IMPROVING BINAURAL AMBISONICS DECODING BY SPHERICAL HARMONICS DOMAIN TAPERING AND COLORATION COMPENSATION

Christoph Hold[1*], *Hannes Gamper*[2], *Ville Pulkki*[3†], *Nikunj Raghuvanshi*[2], *Ivan J. Tashev*[2]

[1]Technical University of Berlin, Germany
[2]Microsoft Research Redmond, WA, USA
[3]Aalto University, Espoo, Finland

ABSTRACT

A powerful and flexible approach to record or encode a spatial sound scene is through spherical harmonics (SHs), or Ambisonics. An SH-encoded scene can be rendered binaurally by applying SH-encoded head-related transfer functions (HRTFs). Limitations of the recording equipment or computational constraints dictate the spatial reproduction accuracy, thus rendering might suffer from spatial degradation as well as coloration. This paper studies the effect of tapering the SH representation of a binaurally rendered sound field in conjunction with its spectral equalization. The proposed approach is shown to reduce coloration and thus improves perceived audio quality.

1. INTRODUCTION

Spherical harmonics (SH) allow describing any spherical sound scene in a representation that is independent of the reproduction system. Unlike object-based audio encoding methods, the SH or Ambisonics-based representation of a sound field does not require a description of the scene in terms of individual sound sources and their locations and is therefore well suited for encoding and transmitting spatial audio recordings of complex acoustic scenes. A comprehensive overview of strategies for decoding an Ambisonics stream at the receiver is found in [1]. One common way of experiencing Ambisonics audio is binaurally over headphones, either by way of simulating an array of virtual speakers or by decoding directly to binaural output signals via SH-encoded head-related transfer functions (HRTFs) [2, 3, 4].

One major advantage of encoding virtual sound scenes in Ambisonics compared to object-based methods is that the rendering cost does not scale with the number of individual sound sources but instead with the SH encoding order, a parameter that can be chosen freely. This allows to trade off computational cost and bandwidth requirements with the desired spatial resolution, e. g. by determining the required encoding order via an adaptive perceptual measure [5]. Furthermore, there are flexible parametric coding approaches specifically designed for efficient binaural enhancement [6, 7].

Decreasing the SH encoding order essentially limits the available bandwidth in the spatial domain, which may result in spatial aliasing and coloration artifacts that negatively affect audio quality [8].

For time-frequency domain signals, applying tapering windows has been widely established. The concepts constitute a way of addressing sidelobes in spatial filtering [9] and Poletti describes some effects of windowing in the SH domain of audio signals [10]. Tapering in the SH domain finds application in the sound field synthesis community and is used in Ambisonics loudspeaker decoding [11] or when synthesizing focused virtual sound sources [12, 5.6.2].

Ben-Hur et al. showed that decoding SH-encoded audio of limited SH order to binaural signals results in a high-frequency roll-off, mainly due to the order truncation of the head-related transfer functions (HRTFs) [13]. To reduce the resulting coloration of the audio signal, the authors propose to equalize spectral distortions by applying an order-dependent compensation filter to the binaural signals. However, while the spectral equalization seems to reduce overall perceived coloration, spatial aliasing due to the truncated HRTFs causes clearly audible angle-dependent artifacts.

Here we analyze the effects of applying a tapering window directly in the SH domain when decoding Ambisonics audio to binaural signals. We show that tapering successfully reduces angle-dependent coloration and expand the order-dependent compensation filter model proposed by Ben-Hur et al. to include a tapering window function.

2. BINAURAL AMBISONICS DECODING

2.1. Ambisonics Representation

Observing a sound field on the unit sphere, the spherical harmonics transform (SHT) allows a compact representation in the spherical harmonics (SH) domain. A point Ω on the unit sphere is given in azimuth φ and colatitude θ. The SHT is also referred to as a spherical Fourier transform, based on the spherical harmonics [14, 1.4], and is defined for any sound field $s(\varphi, \theta) = s(\Omega)$ as

$$\sigma_{nm} = \int_{\Omega} s(\Omega)[Y_n^m(\Omega)]^* d\Omega \,, \tag{1}$$

with the spherical harmonics $Y_n^m(\varphi, \theta) = Y_n^m(\Omega)$. These form an orthogonal and complete set of spherical basis functions [15] and the SH coefficients σ_{nm} can be interpreted as the angular spectrum / space-frequency spectrum on the sphere.

