Immersive Application for Real-Time Interactive Music Performances Using Spatial Audio

Gianluca Fadda, Maurizio Murroni, Marian Alexandru, Nicușor Amarie, Vlad Popescu, Daniele Giusto

***Index Terms*—XR Audio, audio-visual systems, Networked Music Performance, Real-Time Interaction, Quality of Experience, Multimedia Coding.**

***Abstract*—This paper proposes the implementation of an Extended Reality (XR) application that enables real-time streaming of live music concerts, where users equipped with XR headsets can interact with musicians in immersive virtual spaces. Musicians are represented by avatars, each associated with one or more audio tracks captured by a digital mixer. Using spatial audio and real-time interaction mechanisms, the system provides an engaging and dynamic concert experience, simulating the presence of a live audience while offering an user-controlled feedback channel. In the proposed system, multitrack audio streams from the musicians' mixer are transmitted to the XR environment via low-latency streaming protocols. These audio tracks are spatialized in the 3D space using advanced algorithms such as Head-Related Transfer Function (HRTF)-based rendering or Ambisonics within typical 3D game engines (e.g. Unity, Unreal). Each musician's avatar, positioned in the virtual environment, dynamically updates its location and sound directionality relative to the musician's position and orientation. Audience members can interact with musicians in real time by providing audio feedback (e.g. cheering, clapping, singalong), which are reflected on the musician's side by spatialized audio crowd responses.**

I. Introduction

The pandemic conditions witnessed few year ago, pushed a still growing worldwide interest for remote and distributed audio production applications, both at industrial [1], [2], [3] and academic [4], [5], [6] level. Indeed, since 2020, the number of experimental events such as Networked Music Performances (NMP) and live streaming of musical concerts to a remote audience increased exponentially [7], [8], due also to the technological progress of hardware and software technologies supporting NMP [9], [10] (e.g. broadcasting platforms, audio spatial mics, audio routing protocols, 360 and depth cameras, etc.) and higher bandwidth availability [11] (e.g. FTTH, 5G, etc.).

Gianluca Fadda, Maurizio Murroni, and Daniele Giusto are with the University of Cagliari, UdR CNIT of Cagliari, Italy, (e-mail: gianluca.fadda@unica.it, m.murroni@unica.it, ddgiusto@unica.it).

Marian Alexandru, Nicusor Amarie and Vlad Popescu are with the University of Brașov, Romania (e-mail: marian.alexandru@unitbv.ro, nicusor.amarie@gmail.com, vlad.popescu@unitbv.ro)

A. Networked Music Performances

As previously stated, NMP applications have been a real game changer in the past four years, enabling the collaboration in terms of live performances between musicians situated in different geographical locations. The main challenge remains though the level of engagement and audio realism that is still far from the one provided by an in-person performance. Creating a natural and engaging auditory experience in these situations is still a challenging task due to latency, spatial disconnection, and lack of perceptual immersion. Therefore, current research and implementation efforts address the following essential factors: spatial audio, low latency and jitter thresholds to enhance the in-person experience within networked environments.

B. Spatial Audio Techniques

In terms of spatial audio technologies, for the realistic recreation of three-dimensional soundscapes, the HRTF [12] and Ambisonics [13] techniques have recently become prominent in NMP systems. Commercially available systems such as Steam Audio [14] and Resonance Audio [15] deliver the possibility to have real-time spatial rendering that dynamically adapts to user orientation and source positions. The integration of spatial audio into NMP has been discussed in [16], showing how 3D audio increases the sense of presence and immersion for remote performers. Other works implement hardware-optimized solutions, [17] (FPGA-based) and [18] (SOC-based), that look promising for reaching real-time spatial audio rendering with very low computational latency. These systems enable an accurate spatial positioning of audio sources for the musicians, enhancing ensemble synchronization and spatial awareness.

C. Low Latency and Jitter Thresholds

Latency is still the main limitation of NMP systems considering the fact that the perceptual thresholds for acceptable delay are around 20–30 ms for musical synchronization [9]. According to [19], this threshold is critical in real-time performances to maintain rhythmic cohesion. Buffer optimization, packet prioritization, and edge computing are some of the strategies explored to minimize latency in audio transmission. State-of-the-art systems like Soundjack [20] and LoLa [21], which allow very low latencies under optimized network conditions, are among the most used for high-performance musical collaboration. Recent developments in the optimization of 5G networks has led

