Beyond Marketing: a Holistic Analysis of 3D Audio Workflows for Live Music Production

Claudia Rinaldi
CNIT, Research Unit University of L'Aquila
L'Aquila, Italy
claudia.rinaldi@univaq.it

Stefano Giacomelli
DISIM, University of L'Aquila
L'Aquila, Italy
stefano.giacomelli@graduate.univaq.it

Tiziano Salvi
Suonovivo Tecnologia
Bergamo, Italy
tiziano@suonovivo.com

Mauro Galbiati
Independent
Bergamo, Italy
thecomm@jailsound.com

Carlo Centofanti
DISIM, University of L'Aquila
L'Aquila, Italy
carlo.centofanti1@univaq.it

Fabio Graziosi
DISIM, University of L'Aquila
L'Aquila, Italy
fabio.graziosi@univaq.it

***Abstract*—This paper investigates remote spatial audio production by analyzing alternative capture and rendering workflows for live and recorded events. Three microphone configurations are addressed: a compact cardioid array, a diffuse omnidirectional array, and a second-order Ambisonics microphone. Each setup is combined with spatialization strategies including Vector-Based Panning, Ambisonics Equivalent Panning, and KNN interpolation, resulting in a set of workflows that differ in terms of spatial resolution, envelopment, and computational complexity. The framework integrates commercial tools such as SPAT Revolution and MACH1 with open-source solutions including SSR and IEM, ensuring interoperability and replicability across different production contexts. From a networking perspective, the system relies on Dante with PTP-based clock synchronization and VPN-secured transport, providing reliable and low-latency multichannel distribution across geographically distributed sites. The comparative description of these workflows highlights trade-offs in terms of channel mapping, processing load, and rendering flexibility, offering concrete guidelines for system design. While formal perceptual evaluations are planned as future work, the present contribution already provides a systematic technical basis for assessing remote spatial audio pipelines. The proposed workflow is demonstrated through a case study applicable to networked music performances and large-scale public events.**

I. Introduction

In recent years, a growing body of research has converged on the interplay between music, emerging technologies, and networked infrastructures. The concepts of the *Internet of Musical Things* (IoMusT) [1], the broader *Internet of Sounds* (IoS) [2], and the evolution of *Networked Music Performance* (NMP) systems [3], in turn enabled by innovative communication architectures [4], have contributed to redefining the paradigms of music production, experience, and education. These frameworks envision musical instruments, smart devices, and spatial audio interfaces as interconnected nodes within pervasive networks, enabling real-time, multisensory, and often immersive interactions across physical distances.

Such advances open novel opportunities for designing both remote and co-located musical experiences that go beyond mere audio reproduction. The integration of Extended Reality (XR), spatial audio, haptics, and intelligent systems is progressively shaping immersive concert formats and rehearsal practices, both in professional and amateur contexts [5]–[10]. In particular, recent studies have demonstrated that musicians performing through the use of spatial audio systems report significantly higher levels of immersion, spatial awareness, and musical connection compared to conventional stereo setups [11]. Despite this technological momentum and the availability of sophisticated spatial audio solutions, from Ambisonics to object-based rendering [12], [13], a significant gap remains between research developments and field adoption. In practice, spatial audio workflows are still often implemented by technical personnel (sound engineers, studio producers, audio operators) whose expertise and training are predominantly rooted in stereo paradigms. This mismatch leads to a fragile implementation culture where spatial tools are used without a coherent acoustic strategy, and where production decisions are influenced more by commercial trends than by an informed understanding of perceptual and technological implications.

To address this gap, this paper provides a structured and critical overview of spatial audio workflows for live and recorded music streaming. Our goal is to equip practitioners with a decision-making framework that supports methodologically informed choices aligned with artistic and technical objectives. Building upon previous works [6], which demonstrated the computational efficiency and practical accessibility of Spatial PCM Sampling (SPS) over Ambisonics in real-time networked performances, we extend the comparative analysis by evaluating two additional lightweight rendering strategies: Ambisonics Equivalent Panning (AEP) and K-Nearest Neighbors (KNN) interpolation. Furthermore, we compare three immersive microphone acquisition approaches: a *compact coincident* array, a *native* second-order Ambisonics microphone, and a

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") and by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY - CUP E13C22001060006.

custom sparse array. The proposed framework also introduces a preliminary integration of synchronized spatial audio and video routing over DANTE, enabling scalable multi-room configurations. The evaluation focuses on implementation accessibility, computational and communication efficiency, and real-time feasibility across long-range music performances. While perceptual fidelity remains beyond the scope of this paper, this aspect will be explored in an accompanying demonstration.

II. Background

The design of immersive auditory experiences relies on the accurate capture and reproduction of spatial sound fields because the perceived positioning and motion of sound sources are key to realism and engagement [14], [15].

Several approaches have been developed for spatial sound capture. Traditional stereo and surround microphone techniques can be based on coincident, near-coincident, or spaced configurations for offering basic spatial cues; they are widely employed in both live and studio contexts [16], [17]. Microphone arrays, ranging from linear to spherical geometries, allow more precise directional analysis and sound field reconstruction. They are often used in VR or research environments, they also offer support to beamforming and localization, [14]. Binaural recording, achieved with in-ear microphones or dummy heads (e.g., Neumann KU-100, Brüel & Kjær HATS), replicates human auditory perception by preserving interaural differences and head-related spectral cues. While ideal for headphone playback, it tends to be listener-specific and assumes a static head position [14]. Ambisonics encodes the sound field in spherical harmonics and allows post hoc spatial transformations and flexible decoding [15]. Ambisonics is widely adopted in immersive environments for its scalability and scene-based architecture [16], its precision depends on the order. Concerning sound reproduction techniques they can be grouped into channel-based, object-based, and scene-based strategies. Channel-based systems (e.g.: stereo, 5.1) route audio to fixed speaker positions and offer limited adaptability. Object-based systems, such as Dolby Atmos, describe each source with metadata, enabling dynamic rendering depending on playback conditions. Scene-based approaches like Ambisonics or Wave Field Synthesis (WFS) aim to recreate the full sound field around the listener, independent of individual source positions or speaker layout [14], [15].

It is worth noticing that spatial perception arises not only from the physical sound field but through binaural processing in the peripheral and central auditory systems. Achieving ecological validity in spatial audio, especially in simulated environments, requires accurate reproduction of perceptual cues such as interaural time and level differences, reverberation, and spatial continuity [14]. In soundscape studies, where sound is viewed not as a pollutant but as an environmental and cultural component, ISO 12913-1 defines the acoustic environment as a perceptual construct shaped by both physical and contextual factors [14]. Thus, capturing and reproducing spatial audio with perceptual accuracy becomes essential for effective design and evaluation.

