Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Sorbonne University/d'Alembert & IRCAM/STMS, Paris, France

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Interpolation of Sparsely Sampled Array Transfer Functions for Head-Worn Arrays

Konstantin Fontaine[1], Or Berebi[2], Boaz Rafaely[2], and Johannes M. Arend[3]

[1]*Department of Engineering Acoustics, Technische Universität Berlin, Berlin, Germany*
[2]*School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[3]*Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*

Correspondence should be addressed to Konstantin Fontaine (k.fontaine@tu-berlin.de)

ABSTRACT

Integrating microphone arrays into head-worn devices, such as augmented reality (AR) and virtual reality (VR) headsets, as well as hearing aids, has become increasingly popular for capturing and reproducing acoustic scenes. A common requirement in many such systems is a dense set of array transfer functions (ATFs). However, dense ATFs are cumbersome to measure, and practical setups commonly yield sparse grids rather than the uniform dense sampling often required. This motivates the use of interpolation to reconstruct dense ATF sets from sparse measurements. This paper evaluates spherical harmonics and natural-neighbor interpolation, each combined with onset-based time-alignment and post-interpolation magnitude correction, for a head-worn array across sampling densities. To examine how interpolation errors propagate to binaural rendering, the interpolated ATFs are substituted into two recent filter design methods: signal-independent binaural signal matching (BSM) and a signal-dependent method combining COMPASS parametric spatial coding with BSM (COM). Results show that BSM remains largely robust to interpolation errors, while COM substantially degrades under sparse sampling conditions with errors comparable to BSM at the lowest density, but achieves considerably lower errors than BSM as the sampling grid density increases. This is because BSM averages errors across all steering directions, while COM relies on individual steering vectors for source-directed beamforming.

1 Introduction

With the increasing adoption of augmented and virtual reality (AR/VR) technologies, telecommunications, and wearable audio devices, the capture and reproduction of spatial acoustic scenes has become a growing field of research [1]. Spherical microphone arrays provide a well-established framework for this task, as their regular geometry enables sound field processing in the spherical harmonics (SH) domain and facilitates binaural rendering using head-related transfer functions (HRTFs). However, the applicability of such arrays to compact head-worn devices, such as AR/VR headsets and hearing aids, is limited by their

rigid spherical structure. This has driven growing interest in binaural reproduction methods for arbitrary wearable microphone arrays.

One class of such methods is beamforming-based binaural reproduction [2], where the sound field is decomposed into directional components rendered binaurally with corresponding HRTFs. A more general and increasingly adopted approach is binaural signal matching (BSM) [3], which estimates binaural signals directly from the array recordings through linear filters that minimize the mean-square error with respect to a desired binaural reference. At high frequencies, this is commonly extended by magnitude least-squares (MagLS) optimization [4] to relax phase constraints while preserving spectral magnitude cues, with previous studies confirming perceptual improvements over related methods [5, 6]. The classical BSM formulation relies on a diffuse-field assumption that becomes increasingly invalid in directional sound fields and under head rotations. Signal-dependent extensions address this by incorporating parametric sound field models, such as COMPASS [7], that decompose the scene into direct and reverberant components [8], enabling more accurate reproduction in directional conditions at the cost of sensitivity to inaccurate spatial parameter estimates such as direction-of-arrival (DOA).

Both signal-dependent and -independent BSM formulations require a dense set of array transfer functions (ATFs), also termed steering vectors, describing the direction-dependent response of the array under the influence of the listener's head and device geometry. In practice, ATFs are commonly obtained through acoustic measurement [5, 6], while practical setups commonly yield sparse sampling grids. Algorithmic interpolation techniques to reconstruct dense spatial datasets from sparse measurements are well established for HRTFs [9, 10], where time-alignment pre-processing and post-interpolation magnitude correction have been shown to substantially reduce errors and normalize performance across methods. These findings motivate their direct transfer to ATF upsampling. However, the extent to which interpolation errors propagate to downstream binaural reproduction using recent BSM approaches remains insufficiently understood. This is the gap that the present work aims to address.

This study evaluates SH and natural-neighbor (NN) interpolation for ATFs of a head-worn array and assesses their effect on both signal-dependent and -independent BSM rendering. Section 2 provides the mathematical background, Section 3 presents the interpolation and evaluation framework, Section 4 reports the results, and Section 5 discusses the findings and Section 6 concludes the study.

