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 Mar.). https://doi.org/10.17743/jaes.2022.0193.

Issues and Challenges of Audio Technologies for the Musical Metaverse

ALBERTO BOEM,* MATTEO TOMASETTI, AND LUCA TURCHET, *AES Member*
(alberto.boem@unitn.it) (matteo.tomasetti@unitn.it) (luca.turchet@unitn.it)

Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Among all the activities envisioned for the metaverse, music has thus far received comparatively less attention. While virtual concerts and music festivals have been successful in drawing substantial audiences and increasing public attention to the idea of the metaverse, the metaverse is not ready for musicians who decide to take advantage of the distinctive features of socially immersive environments to express themselves and create music together. In this article, the authors analyze the state-of-the-art audio technologies used for the creation of shared, *Audio-First* immersive environments such as the musical metaverse. This work reveals important issues in consumer electronics that currently prevent the realization of a metaverse compatible with musical activities. These include hardware and software limitations used to create and experience shared immersive environments through real-time audio. This work also emphasizes two key challenges: reducing delays in network and audio processing, and addressing the lack of universal standards for spatial audio systems across different platforms. The authors believe that looking at the metaverse from the point of view of musical technologies will provide practitioners in academia and industry with key insights into what is needed to achieve true real-time activities and support human expression in the metaverse in general.

0 INTRODUCTION

Music events, such as concerts and festivals, are among the most popular social activities available in multiuser immersive social environments or in what is now referred to as the metaverse [1, 2]. From the concert of U2 in Second Life to the private clubs in VRChat or the mixed reality (MR) performances of Jean-Michel Jarre in Versailles, various music venues have started to operate in such worlds (i.e., Fortnite, Somnium Space, VRROOM, PatchWorld), drawing attention not only from a growing number of audiences but also from sponsors, investors, and media.

Although these types of musical events and dedicated platforms offer a unique set of experiences, they are primarily designed for audience participation rather than performer interaction. In these environments, audience members, usually represented as 3D avatars, gather in virtual worlds where an artist or a band plays in front of them. In their simplest form, virtual concerts bear a more remarkable resemblance to virtual conferences: a video of a music performance is broadcast directly to users through a 2D projection in a 3D space. In more advanced examples, a musician (usually a singer or a DJ) performs as animated avatars on a virtual stage and synchronized to a prerecorded track. In only a few instances, the audio is actually captured while an artist or a band is performing and transmitted in real time to the audience. On even rarer occasions, the avatar is controlled directly by the artists' movements that are acquired by a motion-capture system while performing.

However, virtual concerts are just one instance of the broader concept known as the musical metaverse (MM) [3]. The MM was proposed as the part of the metaverse explicitly dedicated to musical activities. This emerging space, which blends virtual and augmented worlds, promises to redefine how music is created, shared, and enjoyed, entailing a significant shift in the traditional understanding of musical engagement. The MM will be made possible by combining different technologies, particularly musical extended reality (XR) [4] and the Internet of Musical Things (IoMusT) [5].

Musical XR is a field positioned at the confluence of the domain of interactive music and immersive technologies, augmented reality (AR), mixed reality (MR), and virtual reality (VR). The IoMusT aims to extend the paradigm of the Internet of Things to the creation and use of networked musical devices, fostering innovative interactions among

*To whom correspondence should be addressed, email: alberto.boem@unitn.it

musicians, audiences, and other musical stakeholders. Networked music performance (NMP) systems represent a central component of IoMusT [6–8], which refers to technologies that allow musicians in different geographical locations to play together in real time over a wired or wireless network.

While current research focuses primarily on audience participation [9–11], the broader vision of MM aims to support not just concert attendance but all musical activities. Commercially available and free platforms for immersive music making are not fully optimized to support music production at the level of precision and reliability required by professional musicians, especially when the playing experience is collaborative.

To date, geographically displaced musicians (equipped with their instruments and sound-producing equipment) have minimal means to actively participate in virtual concerts. Activities, such as playing music in real time with other people, rehearsing and recording music, and even teaching music, are not fully supported. Although technologically mediated audience participation in immersive environments has already received a considerable amount of attention from scholars (e.g., [12–16]), technologies for networked music making in immersive environments have received comparatively less attention [17, 18].

Moreover, despite a few prototypes (e.g., [19–25]) or some limited proof of concepts and preliminary studies (e.g., [26–31]), from the point of view of audio technologies, the MM remains mainly in an early phase. However, these works offer a glimpse into what the future of music might hold and present unique challenges and opportunities for technologists and musicians alike. The reason behind these limitations can be found in the current limits of both XR and networking technologies for music.

First, creating a compelling and realistic experience for two or more musicians connected over the Internet represents a significant challenge due to latency, jitter, and bandwidth. While these challenges are common to the field of NMP [6], the inclusion of social aspects and the technologies required for capturing gestures and delivering multimodal signals introduces a distinct set of issues.

Second, although audio is a key component of the metaverse [32], it is currently used primarily to enhance immersion through ambience and auditory feedback. However, this use of sound fails to capture the subtleties and unique attributes inherent in music making. By following Çamcı's and Hamilton's position, the MM should be considered as *Audio-First* to differentiate it from other immersive media where audio represents an ancillary component of the experience and not the main focus [33].

Third, achieving high-quality audio and seamless interaction requires the development of domain-specific tools. The path toward a metaverse oriented to musical activities is not only tightly linked to the progress of networking technologies but also heavily dependent on the characteristics of the consumer electronics used to experience and interact with immersive environments, both at the hardware and software levels.

Through this paper, the state of the art of audio technologies for the metaverse is analyzed. Next, the main challenges and open issues for realizing the MM are identified and explored. Specifically, the challenges analyzed encompass the hardware and software constraints in developing and using shared immersive environments with real-time audio capabilities, such as minimizing latency in network and audio processing and the absence of standardized approaches for synchronization between musical and XR data. The authors posit that examining the metaverse through the lens of musical technologies offers valuable insights for both academic and industry professionals. This perspective is essential for understanding the requirements in facilitating authentic real-time interactions and enhancing human expression within the metaverse. Nevertheless, it is essential to note that the MM is not a mere combination of these technologies. Networked and immersive environments should also support social presence and interactions [34–38].

In the metaverse, users share virtual environments designed for collaboration and interpersonal communication. Musical applications, with their unique requirements for social interactions, coordination, precise timing, and expressive performance, can help address key challenges in general metaverse development.

1 FRAMING THE MUSICAL METAVERSE

The concept of the metaverse was introduced by science fiction writer Neal Stephenson in 1992 with the book *Snow Crash,* which describes a virtual world that users can traverse and where they can engage in different activities and interact with a global community of individuals [39]. Since Stephenson's novel was published, his vision has inspired many virtual social platforms like "Active Worlds" and "Second Life". The term metaverse has since been adapted to serve different industries and their specific needs.

A survey of 54 academic papers conducted in 2022 revealed a wide variety of definitions developed for the metaverse [40]. While these definitions reveal differences of perspectives (mainly depending on their focus, such as tourism [41], education [42], and healthcare [43]), they also exhibit some commonalities regarding environment, interactions, and users. Therefore, the metaverse can be considered as a virtual space connected to or reflecting some aspects of the physical world that supports and facilitates social interactions among users through multisensory and immersive experiences. Notably, some definitions have also highlighted the importance of interoperability, such as letting users interact and exchange assets across different platforms through the seamless integration of different immersive environments [44] and persistence, which is the capability of a virtual world to remain consistent, even after users log off [45].

In these environments, 3D embodied avatars are commonly used as the means through which users manifest and perform their actions, even though volumetric live videos could also be used to enhance realism. Moreover, the metaverse should not be seen as just another term for AR, MR,

VR, XR, or "Spatial Computing." Instead, it represents a convergence of different technologies, including the Internet of Things, edge computing, ultrareliable low-latency communications, multimodal tracking, multisensory rendering, and digital twins. From this perspective, the metaverse is typically described as a "service" that should provide users with a high-quality experience and support for human activities.

Music can also be considered a specific application of this service. However, creating music in the metaverse has unique requirements shaped by its inherent nature and its ties to traditional musical practices, including exploration and experimentation. While previous analyses and propositions of the metaverse have focused on XR and networking [46], they have often treated audio as an ancillary element mostly limited to human speech [47, 48]. Therefore, the analysis here must consider a scenario where people are transported to remote environments to engage in musical interactions such as playing and making music together.

1.1 Interaction Framework

To better understand how the MM functions, it is important to clarify how the interactions and communications between users unfold. This study adopts a framework proposed by Cortés et al. [35] specifically designed for interactions in socially immersive environments. The research also integrates the framework proposed by Turchet [3] that includes a layer dedicated to musical interactions. The framework is presented in Fig. 1.

The interactions in the MM can be represented as an interplay between different connected realities. From a musician's perspective, their local environment can be defined as the *Self Reality*. This reality is composed of different layers.

The first is the *Physical Layer*, which encompasses all the devices and objects associated with the musician's physical reality. These can be divided into three main components: the *Sensor Component, Interaction Component*, and *Musical Component*. The first component regards the sensors used to capture and register musicians' body and movements (i.e., depth cameras [26] or full-body tracking systems [20]). The second component includes all the tools and interfaces for interacting with the virtual worlds, especially for controlling XR instruments, such as hand tracking for interacting with the 3D user interface (UI) of a virtual synthesizer [50], or hand-held controllers for playing a bowed virtual instrument [51]. The third one pertains to the components used to inject audio directly into the metaverse, such as an electric guitar [52] or a microphone [21].

The data produced by all these components are then transmitted to the *Processing Layer*. Within this layer, two processes can be identified: *i*) the *XR Process*, which is dedicated to the data coming from sensors and interaction components and *ii*) the *Audio Process*, which is dedicated to the audio data. The XR Process takes data from head-mounted displays (HMDs), hand-held controllers, and other sensors. Next, it processes them to establish tracking, recognize certain gestures, or control a 3D avatar.

The *Audio Process* is responsible for the auditory part of the experience. Audio can be generated either from a real-time synthesis algorithm driven by a series of gestures from the previous process, or it can be the audio signal of an input instrument, such as an electric guitar, to which one can apply effects through Virtual Studio Technology plugins. This process is also responsible for all the elements involved in the creation of a NMP system, such as packetization or the encoding of data into Musical Instrument Digital Interface (MIDI) and Web MIDI or Open Sound Control. These two processes can happen inside the HMDs, in a host computer, a dedicated device, the cloud, or on edge platforms. The results of the XR and audio processes are then simultaneously sent back to the musicians to render the virtual world both aurally and visually and then sent to the network.

Subsequently, the data of the *Self Reality* of a musician are transmitted to the *Distant Reality*, which represents the reality inhabited by a remote musician. The received audio undergoes two steps: first depacketization, then rendering through spatial audio algorithms.

The data from the two realities are then merged and used to represent musicians and their sounds in an immersive XR environment, the *Immersive Layer*, encompassing AR, MR, and VR. This layer is then synchronized with the physical one, creating the conditions for a persistent MM where actions and music happen in real time.

1.2 Playing Together in the MM

The MM should accommodate different types of musical activities. In the context of NMP systems, Renaud et al. [53] have identified two primary use cases, realistic jam and remote recording. The case of realistic jam has been not only the one that received the most significant attention from researchers, but it is also the one characterized by the most complex challenges. Therefore, it is considered as the focus of this investigation.

The goal of a realistic jam over the network is to get as close as possible to making geographically displaced musicians feel like they are playing in the same physical space. Technically, this involves sending and receiving relevant data as quickly as possible to maintain stability and synchronization among the peers and ensuring that the quality of the audio and data transmitted between them is satisfactory. Additionally, it is essential to consider the sound spatialization (as recently demonstrated in [54]), which must allow for the localization and directionality of the sound sources to be simulated as it occurs in real environments.

