A Decoupled VR and Real-time Audio System for Distributed Musical Collaboration

Alberto Boem*, Ovidiu Costin Andrioaia*, Domenico Stefani*, Alessandra Micalizzi†, and Luca Turchet*
*Department of Information Engineering and Computer Science University of Trento, Trento, Italy
Email: alberto.boem@unitn.it, ovidiu.andrioaia@studenti.unitn.it, domenico.stefani@unitn.it, luca.turchet@unitn.it
†SAE Institute Milano, Milano, Italy
Email: a.micalizzi@sae.edu

***Abstract*—Musical collaboration in virtual reality (VR) faces a fundamental technical challenge: achieving the ultra-low latency required for ensemble synchronization while maintaining the rich spatial interactions characteristic of Shared Virtual Environments. This paper presents a proof-of-concept for a hybrid decoupled architecture that addresses these competing requirements of real-time graphics, interactions, and audio through strategic layer separation. Our system combines dedicated real-time audio hardware for networked music performance (Elk LIVE) with consumer head-mounted displays running a custom-made social VR application (made with the Ubiq framework). These two subsystems are connected through lightweight control messaging infrastructure. We evaluated the system through deployment with 12 musicians across three distributed groups, measuring technical performance metrics and user experience through standardized questionnaires. Results show that the audio subsystem maintained consistent latency with minimal packet loss, while VR layer performance varied significantly. Participants achieved moderate levels of social presence and creativity support, with evidence suggesting that audio consistency enables musical focus even when visual performance degrades. Our findings indicate that decoupled architectures might resolve the tension between musical precision and VR immersion requirements, providing design principles for next-generation Musical Metaverse systems that prioritize temporal consistency over absolute performance optimization.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The emergence of collaborative music systems within virtual reality (VR) environments, where musicians collaborate in real-time across networked immersive environments [1] presents a critical system design challenge: achieving ultra-low audio latency requirements for musical synchronization [2], [3] while maintaining the rich spatial interactions and social presence characteristic of Collaborative and Social VR applications [4]–[6]. The technical requirements for these two domains are fundamentally incompatible [7]. Networked music performance (NMP) research has established that ensemble musical coordination requires audio latencies below 30 ms [2], [3], while social VR applications such as VRChat or Horizon World typically operate effectively with latencies of 100-150 ms [8], [9]. Additionally, musical applications prefer uncompressed audio transmission to minimize codec delay, protocols such as the User Datagram Protocol (UDP) to avoid TCP's reliability overhead, whereas VR platforms prioritize bandwidth-efficient voice codecs such as Opus which are optimized for speech intelligibility rather than musical fidelity. Moreover, in NMP applications network jitter compensation requires intelligent buffering strategies to balance latency with audio quality. All together, these requirements create several technical optimization challenges.

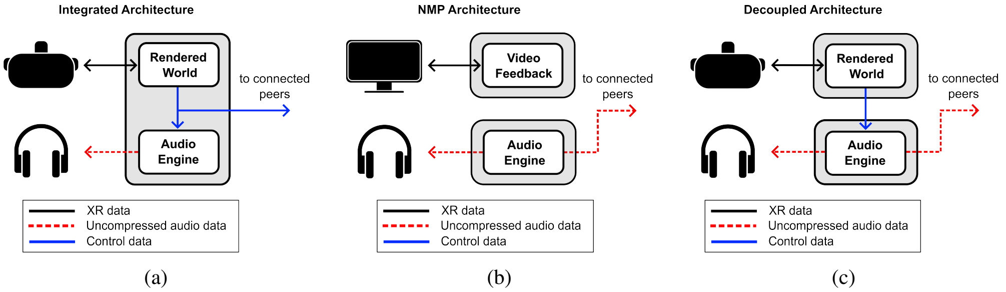
Existing examples of collaborative musical VR systems predominantly employ integrated architectures that combine audio processing, spatial rendering, social interaction, and synchronization within a single VR application framework (see Fig. 1(a)). This is typical of applications such as PatchWorld[1] or CSound Meta [10], as well as several prototypes (e.g., [11]–[13]) that tried to combine social VR with NMP systems. These systems do not exchange audio streams, but control messages that trigger sounds stored locally at each peer node. While ensuring tight audio-visual coupling, this approach forces audio processing to contend with graphics rendering and networking for computational resources, frequently resulting in compromised audio quality or increased latency jitter. Moreover, this makes the integration of electric and acoustic audio sources (singing voice, guitars, brass) unfeasible, since the audio processing (especially synthesis) happens locally. On the other hand, traditional NMP systems like JackTrip [14] and LOLA [15] achieve optimal audio performance through the streaming of uncompressed audio signals, but lack the embodied interaction and spatial presence essential for immersive collaborative experiences [16] (see Fig. 1(b)).

To overcome the limitations of the two approaches, this paper presents a proof-of-concept for a decoupled architecture that addresses these competing requirements through strategic layer separation while preserving perceptual coherence (see Fig. 1(c)). We developed a prototype using Ubiq, a Unity networking library for the VR subsystem [17], and Elk LIVE[2], a hardware–software solution for streaming low-latency, high-

We acknowledge the support from the MUR PNRR PRIN 2022 grant, prot. no. 2022CZWWKP, funded by Next Generation EU.We also acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379) and by the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

[1] PatchWorld: https://patchxr.com/ (accessed: 2025/07/12)

[2] Elk LIVE: https://elk.live/ (accessed: 2025/07/12)



(a) (b) (c)

Fig. 1. (a) An integrated architecture for music making in collaborative VR, where all the processes (Social VR and audio) take place in the same application running in an HMD; (b) A classical NMP architecture, where audio data are streamed between peers but the visual feedback (if present) is a video feed that take place in another application layer; (c) The decoupled architecture proposed in this paper where the two processes (Social VR and audio) take place in two separate subsystems but they are synchronized through a messaging and control layer.

quality audio, for the NMP subsystem. While previous work has demonstrated the viability of decoupled approaches using JackTrip in connection with a simplified immersive framework [18], our work extends this concept by offloading audio synthesis and music networking to a dedicated hardware and software subsystem, while the VR subsystem focuses on visual rendering and social interaction. A lightweight control messaging infrastructure ensures synchronized interaction between the two layers, such as between musical gestures in the VR and audio generation, maintaining the tight feedback loops required for expressive musical performance.

