Live Feedback for Immersive Music Performances - A Case Study

Nicusor Amarie
nicusor.amarie@gmail.com
Universitatea Transilvania din Brașov
Brașov, Romania

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

Vlad Popescu
vlad.popescu@unitbv.ro
Universitatea Transilvania din Brașov
Brașov, Romania

Jean-Marius Ghenta
jean.ghenta@orange.com
Orange Romania
Bucharest, Romania

Maurizio Murroni
maurizio.murroni@unica.it
Università degli Studi di Cagliari
Cagliari, Italy

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

Abstract

This paper presents the architecture of an Extended Reality (XR) system for real-time streaming and spatial rendering of live music performances, along with a comprehensive setup for the acquisition of multichannel audio and three-dimensional performer motion data. The system enables remote users to experience live performances in immersive virtual environments, where each musician is represented by an avatar. Audio input is captured via a digital mixing console, enabling the transmission of synchronized multitrack audio streams using low-latency networking protocols. These streams are spatially rendered within a 3D scene using advanced audio spatialization techniques integrated into standard real-time engines such as Unity. The system also supports bidirectional interaction: audience-generated audio feedback (e.g., cheering, clapping, or singing) is captured, spatially processed, and reintroduced into the performance environment, enhancing performer-audience engagement through immersive, real-time crowd response. The architecture was set-up and initially tested in a real concert environment at the Rockstadt Club in Brașov, Romania, with the occasion of a live blues concert in 2024.

Keywords

Extended Reality, Multi-Channel Audio, Virtual Environment

How to cite this paper:
Nicusor Amarie, Gianluca Fadda, Vlad Popescu, Jean-Marius Ghenta, Maurizio Murroni, and Anderson Augusto Simiscuka. 2025. Live Feedback for Immersive Music Performances - A Case Study. In *Proceedings of ACM IMX Workshops, June 3 - 6, 2025*. SBC, Porto Alegre/RS,Brazil, 4 pages.
https://doi.org/10.5753/imxw.2025.8520

1 Introduction

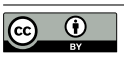
The recent years catalysed a significant and still-growing interest in remote and interactive audio production applications, both among end users and within the audio industry [8, 10, 13]. This trend has also attracted considerable attention from the research community, with several studies investigating the technical, experiential, and artistic implications of such systems [3, 5, 9]. In the last five years there has been a marked increase in experimental events such as Networked Music Performances (NMP) and the live streaming of musical concerts to remote audiences [15, 25]. This growth is largely supported by advances in hardware and software technologies that underpin NMP systems [14, 16], as well as by broader access to high-bandwidth infrastructure [19].

NMP systems have gained substantial traction in both research and industry, especially following the global pandemic, which accelerated the demand for remote, real-time collaboration tools. These systems allow geographically distributed musicians to rehearse and perform synchronously across networks. While the technical feasibility of NMP has improved significantly over the past decade, a core challenge persists: replicating the perceptual quality and interpersonal engagement of co-located performances.

Three main technical domains are central to addressing this challenge: spatial audio rendering, latency minimization, and the enhancement of presence and immersion in virtual environments.

Spatial audio technologies are critical for maintaining the spatial coherence and realism necessary in musical collaboration. Head-Related Transfer Functions (HRTF) [22] and Ambisonics [21] have emerged as dominant paradigms for real-time, three-dimensional audio rendering. Spatial audio engines such as Steam Audio [20] and Resonance Audio [11] enable dynamic scene rendering based on listener orientation and sound source location. Studies such as [2] demonstrate that spatialized sound can significantly enhance the sense of immersion for remote performers. Latency and jitter remain the most significant technical barriers to effective NMP. Research indicates that musical synchronization becomes perceptually unstable at delays above 20–30 ms [4, 16]. To address this, various strategies have been employed, including audio buffer tuning, Quality of Service (QoS) optimization, and edge computing approaches [17]. State-of-the-art platforms like Soundjack [1] and LoLa [6] exemplify systems capable of achieving sub-30 ms round-trip latencies under optimal conditions. The emergence of high-speed network infrastructures, both fixed and mobile, has further enhanced the feasibility of real-time collaboration across wider geographic distances [7, 19].

Last but not least, the question of presence - how performers perceive each other and the performance environment —has also attracted increasing attention. Beyond audio fidelity, systems are beginning to integrate multimodal features such as spatial video, 3D avatars, and haptic feedback to more closely approximate the embodied experience of live, in-person collaboration [12, 18, 23].



In this paper, we present the initial set-up for immersive music performances together with the inherent limitations, the lessons learned, the envisioned architecture and the first performed tests.