However, how realistic jamming is designed and supported depends on the type of instruments used and how music making unfolds. Two main approaches and systems can be identified: one where musicians play virtual instruments (XR musical instruments) and the other where musicians make use of physical acoustic, electric, or digital

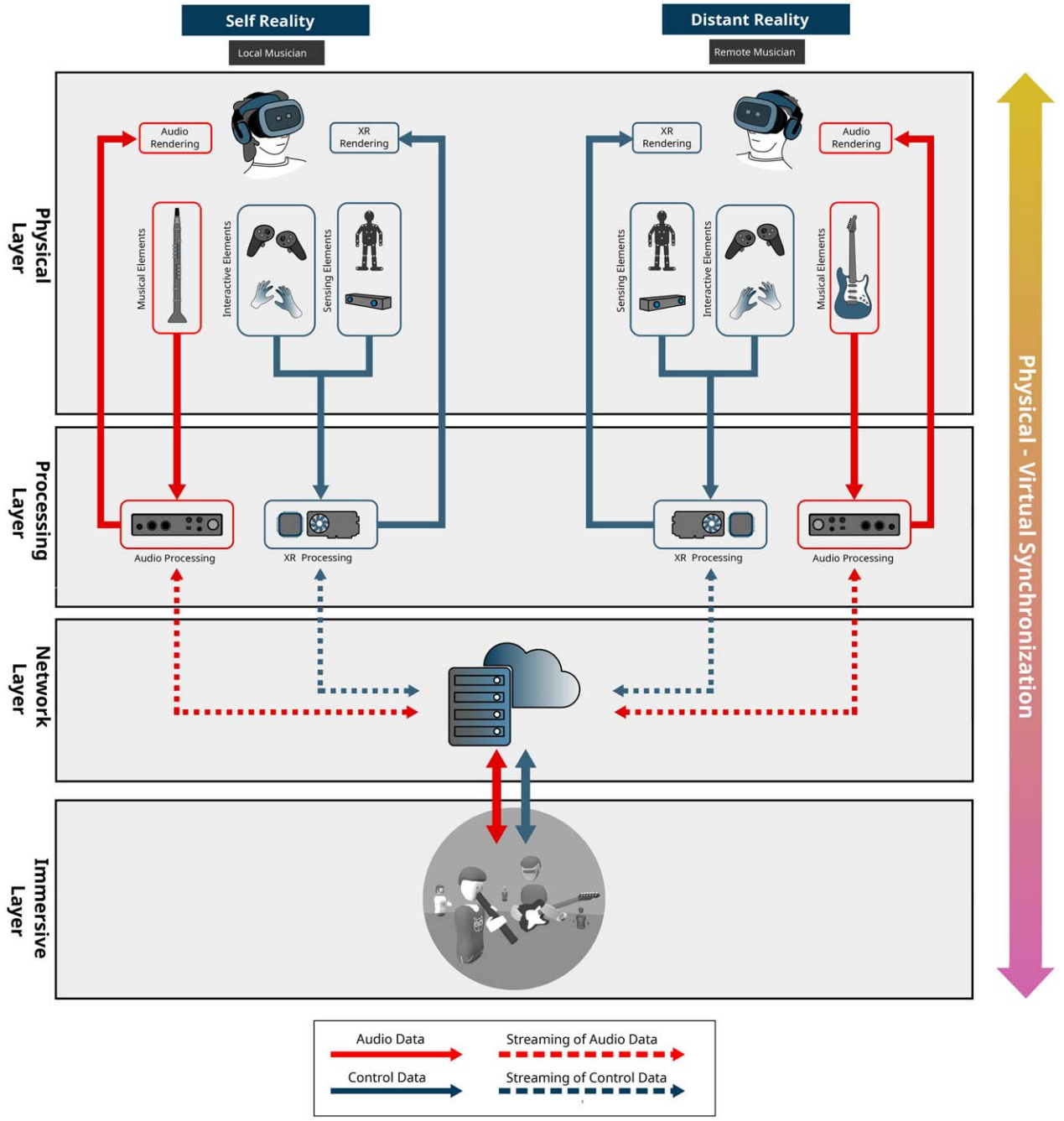
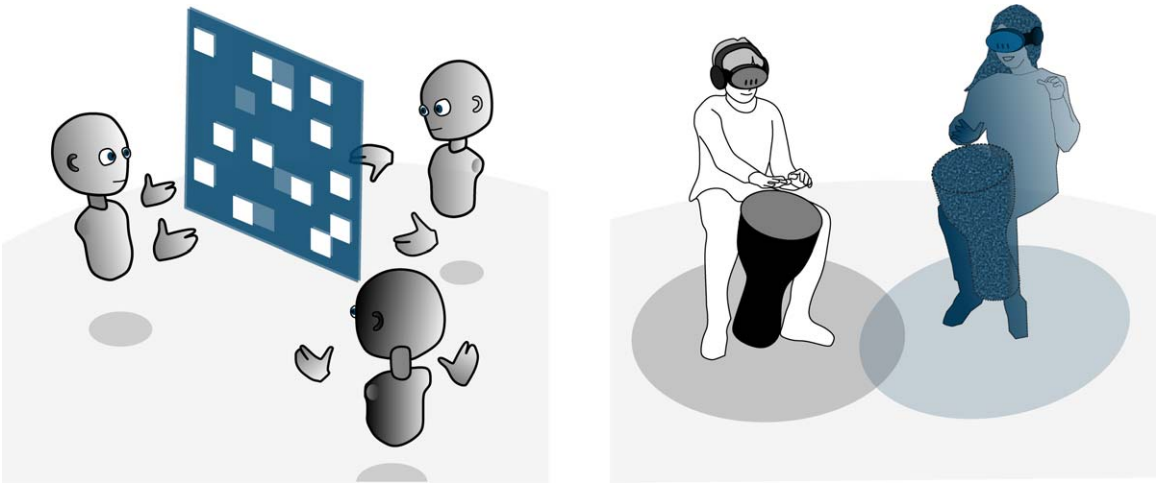


Fig. 1. A representation of the interaction framework for the MM. The diagram exemplifies the interaction components and the technology layers that constitute the MM. To exemplify it, the case of two musicians displaced in different locations is shown. To better understand how this framework can be applied to different use cases, refer to Fig. 4 for XR musical Instruments and Fig. 5 for physical musical Instruments as discussed in 1.2.

instruments (Physical musical instruments). These are detailed hereinafter.

- **XR musical instruments:** This is the most recurrent approach, explored by both researchers and artists, and employed in commercial applications such as PatchWorld.[1] In such systems, connected users can play together instruments that exist only in the form of virtual objects controlled through 3D UIs with gestures or dedicated input devices. XR musical instruments can be 3D sequencers [50], virtual synthesizers [23, 51–53], or new types of collaborative and multiuser musical instruments [21, 24, 25, 57, 55, 56]. An example of this approach is illustrated in Fig. 2(a).
- **Physical musical instruments:** While being the most realistic, this approach has received relatively less attention. Previous work includes compositions for violins [60], rehearsals and recording spaces for rock bands [27, 52], systems for learning and play-

[1] PatchWorld, https://patchxr.com/.



(a) (b)

Fig. 2. The two main approaches for realistic jamming in the MM with XR musical instruments (a) and with physical musical Instruments (b). (a) A representation of an XR musical instrument: three musicians represented as 3D embodied avatars play a virtual collaborative sequencer in VR. The sound is created in real time through a synthesis engine (see [21, 50, 57, 65]). (b) A representation of a physical musical instrument: two musicians play a percussion. They see each other through MR glasses and are represented as volumetric point-cloud videos. They hear the sound produced by the other in real time (see [26, 28, 29, 66]).

ing percussion [29, 61], and group singing [62]. An example of this approach is shown in Fig. 2(b).

These examples cover several applications from group rehearsals to concerts in 3D environments. The difference in these approaches has an impact on the type of data transmitted and on computational and network requirements. They also represent the most complex use of real-time musical interactions in the MM, and they will be analyzed in terms of the issues of audio technologies. After presenting a framework and examples to highlight the key characteristics of the MM, it is essential to recognize that, despite the possibilities it offers, realizing this vision involves a series of technical challenges. In the following sections, the focus is on the ones related to audio technologies.

2 HARDWARE AND SOFTWARE CHALLENGES

The first step for identifying technical challenges is to analyze how current XR and NMP technologies capture and deliver audio to networked musicians, and how existing metaverse systems process audio. Except for a few systems available to the general public, most of the exploration of the MM is composed of prototypes, applications designed for user studies, and proofs of concept.

2.1 Audio Hardware

At the *Self Reality* level of the MM lie the physical devices that enable musicians to experience it. While advancements in embedded and wearable hardware technology have propelled the capabilities of computing devices for XR to unprecedented heights, limitations persist. These limitations manifest in various forms, from processing power and memory constraints to the intricacies of audio input/output interfaces. As such, bridging the gap between the computational demands of immersive audio experiences and the capabilities of existing hardware represents a significant hurdle in the realization of the MM.

HMDs and goggles are the primary devices used to experience the immersive contents of the MM, mainly at the visual level. Various types of devices are available on the market, from professional to entry-level devices. Recent products have transitioned from tethered and wireless headsets (where computation is performed on a host computer) to standalone headsets (where the computation is performed directly on the device itself). Therefore, most of the innovations in this area have focused on reducing the weight and size of HMDs, advancing optics, display resolution, and improving the field of view. While these characteristics are crucial for delivering realistic experiences capable of creating a solid sense of presence and immersion [64], what about audio? As detailed below, this aspect appears to have been overlooked thus far.

For the MM, the hardware used to capture and render audio are as critical as the means used to experience virtual environments at the visual level. Most commercially available HMDs are equipped with integrated audio devices for both input and output. A summary of the audio input and output capabilities of some of the most popular headsets is provided in Fig. 3.

2.1.1 Audio Input Technologies

In current metaverse environments (not only musical), sound is primarily used as a means of communication between users through voice chats [67]. Voice chats represent the way social interaction unfolds in shared immersive environments. For this, the majority of HMDs are equipped with an array of omnidirectional microphones for capturing the voice of the user. These systems feature active noise cancellation and dynamic compression. These arrangements

	Head-Mounted Display		Audio Hardware	
	Product Name	Type	Input	Output
AR	Magic Leap 2	Standalone	2x Integrated Microphones	2x Integrated Stereo Speakers
	Hololens 2	Standalone	4x Integrated Microphones	2x Integrated Stereo Speakers Bluetooth Audio
	Nreal Light	Standalone	5x Integrated Microphones	2x Integrated Stereo Speakers
MR	Meta Quest Pro	Standalone Wired	3x Integrated Microphones	2x Integrated Stereo Speakers 3.5 mm Line Output
	Apple Vision Pro	Standalone	6x Integrated Microphones	Dual-driver audio pods Compatible with Apple AirPods
	Varjo XR-3	Wired	External Microphone Support with 3.5 mm Line Input	3.5 mm Line Output
VR	HTC Vive Pro 2	Wired	2x Integrated Microphones	Hi-Res Certified Headphones
	Meta Quest 2	Standalone Wired	3x Integrated Microphones	2x Integrated Stereo Speakers 3.5 mm Line Output
	Valve Index	Wired	2x Integrated Microphones	2x Near-field Speakers

Fig. 3. A summary of the audio input and output features of the most prominent and widely used HMDs available today in the XR market for AR, MR, and VR applications. The authors distinguish AR and MR devices by defining AR devices as those that use o*ptical see-through displays* and MR devices as those that use v*ideo pass-through displays*. For a comprehensive summary of the differences see [66].

are far from ideal when used for capturing sounds from musical instruments and singing voices. Different from speech-based interactions, musical interactions require a high degree of temporal and spectral resolution.

However, most headsets on the market have neither microphones dedicated to binaural sound capturing nor do they provide classic stereo microphone setups. Moreover, at the moment, external microphones are still not officially supported by any standalone headsets available on the market. Beyond voice, as far as audio input is concerned, there are no current HMDs that provide a dedicated audio input interface comprising analog-to-digital converters with Tip Ring Sleeve or External Line Return inputs and audio drivers for managing audio in real time for XR applications. While these issues can be circumvented with tethered HMDs, it is currently unfeasible to directly connect audio equipment, such as physical musical instruments, to standalone headsets, which are the most widespread and commercially available types of HMDs.

2.1.2 Audio Output Technologies

Audio in immersive systems is usually presented through the use of 2D or 3D audio. While 2D audio represents the prevalent way that music is experienced, 3D audio (or spatial audio) has been identified as a pivotal element for enhancing both the realism and the user's sense of presence [68]. Among modern spatial audio techniques, binaural audio stands as one of the most used in the context of the MM [69] because it is compatible with common audio playback devices such as headphones [70, 71].

In most HMDs, sound is delivered to users through directional transducers embedded in the headset, placed near the ears (e.g., the near-field extra-aural earphones of the Valve Index or the dual-driver audio pods in the Apple Vision Pro). Whereas this represents a practical solution to sound delivery in terms of usage for musicians, it might not represent the best option. Only a few HTC Vive models include actual headphones, although they cannot be compared with professional audio equipment. Notably, most of today's HMDs leave this issue to users by providing a standard 3.5-mm jack output to plug in their own headphones.

From this brief analysis of what is available on the market, it is possible to notice that audio is mainly considered as an ancillary component in XR hardware. While quality has improved over time, the currently available technologies are not ideal for musicians.

2.2 Software for Audio

Building robust audio processing and rendering systems that can seamlessly integrate with the MM requires sophisticated software architectures. Moreover, ensuring cross-platform compatibility and optimizing performance across diverse computing environments add layers of complexity to software development efforts.

When it comes to software, XR applications usually fall into two categories: *i*) native applications, which can be used on a host computer or downloaded directly in a standalone HMD, and *ii*) browser-based applications. This distinction already hints at differences in how audio is handled.

2.2.1 Audio in Native Applications

For native applications, the most common tools used for developing multiuser musical XR applications are game engines, such as Unity 3D and Unreal Engine [69]. In the context of XR, game engines are used as multimedia software for designing and implementing the visual and auditory components of an immersive application. However, they are primarily focused on the former compared with the latter. While both provide sophisticated audio systems (including audio mixers, built-in effects, filters, equalization, and spatial audio), only Unreal has a dedicated high-performance audio utility, named MetaSounds, which gives developers full control of the digital signal processing (DSP), that is especially useful for designing real-time and procedural audio synthesis [72, 73].

