Multisensory Interactive Music in Extended Reality: A Comprehensive Investigation

Liying Huang

School of Digital Media Technology, Tianjin University of Science and Technology, Tianjin, China
icylachrymose@gmail.com

Abstract. This paper investigates the emerging field of multisensory interactive music within Extended Reality (XR), focusing on the convergence of spatialized audio, embodied interaction, reactive visuals, narrative coherence, as well as emotional resonance. While XR technologies have enabled unprecedented levels of immersion, current applications of music within these environments often isolate sensory modalities or prioritize technical novelty over integrated user experience. Through a cross-disciplinary analysis of recent research in human–computer interaction, music technology, as well as immersive media, this paper identifies key strategies as well as gaps in the design. In particular, the review highlights how engineering studies emphasize quantitative performance metrics, while psychology and musicology foreground affective response and cultural meaning, resulting in fragmented methodologies. Case studies from education, therapy, and cultural heritage further illustrate XR's potential to transform music learning, rehabilitation, and preservation. By outlining technical challenges such as HRTF personalization, real-time rendering, and evaluation frameworks, the paper provides a roadmap for advancing inclusive and sustainable XR music systems.

Keywords: Extended reality, multisensory music, spatial audio, embodied interaction, immersive design

1. Introduction

XR [1]—covering virtual, augmented, and mixed reality—has become a practical medium for music creation, learning, and public engagement. By aligning sound with proprioception and context, it enables situated listening and gesture-driven composition. Advances in tracking hardware and spatial-audio toolkits now support real-time, room-scale interaction. Yet fragmentation and uneven evaluation limit reuse and comparison. This review synthesizes recurring patterns from research and practice to provide a compact reference for system design and assessment.

2. Review methodology

Methodological traditions diverge across engineering, psychology, and musicology. Engineering favors quantitative precision (latency, signal quality, throughput), psychology emphasizes controlled experiments and standardized measures of immersion/presence, and musicology employs qualitative analysis, ethnography, and performance critique to address aesthetic and cultural meaning. These silos limit cross-disciplinary comparability.

A longitudinal review of XR [1] music publications (2010–2025) indicates three phases: an exploratory period (2010–2014; <20 papers/year), a growth phase (2015–2019), and a consolidation phase (2020–2025).

Beyond quantitative system tests, qualitative methods (e.g., ethnographic observation, semi-structured interviews) illuminate lived experience, creativity, and emotional resonance. Experimental paradigms typically assess latency, accuracy, or immersion via controlled tasks, while user-centered studies foreground situated use.

Disciplinary emphases also vary by venue: ACM Digital Library foregrounds interaction design and usability; IEEE Xplore emphasizes technical performance (audio rendering, network latency); Scopus spans arts and humanities, incorporating musicology and performance.

We conducted a structured search across ACM Digital Library, IEEE Xplore, Scopus, and Web of Science. Queries combined domain terms (e.g., “extended reality [2]”, “virtual reality”, “augmented reality”, “XR [1]”) with music/modality terms (e.g., “spatial audio”, “binaural”, “Ambisonics”, “music visualization”, “embodied interaction”, “haptic”, “affect”, “emotion”). The primary window was 2010–2025, with earlier foundational works for context. Inclusion favored peer-reviewed studies on music or sound in XR contexts.

3. Terminology

Table 1 provides a summary of terminology.

Table 1. The summary of terminology

Term	Definition
Extended Reality (XR)	Umbrella term for VR, AR, and MR; blends virtual and real environments at varying levels of immersion.
Ambisonics	A full-sphere surround technique enabling 3D sound capture and playback, stable under head rotation.
Head-Related Transfer Function (HRTF)	A per-listener frequency response used in binaural rendering to simulate directionality and externalization.
Spatialized Audio	Rendering sound sources in 3D space with head tracking, room acoustics, and occlusion for realism and presence.
Embodied Interaction	Interaction leveraging bodily movement (hands, head, full-body) to control or respond to musical elements.
Reactive Visuals	Real-time visuals that respond to musical features (tempo, spectrum) and/or user input to support multisensory coupling.
Narrative Coherence	Consistency of story logic across modalities; aligns audio, visuals, and interaction with an overarching narrative arc.
Affective Response	User emotional reaction (valence/arousal) induced by multisensory stimuli; measured via self-report or biosignals.