The inverse spherical harmonics transform is given as the Fourier series

$$s(\Omega) = \sum_{n=0}^{N} \sum_{m=-n}^{+n} \sigma_{nm} Y_n^m(\Omega) \,, \tag{2}$$

where N is referred to as the representation order, yielding to $(N+1)^2$ Ambisonics channels. A perfect reconstruction is achieved for $N = \infty$.

The real spherical harmonics basis functions $Y_{n,m}$ for order n and degree m are given as in [4]:

$$Y_{n,m}(\theta, \varphi) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_{n,|m|}(\cos\theta) y_m(\varphi), \tag{3}$$

*This work was carried out as a research intern at Microsoft Research Labs in Redmond, WA, USA.

†This work was carried out as a consulting researcher at Microsoft Research Labs in Redmond, WA, USA.

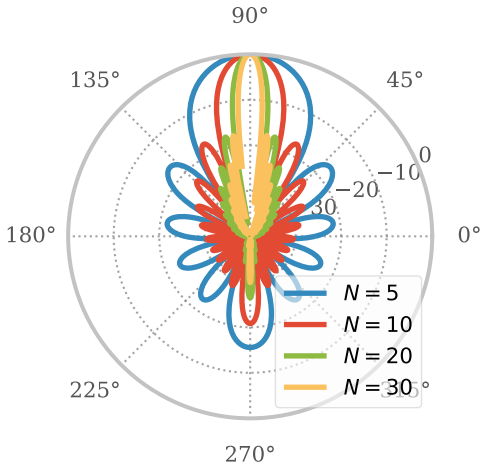

Fig. 1. Cross section ($\theta = 90^\circ$) of a spatial *dirac pulse* magnitude in dB, at $\Omega = [90^\circ, 90^\circ]$, reconstructed from its SH representation by (2) for an increasing SH order N.

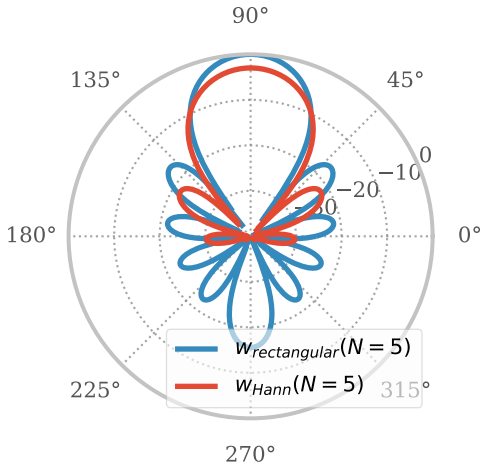

Fig. 2. Cross section ($\theta = 90^\circ$) of a spatial *dirac pulse* magnitude in dB, reconstruction with $N = 5$ at $\Omega = [90^\circ, 90^\circ]$, and with additional Hann tapering function of SH coefficients as in (8).

where $P_{n,|m|}$ is the associated Legendre polynomial and y_m is given as:

$$y_m(\varphi) = \begin{cases} \sqrt{2}\sin(|m|\varphi) & \text{if } m < 0, \\ 1 & \text{if } m = 0, \\ \sqrt{2}\cos(|m|\varphi) & \text{if } m > 0. \end{cases} \tag{4}$$

2.2. Binaural Rendering

To render a point source, the ear input signals s for the left (l) and right (r) ear can be obtained by convolving the source signal x with the head-related impulse response (HRIR) in the desired direction:

$$s^{l,r}(t) = x(t) * h^{l,r}_{\text{HRIR}}(\Omega, t)\,, \tag{5}$$

where $(*)$ denotes the time-domain convolution operation.

In the time-frequency domain, assuming far-field propagation thus plane-wave components $\bar{X}(\Omega)$, the ear input signals are given as

$$S^{l,r}(\omega) = \int_\Omega \bar{X}(\Omega, \omega) H^{l,r}_{nm}(\Omega, \omega) d\Omega\,. \tag{6}$$

Exploiting the orthogonality of the real SH basis functions, this yields [3]

$$S^{l,r}(\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{+n} \breve{X}_{nm}(\omega) \breve{H}^{l,r}_{nm}(\omega)\,. \tag{7}$$

The time domain binaural signals $s^{l,r}(t)$ are obtained from (7) via an inverse time domain Fourier transform.

