Spherical Harmonic Beamforming based Ambisonics Encoding Method in Frequency and Time Domain*

Yuhuan You,[1] Yufan Qian,[1] Tianshu Qu,[1] *AES Member,* **Bin Wang,[2] Xueyang Lv,[3]**

[1]*State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing, China*
[2]*Beijing Xiaomi Mobile Software Co., Ltd, Beijing, China*
[3]*Xiaomi Communications Co., Ltd, Beijing, China*

Implementing Higher-Order Ambisonics (HOA) on consumer devices is hindered by their sparse, irregular microphone arrays, which challenge conventional methods with issues like spatial aliasing and ill-conditioning. Based on the established spherical harmonic beamforming framework, this paper proposes two robust encoding approaches: a frequency-domain (FD) method with compensation for high-frequency artifacts, and a time-domain (TD) method that holistically optimizes broadband FIR filters for enhanced stability. The framework is inherently scalable, allowing on-demand order expansion. Using a measured smartphone array, comprehensive objective and subjective evaluations demonstrate the superiority of the TD method. It excels in signal fidelity, spatial accuracy, and temporal consistency, outperforming baseline and FD approaches. The TD method also maintains its advantage in adverse conditions, showing remarkable robustness against noise and multi-source environments. It provides a practical, high-performance pathway for enabling high-fidelity spatial audio capture on ubiquitous consumer devices.

0 INTRODUCTION

The rapid proliferation of virtual and augmented reality technologies has created a substantial demand for immersive audio experiences, making spatial audio capture and reproduction a pivotal area of research. Ambisonics has emerged as a de facto standard for spatial audio representation, lauded for its inherent flexibility and theoretical elegance [1]. As a scene-based format, Ambisonics signals are independent of any specific playback configuration, enabling them to be decoded for diverse loudspeaker arrays or rendered binaurally for headphones. Furthermore, crucial user interactions such as head-tracking can be implemented with high computational efficiency through direct algebraic manipulations of the spherical harmonic coefficients [2].

The canonical theory of Ambisonics encoding, however, presumes the use of rigid spherical microphone arrays (SMAs) populated with a large number of sensors, which are meticulously arranged to sample the sound field [3]. This requirement poses a significant barrier to implementation on ubiquitous consumer devices like smartphones, smart glasses, and other wearables, which are not purpose-built for spatial audio acquisition. Such devices are typically equipped with sparse, irregularly configured microphone arrays and their form factors introduce complex acoustic scattering and diffraction, all of which confound conventional encoding algorithms.

To bridge this gap, a rich body of research has emerged, proposing various methods to achieve Ambisonics encoding with these non-ideal arrays [1, 3]. The existing approaches can be fundamentally distinguished by a core operational principle: whether the encoding process is signal-independent, utilizing a set of fixed filters, or signal-dependent, dynamically adapting to the characteristics of the incoming sound field.

Signal-independent methods aim to design a single, time-invariant set of encoding filters whose properties are determined solely by the measured or simulated acoustic transfer functions of the array. The dominant technique in this category is mode-matching [3], which formulates a linear system of equations relating the microphone signals to the target Ambisonics coefficients and solves for an optimal encoding matrix, typically in a least-squares sense.

*Corresponding author: Tianshu Qu (email: qutianshu@pku.edu.cn)

This foundational principle has been successfully applied to arrays on complex structures, such as head-worn devices [4]. To enhance numerical stability and perceptual outcomes, these methods often incorporate regularization strategies, such as Tikhonov regularization [5], in their problem formulation [6]. Some approaches have also explicitly leveraged beamforming techniques to design the encoding filters for specific array geometries [7]. Despite their computational efficiency, the performance of these methods is constrained by the physical limitations of the array [3]. They are often challenged by the ill-conditioning of the system matrix, especially when the number of microphones Q is insufficient for the desired Ambisonics order N (i.e., when $Q < (N+1)^2$), and are notoriously susceptible to performance degradation at high frequencies due to spatial aliasing.

In contrast, signal-dependent approaches adapt their processing based on the content of the captured sound field. One major sub-class of this paradigm involves parametric synthesis. These methods first analyze the acoustic scene to estimate key parameters, such as the directions-of-arrival (DOA) of dominant sound sources, and then use this parametric representation to synthesize the target Ambisonics signal. For example, some techniques employ high-resolution DOA estimators like MUSIC to localize sources, which are then extracted via beamforming and encoded, while others may model the field as a composite of directional and ambient components [8]. The primary drawback of this approach is that its final accuracy is fundamentally bottlenecked by the performance of the parameter estimation stage [9]. Indeed, robust multi-source DOA estimation and separation with sparse, irregular arrays remains a challenging open problem, with many advanced separation techniques assuming high-quality Higher-Order Ambisonics (HOA) signals as their input. Another prominent subclass is end-to-end neural mapping. Leveraging the power of deep learning, these methods learn a direct, non-linear transformation from the raw microphone signals to the Ambisonics coefficients. Architectures based on U-Nets, convolutional recurrent neural networks, and other complex models [10] have been proposed. These data-driven methods can implicitly model and compensate for intricate acoustic effects but are highly dependent on the availability of large, diverse training datasets and may face challenges with generalization to novel acoustic environments [10].

This paper revisits the signal-independent paradigm, which remains attractive for its low computational complexity and its independence from explicit sound field assumptions. Adopting the established perspective[11] that linear Ambisonics encoding is fundamentally a spatial filtering or beamforming problem, our goal is to directly address the robustness and performance limitations that have historically constrained this class of methods. This perspective allows us to leverage the mature toolkit of beamformer design to create robust solutions for irregular arrays. The main contributions of this work are the development of two specific realizations within this framework: a frequency-domain (FD) method featuring a novel high-frequency compensation strategy to mitigate aliasing, and a holistically optimized time-domain (TD) method that yields stable, broadband performance. Our proposed methods provide a practical and robust pathway to high-quality HOA encoding on irregular arrays without the need for sound field analysis.

The remainder of this paper is organized as follows. Section 1 introduces the fundamental knowledge of Ambisonics encoding and the proposed method. Section 2 presents experimental validation of the method's effectiveness on measured array manifolds of a smartphone microphone array (SPMA) under multiple conditions. Section 3 concludes the paper.

1 METHODS

1.1 Ambisonics Encoding

Ambisonics is a spatial audio technology that employs spherical harmonic decomposition to represent 3D sound fields mathematically. The pressure field solution for homogeneous media in spherical coordinates is expressed as [1, 3]:

$$p(\mathbf{r}) = \sum_{n=0}^{\infty} (2n+1) b_n(kr) \sum_{m=0}^{n} \sum_{\sigma=\pm 1} B_{nm}^{\sigma} Y_{nm}^{\sigma}(\theta_q, \phi_q) \quad (1)$$

where $b_n(kr) = i^n j_n(kr)$, $j_n(kr)$ is the spherical Bessel function, and Y_{nm}^{σ} denotes Daniel's real spherical harmonics [1]:

$$Y_{nm}^{\sigma}(\theta,\phi) = \begin{cases} \sqrt{\varepsilon_m \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)\cos(m\phi), & \sigma = +1 \\ \sqrt{\varepsilon_m \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)\sin(m\phi), & \sigma = -1 \end{cases} \quad (2)$$

with $\varepsilon_m = 1$ (when $m = 0$) or 2 (when $m > 0$), and P_n^m being normalized associated Legendre polynomials. These functions form a complete orthonormal basis on the unit sphere.

