forum acusticum 2026
8-12 September · Graz, Austria

PERCEPTUAL EVALUATION OF ACOUSTIC-CENTER ALIGNMENT IN SPHERICAL HARMONIC ENCODING OF LOUDSPEAKER DIRECTIVITY FOR BINAURAL REPRODUCTION

Yasen Velchev[1], Lukas Gölles[1], Florian Wendt[2], Manuel Brandner[2], Matthias Frank[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria*
[2] *LEWITT Development GmbH, Vienna, Austria*

Abstract

In virtual and augmented reality (VR/AR), spatial audio rendering often relies on limited-order spherical-harmonic (SH) representations. While SH truncation in head-related transfer functions (HRTFs) is known to introduce high-frequency phase errors that degrade perceptual quality, analogous issues for loudspeaker directivity have received less attention. This work investigates preprocessing strategies for transforming measured loudspeaker directivity into SH form in the presence of direction-dependent time-of-arrival (TOA) inconsistencies, particularly pronounced in multi-way loudspeakers with frequency-dependent acoustic centers. Such inconsistencies, when encoded at finite SH order, can cause spectral artifacts. We compare simple magnitude-oriented approaches (peak alignment, minimum-phase) with methods that aim to preserve physically meaningful phase via broadband and subband acoustic-center compensation, as well as per-driver encoding with crossover reconstruction. A MUSHRA-like listening test assessed perceptual similarity to a reference based on the original measurements for virtual stereo reproduction in anechoic and semi-diffuse conditions. Results show that the simpler preprocessing approaches generally yield higher perceptual similarity than acoustic-center-preserving approaches, while per-driver encoding performs comparably to peak alignment at higher orders. An ERB-band spectral error correlates with the perceptual trends, underscoring the primary importance of magnitude fidelity under SH truncation.

Keywords: *acoustic center, spherical harmonics, virtualization*

Corresponding author*: velchevyasen@gmail.com.*

1. Introduction

Spherical harmonic (SH) representations provide a compact and practical way of describing directional sound fields. In the context of loudspeaker directivity measurements, they allow dense measurement grids to be represented with a limited set of coefficients and facilitate efficient rendering in virtual and augmented reality (VR/AR) applications, where they can be readily converted into Ambisonic signals for spatial audio reproduction [1, 2].

Previous work has extensively investigated SH representations of HRTFs, where finite-order truncation can introduce reconstruction errors and special care must be taken to preserve perceptually relevant localization cues [3–5]. While related challenges can be expected for loudspeaker directivity measurements, this topic has received comparatively little attention. Loudspeaker directivity measurements consist of directional impulse responses that must be represented by a finite-order SH expansion. To reduce truncation artifacts, preprocessing strategies such as peak alignment and minimum-phase transformation simplify the directional impulse responses prior to SH encoding. Peak alignment shifts the measured impulse responses such that their dominant peaks occur at a common sample, whereas minimum-phase transformation removes the excess phase while preserving the magnitude response. Although these methods improve low-order SH representations [3], they do not preserve the original direction-dependent TOA relationships.

Acoustic-center estimation has been proposed as an alternative means of describing loudspeaker radiation [6, 7]. The acoustic center is defined as the “position of the virtual point source from which sound pressure varies inversely as distance” [8]. Building on this concept, this work investigates broadband and frequency-dependent acoustic-center translation for SH encoding, alongside peak alignment, minimum-phase processing, and per-driver encoding. Their perceptual performance is evaluated against the original measured loudspeaker directivities using a MUSHRA-like listening experiment based on virtualized stereo reproduction.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology, including the SH encoding framework, preprocessing strategies, perceptual evaluation, and ERB spectral error metric. Section 3 presents the results, and Section 4 concludes with a discussion and future work.

2. Methods

This section describes the measurement procedure, SH encoding framework, and preprocessing strategies for direction-dependent time-of-arrival (TOA) variations prior to SH encoding. The evaluated methods include peak alignment, minimum-phase representation, and acoustic-center-preserving approaches. The perceptual evaluation and ERB spectral error metric used for assessment are presented subsequently.

