

When Does Content-Awareness Help? An Analysis of Ambisonics Order Adaptation for Networked Immersive Music Performance

Claudia Rinaldi

University of L'Aquila Research Unit
L'Aquila, Italy

ORCID: 0000-0002-1356-8151

claudia.rinaldi@univaq.it

Stefano Giacomelli

DISIM, University of L'Aquila
L'Aquila, Italy

ORCID: 0009-0009-0438-1748

stefano.giacomelli@graduate.univaq.it

Marco Giordano

DISIM, University of L'Aquila
L'Aquila, Italy

ORCID: 0009-0001-1649-6085

marco.giordano3@graduate.univaq.it

Mirco Pezzoli

DEIB, Politecnico di Milano
Milano, Italy

ORCID: 0000-0003-1296-0992

mirco.pezzoli@polimi.it

Alberto Bernardini

DEIB, Politecnico di Milano
Milano, Italy

ORCID: 0000-0001-7973-0134

alberto.bernardini@polimi.it

Abstract—Immersive Networked Music Performance (INMP) streams Higher-Order Ambisonics (HOA) over best-effort networks, where bandwidth fluctuations and packet loss erode spatial fidelity. State-of-the-art schemes adapt the Ambisonics order to the available bandwidth, applying a single global order to the whole scene. Two well-established psychoacoustic results, i.e. the strong dependence of localization accuracy on source incidence, and the early perceptual saturation of diffuse reverberation with low orders, make it tempting to allocate order non-uniformly across the scene by perceptual importance. This paper asks whether such content-awareness actually helps, and characterizes where it does and does not. Our central finding is a negative one: under a spherical-harmonic (SH) coefficient-channel budget, a single global HOA field is coefficient-channel optimal, so decomposing the scene into independently encoded HOA patches and allocating order per component cannot reduce per-source localization distortion at equal budget. We show this and quantify it on a scene whose structure is *estimated* from the signal by Directional Audio Coding (DirAC). Content-awareness can instead be exploited through a different *rendering*: under a constrained transport budget, carrying the higher-importance foreground as objects decouples its localisation from SH-order truncation, while a low-order bed carries the diffuse field. Reallocating rate and error protection by perceptual importance is a complementary direction. We make the signal-availability assumptions explicit, detail the model and estimation pipeline, and delineate the implications for resilient INMP system design.

Index Terms—networked immersive audio, higher-order Ambisonics, spatial audio, packet loss, psychoacoustics, directional audio coding.

I. INTRODUCTION

Networked Music Performance (NMP) lets geographically distributed musicians play together in real time [1]; its immersive variant (INMP) adds spatial rendering so that performers perceive a shared acoustic scene with the directional cues they rely on when co-located [2]–[4]. Higher-Order Ambisonics

(HOA) is a natural encoding for this purpose: it represents the sound field on a spherical-harmonic (SH) basis in a playback-agnostic, scalable form [5], [6]. The *order* N of the expansion sets the spatial resolution and, with it, the number of transmitted coefficient channels, $C(N) = (N + 1)^2$ for three-dimensional reproduction. The number of SH coefficient channels therefore grows directly with order, while the corresponding bitrate depends on the adopted coding and transport scheme.

These impairments are the defining difficulty of INMP. Latency breaks musical synchronisation [7], while a recent subjective study of an intercity INMP link additionally reports that packet loss degrades *sound quality* more than it affects timing [8], [9]. Spatial quality is thus the resource most directly threatened by loss. The state of the art adapts the Ambisonics order to the network: Ostan *et al.* [10] monitor the available bandwidth and select the largest order that fits, signalling order changes to the decoder and applying a single, global order to the entire scene.

Two psychoacoustic results invite a content-aware refinement of this uniform strategy. First, localisation accuracy depends on order *and* on source incidence: frontal and, to a lesser extent, rear sources are reproduced comparatively accurately at low order, whereas lateral sources suffer pronounced under-localisation that only high orders correct [11]. Second, once the direct sound is rendered separately, the perceived quality of the diffuse reverberant field ceases to improve beyond order two or three, far below the order the direct path requires [12]–[14]. Both results say that the order required for transparency is *not uniform* across the scene. It is therefore natural to conjecture that allocating order by perceptual importance—by direction, by direct/diffuse role, by salience—would use a given budget better than a single global order.

This paper tests that conjecture, finds it largely false in the form most directly suggested, and then identifies where content-awareness can still be exploited. Our contributions are analytical:

- we make explicit the two signal-availability regimes—production-side stems versus a post-mix HOA stream—on which any content-aware scheme depends, and show that the object-rendering benefit is fully available only in the former (Sec. III);
- we establish that, for independently encoded per-component HOA patches under a coefficient-channel budget, a single global HOA field is *coefficient-channel optimal*: per-component order allocation—content-aware or not—cannot reduce per-source localisation distortion at equal budget, and we quantify this on a DirAC-estimated scene (Sec. VI);
- we locate where content-awareness *does* help—object rendering of the foreground, and protection and bit allocation rather than order allocation—and bridge the channel-count analysis to bitrate, packetisation and loss (Sec. VII).

The study is objective; subjective validation is future work. We regard the negative core result as the contribution’s value: it identifies a structural limitation of per-component HOA order allocation and redirects attention toward alternative content-aware strategies.

