

SHB-AE: SPHERICAL HARMONIC BEAMFORMING BASED AMBISONICS ENCODING AND UPSCALING METHOD FOR SMARTPHONE MICROPHONE ARRAY

Yuhuan You¹, Yufan Qian¹, Tianshu Qu¹, Bin Wang², Xueyang Lv³

¹Peking University

²Beijing Xiaomi Mobile Software Co., Ltd

³Xiaomi Communications Co., Ltd

*Accepted for presentation at
AES Europe 2025 Convention
(AES 158th Convention)
Warsaw, Poland, May 22-24, 2025.
This is the authors' version prepared
without AES graphical style or template.*

ABSTRACT

With the rapid development of virtual reality (VR) and augmented reality (AR), spatial audio recording and reproduction have gained increasing research interest. Higher Order Ambisonics (HOA) stands out for its adaptability to various playback devices and its ability to integrate head orientation. However, current HOA recordings often rely on bulky spherical microphone arrays (SMA), and portable devices like smartphones are limited by array configuration and number of microphones. We propose SHB-AE, a spherical harmonic beamforming based method for Ambisonics encoding using a smartphone microphone array (SPMA). By designing beamformers for each order of spherical harmonic functions based on the array manifold, the method enables Ambisonics encoding and up-scaling. Validation on a real SPMA and its simulated free-field counterpart in noisy and reverberant conditions showed that the method successfully encodes and up-scales Ambisonics up to the fourth order with just four irregularly arranged microphones.

1. INTRODUCTION

With the rapid development of augmented reality (AR) and virtual reality (VR) technologies, the demand for immersive spatial audio experiences has been continuously rising, leading to growing attention on spatial audio capture and reproduction technologies. Among these, encoding microphone-captured sound pressure signals into Higher-Order Ambisonics (HOA) format offers multiple advantages[1]. HOA signals are device-independent and can be used in various playback environments, such as headphones and loudspeaker arrays. Additionally, HOA signals can be conveniently implemented with head rotation and the head-related transfer function (HRTF) through the Wigner-D function and spherical harmonic coefficients [2].

However, existing Ambisonics encoding algorithms typically rely on bulky spherical microphone arrays to capture sound pressure signals [1], making them difficult to adapt to portable electronic devices not specifically designed for spatial sound capture, such as smartphones, glasses, and other wearable devices [3]. These devices often have irregular microphone arrangements, limited microphone counts, and complex sound scattering, which limits accurate Ambisonics encoding [4].

To address these challenges, several approaches have been proposed for Ambisonics encoding with irregular arrays [5–9]. These methods can be broadly classified into data-driven and model-driven approaches. Heikkinen et al. [5] introduced a data-driven method for Neural Ambisonics Encoding, where a U-net structured neural network is trained to directly encode Ambisonics from sound pressure signals captured by irregular microphone arrays in an end-to-end manner. Building on this, Qiao et al. [6] regularized inter-channel correlations of the Ambisonics signals and utilized a circular microphone array for Ambisonics encoding of multi-speaker speech signals. However, like many data-driven methods, these approaches face challenges related to generalization errors in the absence of actual microphone-captured signals, which are difficult to obtain.

One type of model-driven approach is based on analyzing the captured sound field. For example, McCormack et al. [7] introduced a primary-ambience directional model of the sound field, using spatial filtering techniques to divide the captured sound field into its individual source and directional ambient components, which are then encoded into the Ambisonics format at an arbitrary order. However, these methods rely not only on prior assumptions about the sound field but also on the impact of direction-of-arrival (DOA) estimation and source separation results. Accurate DOA estimation and source separation with SPMA is a challenging task, and most studies assume high-precision HOA recordings as input for these tasks [10, 11].

Another type of model-driven approach is signal-independent encoding, which directly formulates a system of equations linking HOA coefficients and captured sound

pressure signals [8], solving for the optimal encoding matrix in the least squares sense. Bastine et al. [12] directly formulated a linear system of equations between HOA coefficients and the array manifold of head-wearable devices to achieve encoding for microphone arrays on such devices. Geyer et al. [9] further introduced Tikhonov regularization [13] to improve the performance of encoded HOA signals in binaural rendering tasks. However, these methods are still constrained by the number of microphones Q in the array, with the limitation $Q \geq (N + 1)^2$ [4]. Gao et al. [14], correspondingly, designed beamformers to achieve Ambisonics encoding, utilizing 2.5D spherical harmonic functions as target beam patterns.

We propose SHB-AE, a beamforming-based method for Ambisonics encoding and upscaling. Inspired by prior beamforming work [14], we design beamformers corresponding to spherical harmonic functions as target beam patterns, enabling direct output of HOA coefficients for each order without being limited by microphone array configuration or microphone numbers. The limitation of microphone numbers is converted into the limitation of the number of measured steering vectors, which allows the up-scaling of Ambisonics encoding. Experimental results demonstrate superior performance over conventional least-square HOA encoding across various reverberant environments and noise conditions.

The remainder of the paper is organized as follows. Section 2 introduces the fundamental knowledge of Ambisonics encoding and the proposed method. Section 3 presents experimental validation of the method's effectiveness across simulated and measured array manifolds of an SPMA under multiple conditions. Section 4 concludes the paper.