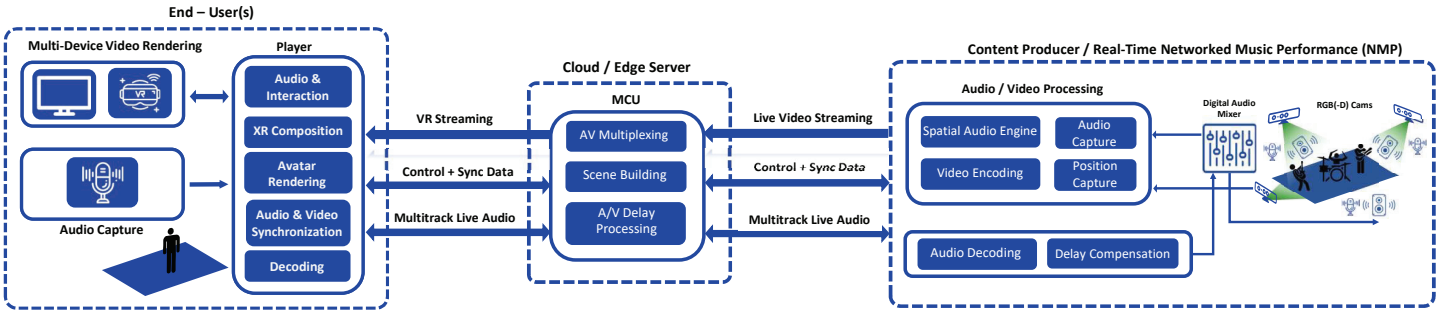

Fig. 1. Proposed Architecture

to the inclusion of this communication standard for further reducing transmission delays and ensure that quality remains consistent across geographically dispersed musicians [22].

D. Enhancing In-Person Engagement

While NMP systems focus mainly on latency reduction, recent studies and applications go further to replicate the physical and emotional engagement of in-person performances. The use of XR technologies has led to the creation of immersive environments where musicians and audiences can interact visually and aurally. XR concerts like those depicted in [3] and projects like MINERVA [23] aim to demonstrate how real-time spatial audio and visual data can enhance presence and engagement, creating a shared virtual stage for performers. Moreover, systems leveraging depth sensors (e.g., Microsoft® Kinect™ and Intel® RealSense™ LiDAR) dynamically adjust spatial audio rendering based on the musicians' movements and orientations, further bridging the gap between physical and networked performances [16], [18].

This paper aims to introduce and discuss the key aspects behind the design and implementation of immersive applications (AR/VR/XR) that enable real-time streaming of live music concerts, where users equipped with suitable hardware (i.e. Smart TV with in-built camera and mic, PC, smartphone, XR headset) can provide feedback to the musicians through the virtual environment. This research has two main scopes to investigate:

1) the bi-directional feedback mechanisms that maximize the user experience on both sides of the application (end-users hearing and seeing the concert and musicians playing the songs);

2) the real time spatial audio generation approaches based on spatial data.

This extended abstract presents the developed architecture, describes its components and draws initial implementation conclusions based on a list of critical challenges. The performed tests, including the Quality of Experience (QoE) evaluation, will be presented in the final paper.

II. Proposed Architecture

A. Architectural Description

The proposed architecture, depicted in Figure 1, is divided into three main layers: 1) *Content producer/ Real Time NMP*, 2) *Cloud/Edge Server*, 3) *End User(s)*. On the *Content Producer* side, the band performance is live video captured with the aid of an array of at least four RBG-Depth cameras, in our specific case the Intel® RealSense™ LiDAR Camera L515. The RGB data, together with the information related to the depth, are fed to the *Position Capture* block, where the data are preprocessed in real-time in order to generate the complete 3D scene, identifying the position of each musician and allowing the extrapolation of the spatial information which is fed to the *Video Encoding* block and also to the *Spatial Audio Engine* block. The multitrack audio data is captured via microphones or directly form the instrument amplifiers through a *Digital Audio Mixer* (e.g. a Behringer X32 Compact mixer) and fed channel-wise to the *Audio Capture* block. Each channel represents one instrument, the singer's voice being captured also on a separate channel. In case of the drums, where more microphones are used to capture the sound, the different sound sources are mixed down internally on a unique channel and fed as such further to the *Audio Capture* block.

The output of the main *Content Producer/Real-Time NMP* layer consists therefore of the live video data, the depth data, the multitrack live audio data, the control data and the synchronization data.

The *Cloud/Edge Sever* layer implements the necessary mechanisms to synchronize the audio and video data (the *A/V Multiplexing* block) and to build the entire 3D scene (the *Scene Building* block). All the mechanisms are implemented in the cloud, using a micro-services based architecture (e.g. Docker/Kubernetes) and a real-time communication protocol (e.g. WebRTC [24]). The avatars of the musicians are dynamically built in terms of texture and position.