However, the audio processing components of these software applications are not comparable with the ones provided by professional digital audio workstations. More advanced and sophisticated sound processing can be achieved with the use of dedicated tool kits and authoring applications (e.g., WWise or FMOD) or third-party libraries for integrating binaural audio (i.e., Atmoky trueSpatial). Similarly, custom plug-ins can also be created using audio frameworks and languages such as Faust, ChucK, CSound, LibPD, or RNBO of Cycling'74 Max. Nonetheless, doing so will require specialized work to ensure compatibility through different hardware and Operating Systems for XR applications such as Windows, Android, or visionOS.

2.2.2 Audio in Web Applications

For browser-based applications, audio is mainly managed by the WebAudio API (WAA) for DSP, real-time audio synthesis, and rendering. The WAA is a standard developed by the World Wide Web Consortium and is supported by most web browsers across various devices, including those available on standalone HMDs (e.g., Meta Quest Browser, Wolvic, Safari). The WAA also provides direct access to audio input devices when available (i.e., microphones), basic audio spatializers, and several libraries and tools have been developed and explored for this purpose [74–76].

Over the years, a variety of web-based tools and frameworks have emerged, building upon the functionalities of the WAA (e.g., Tone.js, Howler.js). These tools have already been used in MM applications leveraging WebXR [21] and have been further enhanced with audio tools specifically designed for browser-based, multiuser immersive applications such as PdXR [22, 24].

Another approach that has been recently explored in the MM is the use of Web Audio Modules. They allow the creation of articulated and high-quality synthesizers in multiuser web-based virtual experiences [25]. Web Audio Modules are units designed for high-level synthesis and processing of audio signals in the browser. They can be considered an equivalent of dedicated audio plug-ins used in digital audio workstations but designed to be compatible with web browsers and web audio [77, 78]. While these approaches represent a promising direction for implementing the MM on the web, it is currently unclear how web browsers, particularly those on commercial HMDs, perform in terms of audio processing.

3 THE LATENCY ISSUE

As prior research has highlighted, creating a functional metaverse requires some foundational prerequisites: low latency and stable jitter in communication networks and systems for seamless audio and visual synchronization [79, 80]. Communication latency refers to the delay between a control action and its reception at the receiver side, while jitter denotes the variability of such delay.

Achieving low-latency, jitter-free audio transmission in networked immersive environments presents a large set of challenges [81]. The inherent limitations of network communication combined with the complexities of real-time audio capture, processing, and rendering, exacerbate latency and synchronization issues. Moreover, fluctuations in network conditions introduce jitter, further complicating the task of maintaining temporal coherence across distributed audio streams. While latency and jitter have been extensively studied and analyzed in the fields of NMP [6] and social XR [35, 48], it has not been systematically addressed in MM systems. Ensuring low communication latency and constant jitter is a staple for very tightly coupled and highly sensitive systems such as the ones oriented toward real-time interactive music.

Communication latency is a critical concern in interactive multimedia systems because can negatively affect users' sense of control and overall Quality of Service. Especially in immersive environments, latency can disrupt the sense of presence [82, 83]. Additionally, latency has a negative impact on task completion time and adverse effects on user social interactions such as mutual understanding between collaborators and perceived workload [84].

Substantial research has consistently demonstrated that when musicians play in the same location (e.g., a room for rehearsals or a concert hall), they can tolerate a latency of up to 25–30 ms (which corresponds to a distance among musicians of 8–10 m). These results become significant in the context of NMP systems [85, 86]. Beyond these thresholds, precise synchronization and rhythmic cohesion, especially in group-based performances, become unattainable. While latency in a physical space is dictated by the speed of sound in the air, in a networked system is based on a much more complex chain of processes needed for capturing, analyzing, and transmitting audio data (for details, the reader is referred to [6]).

Latency can accumulate in various parts of a system. These include the process of analog-to-digital conversion, the buffering of audio-capturing and audio-rendering devices, and the digital-to-analog conversion. Moreover, DSP algorithms can add significant processing latency. In terms of acquisition, audio interfaces typically introduce a delay that depends on their settings of the audio block size [87]. In terms of rendering, the use of algorithmic reverberation

in spatial audio processes also sums to the overall latency [49, 88, 89].

Moreover, the transmission between *Self* and *Distant Reality* and the characteristics of the network (such as its speed and bandwidth) can introduce significant delays. Especially across vast distances, the conditions of the network may also cause high amounts of jitter, which makes it difficult to predict whether certain packets might be delivered or not, resulting in noticeable audio dropouts. A way to cope with this issue is to buffer a certain amount of packets (i.e., using the so-called "jitter buffers"). However, adjusting the buffer size involves trade-offs between reducing jitter, maintaining synchronization, and managing latency: a higher buffer size results in higher latency [6]. Even in ideal conditions, the speed of light represents the ultimate physical limit to the transmission of content over the internet: the network latency adds at least one ms per 300 km of the signal path [90].

Due to these technical limitations, not only does long-distance music performance become challenging, it is also difficult to create conditions to keep latency under the perceptual threshold of 30 ms and maintain the jitter minimally and constant. However, it should be noted that these findings came from systems that concentrated mainly on audio rather than full multimedia experiences. The MM will have other elements that go significantly beyond the sole transmission of sound, including its integration with immersive, multisensory, and interactive technologies.

In the metaverse, creating successful immersive experiences requires virtual sensory inputs that match human expectations. Latency can severely disrupt this experience. The sources of latency occur both in local processing (*Self Reality*) and in remote connections (*Distant Reality*).

3.1 Self Reality Latency

In the MM interaction model, *Self Reality* exhibits several delays that characterize the different elements used to present virtual objects and environments to the user and the means for interacting with them. The first type of delay is referred to as *motion-to-photon*, which is the temporal interval between an input (e.g., a movement performed by the user while rotating the head) and the moment when a full frame of pixels reflecting the associated changes is displayed. This latency encompasses all delays caused by processes involved in rendering the viewport, such as HMD tracking, signal processing, logic updates, and display refresh cycles.

Several studies have shown that to avoid this side effect, the latency between head movements and viewport rendering should not exceed 50 ms [91–93]. However, other studies have established a lower threshold of 20–25 ms [94].

While this is true for VR systems, in MR and AR devices, this type of latency is often referred to as *photon-to-photon*. This represents the time taken for a photon from the real world to hit a camera and a photon from the display to reach the user's eyes. To date, current AR/MR HMDs show a *photon-to-photon* latency between 11 and 40 ms, depending on the system used (pass-through or see-through).[2]

The *motion-to-photon* and *photon-to-photon* latencies are metrics of particular importance because temporal delays can degrade sensorimotor performances, which are fundamental in musical interactions. High latency causes virtual environments to become decoupled from the user's motion and orientation. This not only disrupts the immersive experience but can also lead to motion sickness [95].

Another type of latency that affects perception is *frame delay*, which refers to the time it takes for the Graphics Processing Unit to process a frame and send it to the display. Ideally, frame latency should be half of the motion-to-photon latency [63].

Moreover, in musical applications, another element that is important to consider is *action-to-sound latency* [96], such as the delay occurring between an action performed by the user and its impact at the auditory level on the sound delivered by headphones or loudspeakers. The close coupling of action and sound has been recognized as being of prime importance for building effective interactive digital musical systems. Wessel and Wright described this close coupling [97] as fundamental for maintaining a valid causal link between action and sound that contributes to an embodied engagement with an instrument [98].

While Wessel and Wright suggested that musical systems should aim at latencies of at most 10 ms, McPherson et al. [99] showed that this standard is still not met by many hardware and software systems commonly used by researchers and artists. While there are studies and proposed solutions for reducing *motion-to-photon* latency in the context of XR, knowledge about *action-to-sound* latency is limited to a few studies mostly focusing on VR and fatigue (e.g., [100–102]).

Immersive environments, such as those in the MM, should also support speech communication. The delay between when a sound is uttered and when the listener hears it is known as *mouth-to-ear* latency. Low *mouth-to-ear* latency is essential for maintaining a natural flow of conversation and shares many similarities with musical signals. Ideally, this latency should be less than 150 ms [103], which is nearly five times higher than the threshold for group music.

Another element contributing to latency is spatial audio. Previous research has shown that spatial audio algorithms (including encoding, room simulation, sound scene rotation, and decoding) can increase system latency [88]. This latency must be carefully managed in networked musical applications [49, 90, 104, 105, 106].

3.2 Distant Reality Latency

In the MM, actions and the sounds produced by a musician should be perceived in real time by other connected users, both visually and aurally. In the metaverse, latency is influenced by various factors, as highlighted in recent

[2] Apple Vision Pro Benchmark Test 1, https://www.optofidelity.com/insights/blogs/apple-vision-pro-benchmark-test-1-see-through-latency-photon-to-photon.

studies that examined the *end-to-end latency* of platforms such as Rec Room, VRChat, and Horizon Worlds [48, 107]. These studies found that latency can vary across platforms, depending on factors such as the distance between clients and servers.

Because the metaverse is a social space, each environment must support the presence of a potentially high number of concurrent users interacting together. Studies have shown that latency and throughput increase almost linearly with the number of connected users. For example, when two users are connected, Rec Room and VRChat exhibit an *end-to-end latency* of around 100 ms that increases to approximately 140 ms when up to seven users connect, because each client must receive data related to the newly connected users. In multiuser environments, latency also depends on server processing, with the processing latency on the receiver typically being much higher than on the sender. Furthermore, latency increases when virtual worlds are rendered locally on the headset, compared with systems using remote rendering. Although these results do not specifically pertain to musical environments, they clearly illustrate the impact of latency on applications in general.

Whereas the measurement of the latency exhibited at the different components of *Self Reality* has been conducted multiple times, measuring the latency of *Distant Reality* is more challenging, and only a handful of methods have been proposed [108]. These methods require specialized equipment that should be not only multiplied depending on the number of users to be measured but also synchronized. While there have been some attempts, these are limited to local networks and somewhat simplified environments [27, 28, 109]. Therefore, these measurements as can only be taken as indicative. However, it is worth noticing that the acceptable thresholds reported in such studies are at least three times higher than the ones required for collaborative music making.

While latency and jitter depend on a variety of technical and structural factors, it was also demonstrated that, in the musical context, they can also depend on the tempo of a musical piece, the instruments used, the proficiency of musicians, the number of performers, and the expectations of the users [110, 111]. An examination of the latency-related challenges affecting musical activities in the metaverse revealed several bottlenecks, primarily impacting its users. Such bottlenecks involve the processing capabilities of HMDs, hardware and software used for recognizing gestures performed by the users, and protocols used to encode and decode audio.

Regardless of their source, the latency constraints for musical interactions over the network constitute hard limits. These constraints must be carefully considered when developing consumer electronics tailored for the MM.

3.3 Data Streaming

Latency also depends on the type of data exchanged between connected users. An analysis of the literature reported in [69] has revealed two main approaches used in multiuser XR musical systems. They share similarities with the two use cases outlined in SEC. 1.2: one approach focuses on sending sound control data (such as the ones used for controlling XR musical instruments), while the other focuses on transmitting actual audio data generated by physical musical instruments between peers.

3.3.1 Sound Control Data

In this approach, audio is generated locally in response to a remote event, such as a gesture of a performer or a key event on a control surface, using sound generation techniques. Transmitting sound control data offers several advantages. First, sound control data require less bandwidth to be transmitted. Second, concealing packet loss is easier and can be handled in various ways. For example, a late note can be skipped by the system or, in some instances, can be predicted. Third, this approach allows musicians to exchange data encoded in different communication protocols such as MIDI or Open Sound Control. This approach was explored extensively in the context of NMP (e.g., [112–119]) and was one of the main approaches to be adapted for multiuser musical environments.

An example of this approach is LeMo [50]: the instruments are two sequencers controlled by a 3D UI. The two clients use the same application, which contains the same sound engine. When a client presses a button on the sequencer, it sends control data to its synthesizer. The same control data are then broadcast to the other client to synchronize the two events (see Fig. 4).

This has the advantage of being compatible with several commercial and custom-made controllers and synthesizers. Moreover, sound control data can be encoded in custom messaging systems such as WebRTC data channels, which provide a low-latency way of transmitting data through the internet [21, 22]. Additionally, this can be compatible with live-coding practices [56]. One of the problems of NMP is the difficulty of handling very consistent rhythms and music heavily based on beats and regular pulses. In this case, audio data need to be quantized. However, the use of sound control data in networked systems is constrained by the choice of specific sound generation algorithms and synthesizers. It requires managing several data streams at the same time, which might increase with the number of connected clients. Moreover, sound control data should run in parallel with all of the other control data used to synchronize non-musicals actions and events.

3.3.2 Audio Data

The other approach is based on the actual transmission of audio over Internet Protocol–based networks. Although many metaverse platforms allow for streaming speech data, they currently lack the necessary audio quality and low latency required for musicians to play together.