2. METHODS

2.1. Ambisonics Encoding

Ambisonics is a spatial audio technique that allows for the reproduction of three-dimensional sound fields using spherical harmonics. This method enables immersive sound experiences by representing sound from various directions, which can be accurately reproduced across different listening environments.

Consider a plane wave $e^{i\mathbf{k}\cdot\mathbf{r}_q}$ in spherical coordinates where $\mathbf{k} = (k, \theta_k, \phi_k)$ stands for a plane wave arriving from (θ_k, ϕ_k) , the sound pressure $p(k, \mathbf{r}_q)$ at the q -th microphone $\mathbf{r}_q = (r_q, \theta_q, \phi_q)$ is [4]

$$e^{i\mathbf{k}\cdot\mathbf{r}_q} = p(k, r_q, \theta_q, \phi_q) = 4\pi \sum_{n=0}^N \sum_{m=-n}^n b_n(kr_q) B_n^m(\theta_k, \phi_k) Y_n^m(\theta_q, \phi_q) \quad (1)$$

where $b_n(kr) = i^n j_n(kr)$, $j_n(kr)$ denotes the spherical Bessel function. When the truncation order $N \rightarrow \infty$, the above equation holds. The spherical harmonics $Y_n^m(\theta, \phi)$ are

defined as

$$Y_n^m(\theta, \phi) \equiv \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos \theta) e^{im\phi} \quad (2)$$

with P_n^m being the associated Legendre functions. Within this framework, the theoretical Ambisonics coefficient $B_n^m(\theta_k, \phi_k)$ is given by

$$B_n^m(\theta_k, \phi_k) = [Y_n^m(\theta_k, \phi_k)]^* \quad (3)$$

In practice, Ambisonics signals are obtained using a microphone array and projected onto spherical harmonic bases as

$$B_n^m = \frac{1}{4\pi} \sum_{q=1}^Q w_q p(\theta_q, \phi_q) [Y_n^m(\theta_q, \phi_q)]^* \quad (4)$$

where Q is the number of microphones, $q = 1, \dots, Q$ and w_q denote the microphone index and its sampling weight respectively.

In our simulation and practical scenarios, $Q = 4$, while $p, b_n(kr_q), Y_n^m(\theta_k, \phi_k)$ can be calculated from available data. Therefore, for each sound source direction (θ_k, ϕ_k) , we can use Equation 1 to express the sound pressure at all four microphones, thus establishing a linear equations system for the coefficients B_n^m of different orders. When the truncation order $N = 1$ (First Order Ambisonics), the system is well-determined with an exact solution, though severely affected by spatial aliasing. For higher truncation orders, the system becomes underdetermined, thus we employ pseudo-inverse to obtain the minimum norm solution

$$\hat{B}_n^m = (\mathbf{A}^\dagger \mathbf{p})_{(n,m)}, \quad (5)$$

where $\mathbf{p}$ is the sound pressure vector. $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{A}$, serving as the Ambisonics encoding matrix, while matrix $\mathbf{A}$ has elements

$$A_{q,(n,m)} = 4\pi b_n(kr_q) Y_n^m(\theta_q, \phi_q) \quad (6)$$

2.2. Beamforming Based Ambisonics Encoding

Consider an array of Q omni-directional microphones at (r_q, θ_q, ϕ_q) and a set of K plane waves arriving from (θ_k, ϕ_k) , the array manifold matrix is denoted as $\mathbf{D}(\omega)$ with dimensions $Q \times K$, where each element $[\mathbf{D}(\omega)]_{q,k}$ represents the frequency response of the plane wave source k to microphone q . The sound pressure measured by the array can be expressed as

$$\mathbf{p}(\omega) = \mathbf{D}(\omega) \mathbf{s}(\omega) + \mathbf{n}(\omega), \quad (7)$$

where $\mathbf{p}(\omega) = [p_1(\omega), \dots, p_Q(\omega)]^T$ is the sound pressure vector, $\mathbf{s}(\omega) = [s_1(\omega), \dots, s_K(\omega)]^T$ is the source signal vector and $\mathbf{n}(\omega) = [n_1(\omega), \dots, n_Q(\omega)]^T$ is the noise vector.

Through applying beamforming to the signals, the beamformer outputs

$$Z(\omega) = \mathbf{h}(\omega)\mathbf{p}(\omega), \quad (8)$$

where $\mathbf{h}(\omega) = [h_1(\omega), \dots, h_Q(\omega)]$ is the beamforming weight vector. Comparing equation (8) and equation (4) leads to the finding that Ambisonics encoding can be done through designing a beamformer with the target beam pattern being the spherical harmonics. Neglecting the noise vector, vectoring the spherical harmonics $\mathbf{y}_n^m = [Y_n^m(\theta_1, \phi_1), \dots, Y_n^m(\theta_K, \phi_K)]$, substituting equation (4) into equation (8) and omitting source signal $\mathbf{s}(\omega)$ reveals that

$$\mathbf{h}(\omega)\mathbf{D}(\omega) = \mathbf{y}_n^m. \quad (9)$$

Furthermore, although the Q microphones are not necessarily distributed evenly on a sphere, the steering vectors (each column of $\mathbf{V}$) for K arriving plane waves can be measured following a specific spherical distribution, analogous to the measurement process of head-related transfer functions. This spatial distribution enables us to apply discrete spherical harmonic transform (DSHT) [4] to simplify the right side of equation (9) into a one-hot vector as

