

SHAPE: An Architecture for 6DoF Spatial Audio in Networked Hybrid XR Performances

Gianluca Fadda
Università di Cagliari
Cagliari, Italy
gianluca.fadda@unica.it

Alessia Palladino
Università di Cagliari
Cagliari, Italy
alessia.palladino@unica.it

Massimo Farina
Università di Cagliari
Cagliari, Italy
massimo.farina@unica.it

Maurizio Murrone
Università di Cagliari
Cagliari, Italy
maurizio.murrone@unica.it

Nicutor Amarie
Transilvania University of Braşov
Braşov, Romania
nicutor.amarie@unitbv.ro

Vlad Popescu
Transilvania University of Braşov
Braşov, Romania
vlad.popescu@unitbv.ro

Anderson Augusto Simiscuka
Dublin City University
Dublin, Ireland
andersonaugusto.simiscuka@dcu.ie

Abstract—Live hybrid and remote-attendance XR performances require spatial audio that reacts to listener translation as well as head rotation. SHAPE (Spatial Hybrid Audio Pipeline for XR Events) is an architecture that keeps three audio layers separate from capture to binaural output: object-based performer feeds, a distributed Ambisonic audience/venue bed, and a measured room-response layer. The separation gives each layer its own quality, latency, bandwidth, and privacy requirements and yields explicit rules; e.g., room rendering is applied only to performer objects, since the captured ambience already contains the room. We specify a baseline rendering method for each layer of a 6DoF scene, with position-dependent node selection for the ambience, a latency model separating local, network, and rendering timing domains, and privacy requirements mapped to implementation decisions. Engine-native multichannel I/O and single-node first-order Ambisonic encoding are implemented. Six measured room responses provide broadband decay data, and an evaluation protocol defines how the claimed advantages will be tested. SHAPE is intended as a reusable framework for small and medium cultural venues.

Index Terms—networked immersive audio, 6DoF spatial audio, Ambisonics, object-based audio, room impulse response, hybrid XR performance, binaural rendering, privacy-by-design

I. INTRODUCTION

Live hybrid and remote-attendance extended reality (XR) performances are moving from conventional audio-video streaming towards environments where remote participants can explore a virtual representation of the venue, so the audio scene must react to head rotation and listener translation. This is especially relevant for musical performances, where the perceived position of performers, audience reactions, and room acoustics all contribute to the experience. Previous audience studies on remote music events report reduced sensory richness compared with physical attendance and identify deeper immersion as an unmet need for future platforms [1], [2]. Several parts of this problem have already been addressed in the literature. Object-based audio supports independent rendering of individual sources [3], while six-degrees-of-freedom (6DoF) methods extend spatial reproduction to listener translation [4], [5]. SHAPE therefore does not claim novelty for 6DoF rendering, Ambisonics, object-

based audio, or auralisation from measured rooms considered separately. The aim of SHAPE is instead to organise these elements for networked hybrid performances while keeping three audio components separate: performer objects, the live audience/venue ambience, and the measured acoustic response of the venue. Each component can then be captured, transported, rendered, and protected according to its own requirements. SHAPE is developed within the EU Horizon Europe project HEAT (Hybrid Extended reAliTy), which integrates holographic imaging and multi-sensory media into distributed hybrid XR experiences [6]. The contributions are the SHAPE layer architecture with per-layer requirements and routing rules, a baseline rendering specification for each layer including a hybrid geometric/measured room layer, a latency model separating local, network, and rendering domains, privacy-by-design objectives mapped to implementation decisions, and a room-response analysis informing the room-layer design.

II. RELATED WORK AND POSITIONING

Networked and immersive musical performance has a long history. Early work established immersive audio-visual environments such as CAVE [7], while Remote Media Immersion and Distributed Immersive Performance explored geographically distributed performance over high-bandwidth research networks [8]–[10]. Ambisonic spatialisation was later used in networked musical-performance systems [11]. More recent work spans web-based performance environments, telematic audio for online jamming, low-latency networking, interoperable spatial-audio rendering, and immersive media [12]–[19]. Within HEAT, earlier systems streamed per-instrument mixer tracks spatialised in Unity at performer positions from RGB-D capture, returning audience audio to the musicians [20], [21].