Despite technological advancements, spatial audio techniques remain underutilized in many practical contexts, including music production and immersive installation design. Practitioners often rely on presets or commercial defaults without a clear understanding of spatial encoding, rendering logic, or perceptual foundations. This paper aims to address that gap by offering a structured orientation to spatial audio workflows, enabling informed and purposeful application of these techniques.

III. Spatial Audio Production Workflow

A. Recording techniques for 3D Audio

Microphone techniques for spatial audio reproduction [18] can be initially classified into three broad categories: traditional techniques, compact arrays, and sparse arrays (Table I).

Traditional microphone techniques are rooted in classical stereo and surround recording practices, where microphones are typically positioned in proximity to the sound sources or in semi-coincident arrangements. These setups, e.g. XY, ORTF, MS, AB, or Decca Tree, employ directional capsules to capture a localized sound field and preserve timbral integrity, making them particularly effective for conventional music and broadcast productions. While they offer excellent imaging and tonal control, their spatial coverage is generally limited to the frontal sector or horizontal plane, and they lack the capacity to represent immersive vertical dimensions [15], [17].

In contrast, compact microphone arrays are designed to record the spatial characteristics of a sound field from a single, tight spatial region. These arrays include Ambisonic microphones (e.g., TetraMic, Eigenmike, AMBEO), dummy heads (e.g., Neumann KU-100), and binaural setups (e.g., 3Dio, KEMAR). Their small physical footprint and high phase coherence make them ideal for post-processable formats such as Ambisonics or headphone-based binaural rendering, enabling full-sphere spatial capture and flexible decoding across reproduction systems [14], [16]. Compact arrays are thus the preferred choice in virtual and augmented reality applications, 360° video, and immersive audio research [19]. However, they do not allow for precise control over the individual spatial positions of sound sources and typically require a complex post-production pipeline.

Sparse microphone arrays, on the other hand, consist of widely distributed microphones arranged to mirror or approximate the positions of loudspeakers in multichannel formats (e.g., 5.1.4, 9.1, 22.2). These setups, including techniques like OCT-3D, ORTF-3D, Hamasaki Cube, and Bowles Array, leverage interchannel time and level differences to create natural spatial decorrelation enhancing the sense of envelopment. Often involving both horizontal and vertical spacing, these arrays offer a direct approach to immersive reproduction in physical spaces, particularly suited to concert hall recordings, cinema sound, and spatial music installations [15], [16]. While they provide compelling realism and depth, they are generally less flexible in post-production and require significant physical setup space.

A summary of this tecnhiques is reported in Table I

TABLE I
MICROPHONE ARRAYS/TECHNIQUES FOR SPATIAL AUDIO RECORDING

Traditional	Compact	Sparse
XY	Neumann KU-100	OCT-3D
ORTF	3Dio Free Space	Bowles Array
Blumlein	B&K HATS	Williams Umbrella
AB	Sennheiser AMBEO	2L-Cube
Decca Tree	TetraMic	Spider Tree
MS	Eigenmike	Twin Cube
IRT Cross	Soundfield ST450	Double UFIX
Hamasaki Square	Planar Microphone Array	Hamasaki Cube
OCT Surround		PCMA-3D
Double MS		ORTF-3D
Dummy Head		ESMA-3D
		aud3Dio
		Rec-3D

B. Mixing, Panning & Spatialization

Once spatial audio content has been captured, the mixing phase determines how individual sources are distributed, rendered, and experienced within a target reproduction system. Unlike conventional stereo or surround mixing, spatial audio requires balancing spatial envelopment, localization accuracy, intelligibility, and compatibility across diverse playback formats.

A first strategic decision concerns the *representation model* adopted in the mix: scene-based, object-based, or channel-based. Scene-based formats like Ambisonics allow rotation and re-rendering from any listening orientation and are well-suited to VR and interactive contexts. Object-based audio (OBA), such as Dolby Atmos or MPEG-H, grants per-source spatial positioning and dynamic control, supporting personalization and interactivity. In contrast, channel-based approaches (e.g., 5.1, 7.1.4) rely on fixed speaker layouts and are typically used in cinema or live concert settings [15].

Each representation imposes different constraints on the mixing strategy. *Ambisonic mixing*, for example, relies on encoding input tracks into the B-format using spatial panners and managing their order and gain structure consistently. Head-locked elements (e.g., narration, soloists) can be placed in lower-order components, while ambient or reverberant sources benefit from higher-order spatial encoding. The use of Ambisonic tools (e.g., IEM Plug-in Suite, SPARTA) enables flexible spatialization and dynamic rotation, but demands careful attention to phase coherence and timbral stability [14].

Object-based mixing offers greater granularity by treating each source as an independent entity with metadata (position, trajectory, gain). This enables real-time rendering adapted to playback conditions. However, object counts may be limited in consumer platforms, and spatial interactions between objects (e.g., masking, decorrelation) must still be managed artistically. Object-based strategies are preferred in multimedia and narrative-driven applications, but require robust authoring tools and spatial monitoring [12].

Channel-based mixes, while less flexible, are still relevant in traditional concert hall and fixed speaker installations. Mixing in these formats often involves speaker-targeted panning methods such as Vector Based Amplitude Panning (VBAP [20]) or Distance-Based Amplitude Panning (DBAP [21]) and related variants, which provide good spatial control but depend heavily on array geometry. In these contexts, some authors recommend pre-mixing reverberant and direct sound components separately and re-balancing them based on the expected listener location [16].

Across all formats, immersive mixing must consider *spatial clarity*, *energy distribution*, and *coherence* between the auditory scene and the visual or physical context. This often leads to hybrid workflows combining close-miking, artificial reverberation, multiband panning, and head-tracked binaural rendering for previewing [22].

C. Spatial Audio Reproduction

Spatial audio reproduction techniques aim to faithfully recreate the immersive experience encoded in spatial recordings or synthesized scenes. Depending on the application domain (XR, soundscape design, or multichannel music playback) different reproduction strategies are adopted, each with its own trade-offs in terms of spatial fidelity, computational complexity, and perceptual effectiveness (Table II).

