A SPATIAL AUDIO SYSTEM FOR CO-LOCATED MULTI-PARTICIPANT EXTENDED REALITY EXPERIENCES

Yi Wu[1], *Agnieszka Roginska*[1], *Keru Wang*[2], *Zhu Wang*[2], *Ken Perlin*[2]

[1] Music and Audio Research Laboratory, New York University, New York, USA
{yw5759, roginska}@nyu.edu
[2] Future Reality Lab, New York University, New York, USA
{keru.wang, zhu.wang, perlin}@nyu.edu

ABSTRACT

This paper presents the ongoing development of a spatial audio system for co-located, multi-participant, extended reality (CM-XR) experiences. By integrating spatial audio and informative auditory displays, the system can enhance the sense of immersion and presence among participants and facilitate collaboration. We navigate this development by addressing the challenges encountered during the implementation of spatial audio in server-client frameworks and CM-XR environments, including accurate audio localization, synchronization across multiple users, and designing immersive and informative sounds. The paper covers the architecture of the system, along with technical features, sound design choices, and the use of audio for providing information to participants about the environment. This exploration offers insights into spatial audio integration and suggests directions for future research in the design and implementation of spatial auditory display solutions for CM-XR settings.

1. INTRODUCTION

The emergence of extended reality (XR) has revolutionized our interaction with digital content. Among the various facets of XR technology, co-located, multi-participant XR experiences (CM-XR) stand out for their ability to foster a shared sense of presence and collaboration [1, 2]. These experiences blend shared virtual objects with physical reality, enabling collaborations and interactions within a joint environment.

The shift towards interactive and communal virtual-physical hybrid spaces underscores a transformation in underlying design philosophies. Moving from an egocentric design, which centers on individual XR experiences for personal engagement and focuses on the user's own perspective, there is now an increased emphasis on allocentric design. Allocentric design highlights user/user and user/environment relationships, emphasizing shared experiences that are fundamental to CM-XR settings [3].

In XR experiences, spatial audio and auditory displays are essential for enhancing immersion and providing essential information, such as the locations of other users and virtual objects. When auditory displays are spatialized, they provide more intuitive and effective user interactions with virtual objects and the environment by leveraging the human auditory system's natural spatial hearing abilities [4, 5].

Figure 1: The XR gallery features four thematic areas: 1. "Math and Art," where participants explore the intersection of creativity and logic through educational objects. 2. "Interactive Games and Puzzles," fostering teamwork and problem-solving through collaborative 3D puzzles. 3. "XR Interaction with Physical Objects," integrating tangible objects into the virtual experience. 4. "Experiences of Wonder," transporting participants to magical realms with breathtaking phenomena.

However, adopting an allocentric approach to XR system design, particularly in developing spatial audio systems for CM-XR environments, presents unique challenges. These include accurate audio localization, synchronization among multiple users, and designing sounds that are both immersive and informative. Addressing these challenges requires a careful balance of technical expertise and creative sound design, as well as a deep understanding of how spatial audiotry cues influence user perception and interaction within XR environments.

In this paper, we present the ongoing development of a spatial audio system tailored for WebXR-based CM-XR experiences, featuring an innovative XR gallery paradigm (Figure 1). By integrating spatial audio and informative auditory displays, we aim to amplify the sense of immersion and presence among participants and to facilitate collaboration. We discuss the challenges of implementing spatial audio within server-client frameworks and XR environments. In addition, we detail the system's architecture, technical features, sound design choices, and the role of sounds in informing participants about the elements and collaborators in the environment.

2. SYSTEM ARCHITECTURE

Our CM-XR system is built with the WebXR Application Programming Interface (API) [6], enabling CM-XR experiences directly through web browsers, eliminating the need for app downloads, and ensuring compatibility across different devices (although currently, we are focusing only on the Meta Quest 3 with video pass-through mode [7]).

The CM-XR system's server architecture, designed and developed with Node.js, facilitates client communication. The server utilizes a shared blackboard approach, incorporating a series of state variables that are synchronized across all clients to ensure uniformity. Whenever any client adjusts the value of a state, the server swiftly broadcasts this change to every other client through WebSockets. For example, if a client creates or moves a sound source, the change is broadcast to all clients, allowing them to hear the sound source at the updated location in real time.

