Musical Metaverse Playgrounds: exploring the design of shared virtual sonic experiences on web browsers

Alberto Boem
Dept. of Information Engineering and Computer Science
University of Trento
Trento, Italy
alberto.boem@unitn.it

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

2023 4th International Symposium on the Internet of Sounds | 979-8-3503-8254-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59510.2023.10335297

***Abstract*—The "Musical Metaverse" (MM) promises a new dimension of musical expression, creation, and education through shared virtual environments. However, the research on the MM is in its infancy. Little work has been done to understand the MM musical capabilities and its user experience. One cause can be found in the lack of technologies capable of providing high-quality audio streaming, and complex enough musical interactions within shared virtual worlds. Two promising candidates for bridging these gaps are web technologies such as WebXR and Web Audio, whose combination can potentially allow more accessible and interoperable networked immersive musical experiences. To explore this possibility, we developed two prototypes of MM playgrounds. We leveraged WebXR through Networked-AFrame, and Web Audio with Tone.js and Essentia.js to create and test immersive sonic experiences that conveniently run on the web browsers integrated into commercially available standalone Head-Mounted Displays. The first playground focuses on facilitating musical creation in a multi-user immersive application through real-time sound synthesis and binaural rendering. The second explores real-time audio analysis and music information retrieval for creating audio-reactive virtual shared environments. A preliminary evaluation of these two playgrounds is also presented, which revealed some usability issues such as: accessing the immersive experiences through URLs on a headset, the ambiguity in ownership of shared musical tools, and the impact of real-time audio analysis algorithms on perceived audio-visual latency. Finally, the paper outlines future work and discusses possible developments and applications of web-based MM environments.**

***Index Terms*—Musical XR, networked music performances, Internet of Musical Things, Musical Metaverse, binaural audio.**

I. Introduction

In recent years there has been a lot of discussion around the idea of the "Metaverse". While this term received both critical and enthusiastic support from industries, media, and academia [1], the idea behind the "Metaverse" refers to the establishment of a persistent group of virtual interconnected worlds that runs in parallel to our physical world, where communication is digitally mediated, and where users own and interact through their avatars with the means of immersive technologies, despite being physically distant. As a result of all of this sudden attention – someone might call it *hype* [2] – we witnessed a resurgent interest in what was known until a few years ago as Shared Virtual Environments [3], [4]. Among all the possible applications and uses envisioned for such environments, music occupies an important but less explored area [5]. In this context, the "Musical Metaverse" (MM) [6] emerged as a vision for establishing a part of the Metaverse dedicated primarily to musical activities. This includes attending concerts with personalized avatars [7], performing music in real-time with acoustic and electric instruments [8], [9], or using body gestures to control virtual 3D musical interfaces [10]–[12]. In general, the MM can be seen as the result of the convergence of different fields and technologies, such as Musical eXtended-Reality (XR) [13], Networked Music Performance (NMP) [14], and the Internet of Sounds (IoS) [15].

However, despite all of these efforts, the research on the MM is in its infancy. Little work has been done to understand the user experience, the enabling technologies, and the issues related to accessibility and privacy. While this might be caused by the lack of enabling technologies capable of delivering high-quality audio streaming, effectively supporting the requirements of NMP, it is also affected by a limited understanding of musical interactions within shared and collaborative virtual worlds. Among many, web technologies such as WebXR [16] and Web Audio [17] represent two promising candidates for the MM. Their combination can potentially pave the way to the creation of accessible and interoperable networked immersive musical applications, which can be conveniently experienced using web browsers integrated into commercially available standalone Head-Mounted Displays (HMDs), such as the Meta Quest, Pico Neo, or Apple Vision Pro.

To investigate this potential, we developed two Playground systems (Figure I). We aim to achieve a deeper understanding of how web technologies can be incorporated into the construction of collaborative virtual musical environments, and explore their technical constraints and artistic potential. Furthermore, we envision employing these two Playgrounds in

our long-term agenda, which involves examining user experience facets, such as accessibility and 3D interface design. For creating the VR environments, we leveraged WebXR through Networked-AFrame (NAF) [18], a framework for managing multi-user virtual application design and deployment. Web Audio was integrated using two frameworks (Tone.js [19] and Essentia.js [20], [21]) to enable sound synthesis, processing, and spatialization.