2.1. Loudspeaker Measurements

All measurements were performed in the anechoic chamber of the Institute of Electronic Music and Acoustics (IEM). The impulse-response datasets were obtained from three loudspeakers representative of different topologies and use cases: a Genelec 8341A coaxial studio monitor, a Genelec 8020A two-way studio monitor, and a Yamaha DXR8 two-way public-address loudspeaker.

For each loudspeaker, a spherical measurement grid was acquired using a vertical arc with a radius of 80 cm. 18 microphones were positioned at zenith angles between 5° and 175° in 10° steps. The loudspeaker was mounted on a turntable and rotated in 10° azimuth steps to obtain full horizontal coverage. For the multi-way loudspeakers, the rotation center was chosen midway between the individual drivers, since the effective acoustic center is generally unknown and may vary with frequency. Impulse responses were obtained from 2 s exponential sine sweep measurements recorded at 48 kHz. After deconvolution, the impulse responses were truncated to 1024 samples. The resulting dataset comprises 648 directional impulse responses (18×36 directions), sufficient for approximately uniform sampling up to SH order $N = 15$, and forms the basis for all subsequent SH encoding approaches.

2.2. SH-encoding

Following the measurement procedure, the measured impulse-response grid $h(n, \Omega_q)$ is transformed into a spherical harmonic (SH) representation, where n denotes the discrete time index and $\Omega_q = (\phi_q, \vartheta_q)$ the q-th measurement direction. The SH coefficients are obtained by projecting the measured responses onto the spherical harmonic basis using a least-squares solution:

$$\mathbf{h}_{\mathrm{SH}}(n) = \mathbf{h}(n, \Omega)(\mathbf{Y}^{\dagger})^{T}, \tag{1}$$

where $\mathbf{Y}$ contains the spherical harmonic basis functions evaluated at the measurement directions and $\mathbf{Y}^{\dagger}$ denotes the Moore–Penrose pseudoinverse of it [9]. The measured impulse responses are thereby represented by SH coefficients up to the chosen order N. A direct SH encoding of measured impulse responses is sensitive to direction-dependent TOA variations. These variations increase the spatial complexity of the sound field and can introduce reconstruction artifacts at finite SH orders. Therefore, several preprocessing strategies were investigated prior to SH encoding.

2.3. Time-alignment and Acoustic Center Compensation

Five preprocessing strategies were evaluated. Method A serves as a baseline that removes directional timing information prior to encoding by time alignment. This baseline is perceptually evaluated against other more elaborated methods for specific loudspeakers. While Method A0 can be seen as an alternative baseline based on minimum-phase processing, Methods B-D attempt to preserve the acoustic center of the loudspeaker by compensating for directional TOA differences before SH encoding and restoring them during decoding.

Method A: Peak Alignment

For the baseline Method A the measured impulse responses are shifted such that the absolute maximum of each response occurs at a common sample index. The aligned responses are subsequently encoded into the SH domain without restoring the removed delays during decoding. This preprocessing improves the suitability of the measured data for finite-order SH encoding by reducing truncation artifacts caused by inconsistent arrival times across measurement directions. However, the alignment removes the original directional TOA information and therefore does not preserve the physical propagation delays of the measured loudspeaker.

Method A0: Minimum-Phase Representation

For the alternative baseline Method A0, evaluated using the Genelec 8020A, each measured impulse response is replaced by its minimum-phase equivalent prior to SH encoding. The minimum-phase response is obtained in the cepstral domain by removing the excess-phase component through cepstral liftering and reconstructing the impulse response from the remaining minimum-phase component. This preserves the original magnitude spectrum while replacing the directional phase response, and therefore the original direction-dependent TOA relationships are not retained.