Our spatial audio system is built with the Google Resonance Audio SDK for Web (Resonance Audio) [8]. Resonance Audio is a real-time JavaScript SDK designed to create immersive sound experiences specifically for web applications. It is powered by Google Omnitone [9], an implementation of Ambisonic decoding and binaural rendering written in the Web Audio API, thereby ensuring compatibility with WebXR. This allows for a coherent development process across our system's visual and auditory components.

3. TECHNICAL FEATURES AND SOUND DESIGN

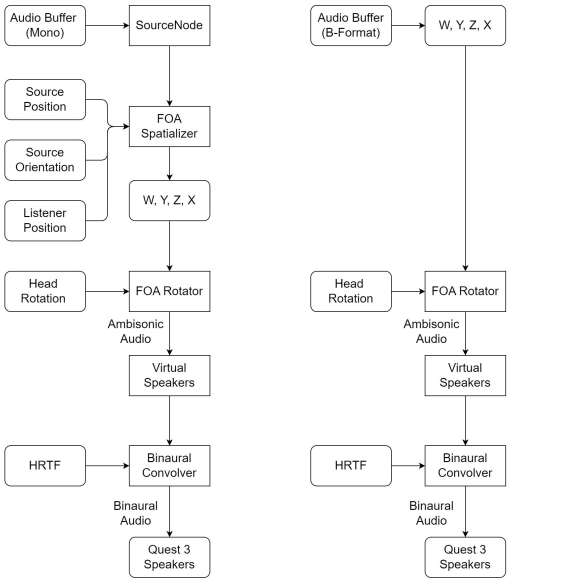
3.1. Spatial Mapping and Calibration

To synchronize all CM-XR users within the same virtual space on the server, we utilize the boundary information stored in each headset to establish a common frame of reference for every user. The Meta Quest 3 headset computes the relative positions of boundary points in relation to its native origin point. By recording this positional data on the server and regularly comparing the changes between previously saved positions and their updates upon each headset's origin point reset, we are able to create and maintain a unified global coordinate system. Our spatial audio subsystem shares the same global coordinate space as our graphical subsystem, which allows us to achieve consistent visual-audio experiences by ensuring that spatial audio cues are consistent with visual locations.

3.2. Localization

In CM-XR settings, being aware of other users and virtual objects' locations is crucial. However, this awareness cannot rely solely on visuals since human vision is not 360 degrees, and current headset technology does not fully match human field of view, limiting location awareness. Therefore, accurately conveying location information through spatialized auditory displays is essential.

Sound localization involves accurately presenting a sound source's location from the listener's perspective within the virtual environment, ensuring that auditory cues align with their spatial relationships. Our system employs an allocentric spatial presentation, using world-space coordinates that are converted to an egocentric reference based on the user's position at rendering time [10], to spatialize the sound source. The system dynamically captures the user's position and head rotation from the headset, calculating each sound source's location relative to the listener in real-



(a) Spatial Emitter Audio Pipeline (b) Ambisonic Ambience Audio Pipeline

Figure 2: Figure 2(a) shows the spatial audio pipeline for the emitter sound source playback. Figure 2(b) shows the Ambisonic audio pipeline for the ambience (Ambisonics B-format) playback.

time (Figure 2). This process considers the positions and orientations of both the source and listener within the global coordinate system, enabling precise rendering of sound directionality.

Additionally, our spatial audio system uses distance-based attenuation to mimic the natural reduction in sound volume with increased distance from the source. It calculates and adjusts attenuation based on the listener's distance from the sound source, thereby enhancing audio realism by simulating distance perception.

By aligning auditory cues with the visual environment, our system ensures an immersive experience. Sound localization becomes an intuitive and effective way to deliver the locations of virtual objects and users, leveraging the human auditory system's innate ability for sound localization.

3.3. Sound Source Directivity

Directivity indicates the orientation of a sound source. Directivity is achieved with Resonance Audio through directivity patterns, including pattern shape (such as Cardioid) and pattern width. Ideally, an omnidirectional sound source emits sound uniformly in all directions. However, most real-life sound sources exhibit directivity, especially at higher frequencies [5, 11]. For instance, a guitar sounds different from the back than from the front due to its body occluding the sound hole and strings. Thus, when orientation information for a virtual object or a user interaction is required, it can be conveyed through the directivity feature of the spatial audio system. Including directivity in the audio representation of virtual objects, especially those with real-life analogs such as musical instruments, enhances immersion as it aligns with the listener's real-world auditory experience. Furthermore, directivity also adds an additional dimension of artistic expression. For example, in one of our scenarios (Figure 3), a window looks out onto alien worlds that are invisible from the reverse side. To complement the views,

Figure 3: A window looks out onto alien worlds that are invisible from the reverse side. Sounds are heard only from the front side of the window.

sounds are heard only from the front side of the window, which enhances the scene's magical essence.