Over the last twenty years, the use of audio data has been extensively explored by the NMP community. In contrast to the approach focused on sound control data, research efforts have led to significant advancements in this area. These efforts have resulted in the development of several

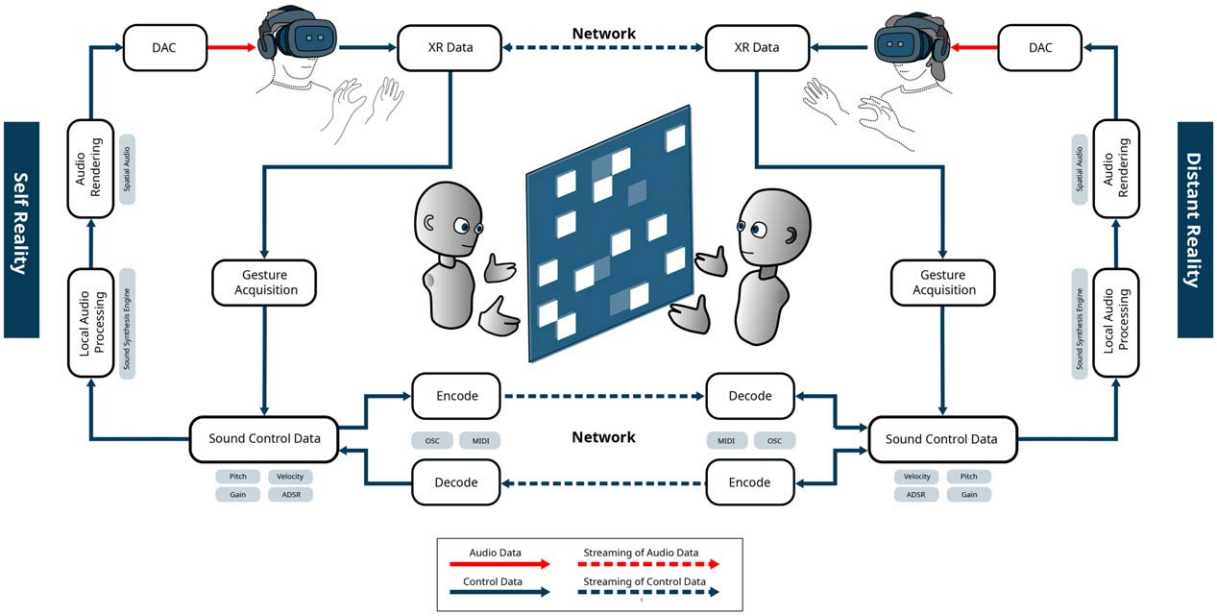


Fig. 4. A breakdown of the main components that contribute to latency in a system that uses sound control data to create a VR musical experience in the metaverse. The example is based on the case illustrated in Fig. 2(a).

commercial and open-source software and hardware systems.

These include applications for stand-alone devices [120] ranging from specific implementations of protocols, such as WebRTC [121], to peer-to-peer solutions, such as 4D Jam Connect[3] and SonoBus,[4] to client/server applications, such as Jamulus.[5] Platforms such as LOLA [90], UltraGrid,[6] and JamKazam,[7] allow the streaming of concurrent live video, or even the integration of videoconferencing tools such as Zoom if properly configured. Widely used tools like JackTrip [122] can be seamlessly operated on both desktop computers and single-board devices, such as the Raspberry Pi. Conversely, the company Elk has released a companion hardware interface for their Elk Live[8] system, which is based on an audio-specific low-latency operating system [123] and allows musicians to better manage latency and jitter locally.

Dante[9] is another widely used system that transmits uncompressed, low-latency audio. Industry standards like AES67 [124] and SMPTE ST 2110,[10] were developed for applications where precise synchronization is essential. These standards were originally designed for media broadcasting, and are mostly used in sports and theater.

An example of this approach is shown in Fig. 5. Here, musicians play together by sending and receiving audio data through a NMP system (i.e., JackTrip, Elk Live) and see each other as volumetric videos captured by one or more stereoscopic and depth cameras [26, 28].

While this approach offers a promising way to connect musicians playing acoustic, electric, or digital instruments, it also presents significant technical challenges. It requires careful integration of hardware, software, and networking, including the streaming of avatar control data or point cloud data, all while ensuring low-latency and high-quality audio.

The applicability of such solutions within the metaverse environment remains a subject of inquiry. Nonetheless, the evolution of a music-oriented metaverse is intricately linked to the availability of high-quality and low-latency network connections. To the authors' knowledge, there are no comprehensive and consistent studies on the impact of the different types of latency in musical immersive multiuser systems beyond just a few performers.

While very few have attempted to integrate current NMP technologies (e.g., JackTrip, Elk Live) with XR systems [28, 125], such integration is not straightforward. There is a lack of systems, both software and hardware, capable of handling both sound control data and audio signal streaming.

Finally, network congestion and limited bandwidth can introduce delays and jitter. Techniques for latency mitigation are discussed in SEC. 3.3.4.

3.3.3 Audio-Visual Data Synchronization

Different from existing NMP technologies, an MM will require musicians to see each other in 3D, either in VR or integrated into their environment with Mixed Reality. Different from purely audio NMP systems, a crucial element

[3] 4D Jam Connect, https://www.4dcreatives.ca/jam.
[4] SonoBus, https://www.sonobus.net/.
[5] Jamulus, https://jamulus.io/.
[6] UltraGrid, https://www.ultragrid.cz/.
[7] JamKazam, https://jamkazam.com/.
[8] Elk Live Bridge, https://www.elk.live/the-bridge.
[9] Dante, https://www.getdante.com/.
[10] ST 2110 Suite of Standards, https://www.smpte.org/standards/st2110.

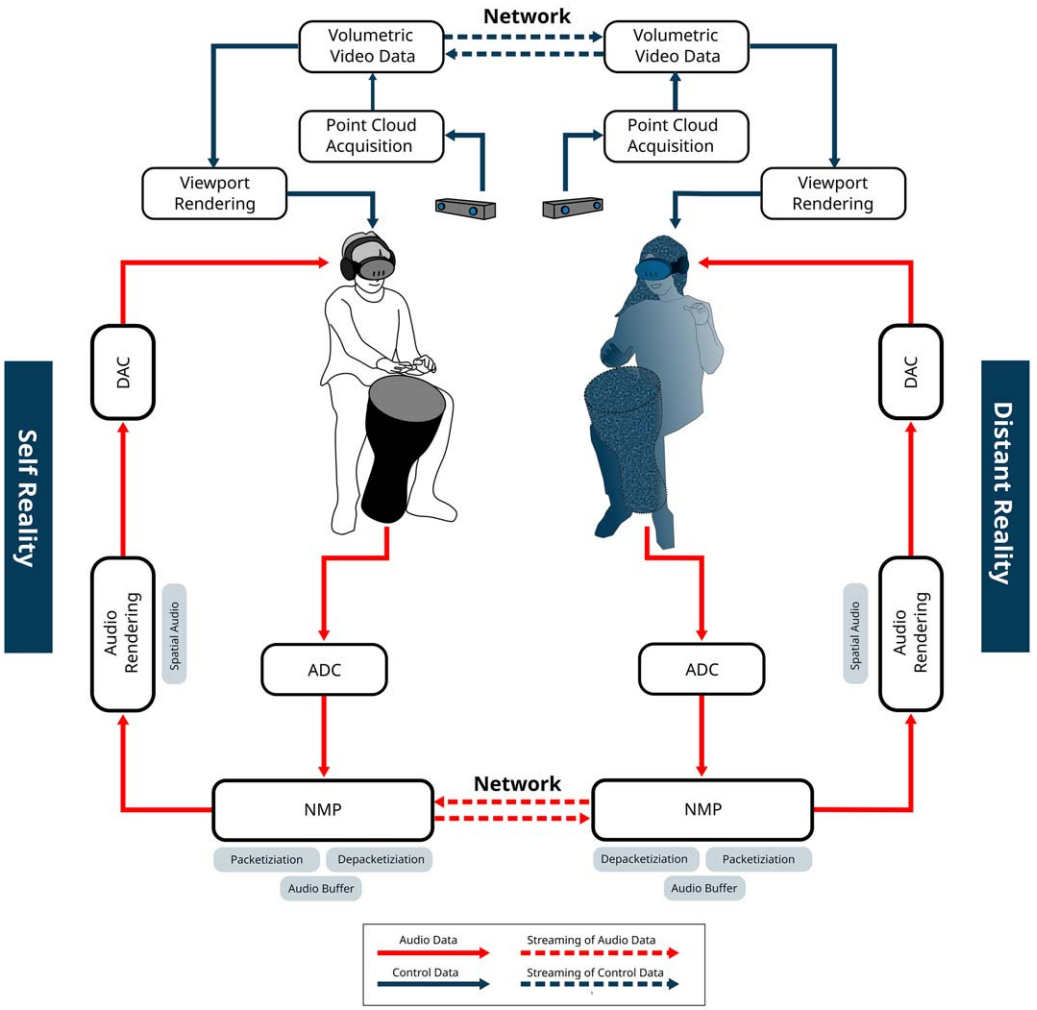


Fig. 5. An analysis of the key factors causing latency in systems that use sound control data to craft MR musical experiences in the metaverse. This example is based on the case illustrated in Fig. 2(b).

is represented by the synchronization between the auditory and visual representation of the connected musicians. Previous studies have shown that mismatches between audio and video can be tolerated up to 250 ms [126].

Recent studies have investigated the impact of audio-visual latency in the context of XR-based NMP, and two main ways to present musicians were explored: through 3D-embodied avatars or real-time 6 Degrees of Freedom volumetric videos. The case of avatars was explored in the study by Cairns et al. [27], which focused on the effect of latency on synchronization. They tested a four-piece band (bass, keyboard, guitar, drum) playing together with avatar control data and audio streamed in a Local Area Network with three different audio latencies (at 19, 24, and 29 ms). They found no significant differences in the perception of the latency between the seen gestures produced by a musician and the corresponding heard sound.

In a similar study, Hunt et al. [52] focused on how avatars are perceived by a group of four musicians (bass, voice, guitar, drum) using a low-latency NMP system ($<$15 ms). They recorded mixed responses from the different musicians regarding the audio and visual latency. At the same time, the drummer relied on visual cues for keeping the tempo, and the bass player reported a lack of them, especially for nonverbal communication.

When musicians are represented as 3D point clouds, the user experience changes significantly. A study tested the feasibility of capturing, streaming, and rendering real-time audio and point cloud data [28]. The results revealed that the audio-visual mismatch was unsustainable for two professional musicians, with a video latency of approximately 400 ms.

However, all of these studies can be considered preliminary because they suffer from a few limitations, such as being conducted within local networks and with two or four musicians connected at the same time. Ensuring precise synchronization of audio data across distributed nodes is critical for maintaining musical cohesion, especially when timing and rhythm are important. Managing synchronization becomes increasingly challenging as the number of participants and geographic distribution of nodes increases.

Taken together, these findings suggest that synchronizing avatars use fewer resources compared with point cloud capturing, resulting in a reduced system and perceived latency;

this is why most multiuser-networked XR systems use 3D-embodied avatars. Volumetric point cloud videos can be beneficial for fostering more natural interactions among remote users [127, 26]. However, previous works have shown that current dedicated systems for capturing, delivering, and rendering point cloud data in real time also show high latency [128, 127] that, even if compatible with thresholds found for 2D video conferencing (i.e., below 200–600 ms) [129], makes them incompatible with the requirements for musical interactions over the network.

3.3.4 Taming the Latency

While NMP systems favor wired Local Area Networks, the metaverse and its musical side will be predominantly experienced via standalone HMDs and goggles equipped with wireless connectivity. The development of 5G and 6G networks plays a pivotal role in the advancement of the metaverse implementation [130–132]. Recent works have shown the feasibility of streaming low-latency audio over 5G [133–136]; however, at the moment there are no specific solutions for multiuser immersive systems [137–139].

It can be hypothesized that coupled with current advancements in edge and cloud computing, an effective metaverse for music should be capable of offloading some of the processing to different cloud services and potentially creating more scalable architectures and low-latency applications. Other solutions might be considered, such as the use of dedicated hardware (e.g., the Elk Live Bridge or the JackTrip Analog Bridge) to offload input and output processes from XR devices that will prioritize visual rendering. However, these avenues have not yet been explored (most existing works have not included musical or Audio-First applications). Still, they represent a more stable solution, especially for mobile networks [140–142]. Furthermore, several approaches have been proposed for integrating spatial audio systems, such as offloading the encoding and decoding processes to edge computing within a 5G network [90, 107, 143].

Several other techniques for latency mitigation have been explored in the NMP community. These include client-side prediction, server reconciliation, and interpolation to ensure smooth and responsive interactions among users [6]. An approach that received particular attention but currently is not explored in the context of the MM employs the mechanism of prediction and anticipation of either performers' gestures [144–146] or musical events [147–149].

It is also necessary to mention other solutions that do not depend on data transmission but have been explored previously, although not yet in the context of the MM. When the first NMPs appeared, several musicians thought of incorporating latency as an inherent characteristic of remote performances [150]. This resulted in several attempts to explore alternative approaches to the realistic jam that focus on experimentation. An example is Ninjam [117], where musicians play asynchronously to the music, receiving the sound delayed of one measure. In his Quintnet.net, Hajdu introduces the idea of a conductor that can control and direct the musical outcome of the connected performers [117].