II. BACKGROUND AND RELATED WORK

Adaptive HOA over networks. HOA is natively scalable: truncating the SH expansion from order N to $N' < N$ yields a lower-order, lower-channel-count version of the same field, discarding the $(N + 1)^2 - (N' + 1)^2$ highest channels [5]. Ostan *et al.* [10] turn this into a real-time mechanism, monitoring throughput and signalling order changes to the decoder through the Real-time Transport Protocol (RTP), with an energy fade on the discarded high-order channels to avoid audible discontinuities. Their adaptation is triggered by bandwidth alone and applied uniformly. NetMusic3D [8] provides the surrounding INMP system—multichannel transport, jitter buffering, selectable order, and low-latency coding—together with the empirical finding that loss harms spatial quality rather than timing. Edge and 5G deployments [15], [16] define the low-latency-yet-lossy regime that is most relevant below, and in which processing may be hosted at a Mobile Edge Computing (MEC) node close to each endpoint.

HOA coding and decomposition. Decomposition has also been exploited in established HOA coding schemes, but with a different goal and transport model from the one analysed here. Scene-based HOA coding separates predominant sound components from an ambient residual and transports them efficiently using side information and conventional audio coding [17], [18]. Other approaches exploit signal-dependent representations to reduce the transport cost, including frequency-domain singular-value decomposition [19] and Higher-Order DirAC parameterisation [20]. These schemes do not represent each scene component as an independent HOA patch and

therefore do not incur a separate SH basis cost per component. They are thus outside the scope of the channel-count result developed in Sec. VI; our question is narrower, namely whether allocating different HOA orders to independent component fields can outperform one shared HOA field at equal coefficient-channel budget.

Loss concealment recovers lost samples after the fact—for NMP, e.g., by hybrid signal- and model-based prediction at low latency [21]. It is orthogonal to, and composable with, order adaptation.

Psychoacoustics of order. Bertet *et al.* [11] study perceptually and objectively the incidence dependence of localisation, reporting an energy-vector source-width measure α_E of $45^\circ, 30^\circ, 22.5^\circ, 18^\circ$ for orders 1 to 4, and showing that lateral sources are pulled toward the median plane (under-lateralisation) at low order. Engel *et al.* [12] show that, when the direct path is rendered separately, the diffuse field can be truncated to order two or three without perceptible loss—a threshold far below that of the direct sound, and dependent on room diffuseness. Related studies confirm the dominant role of the direct sound and early reflections [14], [22], [23].

Decomposition and salience. DirAC [24] parameterises a sound field into a direction of arrival (DoA) and a diffuseness value per time-frequency tile, and can support a partition between directional foreground and diffuse background. The notion of perceptual “importance” that a content-aware scheme would prioritise is formalised, for audition, by auditory-salience models [25], and, for communication, by importance-aware and semantic source-channel coding [26].

III. SCENARIO AND SIGNAL AVAILABILITY

Whether content-awareness can act, and how, depends on what the transmitting endpoint actually holds. We distinguish two regimes; the distinction is architectural, not a detail, and we keep it in view throughout.

Case A – production-side hybrid. Separate stems or spot signals are available at the sender, as is common when each instrument is close-miked. The foreground can then be transmitted as *objects*—mono signals plus DoA and gain metadata—and the ambience as a low-order HOA bed, i.e. the hybrid rendering of [12] applied to transmission. This is a realistic INMP setting: in [8], for instance, a clarinet is captured with a spherical microphone while a piano is encoded from separate stereo sources.

Case B – post-mix decomposition. Only a mixed first-order Ambisonics (FOA) or higher-order stream is available. As shown in Sec. V, DirAC can estimate, per time-frequency tile, the DoA and the diffuseness, and hence a foreground/diffuse partition with per-source energy-based importance weights. Crucially, these are *parameters*, not signals: recovering clean object waveforms additionally requires source separation, an error-prone step that may introduce leakage and residual interference. Object rendering is therefore only approximately available in Case B, and the benefit analysed in Sec. VII is contingent on separation quality. The channel/order analysis of Sec. VI, by contrast, applies unchanged to both cases.

IV. PERCEPTUAL AND CHANNEL MODEL

This section defines every quantity used later. We analyse horizontal-plane localisation, following [11], while accounting for channels with the full three-dimensional count $C(N) = (N+1)^2$ actually transmitted in HOA systems [8], [10].

Ambisonics order and channel count. An order- N HOA signal encodes the sound field on the SH basis up to degree N ; it comprises $C(N) = (N+1)^2$ SH coefficient channels—so $C(N)$ is a channel count—and larger N yields finer spatial resolution at a higher raw channel count. Order $N = 1$ (four channels, W/X/Y/Z) is first-order Ambisonics (FOA).

Decoder and energy vector. A decoder maps the HOA channels to loudspeaker (or virtual-loudspeaker) gains g_l in directions $\mathbf{u}_l$. Gerzon's *energy vector* $\mathbf{r}_E = (\sum_l g_l^2 \mathbf{u}_l) / (\sum_l g_l^2)$ is the energy-weighted barycentre of the active loudspeakers; its magnitude $r_E \in [0, 1]$ measures how concentrated the reproduced energy is toward the intended direction, and the angle $\alpha_E = \arccos r_E$ estimates the resulting source-width blur [11]. A *max- r_E* decoder applies a per-order weight a_n —a gain shared by all components of spherical-harmonic order n —chosen to maximise r_E ; for the horizontal (two-dimensional) case these are $a_n = \cos(n\pi/(2N+2))$, $n = 0, \dots, N$. This is the standard perceptually motivated choice. Because performers in INMP are arranged in the horizontal plane and the localisation data of [11] are horizontal, we model localisation in that plane, while channel budgets retain the full three-dimensional count $C(N)$. For a *max- r_E* decoded order- N horizontal field the magnitude takes the closed form

$$r_E(N) = \cos\left(\frac{90^\circ}{N+1}\right), \quad \alpha_E(N) = \frac{90^\circ}{N+1}, \quad (1)$$

recovering the $45^\circ/30^\circ/22.5^\circ/18^\circ$ energy-vector blur values reported in [11] for orders $N = 1, 2, 3, 4$ (verified numerically in Sec. V).