Listener translation and measured-room auralisation are also established research topics. Cairns et al. already demonstrate server-side 6DoF auralisation with interpolation of measured spatial impulse responses, although their wireless endpoint missed its motion-to-sound target [22]. Stefani et al. demonstrate efficient real-time 6DoF convolution of mea-

sured responses, with response selection and crossfading [23]. McCormack et al. render tracked sound objects and residual ambience from multiple distributed Ambisonic receivers according to listener position [24]. Politis et al. interpolated spatial parameters of multi-point first- and higher-order Ambisonic (FOA/HOA) recordings instead of mixing spatially separated microphones, reducing coloration and ambiguous cues [25]. SHAPE's contribution is a production-oriented organisation: performer objects come directly from the live production chain, a separate distributed Ambisonic layer carries the audience and venue ambience, and a third layer supplies venue-specific acoustics from measured room impulse responses (RIRs). Separate layers can follow their own capture, transport, processing, and privacy rules; in particular, a dry performer feed can receive a measured room response without reverberating the audience recording twice.

III. SHAPE ARCHITECTURE AND IMPLEMENTATION

A. Layers and Implemented Components

SHAPE separates the audio scene into three layers (Fig. 1). Performer signals from close and spot microphones, whose capture has been tested, remain individually addressable objects. The ambience layer is designed around N Ambisonic capture nodes distributed in the audience area. Venue acoustics are represented by RIRs acquired before the event. Two components are implemented in Unity and run in real time, but separately, not yet as one live chain or with the renderer. An engine-native I/O bridge exposes a professional ASIO interface inside the engine audio thread: driver callbacks write into wait-free single-producer/single-consumer (SPSC) ring buffers, and an asynchronous sample-rate converter (ASRC) driven by a proportional-integral fill-level controller reconciles the hardware and engine clock domains. It also provides direct input-to-output monitoring and diagnostics (levels, under/overruns, buffer fill, resampling correction). The capture-node encoder converts the four capsule signals to first-order AmbiX (ACN ordering, Schmidt semi-normalised SN3D) inside the engine audio callback, using an engine-independent core; its mapping between Unity's left-handed frame and the right-handed Ambisonic frame has been verified with reference inputs. For ideal cardioid capsules on a regular tetrahedron, the encoder computes $W = \frac{1}{2} \sum_i A_i$ and $\{X, Y, Z\} = \mp \frac{\sqrt{3}}{2} \{X, Y, Z\}_{\text{raw}}$, with $X_{\text{raw}} = A_{\text{FLU}} + A_{\text{FRD}} - A_{\text{BLD}} - A_{\text{BRU}}$ and sign patterns $(+, -, +, -)$ and $(+, -, -, +)$ for Y_{raw} and Z_{raw} , output in ACN order (W, Y, Z, X) ; the sign accounts for inward-facing capsules [26]. The current node, four Behringer C-2 cardioids without non-coincidence correction, calibration, or measured polar response, has a capsule spacing that lowers the upper limit of valid FOA encoding. The encoder has been run on live input from this array, captured without the bridge.

B. Baseline 6DoF Rendering Specification

Let the listener pose be position $\mathbf{l}$ and orientation $\mathbf{R} \in SO(3)$, mapping head to world coordinates. For performer object i at $\mathbf{p}_i$ (from performer tracking, e.g., RGB-D capture [20], or a fixed stage layout), let $d_i = \|\mathbf{p}_i - \mathbf{l}\|$ and $r_i =$