Method B: Broadband Acoustic Center Translation

Method B assumes a single broadband acoustic-center displacement vector $\mathbf{d}$ relative to the measurement origin. This method is applied to the coaxial loudspeaker Genelec 8341A. The displacement vector is estimated by fitting the measured peak-arrival times across all measurement directions to the geometric model

$$\tau_i \approx \frac{\mathbf{n}_i \cdot \mathbf{d}}{c} + b, \tag{2}$$

where τ_i denotes the measured peak-arrival time for direction i, $\mathbf{n}_i$ is the corresponding unit direction vector, c is the speed of sound, and b is a constant offset accounting for the common propagation delay. The model parameters are obtained using a weighted least-squares fit, where the weights are proportional to the peak amplitudes of the measured impulse responses. The estimated displacement vector is then used to predict the direction-dependent TOA offset

$$\tau_q = \frac{\mathbf{n}_q \cdot \mathbf{d}}{c}, \tag{3}$$

for an arbitrary query direction $\mathbf{n}_q$. Prior to SH encoding, the measured impulse responses are translated by applying the corresponding phase shift $e^{\mathrm{j}\omega\tau_q}$, thereby removing the estimated directional delay field. After SH decoding, the inverse phase shift is applied to restore the estimated directional TOA relationships while preserving the phase implied by the geometric model.

Method C: Subband Acoustic Center Translation

Method C extends the broadband model by allowing the acoustic center to vary with frequency. This method is applied to the two-way Genelec 8020A. Separate acoustic-center displacement vectors, $\mathbf{d}_{\mathrm{LF}}$ and $\mathbf{d}_{\mathrm{HF}}$, are estimated from low-pass- and high-pass-filtered impulse responses, representing the loudspeaker below and above the crossover frequency, respectively. In this work, the transition is centered at $f_{\mathrm{cross}} = 3\,\mathrm{kHz}$ according to the manufacturer's specifications. The effective acoustic-center displacement is obtained by smoothly interpolating between $\mathbf{d}_{\mathrm{LF}}$ and $\mathbf{d}_{\mathrm{HF}}$ as a function of frequency. The resulting frequency-dependent translation is applied before SH encoding and reversed during decoding.

Method D: Per-Driver Encoding

Method D is applied to the multi-way loudspeaker Yamaha DXR8, for which separate measurements, centered at the low- and high-frequency drivers, are available. Each driver is measured independently with the corresponding driver positioned at the measurement origin and encoded into its own SH representation. During decoding, the two SH signals are filtered using a fourth-order Linkwitz–Riley crossover at 3 kHz before being summed to emulate the loudspeaker's active crossover network. Since each driver is encoded at its own rotation center, no phase-ramp correction is required.

2.4. Perceptual evaluation

To evaluate the perceptual relevance of the investigated SH-encoding strategies, a listening experiment was conducted. The study examined whether preserving the estimated acoustic-center-related TOA information present in the measurements yields perceptual benefits over simpler preprocessing strategies when SH encoding is performed at finite order.

Virtualized loudspeaker reproduction was implemented using the MCRoomSim framework [10], which renders loudspeaker directivities via nearest-neighbour interpolation. For each encoding method under test, the measured directivity data of a specific loudspeaker were encoded into the SH domain and decoded back onto the original measurement grid before being supplied to the simulator, ensuring that the unprocessed measurements serve as a well-defined reference condition. Two acoustic environments were evaluated: anechoic and semi-diffuse ($T_{60} = 0.25\,\mathrm{s}$).

To capture the perceptual influence of inter-loudspeaker phase relationships, stimuli were rendered as a virtualized stereo pair. Binaural room impulse responses (BRIRs) were synthesized for a simulated Neumann KU100 artificial head at the listener position and convolved with continuous pink-noise excitation to generate the experimental stimuli. Since the BRIRs were pre-rendered and not updated with listener movement, the evaluation assessed only the on-axis spectral coloration introduced by the investigated preprocessing strategies.