The first playground focuses on musical creation through real-time sound synthesis. Users can generate three-dimensional synthesizers that can be experienced through spatialized audio and can be operated simultaneously by any user present in the environment. The goal of the second playground was to explore the use of users' voice and audio input (captured by the microphone embedded in commercial HMDs). For this purpose, we explored the integration of real-time sound analysis and processing technologies to provide information about the voice to the users. We also employed such an analysis to create sound-reactive scenographic elements in a multi-user virtual environment.

We first present the characteristics of the aforementioned technologies. Then, we describe the implementation of the two playgrounds. Finally, we provide a preliminary evaluation of the two systems with a selected pool of users aimed at better understanding the limitations and possible applications of the two prototypes. Finally, we discuss the findings of such a study and provide insights and considerations for future work in the context of the MM.

II. Web Technologies for the Musical Metaverse

The prospects of employing web technologies such as WebXR and Web Audio for the development of immersive and networked musical systems are highly promising due to several compelling reasons: 1) **Availability**: these technologies enable seamless experiences accessible directly through web browsers; 2) **Interoperability**: web-based projects offer the convenience of accessing content anytime via an Internet connection, without the need for third-party components, and compatibility across various hardware and operating systems; 3) **Accessibility**: by relying on widely-used and supported languages such as HTML, CSS, JavaScript, and open-source APIs web technologies can be easily integrated and extended by practitioners and researchers. This section will detail the technologies used to implement our MM playgrounds.

A. Web Audio API

The **Web Audio API** (WAA) is a JavaScript API that empowers developers to perform audio analysis and synthesis directly within web browsers. Moreover, the WAA supports ambisonics, for spatial audio through the *PannerNode*, using both equal panning and generalized Head Related Transfer Functions (HRTFs). While this API is relatively recent, it is rapidly gaining popularity in the IoS community [22]–[24]. Moreover, popular audio libraries for the web, such as Tone.js, and Howler.js [25], have been built on top of the WAA and used in countless web applications.

B. WebXR Device API

The **WebXR Device API** [16] enables developers to create and host immersive Virtual Reality (VR) and Augmented Reality (AR) experiences within web browsers. The WebXR Device API aims to simplify the development process of immersive applications by providing a unified set of functions for managing 6 degrees of freedom (6-DoF) tracking and 3D scene rendering. Most importantly, WebXR emphasizes interoperability, allowing the same application to be experienced equally on different supported devices and browsers[1]. Among them, the WebXR Device API is supported by Chromium-based web browsers available inside commercial standalone HMDs such as the Meta Quest Browser, or Wolvic[2], an open-source browser dedicated to XR experiences and available for several devices such as HTC Vive, Meta Quest, Pico, Magic Leap 2, and Lynx.

To access these two APIs we employed two frameworks and libraries to facilitate the development of interactive applications in web browsers. Regarding WebXR, we used A-Frame [26] to create the actual 3D virtual worlds and Networked-AFrame, to manage the multi-user experience. However, A-Frame and NAF do not provide real-time audio synthesis and analysis components. Therefore, we had to integrate third-party libraries to access the functionalities of Web Audio. We selected two libraries, Tone.js and Essentia.js. While the first supports audio synthesis, analysis, and spatialization, the second focuses on music information retrieval.

A-Frame: A-Frame is a popular framework used for designing interactive 3D applications on the browser. A-Frame provides direct access to the WebXR Device API and enables the creation of virtual worlds using HTML, CSS, and JavaScript. Initially developed by Mozilla, A-Frame is an open-source web framework built on top of Three.js [27]. One of the strengths of A-Frame is the concept of components that allow developers to extend its core functionalities.

Networked-AFrame: is a component that extends the functionalities of A-Frame to support the creation of shared virtual environments, where users (represented as avatars) can freely move around in a 3D space and interact with each other through gestures and speech. The most important characteristic of NAF is the possibility of synchronizing the different components in a scene across multiple connected users. This can be done through the "adapters", which can be used for data exchange, using a system of messages and callback systems. Among all the available adapters, we choose the standard *easyrtc* adapter, which is maintained by the NAF community and is based on the "open-easyrtc" library[3].