2 Background

2.1 Binaural Signal Matching

Consider a sound field composed of Q far-field sound sources arriving from directions Ω_q, where ϕ denotes the azimuth angle measured counterclockwise from the frontal direction at $\phi = 0^\circ$, and θ denotes the elevation angle with $\theta = 90^\circ$ pointing upwards. The $M \times 1$ array input signal vector is modeled as

$$\mathbf{x}(k) = \mathbf{V}(k)\mathbf{s}(k) + \mathbf{n}(k), \tag{1}$$

where $\mathbf{s}(k)$ is the $Q \times 1$ source signal vector, $\mathbf{V}(k)$ is the $M \times Q$ ATF matrix, and $\mathbf{n}(k)$ represents spatially uncorrelated sensor noise. The ideal target binaural signals at the listener's left and right ears are given by

$$p^{l,r}(k) = [\mathbf{h}^{l,r}(k)]^{\mathrm{T}}\mathbf{s}(k), \tag{2}$$

where $\mathbf{h}^{l,r}(k)$ contains the target HRTFs for the left and right ear, denoted by $(\cdot)^l$ and $(\cdot)^r$, respectively. The estimated binaural signals rendered from the M-channel wearable array are obtained via linear filtering:

$$\hat{p}^{l,r}(k) = [\mathbf{c}^{l,r}(k)]^{\mathrm{H}}\mathbf{x}(k). \tag{3}$$

For brevity, the frequency/wave-number index k is omitted in the subsequent filter formulations.

The signal-independent BSM method derives fixed rendering filters by minimizing the mean-square error between the target and estimated signals over a uniform diffuse field [11]. This yields the closed-form BSM filter

$$\mathbf{c}^{l,r}_{\mathrm{BSM}} = \left(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \frac{\sigma_n^2}{\sigma_s^2}\mathbf{I}_M\right)^{-1}\mathbf{V}[\hat{\mathbf{h}}^{l,r}]^*, \tag{4}$$

where σ_s^2 and σ_n^2 denote the source and noise variances, and $\hat{\mathbf{h}}^{l,r}$ represents an HRTF modified using the MagLS approach [4, 12, 13] to preserve the target magnitude while relaxing phase constraints above a transition frequency, which is usually around 1.5–2 kHz.

2.2 Binaural Signal Matching Incorporating Parametric Spatial Coding

To improve reproduction in highly directional sound fields, signal-dependent extensions of BSM incorporate the parametric spatial coding method COMPASS [7], yielding COMPASS-BSM (COM) [8]. Here, the acoustic scene is first decomposed into discrete direct sources and a diffuse residual. The extraction of the direct sound components relies on the DOAs of the dominant sources. While practical implementations typically employ direction-finding algorithms such as MUSIC or ESPRIT to estimate these DOAs, the present evaluation assumes the true DOAs to be available. This assumption strictly isolates the performance degradation caused by ATF interpolation errors from concurrent DOA estimation inaccuracies.

To isolate the direct sound components, the COMPASS method employs beamforming. While a minimum-variance distortionless response (MVDR) approach is common, this work utilizes a linearly constrained minimum-power (LCMP) formulation. Its $D \times M$ weight matrix $\mathbf{W}$ is computed using the spatial covariance matrix of the microphone signals $\mathbf{R}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{H}}]$ and the $M \times D$ steering matrix $\mathbf{V}_d$ corresponding to the DOAs of the direct sources:

$$\mathbf{W} = \left(\mathbf{V}_d^{\mathrm{H}}\mathbf{R}_x^{-1}\mathbf{V}_d + \lambda \mathbf{I}_D\right)^{-1}\mathbf{V}_d^{\mathrm{H}}\mathbf{R}_x^{-1}, \tag{5}$$

where λ is a regularization parameter.

The overall COM filter is then formulated as a single composite filtering operation that renders these direct sources and a spatial projection of the residual diffuse field [6]:

$$\mathbf{c}_{\mathrm{COM}}^{l,r} = \mathbf{W}^{\mathrm{H}}[\mathbf{h}_d^{l,r}]^* + (\mathbf{I}_M - \mathbf{V}_d\mathbf{W})^{\mathrm{H}}\mathbf{c}_{\mathrm{BSM}}^{l,r}, \tag{6}$$

where $\mathbf{h}_d^{l,r}$ contains the target HRTFs associated with the direct source directions. In this formulation, the subtracted residual is rendered utilizing the signal-independent BSM filters $\mathbf{c}_{\mathrm{BSM}}^{l,r}$ from Eq. (4).

3 Methods

3.1 Array Transfer Function Interpolation

Having established the mathematical basis for BSM and COM, this section presents the ATF interpolation, also referred to as spatial upsampling, based on well-established algorithmic HRTF interpolation methods [9, 10]. HRTF interpolation is performed independently for each channel, and both HRTFs and ATFs involve direction-dependent transfer functions with spatially varying phase and magnitude, which suggests good applicability of the methods across varying head-worn arrays. Two interpolation approaches are investigated: spherical harmonics (SH) as a global and natural-neighbor (NN) as a local interpolation method [14]. Both methods are examined with onset-based time-alignment (OBTA) for preprocessing combined with magnitude correction (MC) for post-processing [9, 10].