$\max(d_i, r_{\min})$ with $r_{\min} > 0$. The head-relative direction is $\mathbf{u}_i = \mathbf{R}^T(\mathbf{p}_i - \mathbf{l})/d_i$ for $d_i > 0$, holding its last value at $d_i = 0$. The direct path is rendered with the head-related transfer function (HRTF) pair interpolated at $\mathbf{u}_i$, inverse-distance gain $g_i = \min(1, r_0/r_i)$ with reference distance r_0 , and propagation delay $\tau_i = r_i/c$ realised as an interpolated fractional delay (yielding Doppler under motion). For the ambience layer, each node signal $\mathbf{b}_n$, captured at $\mathbf{q}_n$, is first rotated into the world frame. The active node a changes to $b = \arg \min_n \|\mathbf{q}_n - \mathbf{l}\|$ only when $\|\mathbf{q}_b - \mathbf{l}\| < \|\mathbf{q}_a - \mathbf{l}\| - \delta$, with hysteresis margin δ ; both are level-matched and delay-aligned using the measured inter-node offset (Section IV-B) and crossfaded jointly on all four channels as

$$\mathbf{b}(t) = \cos\left(\frac{\pi}{2}\phi(t)\right) \mathbf{b}_a(t) + \sin\left(\frac{\pi}{2}\phi(t)\right) \mathbf{b}_b(t), \quad (1)$$

with $\phi(t)$ rising linearly from 0 to 1 over T_x . This law preserves power only for uncorrelated signals: coherent components can rise by up to 3 dB or cancel at mid-transition, and residual delays cause comb filtering. The ambience therefore selects nodes by listener position: it responds only to rotation (3DoF) within a node's region, and translation is conveyed by node switching. The result is rotated by $\mathbf{R}^T$ in the spherical-harmonic (SH) domain and decoded binaurally. FOA is chosen for cost, not as a perceptual optimum: order affects localisation and quality [27]–[29], needs $(\ell + 1)^2$ channels at order ℓ , and can be adapted to the network [30].

C. Room Layer

The planned room layer follows a hybrid scheme. The direct path belongs to the object layer. Above a room-dependent crossover frequency, early reflections are simulated geometrically on an acoustic proxy mesh of the venue, and a measured late tail, shared across listener positions, is rendered through partitioned convolution [23]. The parts are joined around a split time t_s , checked against the room's mixing time [31], with complementary windows and level matching. Below the room's Schroeder transition frequency, where ray-based geometry becomes unreliable [32], the measured response is used from its onset with its direct-sound segment windowed out, and a band-limited crossover avoids double counting. With omnidirectional responses, the binaural tail provisionally uses two decorrelated copies matched to the diffuse-field interaural coherence of the HRTF set. Room rendering is applied only to performer objects: the captured ambience bed already contains the venue acoustics, and processing it again would count the room twice.

D. Transport and Latency Model

The I/O bridge has been validated with a LAN WebRTC transport prototype that uses software clock-offset estimation and whose receiver applies its own spatialisation; its performance for SHAPE's layer streams is not yet characterised. Within a managed local or metro segment, multi-node capture requires bounded jitter and clock alignment; a PTP-clocked AES67 segment [33], [34] is the intended implementation. Wide-area delivery relies on buffering and resynchronisation