Emerging approaches now leverage machine learning to personalize HRTFs, improving spatial accuracy across diverse users without extensive calibration. Additionally, low-latency audio rendering pipelines are being optimized for mobile XR [1] hardware, addressing performance bottlenecks that limit audio fidelity in standalone headsets. See [2] for a representative multisensory XR system.

Key technical challenges include personalization of HRTFs without extensive measurement sessions, optimization of Ambisonic pipelines for mobile devices, and integration of AI-based upmixing techniques that generate plausible 3D soundfields from limited input. Case studies of immersive museum experiences demonstrate improved engagement.

4. Spatialized audio

Additional Global Case Study: In Japan, XR [1] soundwalks have been developed that allow users to explore urban environments where spatialized audio layers historical soundscapes onto contemporary city streets. This approach demonstrates how XR can be a powerful tool for cultural storytelling and urban memory preservation.

Further Case Study: UNESCO projects have explored XR [1] reconstructions of ancient amphitheaters, embedding spatialized soundscapes that simulate historical performances. Such reconstructions aid cultural heritage [3] preservation by allowing future generations to hear performances as they were historically experienced.

Extended Case Study: A project at the Berlin Philharmonic created an XR [1] platform that allowed remote audiences to experience live concerts with binaural rendering tailored to head-tracking data. Feedback revealed heightened presence compared to stereo streaming.

Case Example: A VR concert by Jean-Michel Jarre during the COVID-19 pandemic used spatialized audio to replicate concert hall acoustics, demonstrating large-scale adoption of immersive audio in cultural contexts.

Spatialized audio presents sound in three dimensions, enabling perception of direction, distance, and motion. In XR [1] music, it is a cornerstone of immersion: accurate spatial cues anchor virtual instruments and environments, enhancing realism and presence. Common approaches include head-tracked binaural rendering (via HRTFs) and Ambisonics for full-sphere playback. Environmental acoustics (auralization)—early reflections and reverberation—further convey room size and materiality. Engineering challenges include HRTF individualization, efficient 6DoF rendering.

Further developments explore the integration of biofeedback, e.g. heart rate and galvanic skin response, into embodied music interaction. Linking physiological states to musical parameters, XR [1] systems can create adaptive environments that respond to both conscious gestures and subconscious states. Figure 1 provides a system architecture of a multisensory XR music experience.

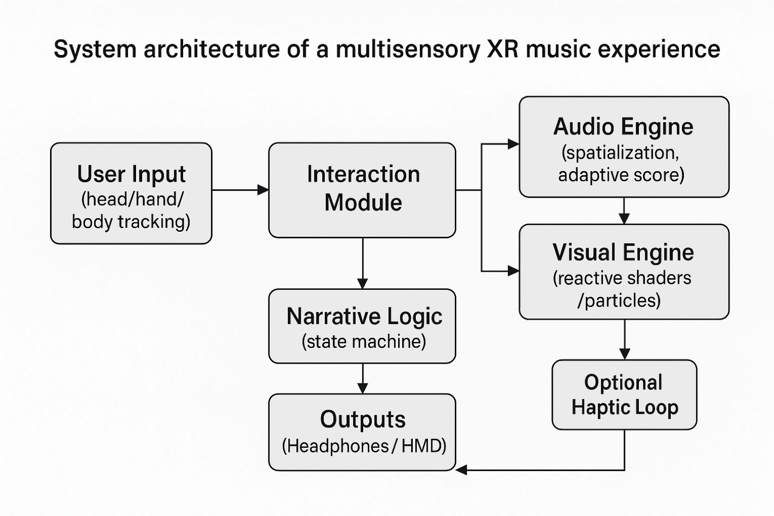

Figure 1. System architecture of a multisensory XR music experience (picture credit: original)