The experiment followed a MUSHRA-like paradigm. For each trial, listeners compared several SH-encoded reproductions against a reference generated from the original measured IR set of the corresponding loudspeaker. Participants rated the similarity of each stimulus relative to the reference on a 0–100 scale (5-point increments), ranging from "very different" to "very similar", using IEM's MUSHRA app[1], which also recorded the responses.

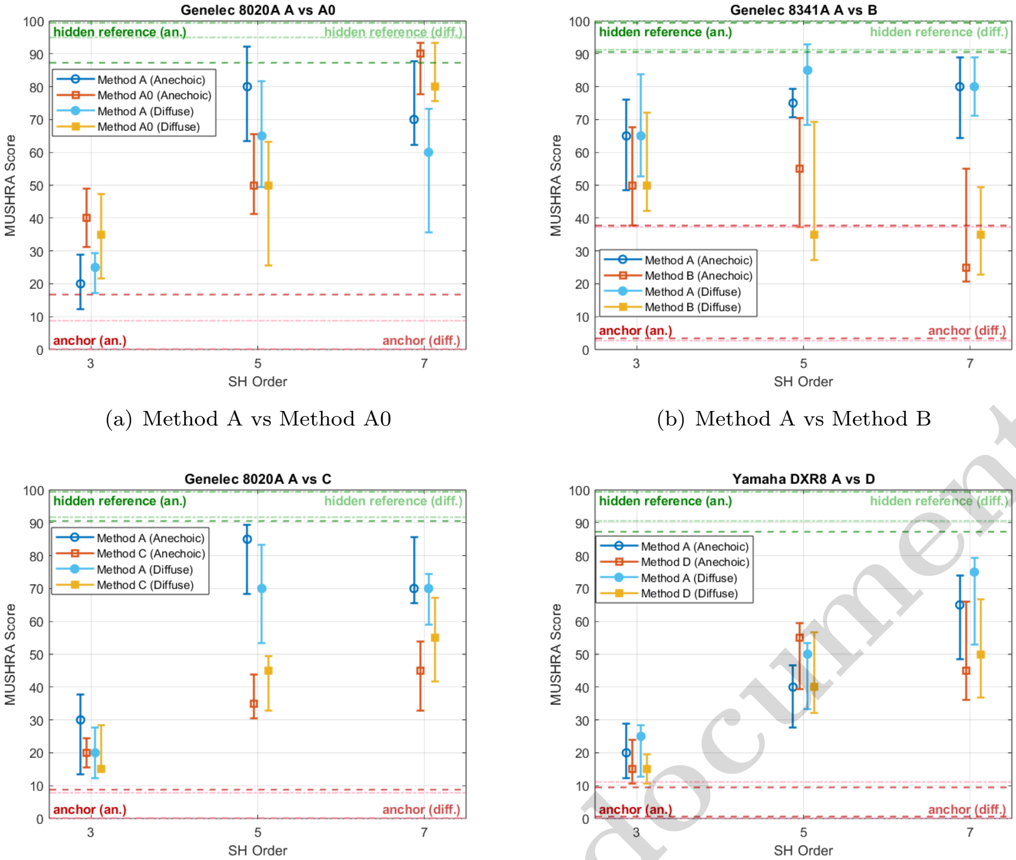
Four comparisons were evaluated in both acoustic environments: Method A0 (Genelec 8020A), Method B (Genelec 8341A), Method C (Genelec 8020A), and Method D (Yamaha DXR8), each against the baseline Method A. SH orders 3, 5, and 7 were tested, yielding six conditions per comparison together with a hidden reference and a first-order anchor, rendered with Method A. A total of 15 participants completed the experiment. Results are reported as medians with 95% confidence intervals.