The primary contributions of this work are: 1) a hybrid architecture that proposes to resolve the tension between musical precision and VR immersion requirements through strategic subsystem decoupling; 2) a working implementation integrating consumer VR hardware with professional audio processing systems; 3) empirical validation through deployment with 12 musicians across three distributed groups; 4) performance analysis demonstrating system resilience; and 5) design principles for multi-modal optimization in VR music systems.

II. System Architecture

Our hybrid architecture addresses the competing demands of musical precision and immersive presence through three fundamental design principles: layer separation with synchronized control, independent performance optimization, and modular integration. Audio synthesis and VR rendering are implemented as independent subsystems based on their respective performance requirements. Audio processing occurs on dedicated real-time hardware, while VR rendering utilizes the full computational capacity of headset graphics processors. These layers maintain synchronization through a low-latency control messaging protocol that preserves tight coupling between musical gestures and audio response. Then, each subsystem operates according to domain-specific performance constraints. While the audio subsystem prioritizes quality and stability, the VR subsystem focuses on consistent frame rates, spatial interaction fidelity, and social presence indicators. This independence should enable sustained musical performance even under VR layer performance variations.

A. Social VR Subsystem

The first subsystem is dedicated to render the immersive VR environment and support embodied, spatial, and social interactions. It was developed using Ubiq a free and open-source networking library for Unity designed to facilitate the construction of social VR systems for research, teaching, and experimentation[3]. Ubiq operates through a component-centric messaging architecture where discrete messages are exchanged directly between Unity Components across networked peers, allowing for flexible network topologies (e.g., peer-to-peer, client-server). Ubiq provides support to both Unity XR Interaction Toolkit and simulated VR controls to be used in the Unity Editor. This simplifies the deployment process and streamlines development and testing. The system provides core social VR functionality including basic avatar management, real-time positional voice communication via WebRTC, object spawning, event logging, and room-based rendezvous services. While Ubiq cannot be compared in terms of scalability to industry-standard tools such as Photon PUN, it can be extended with custom components.

B. NMP and Audio Subsystem

The second subsystem is dedicated to processing and streaming audio data through the network. This was developed using an off-the-shelf NMP solution such as Elk LIVE. The Elk LIVE ecosystem represents a comprehensive solution for NMP, operating as a peer-to-peer audio streaming system that transforms analog audio signals into IP packets for network transport and vice versa. Elk LIVE consists of the Elk LIVE Bridge hardware device and the Elk LIVE Studio web application. The hardware device and Studio webapp are used in conjunction, allowing users to establish an audio and video connection with remote peers with the same Elk LIVE setup.

The Bridge is a dedicated audio and networked interface that handles audio I/O and network transmission. The interface offers several physical audio connectors (e.g., XLR, 6.35 and 3.5mm jacks) along with an Ethernet port for fast wired internet connection. The device operates as a standalone server, functioning as a sender/receiver broadcasting audio

[3] Ubiq: https://github.com/UCL-VR/ubiq (accessed: 2025/07/12)

directly to other Bridges in real-time through the network connection. Each device captures audio input, mixes it with signals from remote peers, and delivers the combined audio to the line and headphone outputs. The Bridge processes 24-bit, 48kHz uncompressed audio, since compression would introduce additional latency. The Elk LIVE Bridge is based on the Elk Audio OS[4] [19], a real-time Linux-based operating system optimized for audio processing. For sound and data processing, Elk Audio OS includes Sushi[5]: an headless (i.e., with no graphical user interface) Digital Audio Workstation (DAW). Sushi can host VST2, VST3, and LV2 plugins. The version present in the Elk LIVE Bridge boards comes with many open-source plugins[6]. Like most of conventional DAWs, Sushi supports the most common protocols for controlling and syncing with external sources and devices such as MIDI, Open Sound Control (OSC), gRPC, and Ableton Link.

Control and management occur through Elk LIVE Studio[7], a browser-based application similar to those used for video-meetings such as ZOOM or Google Meet. Studio allows users to find remote peers, start NMP sessions, establish a video connection as well as the low latency audio communication, mix audio streams and route audio signals. Its Mixer section provides control over input sources and allows participant to create their own headphone mix without affecting others' monitoring. For details and benchmarking of the Elk LIVE system see [20], [21].

C. Communication Infrastructure

The communication infrastructure coordinates the VR and audio subsystems through a lightweight messaging framework that maintains the temporal precision required for interactive musical performance. The system employs OSC protocol for transmitting musical control data from VR gestures to the Elk LIVE Bridge where the audio synthesis happens, achieving sub-10 ms latency on local networks essential for time-critical musical systems [22], [23].

The OSC implementation uses a standardized address space `/keyboard_event/[track_name]`, with parameters for `note_on`, `note_off`, `aftertouch`, and `note_index`. Control messages are transmitted via UDP with a dedicated port per peer, achieving typical end-to-end control latency of 2-3 ms on the local network. By doing so, the architecture separates between processes dedicated to audio and ones to control messaging into independent network channels, preventing high-bandwidth audio data from interfering with time-sensitive gesture commands.

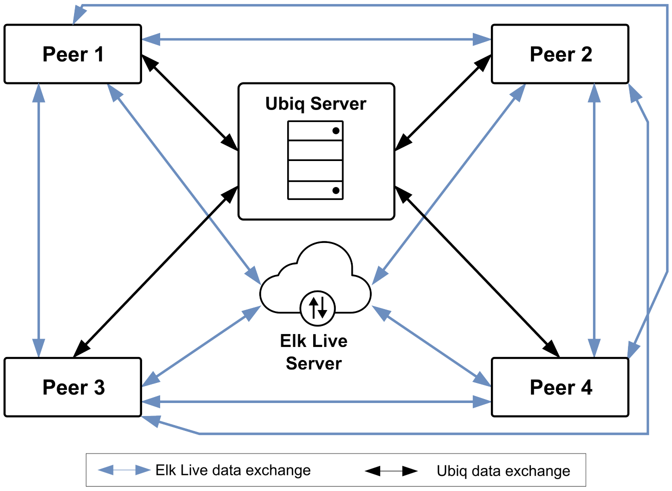


Fig. 2. Networked data exchange between remote peers.

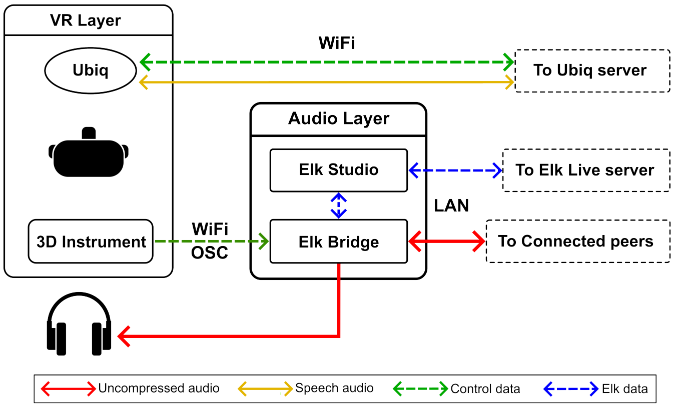


Fig. 3. Detailed view of the system architecture for a single peer.