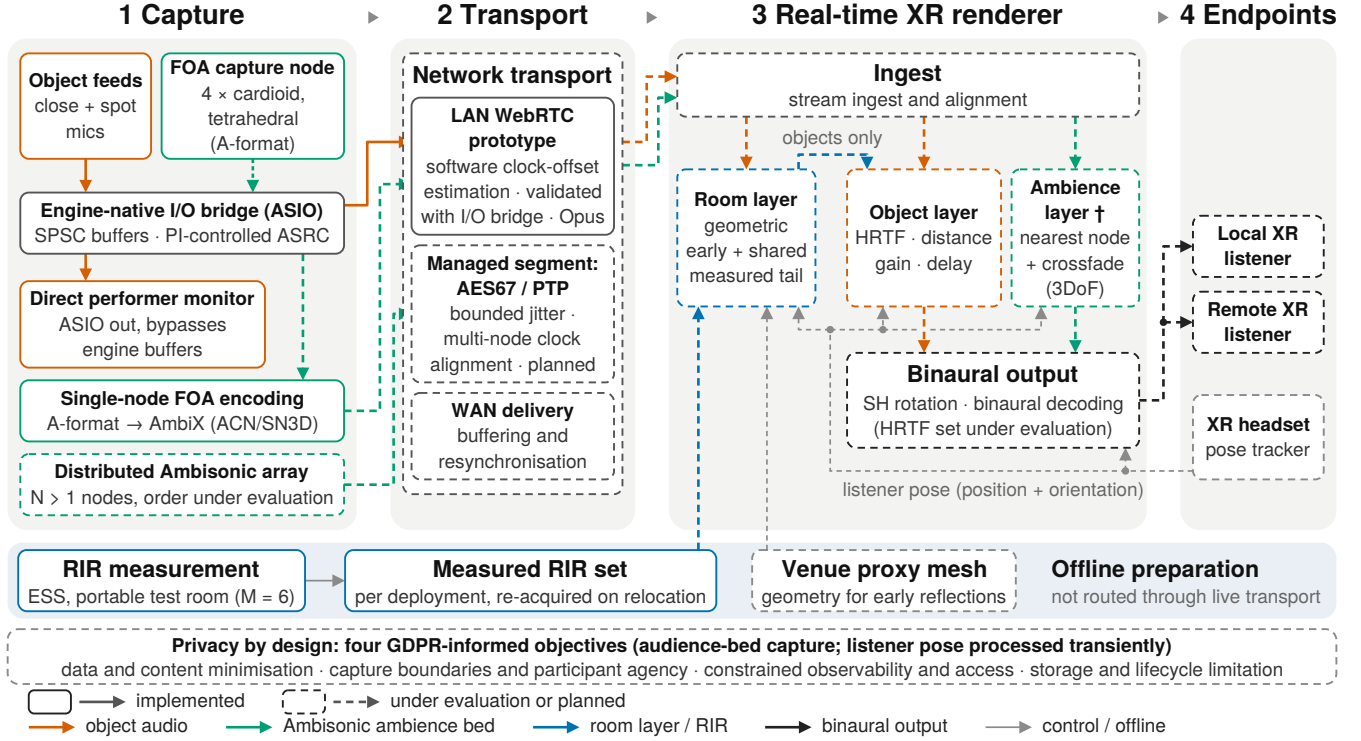


Fig. 1. SHAPE architecture and development status. Performer objects, the Ambisonic ambience bed, and measured room responses remain separate layers; room rendering applies to performer objects only. Implemented (solid): performer capture, the engine-native I/O bridge (validated with a LAN WebRTC prototype), single-node FOA encoding, and offline RIR measurement, with FOA node, bridge, and encoder not yet one live chain. Planned or under evaluation (dashed): multi-node capture, AES67/PTP, the venue proxy mesh, and the integrated renderer. †: node combination under evaluation (Sec. III-B).

rather than sample-accurate alignment. Audio-path latency is modelled as

$$L_{e2e} \approx \underbrace{L_{\text{ASIO}} + L_{\text{ring}} + L_{\text{ASRC}} + L_{\text{eng}} + L_{\text{enc}}}_{\text{local I/O}} + L_{\text{net}} + L_{\text{jib}} + L_{\text{dec}} + L_{\text{rend}} + L_{\text{out}}, \quad (2)$$

where the terms are the ASIO input buffer, ring buffer, ASRC FIFO, engine DSP buffering, encoding and codec framing (at least one 20 ms frame plus look-ahead), one-way network delay, jitter buffer, decoding, rendering block, and output buffering. With the current settings, one 512-sample ASIO buffer at 48 kHz is 10.7 ms, the ASRC target fill is 80 ms, and 4×256 engine samples are 21.3 ms: a nominal allocation of 112 ms, excluding L_{ring} and converter delays, not a measured latency. Listener motion follows a separate path, $L_{m2s} \approx L_{\text{pose}} + L_{\text{rend}} + L_{\text{out}}$, when rendering runs at the listener's endpoint; server-side rendering would add the network, jitter-buffer, and codec terms [22]. Direct performer monitoring bypasses the engine buffers but still adds 21.3 ms of ASIO input and output buffering at 512 samples, so it needs smaller buffers to stay below ensemble thresholds [35]; audience delivery must stay within audio–visual synchronisation limits, and streams are to be aligned at ingest by timestamps and clock offsets. Opus at 64 kb/s per channel (Ambisonic channel mapping family 3) was not significantly different from uncompressed Ambisonic audio in basic audio quality in the tested contexts [29]; with 16 kb/s of