Localisation distortion. Under a *max- r_E* decode, the reproduced lateral component of a source at azimuth ϕ is compressed by $r_E(N)$, yielding the predicted apparent azimuth $\hat{\phi} = \arcsin(r_E(N) \sin \phi)$ in the frontal hemisphere and $180^\circ - \arcsin(\cdot)$ in the rear. The localisation distortion is the resulting angular error

$$\Delta(N, \phi) = |\phi - \hat{\phi}|, \quad (2)$$

which is 0° at $\phi = 0^\circ, 180^\circ$ under this predictor and maximal near $\pm 90^\circ$, capturing the incidence-dependent trend reported in [11]. The zero error at the front and rear is a property of the adopted angular predictor and assumes front/back disambiguation outside the model. At $N = 0$ the field is omnidirectional and carries no directional cue. We use its distortion to normalise,

$$\tilde{\Delta}(N, \phi) = \begin{cases} \frac{\Delta(N, \phi)}{\Delta(0, \phi)}, & \Delta(0, \phi) > 0, \\ 0, & \Delta(0, \phi) = 0, \end{cases} \quad (3)$$

so that $\tilde{\Delta}(N, \phi) \in [0, 1]$.

Diffuse-field cost. Following [12], the perceptual penalty for rendering the *diffuse* component at order N_d rather than

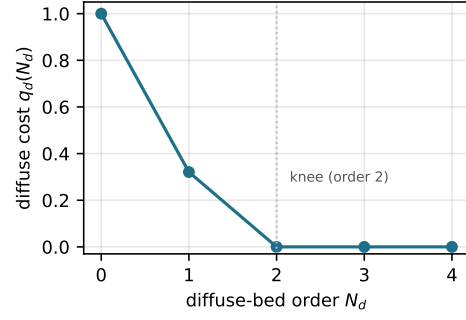


Fig. 1. Diffuse-field perceptual cost $q_d(N_d)$ used in (4): a stylised, saturating deficit consistent with the early perceptual saturation of diffuse reverberation reported in [12], with a knee at order two ($q_d(0)=1$, $q_d(1)\approx 0.32$, $q_d(N_d)=0$ for $N_d \geq 2$).

at full order is a normalised deficit $q_d(N_d) \in [0, 1]$ that falls steeply from order 0 to 1 and saturates by order two (Fig. 1). We use a monotone, saturating curve with a knee at order two, giving $q_d(0) = 1$, $q_d(1) \approx 0.32$, and $q_d(N_d) = 0$ for $N_d \geq 2$; the exact shape is a modelling choice consistent with [12] rather than a measured curve, and could be replaced by a room-dependent (e.g. direct-to-reverberant-ratio-driven) function. The main result (Prop. 1) does not depend on this choice: it concerns per-source localisation and does not involve q_d at all.

Scene, budget, and aggregate distortion. A scene consists of K foreground sources, each with azimuth ϕ_k and *energy-based importance weight* $s_k \in [0, 1]$, plus a diffuse bed with weight $s_d \in [0, 1]$, normalised so that $\sum_{k=1}^K s_k + s_d = 1$ (a partition of modelled perceptual importance). Assigning orders $\mathbf{N} = (N_1, \dots, N_K, N_d)$ to the components incurs the aggregate perceptual distortion (the objective adopted in this work; other perceptual metrics are possible)

$$D(\mathbf{N}) = \sum_{k=1}^K s_k \tilde{\Delta}(N_k, \phi_k) + s_d q_d(N_d), \quad (4)$$

subject to an SH coefficient-channel *budget* B , the maximum number of simultaneously transmitted coefficient channels: $\sum_k C(N_k) + C(N_d) \leq B$. A *decomposed-HOA* allocation encodes each component as its own HOA patch of order N_k (still subject to the blur $\alpha_E(N_k)$ of that patch); the monolithic baseline of [10] is the special case of a single *shared* field, $N_1 = \dots = N_K = N_d = N$, whose cost is $C(N)$ for the whole scene because all sources share the same SH channels.

Equation (4) treats the scene-level distortion as an importance-weighted *sum* of per-component distortions—a deliberately linear, additive model of perceptual importance that ignores non-linear effects such as energetic and informational masking, spatial release from masking, and attentional capture. We adopt it only to summarise a scene as a single scalar for Fig. 4. Crucially, Prop. 1 is stated in terms of *per-source* distortion and does not use D at all, so the coefficient-channel optimality result—and the object-rendering benefit of Sec. VII—are independent of this aggregation model.

V. ANALYTICAL METHODOLOGY

All results are objective and require no listening test. This section specifies how each figure is computed; the pipeline is deterministic and reproducible.¹

A. Energy-vector localisation.

We build a max- r_E decoder for a dense horizontal ring of 72 virtual loudspeakers, compute the panning gains g_l for a source at azimuth ϕ , and evaluate $\mathbf{r}_E$, r_E and α_E directly from their definitions. The numerical α_E reproduces (1) and the corresponding energy-vector blur values reported by Bertet *et al.*. Equation (2) then gives $\Delta(N, \phi)$ for any order and incidence (Fig. 5).