D. Network Architecture

The network architecture implements a dual-channel strategy that optimizes communication pathways for the distinct requirements of audio streaming and control messaging in distributed musical performance (see Figs. 2 and 3). Audio streaming utilizes Elk LIVE's peer-to-peer infrastructure, which has an audio packet transmission protocols specifically tuned for musical latency requirements. To achieve minimal latency, Elk LIVE uses the UDP without audio compression or retransmission. The system employs several solutions to guarantee continuous and quality audio output, including a jitter buffer and packet loss concealment algorithms. The UDP protocol choice reflects the real-time nature of musical performance, where occasional packet loss is preferable to the delays introduced by TCP retransmission mechanisms. Delayed packets are in fact not useful for live musical interaction.

Control messaging operates through Ubiq's WebSocket and WebRTC infrastructure, which handles VR state synchronization and voice communication alongside musical control data. While voice chat leverages Ubiq's established WebRTC implementation, OSC messages for gesture-to-audio communication maintain dedicated low-latency pathways to preserve time-critical musical interactions. This separation allows each communication type to utilize protocols optimized for its

[4] Elk Audio OS: https://elk-audio.github.io/elk-docs/html/ (accessed: 2025/07/12)

[5] Elk Audio Sushi: https://github.com/elk-audio/sushi (accessed: 2025/07/12)

[6] Elk Audio MDA VST2: https://github.com/elk-audio/mda-vst2

[7] Elk LIVE Studio: https://studio.elk.live/ (accessed: 2025/07/12)

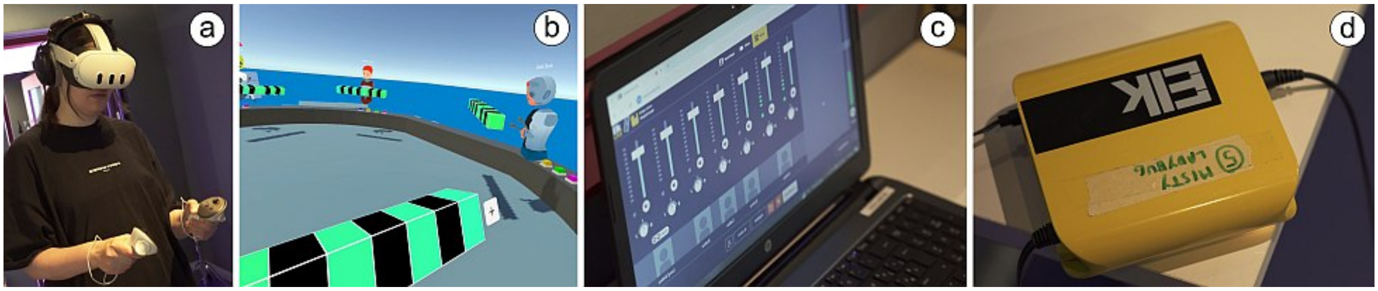


Fig. 4. The main components of the proof-of-concept system: a) a standalone HMD, b) a collaborative virtual world named RythmUS, c) the Elk LIVE system, d) the Elk LIVE Bridge.

specific characteristics. Session coordination is managed by a Node.js server that handles user authentication, session initialization, and NTP-based clock synchronization services. The server additionally provides fallback communication pathways between subsystems, ensuring system resilience when primary communication channels experience disruption. This architecture was validated through real-world deployment scenarios that confirmed its effectiveness in maintaining musical performance quality across distributed environments.

E. Implementation

The main components of the proof-of-concept system are shown in Fig. 4. For the VR subsystem, we developed a VR application named RythmUS[8]. The application was developed using Unity LTS 2022.20.2f1 and tested on Meta Quest 2 and 3 HMDs. The virtual world was designed as an island surrounded by the sea, with a tower, which top part is the performance stage for the users (Fig. 5(a)). Upon entering the virtual world, a user is placed on a bridge. A Ubiq menu allows the user to either create or join an already active room. Inside a room, the users can see each others as 3D embodied avatars. The avatars are randomly assigned. Users can also speak with each others through the microphone available in the HMD.

To reach the stage, users has to climb the tower. There, users can position themselves in the space and through a dedicated menu users can spawn a 3D virtual musical interface (Fig. 5(b)). This is composed of control surface and two mallets (Fig. 5(c)). The control surface is modeled on one octave of a piano, starting from `C3`. Through two arrows placed at the sides, the user can change the octave of the scale. With the mallets, users can interact with the blocks. When the mallet enters, a `note_on` message is sent through OSC to the Elk LIVE Bridge, when the mallet exits, a `note_off` is produced. If the mallet stays inside a block the note selected keeps playing and, depending on the amount of movement, the vibrato can be applied. The audio is synthesized in the Elk LIVE Bridge, through a synthesizer plugin. For our test we selected the MDA JX-10, a lightweight subtractive synthesizer plugin inspired by 1980s Roland synthesizers[9]. The VST synthesizer is loaded in Sushi on the Elk LIVE Bridge, through the dedicated python interface, before the beginning of each session. While the MDA JX-10 includes 52 preset patches, users can select (through a menu placed under the blocks) 4 of them, named Echo Pad, Monosynth, Solid Backing, and Synth Brass.

III. User Study

We conducted a user study out at SAE Institute in Milan, involving 12 participants divided into three groups. The main aim was to evaluate the system's capabilities and gather information on the quality, usability, and sense of social presence provided by the system.

A. Participants

Twelve adult participants (4 identified as women, 8 as man, age: 23 ± 3.25) volunteered to took part in the study. Participants were divided in three groups, each group composed of 3 men and 1 woman. They were all proficient musicians (years: 11.75 ± 4.43), versed in different musical instruments such as guitar, keyboard, and in a variety of tools for digital music production. None of them had previous exposure with tools for music making in VR. Only one reported to have tried once a VR game. Similarly, none of them used regularly NMP systems, and had experienced only a few times ZOOM in this context (see Fig. 6). They were recruited through post and invitations through the network of SAE Institute Milano. All participants received a 50 Euros Amazon Gift Card as compensation. Before the test, they provided informed consent.