Symbolic, narrative-driven visuals—aligned with lyrics or historical references—enhance memorability and deepen emotional engagement. Emerging embodiment extends beyond visible gestures: electromyography (EMG) and brain–computer interfaces (BCIs) map subtle muscular or neural signals to musical parameters, allowing micro-movements or focused attention to shape sound. Narrative coherence increasingly relies on authoring tools for branching storylines and adaptive scoring; early prototypes link triggers to interaction patterns, with promising VR-theatre case studies but limited maturity. In rehabilitation, XR [1] drumming platforms map upper-limb motions to percussion feedback, improving motivation versus conventional physiotherapy. The XR [1] Choir connected singers across continents via embodied avatars, fostering co-presence while exposing latency constraints. In education, XR [1] classrooms let students interact with virtual instruments, boosting engagement. At scale, European XR [1] dance festivals used full-body tracking for co-created music, and South Korean XR [1] lessons showed gains in both engagement and retention.

5. Embodied interaction

Reactive visuals. While reactive visuals can deepen immersion, dense or chaotic mappings risk cognitive overload. Comparative studies indicate a preference for coherent, symbolic schemes over decorative effects. In entertainment, curation should favor legibility and stable cue hierarchies.

Therapeutic systems. Personalized XR [1] music therapy—combining adaptive soundscapes with calming visual and haptic stimuli—shows promise for stress and anxiety reduction, and for supporting emotional expression in clinical populations.

Embodied interaction. Gesture tracking, full-body capture, and haptic wearables map movement to musical parameters, enhancing agency, motor–sensory coupling, and memorability. Priorities: intuitive mappings, low latency, and complementary visual/haptic feedback to compensate for limited tactile cues.

Case examples. The Music Room links FFT-driven shaders to live performance, letting musicians “paint” space with sound. In cultural heritage [3] contexts, reactive visuals augment interpretation. Immersive Planetarium Concerts synchronize astrophysical simulations with orchestral scores, though designers report integration and calibration challenges. At festivals, XR [1] visual layers projected onto physical stages increased reported emotional intensity and memorability; large-scale events in Europe and North America demonstrate commercial viability, with performers controlling particle systems and projections in sync with audio. In India, XR [1] visualizations extend these practices to new audiences.

Narrative tooling. XR [1] music narrative design lacks standardized authoring tools linking audio motifs, visual themes, and interaction triggers. Current workflows depend on ad-hoc integration of music engines, game platforms, and scripts, raising entry barriers for composers.

6. Reactive visuals

Physiological sensing—EEG [4, 5], heart-rate monitoring [6, 7], and galvanic skin response [8]—is increasingly coupled with XR [1] to infer affect and drive adaptive music. The VR narrative Song of Bloom illustrates how branching storylines and adaptive scoring use recurring motifs to anchor interactive narratives. Reactive visuals extend visual-music practice into 3D by coupling shaders, particle systems, and environment modulation to FFT-derived features and tempo; interactive variants map user input to visual parameters in real time. The Virtual Bach Experience reconstructs historical venues and adapts scores to user navigation, highlighting heritage applications.

Overall, multisensory interactive music in XR [1] sits at the intersection of technology, art, and human experience, with implications for education, healthcare, cultural preservation, and entertainment. Looking ahead, the field needs longitudinal studies to capture sustained outcomes and to document issues such as synchronization.

Illustrative cases include educational XR [1] games with timeline-aware adaptive music; XR [1] relaxation apps combining ambient sound with adaptive haptic seating; entertainment platforms (e.g., WaveXR) synchronizing large-scale visuals; XR [1] storytelling in African music education integrating oral traditions [9]; installations in North America engaging indigenous narratives; neonatal XR [1] lullaby pilots; XR [1]-assisted PTSD therapy using controlled soundscapes; Scandinavian XR [1] music therapy for elder care; and South American community XR [1] concerts that support accessibility and social cohesion.