$$\mathbf{h}(\omega)\mathbf{D}(\omega)\mathbf{S}^T = [0, \dots, \underbrace{1}_{l=n^2+n+m+1}, \dots, 0]^T, \quad (10)$$

where $\mathbf{S}$ is the DSHT matrix in [4], consisting of spherical harmonic function values and sampling weights based on sampling scheme, and $l = n^2 + n + m + 1$ stands for the one-dimensional index of spherical harmonics. Since DSHT projects the signal on the basis of spherical harmonics, the orthogonality of spherical harmonics ensures the validity of the formula $\mathbf{y}_n^m \mathbf{S} = [0, \dots, 1, \dots, 0]^T$. Through these derivations, we transform the Ambisonics encoding problem into a simple beamformer design problem. Moreover, this approach is not restricted by specific spherical harmonic orders. For irregular arrays such as SPMA, the complexity of the array manifold can potentially be exploited to achieve higher-order encoding with a limited number of beamforming weights.

As mentioned in Equation 3, which is precisely our beamforming target, we can directly obtain the HOA coefficients for plane waves from any incident direction through the beamforming filter:

$$\hat{\mathbf{B}}_n^m = \mathbf{p}\mathbf{h}^* \quad (11)$$

After determining the (n_s, m_s) -order beamformer, we project the input signal through the obtained beamformer to acquire the Ambisonics component of that order. This process needs to be repeated for each order of Ambisonics, and by combining the results from all orders, we complete the Ambisonics encoding of the entire sound field.

Nevertheless, by incorporating spatial aliasing considerations, we draw inspiration from time-domain alignment techniques in binaural rendering [15], extending this concept to

high-frequency array manifolds. Our approach performs time-domain alignment through replacing the original array manifold with its absolute value above the frequency threshold ω_{th} , which essentially preserves magnitude responses while suppressing erroneous inter-microphone phase differences. This strategy achieves dual benefits: 1) Focusing beamformer optimization on compensating amplitude distortions given the array's inherent performance limits; 2) Reducing the spherical harmonic order required for manifold representation, thereby mitigating spatial aliasing errors.

We hypothesize that the observed anti-aliasing improvement stems from exploiting the array's local density advantage - though traditional spherical harmonic decomposition fails to leverage this benefit due to numerical ill-conditioning in global surface fitting, the localized phase alignment in our method inherently utilizes the higher spatial sampling density within subarray regions. However, such performance gains are likely direction-dependent, as the effective local density varies with the direction of arrival (DOA). This directionality aspect warrants further investigation to quantify how specific DOA patterns interact with the array's geometric configuration.

For implementation, after determining ω_{th} based on array geometry, we replace high-frequency manifold components (above ω_{th}) with their absolute values. While this operation reduces simulated plane-wave manifolds to constant magnitudes (by eliminating pure phase variations), it effectively preserves real-world measured manifold characteristics where scattering-induced amplitude variations remain meaningful.

3. EXPERIMENTS

3.1. Configuration

In our application scenario, four microphones are asymmetrically positioned on a mobile phone: two at the bottom (one primary and one secondary microphone), one at the top, and one on the rear surface. Their specific coordinates and relative positions are illustrated in Fig. 1.

For filter design, as described in [4], we selected 50 source directions following a Gaussian distribution to recover HOA coefficients up to the 4th order ($K = 2(N + 1)^2$): elevation angles $\theta = \{25.02^\circ, 57.42^\circ, 90.00^\circ, 122.58^\circ, 154.98^\circ\}$, with azimuth angles ϕ sampled at 36° intervals.

3.2. Evaluation Metrics

3.2.1. Spatial Correlation & Intensity Difference

Following [16], we evaluate the spatial correlation R :

$$R = \frac{\langle \mathbf{B}, \hat{\mathbf{B}} \rangle}{\|\mathbf{B}\| \cdot \|\hat{\mathbf{B}}\|} \quad (12)$$

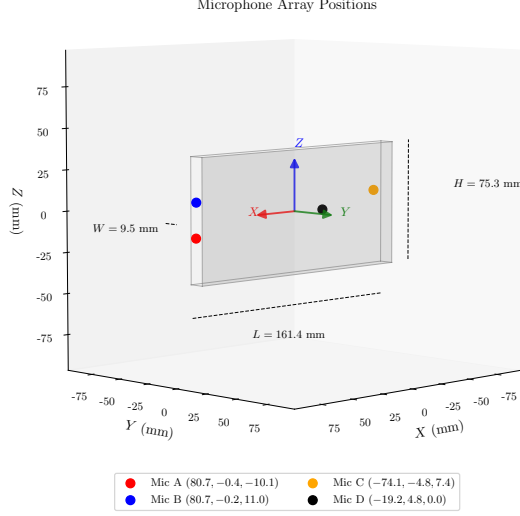


Fig. 1. Microphone Array Configurations of the SPMA utilized.

and the intensity difference L

$$L = 10 \log_{10} \left(\frac{\|\hat{\mathbf{B}}\|}{\|\mathbf{B}\|} \right) \quad (13)$$

where $\mathbf{B} = [B_0^0, B_1^{-1}, B_1^0, B_1^1, \dots, B_N^{-N}, \dots, B_N^N]^T$ is the coefficient vector, $\hat{\mathbf{B}}$ is the estimated HOA coefficients and $\mathbf{B}$ is the ideal HOA coefficients.