For a plane wave incident from direction (θ_k, ϕ_k), its spherical harmonic expansion at position (θ_q, ϕ_q) becomes [12, 3]:

$$\begin{aligned} p(\mathbf{r}) &= s \cdot e^{i\mathbf{k}\cdot\mathbf{r}} \\ &= s \cdot \sum_{n=0}^{\infty} (2n+1) b_n(kr) \sum_{m=0}^{n} \sum_{\sigma=\pm 1} Y_{nm}^{\sigma}(\theta_k, \phi_k) Y_{nm}^{\sigma}(\theta_q, \phi_q) \end{aligned} \quad (3)$$

This expansion forms the theoretical foundation of Ambisonics encoding, requiring only the projection coefficients $Y_{nm}^{\sigma}(\theta_k, \phi_k)$ to fully characterize the spatial properties of sound fields. Ambisonics encoding therefore extracts these coefficients [1]:

$$\begin{cases} B_{nm}^{\sigma}(\omega) = S(\omega) \cdot Y_{nm}^{\sigma}(\theta_k, \phi_k) & \text{(Frequency domain)} \\ b_{nm}^{\sigma}(t) = s(t) \cdot Y_{nm}^{\sigma}(\theta_k, \phi_k) & \text{(Time domain)} \end{cases} \quad (4)$$

By decomposing the acoustic field into a set of orthogonal spherical harmonic basis functions, it achieves comprehensive characterization of sounds from various direc-

tions, thereby enabling accurate reproduction of immersive acoustic experiences in diverse listening environments.

1.2 Spherical Harmonic Beamforming Based Ambisonics Encoding

The theoretical background presented above provides an elegant mathematical foundation for Ambisonics encoding through spherical harmonic decomposition. However, practical implementation faces the challenge of extracting these coefficients B_{nm}^{σ} from real acoustic environments using finite microphone arrays. This section reveals that this extraction process is fundamentally equivalent to spatial beamforming.

Consider a spherical array of Q microphones at positions (r_q, θ_q, ϕ_q). Intuitively, each spherical harmonic channel (n, m, σ) can be viewed as a directional reception pattern with gain $Y_{nm}^{\sigma}(\theta_k, \phi_k)$ across different directions (θ_k, ϕ_k). To extract the coefficient for an incident plane wave from direction (θ_k, ϕ_k), we require the filtered output to satisfy:

$$\hat{B}_{nm}^{\sigma}(t) = s(t) \cdot Y_{nm}^{\sigma}(\theta_k, \phi_k) \tag{5}$$

Through linear filtering of signals $x_q(t)$ with impulse responses $h_{nm,q}^{\sigma}(t)$ from each microphone q, the output becomes:

$$\hat{B}_{nm}^{\sigma}(t) = \sum_{q=1}^{Q} (h_{nm,q}^{\sigma} * x_q)(t) \tag{6}$$

In the frequency domain, this constraint translates to:

$$\sum_{q=1}^{Q} H_{nm,q}^{\sigma}(\omega) \cdot A_q(\Omega, \omega) = Y_{nm}^{\sigma}(\Omega) \tag{7}$$

where $A_q(\Omega, \omega)$ is microphone q's response to a unit plane wave from direction Ω. This is precisely the classical beamforming equation with target pattern $Y_{nm}^{\sigma}(\Omega)$ and weights $H_{nm,q}^{\sigma}(\omega)$ [3]. Each spherical harmonic channel thus corresponds to a beamformer spatially matched to the corresponding spherical harmonic function.

To handle the general case with K arriving plane waves, we define the array manifold matrix $\mathbf{D}(\omega) \in \mathbb{C}^{Q \times K}$, whose elements are the individual microphone responses, $[\mathbf{D}(\omega)]_{qk} = A_q(\Omega_k, \omega)$. The filter weights for all Q microphones can be stacked into a column vector $\mathbf{h}_{nm}^{\sigma}(\omega) = [H_{nm,1}^{\sigma}(\omega), \ldots, H_{nm,Q}^{\sigma}(\omega)]^T$. The optimal beamforming weights that solve the problem in a regularized least-squares sense are given by [3]:

$$\mathbf{h}_{nm}^{\sigma}(\omega) = [\mathbf{D}(\omega)\mathbf{D}(\omega)^H + \lambda \mathbf{I}]^{-1} \mathbf{D}(\omega) \mathbf{y}_{nm}^{\sigma} \tag{8}$$

where $\mathbf{y}_{nm}^{\sigma}$ is a column vector containing the target spherical harmonic values $Y_{nm}^{\sigma}(\Omega_k)$ for all K directions. The encoding process is thus reframed as designing matched spatial filters—beamformers—that approximate spherical harmonic directivity patterns. This formulation unifies traditional encoding and modern array signal processing under a common beamforming framework.

1.3 Frequency Domain Method

Building on the general beamforming solution in Eq. (8), we first describe the Ambisonics Signal-Matching (ASM) baseline [6]. The ASM method implements Eq. (8) by constructing the target vector $\mathbf{y}_{nm}^{\sigma}$ through direct calculation, i.e., evaluating $Y_{nm}^{\sigma}(\Omega_k)$ for each of the K sampling directions. This formulation effectively minimizes the residual error energy in the discrete direction-sample domain ($||\mathbf{e}||_2^2$), penalizing deviations at every sampling point equally regardless of the spatial frequency content. However, using the full complex manifold $\mathbf{D}(\omega)$ across the entire spectrum in this domain can lead to overfitting of grid-specific irregularities and instability at high frequencies. Our proposed Frequency Domain (FD) method [13] improves upon this baseline with two structural modifications.

First, we reformulate the optimization target using the Discrete Spherical Harmonic Transform (DSHT). Unlike the baseline's point-wise matching, we define the target in the spherical harmonic basis as the standard basis vector $\mathbf{e}_l$, and use the DSHT analysis operator $\mathbf{S}$ to project the problem into the truncated modal domain [3]:

$$\mathbf{y}_{nm}^{\sigma} = \mathbf{S}^T \mathbf{e}_l \tag{9}$$

where $\mathbf{e}_l$ is a column vector with a single 1 at the l-th position ($l = n^2 + n + m + 1$). This modification fundamentally alters the cost function to minimize error in the truncated modal domain ($||\mathbf{S}\mathbf{e}||_2^2$). Since the number of sampling directions K typically exceeds the number of spherical harmonic modes L (i.e., $K > L$), $\mathbf{S}$ acts as a dimensionality reduction operator. This effectively filters out spatial aliasing components and grid-specific irregularities that fall into the non-trivial null space of $\mathbf{S}$, providing inherent regularization before any frequency-dependent processing.

Specifically, this formulates the filter design as a regularized least-squares optimization problem. The cost function $J(\mathbf{h}_{nm}^{\sigma})$ and the resulting optimal filter weights are explicitly given by:

$$\begin{aligned} J(\mathbf{h}_{nm}^{\sigma}) &= \left\| \mathbf{D}(\omega)^H \mathbf{h}_{nm}^{\sigma} - \mathbf{S}^T \mathbf{e}_l \right\|_2^2 + \lambda \|\mathbf{h}_{nm}^{\sigma}\|_2^2 \\ \mathbf{h}_{nm}^{\sigma}(\omega) &= \left[\mathbf{D}(\omega)\mathbf{D}(\omega)^H + \lambda \mathbf{I} \right]^{-1} \mathbf{D}(\omega)(\mathbf{S}^T \mathbf{e}_l) \end{aligned} \tag{10}$$

This solution explicitly aligns the beamforming output with the target spherical harmonic mode within the subspace defined by the discrete transform.

Second, we introduce a high-frequency compensation strategy. Above a threshold ω_{th}, we define a high-frequency matrix $\mathbf{D}_{HF}(\omega)$ by taking the element-wise magnitude of the manifold:

$$[\mathbf{D}_{HF}(\omega)]_{qk} = |[\mathbf{D}(\omega)]_{qk}| \tag{11}$$

This strategy is strictly motivated by the physical limitations of the array in the spatial aliasing regime. The half-wavelength frequencies ($f_{1/2} = c/2d$) for the primary microphone pairs in our sparse array fall mostly within the range of 1 to 2 kHz. Above this range, the phase relationships become physically ambiguous due to spatial aliasing, making the least-squares solution highly sensitive to position errors [14, 15]. However, despite the phase ambiguity, the magnitude response remains physically meaningful. Due to the rigid scattering body of the device, high-frequency sound waves create significant acoustic shadow-

ing, ensuring that the magnitude response $[\mathbf{D}(\omega)]_{qk}$ preserves strong directional diversity. Therefore, modeling magnitude only at high frequencies allows the algorithm to exploit these robust spatial differences (Inter-channel Level Differences) instead of unreliable phase information, serving to regularize the ill-conditioned problem and ensuring numerical stability.