B. Methods

For each test, each member of the group (Peer 1-4) was placed in a different room. The rooms were located in the same building. The rooms of Peer 1, 2 and 4 were located at the first floor at a distance of 5 meters between Peer 4 and Peer 1 and 2, were Peer 1 and 2 distanced around 2 meters. Peer 3 was placed one floor below. Each room was equipped with an Elk LIVE Bridge board and a laptop (used

[8] Github repository of RythmUS https://github.com/CIMIL/MusicalMetaverseElkLiveAudio (accessed: 2025/07/12)

[9] Github repository with the tools to reproduce the audio setup with Elk Live and the Bridge. https://github.com/CIMIL/MusicalMetaverse_ElkLive_Bridge (accessed: 2025/07/12)

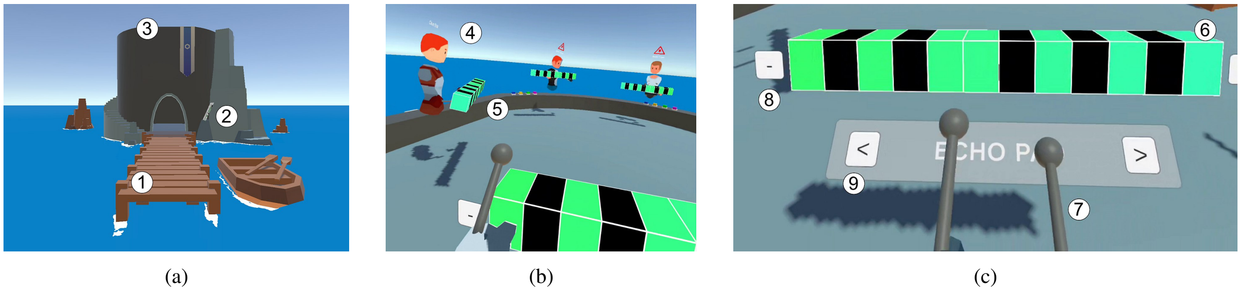


Fig. 5. (a) The RythmUS virtual environment, with the 1) entrance bridge, 2) Ubiq Room Menu, 3) performer's stage; (b) the top of the tower with the different peers represented as avatars (4) playing together their virtual instruments (5); (c) the 3D virtual instrument composed by the 6) block interface; 7) the virtual sticks; 8) buttons to change the octave; 9) menu for selecting the VST presets.

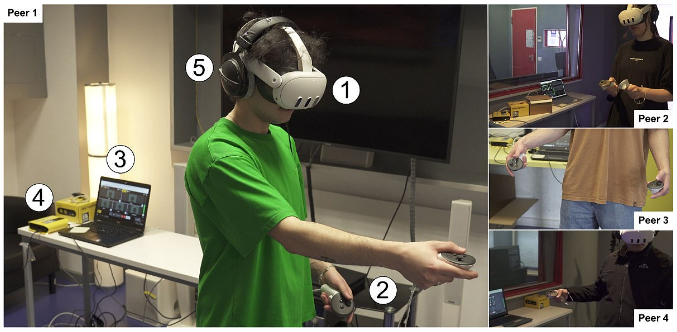


Fig. 6. The equipment used by participants: 1) Meta Quest 3, 2) Tracked controllers; 3) Elk Studio Live; 4) Elk LIVE Bridge; 5) Headphones.

for connecting with Elk LIVE) connected to the LAN. While the Elk LIVE Bridges were connected to the network through LAN cables, the HMDs were connected on the same network using WiFi. For listening to the audio stream, each participant wore a pair of Beyerdynamic DT 770 headphones connected to their Elk LIVE Bridge. For VR, each participant was provided with Meta Quest 3 with hand-held controllers for direct manipulation and interaction. While Elk LIVE connected to the cloud-based server of Elk, the VR applications were connected using the Ubiq local server, which run on the laptop of an experimenter.

C. Procedure

The procedure was the following for each group. First, at the beginning they were instructed about the nature and the objective of the task. Then the experimenter introduced the participants to the system and its basic functionalities. Second, participants were accompanied to a different room by one of the experimenters. There, each participant was assisted by an experimenter for setting up the system, as well as wearing the HMD and headphones. Third, when the test started it was divided into two sessions. In the first session, an experimenter connected from a laptop (using the same setup as the participants), instructed the participants inside the virtual world. Participants were guided to the main functionalities, how to create the virtual instruments, and how to use them. After 15 minutes of free exploration, participants were asked to go through a 30-40 minutes of guided improvisation. After that, the experimenters helped participants in removing the equipment and accompanied them to the initial room, where they were asked to fill in three questionnaires and to report feedback about their experience. The entire session took approximately 90-120 minutes.

D. System Performance Metrics

From a the point of view of our system, we were interested in assessing its characteristics especially in terms of the impact of the network in which the test was made. Using the built-in logging tools of Ubiq we collected a measure of roundtrip latency (Ubiq-RL) between peers and frame time (Ubiq-FT) of each peer. The first measures roundtrip latency between the different peers, in milliseconds. The second represents the amount of time in seconds that has elapsed since the last frame was captured at remote peers during latency probes. This measure reflects the overall application performance and computational load. With the logging tools of Elk LIVE we collected the roundtrip latency between peers and the packet error ratio, which represents the percentage of packets that failed to arrive at destination during network transmission. Additionally, we derived for the four performance metrics the coefficient of variation (CV), which revealed distinct performance profiles across experimental groups. Also for CV, we calculated the mean and standard deviation of each of these measures for each peer in each group, and across all peers in a group.

E. User Experience Assessment

From a subjective level, we were interested in investigating three specific aspects: 1) since VR is a new medium for musicians we wanted to understand the impact of VR as a technology in a collaborative and social musical task; 2) investigate the sense of co-presence and at which level a VR musical application with more than 2 users can efficiently support it; 3) how the VR instrument we developed were effective in supporting musical creativity.

For the first point, we developed ad-hoc questionnaire with four items to be assessed on a 7-point Likert scale. For the second point, we used a 5-point Likert scale modified version of the Networked Minds Social Presence Inventory

TABLE I
UBIQ - MEAN LATENCY BETWEEN PEERS.