2.5. ERB Spectral Error

To complement the perceptual evaluation, an ERB-band spectral error was computed for each SH-encoding strategy. The left-ear BRIR from the anechoic simulation was transformed into the frequency domain and decomposed into equivalent rectangular bandwidth (ERB) bands. The average power level of each band was computed in dB, and the ERB spectral error was calculated as

$$E_{\mathrm{ERB}} = \sqrt{\frac{1}{N_{\mathrm{ERB}}} \sum_{i=1}^{N_{\mathrm{ERB}}} \left(L_i^{\mathrm{ERB}} - L_{i,\mathrm{ref}}^{\mathrm{ERB}} \right)^2}, \tag{4}$$

[1] MUSHRA app, https://git.iem.at/rudrich/mushra



(a) Method A vs Method A0

(b) Method A vs Method B

(c) Method A vs Method C

(d) Method A vs Method D

Figure 1: Perceptual evaluation results. Median similarity ratings and 95% confidence intervals for SH orders 3, 5, and 7 in anechoic (dark colors) and diffuse (light colors) conditions. Comparisons: (a) Method A vs. Method A0 (Genelec 8020A), (b) Method A vs. Method B (Genelec 8341A), (c) Method A vs. Method C (Genelec 8020A), and (d) Method A vs. Method D (Yamaha DXR8). Dashed lines indicate the 95% confidence intervals of the hidden references and first-order anchors.

where L_i^{ERB} and $L_{i,\mathrm{ref}}^{\mathrm{ERB}}$ are the ERB-band power levels (dB) of the processed and reference BRIR, respectively, and N_{ERB} is the number of ERB bands.

3. Results

3.1. Results from Perceptual Evaluation

Similarity ratings obtained from the listening experiment are summarized in Figure 1 as medians with corresponding 95% confidence intervals for all evaluated conditions. Statistical significance was assessed using paired Wilcoxon signed-rank tests with Holm–Bonferroni correction for multiple comparisons.

The comparison between Peak Alignment (Method A) and the Minimum-Phase approach (Method A0) is shown in Figure 1a. Here, the highest ratings were obtained by Method A0 at 7th order in both room conditions. However, the advantage over the best-performing Peak Alignment condition (at SH order 5) was not statistically significant for anechoic conditions ($p = 0.24$). At 5th order, Method A significantly outperformed A0 ($p = 0.001$), whereas the opposite trend was observed at 7th order ($p = 0.0068$). These results indicate that both baseline methods achieve comparable perceptual performance at sufficiently high SH orders.

Figure 1b compares Method A and Broadband Acoustic Center Translation (Method B) using the Genelec 8341A. Across all SH orders and both room conditions, Method A consistently received higher similarity ratings than Method B. For SH orders 5 and 7, these differences were statistically significant in both anechoic and diffuse conditions ($p \leq 0.01$). While Method A showed a tendency towards improved ratings with increasing SH order, Method B exhibited no comparable improvement and, in some cases, received lower ratings at higher orders.

The comparison between Method A and Subband Acoustic Center Translation (Method C) is shown in Figure 1c. Similar to Figure 1b, Method A substantially outperformed Method C at SH orders 5 and 7 in both acoustic environments. For the anechoic condition, the differences between the two methods were significant at both orders ($p \leq 0.002$). Method C

Table 1: ERB spectral error (dB) for the investigated SH-encoding strategies at different SH orders. Lower values indicate greater spectral similarity to the reference.

LSP	Met.	SH1	SH3	SH5	SH7
Gen. 8020A	A	6.78	2.27	0.89	0.64
	A0	-	1.37	1.03	0.73
Gen. 8341A	A	1.61	0.88	0.66	0.58
	B	-	0.95	0.80	0.89
Gen. 8020A	A	6.78	2.27	0.89	0.64
	C	-	2.25	1.47	1.31
Yam. DXR8	A	6.63	4.68	2.31	1.27
	D	-	3.56	1.32	1.48