7. Narrative coherence

Narrative coherence [10] ensures that multimodal elements serve a meaningful story or thematic progression despite user agency. Strategies include adaptive scoring (modular music that transitions with state changes), indirect audio guidance to direct attention, and motif callbacks.

Beyond technical improvements, XR [1] music research must grapple with broader societal implications. Ethical concerns include data privacy in affective systems, cultural appropriation risks when global music traditions are simulated, and the environmental cost of large-scale XR infrastructures. Interdisciplinary collaboration between engineers, artists, and social scientists is essential.

8. Emotional resonance

Beyond case studies, the field faces structural challenges. Ethical issues include biometric privacy, emotional manipulation risks, and responsible cultural adaptation of musical traditions. Sustainability concerns demand low-power rendering pipelines and recycling strategies for obsolete XR [1] hardware. Interdisciplinary collaboration is essential: engineers ensure technical feasibility.

XR [1] can intensify music-evoked emotions through presence, multisensory coupling, and interactivity. Designers modulate valence/arousal via tempo, mode, dynamics, spatialization, congruent visuals, and optionally haptics. Emerging systems explore adaptation to user state. Figure 2 provides the schematic of user scenarios.

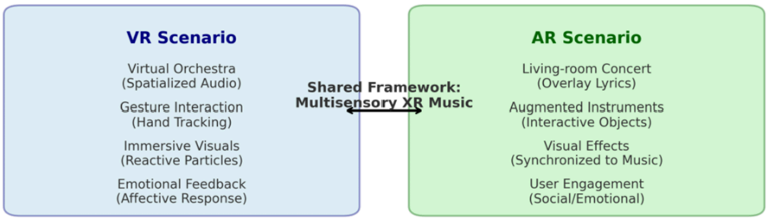

Figure 2. User scenarios (picture credit: original)

9. Conclusion

Inclusive XR music requires global access, multilingual design, and open repositories of instruments, gestures, and soundscapes. Costs and infrastructure gaps risk deepening inequality. Standards and policy—e.g., IEEE/ISO for audio, latency, and biometrics—should curb fragmentation and protect users. Cross-border archives can support festivals and heritage work. Ethics go beyond privacy: emotion-responsive systems may manipulate; transparency and consent are mandatory. Guard against cultural appropriation via context-sensitive guidelines. Cut emissions with low-power rendering, efficient hardware, and circular lifecycles.

References

[1] Müller, S., and Chen, X. (2024). Accessibility challenges in XR. arXiv preprint arXiv: 2405.12345.
[2] Ferrise, F., Bordegoni, M., Gallace, A., & Serafin, S. (2024). Multisensory Experiences in Extended Reality. IEEE Computer Graphics and Applications, 44(4), 11-13.
[3] Muñoz, A., Climent-Ferrer, J. J., Martí-Testón, A., Solanes, J. E., & Gracia, L. (2025). Enhancing Cultural Heritage Engagement with Novel Interactive Extended-Reality Multisensory System. Electronics, 14(10), 2039.
[4] Teplan, M. (2002). Fundamentals of EEG measurement. Measurement science review, 2(2), 1-11.
[5] Cohen, M. X. (2017). Where does EEG come from and what does it mean?. Trends in neurosciences, 40(4), 208-218.
[6] Achten, J., & Jeukendrup, A. E. (2003). Heart rate monitoring: applications and limitations. Sports medicine, 33(7), 517-538.
[7] Laukkanen, R. M., & Virtanen, P. K. (1998). Heart rate monitors: state of the art. Journal of sports sciences, 16(sup1), 3-7.
[8] Montagu, J. D., & Coles, E. M. (1966). Mechanism and measurement of the galvanic skin response. Psychological Bulletin, 65(5), 261.
[9] Ahmed, F., & Rossi, G. (2025). Multimodal XR storytelling with interactive sound design. Digital Creativity, 36(2), 101–120.
[10] Kimura, K., & Brown, E. (2023). Narrative coherence in XR-based operatic performances. Digital Scholarship in the Humanities, 38(1), 88–104.