Group	Latency (ms)	CV
1	76.63 ± 116.07	1.515
2	46.54 ± 45.41	0.976
3	43.35 ± 132.75	2.927

(NMSPI) [24]. Being an exploratory work, which involved only three groups of players, we removed the part of the questionnaire related to “Perception of the other”, and instead focused solely on “Perception of self”. Then, we removed the items related to “Perceived emotional contagion” since this was not part of the main focus of the study, and we deemed it both difficult to apply with the type of task employed, and not directly applicable in a musical context. For the third point, we used the Creativity Support Index (CSI) [25] (7-point Likert scale), a tool used for evaluating how a digital tool can support creativity. Additionally, a series of interviews were conducted with participants to better explore their experience.

IV. RESULTS

For the Ubiq system we found a variation in the roundtrip latency measured for each group. See Table I for a summary of the results, and Fig. 7 for the Cumulative Distribution Function (CDF). In terms of mean latency, Group 3 and 2 showed the lowest (M = 45.35 ms; M = 46.54 ms), while Group 1 showed the largest (M = 76.63). However, by looking at CV we can notice that while Group 2 showed the lowest (0.976), Group 3 showed the highest (2.97), and Group 1 a value in between (1.515). This shows that in terms of latency, Group 2 was the most stable of the three. However, variability can be an issue. Regarding frame time (see Table II for a summary of the results, and Fig. 8 for the CDF), we can see that different groups had different experience. By considering that the target frame-rate for VR applications on the Meta Quest 3 is 72 Frames per Seconds (FPS), we can see that the closest was Group 3 with approximately 72.46 FPS, followed by Group 1 with 68.85 FPS and then by Group 3 with a general performance that lowered to around 46.88 FPS. However, this was probably caused by an anomalous behavior of the HMD of Peer 1. The CV of all three groups is somehow comparable, showing that the variation was generally similar.

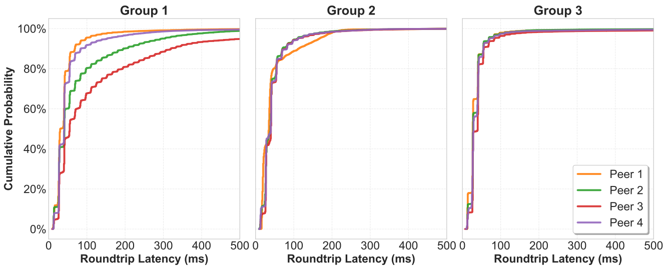


Fig. 7. Cumulative Distribution Function for the Ubiq roundtrip latency measured for each peer.

TABLE II
UBIQ - MEAN FRAME TIME BETWEEN PEERS.

Group	Frame Time (ms)	CV	Frame per second (FPS)
1	14.58 ± 19.87	1.363	68.58
2	21.33 ± 35.16	1.648	46.88
3	13.80 ± 14.61	1.058	72.46

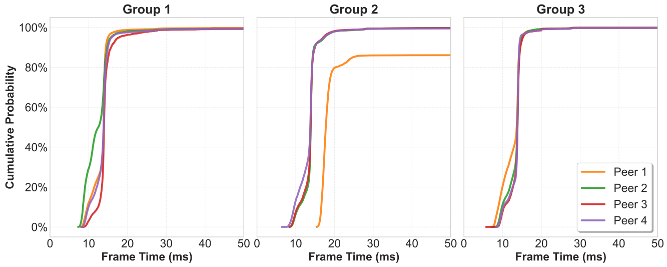


Fig. 8. Cumulative Distribution Function for the Ubiq Frame time measured for each peer.

For the latency of the Elk LIVE system (see Table III for a summary of the results, and Fig. 9 for the CDF) we can see that the results are similar between groups, in the range of 6.68-6.82 ms, with a CV of between 0.141 of Group 2 and 0.105 of Group 3. Regarding the packet error ratio (see Table IV for a summary of the results, and Fig. 10 for the CDF, we also found general similarity between the three groups, and a stable CV between 1.22 and 1.51. Taking together these results confirm the stability and low latency of Elk LIVE, at least in a local network.

Despite identical hardware and network infrastructure, groups exhibited different performance characteristics due to varying computational loads from simultaneous processes, different HMD initialization states, and potential interference from other network traffic during testing sessions. Group 2’s anomalously low frame rate was primarily caused by performance issues on Peer 1’s device, highlighting the sensitivity of VR applications to system state variations even in controlled environments.

TABLE III
ELK LIVE - MEAN LATENCY BETWEEN PEERS.

Group	Latency (ms)	CV
1	6.82 ± 0.30	0.043
2	6.78 ± 0.96	0.141
3	6.80 ± 0.72	0.105

TABLE IV
ELK LIVE - MEAN PACKET ERROR RATIO BETWEEN PEERS.

Group	Packet Error Ratio (%)	CV
1	0.53 ± 0.08	1.51
2	0.58 ± 0.06	1.3
3	0.57 ± 0.07	1.22

Regarding the user experience (see Fig. 11), device awareness scores ranged from 2.75 to 4.75 across the three groups, with Group 1 showing the low device intrusiveness (M =

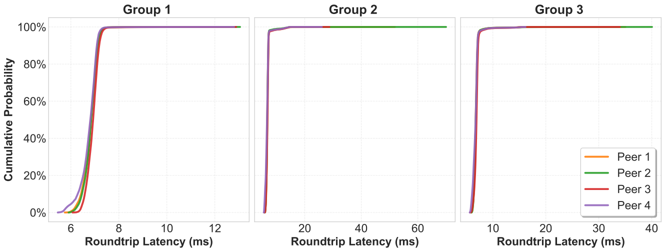


Fig. 9. Cumulative Distribution Function for Elk LIVE roundtrip latency measured for each peer.

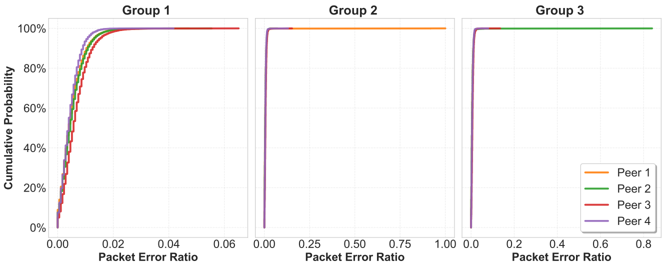


Fig. 10. Cumulative Distribution Function for Elk LIVE Packet Error Ratio measured for each peer.

2.75) and Group 3 the highest of the three (M = 4.75). Focus on the task scores varied among groups, with Group 2 that reported to be more focused on musical execution rather than interface mechanisms (M = 5.25), followed by Group 2 (M = 3.75) and Group 1 (M = 2.50). The visual interference remained relatively consistent across groups. Differently, audio interference showed small variations, with Group 2 reporting low interference (M = 2.75), Group 3 showing a moderate level (M = 3.75), and Group 1 reporting the highest audio interference of the three (M = 4.75).