The remainder of this paper is organized as follows: Section 2 describes the HEAT project and the XR Blues pilot. Section 3 presents the initial set-up while Section 4 describes system architecture and the initial performed measurements. Section 5 illustrates the performed tests and Section 6 draws the conclusions and presents the future research to be performed.

2 The HEAT Project

Funded under the European Union's Horizon Europe programme, the HEAT project investigates novel paradigms for immersive telepresence by enabling users to be virtually transposed into photorealistic, navigable 3D environments. These environments support both remote and co-located users, allowing for shared experiential immersion independent of physical location. The project integrates a range of advanced technologies—including point cloud and holographic rendering, multi-sensory media delivery, and social XR—within a low-latency, multi-user communication framework designed to support real-time interaction, presence, and co-experience across spatially distributed participants. The XR Blues pilot, one of the four of the projects, focuses on the delivery of immersive live concert experiences, both for remote and locally present users.

3 XR Blues Set-Up

The set-up for the XR Blues pilot is the Rockstadt Club in Brasov - Romania, home of various concerts of rock-related genres. It features a 8 X 6 meter raised stage, withe a metal rig on the front of the stage, hosting lighting and sound equipment.

Figure 1: XR Blues Stage Setup

3.1 Audio-video and 3D capture

For the specific XR Blues Pilot, performed by brazilian guitarist Fred Sunwalk and his band, the persons to be captured were three: guitarist/vocalist, bassist and drummer, resulting in four separate audio channels: each instrument plus the voice channel. The drums, although captured with more microphones, were down-mixed for the pilot's purpose to one audio channel and captured as such. During the soundcheck, the guitarist and the bassist were captured from the front and from the back with two Intel Realsense 515 RGB-D cameras, while the drummer, for stage-related reasons, was captured from the side with the same number and type of cameras. Figure 1 depicts the position of the cameras on the stage (the red circles in the left and the top right frame) and also the pointcloud captured by the camera in front of the singer (bottom right frame). Each two cameras corresponing to one performer were connected via a USB cable to a PC where the RGB and the depth data was recorded. All the six cameras onstage were synchronized by an external trigger which allowed the simultaneous start of the six recordings. The three PCs were connected in an Ethernet LAN via a Gigabyte switch.

3.2 3D environment

On the day after the concert, after the entire gear was unmounted, the club environment was scanned using a drone which took a series of 200 pictures to thoroughly map the entire environment, converted subsequently in a mesh and a texture. This allowed us to recreate the club in the Unity VR environment, together with the avatars of the three musicians, appropriately placed on stage. Each of the player's avatars were mapped to the specific sound channel, live captured during the rehearsals and the concert from the stage mixer. Figure 2 illustrates the environment generated from the shot photos (left) and the final result in form of the Unity environment, populated with the specific avatars.

Figure 2: Stage Reconstruction and 3D Environment

3.3 5G Survey

Before and during soundcheck and concert, an extensive survey of the 5G coverage in and around the club was performed by Orange Romania, partners in the HEAT project, in order to assess the feasibility of a 5G link with the Orange core Network. The first results were encouraging both in terms of up- and downlink speed and latency. It was determined that an external CPE (as part of a Fixed Wireless Access solution) placed on the roof of the club would deliver a much better performance in terms of latency and throughput.

4 System Architecture

The proposed and implemented architecture, depicted in Figure 3, is divided into three main layers: 1) *Live Performance*, 2) *Cloud/Edge Server*, 3) *End User*. On the *Live Performance* side, the band is captures as described in section 3.1 with RBG-Depth cameras. The RGB data, together with the information related to the depth, is fed to the *Position Capture* block, where the data is preprocessed in real-time in order to generate the complete 3D scene, allowing the