NMP researchers have also explored the use of visual cues (from video streams) and alternative feedback mechanisms to mitigate the effects of latency. This includes the use of "adaptive metronomes," which are metronomes capable of tracking musicians' tempo and adapting it to the latency variability caused by the network [151]. Similar approaches can be used to accelerate the advancements of more stable and realistic musical interactions in the MM.

Finally, emergence of novel types of frameworks and tools such as CoreLink [125] and Holodeck [152] must be acknowledged. They represent some promising approaches for the transmission and synchronization of various data (audio, video, motion capture) between different nodes in a network. Moreover, Grimm et al. [153] have recently proposed a system for the synchronization of audio data with electroencephalography and other biophysical sensor data for creating multimodal networked experiences, which can also be beneficial for the MM.

4 INTEROPERABILITY AND STANDARDIZATION

The Musical Metaverse will likely consist of interconnected and shared environments containing both virtual objects and physical objects captured digitally. Moreover, the MM should accommodate the use of different hardware, including both HMDs and musical instruments, whether physical or virtual.

Therefore, interoperability becomes a crucial element for the success of the MM, enabling seamless integration and interaction across diverse audio technologies and platforms. The absence of unified industry standards for audio technology in the MM is a significant barrier to interoperability. Different platforms may adopt their own standards, leading to a lack of cohesion and the fragmentation of the field. Establishing industry-wide standards for audio formats, spatial audio protocols, and real-time processing is essential for facilitating interoperability. The following points need to be addressed in future standardization activities.

- **Real-time audio processing and synchronization:** For live performances in the MM, audio streams need to be processed and synchronized with minimal latency. For the technologies that must be used to experience the MM (i.e., HMDs, Operating Systems, web browsers), one of the fundamental requirements is the interoperability between various real-time audio processing systems and protocols. Differences in how platforms such as standalone and browser-based applications or differences between an Android- based operating system (i.e., Meta Quest, Pico) and iOS (i.e., Apple Vision Pro) handle latency, buffering, and synchronization can result in audio lag and timing issues that disrupt the real-time collaborative experience. At the moment, the only existing standard for audio that can be adopted in the MM are Web Audio and Web MIDI. However, this limits immersive worlds to the

ones running on a web browser, and performance can vary between browsers and Operating Systems. This represents an open challenge that requires further investigation.

- **Integration of spatial audio:** The implementation of spatial audio can vary across platforms and also depends on techniques and methods used for encoding, rendering, and delivering 3D sound. The lack of standardized spatial audio protocols complicates the integration of audio experiences in metaverse environments, leading to inconsistent user experiences and potential technical conflicts. While some attempts have been proposed and documented [154, 155], a proper standardization and interoperability are still a distant goal. The only viable standard for spatial audio that was explored in the context of the MM is the one available through the Web Audio API [21, 25, 77].
- **Audio formats and codecs:** To stream, encode, decode, and process audio data, the MM encompasses a wide range of audio formats, each designed for specific purposes and offering different levels of quality and compression. Ensuring that these diverse formats can interact seamlessly is a major technical challenge. The lack of standardized audio formats can lead to compatibility issues, where audio content created on one platform may not be playable or may lose quality when transferred to another platform. However, codecs such as MPEG-I [156, 157] and 3GPP IVAS [158, 159] are emerging as viable solutions for supporting high-quality, low-latency immersive audio across various networked applications. They represent a promising step toward interoperability and standardization for the MM.

5 CONCLUSION

In this article, the leading audio technologies needed for the creation of a functional MM were described. Next, significant issues and challenges in consumer electronics were identified as essential considerations for the realization of a Metaverse that supports musical activities. Through this analysis of state-of-the-art hardware and software used to stream and experience audio in shared virtual environments, it was found that audio technologies play a secondary role in current metaverse environments. Whereas human speech is well supported in XR hardware and software, playing music in real time entails requirements that are far more demanding than those of speech-based interactions.

By looking at how audio is streamed and rendered in current metaverse environments, it was shown how latency represents the factor that more than others hinders the establishment of a metaverse capable of supporting musical activities. From the algorithms used for audio spatialization to the devices used for acquiring and tracking human gestures and the type of Internet connection used, every process involved adds its own latency. Together, all these aspects entail very hard requirements for immersive performances, especially when considering many concurrent users.

However, this paper shows how recent work at the crossroads of NMP and XR have started to tackle these issues, especially the latency of audio and visual components. While this analysis showed that very little research has been conducted in this area, especially on wireless and edge computing systems, this represents an opportunity for future research. Research on perceptual limits, transmission protocols, packet-loss prediction and correction, and strategies for synchronization of audio and visual XR will not only enable the possibility of creating a more musically oriented metaverse but will also significantly impact the development of more user-friendly and realistic support of human activities in massive shared virtual environments in general.

Notably, an element that this study did not investigate in detail is the social aspects and user needs of musicians and audiences. This aspect is poorly understood in the context of music and started only recently to be actively explored [160]. This calls for further research.

Creating a functional MM also opens challenges that have not been clearly addressed, even in the broad research on the metaverse. How will interoperable, social, and persistent immersive worlds manage hundreds of concurrent users occupied both in listening and making music in real time? How will resources regarding computation and networking be managed? Such issues go beyond the focus of this article; however, they highlight that the different disciplines (both scientific and artistic) required to tackle them are varied and require collaboration. It is hoped that this paper will inspire researchers in both academia and industry to address the many challenges and issues identified, enabling the MM to flourish and offer radically novel art forms and musical experiences to different stakeholders.

6 ACKNOWLEDGMENT

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European European (grant agreement n. 101184379). 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. In addition, the authors acknowledge the support of the MUR PNRR PRIN 2022 Grant, prot. num. 2022CZWWKP, funded by Next Generation EU.

7 REFERENCES

[1] B. Loveridge, "An Overview of Immersive Virtual Reality Music Experiences in Online Platforms," *JONMA*, vol. 5, no. 1, p. 5 (2023).

[2] J. Park, Y. Choi, and K. M. Lee, "Research Trends in Virtual Reality Music Concert Technology: A Systematic Literature Review," *IEEE Trans. Visual. Comput.*

Graphics, vol. 30, no. 5, pp. 2195–2205 (2024 May). https://doi.org/10.1109/TVCG.2024.3372069.

[3] L. Turchet, "Musical Metaverse: Vision, Opportunities, and Challenges," *Pers. Ubiquitous Comput.*, vol. 27, pp. 1811–1827 (2023 Oct.). https://doi.org/10.1007/s00779-023-01708-1.

[4] L. Turchet, R. Hamilton, and A. Çamci, "Music in Extended Realities," *IEEE Access*, vol. 9, pp. 15810–15832 (2021 Jan.). https://doi.org/10.1109/ACCESS.2021.3052931.

[5] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of Musical Things: Vision and Challenges," *IEEE Access*, vol. 6, pp. 61994–62017 (2018 Sep.). https://doi.org/10.1109/ACCESS.2018.2872625.

[6] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823–8843 (2016 Dec.). https://doi.org/10.1109/ACCESS.2016.2628440.

[7] L. Gabrielli and S. Squartini, *Wireless Networked Music Performance*, SpringerBriefs in Electrical and Computer Engineering, (Springer Science+Business Media, Singapore, Singapore, 2016). https://doi.org/10.1007/978-981-10-0335-6.

[8] L. Turchet and C. Rottondi, "On the Relation Between the Fields of Networked Music Performances, Ubiquitous Music, and Internet of Musical Things," *Pers. Ubiquitous Comput.*, vol. 27, no. 5, pp. 1783–1792 (2023 Oct.). https://doi.org/10.1007/s00779-022-01691-z.

[9] K. E. Onderdijk, L. Bouckaert, E. Van Dyck, and P.-J. Maes, "Concert Experiences in Virtual Reality Environments," *Virtual Real.*, vol. 27, no. 3, pp. 2383–2396 (2023 Jun.). https://doi.org/10.1007/s10055-023-00814-y.

[10] M. Slater, C. Cabriera, G. Senel, et al., "The Sentiment of a Virtual Rock Concert," *Virtual Real.*, vol. 27, no. 2, pp. 651–675 (2023 Aug.). https://doi.org/10.1007/s10055-022-00685-9.

[11] S. Ppali, V. Lalioti, B. Branch, et al., "Keep the VRhythm Going: A Musician-Centred Study Investigating How Virtual Reality Can Support Creative Musical Practice," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–19 (New Orleans, LA) (2022 Apr./May). https://doi.org/10.1145/3491102.3501922.

[12] A. M. Webb, C. Wang, A. Kerne, and P. Cesar, "Distributed Liveness: Understanding How New Technologies Transform Performance Experiences," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 432–437 (San Francisco, CA) (2016 Feb./Mar.). https://doi.org/10.1145/2818048.2819974.

[13] H. Yakura and M. Goto, "Enhancing Participation Experience in VR Live Concerts By Improving Motions of Virtual Audience Avatars," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 555–565 (Porto de Galinhas, Brazil) (2020 Nov.). https://doi.org/10.1109/ISMAR50242.2020.00083.

[14] A. Munoz-Gonzalez, S. Kobayashi, and R. Horie, "A Multiplayer VR Live Concert With Information Exchange Through Feedback Modulated by EEG Signals," *IEEE Trans. Hum. Mach. Syst.*, vol. 52, no. 2, pp. 248–255 (2022 Jan.). https://doi.org/10.1109/THMS.2021.3134555.

[15] M. Abe, T. Akiyoshi, I. Butaslac, Z. Hangyu, and T. Sawabe, "Hype Live: Biometric-Based Sensory Feedback for Improving the Sense of Unity in VR Live Performance," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 836–837 (Christchurch, New Zealand) (2022 Mar.). https://doi.org/10.1109/VRW55335.2022.00269.

[16] O. Capra, F. Berthaut, and L. Grisoni, "A Taxonomy of Spectator Experience Augmentation Techniques," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 327–330 (Brimingham, UK) (2020 Jul.). https://doi.org/10.5281/zenodo.4813396.

[17] B. Loveridge, "Networked Music Performance in Virtual Reality: Current Perspectives," *JONMA*, vol. 2, no. 1, p. 2 (2020).

[18] V. Zappi, F. Berthaut, and D. Mazzanti, "From the Lab to the Stage: Practical Considerations on Designing Performances with Immersive Virtual Musical Instruments," in M. Geronazzo and S. Serafin (Eds.), *Sonic Interactions in Virtual Environments*, Human–Computer Interaction Series, pp. 383–424 (Springer, Cham, Swizterland, 2023). https://doi.org/10.1007/978-3-031-04021-4_13.

[19] P. Cairns, A. Hunt, J. Cooper, et al., "Recording Music in the Metaverse: A Case Study of XR BBC Maida Vale Recording Studios," *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality* (2022 Aug.), paper 12.

[20] R. Hupke, S. Preihs, and J. Peissig, "Immersive Room Extension Environment for Networked Music Performance," presented at the *153rd Convention of the Audio Engineering Society* (2022 Oct.), paper 28.

[21] A. Boem and L. Turchet, "Musical Metaverse Playgrounds: Exploring the Design of Shared Virtual Sonic Experiences on Web Browsers," in *Proceedings of the 4th International Symposium on the Internet of Sounds*, pp. 1–9 (Pisa, Italy) (2023 Oct.). https://doi.org/10.1109/IEEECONF59510.2023.10335297.

[22] D. Dziwis, H. von Coler, and C. Pörschmann, "Orchestra: A Toolbox for Live Music Performances in a Web-Based Metaverse," *J. Audio Eng. Soc.*, vol. 71, no. 11, pp. 802–812 (2023 Nov.). https://doi.org/10.17743/jaes.2022.0096.

[23] J. Bell, "Networked Music Performance in PatchXR and FluCoMa," in *Proceedings of the International Computer Music Conference (ICMC)* (Shenzhen, China) (2023 Oct.).

[24] 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 *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–10 (Erlangen, Germany) (2024 Sep./Oct.). https://doi.org/10.1109/IS262782.2024.10704079.

[25] M. Buffa, A. Hofr, and D. Girard, "Using Web Audio Modules for Immersive Audio Collaboration in

the Musical Metaverse," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–10 (Erlangen, Germany) (2024 Sep./Oct.). https://doi.org/10.1109/IS262782.2024.10704108.

[26] R. Schlagowski, D. Nazarenko, Y. Can, et al., "Wish You Were Here: Mental and Physiological Effects of Remote Music Collaboration in Mixed Reality," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (Hamburg, Germany) (2023 Apr.). https://doi.org/10.1145/3544548.3581162.

[27] P. Cairns, A. Hunt, D. Johnston, et al., "Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios," *J. Audio Eng. Soc.*, vol. 71, no. 6, pp. 313–325 (2023) Jun.). https://doi.org/10.17743/jaes.2022.0086.