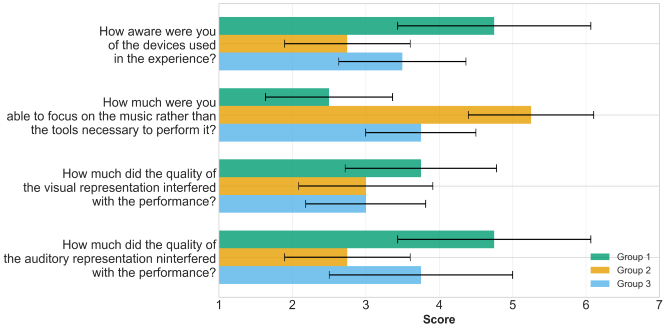


Fig. 11. Mean and standard deviation of the results for the ad-hoc user experience questionnaire for the three groups.

In terms of social presence (see Fig. 12), scores were relatively uniform. Co-presence scores were relatively uniform across groups, ranging from 2.63 to 3.00 (M = 2.84, SD = 0.19). Perceived attentional engagement varied, with Group 2 showing the highest engagement (M = 3.33, SD = 0.360) and Group 1 the lowest (M = 2.58, SD = 0.210) of the three. Perceived comprehension scores ranged from 3.25 for Group 2 to 4.17 for Group 1. Perceived behavioral interdependence remained consistent across all groups, with scores clustering around 3.0.

Collectively, these moderate scores suggest participants experienced adequate social presence for collaborative work, with consistent within-group experiences (SD < 0.5) but noticeable between-group differences in terms of engagement and comprehension.

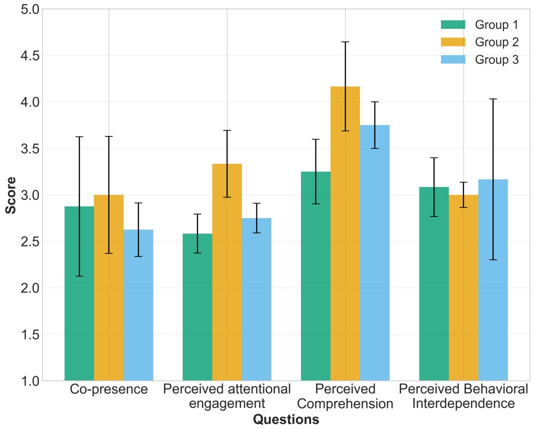


Fig. 12. Mean and standard deviation of the results for the modified version of the Networked Minds Social Presence Inventory for the three groups.

The CSI scores –in a scale from 0% to 100% (see Table V)– show differences between groups. Nevertheless, the overall mean CSI across all groups was 52.22% (SD = 4.54), indicating a global moderate levels of creative support provided by the VR instrument used by participants. The CSI index from Group 1 (48.1%) is below midpoint, while Group 3 (51.7%) and Group 2 (56.9%) are just above midpoint. However, similar standard deviations (M = 2.0-2.3) indicates consistent experiences within each group.

A. *Post-task Interviews*

After the user study, all participants were invited to take part in an in-depth interview conducted by one of the authors. Out of 12, 10 participants agreed to participate. The interviews were conducted online, one week after the conclusion of the user study. The aim was to explore participants' lasting emotional responses and their evaluations of the experience. While, the final part of the interview focused on participants' perspectives on the future of the Musical Metaverse and its impact in their music practices.

The interviews revealed both the appreciation of several technical aspects of the system as well as areas requiring improvement. The majority of comments focused on technical performance, user interface effectiveness, and system reliability during collaborative musical sessions.

At first, participants positively evaluated the spatial audio implementation, noting that "*audio comes from your interface to others... it's much more practical and real.*" According to participants the tracking system of the HMD successfully captured their musical intentions, with one participant observing

TABLE V
THE CREATIVITY SUPPORT INDEX FOR THE THREE GROUPS.

Group	CSI score (%)
1	48.1 ± 2.3
2	56.9 ± 2.3
3	51.7 ± 2.0

that "*if I played softly, the music went softly: it captured the intention.*" When functioning properly, the system achieved fluid real-time interaction, described as "*everything was so fluid and instantaneous. All the moments when it worked perfectly.*"

However, significant technical challenges emerged. First, audio latency from the speech channel consistently impacted musical synchronization and hindered collaborative timing. Being in a social and collaborative immersive environment presented a learning curve for participants which were mostly first time users of VR. Participants especially noted the they felt "*chaotic in my movements because we weren't familiar with the interface... there were some things that didn't work.*" Especially participants from Group 1 reported system stability issues, including frequent crashes and connection problems, disrupted musical flow and were characterized as making the application "*clunky.*"

An important positive aspect appreciated by participants was the design of the virtual environment, which contributed to their sense of immersion, though participants described the graphics as "*Minecraft-like*". However, according to the majority of participants the chosen avatar representation created identity disconnection, affecting participant recognition and authentic presence within the collaborative space.

Despite experiencing significant technical limitations during the sessions, participants acknowledged the system's potential for remote music-making. They specifically identified rehearsal and collaborative production as promising applications, noting that existing web conferencing platforms fail to support adequately musical activities. Notably, post-session interviews revealed more positive perspectives than the feedback recorded after the study, indicating that reflection time allowed participants to better appreciate the system's collaborative possibilities beyond immediate technical frustrations.

In conclusion, participants identified key technical priorities that according to them should be addressed such as latency reduction of speech audio, enhanced interface design for intuitive musical expression, and improved system stability for extended collaborative sessions.

V. DISCUSSION

The results of the technical performance showed that our proof-of-concept developed for creating a social VR music-making experience is feasible but reveals important considerations for system design and deployment.

Our findings suggest that using decoupled systems (Social VR and NMP) may enable selective resilience in collaborative musical interaction, offering preliminary evidence for an alternative architectural approach for VR music systems. Group 2 achieved the highest task focus on musical execution (M = 5.25) while experiencing severe VR layer degradation (46.88 FPS, and 23.27 ms latency), yet maintained an almost identical audio layer performance (6.78 ms peer-to-peer latency) to all other groups. This dissociation demonstrates that musicians can maintain collaborative effectiveness when core musical communication pathways remain intact, even as secondary interaction modalities degrade. Additionally, the two-layer architecture approach reveals that musical communication and VR interaction operate with different performance requirements. While the VR subsystems showed large variation – ranging from 43.35 ms to 76.63 ms of roundtrip latency with coefficients of variation up to 2.927– the dedicated audio subsystem remained consistent across all conditions. This decoupling appears to help musicians to prioritize auditory information when visual-spatial cues become unreliable, suggesting potential modality switching strategies in degraded VR environments. While, interesting, this aspect warrants further investigation.