Scene-based formats, such as Ambisonics, allow the captured B-format signals to be decoded into loudspeaker arrays or binaural renderings through head-related transfer functions (HRTFs). This approach is particularly attractive due to its format flexibility and the possibility of head-tracked rendering, enabling dynamic listener interaction in VR/AR environments [15], [23]. Nevertheless, perceptual studies have highlighted that increasing spatial resolution (e.g., from 2D to 3D arrays) does not necessarily lead to greater immersion or clarity. In fact, 2D horizontal arrays have been rated more natural and spatially well-defined than full 3D layouts, which are often perceived as muffled and overly distant [23].

Object-based rendering engines (e.g., MPEG-H, Dolby Atmos) rely on metadata-driven positioning and often leverage hybrid rendering strategies that mix binaural and speaker-based components. These systems support personalization and adaptability, but their performance heavily depends on accurate room modeling and HRTF selection. Recent work has explored the use of 360° imagery to reconstruct acoustic geometry for interactive spatial rendering [24], improving consistency between visual and auditory cues.

For channel-based reproduction, frameworks such as the SoundScape Renderer (SSR) provide a unified rendering pipeline capable of emulating various panning methods, including Vector-Base Amplitude Panning (VBAP) and Wave Field Synthesis (WFS), across arbitrary loudspeaker geometries [25]. While these systems are particularly suitable for fixed installations or research environments, they demand precise loudspeaker calibration and room acoustic control to ensure perceptual coherence.

Perceptual evaluation remains central to assessing spatial reproduction strategies. Parameters such as presence, readability, distance, localization, coloration, and stability have been identified as key discriminants in listener preference [23]. Notably, the optimal reproduction configuration appears to

TABLE II
SUMMARY OF MAIN SPATIAL AUDIO TECHNIQUES ACROSS RECORDING, MIXING, AND REPRODUCTION STAGES

Stage	Technique / Model	Description	References
Recording	Traditional Stereo/Surround (XY, ORTF, AB, Decca Tree)	Directional microphone setups with limited spatial depth and frontal coverage. Suitable for stereo/5.1 content.	[15], [17]
	Compact Arrays (Ambisonics mics, Dummy heads)	Full-sphere spatial capture with phase coherence. Ideal for Ambisonics and binaural workflows.	[14], [16]
	Sparse Arrays (ORTF-3D, Hamasaki Cube)	Distributed microphone setups emulating playback geometries. High realism for physical spaces.	[15], [16]
Mixing	Scene-based (Ambisonics)	Spatial encoding using B-format and spherical harmonics. Allows flexible rotation and decoding.	[14], [15]
	Object-based (Atmos, MPEG-H)	Sources treated as metadata-rich objects with independent spatial control. Adaptive to playback systems.	[12], [13]
	Channel-based (VBAP, DBAP)	Panning towards fixed loudspeaker positions. Effective in fixed setups and concert reproduction.	[16], [21]
Reproduction	Binaural Rendering	HRTF-based simulation of 3D sound over headphones. Often used for previewing and VR playback.	[14]
	Ambisonics Decoding	Scalable decoding of B-format to speaker arrays or headphones. Enables head-tracking and scene rotation.	[15], [23]
	Object-Based Rendering	Real-time spatialization of sound objects based on metadata. High personalization and flexibility.	[24]
	SSR (VBAP, WFS)	Rendering framework for evaluating and comparing spatial audio algorithms in controlled layouts.	[25]

depend more on the source material and listener expectations than on raw spatial completeness. For example, urban soundscapes are best reproduced with 2D horizontal arrays, while indoor reverberant scenes may benefit from elevated 3D playback.

IV. EXPERIMENTAL SCENARIO DESIGN

Building upon previous work on remote immersive audio streaming infrastructures [6], we designed a distributed audio/video communication setup composed of two physically separated nodes: an *event room*, where a live concert is captured using multiple spatial audio techniques, and a *listening room*, where the performance is reproduced via spatial audio and video projection. This configuration, implemented in a real-world demonstration, reflects practical scenarios for multimedia streaming of musical events.

In the *event room*, the audio is captured by a technical crew using different microphone array configurations. In the *listening room*, the remote listener or producer can dynamically switch between alternative array and mixing pipelines based on artistic or technical preferences. This selection process is orchestrated by a central control node, located in the *listening room*, which automates the switching of microphone groups remotely, DSP configurations for audio/video synchronization, immersive panning, and real-time mixing locally through a pre-programmed timeline. The entire infrastructure ensures interoperability between DANTE-native and non-DANTE audio/video devices, provides precise clock synchronization via GPS and Precision Time Protocol (PTP)v2-aware switches, and enables secure transmission through two VPN tunnels: one for audio signals, the other for DANTE Domain Manager control signals. Audio streams go straight from one location to the other, the DDM position is not influencing. Audio streams are transmitted as 9-channels, 48kHz, 24-bit signals over DANTE, while video is encoded using the H.265-based Dante AV-H protocol, supporting up to 4K resolution with bandwidth adaptation.

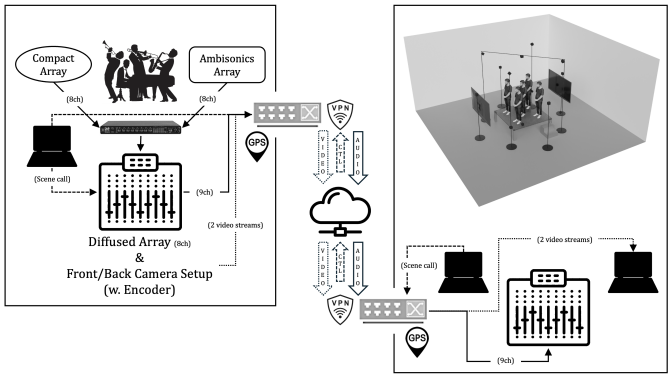

Fig. 1. Experimental setup scenario.

This experimental setup is shown in figure 1 and it allows for a sequential comparison of multiple *microphone array and spatialization* combinations under identical musical content, enabling both technical benchmarking and perceptual evaluation of alternative spatial audio production workflows.

A. Recording Set-up (Event Room)

The experimental scenario involves a chamber music performance, with musicians positioned at the center of the stage and captured using three distinct microphone techniques, each supporting a different immersive reproduction strategy:

(1) *Compact array*: a set of 8 Neumann KM184 cardioid condenser microphones[1], featuring a Signal-to-Noise Ratio (SNR) of 81dB-A, arranged in a compact layout following a 1:1 spatial mapping with the standard 8-channel MACH1 loudspeaker configuration [6]. This setup enables direct spatial

[1]https://www.neumann.com/en-us/products/microphones/km-184-series-180

correspondence between capture and playback, minimizing the need for interpolation or complex panning during mixing.