In SH interpolation, the sparse set of pre-processed ATFs $\tilde{\mathbf{V}}(k)$ is decomposed into SH coefficients via the discrete spherical Fourier transform (SFT). Given Q sampling points forming the matrix $\mathbf{Y} \in \mathbb{C}^{Q\times(N+1)^2}$ of SH basis functions evaluated at the sampling points, the SH coefficients $\mathbf{v}_{nm}$ are obtained as

$$\mathbf{v}_{nm}(k) = \left(\mathbf{Y}^H\mathbf{Y} + \varepsilon\mathbf{D}\right)^{-1}\mathbf{Y}^H\tilde{\mathbf{V}}(k), \tag{7}$$

where ε denotes the Tikhonov regularization parameter and $\mathbf{D}$ the order-weighted diagonal damping matrix [15], with $[\mathbf{D}]_{nn} = (1+n(n+1))$. Tikhonov regularization stabilizes the SFT inversion for irregular grids, for which $\mathbf{Y}$ is often ill-conditioned [16]. Interpolated ATFs at any target direction Ω_q are then reconstructed via the inverse SFT,

$$\left[\hat{V}(k)\right]_{i,q} = \sum_{n=0}^{N}\sum_{m=-n}^{n}\mathbf{v}_{nm}(k)\,Y_n^m(\Omega), \tag{8}$$

where i denotes the microphone index. The maximum resolvable SH order is bounded by the spatial sampling density as $N \leq \lfloor\sqrt{Q/\lambda} - 1\rfloor$, where $\lambda \geq 1$ depends on the sampling scheme [9], such that sparse grids inherently limit both frequency range and spatial accuracy of the interpolation.

As an alternative, without transforming the ATFs to the SH domain, NN interpolation reconstructs ATFs at target directions directly from spatially adjacent measurements with weights derived from Voronoi-diagrams [17]. In contrast to the global nature of SH interpolation, local inaccuracies therefore only affect their spatial neighborhood [14]. The method is particularly robust at very low sampling densities, where the SH representation suffers from undersampling artifacts [18].



Fig. 1: ARIA glasses array geometry.

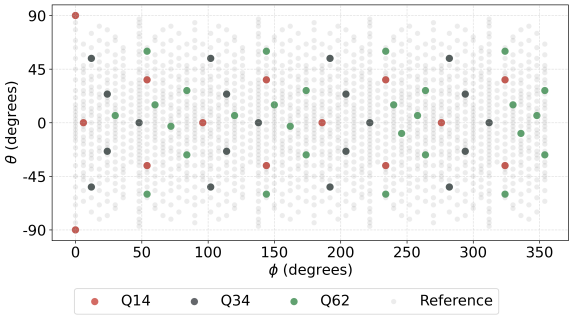


Fig. 2: Sparse sampling grids downsampled from the dense reference grid ($Q_{ref} = 1674$) for $Q = \{14, 34, 62\}$ spatial positions. Each sparse grid forms a subset of the next denser one and is therefore shown jointly as overlaid sampling points on the reference grid.

Across both methods, a primary source of interpolation error stems from rapid spatial phase variation, as direction-dependent propagation delays cause comb-filters and double-peaks when interpolated directly [14]. OBTA mitigates this by estimating and removing the time-of-arrival (TOA) of the ATFs prior to interpolation, interpolating the TOAs separately from the ATFs using the interpolation method employed, and combining the interpolated TOAs with the interpolated ATFs afterward [9]. Following interpolation and TOA reconstruction, a magnitude correction step is applied as proposed in [10], using an auditory-smoothed and interpolated version of the sparse ATF set as a reference to derive magnitude correction filters that reduce remaining magnitude errors in the final dense ATF set. For HRTFs, this post-processing largely normalizes performance differences between SH and NN interpolation [19].

3.2 Materials

The seven-microphone *Project Aria* glasses [20] were used for evaluation, with the array geometry shown in Fig. 1. The ATFs, initially measured for $Q_{\text{ref}} = 1674$ directions in anechoic far-field conditions with the glasses mounted on a KEMAR head-and-torso simulator [21], were spatially subsampled to uniformly distributed grids containing $Q = \{14, 34, 62\}$ directions, as illustrated in Fig. 2. For each grid, ATFs are interpolated to the dense reference using the methods described in Section 3.1, yielding the estimated steering matrices $\hat{\mathbf{V}}_{\text{SH},Q}(k)$ and $\hat{\mathbf{V}}_{\text{NN},Q}(k)$. The reference $\mathbf{V}(k)$ is retained as ground truth throughout.