RTP/UDP/IPv4 headers per stream at 20 ms framing (excluding SRTP and header extensions), objects as separate streams, and each FOA node as one four-channel stream, K objects and N nodes need about $(K + 4N) \times 64 + (K + N) \times 16$ kb/s, while the offline room layer adds none.

IV. MEASUREMENTS, EVALUATION, AND PRIVACY-BY-DESIGN

A. Room-Response Measurements

Measurements were carried out in the project's portable test room: a $3 \times 3 \times 3$ m frame ($V = 27 \text{ m}^3$) lined with black curtains, without a rigid shell, installed in a laboratory room. Room EQ Wizard generated an exponential sine sweep (ESS; 2^{20} samples, 21.8 s at 48 kHz, 20 Hz–20 kHz, -12 dBFS RMS, one repetition) [36]. Responses were deconvolved, aligned with an acoustic timing reference, and truncated with a fixed 1.0 s window. The source was a directional two-way loudspeaker (RCF ART 912-A, $100^\circ \times 60^\circ$ nominal coverage) and the receiver a large-diaphragm studio condenser microphone (Behringer B-2 PRO) set to its omnidirectional pattern. Following the ISO 3382-2 layout [37], two source configurations (triangular corner, asymmetric wall position) were combined with three receiver positions each ($M = 6$), with receiver spacing constrained by the room size. T_{20} and T_{30} are linear fits over -5 to -25 and -35 dB of the Schroeder decay curve. As neither transducer meets the ISO 3382-2 assumptions and the analysis

TABLE I
BROADBAND DECAY PARAMETERS (S) FOR THE $M = 6$ SOURCE-RECEIVER COMBINATIONS.

Receiver	Setup A (corner)			Setup B (wall)		
	1	2	3	4	5	6
EDT	0.842	0.911	0.936	0.861	0.916	0.861
T_{20}	1.104	1.063	1.075	0.964	1.023	1.025
T_{30}	1.118	1.081	1.095	1.037	1.077	1.066

is broadband, the results characterise the acquisition workflow and this configuration’s broadband decay, not an ISO-compliant room characterisation. T_{30} averaged 1.079 s (SD 0.027 s, CV 2.5%), within $\pm 4\%$ across the six source–receiver combinations. The early decay time (EDT) was lower at every position (mean 0.888 s; $\text{EDT}/T_{30} = 0.82$) and showed a larger spread (CV 4.3%), and T_{30} exceeded T_{20} at every position (curvature $C = 100 (T_{30}/T_{20} - 1)$ between 1.3% and 7.6%). Within these six positions, the small spread of broadband late decay motivates testing a shared tail (Section III-C), not perceptual interchangeability. As the curtained frame has no rigid shell, its low-frequency response is expected to depend on the host room, so measured tails are re-acquired whenever the setup is relocated.

B. Evaluation Protocol and Current Evidence

Apart from the six broadband RIR estimates in Table I, the figures in this paper are analytical; no test of SHAPE under listener translation (an earlier XRBlues event used a standard Unity spatial-audio plugin [21]) or comparative network measurement has been performed. The next component-level experiments will (i) measure bridge round-trip latency with an audio loopback; (ii) record a shared impulse at all FOA nodes to measure inter-node offset and drift on a managed LAN; and (iii) profile renderer load against object count, node count, and RIR length. An integrated comparison will then replay the same dry objects, synchronised FOA nodes, and tracked listener path through SHAPE and two baselines, single-node FOA and multipoint interpolation or measured IRs [22], [25], reporting motion-to-sound latency, coloration at node transitions, quality, presence, and cybersickness ratings following the project’s QoE protocol [38], network rate, and computation.