showed a gradual improvement with increasing SH order, suggesting that the performance of the frequency-dependent acoustic-center model benefits from higher spatial resolution. In the diffuse condition, the performance gap between both methods decreased at 7th order, where no statistically significant difference was observed ($p = 0.06$). Figure 1d presents the comparison between Method A and Per-Driver Encoding (Method D) using the Yamaha DXR8. In contrast to the previous comparisons, both methods achieved similar ratings at higher SH orders in both room conditions. In both diffuse and anechoic conditions the highest median score was obtained by Method A at 7th order, though no significant differences were found between the highest-rated conditions, indicating that both approaches provide comparable perceptual performance when sufficient SH order is available. For both methods, 3rd-order representations received substantially lower ratings than 5th- and 7th-order representations. Among all evaluated conditions, seventh-order Minimum-Phase (Method A0) in the anechoic condition was the only tested condition that was not rated significantly different from the hidden reference. 3rd-order representations generally received the lowest ratings regardless of encoding strategy. Increasing the SH order consistently improved perceptual similarity, although the magnitude of this improvement depended on the employed preprocessing method and the measured loudspeaker.

3.2. ERB Spectral Error Analysis

Complementary to the perceptual evaluation, the ERB spectral errors obtained for each preprocessing strategy are summarized in Table 1. For most investigated methods, the ERB spectral error decreases with increasing SH order, indicating improved spectral reconstruction at higher spatial resolutions. Method A consistently achieves lower spectral errors than both Method B and Method C, while the Method A0 reaches errors comparable to Method A at higher SH orders. Likewise, Method A and Method D exhibit similar ERB spectral errors for the Yamaha DXR8. For the Genelec 8341A, the ERB spectral errors of Method A and Method B remain relatively

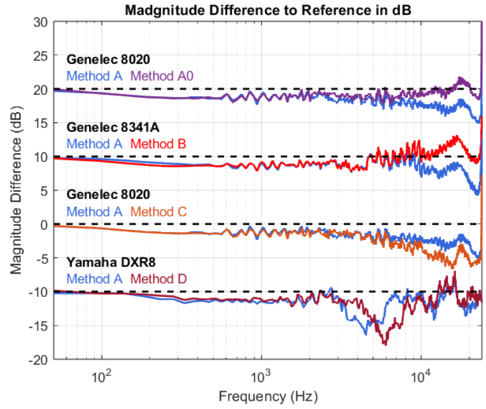


Figure 2: Magnitude difference between the on-axis magnitude responses after seventh-order SH encoding and decoding using each preprocessing method and the native reference response (dashed line).

close across all investigated SH orders, despite the perceptual differences observed in the listening experiment. Table 1 also includes the ERB spectral errors of the first-order SH anchors used during the listening evaluation. While the anchors for the Genelec 8020A and Yamaha DXR8 exhibit errors of approximately 6 dB, the corresponding anchor for the Genelec 8341A shows a substantially lower error of 1.6 dB.

To complement the global ERB spectral error metric, Figure 2 presents the spectral magnitude difference between the native reference and the seventh-order SH reconstruction for each investigated preprocessing strategy. Whereas the ERB spectral error summarizes the average deviation across ERB bands, the magnitude-difference plots reveal how these deviations are distributed across frequency. In particular, Method B exhibits localized deviations at high frequencies for the Genelec 8341A, while both Method A and Method D show pronounced deviations around the crossover region of approximately 3 kHz for the Yamaha DXR8.