(2) Diffuse array: a set of 8 omnidirectional microphones (e.g., AKG C414, AKG P420; SNR ranging from 79 to 88dB-A) distributed throughout the concert hall according to the venue's acoustic and architectural features. This configuration enhances spatial envelopment and immersion by capturing diffuse reverberant energy and ambient cues. Capsule placement was determined through on-site measurements and empirical evaluation to ensure optimal coherence and spatial distribution.

Signals from both arrays and the ambisonic mic are routed to 3 Metric Halo ULN-8 mkIV audio interfaces[2] operating as independent A/D converters and equipped with onboard DSP. Up to 24 analog signals are forwarded to an in-site mixing console via the AES10-standardized Multichannel Audio Digital Interface (MADI) protocol [26].

(3) Ambisonics array: a second-order Ambisonics microphone (OctoMic[3]), composed of eight condenser capsules, whose raw signals are routed to a dedicated laptop for A-to-B Format conversion and Higher Order Ambisonics (HOA) encoding.

The audio signals captured by the compact and diffuse arrays (16 channels), together with the post-processed Ambisonics stream (9 channels), are routed to a DANTE-compatible mixing console (SoundCraft SI series[4]) directly connected to the in-site network infrastructure. As discussed in Sections IV-C and IV-D, the console dynamically manages channel transmission, delivering only the subset of signals required at any given time in the remote *listening room* system (maximum 9 channels per instance). This strategy reduces network bandwidth usage while supporting real-time streaming.

Immersive video recording is conducted using one BG-ADAMO-JRDA20X-B PTZ camera (DANTE native)[5] for audiance framing and a traditional Blackmagic studio camera[6] for the stage, both placed at the center of the event room. Corresponding video signals (1920×1080, 60FPS) are acquired by a DANTE-compatible video encoder[7] and transmitted through the same in-site network switch used for audio, enabling synchronized remote playback and analysis.

B. Spatialization & Mixing Strategies

The mixing strategies adopted in our experimental workflow reflect the demonstrative goals outlined in Section IV, and involve three spatialization techniques applied to each of the three recording setups previously described:

(1) Virtual Vector-Based Panning (VVBP) refers to a spatial audio panning approach derived from conventional Vector Base Amplitude Panning (VBAP), optimized for spatial audio applications without altering the original signal content. Unlike Ambisonics or object-based techniques, VVBP encodes spatial cues through amplitude coefficients only, enabling real-time rendering for 3DoF listening while preserving mix integrity and dynamic range. In our implementation, we adopt the MACH1 Spatial framework[8], which encapsulates VVBP into multichannel deliverables compatible with standard audio containers and codecs (e.g., WAV, AAC, MP3). VVBP avoids distance simulation through filtering or delay, and instead employs a *divergence principle*: panning a signal away from the center increases its energy distribution toward the target direction without affecting perceived proximity — preserving creative intent and dynamic range. Its channel-based design supports stereo coherence, head-tracked decoding, and metering workflows consistent with traditional stereo or surround paradigms. Furthermore, MACH1 content can be transcoded to and from Ambisonics via dedicated conversion tools.

(2) Ambisonics Equivalent Panning (AEP) is an analytical spatialization method that approximates Ambisonics in-phase decoding without requiring explicit B-format encoding or matrix operations [27]. Unlike classical Ambisonics, which relies on spherical harmonics and source/loudspeaker positional matrices, AEP uses a closed-form gain function based on angular distance θ: $f(\theta, p) = \left(\frac{1+\cos(\theta)}{2}\right)^p$ where p is a continuous parameter representing the spatial resolution (analogous to Ambisonics order). In Cartesian coordinates, the formulation becomes: $f(P, P_s, p) = \left(\frac{xx_s+yy_s+zz_s+r^2}{2r}\right)^p$, $r = \sqrt{x^2+y^2+z^2}$.

This technique supports fractional sharpness, smooth control over directional cues, and avoids harmonic truncation artifacts. Due to its efficient implementation and phase-coherent response, AEP is particularly suited for adaptive rendering in VR, AR, and real-time installations.

(3) K-Nearest Neighbors (KNN) is a spatialization algorithm inspired by *machine listening* practices. It selects the K loudspeakers closest to a virtual source position and distributes the signal using inverse-distance weighting. Our implementation, based on SPAT Revolution[9], introduces a *Nearest Neighbor Spreading* parameter, defined as a continuous value $S \in [0, 100\%]$ that determines K dynamically based on the total number of available loudspeakers N. Lower spread values yield focused spatialization ideal for direct sources, while higher spreads result in diffuse spatial fields suitable for ambient material. The continuous nature of S ensures perceptual smoothness during spatial transitions, making KNN attractive for dome-based applications and immersive theatrical settings [28].

The VVBP implementation is hosted on a dedicated laptop in the remote *listening room*, running the MACH1 plugin suite[10] within a REAPER[11] session. AEP and KNN are

[2]https://mhsecure.com/products/mkIV/ULN8mkIV.html
[3]https://www.core-sound.com/products/octomic
[4]https://www.soundcraft.com/en/product_families/si-series
[5]https://zzipp.com/gb/camere-ptz/11862-bg-adamo-jrda20x-b.html
[6]https://www.blackmagicdesign.com/it/products/blackmagicstudiocamera
[7]https://www.amx.com/en/products/nmx-enc-n3312d
[8]https://mach1.tech
[9]https://doc.flux.audio/spat-revolution/Spatialisation_Technology_Panning_Algorithms.html
[10]https://www.mach1.tech/spatial-system
[11]http://reaper.fm

managed via a standalone instance of SPAT Revolution[12]. Upstream control is handled by a Supercollider patch[13], which coordinates parameter switching and activation across software components using Open Sound Control (OSC)/MIDI/rtpMIDI, and proprietary protocols. This timeline orchestrates: *(I)* the muting/unmuting of input channels at the SoundCraft via rtpMIDI in the *event room*; *(II)* the routing and activation of specific panning algorithms; and *(III)* the overlay of a visual marker on the video feed indicating the currently active mic+panning configuration. The result is a fully automated, repeatable sequence for testing alternative workflows under identical musical content.

The combination of the three recording setups (CMA, DMA, AM) with the three spatialization strategies (VVBP, AEP, KNN) results in nine possible workflows. Table III summarizes the main advantages and limitations of each option, highlighting the trade-offs that informed the design of the demonstration scenarios presented in Section IV. We will alternatively use either AEP or KNN, thus experimenting only six listening conditions as shown in Table IV.