Group 1 demonstrated a distinctive adaptation mechanism: despite experiencing the worst performance of the VR subsystem (76.63 ms roundtrip latency, CV = 1.515) and low task focus (2.50), they maintained high perceived comprehension (M = 4.17). At the same time, their audio performance (6.82 ms peer-to-peer latency, CV = 0.043) remained consistent with the other groups. This dissociation might indicate that even in VR contexts musical comprehension operates primarily through audio channels, with the VR subsystem affecting secondary spatial-visual interactions. As shown by previous research (in non-VR settings) musicians can collectively adjust temporal expectations to accommodate system constraints, maintaining comprehension through coordinated adaptation [26], [27]. Additionally, these results align with the ones of Cairns et al. [18] such as musicians could maintain focus when audio remains stable despite visual issues. Taken together, this suggests that VR music system design should prioritize audio modalities with significant implications for resource allocation and system architecture. This adaptation mechanism aligns also with the results by Van Kerrebroeck et al. [28] that temporal rather than visual synchronization drives musical *liveness*, but extends this insight to reveal that temporal predictability may be more critical than temporal optimization. Even if preliminary, the design implications of these are profound rather than pursuing minimal latency, systems should prioritize consistency. However, the understanding of these strategies in collaborative VR contexts with larger ensembles require further investigation.

Beyond such considerations, our results reveal a critical relationship between network stability and voice communication quality that extends beyond musical audio performance. Musicians experienced significant voice communication interference due to speech being integrated within the VR subsystem rather than the dedicated audio system. The relatively high stability of the VR subsystem experienced by Group 2 (CV = 0.976) corresponded with lowest voice interference (M = 2.75), while Group 1 network variability was accompanied with a high perceived voice interference (M = 4.75). Critically, this occurred while musical audio remained consistently high-quality across all groups through the dedicated NMP system.

This dissociation shows that voice communication and musical audio serve distinct collaborative functions that cannot be substituted for each other. Even when musical audio transmission remained optimal, degraded voice communication significantly impacted the collaborative experience, suggesting that musicians rely on voice channels for coordination, planning, and social connection while using the audio layer for musical interaction. The sensitivity to voice interference despite stable musical audio reveals that VR music collaboration requires dual-channel reliability. It is possible that musicians monitor voice communication quality as an indicator of overall system trustworthiness.

Although music is inherently non-verbal, speech was used by participants for negotiating musical decisions when non-verbal communication failed and for problem-solving during rehearsals. Voice degradation was particularly important since the experimenter guided sessions through verbal instructions, acting as a sort of conductor. Previous research found that verbal communication in music was more effective than gestures when used by conductors during rehearsals [29]. Yarbrough [30] showed that singers responded better to verbal instructions about articulation than gestural cues. Unlike traditional NMP settings, which focus primarily on musical data and audio transmission, social and collaborative VR environments introduce this additional communication channels critical for effective collaboration. This aspect warrants further research, since voice communication integration in musical XR environments has not been deeply studied before. Our findings suggest that future VR music collaboration systems must prioritize the reliability of both musical and communicative channels to support more effective distributed musical collaboration.

Taken together, our findings provide several evidences for hybrid architectural approaches that optimize multiple communication channels independently. While Cairns et al. [18] have already showed the feasibility of decoupled architectures using dedicated NMP systems with simplified VR interfaces, our findings extend this approach by demonstrating that also a dedicated hardware can provide even greater consistency in the audio layer while supporting more complex VR interactions. We can then suggest that a successful VR music collaboration systems should require at least three distinct networking optimizations: musical audio, voice, and visual-spatial interaction including gestures and body sway which have been extremely important for synchronization [31], [32].

A. Limitations and Future Work

The small sample size (n = 12) and the controlled local network environment limit the generalizability of our findings to real-world distributed collaborations, particularly those operating under wide area network conditions that characterize most distributed musical collaboration systems.

In addition, the lack of prior VR experience among participants may have influenced adaptation patterns and user experience ratings. Future studies should investigate how extended exposure and familiarity affect both technical performance tolerance and creative engagement.

Further investigation is needed into how individual hardware variations affect group performance, especially in the context of developing systems capable of maintaining synchronization across heterogeneous user equipment.

Additionally, the moderate creative support scores suggest significant opportunity for instrument design improvements that could enhance the collaborative music-making experience.

Finally, although we demonstrated a proof-of-concept version of the decoupled architecture, it depended on a custom-built integration of diverse technologies. Future works might focus on fully integrated solutions that might reduce the complexity and be more efficient in integrating and accommodating different types of technologies, as started to be explored by the Internet of Sounds community [33].

VI. Conclusion

Musical Metaverse applications fundamentally represent ultimately immersive social spaces where all aspects of musical creativity and collaboration must be supported –not only in performance execution but also in group ideation, brainstorming, and peer support. These environments require optimization across multiple dimensions, where both latency and audio quality must be carefully balanced to preserve the essential characteristics of musical interaction.

Our results show that dedicated hardware for audio processing can help in maintaining stability compared to general-purpose solutions and current architectures used in systems such as PatchWorld or CSoundMeta, as it operates without competing system demands. This finding suggests that future architectures should also relocate voice communication channels to the dedicated NMP subsystem, instead of leaving it to the VR subsystem, where it is usually located. Through this it will be possible to leverage the specialized characteristics of integrated solutions such as Elk Live and Elk LIVE Bridge, while leaving visual and gestural processing to the HMD. However, this leaves open a problem related to the synchronization between the two layers.

Furthermore, our observations show that musicians can effectively navigate multi-layer system performance when they understand which communication channels remain reliable. This principle directly informs VR music interface design, where distinct status indicators for different system layers could enhance collaborative effectiveness.

While these findings represents a challenge to the assumption that technical excellence automatically translates to musical excellence in VR environments. Instead, musical collaboration thrives on predictability, consistency, and selective optimization—principles that should guide next-generation Musical Metaverse system development.

Acknowledgment

We acknowledge the support from the MUR PNRR PRIN 2022 grant, prot. no. 2022CZWWKP, funded by Next Generation EU.

We also acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379) and by the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

We thank Francesco Manenti for the help provided during the user study, the technical staff of SAE Institute Milano, and Pietro Gola for the help provided during the first stages of development.