C. Privacy-by-Design

Ambient audience-bed capture raises specific privacy challenges, because acoustic signals may be acquired simultaneously from multiple and dynamically changing attendees. Privacy-by-design has been identified as an open challenge for the Musical Metaverse [39], and research on immersive environments has highlighted the limits of purely procedural safeguards [40]; audience audio was already captured and re-spatialised in an earlier XRBlues event [21]. Listener pose streams are also personal data, since position and movement can enable indirect identification [41], and are processed transiently for rendering only. SHAPE therefore treats privacy as a design constraint rather than a post-hoc compliance layer, translating relevant GDPR principles [42] into four design objectives (Table II). The objectives guide design but do not

TABLE II
PRIVACY-BY-DESIGN OBJECTIVES AND PROPOSED SHAPE MEASURES.

Objective	Measures
Data and content minimisation (Art. 5(1)(c))	Acquisition limited to what ambience rendering requires; auxiliary recording paths disabled by default; raw audience signals kept transient.
Capture boundaries and participant agency	Capture limited to required areas, excluding back-of-house, transit, and rest areas, with boundaries assessed acoustically; attendees informed of capture; where consent is the lawful basis, it must be valid, not replaced by notices or passive opt-out.
Constrained observability and access (Arts. 5(1)(f), 25, 32)	Routine monitoring via telemetry (levels, synchronisation, packet loss, clock drift) rather than playback; access to intelligible content only when technically necessary, with risk-based safeguards (privileges, authentication, logging, time-limited diagnostic access).
Storage and lifecycle limitation (Art. 5(1)(b), (e))	Transient buffers preferred; retention only for a predefined purpose and duration; processing data, telemetry, diagnostics, research datasets, and archives kept as separate categories.

by themselves establish GDPR compliance, which depends on each deployment’s configuration, purpose, legal basis, and risk profile; the framework remains a design-oriented model pending validation with an audience present. FOA encoding is not an anonymisation mechanism, since encoded streams may still contain intelligible speech. Performer feeds need their own purpose, access, and retention rules; the implemented I/O bridge already exposes the indicators needed for monitoring through telemetry. A pre-deployment review covers (i) capture and non-capture areas, (ii) transparency and, where applicable, consent, (iii) unnecessary recording paths, (iv) signal persistence, (v) retention and deletion, and (vi) access privileges, with security assessed on a risk basis. Applied preliminarily to the empty-room measurement, it found the attendee items not applicable but retention and access responsibilities unspecified in advance, showing how the framework exposes governance gaps before a deployment with an audience.

V. DISCUSSION AND CONCLUSION

SHAPE defines a reusable framework for networked hybrid XR performances in which performer objects, ambience, and measured venue acoustics remain separate from capture to binaural output, with per-layer requirements and explicit rules such as applying room rendering to performer objects only. Its interfaces and baseline methods allow alternatives to be compared within it and the architecture to be transferred to other venues; the HEAT XRBlues pilot [43] provides the setting for the evaluation protocol of Section IV-B.

ACKNOWLEDGMENT

The research activities described in this paper were carried out within the “HEAT – Hybrid Extended reAliTy” Project GA 101135637 funded by the EU Horizon Europe Framework Programme (HORIZON) and supported by a grant of the Romanian Ministry of Research, Innovation and Digitization, CNCS/CCCDJ - UEFISCDI, project number PN-IV-P8-8.1-PRE-HE-ORG-2025-0331, within PNCDI IV.