C. Control & Signals Networking

The adopted network architecture ensures reliable, low-latency transmission and synchronization of audio and video signals between geographically distributed locations. Building on the modular infrastructure in [6], it integrates PTPv2-aware switching, GPS-based timecode synchronization, and secure VPN tunneling, with enhancements tailored for *AV-over-IP* workflows and hybrid control. Both the *event room* and *listening room* are equipped with a Mikrotik RB5009 router[14] and a CRS326-24S+2Q+RM switch[15] (RouterOS 7.19.3), supporting IEEE 1588 (PTPv2). GPS modules provide UTC-based timecode, and a Raspberry Pi 4 at each site acts as a *grandmaster clock*. Switches operate as boundary clocks, reducing jitter and ensuring consistent packet timing. Inter-site communication is secured via WireGuard VPN tunnels, transporting audio, video, and control data in real time.

DANTE (by Audinate) serves as the core transport layer for digital audio and video over Ethernet. It encapsulates 24-bit/48kHz audio into IP packets, requiring DANTE-enabled devices or compatible interfaces. Operating on switched 100Mbps or Gigabit Ethernet with DiffServ-based QoS, DANTE supports up to 48 bi-directional channels on Fast Ethernet and up to 512 on Gigabit networks, depending on stream settings. Latency is configurable through DANTE Domain Manager. Clocking is handled via PTP, and the system supports redundancy, AES67 interoperability, unicast/multicast modes, and up to 10 network hops, making it suitable for scalable and reliable professional deployments. Unlike the setup in [6], the DANTE Domain Manager is hosted in the cloud and accessed via a dedicated VPN, decoupling configuration from local machines and improving fault tolerance without compromising AV performance. All audio streams follow a high-resolution profile: up to 9 channels, 48 kHz, 24-bit, grouped into 3 Dante flows (4 channels per flow), requiring approximately 13.8Mbps total bandwidth. Streams are timestamped and clock-aligned via PTPv2, supporting stable low-latency transmission with an upper bound of 40ms. Unicast mode is used to maintain predictable bandwidth allocation and simplify setup. To enable synchronized visual monitoring in the *listening room*, video is transmitted using DANTE AV-H, an H.265-based protocol optimized for high-efficiency streaming. The video signal from studio camera is routed as: camera $\rightarrow$ encor $\rightarrow$ decoder $\rightarrow$ video mixer for overlay, while PTZ camera follows $\rightarrow$ DANTE studio (in the listening room) $\rightarrow$ video mixer. AV-H streams are not timestamped; empirical evaluation shows end-to-end latency around 50ms (three frames at 60FPS), including encoding, network transmission, and decoding. Although a single stream can peak at 50Mbps, typical usage under *automatic bitrate mode* ranges from 8 to 25Mbps depending on scene complexity. Two concurrent streams result in a maximum of 100Mbps for the video layer. Audio carried by AV-H devices is not amplified due to limited channel support but is transmitted to provide a reference signal for manual alignment with microphone array tracks in the *listening room*.

The combined network load reaches approximately 114Mbps (13.8Mbps for audio and up to 100Mbps for video), leaving a safe margin on a Gigabit backbone. Audio latency is tightly bounded by PTP clocking, while video latency remains more flexible due to lack of timestamping.

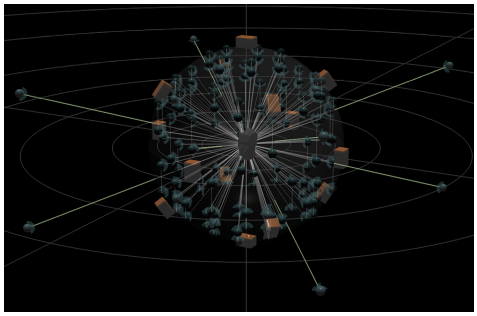

Fig. 2. Loudspeaker System (*orange*) and Ambisonics Virtual Sources (*blue*) 3D Rendering.

D. Reproduction System & Playback (Listening Room)

The *listening room* features a 14-channel loudspeaker array[16] (Figure 2) designed for 3DoF spatial audio reproduction. The layout adheres to an upcoming MACH1 Spatial framework specifications update, and consists of a geometrically uniform distribution:

- 8 loudspeakers positioned at the cube's vertices, centered on the listener;
- 6 additional loudspeakers located at the center of each face.

[12]https://www.flux.audio/project/spat-revolution/
[13]https://supercollider.github.io/
[14]https://mikrotik.com/product/rb5009ug_s_in
[15]https://mikrotik.com/product/crs326_24s_2q_rm
[16]https://www.kgear.it/en/product/GH4

TABLE III
MATRIX OF CAPTURE–RENDERING WORKFLOWS WITH MAIN ADVANTAGES AND LIMITATIONS .

Capture	VVBP (MACH1)	AEP (SPAT)	KNN (SPAT)
Compact Array (CMA)	Direct 1:1 mapping, low latency, coherent timbre. Limited immersion, poor lateral/rear coverage.	Analytical panning with sharpness control, homogeneous field. Higher computational load, possible inter-channel incoherence.	Flexible interpolation, good balance between focus and diffusion. Strong dependence on spread parameter, risk of spatial blur.
Diffuse Array (DMA)	Natural reverberation and envelopment, simple pipeline. Weak directionality, "wash out" effect.	Analytical spatial control, improved phase consistency. Diffuse capture reduces directional impact.	Wide, enveloping reproduction, adaptive focus/diffusion. Lower localization precision, possible incoherence.
Ambisonics Mic (AM)	Standardized HOA, flexible decoding, good frontal localization. Complex pipeline (A→B + decode), higher latency.	Phase-coherent analytical panning, balanced accuracy/immersion. Computationally demanding, decoding-sensitive.	Immersive diffuse reproduction, suited for large arrays. Less effective for point sources, potential loss of sharpness.