For binaural reproduction, target HRTFs are based on Neumann KU100 measurements from the Cologne database [22], sampled at 48 kHz on a Lebedev grid with 2702 positions. To evaluate reproduction accuracy, a separate single-source anechoic simulation was conducted independently for every direction in the dense reference grid. For a target source arriving from direction Ω_q, the captured microphone array signals were modeled using the corresponding ground-truth reference ATF, such that the input signal vector was $\mathbf{x} = \mathbf{v}_{\text{ref}}(\Omega_q)$.

The binaural rendering filters were constructed using the interpolated steering matrices $\hat{\mathbf{V}}(k) \in \{\hat{\mathbf{V}}_{\text{SH}}(k), \hat{\mathbf{V}}_{\text{NN}}(k)\}$ and the measured reference $\mathbf{V}(k)$, while assuming a signal-to-noise ratio of 20 dB. For BSM, the full matrix $\hat{\mathbf{V}}(k)$ was substituted into Eq. (4) to yield the filter coefficients $\mathbf{c}^{l,r}_{\text{BSM}}$. For COM, the composite rendering filter $\mathbf{c}^{l,r}_{\text{COM}}$ was formulated according to Eq. (6). In this formulation, the beamformer weights were computed using the specific interpolated steering vector $\hat{\mathbf{v}}_d(\Omega_q)$ corresponding to the target source direction. Meanwhile, the direct-sound HRTF selection relied on the reference DOA to strictly isolate the effect of steering vector inaccuracies from DOA estimation errors. Finally, the evaluated binaural signals $\hat{p}^{l,r}$ were reproduced by applying these respective interpolated filters to the ground-truth array input $\mathbf{x}$, yielding an independent performance measurement for each individual source direction.

3.3 Evaluation Metrics

Interpolation performance is evaluated with respect to both ATF reconstruction accuracy and downstream binaural reproduction quality, employing two metrics for each. ATF reconstruction is assessed using the normalized mean square error (NMSE) and the logarithmic magnitude error (LME), both evaluated from

20 Hz to 20 kHz operating on the reference measurement $V_{\mathrm{ref},i}(k,\Omega)$ and interpolated transfer functions $\hat{V}_i(k,\Omega)$, for microphone channel i, wavenumber k, and direction Ω. The NMSE quantifies the complex-valued reconstruction error, thereby capturing both spectral magnitude and phase deviations,

$$\varepsilon_{\mathrm{NMSE}}(k,\Omega,i) = 10\log_{10}\frac{|V_{\mathrm{ref},i}(k,\Omega)-\hat{V}_i(k,\Omega)|^2}{|V_{\mathrm{ref},i}(k,\Omega)|^2}. \tag{9}$$

To separately assess magnitude reconstruction independent of phase variations, the LME is computed as the unsigned logarithmic deviation between the reference and interpolated ATF magnitudes,

$$\varepsilon_{\mathrm{LME}}(k,\Omega,i) = 10\log_{10}\frac{|V_{\mathrm{ref},i}(k,\Omega)|^2}{|\hat{V}_i(k,\Omega)|^2}. \tag{10}$$

For binaural reproduction quality, the LME is additionally applied to the reference $P^{l,r}(k,\Omega)$ and reproduced binaural spectra $\hat{P}^{l,r}(k,\Omega)$ by substituting these into Eq. (10) for $V_{\mathrm{ref},i}$ and $\hat{V}_i$, where the superscripts $(.)^{l,r}$ replace the channel index i and denote the left and right ear, respectively.

To obtain a perceptually motivated estimate of coloration, the predicted binaural coloration (PBC-2) [23] is computed by summing band-weighted absolute level differences between reference and test signals across K equivalent rectangular bandwidth (ERB) bands, yielding $C^{l,r}[k]$ for the left and right ear. With level-based binaural weights $g^{l,r}[k]$ that assign a higher perceptual relevance to the ipsilateral ear,

$$\text{PBC-2} = \sum_{k=1}^{K}\left(g^{l}[k]\,C^{l}[k] + g^{r}[k]\,C^{r}[k]\right), \tag{11}$$

yielding a scalar estimate of perceived binaural coloration between the reference and test signal on a MUSHRA scale from 0 to 100, with lower values indicating smaller perceptual coloration differences.