References

[1] L. Turchet, “Musical metaverse: vision, opportunities, and challenges,” *Personal and Ubiquitous Computing*, vol. 27, no. 5, pp. 1811–1827, 2023.

[2] A. Renaud, A. Carôt, and P. Rebelo, “Networked music performance: State of the art,” in *Proceedings of the AES 30th International Conference*. Finland Saariselkä, 2007, pp. 16–22.

[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] E. F. Churchill and D. Snowdon, “Collaborative virtual environments: an introductory review of issues and systems,” *virtual reality*, vol. 3, pp. 3–15, 1998.

[5] J. G. Tromp, A. Steed, and J. R. Wilson, “Systematic usability evaluation and design issues for collaborative virtual environments,” *Presence: Teleoperators & Virtual Environments*, vol. 12, no. 3, pp. 241–267, 2003.

[6] C. Cortés, P. Pérez, and N. García, “Understanding latency and qoe in social xr,” *IEEE Consumer Electronics Magazine*, vol. 13, no. 3, pp. 61–72, 2023.

[7] A. Boem, M. Tomasetti, and L. Turchet, “Issues and Challenges of Audio Technologies for the Musical Metaverse,” *J. Audio Eng. Soc*, vol. 73, no. 3, pp. 94–114, 2025.

[8] R. Cheng, N. Wu, M. Varvello, S. Chen, and B. Han, “Are we ready for metaverse? a measurement study of social virtual reality platforms,” in *Proceedings of the 22nd ACM internet measurement conference*, 2022, pp. 504–518.

[9] H. Wang, R. Martinez-Velazquez, H. Dong, and A. El Saddik, “Experimental studies of metaverse streaming,” *IEEE Consumer Electronics Magazine*, vol. 14, no. 1, pp. 26–36, 2024.

[10] H. Vo and R. Boulanger, “Csound in the metaverse – from cabbage to csoundunity and beyond: Developing a working environment for soundscapes, soundcollages, and collaborative soundplay,” in *Proceedings of the 7th International Csound Conference*. Vienna, Austria: ICSC 2024, September 2024, pp. 115–119.

[11] L. Men and N. Bryan-Kinns, “LeMo: supporting collaborative music making in virtual reality,” in *2018 IEEE 4th VR workshop on sonic interactions for virtual environments (SIVE)*. IEEE, 2018, pp. 1–6.

[12] A. Boem, D. Dziwis, M. Tomasetti, S. Etezazi, and L. Turchet, ““it takes two”-shared and collaborative virtual musical instruments in the musical metaverse,” in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

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

[13] M. Buffa, D. Girard, and A. Hofr, “Using web audio modules for immersive audio collaboration in the musical metaverse,” in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[15] C. Drioli, C. Allocchio, and N. Buso, “Networked performances and natural interaction via LOLA: Low latency high quality A/V streaming system,” in *International conference on information technologies for performing arts, media access, and entertainment*. Springer, 2013, pp. 240–250.

[16] G. Hajdu, “Embodiment and disembodiment in networked music performance,” in *Body, Sound and Space in Music and Beyond: Multimodal Explorations*. Routledge, 2017, pp. 257–278.

[17] S. J. Friston, B. J. Congdon, D. Swapp, L. Izzouzi, K. Brandstätter, D. Archer, O. Olkkonen, F. J. Thiel, and A. Steed, “Ubiq: A System to Build Flexible Social Virtual Reality Experiences,” in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST ’21. New York, NY, USA: Association for Computing Machinery, 2021.

[18] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, “Evaluation of metaverse music performance with bbc maida vale recording studios,” *Journal of the Audio Engineering Society*, pp. 313–325, 2023.

[19] L. Turchet and C. Fischione, “Elk Audio OS: an open source operating system for the internet of musical things,” *ACM Transactions on Internet of Things*, vol. 2, no. 2, pp. 1–18, 2021.

[20] L. Turchet and P. Casari, “On the Impact of 5G Slicing on an Internet of Musical Things System,” *IEEE Internet of Things Journal*, vol. 11, no. 19, pp. 32 079–32 088, 2024.

[21] ——, “Latency and Reliability Analysis of a 5G-Enabled Internet of Musical Things System,” *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 1228–1240, 2024.

[22] D. Wessel and M. Wright, “Problems and prospects for intimate musical control of computers,” in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2001, pp. 11–14.

[23] S. Fels, “Designing for intimacy: Creating new interfaces for musical expression,” *Proceedings of the IEEE*, vol. 92, no. 4, pp. 672–685, 2004.

[24] F. Biocca and C. Harms, “Networked Minds Social Presence Inventory:—(Scales only, Version 1.2) Measures of co-presence, social presence, subjective symmetry, and intersubjective symmetry,” 2003.

[25] E. A. Carroll and C. Latulipe, “The creativity support index,” in *CHI’09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2009, pp. 4009–4014.

[26] C. Bartlette, D. Headlam, M. Bocko, and G. Velikic, “Effect of network latency on interactive musical performance,” *Music Perception*, vol. 24, no. 1, pp. 49–62, 2006.

[27] I. R. Roman, A. Washburn, E. W. Large, C. Chafe, and T. Fujioka, “Delayed feedback embedded in perception-action coordination cycles results in anticipation behavior during synchronized rhythmic action: A dynamical systems approach,” *PLoS computational biology*, vol. 15, no. 10, p. e1007371, 2019.

[28] B. Van Kerrebroeck, K. Crombé, S. M. de Leymarie, M. Leman, and P.-J. Maes, “The virtual drum circle: polyrhythmic music interactions in mixed reality,” *Journal of New Music Research*, vol. 52, no. 4, pp. 316–336, 2023.

[29] J. Napoles, “Verbal instructions and conducting gestures: Examining two modes of communication,” *Journal of Music Teacher Education*, vol. 23, no. 2, pp. 9–20, 2014.

[30] C. Yarbrough and K. Madsen, “The evaluation of teaching in choral rehearsals,” *Journal of Research in Music Education*, vol. 46, no. 4, pp. 469–481, 1998.

[31] L. Bishop, “Collaborative musical creativity: How ensembles coordinate spontaneity,” *Frontiers in psychology*, vol. 9, p. 1285, 2018.

[32] ——, “Togetherness in musical interaction,” *Routledge Open Research*, vol. 3, no. 16, p. 16, 2024.

[33] G. Grimm, M. Daeglau, V. Hohmann, and S. Debener, “Eeg hyperscanning in the internet of sounds: Low-delay real-time multi-modal transmission using the ovbox,” in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–8.