3.2.2. Signal-to-Distortion Ratio

The Signal-to-Distortion Ratio (SDR) evaluates the quality of HOA coefficient estimation[17]. For each spherical harmonic coefficient of order n and degree m , SDR is defined as

$$\text{SDR}(B_n^m) = 10 \log_{10} \frac{|B_n^m|^2}{|B_n^m - \hat{B}_n^m|^2} \quad (14)$$

where B_n^m represents the ideal HOA and $\hat{B}_n^m$ is its estimated value. This metric provides a coefficient-wise assessment of the estimation accuracy, complementing the global metrics of spatial correlation and intensity difference.

3.2.3. Pressure Reconstruction Error

Inspired by [17], the normalized mean square error (NMSE) between the truncated theoretical sound pressure p_{trunc} and the reconstructed approximated pressure $\hat{p}$ on a spherical surface of radius R is defined as

$$\varepsilon_{\text{error}}(k, R) := \frac{\|\hat{p} - p_{\text{trunc}}\|_S^2}{\|p_{\text{trunc}}\|_S^2}, \quad (15)$$

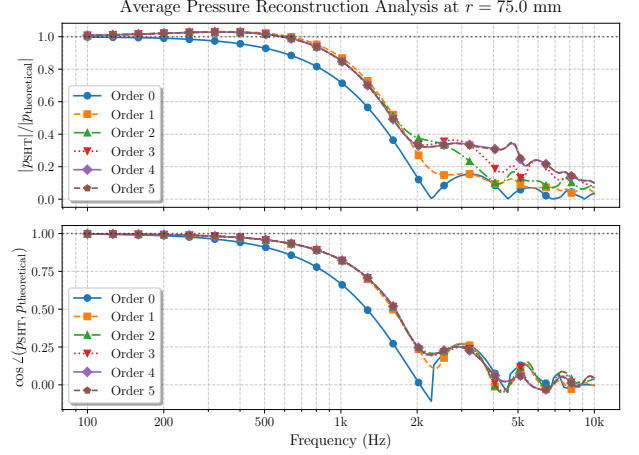


Fig. 2. Sound field reconstruction accuracy for different upscaling orders. Top: Pressure amplitude ratio $|p_{\text{new}}|/|p_{\text{theor.}}|$; Bottom: Pressure vector cosine similarity $\cos \angle(p_{\text{new}}, p_{\text{theor.}})$.

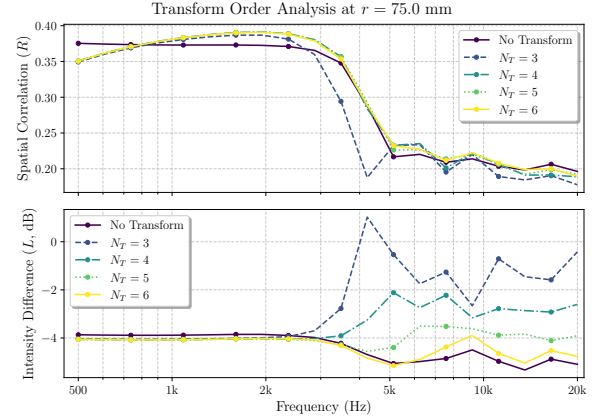


Fig. 3. Spatial correlation R and intensity difference L between HOA vector $\hat{\mathbf{B}}$ and $\mathbf{B}$ for different transform orders.

where $\|\cdot\|_S^2$ represents the squared L^2 -norm over the spherical surface S with $dS = R^2 \sin \theta d\theta d\phi$.

Leveraging spherical harmonic orthogonality, NMSE can be expressed as

$$\varepsilon_{\text{error}}(k, R) = \frac{\sum_{n=0}^4 \sum_{m=-n}^n |\hat{B}_n^m - B_n^m|^2 |j_n(kR)|^2}{\sum_{n=0}^4 \sum_{m=-n}^n |B_n^m|^2 |j_n(kR)|^2}. \quad (16)$$

3.3. Ablation Experiment

3.3.1. Order Upscaling

As described in Section 2.2, beamforming filter $\mathbf{h}_{n_s}^{m_s}(\omega)$ is designed for each spherical harmonic function separately. This architecture ensures mutual independence between filters across upscaling orders—even with limited number of microphones

or sparse source directions in the array manifold matrix. This leads us to investigate how our method affects sound field reconstruction as the maximum expansion order increases.

We employ the methodology outlined in Section 2.2 to an SPMA as depicted in Fig. 1. Different maximum orders of upscaling are set and the sound pressure is reconstructed on a $r = 0.075\text{m}$ sphere to judge the performance. Positions of reconstructed sound pressure θ_{Rec}, ϕ_{Rec} and the directions of arriving plane wave θ_k, ϕ_k are both obtained through Fibonacci sphere uniform sampling[18] of one thousand points. We average the corresponding metrics to produce the figures below. Evaluation metrics include the ratio of the reconstructed sound pressure amplitude to the ideal sound pressure amplitude $\langle |p_{new}|/|p_{theoretical}| \rangle$ and the cosine angle between the reconstructed sound pressure vector and the ideal sound pressure vector $\langle \cos \angle(p_{new}, p_{theoretical}) \rangle$.