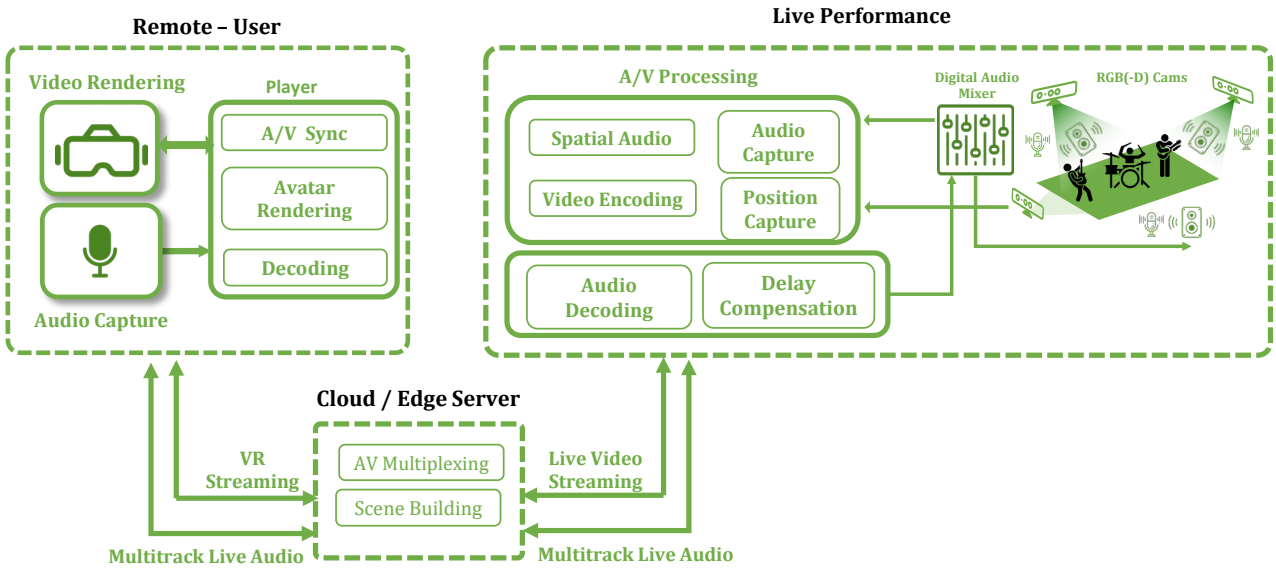


Figure 3: System Architecture

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 from the a *Digital Audio Mixer* (in this case a Behringer X32 Compact mixer) and fed channel-wise to the *Audio Capture* block.

The output of the main *Live Performance* layer consists therefore of the live video data, the depth data and the multitrack live audio 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 and the WebRTC communication protocol [24]).

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,m for example 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 input from the keyboard or mouse is considered. 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.

In the *Live Performance* layer the audio is decoded, the delay is again compensated to overcome network delays and then fed to the *Digital Audio Mixer* as a dedicated input feedback channel. Based on the spatial metadata, 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.

5 Testing and Results

To assess the proposed architecture with respect to connection quality in ultra-low-latency immersive audio scenarios, we implemented a subset of the system, specifically targeting the audio transmission path between the content producer and the end user. A Unity-based server was deployed in the cloud and configured to stream four pre-recorded, lossless WAV audio tracks. On the client side, users were able to explore a 3D environment hosted by the Unity server, where each avatar corresponded to one of the three musicians (bass, guitar and vocals, drums). The audio sources were spatially rendered in real time, dynamically mapped to the listener's position relative to the virtual performers.

WebRTC was employed for audio transmission, and network performance was monitored using Chrome's WebRTC Internals diagnostic tools. Two key metrics—total round-trip time and jitter—were used to evaluate system performance under the defined low-latency constraints. The obtained values are in line with the ones mentioned in the literature, but do not take into consideration also other processing delays which are very much depending on the hardware and software installed at the end points. Parallel measures were done on the server's side using Wireshark, yielding much higher values for the round trip time, for example for the a value of 82 ms vs. the 3.68 ms of the WebRTC internals value.

6 Conclusion

This paper illustrates the first teste performed in the frame of the XR Blues pilot of the HEAT research project. The initial set-up allowed us to assess the inherent limitations related to recording of

live RGB-D data from multiple cameras and the subsequent data processing in terms of avatar creation and. Nevertheless, the first test allowed us to concentrate on the immersive audio part and the development of a suitable architecture for live feedback for immersive music performances. First test in a controlled network environment, demonstrated the initial feasibility of the architecture. Future test will be performed using a 5G network in a controlled end-to-end scenario and will include live streaming of audio and video data.

Acknowledgments

This work was supported by the European Union (EU) Horizon Europe grant 101135637 (HEAT Project). We express our gratitude to Fred Sunwalk and his band for performing and allowing us to test the entire setup during soundcheck and concert. Many thanks also to the Rockstadt club for the usage of the entire environment.

References

[1] 2024. Soundjack: low-latency p2p and server streaming application. Online Resource. Available at https://www.soundjack.eu.

[2] Patrick Cairns. 2021. *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.

[3] Patrick Cairns, Helena Daffern, and Gavin Kearney. 2024. Investigation of Server–Based Spatial Audio for Metaverse Concert Distribution. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2).* 1–8. doi:10.1109/IS262782.2024.10704097

[4] Alexander Carôt and Christian Werner. 2007. Network music performance-problems, approaches and perspectives. In *Proceedings of the "Music in the Global Village"-Conference, Budapest, Hungary*, Vol. 162. 23–10.

[5] Luca Comanducci. 2023. *Intelligent Networked Music Performance Experiences.* Springer International Publishing, Cham, 119–130. doi:10.1007/978-3-031-15374-7_10

[6] Carlo Drioli, Claudio Allocchio, and Nicola Buso. 2013. 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*, Paolo Nesi and Raffaella Santucci (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 240–250.