The complete FD filter, which substitutes Eq. (9) into the optimization framework and applies this frequency-dependent manifold, is defined as:

$$\mathbf{h}_{nm}^{\sigma}(\omega) = \begin{cases} \mathbf{h}_{nm,LF}^{\sigma}(\omega), & \omega \le \omega_{th} \\ \mathbf{h}_{nm,HF}^{\sigma}(\omega), & \omega > \omega_{th} \end{cases} \tag{12}$$

where $\mathbf{h}_{nm,LF}^{\sigma}(\omega)$ and $\mathbf{h}_{nm,HF}^{\sigma}(\omega)$ are both derived using the solution in Eq. (10) but use $\mathbf{D}(\omega)$ and $\mathbf{D}_{HF}(\omega)$ respectively.

Design Synergy Analysis. The proposed FD method integrates the sparse grid, DSHT projection, and magnitude-only processing into a coherent strategy to address spatial aliasing. First, replacing the complex manifold with magnitude values at high frequencies introduces a fundamental mechanism shift: beamforming no longer relies on phase interference (which fails in the aliasing regime) but transitions to shadowing-based spatial amplitude selection, utilizing the robust directional information preserved by the device's scattering body. Second, the sparse grid of $K = 50$ directions represents the critical sampling density for the target order ($K \approx 2(N+1)^2$). While this density is insufficient to capture rapid high-frequency phase variations (aliasing), it is sufficient to capture the smoother topology of the magnitude envelopes created by acoustic shadowing. Finally, the DSHT projection acts as a necessary subspace regularizer. It compensates for the non-uniformity of the sparse grid and prevents overfitting to grid-specific irregularities, ensuring that the beam patterns synthesized from the magnitude features remain smooth and continuous in the spherical harmonic domain.

1.4 Time Domain Method

While frequency-domain approaches optimize each frequency bin independently, the time-domain filtering method (TD) directly designs multichannel Finite Impulse Response (FIR) filter banks through holistic optimization across the entire bandwidth. This approach formulates Ambisonics encoding as a broadband beamforming problem solved via time-domain least squares.

Each encoding filter $h_{nm;q}^{\sigma}[k]$ with length L extracts spherical harmonic component (n,m,σ) from microphone q. Given training directions $\{\Omega_i\}_{i=1}^{K}$ and microphone impulse responses $d_{q,i}[n]$ (length L_{IR}) for direction i, we require the actual output to match the ideal response:

$$\sum_{q=1}^{Q} (h_{nm;q}^{\sigma} * d_{q,i})[n] = Y_{nm}^{\sigma}(\Omega_i)\delta[n-n_0] \tag{13}$$

for all directions $i = 1,\dots,K$, where n_0 is the reference alignment time.

To solve this system, we transform convolution constraints into linear algebraic form using Toeplitz matrices. For each microphone q and direction i, we construct the array manifold matrix $\mathbf{D}_{i,q} \in \mathbb{R}^{(L_{IR}+L-1)\times L}$ in a convolution form:

$$\mathbf{D}_{i,q} = \begin{pmatrix} d_{q,i}[0] & 0 & \cdots & 0 \\ d_{q,i}[1] & d_{q,i}[0] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ d_{q,i}[L_{IR}-1] & d_{q,i}[L_{IR}-2] & \cdots & d_{q,i}[L_{IR}-L] \\ 0 & d_{q,i}[L_{IR}-1] & \cdots & d_{q,i}[L_{IR}-L+1] \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{q,i}[L_{IR}-1] \end{pmatrix} \tag{14}$$

where $d_{q,i}[n] = 0$ outside $[0, L_{IR}-1]$. Horizontally concatenating matrices for all microphones gives:

$$\mathbf{D}_i = \begin{bmatrix} \mathbf{D}_{i,1} & \mathbf{D}_{i,2} & \cdots & \mathbf{D}_{i,Q} \end{bmatrix} \tag{15}$$

The filter coefficients for channel (n,m,σ) are concatenated as a single column vector:

$$\mathbf{h}_{nm}^{\sigma} = \begin{bmatrix} \mathbf{h}_{nm;1}^{\sigma} \\ \mathbf{h}_{nm;2}^{\sigma} \\ \vdots \\ \mathbf{h}_{nm;Q}^{\sigma} \end{bmatrix} \tag{16}$$

Vertically stacking constraints from all K directions as $\mathbf{G}, \mathbf{y}_{nm}^{\sigma}$ forms the global overdetermined system:

$$\underbrace{\begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_K \end{bmatrix}}_{\mathbf{G}} \mathbf{h}_{nm}^{\sigma} \approx \underbrace{\begin{bmatrix} \mathbf{y}_{nm;1}^{\sigma} \\ \mathbf{y}_{nm;2}^{\sigma} \\ \vdots \\ \mathbf{y}_{nm;K}^{\sigma} \end{bmatrix}}_{\mathbf{y}_{nm}^{\sigma}} \tag{17}$$

The optimal solution is obtained through least squares minimization:

$$\mathbf{h}_{nm}^{\sigma} = \mathbf{G}^{\dagger}\mathbf{y}_{nm}^{\sigma} \tag{18}$$

where $\mathbf{G}^{\dagger}$ is the Moore-Penrose pseudoinverse. This process repeats for each spherical harmonic channel to obtain the complete filter set.

The time-domain approach offers three key advantages over frequency-domain methods. First, holistic frequency optimization: by constraining filter length, it balances errors across all frequencies rather than independently matching each frequency bin, typically producing smoother frequency responses with shorter FIR filters. Second, adaptive frequency trade-off: when arrays exhibit unavoidable distortions (e.g., spatial aliasing), the least squares formulation automatically reduces weights at problematic frequencies to minimize overall error, sacrificing accuracy in poorly reconstructible bands for better performance in perceptually relevant ranges. Third, implementation efficiency: the method provides a practical balance between complexity and performance, enabling real-time spatial audio encoding through finite-length filters while maintaining reasonable approximation quality within targeted frequency bands.

1.5 Framework Realization and Order Extension

Despite their implementation differences, both the FD and TD approaches are realized as solutions to the same fundamental least-squares optimization problem [3]. This formulation leverages the established perspective of Ambisonics encoding as a spatial filtering problem, where each spherical harmonic channel (n,m,σ) implements a beamformer designed to match the target directivity pattern $Y_{nm}^{\sigma}(\Omega)$.

The primary divergence lies in the optimization domain, which fundamentally dictates performance. The FD method optimizes $\mathbf{H}(\omega)$ independently at each frequency bin. While this minimizes error at specific frequencies, it enforces no constraints on phase consistency across adjacent bins, leading to group delay discontinuities that manifest as temporal artifacts (pre-echo). In contrast, the TD method constrains $\mathbf{H}(\omega)$ to be the response of a single finite-length FIR filter. By performing a holistic broadband optimization, the TD method implicitly enforces cross-frequency coherence, preventing group delay chaos and ensuring a strictly causal, artifact-free temporal response.

A key advantage of this component-wise beamforming realization is its inherent modularity, allowing for flexible order extension. Because a distinct beamformer is designed independently for each spherical harmonic channel (n,m,σ), an existing N th-order encoding system can be augmented to a higher order $(N+1)$ by simply computing the new filters for the additional channels, without altering the previously designed lower-order filters. It is important to note that while the number of microphones Q limits the number of linearly independent channels (rank limitation), extending the encoding order serves as a spatial interpolation. This provides a more accurate approximation of the continuous sound pressure field compared to the spatial aliasing inherent in a truncated low-order representation, offering a practical pathway to higher-fidelity spatial imaging on sparse arrays.