Fig. 2 shows notable improvement through upscaling to order 4. Further upscaling yields negligible improvements. Therefore, we will focus on order 4 as an example for the analysis of HOA coefficient estimation.

3.3.2. DSHT Order of Array Manifold Matrix

Next, we need to determine the order of DSHT utilized in Section 2.2, incorporating the array manifold matrix into the SHD. Though mathematically the order of DSHT is only limited by the spatial sampling scheme [4], which in this section corresponds to the number of steering vectors simulated or measured. In real applications, appropriate truncation helps prevent spatial aliasing in steering vectors, thus improving the following Ambisonics encoding process. As depicted in Fig. 3, we compare the spatial correlation and intensity difference for different DSHT orders $N_T = 3, 4, 5, 6$ and determine that an appropriate DSHT order is $N = 4$, as both R and L metrics show no notable improvement with further increasing N .

3.4. Simple Condition

3.4.1. Simulation Experiments

In this section, we conduct a detailed comparison between the baseline method in Section 2.1 and our method across different metrics. We simulate the theoretical sound pressure at each microphone for a total of K plane wave sound sources with different directions of arrival θ_k, ϕ_k , resulting in the steering vector given by

$$\mathbf{d}(\omega, \theta_k, \phi_k) = [p_1(\theta_k, \phi_k), \dots, p_Q(\theta_k, \phi_k)]^T. \quad (17)$$

Through substituting $\mathbf{d}(\omega, \theta_k, \phi_k)$ into $\mathbf{D}$ and solving Equation 10 for each spherical harmonic function, the beamforming weight $\mathbf{h}_n^m(\omega)$ is obtained and Equation 11 can be used for Ambisonics encoding.

As illustrated in Fig. 4, both the baseline and our method exhibit similar estimates of the vector angles in the overall

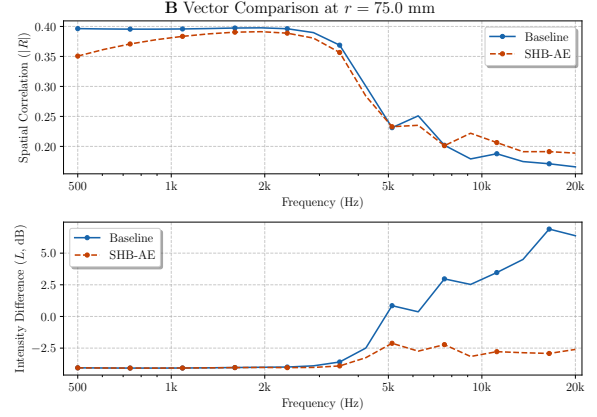


Fig. 4. $\hat{\mathbf{B}}$ vector estimation comparison between baseline method and our method.

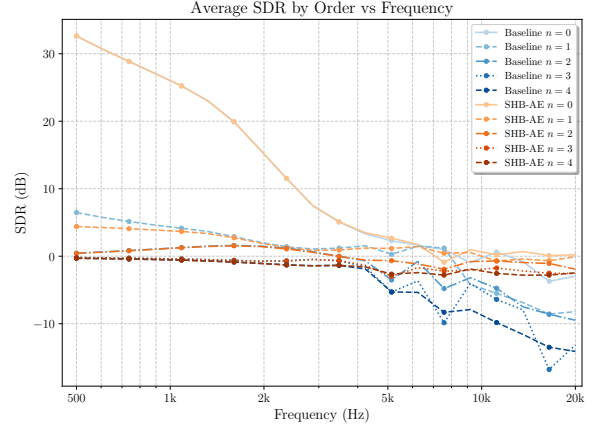


Fig. 5. Average SDR by order(n).

HOA estimation. However, the baseline begins to exhibit significant amplitude deviations in the mid-to-high frequency range. It is also reflected in Fig. 5, where the SDR metric of the baseline method is not only significantly lower than that of our approach in terms of overall mean but also displays a more pronounced downward trend in the mid to high frequencies, indicating its divergence from the theoretical HOA coefficients. We believe that the reason for the phenomenon lies on the aliasing frequency of the SPMA, indicating the difficulty of encoding high frequency HOA. By calculating the half-wavelength frequencies corresponding to the distances between microphones in our setup, we find that four out of the six distances correspond to frequencies below 2 kHz, resulting in significant spatial aliasing in the high-frequency range. In contrast, our method focuses on the number of sound source directions (steering vectors) rather than the number of microphones, making our system of linear equations overdetermined. Furthermore, as the equations are transformed into the spherical harmonic domain (SHD), our method exhibits improved robustness.