[28] L. Turchet, N. Garau, and N. Conci, "Networked Musical XR: Where's the Limit? A Preliminary Investigation on the Joint Use of Point Clouds and Low-Latency Audio Communication," in *Proceedings of the 17th International Audio Mostly Conference*, pp. 226–230 (St. Polten, Austria) (2022 Sep.). https://doi.org/10.1145/3561212.3561237.

[29] 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," *J. New Music Res.*, vol. 52, no. 4, pp. 316–336 (2024 Apr.). https://doi.org/10.1080/09298215.2024.2339244.

[30] R. Hupke, S. Preihs, and J. Peissig, "Immersive Networked Music Performance: Impact of Extended Reality on the Quality of Experience," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–9 (Erlangen, Germany) (2024 Sep./Oct.). https://doi.org/10.1109/IS262782.2024.10704092.

[31] L. Bruns, B. Saurbier, T. M. Voong, and M. Oehler, "Presence and Flow in Virtual and Mixed Realities for Music-Related Educational Settings," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–7 (Erlangen, Germany) (2024 Sep./Oct.). https://doi.org/10.1109/IS262782.2024.10704115.

[32] H. Dong and Y. Liu, "Metaverse Meets Consumer Electronics," *IEEE Consum. Electron. Mag.*, vol. 12, no. 3, pp. 17–19 (2023 May). https://doi.org/10.1109/MCE.2022.3229180.

[33] A. Çamci and R. Hamilton, "Audio-First VR: New Perspectives on Musical Experiences in Virtual Environments," *J. New Music Res.*, vol. 49, no. 1, pp. 1–7 (2020 Jan.). https://doi.org/10.1080/09298215.2019.1707234.

[34] R. C. Waters and J. W. Barrus, "The Rise of Shared Virtual Environments," *IEEE Spectr.*, vol. 34, no. 3, pp. 20–25 (1997 Mar.). https://doi.org/10.1109/6.576004.

[35] C. Cortés, P. Pérez, and N. García, "Understanding Latency and QoE in Social XR," *IEEE Consum. Electron. Mag.*, vol. 13, no. 3, pp. 61–72 (2024 May). https://doi.org/10.1109/MCE.2023.3338130.

[36] S. Mann, Y. Yuan, F. Lamberti, A. El Saddik, R. Thawonmas, and F. G. Prattico, "eXtended meta-uni-omni-Verse (XV): Introduction, Taxonomy, and State-of-the-Art," *IEEE Consum. Electron. Mag.*, vol. 13, no. 3, pp. 27–35 (2024 May). https://doi.org/10.1109/MCE.2023.3283728.

[37] D. Friedman, A. Steed, and M. Slater, "Spatial Social Behavior in Second Life," in C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé (Eds.), *Intelligent Virtual Agents*, Lecture Notes in Computer Science, pp. 252–263 (Springer-Verlag, Berlin, Germany, 2007). https://doi.org/10.1007/978-3-540-74997-4_23.

[38] S. J. Friston, B. J. Congdon, D. Swapp, et al., "Ubiq: A System to Build Flexible Social Virtual Reality Experiences," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Osaka, Japan) (2021 Dec.), https://doi.org/10.1145/3489849.3489871.

[39] N. Stephenson, *Snow crash* (Penguin Boods, London, UK, 1994).

[40] S.-M. Park and Y.-G. Kim, "A Metaverse: Taxonomy, Components, Applications, and Open Challenges," *IEEE Access*, vol. 10, pp. 4209–4251 (2022 Jan.). https://doi.org/10.1109/ACCESS.2021.3140175.

[41] C. Koo, J. Kwon, N. Chung, and J. Kim, "Metaverse Tourism: Conceptual Framework and Research Propositions," *Curr. Issues Tourism*, vol. 26, no. 20, pp. 3268–3274 (2023 Oct.). https://doi.org/10.1080/13683500.2022.2122781.

[42] X. Zhang, Y. Chen, L. Hu, and Y. Wang, "The Metaverse in Education: Definition, Framework, Features, Potential Applications, Challenges, and Future Research Topics," *Front. Psychol.*, vol. 13, p. 1016300 (2022 Oct.). https://doi.org/10.3389/fpsyg.2022.1016300.

[43] R. Chengoden, N. Victor, T. Huynh-The, et al., "Metaverse for Healthcare: A Survey on Potential Applications, Challenges and Future Directions," *IEEE Access*, vol. 11, pp. 12765–12795 (2023 Feb.). https://doi.org/10.1109/ACCESS.2023.3241628.

[44] Y. Wang, Z. Su, N. Zhang, et al., "A Survey on Metaverse: Fundamentals, Security, and Privacy," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 1, pp. 319–352 (2023 Feb.). https://doi.org/10.1109/COMST.2022.3202047.

[45] A. Abilkaiyrkyzy, A. Elhagry, F. Laamarti, and A. Elsaddik, "Metaverse Key Requirements and Platforms Survey," *IEEE Access*, vol. 11, pp. 117765–117787 (2023 Oct.). https://doi.org/10.1109/ACCESS.2023.3325844.

[46] H. Yu, M. Shokrnezhad, T. Taleb, R. Li, and J. Song, "Toward 6G-Based Metaverse: Supporting Highly-Dynamic Deterministic Multiuser Extended Reality Services," *IEEE Network*, vol. 37, no. 4, pp. 30–38 (2023 Jul./Aug.). https://doi.org/10.1109/MNET.004.2300101.

[47] F. Tang, X. Chen, M. Zhao, and N. Kato, "The Roadmap of Communication and Networking in 6G for the Metaverse," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 72–81 (2023 Aug.). https://doi.org/10.1109/MWC.019.2100721.

[48] 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*, pp. 504–518 (New York, NY) (2022 Oct.). https://doi.org/10.1145/3517745.3561417.

[49] L. Turchet and M. Tomasetti, "Immersive Networked Music Performance Systems: Identifying Latency Factors," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–6 (Bologna, Italy) (2023 Sep.). https://doi.org/10.1109/I3DA57090.2023.10289169.

[50] L. Men and N. Bryan-Kinns, "LeMo: Supporting Collaborative Music Making in Virtual Reality," in *Proceedings of the IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, pp. 1–6 (Reutlingen, Germany) (2018 Mar.). https://doi.org/10.1109/SIVE.2018.8577094.

[51] R. Hamilton, "Trois Machins de la Grâce Aimante: A Virtual Reality String Quartet," in *Proceedings of the International Computer Music Conference*, pp. 202–206 (New York, NY) (2019 Jun.).

[52] A. Hunt, H. Daffern, and G. Kearney, "Avatar Representation in Extended Reality for Immersive Networked Music Performance," in *Proceedings of the AES International Conference on Spacial and Immersive Audio* (Huddersfield, UK) (2023 Aug.), paper 35.

[53] A. Renaud, A. Carôt, and P. Rebelo, "Networked Music Performance: State of the Art," in *Proceedings of the AES 30th International Conference* (Saariselkä, Finland) (2007 Mar.), paper 16.

[54] M. Tomasetti and L. Turchet, "Playing With Others Using Headphones: Musicians Prefer Binaural Audio With Head Tracking Over Stereo," *IEEE Trans. Hum. Mach. Syst.*, vol. 53, no. 3, pp. 501–511 (2023 Jun.). https://doi.org/10.1109/THMS.2023.3270703.

[55] A. MacLean, "Immersive Dreams: A Shared VR Experience," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 380–381 (Birmingham, UK) (2020 Jul.). https://doi.org/10.5281/zenodo.4813426.

[56] D. Dziwis, H. von Coler, and C. Porschmann, "Live Coding in the Metaverse," in *Proceedings of the 4th International Symposium on the Internet of Sounds*, pp. 1–8 (Pisa, Italy) (2023 Oct.). https://doi.org/10.1109/IEEECONF59510.2023.10335358.

[57] R. Hamilton, "Collaborative and Competitive Futures for Virtual Reality Music and Sound," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1510–1512 (Osaka, Japan) (2019 Mar.). https://doi.org/10.1109/VR.2019.8798166.

[58] R. Hamilton, J.-P. Caceres, C. Nanou, and C. Platz, "Multi-Modal Musical Environments for Mixed-Reality Performance," *J. Multimodal User Interfaces*, vol. 4, pp. 147–156 (2011 Nov.). https://doi.org/10.1007/s12193-011-0069-1.

[59] G. Martín, "Social and Psychological Impact of Musical Collective Creative Processes in Virtual Environments; The Avatar Orchestra Metaverse in Second Life," *Musica Tecnol.*, vol. 12, no. 1, pp. 75–87 (2018 Aug.). https://doi.org/10.13128/Music_Tec-23801.

[60] D. Dziwis and H. von Coler, "The Entanglement: Volumetric Music Performances in a Virtual Metaverse Environment," *JONMA*, vol. 5, no. 1, p. 3 (2023 May.).

[61] T. Hopkins, S. C. C. Weng, R. Vanukuru, et al., "AR Drum Circle: Real-Time Collaborative Drumming in AR," *Front. Virtual Real.*, vol. 3, p. 847284 (2022 Aug.). https://doi.org/10.3389/frvir.2022.847284.

[62] T. Di, D. Medeiros, M. Sousa, and T. Grossman, "VRChoir: Exploring Remote Choir Rehearsals via Virtual Reality," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 895–896 (Shanghai, China) (2023 Mar.). https://doi.org/10.1109/VRW58643.2023.00290.

[63] J. Jerald, *The VR Book: Human-Centered Design for Virtual Reality* (Morgan & Claypool Publishers, San Rafael, CA, 2015).

[64] L. Men and N. Bryan-Kinns, "LeMo: Supporting Collaborative Music Making in Virtual Reality," in *Proceedings of the IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, pp. 1–6 (Reutlingen, Germany) (2018 Mar.). https://doi.org/10.1109/SIVE.2018.8577094.

[65] T. Hopkins, S. C.-C. Weng, R. Vanukuru, et al., "Studying the Effects of Network Latency on Audio-Visual Perception During an AR Musical Task," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 26–34 (Singapore, Singapore) (2022 Oct.). https://doi.org/10.1109/ISMAR55827.2022.00016.

[66] J. Bailenson, B. Beams, J. Brown, et al., "Seeing the World Through Digital Prisms: Psychological Implications of Passthrough Video Usage in Mixed Reality," *TMB*, vol. 5, no. 2, pp. 1–17 (2024 Jun.). https://doi.org/10.1037/tmb0000129.

[67] A. M. Al-Ghaili, H. Kasim, N. M. Al-Hada, et al., "A Review of Metaverse's Definitions, Architecture, Applications, Challenges, Issues, Solutions, and Future Trends," *IEEE Access*, vol. 10, pp. 125835–125866 (2022 Nov.). https://doi.org/10.1109/ACCESS.2022.3225638.

[68] J. Paterson and H. Lee, *3D Audio* (Routledge, Abingdon, UK, 2021).

[69] A. Boem, M. Tomasetti, and L. Turchet, "Harmonizing the Musical Metaverse: Unveiling Needs, Tools, and Challenges From Experts' Point of View," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 206–214 (Utrecht, the Netherlands) (2024 Oct.). https://doi.org/10.5281/zenodo.13904834.

[70] T. Walton, "The Overall Listening Experience of Binaural Audio," in *Proceedings of the 4th International Conference on Spatial Audio* (Graz, Austria) (2017 Sep.).

[71] R. Gupta, J. He, R. Ranjan, et al., "Augmented/Mixed Reality Audio for Hearables: Sensing, Control, and Rendering," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 63–89 (2022 May.). https://doi.org/10.1109/MSP.2021.3110108.

[72] D. Menexopoulos, P. Pestana, and J. Reiss, "The State of the Art in Procedural Audio," *J. Audio Eng. Soc.*, vol. 71, no. 12, pp. 826–848 (2023 Dec.). https://doi.org/10.17743/jaes.2022.0108.

[73] Ç. Anil, "Modern Workflows for Procedural Audio at the Intersection of Gaming and Music Performance in Virtual Reality," in *Proceedings of the AES International*

Conference on Audio for Games (Tokyo, Japan) (2024 Apr.), paper 12.

[74] C. Çakmak and R. Hamilton, "od: Composing Spatial Multimedia for the Web," *J. Audio Eng. Soc.*, vol. 68, no. 10, pp. 747–755 (2020 Oct.). https://doi.org/10.17743/jaes.2020.0017.

[75] C. Çakmak and R. Hamilton, "Composing Spatial Music With Web Audio and WebVR," in *Proceedings of the Web Audio Conference*, pp. 1–5 (Trondheim, Norway) (2019 Dec.).

[76] M. Tomasetti, A. Boem, and L. Turchet, "How to Spatial Audio with the WebXR API: A Comparison of the Tools and Techniques for Creating Immersive Sonic Experiences on the Browser," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–9 (Bologna, Italy) (2023 Sep.). https://doi.org/10.1109/I3DA57090.2023.10289525.

[77] J. Kleimola and O. Larkin, "Web Audio Modules," in *Proceedings of the 12th Sound and Music Computing Conference*, pp. 249–256 (Maynooth, Ireland) (2015 Jul.).