4 Results

4.1 Array Transfer Function Reconstruction

Figure 3 shows the frequency-dependent LME averaged over spatial positions and microphone channels, while Table 1 summarizes mean, standard deviation, and peak errors for LME and NMSE. As expected, errors increase monotonically toward higher frequencies and with decreasing sampling density. Overall, SH and NN interpolation perform comparably well, though the deviation between the two methods increases toward higher frequencies, with NN showing a slightly improved mean LME across all conditions. The spatial distribution in Fig. 4 further shows that errors concentrate in the lower hemisphere, in particular around rear directions at $\theta \approx -70°$, and spatially diminish as sampling density increases. Informal, channel-wise inspection further showed elevated errors for the rear microphones at corresponding contralateral incident angles. NMSE results follow the same trend, with SH exhibiting consistently lower NMSE peak errors and standard deviations across all grids.

Table 1: Summary of the mean, standard deviation and peak errors in ATF reconstruction across spatial positions and microphones.

	LME (dB)			NMSE (dB)		
Condition	**Mean**	**Std**	**Peak**	**Mean**	**Std**	**Peak**
NN						
Q_{14}	3.1	1.4	10.9	-2.6	7.1	10.7
Q_{34}	2.6	1.3	9.7	-6.6	9.1	8.7
Q_{62}	**2.2**	**1.3**	9.4	**-10.3**	10.9	9.2
SH						
Q_{14}	3.3	**1.3**	11.6	-3.2	**3.9**	10.6
Q_{34}	2.9	**1.3**	**8.9**	-6.4	4.5	7.3
Q_{62}	2.6	**1.3**	9.7	-8.8	4.9	**5.5**

4.2 Binaural Reproduction

Figure 5 shows the frequency-dependent left-ear LME for both interpolation and reproduction methods, av-

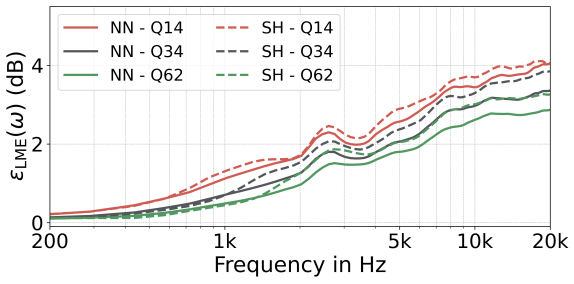


Fig. 3: Frequency-dependent LME for NN and SH interpolation across the sampling densities Q_{14}, Q_{34}, and Q_{62}.

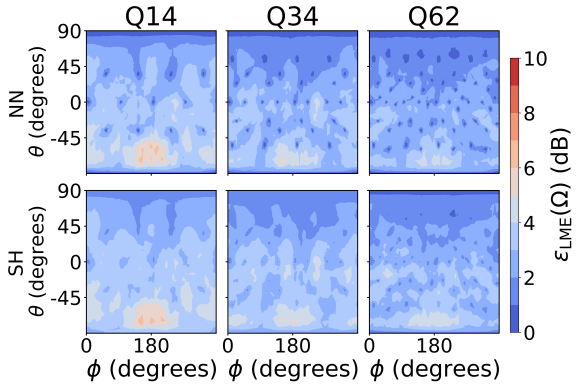


Fig. 4: Spatial distribution of the channel-averaged LME for both interpolation methods across the sampling densities Q_{14}, Q_{34}, and Q_{62}.

eraged over incident angles Q_{ref}. Additionally, Table 2 summarizes the mean, standard deviation and peak errors throughout all conditions, averaged over frequencies. Across the conditions, binaural reproduction quality is largely independent of the interpolation method but highly dependent on reproduction method. BSM errors are mostly invariant across sampling densities with a mean LME of approximately 5 dB, indicating that the interpolated ATFs introduce no additional degradation relative to the dense reference. COM, in contrast, exhibits strong density dependence above 1 kHz. At Q_{14}, COM yields a mean LME of 5.1 dB, matching BSM levels across all frequencies, and improves to 3.0 dB at Q_{62} with NN interpolation.

The spatial distributions of the LME in Fig. 6 (left) reflect dependency trends on sampling grid and interpolation method. BSM errors concentrate at the lower hemisphere cap below $\theta = -70°$ and at contralateral incident angles, with negligible differences across grid densities. In contrast, COM errors decrease with increasing grid density, while peak errors are located at the same lower rear region and contralateral incident angles. Informal right-ear assessment showed that the critical regions in both BSM and COM mirror across the interaural axis.

The PBC-2 distributions in Fig. 6 (right) and values reported in Table 2 follow the LME patterns closely. BSM yields stable mean coloration around 43 across all conditions, while COM decreases from approximately 42 at Q_{14} to 21 at Q_{62}. At Q_{14}, COM coloration

Table 2: Summary of the mean, standard deviation, peak errors in binaural reproduction across the spatial positions.