We can also judge the estimation of the HOA coefficients

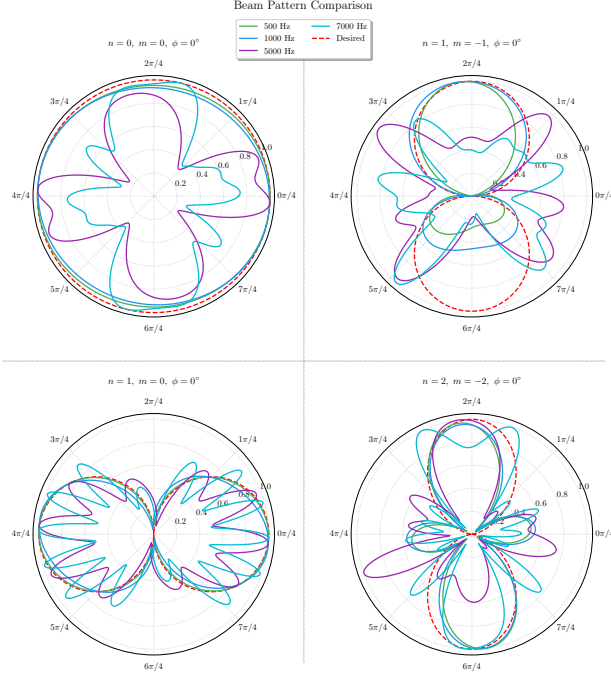


Fig. 6. Beam pattern examples. n, m denotes the order and degree of the harmonic function, while ϕ denotes the azimuth angle.

and beamforming in a more intuitive way. Fig. 6 presents beam patterns for different spherical harmonic orders at various frequencies within the same plane, illustrating that our method effectively achieves the task of mitigating the spherical harmonics through beamforming. Furthermore, the beam patterns at low frequencies are generally smoother, while rougher at higher frequencies, because of the constraints corresponding to the fixed source direction becoming more pronounced, indicating a weakening of the smoothing effect of spherical harmonic transformations on beam formation. This phenomenon can also be attributed to high-frequency aliasing caused by insufficient element spacing. Additionally, it is a plausible theoretical explanation that the low-order spherical harmonic basis functions exhibit slow spatial variation, rendering them incapable of adequately expressing the details of high-frequency acoustic fields.

It is important to note that both the baseline method and our approach have limited frequency-dependent characteristics in their least squares solutions, which cannot be simply attributed to the high condition number at high frequencies. In fact, the condition numbers of the two linear systems fluctuate with frequency and do not exhibit a clear monotonicity.

3.4.2. Practical Experiments

We also conduct experiments using data collected from real microphones. In the anechoic chamber depicted in Fig. 7, impulse responses of the SPMA are measured through expo-



Fig. 7. The acoustic system utilized for measuring the impulse responses of the SPMA

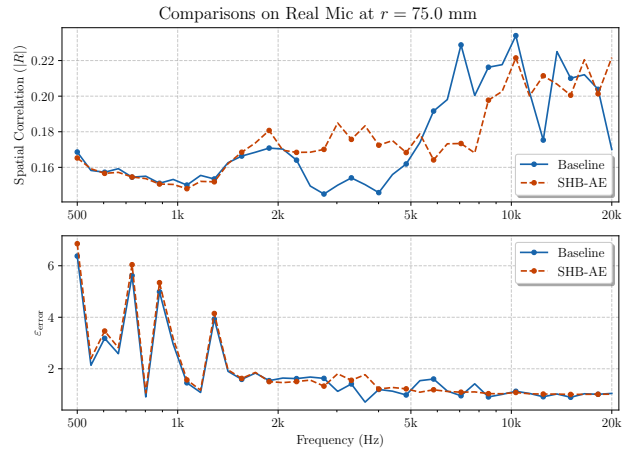


Fig. 8. R and ϵ_{error} evaluation on real microphone arrays.

nential sine sweep signals exciting a loudspeaker [19].

We obtain the sound pressure vector through convolution of the impulse responses with sound source signals. In simulation, the baseline construction of the encoding matrix relies solely on theoretical calculations from Equation 6, which fails to utilize the manifold information from the real SPMA, thus preventing effective adjustment of actual sound pressure encoding.

Therefore, in our actual experiments, the baseline method is considered to further separate the radial and angular components in the encoding matrix A , which is essentially equivalent to the beamforming assumption we introduced. Our method performs DSHT and frequency division operations based on this assumption, and our actual experiments focus on this comparison. Figures 8 and 9 demonstrate the performance metrics in our physical experiments. We set $\omega_{th} = 2\text{kHz}$ by experiments. Due to the influence of various external factors on the actual collected data, we performed a uniform energy normalization when calculating the HOA coefficients. This normalization renders the L metric ineffective, but does not affect the evaluation results of other metrics.

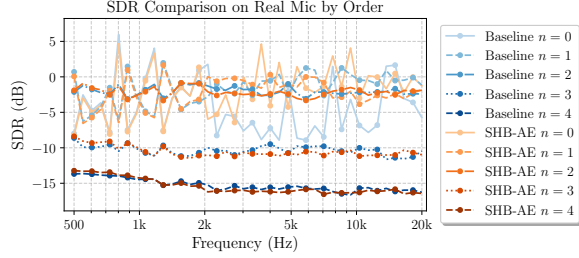


Fig. 9. SDR evaluation on real microphone arrays.

Table 1. Simulation results under noise with different SNR(dB).