[7] dürre jan, werner norbert, hämäläinen seppo, lindfors oscar, koistinen janne, saarenmaa miro, and hupke robert. 2022. in-depth latency and reliability analysis of a networked music performance over public 5g infrastructure. *journal of the audio engineering society* 10621 (october 2022).

[8] Melanie Fernandes, Nicole Mallmann, and Sangwoo Shin. 2024. The Rise of Augmented Reality in Live Music Events: The Cases of Snapchat and Gorillaz. *Business Communication Research and Practice* 7, 1 (2024), 58–63. doi:10.22682/bcrp.2024.7.1.58

[9] Andrea F. Genovese, Marta Gospodarek, Zack Nguyen, Robert Pahle, and Agnieszka Roginska. 2024. Locally Adapted Immersive Environments for Distributed Music Performances in Mixed Reality. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2).* 1–10. doi:10.1109/IS262782.2024.10704217

[10] Stefano Giacomelli, Carlo Centofanti, Jose Santos, Mauro Galbiati, Tiziano Salvi, Fabio Graziosi, and Claudia Rinaldi. 2024. Remote Immersive Audio Production: State of the Art Implementation, Challenges, and Improvements. 1–10. doi:10.1109/IS262782.2024.10704192

[11] Google Inc. 2018. Resonance Audio: Spatial Audio for VR, AR, and 360-degree video. Online Resource. Available at https://developers.google.com/resonance-audio.

[12] Michael Gurevich, Adam Fyans, and Paul Stapleton. 2004. JamSpace: Interactive Music Collaboration on the Web. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).* 109–112.

[13] Rob Hamilton. 2023. Real-time musical performance across and within extended reality environments. *The Journal of the Acoustical Society of America* 153 (03 2023), A35–A35. doi:10.1121/10.0018060

[14] Rory Hoy and Doug Van Nort. 2021. A Technological and Methodological Ecosystem for Dynamic Virtual Acoustics in Telematic Performance Contexts. 169–174. doi:10.1145/3478384.3478425

[15] Bartłomiej Mróz, Piotr Odya, Przemyslaw Danowski, and Marek Kabaciński. 2023. A commonly-accessible toolchain for live streaming music events with higher-order ambisonic audio and 4k 360 vision.

[16] Cristina Rottondi, Chris Chafe, Claudio Allocchio, and Augusto Sarti. 2016. An Overview on Networked Music Performance Technologies. *IEEE Access* 4 (2016), 8823–8843. doi:10.1109/ACCESS.2016.2628440

[17] Jens Schuett, Samuel Le Groux, and Pieter-Jan Maes. 2021. Reducing Latency in Remote Musical Interaction: An Edge Computing Perspective. *IEEE Transactions on Multimedia* 23 (2021), 4253–4264. doi:10.1109/TMM.2021.3051963

[18] Barry Truax. 2012. Sound, Listening and Place: The Aesthetic Dilemma. *Organised Sound* 17, 3 (2012), 193–201. doi:10.1017/S135577181200005X

[19] Luca Turchet, Claudia Rinaldi, Carlo Centofanti, Luca Vignati, and Cristina Rottondi. 2024. 5G-Enabled Internet of Musical Things Architectures for Remote Immersive Musical Practices. *IEEE Open Journal of the Communications Society* 5 (2024), 4691–4709. doi:10.1109/OJCOMS.2024.3407708

[20] Valve Corporation. 2023. Steam Audio: Audio Solutions for Virtual and Augmented Reality. Online Resource. Available at https://valvesoftware.github.io/steam-audio/.

[21] Wikipedia contributors. 2024. Ambisonics. https://en.wikipedia.org/wiki/Ambisonics. Accessed: 2024-12-11.

[22] Wikipedia contributors. 2024. Head-Related Transfer Function. https://en.wikipedia.org/wiki/Head-related_transfer_function. Accessed: 2024-12-11.

[23] Elise Wilk, Daan Niermann, and Julia Hentschel. 2023. Embodiment in XR Performances: Towards Situated, Responsive Stages. *Journal of New Music Research* 52, 1 (2023), 45–61. doi:10.1080/09298215.2023.2172846

[24] World Wide Web Consortium (W3C). 2023. WebRTC: Real-Time Communication in Web Browsers. https://webrtc.org/. Accessed: 2024-12-11.

[25] Gareth W. Young, Néill O'Dwyer, Matthew Moynihan, and Aljosa Smolic. 2022. Audience Experiences of a Volumetric Virtual Reality Music Video. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).* 775–781. doi:10.1109/VR51125.2022.00099