ramps) and four source directions $\phi_q = \{10°, -45°, 80°, -130°\}$ were tested in a random sequence. The median and 95% confidence interval of ratings (7 participants, all male, average age 32 years) per source direction for the drum signal are depicted in Fig. 4. While for frontal directions cut-on frequencies above $\omega_c > 1.5$kHz are not significantly different to the reference (Kruskal Wallis test, $p = 0.51$), a minimum cut-on frequency of $\omega_c > 2$kHz is required for lateral directions ($p = 0.84$). Results for speech and pulsed noise showed similar direction-dependent behavior and are therefore not depicted here.

Results of the pooled data per source signal (all four directions) are presented in Fig. 5. The lowest cut-on frequencies which are not significantly different to the reference are $\omega_c = 2$kHz ($p = 0.164$), $\omega_c = 3$kHz ($p = 0.326$), and $\omega_c = 4$kHz ($p = 0.413$) for speech, drums, and pulsed noise, respectively. The increased phase-sensitivity for pulsed noise is explained by onset ITD and envelope ITD evaluation. However, for natural signals (speech, drums) a cut-on frequency as low as $\omega_c = 2$kHz is considered to be sufficient (well inline with duplex theory [7]).

Conclusion

We proposed a method to design binaural rendering filters for Ambisonic signals based on magnitude-only optimization at high frequencies. It has been shown that errors related to the perceived timbre of order-limited Ambisonic signals can be significantly reduced by disregarding interaural phase differences above some cut-on frequency. In a formal listening experiment we found, that this cut-on frequency can be as low as 2 kHz for most signals.

References

[1] B. Bernschütz, A. Vazquez Giner, C. Pörschmann, and J. Arend, “Binaural reproduction of plane waves with reduced modal order,” *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[2] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, “Spectral equalization in binaural signals represented by order-truncated spherical harmonics,” *J. Acoust. Soc. Am.*, vol. 141, no. 6, 2017.

[3] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, “Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint,” *J. Acoust. Soc. Am., submitted*, 2018.

[4] E. A. Macpherson and J. C. Middlebrooks, “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited,” *J. Acoust. Soc. Am.*, vol. 111, no. 5, p. 2219, 2002.

[5] E. J. Candès, T. Strohmer, and V. Voroninski, “PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming,” *Comm. on Pure and App. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[6] Z.-Q. Luo, A. M.-C. So, Y. Ye, and S. Zhang, “Semidefinit relaxation of quadratic optimization problems,” *IEEE Signal Processing Magazine*, no. May, pp. 20–34, 2010.

[7] W. M. Hartmann, B. Rakerd, Z. D. Crawford, and P. X. Zhang, “Transaural experiments and a revised duplex theory for the localization of low-frequency tones,” *J. Acoust. Soc. Am.*, vol. 139, no. 2, p. 968, 2016.