3. SPHERICAL HARMONICS TAPERING

3.1. Tapering functions

As introduced in Section 2, the spherical harmonics domain constitutes a spherical Fourier domain. Hence, any window function applied in the SH domain introduces spatio-spectral leakage on the sphere. The resulting sidelobes exhibit a periodic pattern. In the case of HRTFs, with two receivers positioned symmetrically on the sphere, these sidelobes may be especially critical as they may lead to unwanted crosstalk between the ears. A common trade-off for selecting a particular window function is between side-lobe suppression and main-lobe widening. We analyze two representative windowing functions in the following.

We extend (2) to include the windowing function w_N as

$$s(\Omega) = \sum_{n=0}^{N} \sum_{m=-n}^{+n} w_N(n) \sigma_{nm} Y^m_n(\Omega)\,, \tag{8}$$

and (7), accordingly:

$$S^{l,r}(\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{+n} w_N(n) \breve{X}_{nm}(\omega) \breve{H}^{l,r}_{nm}(\omega)\,. \tag{9}$$

A hard truncation of the spherical order to N by dropping the higher-order coefficients is equivalent to applying a rectangular window.

To fade out higher-order modes and suppress side-lobes, a tapering function can be applied instead of a rectangular window. The tapering is implemented by multiplying the SH coefficients with a decreasing weight *per order* n, derived from a half-sided window function. As an example, a *Hann* tapering window w_N up to SH order N would be $w_3(n) = [1., 1., 1., 0.5]$, $w_4(n) = [1., 1., 1., 1., 0.5]$, and $w_5(n) = [1., 1., 1., 1., 0.75, 0.25]$, while zero everywhere else.

3.2. Extended coloration compensation filter

Assuming a spherical scatterer object of radius r_0 in a diffuse sound field, the order dependent frequency response on the sphere can be derived analytically [13]. Observing the spherical scatterer pressure response of wavenumber $k = 2\pi f/c$, we expand on this work by introducing a tapering function $w_N(n)$ weighting each mode n to

$$\bar{p}_w(kr_0)|_N = \frac{1}{4\pi} \sqrt{\sum_{n=0}^{N} w_N(n)(2n+1)|b_n(kr_0)|^2}\,. \tag{10}$$

The mode strength on the rigid sphere is given as [14, 2.62]

$$b_n(kr_0) = 4\pi i^n \left[j_n(kr_0) - \frac{j'_n(kr_0)}{h'_n(kr_0))} h_n(kr_0) \right], \tag{11}$$

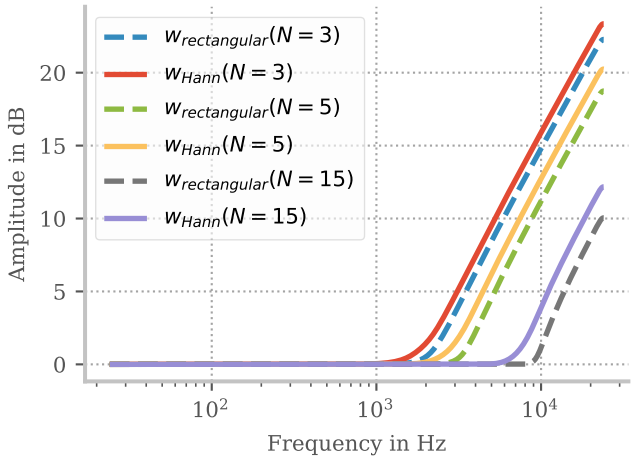

Fig. 3. Coloration compensation filter for SH order truncation, for multiple truncated SH orders, with Hann and rectangular windowing functions applied.

where j_n is the spherical Bessel function, h_n the spherical Hankel function of the second kind, and $(')$ is the derivative with respect to the argument.

Comparing the scatterer response above the spatial aliasing limit $N_{\text{full}} > kr_0$ to the desired truncated order N leads to the extended spectral equalization filter magnitude response

$$G(kr)|_N = \frac{\bar{p}(kr_0)|_{N_{\text{full}}}}{\bar{p}_w(kr_0)|_N}. \tag{12}$$

This spectral filter equalizes signals of order N to the frequency response of a signal at order $N_{\text{full}} > kr_0$ [13]. Extending the design to account for a window function compensates for the spectral effects of applying SH tapering, which in turn suppresses sidelobes. A reference implementation and sound samples are available online[1].