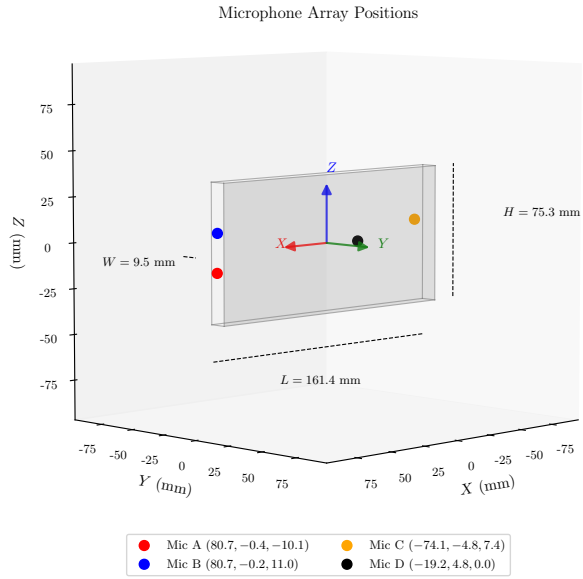

Fig. 1: Microphone Array Configurations of the SPMA utilized.

Fig. 2: The acoustic system utilized for measuring the impulse responses of the SPMA.

2 EXPERIMENTS

2.1 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 [3], we selected $K = 50$ source directions following a Gaussian distribution to recover HOA coefficients up to the 4th order, where K should be greater than $(N+1)^2$: 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 starting from 18°.

We conduct the experiments using data collected from real microphones. In the anechoic chamber depicted in Fig. 2, impulse responses of the SPMA are measured through exponential sine sweep signals exciting a loudspeaker [16]. We obtain the sound pressure vector through convolution of the impulse responses with sound source signals.

With the exception of the order extension analysis, which uses an ideal simulated manifold for theoretical evaluation, all experiments use data collected from real microphones to assess performance in a realistic setting.

For performance evaluation, we analyze the system impulse responses, which represent the end-to-end response of the designed filters to an ideal impulse signal from each test direction. For the time-domain method, the resulting system impulse response for a given direction k and Ambisonics channel (n,m,σ) is computed by applying the solved filter vector $\mathbf{h}_{nm}^{\sigma}$ to the array's measured impulse responses for that direction $\mathbf{D}_k$ as $\mathbf{z}_{nm,k}^{\sigma} = \mathbf{D}_k \mathbf{h}_{nm}^{\sigma}$ where $\mathbf{z}_{nm,k}^{\sigma}$ is the final system impulse response, $\mathbf{D}_k$ is the direction-specific convolution matrix, and $\mathbf{h}_{nm}^{\sigma}$ is the corresponding filter coefficient vector. For the frequency-domain method,

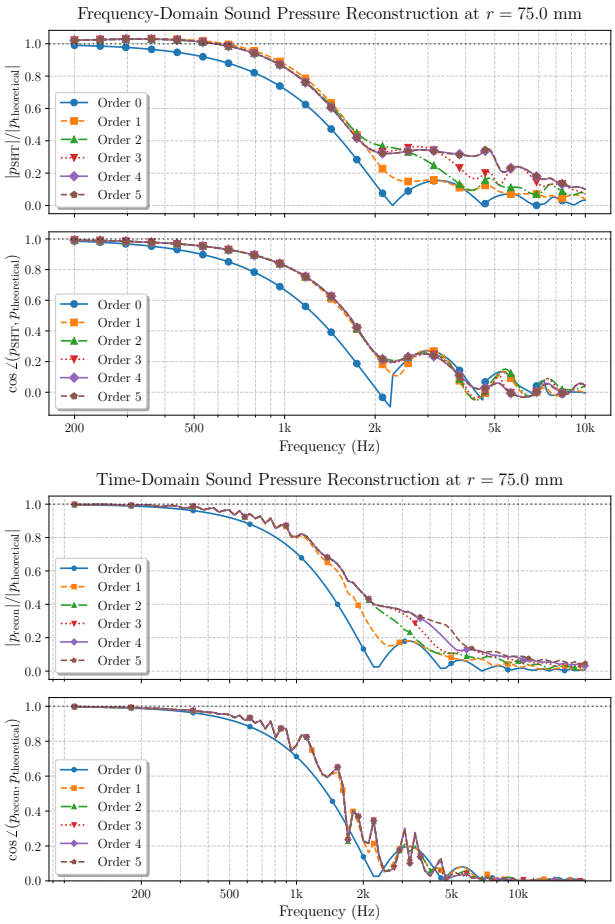

Fig. 3: Sound field reconstruction accuracy for different extended orders for the frequency-domain filters (top panel) and time-domain filters (bottom panel). Upper subplots: pressure amplitude ratio $|p_{\text{new}}|/|p_{\text{theor.}}|$. Lower subplots: pressure vector cosine similarity $\cos\angle(p_{\text{new}}, p_{\text{theor.}})$.

the system response is obtained by encoding the array impulse responses via Short-Time Fourier Transform (STFT), applying the frequency-domain filters, and reconstructing the time-domain signal via inverse STFT.

We compare three methods in our evaluation: the ASM baseline [6] (whose implementation is defined in Section 1.3), our proposed FD method (Section 1.3), and our proposed TD method (Section 1.4). To ensure a fair comparison, the same Tikhonov regularization strength of $\lambda = 10^{-7}$ was used for all frequency-domain filter designs.

2.2 Order Extension Analysis

To validate the efficacy of order extension, we investigate how the reconstruction accuracy evolves as the maximum expansion order increases. Although the system rank is physically limited by the number of microphones ($Q = 4$), targeting a higher order functions as a spatial interpolation strategy to minimize the approximation error of the continuous sound field.

We evaluate this performance by reconstructing the sound pressure on a sphere of radius $r = 0.075$m, using 1000 Fibonacci-sampled points for both the incident plane waves and the reconstruction locations [17]. For the TD method, a dynamic compensation scheme is applied to correct for direction-dependent group delay variations, a step not required by the FD method. We use two metrics averaged over all directions: the amplitude ratio $\langle|p_{\text{new}}|/|p_{\text{theoretical}}|\rangle$ and the cosine of the phase difference after compensation. As shown in Fig. 3, the results indicate notable improvement up to order 4. This confirms that extending the representation to $N = 4$ effectively reduces spatial aliasing errors compared to a lower-order truncation. Since further extension beyond order 4 yields negligible gains, we focus on order 4 for the subsequent analysis.

2.3 Objective Evaluation

To rigorously evaluate the proposed methods, all experiments were conducted at a sampling rate of $F_s = 48000$ Hz and a target Ambisonics order of $N = 4$. The DSHT order for the FD method was set to $N_T = 4$, as determined to be optimal in previous work [13]. We established a unifying design parameter to ensure a fair comparison: the FIR filter length for the TD method was set equal to the STFT hop size for the FD and Baseline methods, with the STFT frame size being double this length. This ensures all methods operate at a similar computational complexity.

2.3.1 Evaluation Metrics

We evaluate the encoding methods by comparing the estimated system impulse response $z^{\text{est}}(t)$ against the ideal response, which is a delayed Dirac delta function weighted by the target spherical harmonic value:

$$z^{\text{ideal}}_{n,m,\sigma,k}(t) = Y^{\sigma}_{n,m}(\theta_k, \phi_k) \cdot \delta(t - t_d) \tag{19}$$

where $\delta(t)$ is the Dirac delta function, and t_d is the target delay, set to the center of the corresponding analysis window, $t_d = L/2$, to ensure a fair and optimal comparison.