REFERENCES

- [1] S. Ppali, M. Scorer, E. Ppali, B. Branch, and A. Covaci, "Remote rhythms: Audience-informed insights for designing remote music performances," in *Proc. 2024 ACM Designing Interactive Systems Conf. (DIS '24)*. ACM, 2024, pp. 2675–2690.
- [2] S. Ppali *et al.*, "The virtual concert-goer: Audience perspectives on remote music performances," *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. 7, pp. 1–30, 2025, article CSCW250.
- [3] P. Coleman *et al.*, "An audio-visual system for object-based audio: From recording to listening," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 1919–1931, 2018.
- [4] B. Rafaely *et al.*, "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges," *Acta Acustica*, vol. 6, p. 47, 2022.
- [5] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information," in *AES Int. Conf. on Audio for Virtual and Augmented Reality*, 2018, aES E-Library ID 19684.
- [6] M. Murrioni, G.-M. Muntean, M. Montagud, and V. Popescu, "Toward interactive and distributed hybrid extended reality experiences," in *Proc. 2024 ACM Int. Conf. on Interactive Media Experiences Workshops (IMXw)*, 2024, pp. 127–130.
- [7] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The cave: Audio visual experience automatic virtual environment," *Commun. ACM*, vol. 35, no. 6, pp. 64–72, 1992.
- [8] R. Zimmermann, C. Kyriakakis, C. Shahabi, C. Papadopoulos, A. A. Sawchuk, and U. Neumann, "The remote media immersion system: Internet meets the future home theater," *IEEE MultiMedia*, vol. 11, no. 2, pp. 48–57, 2004.
- [9] E. Chew *et al.*, "Musical interaction at a distance: Distributed immersive performance," in *Proc. 4th Open Workshop of MUSICNETWORK: Integration of Music in Multimedia Applications*, 2004.
- [10] A. A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," in *Proc. ACM SIGMM 2003 Workshop on Experiential Telepresence (ETP '03)*. ACM, 2003, pp. 110–120.
- [11] M. Gurevich, D. Donohoe, and S. Bertet, "Ambisonic spatialization for networked music performance," in *Proc. 17th Int. Conf. on Auditory Display (ICAD 2011)*, June 2011, pp. 1–8.
- [12] D. Dziwis, H. von Coler, and C. Pörschmann, "Orchestra: A toolbox for live music performances in a web-based metaverse," *J. Audio Eng. Soc.*, vol. 71, no. 11, pp. 802–812, November 2023.
- [13] P. Müller, R. Häfeli, J. M. Schütt, and M. Ziegler, "Telemersive audio systems for online jamming," in *2024 IEEE 5th Int. Symp. on the Internet of Sounds (IS2)*, 2024, pp. 1–15.
- [14] 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 J. Commun. Soc.*, vol. 5, pp. 4691–4709, 2024.
- [15] C. Rinaldi *et al.*, "Immersive acoustics via next-generation networks: Achieving high-fidelity audio in the metaverse," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 7899–7919, 2025.
- [16] A. F. Genovese, Z. Nguyen, M. Gospodarek, R. Pahle, C. Brenner, and A. Roginska, "Holodeck: A research framework for distributed multimedia concert performances," in *2024 IEEE 5th Int. Symp. on the Internet of Sounds (IS2)*, 2024, pp. 1–10.
- [17] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally adapted immersive environments for distributed music performances in mixed reality," in *2024 IEEE 5th Int. Symp. on the Internet of Sounds (IS2)*, 2024, pp. 1–10.
- [18] R. Hupke, S. Preihs, and J. Peissig, "Immersive networked music performance: Impact of extended reality on the quality of experience," in *2024 IEEE 5th Int. Symp. on the Internet of Sounds (IS2)*, 2024, pp. 1–9.
- [19] J.-M. Jot, R. Audfray, M. Hertensteiner, and B. Schmidt, "Rendering spatial sound for interoperable experiences in the Audio Metaverse," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021, pp. 1–15.
- [20] N. Amarie *et al.*, "Immersive application for real-time interactive music performances using spatial audio," in *2025 IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2025.
- [21] N. Amarie, G. Fadda, V. Popescu, J.-M. Ghenta, M. Murrioni, and A. A. Simiscuka, "Live feedback for immersive music performances – a case study," in *Proc. ACM Int. Conf. on Interactive Media Experiences Workshops (IMXw)*. SBC, 2025, pp. 26–29.
