Networked Musical XR: where's the limit? A preliminary investigation on the joint use of point clouds and low-latency audio communication

Luca Turchet
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy
luca.turchet@unitn.it

Nicola Garau
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy
nicola.garau@unitn.it

Nicola Conci
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy
nicola.conci@unitn.it

Figure 1: A picture taken during the testing phases of the proposed networked musical XR solution. The participant, while playing the instrument is able to visualize the virtual scene with the other musicians. As it can be seen the whole architecture consists of off-the-shelf hardware, that can be easily configured to suit the application scenario.

ABSTRACT

As of today, the field of networked musical XR is in its infancy. While the next generation networks keep pushing the available bandwidth towards new frontiers, promoting the deployment of new services and applications, a limited amount of residual latency still hinders the possibility for musicians to seamlessly interact over the Internet. In fact, while audiovisual (and audio in particular) streaming has reached high performances, allowing for smooth interaction in many application scenarios such as video calls and dialogues, the study of tools to ensure a flawless immersive interplay experience among musicians is a rather unexplored area. This paper reports a preliminary investigation on a technical setup that couples a networked music performance system with an XR system, conceived to interconnect geographically displaced musicians. The setup we have envisaged has allowed us to identify the existing technical issues in current technologies used for networked musical XR. We discuss such issues and reason about possible future research directions in this area that should be covered to advance the current state of the art.

AM '22, September 6–9, 2022, St. Pölten, Austria

ACM ISBN 978-1-4503-9701-8/22/09…$15.00
https://doi.org/10.1145/3561212.3561237

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; • **Human-centered computing** → *Mixed / augmented reality*; • **Information systems** → Internet communications tools.

KEYWORDS

Musical XR, Internet of Musical Things, Augmented Reality, Networked Music Performance Systems

ACM Reference Format:

Luca Turchet, Nicola Garau, and Nicola Conci. 2022. Networked Musical XR: where's the limit? A preliminary investigation on the joint use of point clouds and low-latency audio communication. In *AudioMostly 2022 (AM '22), September 6–9, 2022, St. Pölten, Austria.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3561212.3561237

1 INTRODUCTION

The recent COVID-19 pandemic has promoted new forms of network-based communication in various domains, and music is no exception. Latency (i.e., the time it takes for data to be transmitted between two musicians) and jitter (i.e., the variation of latency over time) represent crucial issues for networked musical communication. In fact, in order for musicians to play synchronously, an empirical threshold in the order of 20-30 ms is considered as the maximum acceptable latency, making it a quite tight constraint. This has been proven in several perceptual studies (for a review see [12]). Various networked music performance (NMP) systems have been developed to interconnect musicians at both auditory and visual level. Noticeable examples of NMP systems are LoLa [6], Jacktrip [4], and Elk LIVE [14]. Most of research efforts in this space have focused on the delivery of auditory content (e.g., in terms of protocols, codecs, and packet loss concealment methods). As far as the visual information is concerned, this is generally provided in the form of video streaming. On the one hand, video streaming relies on the adoption of established TCP- or UDP-based protocols; on the other hand, it requires a considerably higher bandwidth in transmission, as the raw data can be orders of magnitude more demanding that the auditory counterpart. This necessarily implies an increased delay and, as a consequence, a de-syncrhonization of the two flows.

Unlike video streams, which tend to replicate a video conferencing toolkit, eXtended Reality (XR) technologies have the ability to provide immersive experiences. The term XR itself, encompasses both Virtual Reality (VR) and Augmented Reality (AR) [16], navigating the spectrum that spans from reality to virtuality in a continuous fashion. In particular, the "Musical XR" term has recently emerged to indicate the adoption of XR technologies in musical contexts [15]. AR can create a 3-dimensional environment overlaid on a user's current surrounding. Therefore, AR is an ideal medium for musical communication over the network, given its ability to connect musicians in ways that can recreate the sensation of sharing the same physical space, generating the so-called sense of social presence [3]. Indeed, in AR it is possible to visualize the remotely connected musicians as fully-articulated avatars which move in the 3D space (as opposed to miniatures of faces and bodies on a 2D screen). However, to date, scarce research has been conducted on the use of AR techniques in conjunction with NMP systems [12]. This is a common issue with VR, as evidenced by different authors [11, 15]. As a result it is still unclear what are the exact challenges in transmitting AR information in real-time over the network for musical collaboration.