[78] M. Buffa, S. Ren, T. Burns, A. Vidal-Mazuy, and S. Letz, "Evolution of the Web Audio Modules Ecosystem," in *Proceedings of the Web Audio Conference* (West Lafayette, IN) (2024 Mar.). https://doi.org/10.5281/zenodo.10825647.

[79] H. Dong and J. S. A. Lee, "The Metaverse From a Multimedia Communications Perspective," *IEEE MultiMedia*, vol. 29, no. 4, pp. 123–127 (2022 Oct.-Dec.) https://doi.org/10.1109/MMUL.2022.3217627.

[80] J. Santos, T. Wauters, B. Volckaert, and F. De Turck, "Towards Low-Latency Service Delivery in a Continuum of Virtual Resources: State-of-the-Art and Research Directions," *IEEE Commun Surv Tutor.*, vol. 23, no. 4, pp. 2557–2589 (2021 Nov.). https://doi.org/10.1109/COMST.2021.3095358.

[81] N. P. Lago and F. Kon, "The Quest for Low Latency," in *Proceedings of the International Computer Music Conference*, pp. 33–36 (Miami, FL) (2004 Nov.).

[82] I. S. MacKenzie and C. Ware, "Lag as a Determinant of Human Performance in Interactive Systems," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, pp. 488–493 (Amsterdam, the Netherlands) (1993 Apr.). https://doi.org/10.1145/169059.169431.

[83] M. Meehan, S. Razzaque, M. C. Whitton, and F. P. Brooks, "Effect of Latency on Presence in Stressful Virtual Environments," in *Proceedings of the IEEE Virtual Reality*, pp. 141–148 (Los Angeles, CA) (2003 Mar.). https://doi.org/10.1109/VR.2003.1191132.

[84] S. Van Damme, J. Sameri, S. Schwarzmann, et al., "Impact of Latency on QoE, Performance, and Collaboration in Interactive Multi-User Virtual Reality," *Appl. Sci.*, vol. 14, no. 6, p. 2290 (2024 Mar.). https://doi.org/10.3390/app14062290.

[85] C. Chafe and M. Gurevich, "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry," presented at the *117th Convention of the Audio Engineering Society* (San Francisco, CA) (2004 Oct.), paper 6208.

[86] C. Chafe, J.-P. Caceres, and M. Gurevich, "Effect of Temporal Separation on Synchronization in Rhythmic Performance," *Perception*, vol. 39, no. 7, pp. 982–992 (2010 Jul.). https://doi.org/10.1068/p6465.

[87] A. Carôt, U. Krämer, and G. Schuller, "Network Music Performance (NMP) in Narrow Band Networks," presented at the *120th Convention of the Audio Engineering Society* (Paris, France) (2006 May), paper 6724.

[88] M. Tomasetti, A. Farina, and L. Turchet, "Latency of Spatial Audio Plugins: A Comparative Study," in *Proceedings of the Immersive and 3D Audio: From Architecture to Automotive (I3DA)*, pp. 1–10 (Bologna, Italy) (2023 Sep.). https://doi.org/10.1109/I3DA57090.2023.10289279.

[89] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-Enabled Internet of Musical Things Architectures for Remote Immersive Musical Practices," *IEEE Open Commun. Soc.,* vol. 5, pp. 4691–4709 (2024 May). https://doi.org/10.1109/OJCOMS.2024.3407708.

[90] C. Drioli, C. Allocchio, and N. Buso, "Networked Performances and Natural Interaction via LOLA: Low Latency High Quality A/V Streaming System," in P. Nesi and R. Santucci (Eds.), *Information Technologies for Performing Arts, Media Access, and Entertainment, Media Access, and Entertainment*, pp. 240–250 (Springer-Verlag, Berlin, Germany, 2013).

[91] L. M. Batteau, A. Liu, J. A. Maintz, Y. Bhasin, and M. W. Bowyer, "A Study on the Perception of Haptics in Surgical Simulation," in *Medical Simulation: International Symposium, ISMS, ISMS* pp. 185–192 (Cambridge, MA) (2004 Jun.). https://doi.org/10.1007/978-3-540-25968-8_21.

[92] R. S. Allison, L. R. Harris, M. Jenkin, U. Jasiobedzka, and J. E. Zacher, "Tolerance of Temporal Delay in Virtual Environments," in *Proceedings IEEE Virtual Reality*, pp. 247–254 (Yokaham, Japan) (2001 Mar.). https://doi.org/10.1109/VR.2001.913793.

[93] S. Agarwal and J. R. Lorch, "Matchmaking for Online Games and Other Latency-Sensitive P2P Systems," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 315–326 (2009 Oct.). https://doi.org/10.1145/1594977.1592605.

[94] Z. Tan, Y. Li, Q. Li, Z. Zhang, Z. Li, and S. Lu, "Supporting Mobile VR in LTE Networks: How Close Are We?" in *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 2, no. 1, pp. 1–31 (2018 Mar.). https://doi.org/10.1145/3219617.3219647.

[95] J.-P. Stauffert, F. Niebling, and M. E. Latoschik, "Latency and Cybersickness: Impact, Causes, and Measures. A Review," *Front. Virtual Real.*, vol. 1, p. 582204 (2020 Nov.). https://doi.org/doi:10.3389/frvir.2020.582204.

[96] R. H. Jack, A. Mehrabi, T. Stockman, and A. McPherson, "Action-Sound Latency and the Perceived Quality of Digital Musical Instruments: Comparing Professional Percussionists and Amateur Musicians," *Music Percept.*, vol. 36, no. 1, pp. 109–128 (2018 Sep.). https://doi.org/10.1525/mp.2018.36.1.109.

[97] D. Wessel and M. Wright, "Problems and Prospects for Intimate Musical Control of Computers," *Com-*

put. Music J., vol. 26, no. 3, pp. 11–22 (2002 Sep.). https://doi.org/10.3389/frvir.2020.582204.

[98] G. Essl and S. O'modhrain, "An Enactive Approach to the Design of New Tangible Musical Instruments," *Organised Sound*, vol. 11, no. 3, pp. 285–296 (2006 Dec.). https://doi.org/10.1017/s135577180600152X.

[99] A. P. McPherson, R. H. Jack, G. Moro, et al., "Action-Sound Latency: Are Our Tools Fast Enough?" in *Proceedings of the International Conference on New Interfaces for Musical Expression* pp. 20–25 (Brisbane, Australia) (2016 Jul.). https://doi.org/10.5281/zenodo.3964611.

[100] T. Mäki-Patola, "User Interface Comparison for Virtual Drums," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 144–147 (Vancouver, Canada) (2005 Jun.). https://doi.org/10.5281/zenodo.1176784.

[101] V. Reynaert, Y. Rekik, F. Berthaut, and L. Grisoni, "The Effect of Hands Synchronicity on Users Perceived Arms Fatigue in Virtual Reality Environment," *Int. J. Hum.-Comput. Stud.*, vol. 178, p. 103092 (2023 Oct.). https://doi.org/10.1016/j.ijhcs.2023.103092.

[102] A. Boem and L. Turchet, "Selection as Tapping: An Evaluation of 3D Input Techniques for Timing Tasks in Musical Virtual Reality," *Int. J. Hum.-Comput. Stud.*, p. 103231 (2024 May). https://doi.org/10.1016/j.ijhcs.2024.103231.

[103] Y. Chen, T. Farley, and N. Ye, "QoS Requirements of Network Applications on the Internet," *Info. Knowl. Syst.*, vol. 4, no. 1, pp. 55–76 (2004 Jan.).

[104] M. Gurevich, D. Donohoe, and S. Bertet, "Ambisonic Spatialization for Networked Music Performance," in *Proceedings of the 17th International Conference on Auditory Display* (Budapest, Hungary) (2011 Jun.).

[105] G. Hajdu, "Embodiment and Disembodiment in Networked Music Performance," in *Body, Sound and Space in Music and Beyond: Multimodal Explorations Sound and Space in Music and Beyond: Multimodal Explorations*, pp. 257–278 (Routledge, Abingdon, UK, 2017).

[106] F. Martusciello, C. Centofanti, C. Rinaldi, and A. Marotta, "Edge-Enabled Spatial Audio Service: Implementation and Performance Analysis on a MEC 5G Infrastructure," in *Proceedings of the 4th International Symposium on the Internet of Sounds*, pp. 1–8 (Pisa, Italy) (2023 Oct.). https://doi.org/10.1109/IEEECONF59510.2023.10335480.

[107] H. Wang, R. Martinez-Velazquez, H. Dong, and A. E. Saddik, "Experimental Studies of Metaverse Streaming," *IEEE Consum. Electron. Mag.*, vol. 14, no. 1, pp. 26–36 (2024 Feb.). https://doi.org/10.1109/MCE.2024.3364118.

[108] A. Becher, J. Angerer, and T. Grauschopf, "Novel Approach to Measure Motion-to-Photon and Mouth-to-Ear Latency in Distributed Virtual Reality Systems," *arXiv preprint arXiv:1809.06320* (2018).

[109] B. Van Kerrebroeck, G. Caruso, and P.-J. Maes, "A Methodological Framework for Assessing Social Presence in Music Interactions in Virtual Reality," *Front. Psychol.*, vol. 12 (2021 Jun.). https://doi.org/10.3389/fpsyg.2021.663725.

[110] M. Annett, A. Ng, P. Dietz, W. F. Bischof, and A. Gupta, "How Low Should We Go? Understanding the Perception of Latency While Inking," in *Proceedings of the Graphics Interface Conference*, pp. 167–174 (AK Peters/CRC Press, New York, NY, 2014).

[111] A. Washburn, M. J. Wright, C. Chafe, and T. Fujioka, "Temporal Coordination in Piano Duet Networked Music Performance (NMP): Interactions Between Acoustic Transmission Latency and Musical Role Asymmetries," *Front. Psychol.*, vol. 12, p. 707090 (2021 Sep.). https://doi.org/10.3389/fpsyg.2021.707090.

[112] J. Lazzaro and J. Wawrzynek, "A Case for Network Musical Performance," in *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 157–166 (Port Jefferson, NY) (2001 Jun.) https://doi.org/10.1145/378344.378367.

[113] Á. Barbosa, "Displaced Soundscapes: A Survey of Network Systems for Music and Sonic Art Creation," *Leonardo Music J.*, vol. 13, no. 1, pp. 53–59 (2003 Dec.). https://doi.org/10.1162/096112104322750791.

[114] A. Kapur, G. Wang, P. Davidson, and P. R. Cook, "Interactive Network Performance: A Dream Worth Dreaming?" *Organised Sound*, vol. 10, no. 3, pp. 209–219 (2005 Nov.). https://doi.org/10.1017/S1355771805000956.

[115] G. Weinberg, "Interconnected Musical Networks: Toward a Theoretical Framework," *Comput. Music J.*, vol. 29, no. 2, pp. 23–39 (2005 Jul.).

[116] G. Hajdu, "Quintet.net: An Environment for Composing and Performing Music on the Internet," *Leonardo*, vol. 38, no. 1, pp. 23–30 (2005 Feb.). https://doi.org/10.1162/leon.2005.38.1.23.

[117] M. Anderson, "Resources: Virtual Jamming," *IEEE Spectrum*, vol. 44, no. 7, pp. 53–56 (2007 Jul.). https://doi.org/10.1109/MSPEC.2007.376609.

[118] S. Ariyani, A. I. Wuryandari, and Y. Priyana, "Design and Implementation of BeatME Server for Networked Musical Performance," in *Proceedings of the International Conference on System Engineering and Technology*, pp. 1–5 (Bandung, Indonesia) (2012 Sep.). https://doi.org/10.1109/ICSEngT.2012.6339310.

[119] R. Hupke, J. Dürre, N. Werner, and J. Peissig, "Latency and Quality-of-Experience Analysis of a Networked Music Performance Framework for Realistic Interaction," presented at the *152nd Convention of the Audio Engineering Society* pp. (The Hague, the Netherlands) (2022 May), paper 10546.

[120] F. Meier, M. Fink, and U. Zölzer, "The Jamberry-A Stand-Alone Device for Networked Music Performance Based on the Raspberry Pi," in *Proceedings of the Linux Audio Conference*, pp. 31–39 (2014 May).

[121] M. Sacchetto, P. Gastaldi, C. Chafe, C. Rottondi, and A. Servetti, "Web-Based Networked Music Performances via WebRTC: A Low-Latency PCM Audio Solution," *J. Audio Eng. Soc.*, vol. 70, no. 11, pp. 926–937 (2022 Nov.). https://doi.org/10.17743/jaes.2022.0021.

[122] J.-P. Cáceres and C. Chafe, "JackTrip: Under the Hood of an Engine for Network Audio," *J. New*

Music Res., vol. 39, no. 3, pp. 183–187 (2010 Sep.). https://doi.org/10.1080/09298215.2010.481361.

[123] L. Turchet and C. Fischione, "Elk Audio OS: An Open Source Operating System for the Internet of Musical Things," *ACM Trans. Internet Things*, vol. 2, no. 2, pp. 1–18 (2021 Mar.). https://doi.org/10.1145/3446393.