- [22] P. Cairns, H. Daffern, and G. Kearney, "Investigation of server-based spatial audio for metaverse concert distribution," in *2024 IEEE 5th Int. Symp. on the Internet of Sounds (IS2)*, 2024, pp. 1–8.
- [23] D. Stefani, M. Binelli, A. Farina, and L. Turchet, "Improved real-time six-degrees-of-freedom dynamic auralization through nonuniformly partitioned convolution," *J. Audio Eng. Soc.*, vol. 73, no. 10, pp. 671–681, 2025.
- [24] L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, "Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple ambisonic receivers," *J. Audio Eng. Soc.*, vol. 70, no. 5, pp. 355–372, May 2022.
- [25] A. Politis, L. Pajunen, J. Leppanen, S. Mate, and A. Eronen, "Wide-area 6dof rendering of multi-point ambisonic recordings based on interpolation of spatial parameters," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023.
- [26] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, ser. Springer Topics in Signal Processing. Springer, 2019, vol. 19.
- [27] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney, "Auditory localization in low-bitrate compressed ambisonic scenes," *Applied Sciences*, vol. 9, no. 13, p. 2618, 2019.
- [28] B. Mróz and P. Kosior, "Comparison of spatial sound recording techniques with usage of ambisonics and object-based audio," *Int. J. Electron. Telecommun.*, vol. 71, no. 4, pp. 1–6, 2025.
- [29] B. Lee, T. Rudzki, J. Skoglund, and G. Kearney, "Context-based evaluation of the opus audio codec for spatial audio content in virtual reality," *J. Audio Eng. Soc.*, vol. 71, no. 4, pp. 145–154, April 2023.
- [30] 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 Conf. (EUSIPCO)*. IEEE, 2025, pp. 1258–1262.
- [31] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model- and signal-based predictors of the mixing time in Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898, 2012. [Online]. Available: <https://www.aes.org/e-lib/browse.cfm?elib=16633>
- [32] H. Kuttruff, *Room Acoustics*, 6th ed. CRC Press, 2016.
- [33] IEEE Standards Association, "IEEE standard for a precision clock synchronization protocol for networked measurement and control systems (IEEE std 1588-2019)," IEEE, Standard, 2020.
- [34] Audio Engineering Society, "AES standard for audio applications of networks – high-performance streaming audio-over-IP interoperability," Audio Engineering Society, Standard AES67-2023, 2023. [Online]. Available: <https://aes.org/publications/standards-store/>
- [35] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
- [36] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Convention of the Audio Engineering Society*, February 2000, convention Paper 5093.
- [37] International Organization for Standardization, "ISO 3382-2:2008 acoustics — measurement of room acoustic parameters — part 2: Reverberation time in ordinary rooms," ISO, Tech. Rep., 2008.
- [38] S. Baldoni, G. Fadda, J. Wu, M. Murrioni, I. Ghergulescu, and F. Battisti, "Beyond the stage: A QoE protocol for immersive XR theater experiences," in *Proc. IEEE Conf. on Telepresence*, 2026.
- [39] L. Turchet, "Musical metaverse: Vision, opportunities, and challenges," *Pers. Ubiquitous Comput.*, vol. 27, no. 5, pp. 1811–1827, 2023.
- [40] E. Hine, I. N. Rezende, H. Roberts, D. Wong, M. Taddeo, and L. Floridi, "Safety and privacy in immersive extended reality: An analysis and policy recommendations," *Digital Society*, vol. 3, no. 2, pp. 1–40, 2024.
- [41] G. Fadda *et al.*, "EnviLoop: A privacy-aware edge-assisted IoT–XR multimedia service architecture for smart venues," in *2026 IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2026.
- [42] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," 2016, official Journal of the European Union, L 119, 4 May 2016.
- [43] A. A. Simiscuka, G. Fadda, V. Popescu, M. Murrioni, and G.-M. Muntean, "Immersive live concert: A multi-sensory experience based on real-time lyrics detection from spatial audio data," in *Proc. ACM Int. Conf. on Interactive Media Experiences Workshops (IMXw)*. SBC, 2025, pp. 7–11.