In this paper we present the evaluation of a case study, aimed to assess the performances and limitations of an XR-based NMP framework, from a technical standpoint. We report preliminary results obtained using the prototype we have developed, which couples the networked music performance system with an AR-based framework to make possible the interconnection among geographically displaced musicians. At the current level of investigation, the main goal of our study is the identification of the technical limitations, when relying on off-the-shelf state of the art devices when deployed in an ideal scenario, namely communicating over a LAN network. The study has allowed us to gather information about the pros and cons of such an architecture, paving the way for the definition of future research directions and the investigation of novel communications paradigms to virtually bring musicians nearer. This work aims at answering the call of the authors of [15] in progressing the networked Musical XR field.

2 SYSTEM ARCHITECTURE AND COMPONENTS

For our study we have set up an audio-visual communication framework in which the participants are located in different rooms, and are asked to perform a simple jam session. The system is composed by the following hardware components that were used by each musician:

- **NMP system.** We used the board of the Elk LIVE NMP system to transmit and receive the auditory signals [14];
- **XR headset.** An HTC Vive Cosmos Elite was utilized to render the AR content at visual level;
- **High-resolution camera.** A ZED mini camera was attached to the HTC headset in order for the musicians to see their own body and instrument;
- **External high-resolution camera**. A ZED 2 camera was mounted on a tripod and served the purpose of detecting the musicians' body and instrument to be transmitted over the network.

Software-wise, the components required to run the system are listed hereafter:

- **Point cloud rendering engine.** The point cloud is rendered and displayed using Unity 3D.
- **NMP-audio software.** Elk LIVE NMP system based on the ELK Audio low-latency operating system [14]
- **NMP-AR software.** AR content was streamed via a P2P network based on the Unity Render Streaming platform.

In Figure 1 an overview of the proposed hardware architecture is shown. The musician's 3D avatar is captured by the ZED 2 camera mounted on a tripod. The point cloud resembling the avatar is then processed and filtered, before being sent via a P2P dedicated communication protocol to the second musician, who can experience it in a custom 3D environment rendered through the HTC Vive XR headset. The customization is meant to let the participants decide the configuration of recreated virtual scene based on personal taste or application requirements.

Figure 2 depicts a detailed overview of the architecture of the implemented networked musical XR framework. The system was conceived and tested for two musicians, although it can be easily