TABLE IV
TIMELINE AND SIGNAL ROUTING ACROSS CONTROL STATIONS AND SPATIALIZATION MODES

Time (mm:ss)	CTRL to *Concert Room*	CTRL to *Listening Room* (FOH)	CMA	DMA	AM	VVBP	AEP/KNN
00:00 - 01:00	Desk CUE 1	Desk (Ch. 1–14), SPAT CMA-On, Reaper MUTE-All	✓				✓
01:00 - 02:00	Desk CUE 2	SPAT CMA-Off, SPAT DMA-On		✓			✓
02:00 - 03:00	Desk CUE 3	SPAT DMA-Off, SPAT AM-On			✓		✓
03:00 - 04:00	Desk CUE 1	Desk (Ch. 15–28), SPAT MUTE-All - Reaper CMA-On	✓			✓	
04:00 - 05:00	Desk CUE 2	Reaper CMA-Off, Reaper DMA-On		✓		✓	
05:00 - 06:00	Desk CUE 3	Reaper DMA-Off, Reaper AM-On			✓	✓	

CMA = Compact microphones array, **DMA** = Diffused microphones array, **AM** = Ambisonics microphone
Desk CUE commands refer to SoundCraft SI scene calls for appropriate output bus routing of input signals

TABLE V
SPEAKER LAYOUT IN 3D CARTESIAN COORDINATES RELATIVE TO THE LISTENER'S HEAD AT (0,0,0)

Speaker ID	Position (x, y, z) [m]	Elevation from ground [m]	Support
V1	(+1.27, +1.27, +1.27)	3.47	Shared stand (1/4)
V2	(+1.27, +1.27, –1.27)	0.93	Shared stand (1/4)
V3	(+1.27, –1.27, +1.27)	3.47	Shared stand (2/4)
V4	(+1.27, –1.27, –1.27)	0.93	Shared stand (2/4)
V5	(–1.27, +1.27, +1.27)	3.47	Shared stand (3/4)
V6	(–1.27, +1.27, –1.27)	0.93	Shared stand (3/4)
V7	(–1.27, –1.27, +1.27)	3.47	Shared stand (4/4)
V8	(–1.27, –1.27, –1.27)	0.93	Shared stand (4/4)
F1	(+2.20, 0.00, 0.00)	2.20	Individual stand
F2	(–2.20, 0.00, 0.00)	2.20	Individual stand
F3	(0.00, +2.20, 0.00)	2.20	Individual stand
F4	(0.00, –2.20, 0.00)	2.20	Individual stand
F5	(0.00, 0.00, +2.20)	4.40	Suspended
F6	(0.00, 0.00, –2.20)	0.00	Underfloor

Loudspeakers **V1–V8** define the cube's vertices. The listener's head is centered at 1.8m above the ground.
F1–F4 correspond to the face centers; **F5** and **F6** align vertically with the listener.

The coordinate system follows the MACH1 convention[17]. This configuration extends the classical 8-channel cube [6] by adding face-centered transducers to increase spatial density and improve divergence rendering. It preserves compatibility with octahedral layouts, enabling smooth transcoding from formats such as surround or Atmos, and offers perceptual benefits for dynamic scenes requiring accurate localization. The design reflects the principles of Spatial PCM Sampling (SPS) [29], which conceptualizes spatial audio as a discrete sampling process analogous to temporal PCM. While MACH1 does not implement SPS explicitly, its regular geometry supports the same goal: minimizing spatial aliasing and enhancing perceptual coherence via uniform speaker distribution. This facilitates both artistic flexibility and integration with head-tracked and scene-based pipelines.

For Ambisonics input, encoding and panning are handled as follows:

- The A-format signal is transmitted to a laptop in the *event room*.
- The A-to-B format converter plugin performs A-to-B format encoding for HOA compatibility.

In the *listening room*, the IEM AllRADecoder[18] applies All-Round Ambisonics panning [30], [31], following AmbiX ordering [32], preparing the signal for VVBP-compatible

[17] X = Left to Right, Y = Front to Back, Z = Top to Bottom.

[18] https://plugins.iem.at/docs/allradecoder

playback. B-format Ambisonics signals are processed by SPAT using a *uniform Sloane* grid to simulate up to 100 virtual loudspeaker directions. This virtual decoding layer enables fine-tuned spatial accuracy, based on AllRAD principles, and can be scaled to trade off spatial resolution against processing load. This step is applied only to Ambisonics streams.

All signals are routed via USB from a second SoundCraft SI console to a Front-of-House (FOH) laptop, where two spatialization sessions operate independently:

- REAPER, for VVBP rendering;
- SPAT Revolution, for AEP and KNN rendering.

Panned signals are sent back to the console for final amplification and playback through the 14-channel loudspeakers array.

A second laptop handles video playback and control: it receives DANTE AV-H video streams via a dedicated decoder[19], overlays a color-coded frame for synchronization, manages timeline parameters, and triggers remote muting via OSC/MIDI/rtpMIDI back to the *event room* mixing console (see Table IV).

V. Open Implementation Alternatives

While the proposed setup integrates professional-grade commercial tools — such as DANTE for audio/video-over-IP, SPAT Revolution for spatial rendering, and MACH1 for vector-based panning — the same architectural principles can be replicated using open-source and freely available alternatives. This section outlines substitution pathways for each major component of the signal chain, providing a blueprint adaptable to pedagogical, artistic, and budget-conscious installations.

A. Audio-over-IP and Clock Synchronization

DANTE offers a robust framework for multichannel audio streaming with low latency and PTPv2-based clocking. Comparable open-source solutions exist that support spatial and distributed audio workflows with flexible configuration.

`JACK Audio Connection Kit`[20] is a low-latency audio server widely adopted in Linux-based environments for routing audio between applications and networked devices via its `NetJACK` extension. `JackTrip` [33], developed at Stanford's CCRMA, enables uncompressed, bidirectional audio streaming over UDP with jitter buffering and sample-accurate synchronization only through external clock synchronization, ideal for remote music collaboration. Both tools are agnostic to content encoding and channel layout, with routing decoupled from rendering logic. `SonoBus`[21] provides a cross-platform, Peer-to-Peer (P2P) UDP streaming to stream audio directly between users, with a central server only used to help users find each other and establish connections. It supports manual channel grouping (e.g., stereo, 5.1, 7.1), configurable encoding (PCM or Opus), and real-time diagnostics via a graphical interface. It is also available as a plugin for DAW integration, making it suitable for experimental and semi-professional contexts.

`RAVENNA`[22] is an open audio-over-IP protocol based on RTP, RTSP, and IEEE 1588 (PTPv2). Originally designed for broadcast, it supports uncompressed multichannel PCM with sub-millisecond latency over IP networks. While primarily used in hardware setups, software stacks like ALSA, JACK, or GStreamer can interface with RAVENNA nodes using standard endpoints. `Zita-AJBridge`[23] and `Zita-NJBridge`[24] complement JACK-based systems by bridging ALSA hardware and enabling multichannel IP streaming between servers, with built-in sample rate adaptation. Although these tools lack native PTPv2 support, synchronization can be achieved through system-level NTP or auxiliary timecode mechanisms. When stricter alignment is required, GPIO-based triggers or external word clocks can be integrated.