2.3.1.1 Signal Fidelity Metrics

Scale-Invariant Signal-to-Distortion Ratio (SI-SDR): To provide a normalized measure of signal quality that is insensitive to absolute amplitude differences, we use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [18]. The metric is computed by first finding the optimal scaling factor α that minimizes the energy of the error between the scaled ideal signal $z^{\text{ideal}}(t)$ and the estimated signal $z^{\text{est}}(t)$. The SI-SDR is then defined as the ratio of the energy of the scaled target component to the energy of the error component:

$$\text{SI-SDR} = 10\log_{10}\left(\frac{\|\alpha z^{\text{ideal}}\|^2}{\|z^{\text{est}} - \alpha z^{\text{ideal}}\|^2}\right), \alpha = \frac{\langle z^{\text{est}}, z^{\text{ideal}}\rangle}{\|z^{\text{ideal}}\|^2} \tag{20}$$

Log-Spectral Distance (LSD): LSD measures the spectral deviation by calculating the root-mean-square error between the log-magnitude spectra of the estimated and ideal responses [19]. This provides a perceptually relevant measure of timbral difference.

$$\text{LSD} = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f}\left[20\log_{10}|Z^{\text{est}}(\omega_i)| - 20\log_{10}|Z^{\text{ideal}}(\omega_i)|\right]^2}$$

(21)

where $Z(\omega) = \mathscr{F}\{z(t)\}$ is the Fourier transform of the response and N_f is the number of frequency bins.

2.3.1.2 Temporal Consistency Metric

Directional Power Ratio (DPR): To quantify temporal smearing artifacts, we define the Directional Power Ratio (DPR). This metric is specifically designed to measure pre-echo, a non-causal artifact where signal energy appears before the true impulse arrival. Pre-echo serves as an unambiguous indicator of algorithmic artifacts, as no energy can physically arrive from a source before the direct-path impulse. This provides a clean, objective measure of temporal distortion introduced by the encoder. DPR is therefore defined as the ratio of the energy in the desired causal portion of the impulse response (from the main peak onwards) to the energy in the spurious pre-echo portion. A higher DPR indicates superior temporal precision.

$$\mathrm{DPR} = 10\log_{10}\left(\frac{\sum_{t=t_{\mathrm{peak}}}^{L-1}|z^{\mathrm{est}}(t)|^2}{\sum_{t=0}^{t_{\mathrm{peak}}-1}|z^{\mathrm{est}}(t)|^2+\varepsilon}\right) \tag{22}$$

where $t_{\mathrm{peak}} = \arg\max_t |z^{\mathrm{est}}(t)|$ is the position of the main peak, L is the response length, and ε is a small constant (e.g., 10^{-10}) to prevent division by zero.

2.3.1.3 Spatial Accuracy Metrics

Spatial Pattern Mismatch (SPM-KL): This evaluates the accuracy of the spatial energy distribution using KL-Divergence.

$$D_{KL}\left(P_{\mathrm{ideal}}\,\|P_{\mathrm{est}}\right) = \sum_{i=1}^{N_g} P_{\mathrm{ideal}}\,(i)\log\left(\frac{P_{\mathrm{ideal}}\,(i)}{P_{\mathrm{est}}\,(i)+\varepsilon}\right) \tag{23}$$

where $P(i)$ is the normalized spatial power at grid point i, N_g is the number of spherical grid points, and $\varepsilon = 10^{-8}$ prevents numerical instability. The spatial power is computed as $P_{\mathrm{est}}\,(i) = \sum_t |s_i(t)|^2$ with $s_i(t) = \sum_{n,m,\sigma} z^{\mathrm{est}}_{n,m,\sigma,k}(t)\cdot Y^{\sigma}_{n,m}(\theta_i,\phi_i)$.

Directional Gain Consistency (DGC): This metric measures the uniformity of the system response across different directions. A lower DGC value indicates a more uniform spatial response.

$$\mathrm{DGC} = \mathrm{std}\,(G_1, G_2, \ldots, G_K) \tag{24}$$

where $G_k = 10\log_{10}\left(\sum_{n,m,\sigma}\sum_t \left|z^{\mathrm{est}}_{n,m,\sigma,k}(t)\right|^2\right)$ is the total energy gain for direction k, and $\mathrm{std}(\cdot)$ denotes standard deviation.

2.3.2 Simple Condition

We first evaluate the methods under a simple condition, defined as an anechoic, noise-free environment with a single sound source treated as a far-field plane wave. We analyze a typical configuration using the FIR filter length of 1025 samples. The assessment is based on analyzing the end-to-end system impulse response against an ideal, delayed Dirac delta function.

It is important to note that because this ideal reference is a mathematically perfect impulse (with zero temporal smearing), any energy introduced by the filters (e.g., pre-echo or post-ringing) is treated as distortion by the SI-SDR metric. This strictness naturally results in negative dB values for all methods. The critical finding is therefore not the absolute values, but the large relative differences between methods.

Table 1: Objective performance evaluation for the 1025-point configuration. Best performance in each column is highlighted in bold.

Method	SI-SDR (↑)	LSD (↓)	DPR (↑)	DGC (↓)	SPM-KL (↓)
Baseline	-13.98	7.63	–	1.85	1.41
FD	-10.82	8.49	3.11	1.56	1.26
TD	**-2.65**	**4.59**	**9.35**	**1.06**	**0.81**

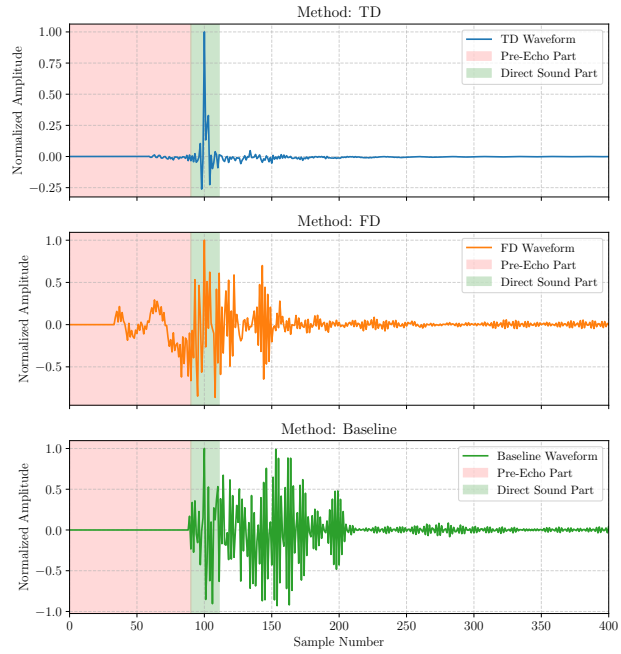

Fig. 4: System impulse response waveforms for the TD, FD, and Baseline methods.

The results for this configuration, summarized in Table 1, establish a clear and decisive performance hierarchy. The proposed TD method significantly outperforms both the FD and Baseline approaches across every evaluated metric. In terms of signal fidelity, the TD method demonstrates a commanding lead, with an SI-SDR score over 8 dB higher than the FD method and a more accurate spectral shape (lower LSD). Furthermore, the TD method shows remarkable suppression of temporal artifacts, with a DPR value more than three times higher than the FD method, indicating a much cleaner impulse response with minimal pre-echo. Finally, the TD method provides superior spatial accuracy, yielding a more uniform directional gain (lower DGC) and a more precise spatial energy distribution (lower SPM-KL).

To provide a more intuitive understanding of these results, Figure 4 illustrates the typical system impulse responses for each method. The TD method's waveform is exceptionally clean, with a sharp, well-defined peak and virtually no energy in the pre-echo region, visually con-

firming its superior performance scores. The FD method, in contrast, exhibits noticeable pre-echo and subsequent post-ringing, which aligns with its intermediate DPR score. The waveform for the Baseline method reveals why the DPR metric is not applicable to it. While some pre-echo is present, the signal is overwhelmingly dominated by a severe and prolonged post-echo that makes the concept of a distinct main peak almost meaningless.