3.4. Acoustic Simulation (Early Reflections and Late Reverberation)

Acoustic simulation plays a key role in integrating virtual objects with the physical world, enhancing the realism of auditory delivery and conveying spatial information like room size. Elements like early reflections and late reverberation are crucial in delivering a perception of the room's size and characteristics [5, 12]. Early reflections are the initial sound reflections that reach the listener's ears after the direct sound, coming from surfaces close to the sound source, such as walls, ceilings, and floors, helping the listener locate the sound source in space. Late reverberation, or late reflections, are multiple diffuse reflections occurring after many bounces, providing a sense of the space's size and liveliness of the environment [5].

Our spatial audio system utilizes Resonance Audio's geometry-based reverb engine to accurately simulate these acoustic properties, allowing for realistic space emulation. Prior to the XR experience, the dimensions of the room (a rectangular cuboid) and materials of the walls (e.g., plywood-panel, concrete-block-coarse, curtain-heavy) are manually input into the reverb engine. This enables customization of the environment's acoustic properties to align with visual cues, ensuring a seamless integration of virtual objects with the physical space.

3.5. Ambisonic Ambience Playback

Ambience (ambient sound) encompasses all secondary sounds in an environment or scene, excluding the primary sounds being focused on. It creates an auditory backdrop that delivers the sense and atmosphere of the place. Ambisonics is a sound recording and reproduction technique that captures and reproduces a full and accurate representation of sound waves. This involves decomposing the sound field into spherical harmonics and encoding sounds from all directions [5, 13]. Unlike emitter sounds, ambient sounds come from everywhere, making Ambisonic recording reproduction well-suited for ambient sound playback, thus providing a fully immersive environment. Ambience playback in our system follows a different pipeline than emitters playback (Figure 2); the Ambisonic audio files are fed directly to the rotator (where the entire sound field is allowed to rotate based on the user's head orientation). Our CM-XR experience can support scenes such as "snowfall", which immerses users within a volume of snowflakes dancing in the room. This requires Ambisonic ambience playback to complement the 360-degree visuals.

3.6. Music Playback

Music significantly enhances the CM-XR experience by adding artistic depth, emotional resonance, and interactive feedback. Additionally, music serves as an intuitive auditory cue within the virtual environment to guide users, signal changes, or indicate other users' actions. In our audio system, music is primarily played in stereo through the Web Audio API but can also be spatialized using Ambisonic audio files to meet specific design requirements.

3.7. Binaural Rendering

Human auditory spatial perception relies on interaural time and level differences (ITDs and ILDs) [14] and spectral cues [15, 16] from sound interaction with head, torso and pinnae. These cues are captured in Head-Related Transfer Functions (HRTFs) [5, 17], which are essential for precise sound localization in headphone-based binaural rendering.

Our spatial audio utilizes Resonance Audio for binaural rendering. After the rotators, the Ambisonic stream (whether from the spatial emitter or the Ambisonic ambience audio pipeline) is reproduced with multiple virtual speakers. Each speaker channel is then convolved with Resonance Audio's built-in HRTFs, based on its virtual location, to render the final binaural audio stream (Figure 2). This final stream is then played through the built-in speakers of the Meta Quest 3 (Figure 2).

3.8. Sound Design

For interaction sounds, such as creating objects in specific locations, we choose short, full-spectrum sounds because simple harmonic sounds, such as pure tone sine waves, can challenge user localization. As a result, we craft these sound effects from broadband sounds combined with organic samples, which increases familiarity and eases localization. For instance, the "create ball" sound in the "Construct" scene (Figure 4) combines a synthesized click with a champagne bottle popping sound. With sound effects that are easily localized, users can more readily discern other participants' interactions and the locations of these interactions, even outside their field of view.

We design stationary objects that need locational indication (such as helping users to distinguish between virtual and real objects), whether animated or not, to emit subtle humming sounds. These sounds, quiet yet full-spectrum, help users to easily identify object positions without causing annoyance.

Ambience utilizes real-world Ambisonic recordings for a more realistic and immersive experience. These files are amplified beyond background levels to either dominate or balance with environmental sounds, thereby augmenting the user's experience of the physical space.