B. Allocation policies.

For a scene and budget B we compare three policies, each choosing orders to minimise D in (4) under the budget: (i) *single-field HOA*—one shared order N , the largest with $C(N) \leq B$, applied to all components; (ii) *decomposed HOA, content-blind*—one common order across the K foreground patches and the bed, the largest with $(K+1)C(N) \leq B$; (iii) *decomposed HOA, content-aware*—per-component orders $\mathbf{N}$ minimising D under $\sum_k C(N_k) + C(N_d) \leq B$, obtained by exhaustive search over the small order grid $N \in \{0, \dots, 4\}$. Sweeping B yields the curves of Fig. 4. For the object-rendering comparison (Fig. 6), the foreground is instead rendered as objects. For illustration, we assume a 5° directional uncertainty, consistent with the DoA-estimation residual observed in the DirAC analysis below; the single field, by contrast, renders the same source at $\Delta(N(B), \phi)$.

C. Scene decomposition (Case B).

The following standard DirAC analysis is used only to estimate, from a mixed FOA stream, the scene parameters required by the numerical illustration; no new decomposition method is proposed.

a) *Simulation*: We synthesise a physically-based FOA scene with the image-source method [27], [28]: a shoebox room of $8 \times 6 \times 3.5$ m with uniform wall absorption 0.38 (Sabine reverberation time ≈ 0.4 s) and image sources up to order 10; a virtual FOA receiver at the room centre; and two source proxies—a louder “lead” at azimuth 78° and 2.0 m, and a quieter “accompaniment” at 12° and 2.4 m. Both are driven by temporally interleaved broadband (white, 150–6000 Hz) noise bursts of 0.16 s with a Hann envelope, one source every 0.5 s, offset by 0.25 s, with the lead 2.5 dB above the accompaniment. No independent noise floor is added, so the diffuse energy is the room’s reverberation of the sources and the direct-to-reverberant ratio (DRR) is set by the source distances (2.0/2.4 m) and the wall absorption. Each image source is encoded as a plane wave into the FOA channels ($W = s$, $[X, Y, Z] = s \mathbf{n}$ for amplitude s from unit direction

$\mathbf{n}$), with gain and delay from its attenuation and distance, and convolved with the dry signals to form the FOA mix. A sampling rate of 16 kHz is used throughout the analysis, covering the 150–6000 Hz band employed here; the analysis targets spatial parameters rather than full-band audio quality.

Fig. 2 shows the resulting scene: its geometry with the decoder’s virtual-loudspeaker ring (a), the directional energy over time (b), and the spectrogram of the omnidirectional channel (c). We drive the two sources with broadband noise bursts rather than recorded instruments deliberately: the study concerns spatial structure and its estimation, not timbre, and broadband, temporally non-overlapping excitation increases the time–frequency sparsity that facilitates reliable per-tile DoA and diffuseness estimates. The two sources therefore stand in for a louder lateral *lead* (78°) and a quieter frontal *accompaniment* (12°); substituting real instrument stems would enrich the spectrogram but leave the geometry-driven conclusions unchanged.

b) *DirAC analysis*: On a short-time Fourier transform (1024-point window, 50% overlap) we compute, per tile, the active intensity $\mathbf{I} = \text{Re}\{W^*[X, Y, Z]\}$ and the energy density $E = \frac{1}{2}(|W|^2 + |X|^2 + |Y|^2 + |Z|^2)$; the DoA is the direction of $\mathbf{I}$ and the diffuseness is $\Psi = 1 - \|\langle \mathbf{I} \rangle\|/\langle E \rangle \in [0, 1]$, with $\langle \cdot \rangle$ a short temporal average.

c) *Partition and importance*: Among tiles above the 70th energy percentile, those with $\Psi < 0.5$ form the foreground and those with $\Psi \geq 0.5$ the diffuse bed; the energy percentile is a silence gate discarding near-silent inter-burst tiles, while $\Psi = 0.5$ is the threshold adopted here for the foreground/diffuse partition. The split is therefore a direct/diffuse (DRR-governed) partition, not a signal-to-noise one, as no separate noise floor is present. Source directions are the peaks (separated by $\geq 20^\circ$) of an energy-weighted DoA histogram over the direct-dominated tiles ($\Psi < 0.2$); each source’s importance weight s_k is the foreground energy nearest its peak and s_d the diffuse energy, all normalised to sum to one.

d) *Result*: The pipeline recovers the two sources at $\sim 78^\circ$ and $\sim 8^\circ$ (ground truth $78^\circ, 12^\circ$), separates a diffuse bed of weight $s_d = 0.23$, and assigns the louder lateral lead the largest importance weight ($s_1 = 0.59$, $s_2 = 0.19$); see Fig. 3. These estimated (ϕ_k, s_k, s_d) are the exact inputs to Figs. 4 and 6, so the analysis rests on estimated, not oracle, structure. Obtaining clean object *signals* for the hybrid rendering of Sec. VII, however, requires source separation—the boundary between Case B and Case A. The estimated source DoAs and the qualitative scene structure are stable to these thresholds, whereas the diffuse weight varies with the Ψ boundary and the DRR (e.g. $s_d \in [0.14, 0.36]$ for $\Psi \in [0.6, 0.4]$); the main order-allocation conclusion is independent of the exact importance weights, since Prop. 1 does not use them. The magnitude of the object-rendering advantage instead depends on source incidence and directional uncertainty, as discussed in Sec. VII. On measured recordings, a gate adaptive to the signal-to-noise ratio (SNR) and DRR could replace these fixed thresholds.

¹The analysis is implemented in Python with NumPy, SciPy and Matplotlib; the FOA room simulation uses the `pyroomacoustics` image-source model [27]. Code and scripts to regenerate every figure will be released by the conference.