	LME			PBC-2		
Condition	**Mean**	**Std**	**Peak**	**Mean**	**Std**	**Peak**
BSM_NN						
ARIA	4.9	1.6	11.2	43.5	10.9	75.2
Q_{14}	5.2	1.6	10.7	42.6	9.1	74.6
Q_{34}	5.1	1.6	10.5	42.2	9.6	74.6
Q_{62}	4.9	1.6	10.8	42.4	10.1	74.4
BSM_SH						
ARIA	4.9	1.6	11.2	43.5	10.9	75.2
Q_{14}	5.1	**1.4**	10.2	43.5	11.1	75.3
Q_{34}	5.1	**1.5**	9.7	44.4	11.2	77.9
Q_{62}	4.9	1.6	11.3	43.8	11.3	79.8
COM_NN						
ARIA	0.0	0.0	0.2	0.0	0.1	1.3
Q_{14}	5.1	2.1	11.5	42.4	11.5	86.1
Q_{34}	3.9	1.9	11.1	30.0	10.7	70.1
Q_{62}	**3.0**	1.7	10.4	21.4	8.9	61.0
COM_SH						
ARIA	0.0	0.0	0.2	0.0	0.1	1.3
Q_{14}	5.0	2.0	12.2	39.1	**8.0**	65.4
Q_{34}	3.7	1.7	9.0	28.1	9.3	61.3
Q_{62}	3.1	**1.4**	**8.9**	**21.3**	8.5	**57.8**

is comparable to BSM, whereas lower values are registered at Q_{34} and Q_{62}, which indicate that the perceptual advantage of signal-dependent rendering emerges only above a minimum ATF sampling density. Notably, it could be observed that spatially critical LME

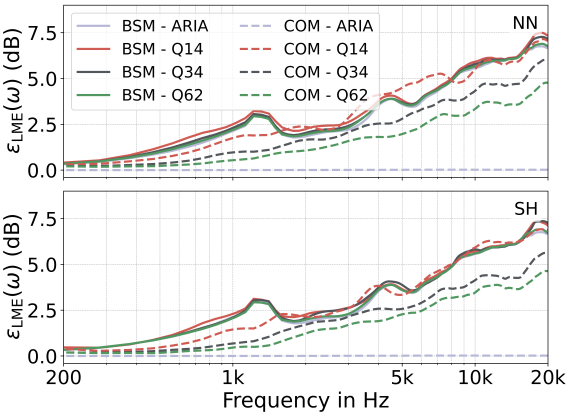


Fig. 5: Frequency-dependent left-ear LME for both reproduction and interpolation methods across the sampling densities Q_{14}, Q_{34} and Q_{62}