To verify if these findings hold universally, we evaluated the methods across a full range of design parameters: $\{257, 513, 1025, 2049, 4097\}$ samples. The results, shown in Figure 5, confirm the general superiority of the TD method while revealing more complex dynamics.

As illustrated in the figure, the leadership of the TD method in the core fidelity metric SI-SDR is robust across all parameter values, validating its holistic optimization strategy. The SI-SDR of the TD method is consistently 4-8 dB higher than its counterparts.

In terms of spectral and spatial accuracy, the TD method also maintains a clear advantage, consistently achieving the lowest (best) LSD and DGC scores. All methods exhibit a "U-shaped" performance curve, indicating an optimal region around the 1025-sample parameter length. This arises from a fundamental trade-off in model complexity: shorter filters lack the degrees of freedom to capture the array's full acoustic response (underfitting), while longer filters risk introducing ringing artifacts due to overfitting. The 1025-sample length appears to provide the best balance by matching the intrinsic timescale of the array's physical response. Interestingly, the TD method's pre-echo suppression (DPR) also performs best in this optimal region, highlighting a further internal trade-off within its optimization between temporal fidelity and other constraints.

In summary, the comprehensive evaluation shows that the TD method holds an overwhelming and robust advantage in dimensions critical to audio fidelity—signal waveform, temporal dynamics, and spatial accuracy—making it the preferred choice for high-fidelity applications. However, the multi-parameter analysis reveals that no single method is perfect in all aspects. The FD method may have operational sweet spots, and the TD method's pre-echo suppression is dependent on the specific filter design length, revealing a complex performance trade-off that provides deeper insight for selecting an optimal method based on specific application priorities.

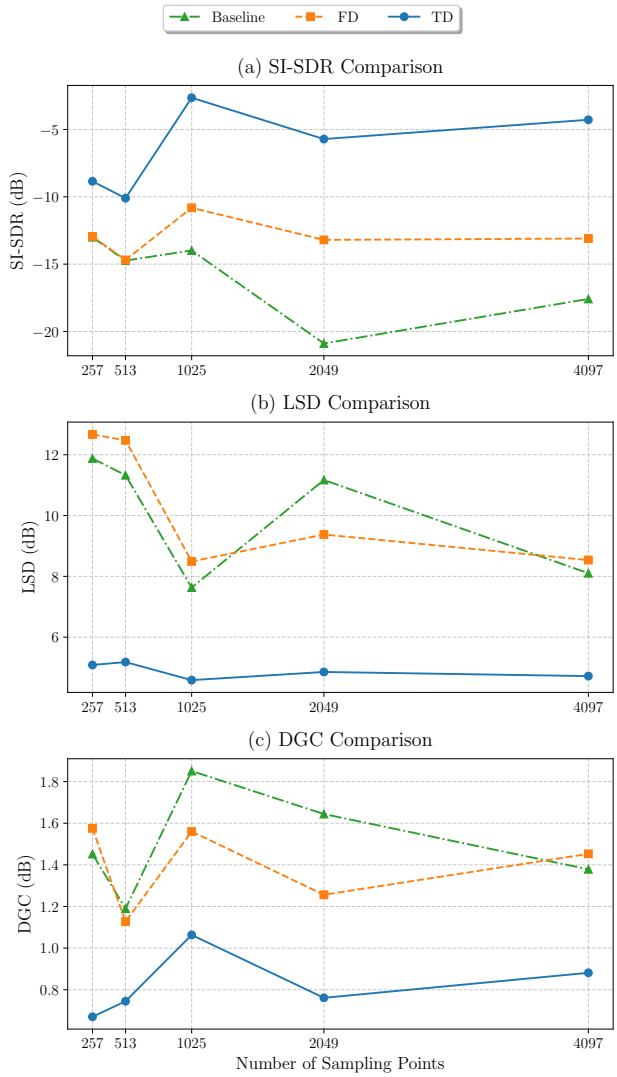

Fig. 5: Performance comparison across evaluation metrics under different hyper-parameters. All methods use equal computational complexity: TD filter length = FD hop size. The STFT window length is twice the hop size.

2.3.3 Complex Conditions

To validate the practical robustness of the proposed methods, we extend the evaluation to a series of challenging acoustic scenarios that emulate real-world conditions. We analyze the performance of each encoder when subjected to varying levels of ambient noise and in complex sound fields with multiple concurrent sources. The primary objective is to determine if the performance advantages and characteristics identified in the anechoic case remain consistent under these common acoustic impairments.

2.3.3.1 Noise To assess the encoders' robustness against ambient noise, the experiment was conducted by adding diffuse white Gaussian noise to the four-channel anechoic microphone signals. This simulation of noisy environments was performed across four distinct signal-to-noise ratios (SNRs): 30 dB, 20 dB, 10 dB, and 0 dB. The resulting noisy signals were then processed by each encoder, and the output was compared against the clean (noise-free) ideal HOA response.

The evaluation results, presented in Table 2, demonstrate the superior robustness of the TD method in noisy conditions. As expected, the performance of all methods degrades as the SNR decreases. However, the TD method consistently maintains a significant advantage in most key areas. Its lead in SI-SDR and DPR is substantial across all noise levels, indicating a much better preservation of waveform fidelity and greater resilience to temporal artifacts. Furthermore, its consistently lower LSD and SPM-KL scores show that it more effectively maintains the sound field's crucial spectral and spatial characteristics. In-

Table 2: Performance evaluation under noise with different SNR(dB).

SNR	Method	SI-SDR (↑)	LSD (↓)	DPR (↑)	DGC (↓)	SPM-KL (↓)
0	Baseline	-16.55	9.09	–	0.88	1.30
	FD	-11.10	7.45	0.18	**0.87**	1.16
	TD	**-8.44**	**5.35**	**0.63**	0.94	**1.14**
10	Baseline	-13.00	7.63	–	1.00	1.28
	FD	-9.85	7.52	2.30	**0.93**	1.17
	TD	**-5.18**	**4.95**	**6.29**	0.95	**1.10**
20	Baseline	-12.72	7.30	–	1.50	1.31
	FD	-9.70	7.83	2.81	**1.08**	1.21
	TD	**-4.60**	**5.52**	**8.56**	1.08	**1.05**
30	Baseline	-12.65	7.34	–	1.63	1.32
	FD	-9.68	8.04	2.91	**1.10**	1.22
	TD	**-4.54**	**6.25**	**8.91**	1.12	**1.04**

terestingly, while the TD method excels in almost all aspects, the FD method exhibits a marginal but consistent advantage in DGC, suggesting its frequency-specific optimization may offer slightly more uniform gain across directions in the presence of noise. Nevertheless, considering the overwhelming advantages in signal fidelity and spatial accuracy, the TD method's holistic optimization proves to be more robust, providing a more graceful performance degradation in noisy environments.

2.3.3.2 Multiple Sound Sources The second experiment assesses performance in complex acoustic scenes by linearly mixing anechoic signals from multiple, uncorrelated sources. Two conditions were evaluated: scenes with two concurrent sources (pairs) and three concurrent sources (triplets), with results averaged over numerous directional combinations. As the sound field becomes more complex, the evaluation criteria shift and certain single-source metrics become less informative. Consequently, two metrics are omitted here: DGC is ill-defined for a multi-source stimulus, and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) is also excluded. The reason for excluding SI-SDR is that its reliance on a specific, phase-sensitive reference waveform is less meaningful for evaluating a mixture of multiple uncorrelated sources. The consistently low SI-SDR scores (below -60 dB) across all methods in preliminary tests confirmed the metric's lack of discriminatory power in this scenario. Thus, the evaluation focus shifts to spectral, temporal, and spatial characteristics.

The results, presented in Table 3, reveal a significant performance trade-off. The TD method continues to assert its dominance in temporal precision (highest DPR) and maintains the best overall spatial accuracy (lowest SPM-KL). However, a notable characteristic emerges in spectral fidelity. While the TD method's spectral performance (LSD) is superior to the FD method's, neither matches the raw spectral balance of the signal mixture that is incidentally preserved by the simplistic Baseline (ASM) approach.