SNR	Method	$\varepsilon_{\text{error}} \downarrow$	$R \uparrow$	$L \rightarrow 0$	SDR $\uparrow$
30	Baseline	1.90	0.32	-1.38	-1.74
	SHB-AE	0.70	0.30	-3.43	0.02
20	Baseline	1.98	0.29	-0.76	-2.29
	SHB-AE	0.75	0.28	-2.82	-0.46
10	Baseline	2.79	0.25	0.71	-3.45
	SHB-AE	1.29	0.25	-1.36	-1.52
0	Baseline	10.90	0.20	3.50	-5.97
	SHB-AE	6.69	0.20	1.42	-3.59

The experimental results shows that our method achieves significant improvements in spatial correlation within the 2-5 kHz range, accompanied by a notable reduction in reconstruction error. In the high-frequency band, the frequency-divided curves exhibit relatively stable behavior, while the baseline method shows substantial fluctuations. These findings confirm that our frequency division strategy above 2 kHz achieves its intended purpose—maintaining frequency consistency is more effective when withholding phase information rather than providing incorrect phase information. SHT also contributes to this high-frequency stability characteristic, which is well reflected in the SDR metrics.

3.5. Complex Conditions

Furthermore, we conduct additional simulation and practical experiments in different noise levels and reverberation time scenarios to judge the robustness of methods. Consistent with the previous sections, in the simulation experiments, we compared the results of the baseline with those obtained through beamforming and spherical harmonic transformation. In the practical experiments, we compared the outcomes of beamforming alone with those of spherical harmonic transformation coupled with frequency division.

3.5.1. Noise

To test the robustness of methods against additive, spatially white Gaussian noise under different Signal-to-Noise Ratio (SNR) is utilized. The experimental results demonstrate that the proposed method exhibits strong robustness across various

Table 2. Practical experimental results under noise with different SNR(dB).

SNR	Method	$\varepsilon_{\text{error}} \downarrow$	$R \uparrow$	SDR $\uparrow$
30.0	Baseline	2.02	0.18	-9.53
	SHB-AE	2.16	0.18	-9.54
20.0	Baseline	1.86	0.18	-9.49
	SHB-AE	1.99	0.18	-9.52
10.0	Baseline	1.66	0.18	-9.39
	SHB-AE	1.76	0.19	-9.42
0.0	Baseline	1.63	0.19	-9.37
	SHB-AE	1.71	0.18	-9.36

Table 3. Simulation results under reverberation with different RT60(s).

RT60	Method	$\varepsilon_{\text{error}} \downarrow$	$R \uparrow$	$L \rightarrow 0$	SDR $\uparrow$
0.2	Baseline	3.36	0.21	-10.29	-0.50
	SHB-AE	1.89	0.21	-12.30	-0.29
0.5	Baseline	156.53	0.21	-1.65	-2.78
	SHB-AE	48.39	0.22	-3.75	-1.27
1.0	Baseline	551.44	0.22	0.91	-4.61
	SHB-AE	163.68	0.22	-1.20	-2.26
2.0	Baseline	1020.52	0.22	2.16	-5.75
	SHB-AE	300.26	0.22	0.04	-2.99

noise levels. In simulation experiments, as SNR decreases from 30 dB to 0 dB, the proposed method consistently outperforms the baseline method in terms of $\varepsilon_{\text{error}}$, particularly under low SNR conditions (0 dB), where the proposed method's error (6.69) is significantly lower than that of the baseline method (10.90), indicating its excellent noise resistance capability. Regarding the correlation coefficient R , both methods show similar performance that R declines as the SNR decreases as expected. In terms of the SDR, the proposed method outperforms the baseline across all SNR levels, indicating superior signal fidelity.

In practical experiments, the performance difference between the proposed method and traditional beamforming techniques is not as obvious, suggesting that the introduction of SHT and frequency division only leads to limited advantages. In fact, results of different frequencies fluctuate significantly for both methods but the reason remains unknown. The comparable measurement metrics indicate that both methods exhibit similar noise resistance performance in practical applications.

3.5.2. Reverberation

In the experiments conducted under reverberation conditions, a typical 3D shoebox room with different Reverberation Time 60 (RT60) is utilized. After simulating the room impulse responses with pyroomacoustics [20], convolution between them and simulated or measured steering vectors is done for mitigating the room reverberation nature.

The simulation results demonstrate that system performance degrades with increasing RT60, primarily manifested

Table 4. Practical experimental results under reverberation with different RT60(s).

RT60	Method	$\varepsilon_{\text{error}} \downarrow$	$R \uparrow$	SDR $\uparrow$
0.2	Baseline	2.39	0.20	-9.91
	SHB-AE	1.75	0.19	-9.40
0.5	Baseline	2.00	0.19	-9.83
	SHB-AE	1.79	0.18	-9.45
1.0	Baseline	1.89	0.19	-9.81
	SHB-AE	1.83	0.18	-9.46
2.0	Baseline	1.86	0.19	-9.81
	SHB-AE	1.85	0.18	-9.46

in the significant rise of the error metric $\varepsilon_{\text{error}}$. However, compared to the baseline method, the proposed approach exhibits superior robustness against reverberation. Under mild reverberation conditions (RT60=0.2s), the proposed method reduces the error from 3.36 to 1.89 and in highly reverberant environments (RT60=2.0s), the proposed method maintains the error at 300.26 compared to 1020.52 of the baseline method, showcasing substantial performance advantages.