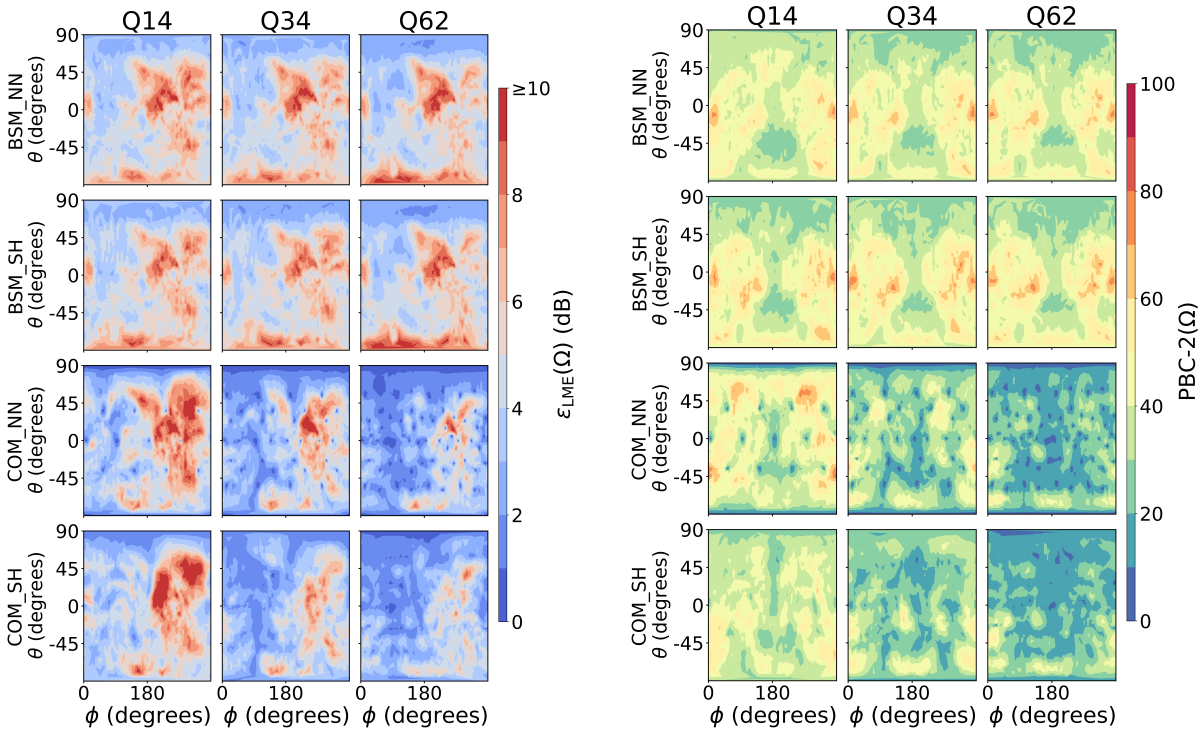


Fig. 6: Spatial distribution of the left-ear LME (left) and PBC-2 (right) for both reproduction and interpolation methods across the sampling densities Q_{14}, Q_{34}, and Q_{62}.

regions do not necessarily translate to correspondingly high coloration, as elevated LME values in several areas yield only moderate PBC-2 values.

5 Discussion

5.1 Array Transfer Function Reconstruction

The closely matched LME of NN and SH at each sampling density confirms that systematic differences between the two interpolation methods are largely equalized, consistent with findings for HRTF interpolation [10, 19]. The slightly lower errors of NN at sparse grids are expected given its local interpolation scheme, which avoids the global undersampling artifacts that affect SH at low sampling densities [14]. The elevated errors in the lower rear hemisphere and at contralateral incident angles of the rear microphones reflect the acoustic shadow cast by the head, which introduces spatially localized spectral complexity that is inherently difficult to reconstruct from sparse measurements and reflects a pattern well established in HRTF interpolation [9, 10].

5.2 Binaural Reproduction

The differing sensitivity of BSM and COM to ATF interpolation errors follows directly from their filter designs. Because BSM minimizes the mean-squared error averaged over all steering directions, it inherently suppresses the influence of direction-specific ATF inaccuracies, explaining its robust and stable performance across all sampling densities [3]. COM, by contrast, relies on accurate individual steering vectors for source-directed beamforming. At sparse grids (e.g., Q_{14}), localized ATF interpolation errors severely compromise the beamformer's ability to cleanly isolate the direct sound. Because the COM framework relies on a subtraction operation to estimate the diffuse field (i.e., the residual), an erroneous estimate of the direct signal causes the remaining direct sound energy to leak into the residual rendering path. In the anechoic conditions considered here, any leaked energy is therefore rendered entirely through the BSM diffuse-field path, which introduces phase distortion and musical noise [24] into the composite binaural signal.

It is important to note that the present evaluation strictly isolates the impact of ATF mismatch by utilizing true DOAs and single-source anechoic conditions. In practical applications, concurrent challenges such as DOA estimation errors and dense reverberation will further impair the beamformer's ability to isolate the direct source and further compound these subtraction-induced artifacts, while elevating errors above the near-zero error observed here for the reference condition. Furthermore, such complex, nonlinear distortions are often inadequately captured by standard objective metrics, yet they can significantly degrade perceived spatial and timbral quality in subjective listening tests.

The near-equivalence of NN and SH across all binaural metrics confirms that the choice of interpolation method remains secondary to applying appropriate pre- and post-processing, which is consistent with HRTF interpolation [10]. Overall, these findings demonstrate a practical trade-off: signal-dependent parameterization offers significant perceptual benefits when the ATF grid is sufficiently dense before interpolation, and scene estimation is accurate, whereas signal-independent BSM provides a robust and reliable baseline under conditions of low grid density.

6 Conclusion

This work evaluated SH and NN interpolation of ATFs from a head-worn glasses array across three sampling densities, assessing their effect on signal-dependent and signal-independent BSM rendering under anechoic conditions. Both interpolation methods yield comparable reconstruction accuracy across all densities. Errors are shown to concentrate in the lower hemisphere and at contralateral directions and increase toward higher frequencies at sparser grids. BSM remains robust to interpolation errors across all conditions, while COM degrades substantially at sparse grids but recovers with increasing density. Despite its sensitivity, COM yields coloration comparable to BSM at the sparsest grid but maintains lower coloration at moderate and dense sampling densities. This suggests that, once a minimum spatial sampling density is met, signal-dependent rendering retains its perceptual advantages.

References

[1] Rafaely, B., Tourbabin, V., Habets, E., Ben-Hur, Z., Lee, H., Gamper, H., Arbel, L., Birnie, L., Abhayapala, T., and Samarasinghe, P., "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes–review and challenges," *Acta Acustica*, 6, p. 47, 2022.