4. EXPERIMENTAL EVALUATION

4.1. Head-related Transfer Functions

Experiments are carried out using a set of spherical anechoic far-field measurements of a Neumann KU100 dummy-head, available publicly [16]. The set comprises measurements on an equidistant spherical Lebedev grid with 2354 nodes, which should enable a stable transform into the SH domain with low spatial aliasing over the entire audio frequency range. The spherical harmonics transform of the HRTFs is carried out by a least mean square fit with Tikhonov regularization directly to the target order. When deriving the compensation FIR filter we used the time sampling frequency $f_s = 48\,\text{kHz}$, leading to $N_{full} = 39$ and a scatterer radius of $r_0 = 0.0875\,\text{m}$.

4.2. Coloration Model

To model the coloration error (CE) between the reference HRIRs (time-domain) and the reconstructed HRIRs (after order-truncation in the SH-domain), a model proposed by Brinkmann and Weinzierl was used [17]:

$$CE = w_l \Delta L_l + w_r \Delta L_r\,, \tag{13}$$

[1] https://github.com/chris-hld/spaudiopy

where w_l and w_r are binaural weighting factors. The domain level differences $\Delta L_{l/r}$ per auditory filter band from 50 Hz to 20 kHz for each ear are calculated by May's localization model implementation [18], which includes rectification, compression, and an auditory filter bank. The binaural weights are given as [17]

$$w_l = \frac{2^{\Delta L_{lr}/10}}{1 + 2^{\Delta L_{lr}/10}}\,, \qquad w_r = 1 - w_l\,. \tag{14}$$

The weights account for the fact that coloration errors are perceptually more relevant for the ear receiving a louder signal [17].

5. RESULTS

An ideal representation of a point source on the sphere is a spatial *dirac pulse*, which exhibits infinite spatial bandwidth. Truncating the Fourier series (2) to an order $N < \infty$ causes a non-ideal reconstruction, as shown in Fig. 1, resulting e. g. in spatial blur and coloration.

In the case of a simple truncation to $N = 5$, which is equivalent to applying a rectangular window, Fig. 2 shows the most prominent sidelobe is the backlobe, suppressed only by about 15 dB. Introducing a half-sided *Hann* tapering function, as described in Section 3, improves the backlobe suppression drastically to more than 40 dB. However, the sidelobe suppression comes at the expense of a slightly quieter and widened mainlobe.

When applying tapering coefficients to auralizations, it is important to compensate for the spectral distortion introduced by any window, as described in Section 3.2. Figure 3 shows the frequency response of that filter, which equalizes the diffuse field response of an order truncated soundfield. Compared to the simple rectangular truncation, the *Hann* function requires only marginally more high frequency boosting.

The error between the reference time-domain HRTF and its third order SH representation is visualized in Fig. 4 and detailed in Table 1, with negligible error below 2.5 kHz. As can be seen, applying the compensation filter even with a non-tapered window reduces the overall coloration error (CE) in terms of the root-mean-squared error (RMSE), which averages over frequency and angle. However, for both untapered SH representations, the reconstruction error reveals a strong angle dependence, with excess energy especially at the contralateral side (cf. Fig. 4, (left) and (center)). This manifests itself in a large maximum CE (cf. Table 1, $\max(\text{CE}(\Omega))$ and $\max(\text{CE}(\Omega, f))$). The proposed tapering seems to reduce the error maxime and improve the contralateral ear signals (cf. Fig. 4, (right)).

The CE for a point source moving in the horizontal plane is shown in Fig. 5 for a truncation to third order with a rectangular window without any spectral equalization, a rectangular window with its spectral compensation, and a Hann tapering window with its spectral compensation. The spectral compensation of the rectangular window reduces coloration in the front and in the back at the expense of stronger coloration to the sides. Applying an additional tapering and the corresponding compensation filter, the variance of estimated coloration is lower and it is distributed more evenly across directions. This can also be seen in Table 1 in a reduction of the $\max(\text{CE}(\Omega))$ and $\max(\text{CE}(\Omega, f))$ coloration estimate. This indicates that, anywhere on the sphere, the maximum coloration introduced by the order truncation of the HRTFs is greatly reduced by applying a tapering window together with the proposed extended coloration compensation filter. Informal listening tests confirmed a clearly audible effect of the tapering window on the binaural decoding of Ambisonics signals at third and fifth order, and indicate improvements for even higher orders.