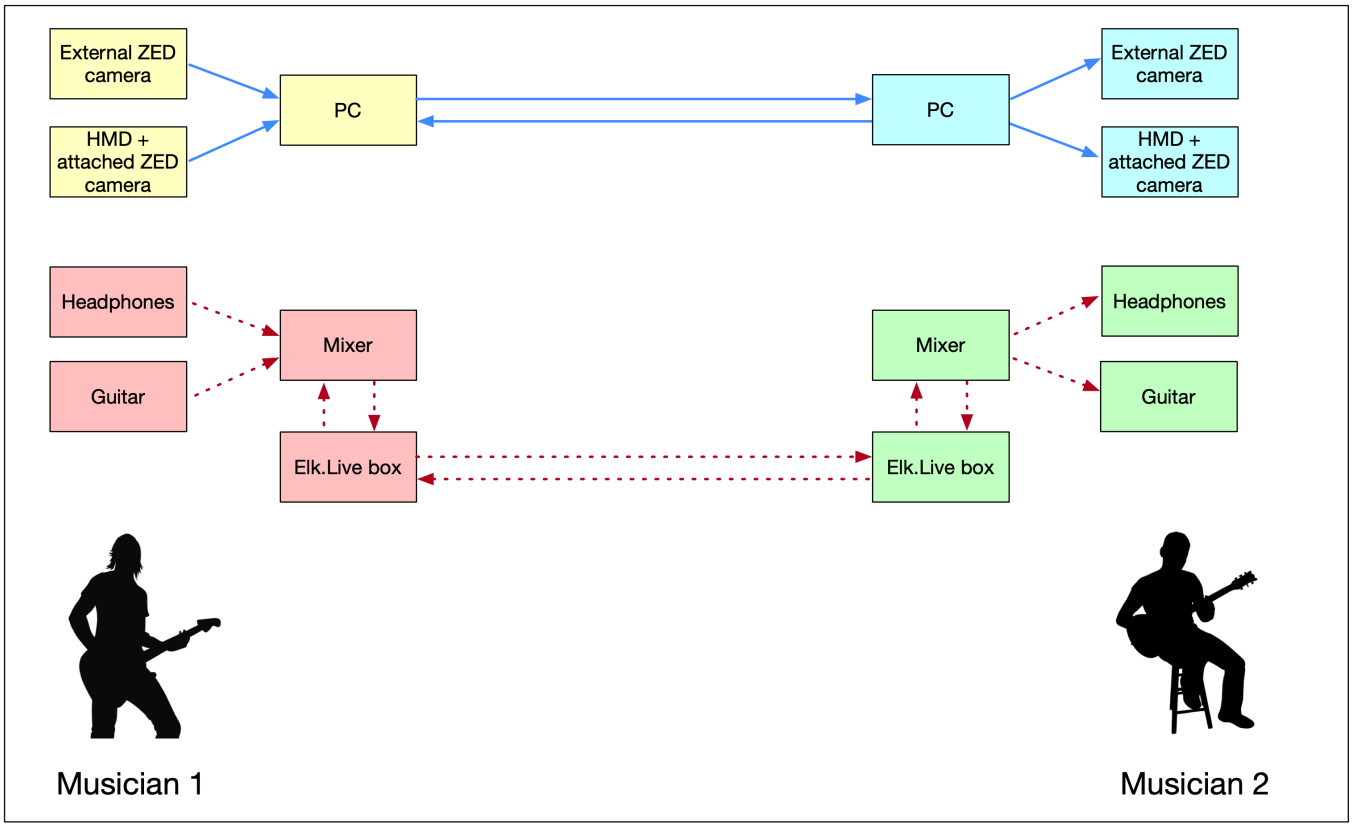

Figure 2: The proposed framework. The two musicians, geographically displaced, are provided with a dedicated audio-visual streaming equipment. The video flow is handled by a regular PC and a stereo camera; the audio streaming is implemented using the Elk LIVE low-latency NMP system.

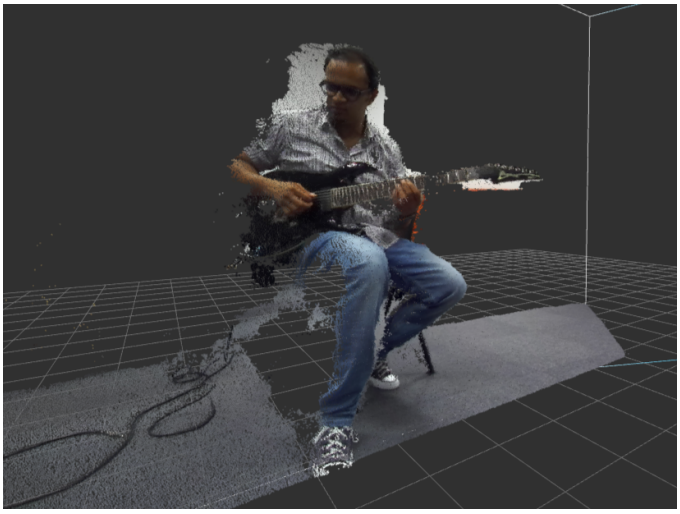

Figure 3: Example of a point cloud captured by the ZED camera during a live performance

extended to four participants without loss of generality (currently Elk LIVE supports the interconnection of four musicians).