The *End User* layer receives the data from the cloud through the *Decoding* block and builds the 3D scene for the final user, rendering the content for multiple devices, such as 3D headset or 2D screen, by means of the *XR Composition* and the *Avatar Rendering* blocks. The user is interacting with the scene through the rendering device, as follows: for the 3D headset the movement of the user is extrapolated directly by using the

data delivered by the headset, while for the 2D screen, the eyegaze of the viewer is analyzed by a 2D camera (dedicated camera or cellular phone), using specific algorithms [25], [26]. The end user's audio is captured by using the 3D headset's or the built-in camera's microphone, correlated with the user's movement in the *Audio-Interaction* block and synchronized by the *Audio-Video Synchronization* block. The user's audio feedback, as a separate audio channel, is subsequently forwarded to the *Cloud/Edge Server* layer where the delay is furthermore compensated by the *A/V Delay Processing* block and then forwarded to the live performance. The position of the user is transmitted as spatial metadata through the *Control+Sync Data* channel.

In the *Content Producer/Real-Time NMP* layer the audio is decoded, the delay is again compensated to overcome network delays and then fed to the *Digital Audio Mixer* as a separate input feedback channel. Based on the spatial metadata, received via the the *Control+Sync Data* channel, the audio feedback is panned by the digital mixer (controlled by the mixer's SDK) and fed to each musician's monitoring line (e.g. monitor loudspeakers or in-ear monitors) to simulate the movement of the user, creating a realistic audience representation also on the musician's side.

B. Architectural Challenges

Keeping latency below acceptable thresholds (e.g. 30 ms) between the two endpoints is one of the key challenges, since this technical condition is essential for reaching real-time interaction and synchronization between performers and audiences. High-latency environments disrupt the flow of bi-directional feedback and diminish the immersive experience. To tackle this, the architecture incorporates low-latency streaming and coding protocols, (e.g. RTP and Opus Codec [27]), which have demonstrated effective real-time transmission capabilities in platforms like JackTrip [28] and LoLa [21]. Additionally, the use of 5G networks and edge computing solutions ensures faster data transmission, minimizing latency for spatial audio and multi-sensory feedback.

Bandwidth and scalability present another challenge, as high-quality spatial audio and volumetric video demand significant bandwidth, especially in large-scale implementations. Object-based audio rendering technologies like MPEG-H 3D Audio [29] and Ambisonics [13] provide scalable solutions, enabling individual sound sources to be dynamically spatialized while maintaining efficient bandwidth usage. For example, Ambisonics can capture 360° soundfields, ensuring that remote listeners perceive audio spatially aligned with visual cues in XR settings. The integration of 5G technology further enhances scalability by offloading processing to edge servers, reducing the computational burden on local devices.

The synchronization of multi-sensory feedback, including audio, visual, and haptic elements, is another technical hurdle. Bi-directional audio feedback systems can further enhance interaction by spatializing audience responses (e.g., clapping and cheering) and transmitting them as real-time cues to performers. Additionally, haptic devices can simulate remote applause or heartbeat feedback, providing a richer sensory connection between participants.

Through the combined use of spatial audio rendering, real-time audio processing, and XR integration, the proposed architecture aims to create an immersive and scalable platform that overcomes latency, bandwidth, and synchronization challenges while fostering bi-directional interaction between performers and audiences. These innovations lay the groundwork for a transformative experience in interactive music performances.

III. Conclusions

The present paper introduce and discuss the key aspects behind the design and implementation of immersive applications (AR/VR/XR) that enable real-time streaming of live music concerts, where users equipped with suitable hardware (i.e. Smart TV with in-built camera and mic, PC, smartphone, XR headset) can interact with musicians in a virtual environment. A first proof-of-concept of the system architecture capable of delivering real-time, spatially accurate audio experiences with acceptable latency, scalable to support multiple participants is presented. The study aims to develop innovative feedback loops for audience engagement, such as spatialized applause and dynamic performer-audience interactions. By addressing these challenges, this research contributes to advancing immersive networked music performances, paving the way for highly interactive and inclusive musical experiences, new content delivery and broadcasting-related services.

The extended abstract introduces and describes the proposed architecture, outlines its components, and shares preliminary implementation insights as well as the main challenges to be addressed. The final paper will present the conducted tests, including an evaluation of the user-perceived QoE, both for the remote users, as for the content producers.

Acknowledgment

The research activities described in this paper were carried out within the "HEAT – Hybrid Extended reAliTy" Project GA 101135637 funded by the EU Horizon Europe Framework Programme (HORIZON).

References

[1] R. Hamilton, "Real-time musical performance across and within extended reality environments," *The Journal of the Acoustical Society of America*, vol. 153, pp. A35–A35, 03 2023.

[2] S. Giacomelli, C. Centofanti, J. Santos, M. Galbiati, T. Salvi, F. Graziosi, and C. Rinaldi, "Remote immersive audio production: State of the art implementation, challenges, and improvements," 09 2024, pp. 1–10.