This suggests a clear divergence in how the algorithms handle complex scenes. The TD method appears to prioritize the accurate reconstruction of the spatial field and the maintenance of a clean temporal structure, a complex optimization that comes at the cost of some spectral coloration when compared to a non-optimized baseline. This highlights a critical choice for designers based on application priorities: the TD method is superior for applications requiring precise spatial imaging and artifact-free audio, which are the primary goals of advanced HOA encoding.

Table 3: Average performance under multi-source conditions.

Condition	Method	LSD (↓)	DPR (↑)	SPM-KL (↓)
2 sources	Baseline	**5.98**	–	0.63
	FD	8.47	5.80	0.61
	TD	7.22	**12.54**	**0.61**
3 sources	Baseline	**5.86**	–	0.43
	FD	8.42	4.98	0.42
	TD	7.15	**12.78**	**0.42**

2.4 Subjective Evaluation

To complement the objective evaluations, formal listening tests were conducted to assess the perceptual performance of the proposed methods. The experiment evaluated two key attributes: overall sound quality and spatial localization (horizontal and vertical). The methods under evaluation were the baseline, our proposed TD method, and the FD method.

A total of 16 listeners (12 male, 4 female), aged between 20 and 25, participated in the experiments, all with no self-reported hearing loss. **The listening tests were conducted in an anechoic chamber** to ensure a quiet environment free from external distractions. The stimuli were presented over Sennheiser HD 600 headphones using the webMUSHRA framework [20] following the MUSHRA paradigm (ITU-R BS.1534). The test consisted of 80 trials (20 for quality, 30 for horizontal, and 30 for vertical localization), with the order of trials randomized for each participant. The experiment was split into three independent sub-experiments (Sound Quality, Horizontal Localization, Vertical Localization). The approximate duration was $\sim$40 minutes for the quality task and $\sim$80 minutes ($\sim$40 min per sub-task) for the spatial tasks, which participants could complete separately.

In each trial, listeners were presented with the hidden reference (ideal HOA signal), the methods being evaluated, and a corresponding anchor. For the sound quality assessment, the anchor was the reference signal low-pass filtered at 3.5 kHz. For the spatial localization tests, the anchor was an ideal HOA signal rendered from the diametrically opposite direction. The source signals, consisting of pink noise and excerpts of speech, symphony, clapper and dog bark from [21], were encoded into the 4th-order HOA domain ($N = 4$) using the transfer functions measured from the smartphone array. Subsequently, these signals were decoded to binaural format using the Neumann KU100 HRTF database from [22]. We deliberately utilized a standard linear decoding chain, avoiding parametric spatial enhancement techniques, to strictly isolate the perceptual performance of the encoders themselves. This choice clarifies that the aim of this test is to relatively compare the performance of the proposed encoding implementations, rather than to evaluate the absolute state-of-the-art listening qual-

Table 4: Subjective Audio Quality Evaluation (Paired t-test, trial-level analysis). All comparisons are statistically significant ($p < 0.001$).

Angle	FD vs. Baseline			TD vs. Baseline			TD vs. FD		
	Mean Diff	95% CI	Cohen's d_z	Mean Diff	95% CI	Cohen's d_z	Mean Diff	95% CI	Cohen's d_z
Overall	-15.78	[-18.35, -13.20]	-0.67	9.14	[6.70, 11.57]	0.41	24.91	[22.58, 27.24]	1.17
18°	-17.10	[-22.48, -11.72]	-0.71	4.35	[-0.80, 9.50]	0.19	21.45	[16.89, 26.01]	0.96
90°	-21.60	[-26.05, -17.15]	-1.08	1.65	[-2.97, 6.27]	0.08	23.25	[18.99, 27.51]	1.17
198°	-14.31	[-20.06, -8.57]	-0.58	14.54	[8.98, 20.09]	0.59	28.85	[22.75, 34.95]	1.13
270°	-10.09	[-15.80, -4.38]	-0.41	16.00	[10.45, 21.55]	0.66	26.09	[21.05, 31.13]	1.25

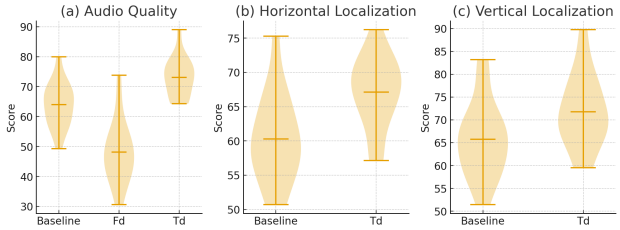

Fig. 6: Overall subjective evaluation results for Quality, Azimuth (Horizontal), and Elevation (Vertical) localization tasks. Bars show mean scores (based on per-subject means, $n = 16$), and error bars represent 95% confidence intervals.

ity that could be achieved with advanced parametric renderers (e.g., MagLS[23]).

For statistical analysis, a repeated measures analysis of variance (RM-ANOVA) was first conducted to assess the main effect of the encoding method. Subsequently, paired t-tests were performed at the trial level (i.e., treating each listener $\times$ angle $\times$ material rating as a paired record, $n \geq 320$ for overall, $n \geq 80$ for by-angle) to maintain statistical consistency with the original Table 5 and 6. This approach provides a robust analysis of the overall variance. In line with modern reporting standards, we report effect sizes (Cohen's d_z) and the 95% confidence interval (CI) of the mean difference alongside t-values and p-values.

The overall results, visualized with descriptive statistics (based on per-subject means, $n = 16$) in Fig. 6, show a clear performance hierarchy. For overall sound quality, the TD method (Mean = 73.09, SD = 6.62) was rated significantly higher than both the Baseline (Mean = 63.96, SD = 7.50) and the FD method (Mean = 48.18, SD = 10.51). The subsequent inferential statistics (Table 4) confirm these differences are significant.

The detailed quality results in Table 4 reveal two key insights. First is the discrepancy between subjective quality (Baseline $>$ FD) and objective metrics like SI-SDR (FD $>$ Baseline, Table 1). This is attributable to their different methodological trade-offs. The FD method's frequency-bin-specific optimization (Sec 2.3), particularly its high-frequency magnitude-only matching (Eq. 10), introduces non-causal pre-echo artifacts (Fig. 4). The Baseline's broadband optimization (Eq. 8), in contrast, produces significant post-echo. Psychoacoustically, human hearing has tolerance for post-echo, which can be masked or perceived as "natural" reverberation, but is highly sensitive to "unnatural" pre-echo [24, 25]. Listeners thus strongly preferred the "naturally" distorted Baseline over the "unnaturally" artifacted FD (Mean Diff = -15.78, $d_z = -0.67$).

Second, the TD method's quality advantage over the Baseline (Overall Mean Diff = 9.14, $d_z = 0.41$) is most pronounced for rear-hemisphere sources (e.g., 198° $d_z = 0.59$; 270° $d_z = 0.66$). This suggests the TD method's holistic time-domain optimization better preserves the fragile, high-frequency spectral and transient cues shaped by the pinna, which are critical for resolving front-back ambiguity [26].

For the spatial localization tasks, the FD method was excluded from the comparison. Its optimization goal (spectral magnitude) and resultant pre-echo artifacts methodologically interfere with the subtle temporal and spectral cues that the human auditory system relies on for localization [24, 26, 25], making a direct comparison of its spatial fidelity untenable.

The spatial evaluation therefore compared the TD method against the Baseline. Overall scores (Fig. 6) favored the TD method for both horizontal (Mean=67.14 vs. 60.28) and vertical (Mean=71.35 vs. 65.52) tasks. The detailed pairwise results in Table 5 quantify the practical significance of these differences. For horizontal localization (Table 5a), the TD method's overall advantage is clear (Mean Diff = 6.87, $p < .001$, $d_z = 0.26$). This advantage is driven almost entirely by its superior performance in the rear-hemisphere, showing large effect sizes (e.g., 126° $d_z = 0.80$; 162° $d_z = 0.83$), confirming its strength in resolving front-back ambiguity. For frontal sources (18°, 54°, 90°), the differences were not statistically significant ($p > 0.14$).