B. Spatialization and Rendering Engines

SPAT Revolution integrates VBAP, Ambisonics, and WFS rendering into a unified engine with control over source trajectory, perceptual parameters, and loudspeaker layout. Equivalent functionalities can be implemented through modular open-source plugins and toolkits supporting both scene- and object-based paradigms.

The `IEM Plugin Suite` [25] provides a comprehensive collection of free VST/AU tools for Higher-Order Ambisonics (HOA), including the `AllRADecoder` for customizable loudspeaker geometries. `SPARTA` [26] adds beamforming, spatial reverberation, and visualization, while `AmbiX` and `MCFX` [27] include Ambisonics rotation, panning, and convolution tools compatible with AmbiX conventions. The `ICST Ambisonics Tools` [28] offer Max/MSP integration and flexible routing for spherical panning in graphical environments.

For channel-based panning and WFS, the `SoundScape Renderer (SSR)` [29] is a modular real-time engine supporting VBAP, WFS, and distance-based panning over arbitrary loudspeaker arrays, suitable for reproducible research and fixed installations.

For strategies conceptually aligned with MACH1 VVBP model, we note that MACH1 has opened the code and just sells plugins. However, the Spatial PCM Sampling (SPS) framework [29], developed by A. Farina[30], enables encoding from mono/stereo to SPS-compatible formats. Combined with `Xvolver` [31], a multichannel FIR convolver, these tools enable

[19] https://www.amx.com/en/products/nmx-dec-n3322d
[20] https://jackaudio.org/
[21] https://sonobus.net/
[22] https://www.ravenna-network.com/
[23] https://kokkinizita.linuxaudio.org/linuxaudio/zita-ajbridge-doc/quickguide.html
[24] https://github.com/digital-stage/zita-njbridge
[25] https://plugins.iem.at/
[26] https://leomccormack.github.io/sparta-site/
[27] https://www.matthiaskronlachner.com/?p=2015
[28] https://ambisonics.ch/icst-ambisonics-tools/
[29] https://spatialaudio.net/ssr/
[30] https://www.angelofarina.it/SPS-conversion.htm
[31] https://www.angelofarina.it/X-volver.htm

flexible spatial mixes across arbitrary speaker layouts. Though not implementing VVBP explicitly, they offer high-fidelity rendering and are well-suited to artistic or research contexts.

C. Video Streaming and Synchronization

DANTE AV-H provides H.265-based video-over-IP with hardware-level PTPv2 synchronization, enabling frame-accurate audio-video alignment. Open-source frameworks can replicate these functionalities using standardized protocols and clock-sharing methods.

`OBS Studio`[32] supports real-time AV capture and streaming via RTMP and NDI. With plugins support, it allows low-latency LAN transmission and local recording. For fine-grained pipeline control, `GStreamer`[33] supports RTP, SRT, and RTSP streaming over UDP/TCP with explicit timecode negotiation and custom pipeline buffering. `FFmpeg`[34] offers a flexible backend for batch or headless streaming via RTP or SRT. Although native PTPv2 support is absent, synchronization can be managed through shared system clocks (e.g., NTP) or timestamp-based buffering, especially within GStreamer pipelines. Integration with JACK-based audio systems ensures coherent AV performance in distributed environments.

D. Control and Command Routing

Besides used commercial tools equivalent open-source platforms provide modular alternatives with comparable expressiveness. `Open Stage Control`[35] is a browser-based interface builder supporting OSC/MIDI panels and WebSocket integration, suitable for DAWs, media servers, and spatial engines. `QLab`[36] (free tier) supports timeline scripting via MIDI/OSC for coordinating cues across lighting, video, and audio.

For custom logic and interactive systems, `PureData`[37] and `Cycling '74 Max`[38] provide programmable environments for DSP, sequencing, and multichannel routing, supporting OSC, MIDI, and serial protocols. These are widely used in responsive stage setups, generative installations, and networked music systems.

VI. Evaluating Immersion and Presence in Immersive Audio Experiences

Immersion and presence are key experiential dimensions in spatial audio systems, yet conceptually distinct. *Immersion* describes a psychological state of deep engagement, shaped by sensory fidelity, narrative coherence, and individual predisposition [34], while *presence* refers to the illusion of "being there" in a mediated environment, modulated by sensory-consistency and contextual expectations [35]. Presence may be physical, social, or both [36], and the two constructs can occur independently.

[32]https://obsproject.com/
[33]https://gstreamer.freedesktop.org/
[34]https://ffmpeg.org/
[35]https://openstagecontrol.ammd.net/
[36]https://qlab.app/
[37]https://puredata.info/
[38]https://cycling74.com/

Subjective evaluation methods, especially post-experience questionnaires such as ITC-SOPI and MEC-SPQ, are widely used to assess spatial presence, engagement, and naturalness [36], [37]. Complementary approaches include slider-based continuous ratings and open-ended interviews to capture temporal and qualitative nuances of the experience.

A within-subject listening test will be conducted with four to six participants per session, exposed to a remote string quartet performance reproduced in the six capture–mix–render configurations. Seated in a circular loudspeaker array, participants will evaluate each condition using a 7-point Likert scale on:

1) the degree of immersion (e.g., "How absorbed were you in the listening experience?"),
2) physical presence (e.g., "To what extent did you feel located in the same space as the performers?"),
3) social presence (e.g., "Did you feel as if the musicians were performing for you personally?"),
4) audio naturalness (e.g., "How natural did the performance sound to you?"),
5) envelopment (e.g., "Did the sound feel as if it surrounded you spatially?").

The six conditions will be presented in a counterbalanced Latin Square design, divided into two blocks with intermission to reduce fatigue. At the end, participants will rank the configurations globally and provide open-ended feedback on factors contributing to immersion and co-presence.

This mixed-method design combines quantitative and qualitative data to assess perceptual outcomes across technical configurations, ensuring both ecological validity and experimental control. Literature supports this approach as suitable for evaluating cognitive-affective constructs in non-interactive music listening [36], [37].

VII. Conclusions

This paper presented a comparative analysis of end-to-end 3D audio workflows for live music production, evaluating the technical and experiential implications of different microphone arrays, spatialization strategies, and reproduction configurations. By integrating compact, diffused, and Ambisonic capture methods with lightweight panning techniques (VVBP, AEP, KNN) and real-time audio-video transport over DANTE networks, the proposed framework supports reproducible, low-latency, and perceptually coherent immersive experiences.