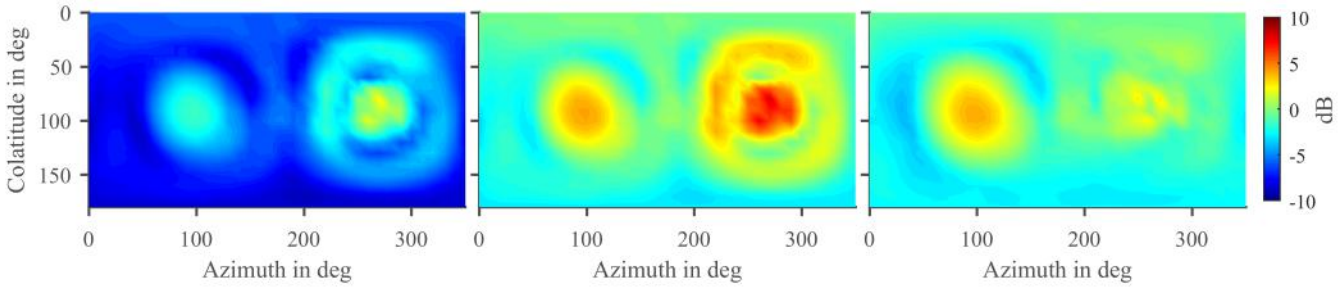

Fig. 4. Error between time domain HRTF and spherical harmonics reconstruction of third order, averaged across 39 auditory filter bands in dB; (left) truncation using a rectangular window without compensation, (center) truncation using a rectangular window with coloration compensation, (right) Hann tapering window with proposed extended coloration compensation.

	Full Band			Above 2.5 kHz		
	RMSE (dB)	$\max(\mathrm{CE}(\Omega))$ (dB)	$\max(\mathrm{CE}(\Omega, f))$ (dB)	RMSE (dB)	$\max(\mathrm{CE}(\Omega))$ (dB)	$\max(\mathrm{CE}(\Omega, f))$ (dB)
no tapering, no compensation	2.0234	4.0425	20.8375	6.3004	13.1143	20.8375
no tapering, with compensation	1.7614	4.8412	22.6504	3.8908	14.9174	22.6504
Hann tapering, with compensation	**1.7199**	**3.1641**	**13.4945**	**3.3664**	**8.7494**	**13.4945**

Table 1. Coloration errors (CE) estimated from a 20 ms white noise burst convolved with third-order reconstructed HRIRs, for 1024 directions distributed uniformly on the sphere. RMSE shows root-mean-squared error over frequency and angle, $\max(\mathrm{CE}(\Omega))$ the maximum frequency-averaged CE, and $\max(\mathrm{CE}(\Omega, f))$ the maximum CE at any filter band frequency and angle.

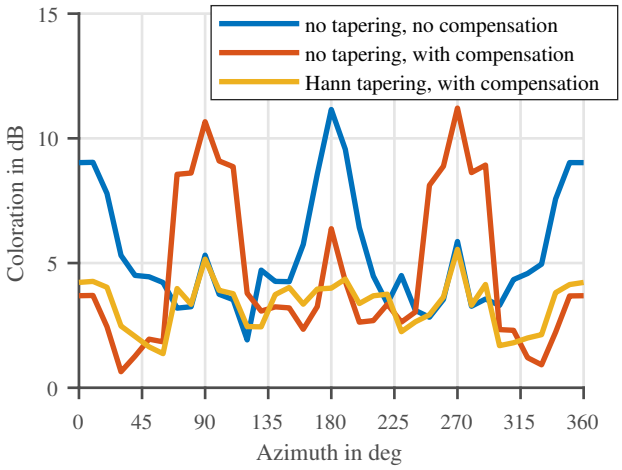

Fig. 5. Coloration estimation of third order HRIR reconstruction above 2.5 kHz for impulse responses moving around the listener in ten degree steps in the horizontal plane ($\theta = 90°$).