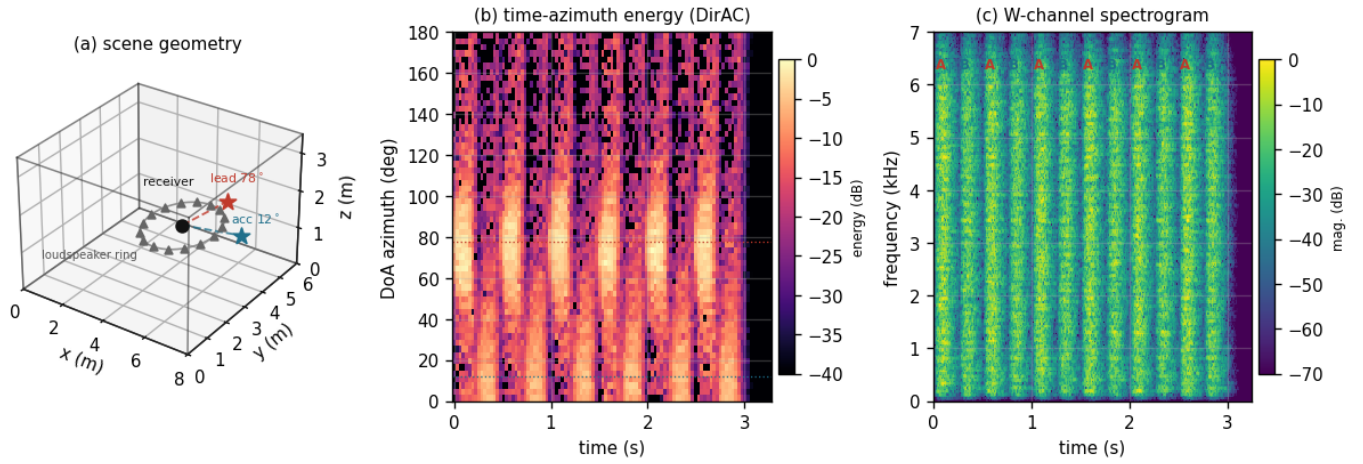


Fig. 2. The simulated first-order Ambisonics scene of Sec. V. (a) Geometry: the shoebox room, the FOA receiver (black dot), the horizontal virtual-loudspeaker ring of the max- r_E decoder (grey), and the two sources, with dashed lines marking their azimuths. (b) Time-azimuth energy map from the DirAC analysis: energy alternates in time between the two sources, at $\sim 78^\circ$ (lead) and $\sim 12^\circ$ (accompaniment), over a diffuse floor. (c) Spectrogram of the omnidirectional (W) channel; letters mark which source is active (A: lead, B: accompaniment). The two sources are temporally interleaved broadband bursts standing in for foreground objects, so most active time-frequency tiles are dominated by a single source, facilitating the DirAC estimation.

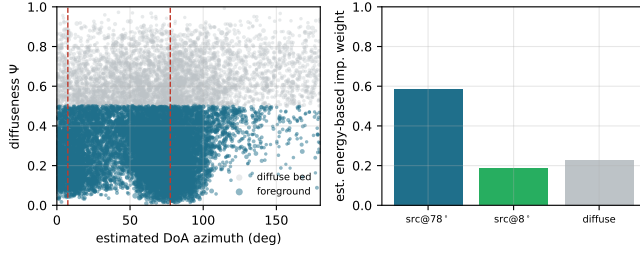


Fig. 3. DirAC decomposition of the simulated FOA scene (Case B). Left: per-tile DoA vs. diffuseness Ψ , with the foreground/diffuse partition and the estimated source DoAs (dashed). Right: estimated energy-based importance weights. Parameters are estimated from the signal; clean object signals are not.

VI. ANALYSIS I: SINGLE-FIELD HOA IS COEFFICIENT-CHANNEL OPTIMAL

Proposition 1: For any scene and any SH coefficient-channel budget B , no order allocation across independently encoded per-component HOA fields achieves lower per-source localisation distortion than a single-field HOA of the largest order N with $C(N) \leq B$.

Proof. Let N be the largest order with $C(N) \leq B$. Any HOA patch of order $N+1$ already requires $C(N+1) > B$ coefficient channels, so within the budget no component of any decomposition can exceed order N ; every source is therefore reproduced at order $\leq N$, hence at blur $\geq \alpha_E(N)$ (as α_E decreases with order). The single field reproduces every source at exactly $\alpha_E(N)$ using $C(N) \leq B$ coefficient channels because all sources share the same SH basis. No decomposed-HOA allocation can therefore achieve lower per-source distortion. $\square$

The result is elementary once channel sharing is made explicit; its interest is that the psychoacoustic motivation of Sec. I

makes the opposite design choice tempting, and practitioners may well reach for it. It is worth noticing that the claim is specifically about *per-component HOA order allocation using independently encoded HOA fields*, one particular content-aware scheme—not about content-awareness in general, nor about joint coding, quantisation, or rate allocation within a shared HOA representation.

Fig. 4 quantifies the proposition on the DirAC-estimated scene of Sec. V. The single field attains the lowest aggregate distortion (4) at every budget: at $B = 16$ coefficient channels it reaches 0.11, against 0.38 for either decomposed policy; the gap narrows but never closes as B grows. More generally, because $\tilde{\Delta}(N, \phi)$ and $q_d(N)$ are non-increasing with order and all weights in (4) are non-negative, the same order-wise dominance implies that the shared field cannot have higher D than any independently encoded decomposed-HOA allocation under the same budget, irrespective of source count or importance-weight distribution. The numerical gap shown in Fig. 4, however, remains scene-dependent.