[124] G. F. Shay, "How AES-67, the New Audio-Over-IP Standard, Will Bring the Convergence of Telecommunications, Intercom, Radio and Television Broadcast Studio Audio," in *Proceedings of the SMPTE 15: Persistence of Vision Defining the Future*, pp. 1–10 (Sydney, Australia) (2015 Jul.). https://doi.org/10.5594/M001620.

[125] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally Adapted Immersive Environments for Distributed Music Performances in Mixed Reality," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–10 (Erlangen, Germany) (2024 Oct.). https://doi.org/10.1109/IS262782.2024.10704217.

[126] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor, "Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration," *J. Neurosci.*, vol. 36, no. 38, pp. 9888–9895 (2016 Sep.). https://doi.org/10.1523/JNEUROSCI.1396-16.2016.

[127] I. Viola, J. Jansen, S. Subramanyam, I. Reimat, and P. Cesar, "Vr2gather: A Collaborative Social VR System for Adaptive Multi-Party Real-Time Communication," *IEEE MultiMed.*, vol. 30, no. 2, pp. 48–59 (2023 Apr.-Jun.). https://doi.org/10.1109/MMUL.2023.3263943.

[128] J. Jansen, S. Subramanyam, R. Bouqueau, et al., "A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds Over Low Latency DASH," in *Proceedings of the 11th ACM Multimedia Systems Conference*, pp. 341–344 (Istanbul, Turkey) (2020 May). https://doi.org/10.1145/3339825.3393578.

[129] J. Tam, E. Carter, S. Kiesler, and J. Hodgins, "Video Increases the Perception of Naturalness During Remote Interactions With Latency," in *Proceedings of the CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pp. 2045–2050 (Austin, TX) (2012 May). https://doi.org/10.1145/2212776.2223750.

[130] T. Taleb, Z. Nadir, H. Flinck, and J. Song, "Extremely Interactive and Low-Latency Services in 5G and Beyond Mobile Systems," *IEEE Commun. Stand. Mag.*, vol. 5, no. 2, pp. 114–119 (2021 Jun.). https://doi.org/10.1109/MCOMSTD.001.2000053.

[131] F. Tang, X. Chen, M. Zhao, and N. Kato, "The Roadmap of Communication and Networking in 6G for the Metaverse," *IEEE Wirel. Commun.*, vol. 30, no. 4, pp. 72–81 (2023 Jun.). https://doi.org/10.1109/MWC.019.2100721.

[132] H.-W. Kao and E. H.-K. Wu, "QoE Sustainability on 5G and Beyond 5G Networks," *IEEE Wirel. Commun.*, vol. 30, no. 1, pp. 118–125 (2023 Feb.). https://doi.org/10.1109/MWC.007.2200260.

[133] A. Carôt, M. Dohler, S. Saunders, F. Sardis, R. Cornock, and N. Uniyal, "The World's First Interactive 5G Music Concert: Professional Quality Networked Music Over a Commodity Network Infrastructure," in *Proceedings of the Sound and Music Computing Conference*, pp. 407–412 (Torino, Italy) (2020 Jun.).

[134] J. Dürre, N. Werner, S. Hämäläinen, O. Lindfors, J. Koistinen, M. Saarenmaa, et al., "In-Depth Latency and Reliability Analysis of a Networked Music Performance over Public 5G Infrastructure," in *Proceedings of the 153rd Convention of the Audio Engineering Society* (New York, NY) (2022 Oct.), paper 10621.

[135] L. Turchet and P. Casari, "Latency and Reliability Analysis of a 5G-Enabled Internet of Musical Things System," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1228–1240 (2024 Jan.). https://doi.org/10.1109/JIOT.2023.3288818.

[136] L. Vignati, G. Nardini, M. Centenaro, et al., "Is Music in the Air? Evaluating 4G and 5G Support for the Internet of Musical Things," *IEEE Access*, vol. 12, pp. 38081–38101 (2024 Mar.) https://doi.org/10.1109/ACCESS.2024.3374641.

[137] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward Low-Latency and Ultra-Reliable Virtual Reality," *IEEE Netw.*, vol. 32, no. 2, pp. 78–84 (2018 Mar.-Apr.). https://doi.org/10.1109/MNET.2018.1700268.

[138] F. Tang, X. Chen, M. Zhao, and N. Kato, "The Roadmap of Communication and Networking in 6G for the Metaverse," *IEEE Wirel. Comm.*, vol. 30, no. 4, pp. 72–81 (2023 Aug.). https://doi.org/10.1109/MWC.019.2100721.

[139] Z. Huang, C. Xiong, H. Ni, D. Wang, Y. Tao, and T. Sun, "Standard Evolution of 5G-Advanced and Future Mobile Network for Extended Reality and Metaverse," *IEEE Internet Things M.*, vol. 6, no. 1, pp. 20–25 (2023 Mar.). https://doi.org/10.1109/IOTM.001.2200261.

[140] Y. Wang and J. Zhao, "Mobile Edge Computing, Metaverse, 6G Wireless Communications, Artificial Intelligence, and Blockchain: Survey and Their Convergence," in *Proceedings of the IEEE 8th World Forum on Internet of Things (WF-IoT)*, pp. 1–8 (Yokahama, Japan) (2022 Oct.-Nov.). https://doi.org/10.1109/WF-IoT54382.2022.10152245.

[141] N. T. Hoa, B. D. Son, N. C. Luong, and D. Niyato, "Dynamic Offloading for Edge Computing-Assisted Metaverse Systems," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1749–1753 (2023 Jul.). https://doi.org/10.1109/LCOMM.2023.3274649.

[142] M. Sharma, A. Tomar, and A. Hazra, "Edge Computing for Industry 5.0: Fundamental, Applications and Research Challenges," *IEEE Internet Things J.,* vol. 11, no. 1, pp. 19070–19093 (2024 Jun.). https://doi.org/10.1109/JIOT.2024.3359297.

[143] C. Rinaldi, F. Franchi, A. Marotta, F. Graziosi, and C. Centofanti, "On the Exploitation of 5G Multi-Access Edge Computing for Spatial Audio in Cultural Heritage Applications," *IEEE Access*, vol. 9, pp. 155197–155206 (2021 Nov.). https://doi.org/10.1109/ACCESS.2021.3128786.

[144] R. Oda, A. Finkelstein, and R. Fiebrink, "Towards Note-Level Prediction for Networked Music Performance," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 94–97 (Seoul, South Korea) (2013 May). https://doi.org/10.5281/zenodo.1178624.

[145] Z. Jin, R. Oda, A. Finkelstein, and R. Fiebrink, "MalLo: A Distributed Synchronized Musical Instrument Designed For Internet Performance," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 293–298 (Baton Rouge, LA) (2015 May-Jun.). https://doi.org/10.5281/zenodo.1179102.

[146] T. Vets, J. Degrave, L. Nijs, F. Bressan, and M. Leman, "PLXTRM: Prediction-Led eXtended-Guitar Tool for Real-time Music Applications and Live Performance," *J. New Music Res.*, vol. 46, no. 2, pp. 187–200 (2017 Jan.). https://doi.org/10.1080/09298215.2017.1288747.

[147] C. Alexandrak and R. Bader, "Using Computer Accompaniment to Assist Networked Music Performance," in *Proceedings of the 53rd AES International Conference on Semantic Audio*, pp. 27–29 (London, UK) (2014 Jan.).

[148] B. Vera and E. Chew, "Towards Seamless Network Music Performance: Predicting an Ensemble's Expressive Decisions for Distributed Performance," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pp. 489–494 (Taipei, Taiwan) (2014 Oct.).

[149] C. Alexandraki and R. Bader, "Anticipatory Networked Communications for Live Musical Interactions of Acoustic Instruments," *J. New Music Res.*, vol. 45, no. 1, pp. 68–85 (2016 Jan.). https://doi.org/10.1080/09298215.2015.1131990.

[150] A. Tanaka and B. Bongers, "Global String: A Musical Instrument for Hybrid Space," in *Proceedings of the 28th International Computer Music Conference*, pp. 299–304 (Gothenburg, Sweeden) (2002 Sep.)

[151] R. Battello, L. Comanducci, F. Antonacci, G. Cospito, and A. Sarti, "Experimenting with Adaptive Metronomes in Networked Music Performances," *J. Audio Eng. Soc.*, vol. 69, no. 10, pp. 737–747 (2021 Oct.). https://doi.org/10.17743/jaes.2021.0034.

[152] A. F. Genovese, Z. Nguyen, M. Gospodarek, R. Pahle, C. Brenner, and A. Roginska, "Holodeck: A Research Framework for Distributed Multimedia Concert Performances," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–10 (Erlangen, Germany) (2024 Sep.).. https://doi.org/10.1109/IS262782.2024.10704113.

[153] 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 *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–8 (Erlangen,Germany) (2024 Sep.) https://doi.org/10.1109/IS262782.2024.10704205.

[154] J.-M. Jot, R. Audfray, M. Hertensteiner, and B. Schmidt, "Rendering Spatial Sound for Interoperable Experiences in the Audio Metaverse," in *Proceedings of the Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–15 (Bologna, Italy) (2021 Sep.). https://doi.org/10.1109/I3DA48870.2021.9610971.

[155] C. Rinaldi and C. Centofanti, "The Musical Metaverse: Advancements and Applications in Networked Immersive Audio," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–7 (Erlangen, Germany) (2024 Sep.). https://doi.org/10.1109/IS262782.2024.10704154.

[156] J. Herre and S. Disch, "MPEG-I Immersive Audio–Reference Model For The Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 229–240 (2023 May).

[157] J. Herre, S. Disch, C. Borß, A. Silzle, A. Adami, and N. Peters, "MPEG-I Immersive Audio: A Versatile and Efficient Representation of VR/AR Audio Beyond Point Source Rendering," in *Proceedings of the AES International Conference on Audio for Games* (Tokyo, Japan) (2024 Apr.), paper 19.

[158] J. Paulus, L. Laaksonen, T. Pihlajakuja, M.-V. Laitinen, J. Vilkamo, and A. Vasilache, "Metadata-Assisted Spatial Audio (MASA) - An Overview," in *Proceedings of the IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–10 (Erlangen, Germany) (2024 Sep.). https://doi.org/10.1109/IS262782.2024.10704105.

[159] E. Fotopoulou, K. Sagnowski, K. Prebeck, M. Chakraborty, S. Medicherla, and S. Döhla, "Use-Cases of the new 3GPP Immersive Voice and Audio Services (IVAS) Codec and a Web Demo Implementation," in *Proceedings of the 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pp. 1–6 (Erlangen, Germany) (2024 Sep.). https://doi.org/10.1109/IS262782.2024.10704170.

[160] A. Boem, M. Tomasetti, A. Gabriele, A. D. Scipio, and L. Turchet, "User Needs in the Musical Metaverse: A Case Study With Electroacoustic Musicians," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 221–229 (Utrecht, the Netherlands) (2024 Sep.). https://doi.org/10.5281/zenodo.13904838.

THE AUTHORS


Alberto Boem


Matteo Tomasetti


Luca Turchet

Alberto Boem is currently a postdoctoral researcher in the Department of Information Engineering and Computer Science of the University of Trento, Italy. In 2010, he received his master's degree in languages and technologies of new media from the University of Udine, and in 2014, he received a Master of Arts in Interface Culture from the University of Art and Design Linz. He received his Ph.D. in human informatics (2019) from the University of Tsukuba. His research and artistic work focus on interactive, embodied, and immersive systems. During the years, he explored areas such as deformable and shape-changing interfaces, multisensory interactions, and musical XR. His research activities have been presented in venues such as ACM DIS, IEEE VR, IS2, and NIME. His artistic work has been presented at several international venues such as Ars Electronica Festival, Sónar+D, STEIM, YCAM, MNAC, and the Guthman Musical Instrument Competition.

•

Matteo Tomasetti is a Ph.D. student at the Department of Information Engineering and Computer Science of the University of Trento, Italy. He studied digital signal processing and electroacoustic composition (B.A.) at the Conservatory of Pesaro and audiovisual composition, 3D audio, and virtual reality (M.A) at the Conservatory of Frosinone and at the Institut für Elektronische Musik und Akustik (IEM) in Graz (Austria), with a thesis focused on interactive music composition experiences in virtual reality environments. His research interests are immersive audio, eXtended Reality, music composition and performance, and human-computer interaction.

•

Luca Turchet is an associate professor in the Department of Information Engineering and Computer Science, University of Trento, Italy. He holds a master's degree in computer science (2006) from the University of Verona, degrees in classical guitar (2007) and composition (2009) from the Music Conservatory of Verona, as well as in electronic music (2015) from the Royal College of Music in Stockholm. He earned his Ph.D. in media technology (2013) from Aalborg University in Copenhagen. His research was supported by international funding agencies such as the European Commission and the European Space Agency. He cofounded the company Elk. He is the chair of the IEEE Emerging Technology Initiative on the Internet of Sounds and the founding president of the Internet of Sounds Research Network. He serves as an associate editor for the Journal of the Audio Engineering Society and IEEE Transactions on Human-Machine Systems.