4. CONCLUSION AND FUTURE DIRECTIONS

The spatial audio system significantly enriches the XR experience by enhancing immersion and providing essential information

Figure 4: The "Construct" scene allows users to collaboratively create and move balls in 3D space.

through spatialized auditory displays. These displays enable more intuitive and effective user interactions with both the environment and other users.

While the system can simulate occlusion, this feature has not yet been implemented for participant interactions. Occlusion could introduce more realism by allowing users' bodies to affect sound propagation. This addition could enrich the realism and introduce new interactive possibilities by more accurately informing users about the locations of other users and sound sources.

We observed that the effectiveness of our spatial audio system's sound localization was compromised during testing due to audio leakage from the Meta Quest 3's built-in speakers. This issue can be readily addressed by using open-back headphones to enhance the experience, although this solution may introduce potential discomfort due to additional weight.

Designing spatial audio systems and auditory displays showcases distinct approaches between CM-XR and single-player XR experiences. In single-player XR, the priority is to enhance the individual user's immersion with audio that enriches the narrative. For CM-XR, beyond immersion and engagement, spatial audio and auditory displays must facilitate user interaction and collaboration, requiring spatial auditory cues to signal the presence and actions of other participants. Such systems need to balance the global experience with personalization. In multi-participant settings, it is important to identify which sounds are relevant to specific users, thus categorizing auditory displays. For example, a menu sound might only be audible to the user who activates it, to avoid disruption in environments where multiple people interact with the menu. Furthermore, multi-participant experiences can benefit from different levels of auditory display tailored to distinct groups, enhancing the customization of categorical auditory displays. This ensures that the auditory display is affected only for those needing the information, minimizing clutter. For instance, in a workshop scenario with student and instructor groups, some sounds might be exclusive to students, while others are only for instructors, thereby providing clear and relevant auditory cues for each group. We are exploring these categorical approaches, aiming to unlock new possibilities for auditory displays in CM-XR environments.

5. REFERENCES

[1] V. Pereira, T. Matos, R. Rodrigues, R. Nóbrega, and J. Jacob, "Extended reality framework for remote collaborative interactions in virtual environments," pp. 17–24.

[2] K. Wang, Z. Wang, K. Rosenberg, Z. He, D. W. Yoo, U. J. Christopher, and K. Perlin, "Mixed reality collaboration for complementary working styles," publisher: Association for Computing Machinery, Inc.

[3] S. Bae, H. Lee, H. Park, H. Cho, J. Park, and J. Kim, "The effects of egocentric and allocentric representations on presence and perceived realism: Tested in stereoscopic 3d games," vol. 24, no. 4, pp. 251–264. [Online]. Available: https://doi.org/10.1016/j.intcom.2012.04.009

[4] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," vol. 42, pp. 135–159.

[5] Agnieszka Roginska, Paul Geluso, *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. Routledge.

[6] Immersive web developer home. [Online]. Available: https://immersiveweb.dev/

[7] Full-color passthrough on meta quest | meta store. [Online]. Available: https://www.meta.com/help/quest/articles/getting-started/getting-started-with-quest-pro/full-color-passthrough/

[8] Resonance audio - resonance audio SDK for web. [Online]. Available: https://resonance-audio.github.io/resonance-audio/develop/web/getting-started.html

[9] "GoogleChrome/omnitone," original-date: 2016-06-02T17:21:26Z. [Online]. Available: https://github.com/GoogleChrome/omnitone

[10] S. M. Town, W. O. Brimijoin, and J. K. Bizley, "Egocentric and allocentric representations in auditory cortex," vol. 15, no. 6, p. e2001878, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2001878

[11] R. Mehra, L. Antani, S. Kim, and D. Manocha, "Source and listener directivity for interactive wave-based sound propagation," vol. 20, no. 4, pp. 495–503, conference Name: IEEE Transactions on Visualization and Computer Graphics. [Online]. Available: https://ieeexplore.ieee.org/document/6777442

[12] A. Rungta, N. Rewkowski, R. Klatzky, and D. Manocha, "P-reverb: Perceptual characterization of early and late reflections for auditory displays," pp. 455–463.

[13] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*.

[14] L. Rayleigh, "XII. on our perception of sound direction," vol. 13, no. 74, pp. 214–232.

[15] W. Yost and R. Dye, "Fundamentals of directional hearing," vol. 18, pp. 321–344.

[16] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," vol. 56, no. 6, pp. 1829–1834.

[17] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," vol. 94, no. 1, pp. 111–123.