[2] Davis, L. S., Duraiswami, R., Grassi, E., Gumerov, N. A., Li, Z., and Zotkin, D. N., "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues," in *Audio Engineering Society Convention 119*, Audio Engineering Society, 2005.

[3] Madmoni, L., Ben-Hur, Z., Donley, J., Tourbabin, V., and Rafaely, B., "Design and analysis of binaural signal matching with arbitrary microphone arrays and listener head rotations," *EURASIP Journal on Audio, Speech, and Music Processing*, 2025(1), p. 11, 2025.

[4] Schörkhuber, C., Zaunschirm, M., and Holdrich, R., "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of 44th DAGA*, pp. 339–342, Munich, Germany, 2018.

[5] Lübeck, T., Amengual Garí, S. V., Calamia, P., Alon, D. L., Crukley, J., and Ben-Hur, Z., "Perceptual evaluation of approaches for binaural reproduction of non-spherical microphone array signals," *Frontiers in Signal Processing*, 2, p. 883696, 2022.

[6] McCormack, L., Meyer-Kahlen, N., Alon, D. L., Ben-Hur, Z., Amengual Garí, S. V., and Robinson, P., "Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture," *Journal of the Audio Engineering Society*, 71(10), pp. 638–649, 2023.

[7] Mc Cormack, L. and Politis, A., "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *Proceedings of the AES Conference on Immersive and Interaktive Audio*, pp. 1–12, York, UK, 2019.

[8] Berger, A., Yifrach, M., Berebi, O., Tourbabin, V., Donley, J., Ben-Hur, Z., and Rafaely, B., "Performance and robustness of signal-dependent vs. signal-independent binaural signal matching with wearable microphone arrays," *Journal on Audio, Speech, and Music Processing*, 2026.

[9] Arend, J. M., Brinkmann, F., and Pörschmann, C., "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *Journal of the Audio Engineering Society*, 69(1/2), pp. 104–117, 2021.

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

[11] Madmoni, L., Donley, J., Tourbabin, V., and Rafaely, B., "Beamforming-based binaural reproduction by matching of binaural signals," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society, 2020.

[12] Zotter, F. and Frank, M., *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer-Verlag, 2019, doi:10.1007/978-3-030-17207-7.

[13] Madmoni, L., Donley, J., Tourbabin, V., and Rafaely, B., "Binaural reproduction from microphone array signals incorporating head-tracking," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–5, IEEE, 2021.

[14] Pörschmann, C., Arend, J. M., Bau, D., and Lübeck, T., "Comparison of spherical harmonics and nearest-neighbor based interpolation of head-related transfer functions," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality, Redmond, WA, USA*, pp. 1–10, Audio Engineering Society, 2020.

[15] Duraiswami, R., Zotkin, D. N., and Gumerov, N. A., "Interpolation and range extrapolation of HRTFs [head related transfer functions]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pp. iv–iv, IEEE, 2004.

[16] Bau, D., Arend, J. M., and Pörschmann, C., "Estimation of the optimal spherical harmonics order for the interpolation of head-related transfer functions sampled on sparse irregular grids," *Frontiers in Signal Processing*, 2, p. 884541, 2022.

[17] Sibson, R., "A brief description of natural neighbour interpolation," *Interpreting multivariate data*, pp. 21–36, 1981.

[18] Bernschütz, B., Giner, A. V., Pörschmann, C., and Arend, J., "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, 100(5), pp. 972–983, 2014.

[19] Arend, J. M., Pörschmann, C., Weinzierl, S., and Brinkmann, F., "Magnitude-corrected and time-aligned HRTF interpolation: Effect of interpolation and alignment method," *Proc. 49th DAGA*, pp. 1098–1101, 2023.

[20] Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al., "Project aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[21] Žmolíková, K., Merello, S., Kalgaonkar, K., Lin, J., Moritz, N., Ma, P., Sun, M., Chen, H., Saliou, A., Petridis, S., et al., "The CHiME-8 MMCSG Challenge: Multi-modal conversations in smart glasses," in *Proc. CHiME 2024*, pp. 7–12, 2024.

[22] Bernschütz, B., "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*, volume 29, German Acoustical Society (DEGA) Berlin, 2013.

[23] McKenzie, T. and Brinkmann, F., "Toward an improved auditory model for predicting binaural coloration," *Journal of the Audio Engineering Society*, 73(3), pp. 115–126, 2025.

[24] McCormack, L., Politis, A., Gonzalez, R., Lokki, T., and Pulkki, V., "Parametric ambisonic encoding of arbitrary microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 2062–2075, 2022.