In practical experiments, although the overall error levels are lower than simulation results due to normalization, our method consistently maintains its performance advantages. Specifically, across all RT60 conditions, our method outperforms the beamforming approach in both $\varepsilon_{\text{error}}$ and SDR metrics. Notably, the impact of RT60 variations on system performance in real environments is considerably less pronounced than observed in simulations, which may be attributed to the more complex acoustic characteristics in practical settings. Additionally, the R -value remains relatively stable across different conditions, indicating its insensitivity to reverberant conditions.

4. SUMMARY

For the task of Ambisonics encoding and upsampling on SPMA, we propose SHB-AE, a spherical harmonic beamforming-based method which designs beamformers for each spherical harmonic function based on the array manifold and introducing DSHT together with frequency division operations. Experimental validation was conducted on both simulated and real SPMA across varying reverberant and noisy environments. Results demonstrated that our method achieved superior performance in various metrics and successfully encodes and up-scales HOA. Its robustness in high-frequency ranges and complex conditions is also obvious. This approach offers a practical solution for spatial audio recording using portable devices, potentially enabling wider adoption of immersive audio capture in AR/VR applications.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No.2024YFB2808902), and the High-performance Computing Platform of Peking University.

References

- [1] Franz Zotter and Matthias Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*, Springer Nature, 2019.
- [2] Tetsu Magariyachi and Yuki Mitsufuji, “Analytic error control methods for efficient rotation in dynamic binaural rendering of ambisonics,” *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 218–230, Jan 2020.
- [3] Dumidu S. Talagala, Wen Zhang, and Thushara D. Abhayapala, “Broadband doa estimation using sensor arrays on complex-shaped rigid bodies,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1573–1585, Aug 2013.
- [4] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.
- [5] Mikko Heikkinen, Archontis Politis, and Tuomas Virtanen, “Neural ambisonics encoding for compact irregular microphone arrays,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 701–705.
- [6] Yue Qiao, Vinay Kothapally, Meng Yu, and Dong Yu, “Neural ambisonic encoding for multi-speaker scenarios using a circular microphone array,” *arXiv preprint arXiv:2409.06954*, 2024.
- [7] Leo McCormack, Archontis Politis, Raimundo Gonzalez, Tapio Lokki, and Ville Pulkki, “Parametric ambisonic encoding of arbitrary microphone arrays,” *CERN European Organization for Nuclear Research - Zenodo*, CERN European Organization for Nuclear Research - Zenodo, Mar 2022.
- [8] Amy Bastine, Lachlan Birnie, Thushara D. Abhayapala, Prasanga Samarasinghe, and Vladimir Tourbabin, “Ambisonics capture using microphones on head-worn device of arbitrary geometry,” in *2022 30th European Signal Processing Conference (EUSIPCO)*, Aug 2022, p. 309–313.
- [9] Yhonatan Gayer, Vladimir Tourbabin, Zamir Ben-Hur, Jacob Donley, and Boaz Rafaely, “Ambisonics encoding for arbitrary microphone arrays incorporating residual channels for binaural reproduction,” *arXiv preprint arXiv:2402.17362*, 2024.

- [10] Donghang Wu, Xihong Wu, and Tianshu Qu, “Leveraging moving sound source trajectories for universal sound separation,” *arXiv preprint arXiv:2409.04843*, 2024.
- [11] Donghang Wu, Xihong Wu, and Tianshu Qu, “Exploiting motion information in sound source localization and tracking,” *Journal of the Audio Engineering Society*, , no. 10687, january 2024.
- [12] Amy Bastine, Lachlan Birnie, Thushara D Abhayapala, Prasanga Samarasinghe, and Vladimir Tourbabin, “Ambisonics capture using microphones on head-worn device of arbitrary geometry,” in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 309–313.
- [13] Gene H. Golub, Per Christian Hansen, and Dianne P. O’Leary, “Tikhonov regularization and total least squares,” *SIAM Journal on Matrix Analysis and Applications*, p. 185–194, Jan 1999.
- [14] Shan Gao, Xihong Wu, and Tianshu Qu, “High order ambisonics encoding method using differential microphone array,” in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.
- [15] Michael J Evans, James AS Angus, and Anthony I Tew, “Analyzing head-related transfer function measurements using surface spherical harmonics,” *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2400–2411, 1998.
- [16] Stéphanie Bertet, Jérôme Daniel, and Sébastien Moreau, “3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone,” in *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.
- [17] Lingkun Zhang, Xiaochen Wang, Ruimin Hu, Dengshi Li, and Weipin Tu, “Estimation of spherical harmonic coefficients in sound field recording using feed-forward neural networks,” *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 1–16, 2021.
- [18] A Gonzalez, “Measurement of areas on a sphere using fibonacci and latitude-longitude lattices,” *Mathematical geosciences*, , no. 1, pp. 42, 2010.
- [19] Angelo Farina, “Advancements in impulse response measurements by sine sweeps,” in *Audio engineering society convention 122*. Audio Engineering Society, 2007.
- [20] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, “Py-roomacoustics: A python package for audio room simulation and array processing algorithms,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.