4. Discussion and Outlook

The perceptual evaluation showed that Peak Alignment (Method A) generally achieved higher similarity ratings than Broadband Acoustic Center Translation (Method B) and Subband Acoustic Center Translation (Method C). In contrast, Per-Driver Encoding (Method D) yielded perceptual performance comparable to Peak Alignment for the Yamaha DXR8, while additionally preserving the measured phase relationships between the individually encoded drivers. These findings suggest that, within the investigated SH orders and listening conditions, the proposed acoustic-center-preserving strategies did not provide a perceptual advantage over the simpler peak-alignment preprocessing. The ERB spectral error generally reproduced the ranking of the investigated preprocessing strategies observed in the listening experiment. While the overall trends agreed well, the numerical differences between conditions were not always proportional to the corresponding perceptual ratings. A notable observation was the comparison between Method A and Method B for the Genelec 8341A. Although the ERB metric correctly predicted the superior performance of Method A, the difference between the two methods remained comparatively small. The ERB errors of the first-order anchors provide additional context for this observation. Whereas the anchors for the Genelec 8020A and Yamaha DXR8 exhibited spectral errors of approximately 6 dB, the corresponding anchor for the Genelec 8341A showed an error of only 1.6 dB, indicating that the anchor itself constituted a relatively accurate spectral approximation of the reference. Consequently, listeners may have distributed their ratings differently in this trial compared to the remaining comparisons. The magnitude-difference plots complement the ERB spectral error by revealing the frequency regions contributing to the observed deviations. For Method B, the seventh-order reconstruction of the Genelec 8341A shows localized deviations at high frequencies, despite the relatively small overall ERB spectral error. These localized differences may contribute to the perceptual differences observed in the listening experiment, particularly since broadband noise was used as the stimulus and exposes spectral deviations across a wide frequency range. For the Yamaha DXR8, both Method A and Method D exhibit pronounced deviations around the crossover region, consistent with their comparatively larger ERB spectral errors. While this may indicate that larger multi-way loudspeakers require higher SH orders to accurately reproduce the transition between the individual drivers, the performance of Method D also depends on the assumed fourth-order Linkwitz–Riley crossover used to combine the separately encoded driver responses.

Within the investigated SH orders and listening conditions, none of the proposed acoustic-center-preserving preprocessing strategies provided a consistent perceptual advantage over simpler preprocessing methods. Peak Alignment proved to be a robust baseline across all investigated loudspeakers, while the Minimum-Phase approach achieved the closest perceptual agreement with the reference, with the seventh-order anechoic condition being the only configuration not rated significantly different from the hidden reference. One possible explanation is that the effective acoustic center of a loudspeaker cannot always be represented sufficiently by either a single broadband displacement or a two-region frequency-dependent model. Future work could therefore investigate more flexible acoustic-center translation models employing multiple frequency-dependent displacement vectors. In addition, simulated loudspeakers with known driver positions and crossover characteristics could eliminate measurement uncertainties and provide a controlled environment for evaluating acoustic-center-preserving preprocessing strategies. Finally, future perceptual evaluations could consider dynamic listening scenarios, off-axis reproduction, higher SH orders, and alternative preprocessing methods such as Magnitude Least Squares (MagLS), which may better preserve loudspeaker directivity while reducing the SH order required for perceptually transparent reproduction.

References

[1] A. Szwajcowski, D. Krause, and A. Snakowska, "Error analysis of sound source directivity interpolation based on spherical harmonics," *Archives of Acoustics*, vol. 46, no. 1, 2021.

[2] J. Ahrens and S. Bilbao, "Computation of Spherical Harmonics Based Sound Source Directivity Models from Sparse Measurement Data," in *Forum Acusticum*, Lyon, France, Dec. 2020, pp. 2019–2026.

[3] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 69, no. 1/2, pp. 104–117, 2021.

[4] 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 Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2249–2262, 2019.

[5] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-corrected and time-aligned interpolation of head-related transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3783–3799, 2023.

[6] C. Imbery, S. Franz, S. Van De Par, and J. Bitzer, "Method to estimate the acoustic center of directional sources and its psychoacoustic evaluation," *Journal of the Audio Engineering Society*, vol. 66, no. 12, pp. 1062–1071, 2018.

[7] I. Ben Hagai, M. Pollow, M. Vorländer, and B. Rafaely, "Acoustic centering of sources measured by surrounding spherical microphone arrays," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2003–2015, 2011.

[8] *Acoustical terminology*, Melville, NY: Acoustical Society of America, 2013.

[9] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality.* Springer, 2019.

[10] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, Citeseer, 2010, pp. 1–6.