Regarding the communication of the signals generated by the musicians, the Elk LIVE NMP system is based on a P2P architecture. Musicians simply plug their instrument in a box to stream the content they produce while also receiving from it the sound of the connected musicians (via headphones). Optionally a mixer can be used to connect more musical instruments to a single box.

As far as the visual information is concerned, we used a high-resolution rendering pipeline inside Unity 3D for the real-time rendering of the *point cloud* [8] (see Figure 3). The *point cloud* is captured directly inside Unity via the ZED SDK, with an expected latency of 20ms-30ms (depending on the underlying hardware configuration). The point cloud streaming is handled by the Unity Render Streaming platform, based on the fast WebRTC protocol with hardware encoding/decoding, which guarantees a sub–500ms delivery latency. Overall, the upper bound for the streaming of the virtual 3D avatar in normal streaming conditions should not be greater than 530ms. In practice, during preliminary evaluation with optimal hardware and network conditions (the LAN made available at the university premises) we observed a lower video delay, in the order of about 400ms. This quantity has been measured experimentally over different trials, at different times of the day, in different network traffic conditions. The latency has proven to be stable across the different tests.

For the considered avatar streaming scenario, point clouds offer a more efficient representation over other data types such as 3D meshes, since they require no additional pre-processing and are easier to filter and downsample. In addition, the ZED camera further allows for the detection and tracking of human body joints in real-time, similarly to other time-of-flight cameras (e.g., Kinect [13]), but with higher resolution (up to 2K) and better skeleton tracking. The skeleton information can be used to filter out points not belonging to the tracked avatar, further reducing the amount of data to be delivered. Moreover, by rigging a pre-computed 3D mesh of the avatar using well known mesh reconstruction algorithms in literature [2, 7, 10], only a few keypoints of the skeleton rig would need to be transmitted over the network, unlocking new possibilities for networked musical XR real-time applications.

Concerning the audio signal communication, the preliminary evaluation conducted using the LAN at the university premises showed an average round-trip latency of about 4 ms. This meant that in our system the visual information arrived to the connected musician with a delay of about 400 ms.

3 PRELIMINARY EVALUATION

The goal of the evaluation was to assess whether the proposed system was suitable for networked musical communications. The evaluation involved two expert guitar players (both males, aged 44 and 39, each with more than 30 years of musical experience). The evaluation was conducted at the facilities of the authors' institution. Specifically, the two musicians were placed in different rooms of adjacent buildings, interconnected through the same LAN network. This provided ideal conditions with respect to network latency and jitter.

Musicians were instructed to wear the XR headset, the headphones and afterwards to hold their instrument (see Figure 1). The assigned task was to play together improvising on a blues in E minor, while paying attention to the other musician's gestures at visual and auditory levels. Results were not encouraging. The latency was well above the perceptual threshold for noticing an audio-visual mismatch. While a specific measurement was not performed we quantified it as about 400 ms of delay. This estimate found empirically by playing strumming chords at 150 BPM (i.e., 400 ms between beats), and noticing that there was a delay of one strumming chord between audio and video). The perceived delay was not tolerable by musicians, who reported to feel disoriented while using the system. For instance, the delay was immediately perceived when a musician played rhythmic patterns such as chords strumming. Moreover, notwithstanding the idea of a Musical XR system was appreciated for its potential, playing with the XR headset was deemed cumbersome and impractical.

While network conditions were ideal, summing up to a few milliseconds of latency for the transmission of the visual information, the tracking of the connected musician's gestures and their visual rendering as a point cloud hologram were the main bottlenecks in the implemented system. On the other hand, the latency of the auditory communication was ideal, ensured by the Elk LIVE system used over a LAN. Mismatches between audio and video can be tolerated by humans to a certain extent, as shown in various studies from the field of multisensory integration [1, 5, 9] (e.g., for speech is about 250 ms according to the study reported in [5]). However, as of today to our best knowledge no study has investigated to which extent a delay between auditory and visual information can be perceived or tolerated in the context of networked musical interactions in AR.