Content-aware allocation does beat content-blind allocation—e.g. 0.20 vs. 0.38 at $B = 25$, where it raises the order of the higher-importance lateral source and assigns lower orders to the frontal source and the bed—but only *within* the decomposed representation, and always above the coefficient-channel-optimal frontier set by the single field. The reason the decomposition loses is structural: encoding each foreground component as its own HOA patch pays the SH basis cost per component and still leaves each patch subject to α_E . The appeal of non-uniform order is thus misleading once channel sharing is accounted for. The different treatment that makes decomposition worthwhile—rendering the foreground as objects, which are not HOA patches and so are not bound by α_E —is analysed in Sec. VII. The limitation is thus a property of the per-component HOA *representation*, not of

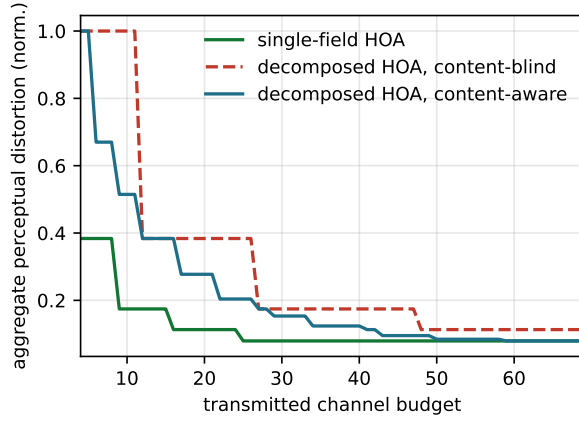


Fig. 4. Aggregate perceptual distortion (4) vs. SH coefficient-channel budget B for the three policies of Sec. V. Single-field HOA is coefficient-channel optimal (Prop. 1); per-component decomposed-HOA order allocation cannot beat it. Content-aware allocation improves only on content-blind allocation *within* the decomposed representation.

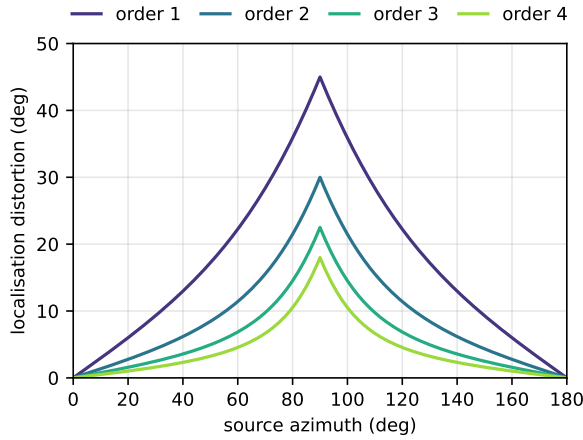


Fig. 5. Localisation distortion $\Delta(N, \phi)$ (2) vs. source azimuth for orders 1–4. Under the adopted predictor, distortion is lowest at frontal and rear incidences and largest near 90° , where increasing the order provides the greatest benefit. This incidence dependence motivates non-uniform order allocation, but Prop. 1 shows that implementing it through independently encoded HOA component fields cannot outperform a shared field at equal coefficient-channel budget.

content-awareness itself.

VII. ANALYSIS II: WHERE CONTENT-AWARENESS HELPS

Proposition 1 concerns HOA *order* allocation under a lossless SH coefficient-channel budget. Content-awareness can still be exploited along axes the proposition does not cover: a different *rendering* of the foreground, and the use of the rate for protection and quantisation rather than for order. In particular, content-aware bit or quantisation allocation within a shared HOA representation does not replicate the SH basis per component and is therefore outside the scope of Prop. 1, consistently with existing HOA coding approaches [19], [20].

Object rendering escapes the isotropic blur. A limitation of single-field adaptation is that reducing the global order under pressure increases the spatial blur of *every* source (Fig. 5),

while the localisation effect falls disproportionately on lateral foreground sources [11]; a single field cannot spare them, because the order is global. In Case A, rendering the foreground as objects—mono signals with DoA metadata, reproduced by amplitude panning or head-related transfer function (HRTF) convolution as in the hybrid scheme of [12]—removes the foreground from the SH truncation entirely. Concretely, at the receiver each object is rendered at its available DoA—by vector-based amplitude panning (VBAP) to a loudspeaker layout, or by HRTF convolution for binaural playback in headphone-based INMP [3]—while the diffuse bed is decoded by a standard HOA decoder (e.g. max- r_E or AllRAD [5]); the two are then summed. Its localisation error is therefore governed by the accuracy of the directional metadata and rendering rather than by $\alpha_E(N)$.

Fig. 6 contrasts the two renderings for the higher-importance lateral source. For illustration, the object foreground is assigned a 5° directional uncertainty, consistent with the DoA-estimation residual observed in the DirAC analysis of Sec. V. As the budget shrinks, the single-field error grows from approximately 9.5° (order 4) to 34.2° (order 1), whereas the assumed object-direction error remains fixed as long as the object stream is retained. More generally, if ϵ_{obj} denotes the effective angular uncertainty of the object rendering, it improves localisation over the single field whenever $\epsilon_{\text{obj}} < \Delta(N(B), \phi)$. For the source at $\phi = 78^\circ$, the corresponding thresholds are approximately 34.2° , 20.1° , 13.4° , and 9.5° for orders 1–4, respectively. In Case B this directional criterion does not account for additional artefacts such as source leakage or residual interference, so the benefit remains contingent on source-separation quality.

Meanwhile the diffuse bed remains a low-order HOA stream, where order reduction is perceptually cheap [12]. This is a *hybrid object/bed* scheme; the analytical point is that the locus of benefit is the *rendering* of the foreground, not the *allocation* of HOA order.