For vertical localization (Table 5b), the TD method also showed a significant overall advantage (Mean Diff = 6.05, $p < .001$, $d_z = 0.26$). The effect was largest and most consistent at the vertical extremes (e.g., 25° $d_z = 0.82$; 155° $d_z = 0.83$), which are typically challenging for irregular arrays.

In conclusion, the subjective evaluations, supported by comprehensive trial-level analysis including effect sizes and confidence intervals, corroborate the objective findings. The TD method provides a statistically significant and perceptually meaningful improvement in both sound quality and spatial accuracy. Its strength is particularly evident in preserving the fragile acoustic cues required for challenging spatial tasks, validating its holistic optimization approach for high-fidelity spatial audio capture.

Table 5: Spatial Localization Evaluation (Paired t-test: Baseline vs. Time-Domain, trial-level analysis). All comparisons are TD vs. Baseline.

(a) Horizontal Localization (Azimuth)

Angle (°)	**Mean Diff**	**95% CI**	t**-value**	p**-value**	**Cohen's** d_z
Overall	6.87	[4.46, 9.28]	5.594	<**0.001**	0.255
18	-7.92	[-19.21, 3.38]	-1.410	0.165	-0.203
54	-4.25	[-10.02, 1.52]	-1.466	0.149	-0.212
90	-4.29	[-10.07, 1.48]	-1.481	0.145	-0.214
126	18.71	[12.01, 25.41]	5.569	<**0.001**	0.804
162	23.52	[15.11, 31.93]	5.780	<**0.001**	0.834
198	4.83	[-2.42, 12.09]	1.328	0.191	0.192
234	-0.27	[-6.21, 5.67]	-0.093	0.926	-0.013
270	5.92	[-0.49, 12.32]	1.838	0.072	0.265
306	17.29	[11.83, 22.75]	6.376	<**0.001**	0.920
342	4.23	[-1.75, 10.21]	1.411	0.165	0.204

(b) Vertical Localization (Elevation)

Angle (°)	**Mean Diff**	**95% CI**	t**-value**	p**-value**	**Cohen's** d_z
Overall	6.05	[3.96, 8.14]	5.689	<**0.001**	0.260
25	18.35	[8.14, 28.57]	3.830	<**0.001**	0.827
57	-0.38	[-8.64, 7.89]	-0.098	0.923	-0.021
90	2.06	[-3.18, 7.30]	0.788	0.434	0.170
123	9.79	[3.65, 15.93]	3.397	**0.001**	0.733
155	-1.40	[-8.08, 5.29]	-0.445	0.658	-0.096

3 CONCLUSION

In this paper, we presented robust encoding implementations based on the Spherical Harmonic Beamforming (SHB-AE) perspective to address the challenges of sparse, irregular microphone arrays. Specifically, we developed a frequency-domain method with high-frequency magnitude matching and a time-domain (TD) method characterized by holistic broadband optimization. Our comprehensive evaluations, using a real smartphone array, conclusively demonstrated that the proposed TD method is the superior approach. By implicitly enforcing cross-frequency coherence, the TD method yielded significant objective improvements in signal fidelity, spectral accuracy, and spatial representation. These findings were validated by subjective listening tests, which confirmed higher perceived sound quality and more accurate localization, especially for challenging rear-hemisphere sources. The method also proved robust across noisy and multi-source conditions. The proposed TD implementation offers a practical and computationally efficient solution for bringing high-fidelity spatial audio capture to a wide range of consumer devices.

For future work, we plan to explore adaptive filtering techniques that allow the TD filters to adjust to changing acoustic environments, such as varying noise levels. Another direction is the integration of machine learning, where a lightweight neural network could predict optimal filter parameters for specific acoustic scenes, combining the robustness of our physics-based model with data-driven adaptability. Finally, optimizing the filters using a perceptually motivated cost function could further align the encoder's output with the characteristics of human hearing, potentially leading to even greater subjective quality.

4 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.

5 REFERENCES

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

[2] Y. Gayer, V. Tourbabin, Z. B. Hur, D. L. Alon, and B. Rafaely, "Array-Aware Ambisonics and HRTF Encoding for Binaural Reproduction With Wearable Arrays," (2025), URL `https://arxiv.org/abs/2507.11091`.

[3] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8 (Springer, 2015).

[4] A. Bastine, L. Birnie, T. D. Abhayapala, P. Samarasinghe, and V. Tourbabin, "Ambisonics capture using microphones on head-worn device of arbitrary geometry," presented at the *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 309–313 (2022).

[5] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov Regularization and Total Least Squares," *SIAM Journal on Matrix Analysis and Applications*, p. 185–194 (1999 Jan), doi:10.1137/

s0895479897326432, URL http://dx.doi.org/10.1137/s0895479897326432.

[6] Y. Gayer, V. Tourbabin, Z. Ben-Hur, J. Donley, and B. Rafaely, "Ambisonics Encoding For Arbitrary Microphone Arrays Incorporating Residual Channels For Binaural Reproduction," *arXiv preprint arXiv:2402.17362* (2024).

[7] S. Gao, X. Wu, and T. Qu, "High order ambisonics encoding method using differential microphone array," presented at the *Audio Engineering Society Convention 144* (2018).

[8] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2062–2075 (2022), doi:10.1109/TASLP.2022.3182857.

[9] D. S. Talagala, W. Zhang, and T. 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, p. 1573–1585 (2013 Aug), doi:10.1109/tasl.2013.2255282, URL https://doi.org/10.1109/tasl.2013.2255282.

[10] Y. Qiao, S. Chakrabarty, and E. A. Habets, "Neural ambisonics encoding of multi-speaker speech from a circular microphone array," *arXiv preprint arXiv:2403.02450* (2024).

[11] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the audio engineering society*, vol. 53, no. 11, pp. 1004–1025 (2005).

[12] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography* (Academic press, 1999).

[13] Y. You, Y. Qian, T. Qu, B. Wang, and X. Lv, "Spherical harmonic beamforming based Ambisonics encoding and upscaling method for smartphone microphone array," presented at the *Audio Engineering Society Convention 158* (2025).

[14] K. F. Warnick, R. Maaskant, M. V. Ivashina, D. B. Davidson, and B. D. Jeffs, *Array Signal Processing, Calibration, and Beamforming*, p. 325–395, EuMA High Frequency Technologies Series (Cambridge University Press) (2018).

[15] J. Yu and K. D. Donohue, "Geometry descriptors of irregular microphone arrays related to beamforming performance," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 249 (2012).

[16] A. Farina, "Advancements in impulse response measurements by sine sweeps," presented at the *Audio engineering society convention 122* (2007).

[17] Á. González, "Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices," *Mathematical geosciences*, vol. 42, no. 1, pp. 49–64 (2010).

[18] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630 (2019).

[19] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391 (1976), doi:10.1109/TASSP.1976.1162849.

[20] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, *et al.*, "webMUSHRA—A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1 (2018).

[21] C. Bauer and M. Vinton, "Joint optimization of scale factors and Huffman code books for MPEG-4 AAC," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 177–189 (2005), doi:10.1109/TSP.2005.861092.

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

[23] A. Bastine, L. Birnie, T. D. Abhayapala, P. Samarasinghe, and V. Tourbabin, "Magnitude Least-Squares Based Ambisonics Estimation of Head-Worn Device Microphone Measurements for Binaural Reproduction," presented at the *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 474–478 (2024), doi:10.1109/IWAENC61483.2024.10693997.

[24] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. (Brill, 2013).

[25] J. Blauert, *Spatial hearing: The psychophysics of human sound localization* (MIT Press, 1997).

[26] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I: Stimulus synthesis," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867 (1989).

THE AUTHORS