6. DISCUSSION

As noted in prior work, order truncation of HRTFs results in high-frequency roll-off that an order-dependent compensation filter can equalize [13]. As expected, a higher SH representation order generally results in less residual error, nonetheless, sidelobes and spectral distortion are still noticeable at high orders. However, due to the angle-dependence of the coloration error, the inherently static compensation filter may boost erroneous aliased components and tends to sound excessively bright for lateral sources, as shown in Fig. 4.

By applying a tapering function in the SH domain, e. g. the proposed half sided *Hann*-function, the coloration error could be reduced through a suppression of the sidelobes on the sphere and thus enhancing the directional pattern. In the case of binaural rendering, suppressing the backlobe appears to be particularly critical to reducing crosstalk between the ear signals due to the symmetric arrangement of the ears. Here, tapering seems to mitigate various perceptual artifacts, most notably it helps restoring interaural level differences (ILDs) degraded by the crosstalk.

We proposed an extension of the spectral roll-off equalization of order-limited signals proposed in [13] to account for a tapering function, thus combining the roll-off compensation with the sidelobe suppression of a tapering window.

The coloration compensation can also be applied when encoding the HRTFs, i.e., no additional processing is required at run time when decoding Ambisonics audio to binaural signals.

Future work includes comparing various window design methods, e. g. the max-$\mathbf{r}_E$ weighting and a formal listening test.

7. SUMMARY AND CONCLUSION

This work investigated the effect of truncating the spherical harmonics order, in particular when applied to head-related transfer functions (HRTFs). It was observed that the truncation causes both spatial degradation and also coloration. While applying an order-dependent compensation filter reduces high-frequency roll-off due to order truncation, it does not compensate for angle-dependent artifacts. Tapering in the spherical harmonics domain in combination with an extended order-dependent coloration compensation filter was shown to improve binaural Ambisonics rendering significantly without increasing computational complexity at run time. Both informal listening and a binaural model confirmed the perceptual quality.

8. REFERENCES

[1] Matthias Frank, Franz Zotter, and Alois Sontacchi, "Producing 3D Audio in Ambisonics," in *AES 57th International Conference*, 2015, pp. 1–8.

[2] Ramani Duraiswami, Dmitry N. Zotkin, Zhiyun Li, Elena Grassi, Nail A. Gumerov, and Larry S. Davis, "High Order Spatial Audio Capture and its Binaural Head-Tracked Playback over Headphones with HRTF Cues," in *119th AES Convention*, 2005.

[3] Boaz Rafaely and Amir Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 823–828, 2010.

[4] Archontis Politis, *Microphone array processing for parametric spatial audio techniques*, PhD Thesis, Aalto University, 2016.

[5] Carl Schissler, Peter Stirling, and Ravish Mehra, "Efficient Construction of the Spatial Room Impulse Response," *Proceedings of the IEEE Virtual Reality (VR)*, pp. 122–130, 2017.

[6] Archontis Politis, Leo McCormack, and Ville Pulkki, "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, vol. 2017-Octob, pp. 379–383.

[7] Archontis Politis, Sakari Tervo, and Ville Pulkki, "COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 6802–6806.

[8] Amir Avni, Jens Ahrens, Matthias Geier, Sascha Spors, Hagen Wierstorf, and Boaz Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.

[9] Barry D. Van Veen and Kevin M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[10] Mark A. Poletti, "A unified theory of horizontal holographic sound systems," *Journal of the Audio Engineering Society*, vol. 48, no. 12, pp. 1155–1182, 2000.

[11] Franz Zotter and Matthias Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, pp. 807–820, Oct. 2012.

[12] Jens Ahrens, *Analytic Methods of Sound Field Synthesis*, Springer, 2012.

[13] Zamir Ben-Hur, Fabian Brinkmann, Jonathan Sheaffer, Stefan Weinzierl, and Boaz Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[14] Boaz Rafaely, *Fundamentals of Spherical Array Processing*, Springer, 2015.

[15] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics*, Chelsea Pub. Co., 1955.

[16] Benjamin Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU100," in *AIA/DAGA Meran (Italy)*, 2013.

[17] Fabian Brinkmann and Stefan Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *AES Conference on Audio for Virtual and Augmented Reality (AVAR)*, 2018.

[18] Tobias May, Steven Van De Par, and Armin Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.