From channels to bitrate, packetisation and loss. We adopt a representative professional-audio format, 48 kHz and 16-bit pulse-code modulation (PCM; as in NetMusic3D), solely for illustrative transport accounting. These bitrate figures are computed from the number of audio streams and the 48 kHz/16-bit format, independently of the 16 kHz signals used for the spatial analysis. One channel is 0.77 Mbit/s; Table I lists the resulting architectures. These figures assume uncompressed PCM; a perceptual codec such as Opus (used in NetMusic3D [8]) would lower the absolute rates, but could also alter the relative efficiency of the different representations. Its interaction with the object/bed split and with UEP is left to future work.

Consider, under this uncompressed PCM accounting, a link whose usable rate falls from 12.3 to ~ 5 Mbit/s: single-field HOA must drop from order 3 (16 channels) to order 1 (4 channels, 3.1 Mbit/s), increasing the energy-vector blur to $\alpha_E = 45^\circ$; a hybrid stream of a 4-channel order-1 bed plus $K = 2$ objects fits in 4.6 Mbit/s (plus negligible DoA metadata) and keeps the foreground free from SH-order truncation.

TABLE I
ILLUSTRATIVE MAPPING FROM SPATIAL-AUDIO ARCHITECTURE TO NETWORK IMPLICATIONS (UNCOMPRESSED PCM AT 48 KHz, 16 BIT: 0.77 Mbit/s PER AUDIO STREAM).

Architecture	Audio streams	Bitrate (PCM)	Network implication
Single-field HOA, order 3	16 channels	12.3 Mbit/s	high raw rate; loss affects the shared scene representation
Single-field HOA, order 1	4 channels	3.1 Mbit/s	low raw rate; isotropic spatial blur ($\alpha_E = 45^\circ$)
Hybrid object/bed	4-ch bed + K objects	$(4+K) \times 0.77$ Mbit/s	foreground separately renderable/protectable; bed can be deprioritised
Layered HOA + UEP	order layers $0..N$	order- N payload + protection overhead	graceful degradation; low-order layers can be preferentially protected

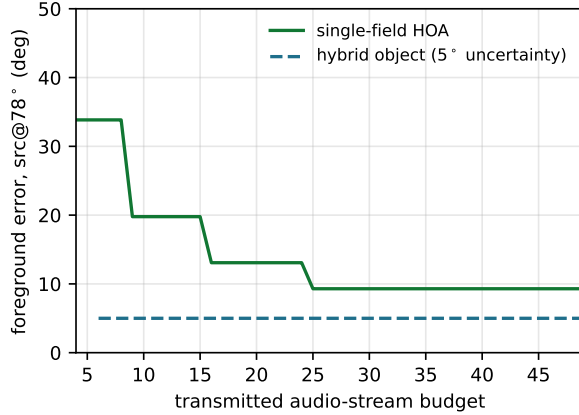


Fig. 6. Localisation error of the higher-importance lateral source ($\phi \approx 78^\circ$) vs. transmitted audio-stream budget. Single-field HOA blurs the foreground as the budget shrinks (isotropic order drop, $\Delta(N(B), \phi)$); the hybrid object foreground is shown with an assumed 5° directional uncertainty, representative of the DirAC residual observed in Sec. V. For $K = 2$ objects and a four-channel order-1 bed, the hybrid configuration is feasible from $B = 6$ audio streams.

Layered transport can further enable graceful degradation under bursty (e.g. Gilbert–Elliott) loss. Scalable HOA order truncation [10] and related layered scene-based coding approaches [17], [18] provide a basis for protecting different parts of the representation differently. Unequal error protection (UEP) can preferentially protect low-order layers and, in the hybrid case, the object streams, at the cost of additional protection overhead, so that less critical spatial information can be allowed to degrade first. These are content-aware uses of *rate and protection*, orthogonal to per-component order allocation, and are particularly relevant to loss-prone edge/5G deployments [15], [16]. They are design implications rather than results of a loss/FEC simulation in the present work.

Latency, dominated by the network path and jitter buffer rather than by the Ambisonics order [8], constrains any on-line DirAC front-end. More generally, estimating the scene parameters—by DirAC or any alternative—adds computational load and may increase the end-to-end latency of the stream, a cost that must be weighed against the benefits above and that lies outside this analysis.

VIII. IMPLICATIONS AND LIMITATIONS

The analysis yields concrete guidance for resilient INMP, together with its current limitations.

Implications.

- Per-component HOA *order* allocation through independently encoded HOA fields does not improve on a shared global-order scheme [10] at equal SH coefficient-channel budget (Prop. 1).
- In the bandwidth/loss-constrained edge/5G regime, and when stems are available (Case A), render the higher-importance foreground as *objects* to decouple its localisation accuracy from the transmitted HOA order (Fig. 6).
- Keep the diffuse field as a *low-order* bed, exploiting its early perceptual saturation [12], while allowing its coding and protection priority to be lower than that of the foreground when appropriate.
- Unequal error protection can preferentially protect low-order layers and object streams, allowing less critical spatial information to degrade first under loss. This is a complementary content-aware design dimension rather than a consequence of Prop. 1.
- When only a mixed stream is available (Case B), DirAC can supply the spatial parameters needed for decomposition, but the object-rendering benefit remains contingent on source-separation quality.

Limitations.