[3] M. Fernandes, N. Mallmann, and S. Shin, "The rise of augmented reality in live music events: The cases of snapchat and gorillaz," *Business Communication Research and Practice*, vol. 7, no. 1, pp. 58–63, 2024.

[4] L. Comanducci, *Intelligent Networked Music Performance Experiences*. Cham: Springer International Publishing, 2023, pp. 119–130. [Online]. Available: https://doi.org/10.1007/978-3-031-15374-7_10

[5] P. Cairns, H. Daffern, and G. Kearney, "Investigation of server–based spatial audio for metaverse concert distribution," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–8.

[6] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally adapted immersive environments for distributed music performances in mixed reality," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, 2024, pp. 1–10.

[7] G. W. Young, N. O'Dwyer, M. Moynihan, and A. Smolic, "Audience experiences of a volumetric virtual reality music video," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2022, pp. 775–781.

[8] B. Mróz, P. Odya, P. Danowski, and M. Kabaciński, "A commonly-accessible toolchain for live streaming music events with higher-order ambisonic audio and 4k 360 vision," 08 2023.

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

[10] R. Hoy and D. Van Nort, "A technological and methodological ecosystem for dynamic virtual acoustics in telematic performance contexts," 09 2021, pp. 169–174.

[11] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5g-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 4691–4709, 2024.

[12] Wikipedia contributors, "Head-related transfer function," https://en.wikipedia.org/wiki/Head-related_transfer_function, 2024, accessed: 2024-12-11.

[13] ——, "Ambisonics," https://en.wikipedia.org/wiki/Ambisonics, 2024, accessed: 2024-12-11.

[14] Valve Corporation, "Steam audio: Audio solutions for virtual and augmented reality," Online Resource, 2023, available at https://valvesoftware.github.io/steam-audio/.

[15] Google Inc., "Resonance audio: Spatial audio for vr, ar, and 360-degree video," Online Resource, 2018, available at https://developers.google.com/resonance-audio.

[16] P. Cairns, "Viiva-nmp audio system: The design of a low latency and naturally interactive ambisonic audio system for immersive network music performance," Master's thesis, University of York, 2021.

[17] D. Joy, S. Kiran, A. Ponnachan, R. Ashok, and M. Nikhil, "Real-time implementation of spatial audio on fpga using interaural time difference and amplitude-driven perceptual depth," in *2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP)*, 2024, pp. 1–6.

[18] W. Fohl, J. Reichardt, and J. Kuhr, "A system-on-chip platform for hrtf-based realtime spatial audio rendering," in *Second International Conference on Creative Content Technologies (CONTENT)*, 2010, pp. 12–17.

[19] A. Carôt and C. Werner, "Network music performance-problems, approaches and perspectives," in *Proceedings of the "Music in the Global Village"-Conference, Budapest, Hungary*, vol. 162, 2007, pp. 23–10.

[20] "Soundjack: low-latency p2p and server streaming application," Online Resource, 2024, available at https://www.soundjack.eu.

[21] C. Drioli, C. Allocchio, and N. Buso, "Networked performances and natural interaction via lola: Low latency high quality a/v streaming system," in *Information Technologies for Performing Arts, Media Access, and Entertainment*, P. Nesi and R. Santucci, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 240–250.

[22] dürre jan, werner norbert, hämäläinen seppo, lindfors oscar, koistinen janne, saarenmaa miro, and hupke robert, "in-depth latency and reliability analysis of a networked music performance over public 5g infrastructure," *journal of the audio engineering society*, no. 10621, october 2022.

[23] "Musical interactions in networked experiences using real-time virtual audio (minerva)," Online Resource, 2024, available at https://audiolab.york.ac.uk/minerva/.

[24] World Wide Web Consortium (W3C), "Webrtc: Real-time communication in web browsers," https://webrtc.org/, 2023, accessed: 2024-12-11.

[25] S. Porcu, A. Floris, M. Anedda, V. Popescu, M. Fadda, and L. Atzori, "Quality of experience eye gaze analysis on hbbtv smart home notification system," in *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2020, pp. 1–6.

[26] S. Porcu, A. Floris, and L. Atzori, "Towards the prediction of the quality of experience from facial expression and gaze direction," in *2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 2019, pp. 82–87.

[27] J. Spittka, K. Vos, and J.-M. Valin, "RTP Payload Format for the Opus Speech and Audio Codec," RFC 7587, Jun. 2015. [Online]. Available: https://www.rfc-editor.org/info/rfc7587

[28] Chris Chafe and Juan Pablo Cáceres, "Jacktrip: A system for real-time network audio streaming," https://ccrma.stanford.edu/groups/soundwire/software/jacktrip/, 2023, accessed: 2024-12-11.

[29] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "Mpeg-h 3d audio—the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, 2015.