Another aspect that emerged during the evaluation is that musicians often rely on facial visual cues. However, the use of a cumbersome XR headset such as the one used in the present study, occludes the tracking of a large part of the face of the connected musician, thus limiting an important part of the non-verbal communication typically occurring during a real collaborative music making session. Therefore, the study highlighted also the need for new and less occluding XR devices, such as glasses.

4 CONCLUSIONS

Our proposed system was found not usable by the involved expert musicians given a too high delay between auditory and visual information. We believe that reporting negative results is important to communicate to the musical XR community the attempted solutions that are not working. More importantly, our attempt, which involved consumer grade equipment, highlights the need to advance hardware and software technologies for the specific networked musical XR application.

Notably, the study was conducted under ideal network conditions where latency and jitter introduced by the communication were low and constant, and did not reflect fluctuations and packet losses of the Internet. This allowed us to isolate the components responsible for the audio-visual mismatch using the adopted hardware. The study pinpointed the need for cameras able to acquire data with higher sample rates, software for processing gestural information, as well as new methods to render the tracked musician in the form of avatar in an XR environment. Furthermore, our study highlighted the need for less cumbersome and lighter XR headsets. It is worth noticing that the utilized setup involved state-of-the-art consumer grade equipment, which overall is not cost-effective (about 1500 euros), especially considering that the setup has to be replicated for each musician. Therefore, there is also the need for more affordable technologies.

In future work, we plan to investigate to which extent audio-visual latency mismatch can be perceived and tolerated by musicians while playing (using a Just Noticeable Difference approach), as well as how tolerable latencies affect musical communication. This will provide effective guidelines for designers of networked musical XR systems.

REFERENCES

[1] W.J. Adams. 2016. The development of audio-visual integration for temporal judgements. *PLoS Computational Biology* 12, 4 (2016), e1004865.
[2] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. 1999. The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics* 5, 4 (1999), 349–359.
[3] S.T. Bulu. 2012. Place presence, social presence, co-presence, and satisfaction in virtual worlds. *Computers & Education* 58, 1 (2012), 154–161.
[4] J.P. Cáceres and C. Chafe. 2010. JackTrip: Under the hood of an engine for network audio. *Journal of New Music Research* 39, 3 (2010), 183–187.
[5] M.J. Crosse, G.M. Di Liberto, and E.C. Lalor. 2016. Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience* 36, 38 (2016), 9888–9895.

[6] C. Drioli, C. Allocchio, and N. Buso. 2013. 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, 240–250.
[7] M. Kazhdan, M. Bolitho, and H. Hoppe. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, Vol. 7.
[8] L. Linsen. 2001. *Point cloud representation*. Technical Report. University of Karlsruhe, Faculty of Computer Science.
[9] J. Liu, V. Drga, and I. Yasin. 2021. Optimal Time Window for the Integration of Spatial Audio-Visual Information in Virtual Environments. In *Proceedings of IEEE Virtual Reality and 3D User Interfaces*. IEEE, 723–728.
[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.
[11] B. Loveridge. 2020. Networked Music Performance in Virtual Reality: Current Perspectives. *Journal of Network Music and Arts* 2, 1 (2020), 2.
[12] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti. 2016. An Overview on Networked Music Performance Technologies. *IEEE Access* 4 (2016), 8823–8843.
[13] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. Ieee, 1297–1304.
[14] L. Turchet and C. Fischione. 2021. Elk Audio OS: an open source operating system for the Internet of Musical Things. *ACM Transactions on the Internet of Things* 2, 2 (2021), 1–18.
[15] L. Turchet, R. Hamilton, and A. Çamci. 2021. Music in Extended Realities. *IEEE Access* 9 (2021), 15810–15832.
[16] M. Vasarainen, S. Paavola, and L. Vetoshkina. 2021. A systematic literature review on extended reality: Virtual, augmented and mixed reality in working life. *Internationa Journal of Virtual Reality* 21, 2 (2021), 1–28.