- The study is objective: the localisation model follows the energy-vector formulation and incidence-dependent trends reported in [11], while the diffuse-cost curve q_d is a stylised model consistent with [12]. The perceptual consequences of the proposed strategies still require subjective validation, including localisability and basic audio quality under loss.
- Prop. 1 holds for *any* scene under the assumed independently encoded per-component HOA representation and SH coefficient-channel budget. The single, first-order, *simulated* (not measured) scene only illustrates the result. The numerical size of the gap in Fig. 4, as well as generalisation of the *estimation* pipeline and its energy-based importance weights, requires measured HOA recordings and a broader range of source configurations and DRRs.
- The network account is an illustrative bridge (Table I), not a network simulation with a real loss process and forward error correction (FEC).
- The bitrate figures assume uncompressed PCM. Perceptual codecs, joint HOA coding, and content-aware quantisation or bit allocation can alter the relation between SH coefficient count and transmitted bitrate and are not modelled here.

- The aggregate distortion D (4) is a linear, additive model that omits masking and attentional interactions; it affects only the illustrative Fig. 4, not Prop. 1.
- Clean object extraction from FOA, source-separation artefacts such as leakage and residual interference, the latency of an online DirAC front-end, and integration into a live system such as NetMusic3D [8] remain open engineering questions.

IX. CONCLUSION

Motivated by two psychoacoustic results that make non-uniform Ambisonics order seem attractive, we asked whether content-aware order adaptation helps INMP, and found that it does not in the form most directly suggested: under an SH coefficient-channel budget, a single-field HOA is optimal with respect to per-source localisation distortion, so allocating different orders to independently encoded per-component HOA fields cannot beat it at equal budget. Content-awareness can instead be exploited along dimensions not covered by this result: given production-side stems, through object rendering of the foreground, and through protection and bit allocation rather than per-component order allocation. When only a mixed HOA stream is available, DirAC can estimate the spatial parameters needed for decomposition, while obtaining clean object signals additionally requires source separation. Subjective validation and integration of these strategies into a live INMP system are the natural next steps.

REFERENCES

- [1] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
- [2] L. Turchet and M. Tomasetti, "Immersive networked music performance systems: Identifying latency factors," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2023, pp. 1–6.
- [3] M. Tomasetti and L. Turchet, "Playing with others using headphones: Musicians prefer binaural audio with head tracking over stereo," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 501–511, 2023.
- [4] C. Rinaldi, C. Centofanti, L. Comanducci, M. Pezzoli, A. Bernardini, A. Marotta, and F. Antonacci, "Immersive acoustics via next-generation networks: Achieving high-fidelity audio in the metaverse," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 7899–7919, 2025.
- [5] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, 2019.
- [6] B. Rafaely, V. Tourbabin, E. Habets, Z. Ben-Hur, H. Lee, H. Gamper, L. Arbel, L. Birnie, T. Abhayapala, and P. Samarasinghe, "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges," *Acta Acustica*, vol. 6, p. 47, 2022.
- [7] S. Giacomelli, C. Centofanti, J. Santos, M. Galbiati, T. Salvi, F. Graziosi, and C. Rinaldi, "Remote immersive audio production: State of the art implementation, challenges, and improvements," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–10.
- [8] F. Miotello, P. Ostan, G. Elli, C. Centofanti, M. Pezzoli, A. Bernardini, C. Rinaldi, and F. Antonacci, "Latency and quality of experience in immersive networked music performances using higher-order Ambisonics spatial audio," in *2025 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2025.
- [9] A. I. Mezza, A. Bernardini, and C. Rinaldi, "The ieee-is2 2025 music packet loss concealment challenge," in *2025 IEEE 6th International Symposium on the Internet of Sounds (IS2)*, 2025, pp. 1–10.
- [10] P. Ostan, C. Centofanti, M. Pezzoli, A. Bernardini, C. Rinaldi, and F. Antonacci, "Dynamic real-time Ambisonics order adaptation for immersive networked music performances," in *2025 33rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2025, arXiv:2508.00509.
- [11] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.
- [12] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual implications of different Ambisonics-based methods for binaural reverberation," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 895–910, 2021.
- [13] B. Bernschütz, "Microphone arrays and sound field decomposition for dynamic binaural recording," Ph.D. dissertation, Technische Universität Berlin, 2016.
- [14] J. Ahrens and C. Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2783–2794, 2019.
- [15] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, 2024.
- [16] F. Martusciello, C. Centofanti, C. Rinaldi, and A. Marotta, "Edge-enabled spatial audio service: Implementation and performance analysis on a MEC 5G infrastructure," in *2023 4th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2023, pp. 1–8.
- [17] N. Peters, D. Sen, M.-Y. Kim, O. Wuebbolt, and S. M. Weiss, "Scene-based audio implemented with higher order ambisonics," *SMPTE Motion Imaging Journal*, vol. 125, no. 9, pp. 16–24, 2016.
- [18] D. Sen, N. Peters, M.-Y. Kim, and M. Morrell, "Efficient compression and transportation of scene-based audio for television broadcast," in *2016 AES International Conference on Sound Field Control*. Audio Engineering Society, 2016, aES E-Library, paper 18329.
- [19] S. Zamani, T. Nanjundaswamy, and K. Rose, "Frequency domain singular value decomposition for efficient spatial audio coding," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 126–130.
- [20] C. Hold, V. Pulkki, A. Politis, and L. McCormack, "Compression of higher-order ambisonic signals using directional audio coding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 651–665, 2024.
- [21] A. I. Mezza, M. Amerena, A. Bernardini, and A. Sarti, "Hybrid packet loss concealment for real-time networked music applications," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 266–273, 2024.
- [22] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2711–2721, 2013.
- [23] C. Schissler, P. Stirling, and R. Mehra, "Efficient construction of the spatial room impulse response," in *2017 IEEE Virtual Reality (VR)*. IEEE, 2017, pp. 122–130.
- [24] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.
- [25] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *2012 46th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2012, pp. 1–6.
- [26] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2023.
- [27] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: 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.
- [28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.