The experimental setup, designed for distributed performance monitoring and critical A/B testing, highlights the trade-offs between spatial fidelity, implementation complexity, and network scalability. More importantly, it offers a practical reference for audio professionals, sound designers, and live engineers seeking to adopt immersive practices without relying exclusively on proprietary or cinema-oriented solutions. The inclusion of synchronized video streaming further supports coherent audiovisual experiences in hybrid and remote setups. The proposed methodology emphasizes accessibility, interoperability, and perceptual relevance, thus enabling technically grounded decisions aligned with artistic and production goals.

By including a listener-centered evaluation protocol, the work addresses the need to connect system design with subjective dimensions of experience, providing tools to assess immersion and presence in realistic contexts.

Future work will extend this framework with formal perceptual results and explore adaptive streaming mechanisms and hybrid spatialization models optimized for remote collaborative scenarios over heterogeneous infrastructures, including 5G and edge-assisted environments.

ACKNOWLEDGEMENTS

The authors would like to thank all the technical partners for their valuable support in providing cutting-edge tools and technologies that enabled the implementation of the immersive and distributed audio experience showcased during the event. In particular, we acknowledge: Audinate, Harman Industries/FLUX, MACH1, Metric Halo and K-array.

REFERENCES

[1] L. Turchet et al., "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.

[2] ——, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[3] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[4] L. Turchet et al., "5g-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4691–4709, 2024.

[5] A. Hunt, H. Daffern, and G. Kearney, "Avatar Representation in Extended Reality for Immersive Networked Music Performance," *Journal of the Audio Engineering Society*, no. 35, August 2023.

[6] S. Giacomelli et al., "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.

[7] L. Turchet, T. West, and M. M. Wanderley, "Touching the audience: Musical haptic wearables for augmented and participatory live music performances," *Personal and Ubiquitous Computing*, vol. 25, pp. 749–769, 2021.

[8] O. Pavlenko et al., "Development of music education in virtual and extended reality," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol. 13, no. 3, 2022.

[9] D. Dziwis, H. von Coler, and C. Pörschmann, "Orchestra: A Toolbox for Live Music Performances in a Web-Based Metaverse," *Journal of the Audio Engineering Society*, vol. 71, no. 11, pp. 802–812, 2023.

[10] 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 International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–10.

[11] 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.

[12] P. Małecki, J. Stefanska, and M. Szydłowska, "Assessing spatial audio: A listener-centric case study on object-based and ambisonic audio processing," *Archives of Acoustics*, vol. vol. 49, no. No 3, 2024.

[13] P. Kosior and B. Mróz, "Comparison of ambisonic and object-based spatial sound recording techniques," in *156th Audio Engineering Society Convention*, 06 2024.

[14] J.Y. Hong et al., "Spatial audio for soundscape design: Recording and reproduction," *Applied Sciences*, vol. 7, no. 6, p. 627, 2017.

[15] T. Harada et al., "Surround by sound: A review of spatial audio recording and reproduction," *Acoustical Science and Technology*, vol. 41, no. 4, pp. 164–173, 2020.

[16] C. J. McGrath, A. Franck, and H. Lee, "Multichannel microphone array recording for popular music production in virtual reality," *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 538–547, 2019.

[17] H. Engel and G. Theile, "Microphone techniques in stereo and surround recording," *Tonmeistertagung*, 2013.

[18] L. Hyunkook, "multichannel 3d microphone arrays: a review," *Journal of the Audio Engineering Society*, vol. 69, pp. 5–26, january 2021.

[19] B. Mróz, P. Odya, P. Danowski, and M. Kabacinski, "A Commonly-Accessible Toolchain for Live Streaming Music Events with Higher-Order Ambisonic Audio and 4K 360 Vision," in *AES Convention Papers*, 2023, p. Paper 14.

[20] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of The Audio Engineering Society*, vol. 45, pp. 456–466, 1997.

[21] T. Lossius, P. Baltazar, and T. de la Hogue, "Dbap–distance-based amplitude panning," in *ICMC*, 2009.

[22] S. Bech and N. Zacharov, *Perceptual Audio Evaluation-Theory, Method and Application*. Wiley, 2007.

[23] C. Guastavino and B. F. G. Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 1105–1115, 08 2004. [Online]. Available: https://doi.org/10.1121/1.1763973

[24] H. Kim et al., "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360° images," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, pp. 120–126.

[25] J. Ahrens, M. Geier, and S. Spors, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," *Journal of the Audio Engineering Society*, no. 7330, May 2008.

[26] Audio Engineering Society, *AES10-2020: AES Recommended Practice for Digital Audio Engineering — Serial Multichannel Audio Digital Interface (MADI)*, Audio Engineering Society Std., 2020, aES Standard, accessed July 2025. [Online]. Available: https://www.aes.org/publications/standards/search.cfm?docID=17

[27] M. Neukom and J. C. Schacher, "Ambisonics equivalent panning (aep)," in *Proceedings of the International Computer Music Conference (ICMC)*. Belfast, Northern Ireland: International Computer Music Association, 2008.

[28] C. Plummer, "Affordable sound field panning in theatre," *USITT Current Practices and Research in Sound*, Apr. 2020.

[29] A. Farina et al., "Spatial pcm sampling: A new method for sound recording and playback," in *AES 52nd International Conference*, 09 2013, pp. 1–12.

[30] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, Oct. 2012. [Online]. Available: https://www.aes.org/tmpFiles/elib/20250712/16554.pdf

[31] ——, "Ambisonic decoding with panning-invariant loudness on small layouts (allrad2)," in *AES 144th Convention*, no. 9943, Milan, Italy, May 2018, convention Paper.

[32] C. Nachbar et al., "AmbiX: A suggested ambisonics format," in *Proceedings of the Ambisonics Symposium 2011*. Lexington, Kentucky, USA: Ambisonics Symposium, June 2011, p. –.

[33] J.-P. Cáceres and C. Chafe, "Jacktrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010.

[34] S. Agrawal et al., "A Method for Subjective Assessment of Immersion in Audiovisual Experiences," *Journal of the Audio Engineering Society*, vol. 69, no. 11, pp. 860–873, 2021.

[35] J. W. Kelly, "Presence and the Reality of Experience," *Frontiers in Psychology*, vol. 11, p. 1468, 2020.

[36] S. Agrawal and G. Fazekas, "Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences," in *Proceedings of the Audio Engineering Society Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, 2020.

[37] E. Bates et al., "Are You There? A Literature Review of Presence for Immersive Music Reproduction," in *Proceedings of the Audio Engineering Society Conference on Immersive and Interactive Audio*, Online, 2021.