The *easyrtc* adapter transfers data from different connected users with a peer-to-peer (P2P) network through UDP. This adapter leverages the functionalities of **WebRTC** (Web Real-Time Communication) [28] to allow not only the exchange

[1]For a complete list of WebXR-supported browsers, the reader is referred to https://immersiveweb.dev/#supporttable, accessed: 2023-07-16

[2]https://www.wolvic.com/, accessed: 2023-07-12

[3]https://github.com/open-easyrtc/open-easyrtc, accessed: 2023-07-12

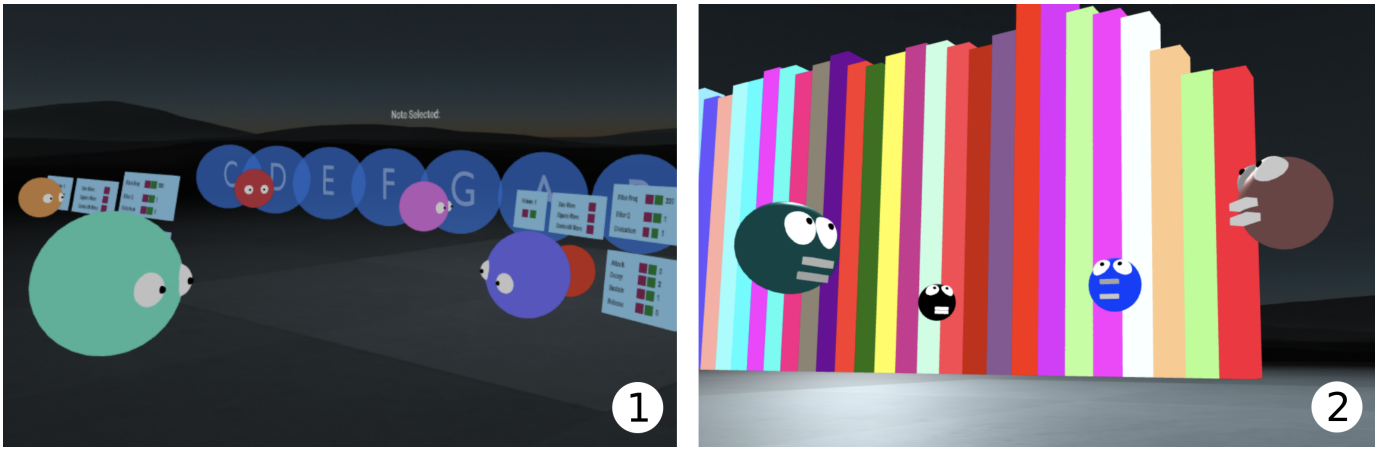

Fig. 1. Screenshots taken by a user inside the two Musical Metaverse Playgrounds. The users are represented as simple spherical avatars. Playground 1 (on the left) involves the use of the *shared sound generators*. Playground 2 (on the right) is designed as *multi-user audio reactive environment*. A detailed examination of the two Playgrounds will be provided in Subsections IV.A and IV.B, respectively.

of messages between connected clients but also real-time streaming of audio. For broadcasting audio through the *easyrtc* adapter, NAF applications need to use a safe context through HTTPs. Interestingly, NAF provides a built-in function to spatialize audio captured from users' microphones[4]. Voices can be heard as coming directly from the avatar's position. This is an important feature that can be used to increase social presence and feelings of closeness among users. Besides synchronization and media streaming, NAF allows the possibility for users to share the ownership of virtual objects among themselves. A-Frame and NAF are the backbones of the two playgrounds we developed.

Tone.js: is an open-source JavaScript framework based on the WAA that provides a comprehensive set of tools for creating interactive music applications on the web. Tone.js simplifies audio programming by enabling developers to integrate and perform audio synthesis and analysis into their web projects through a set of pre-defined instruments, effects, and control systems. We used Tone.js for real-time sound synthesis and binaural rendering of spatialized sound in the first playground and for real-time sound analysis in the second playground.

Essentia.js: refers to a JavaScript implementation of the Essentia library [29]. Essentia is an open-source audio analysis and processing library developed by the Music Technology Group at the Universitat Pompeu Fabra in Barcelona. Essentia provides a collection of algorithms and tools for audio feature extraction, audio analysis, and audio signal processing. We used this library in the second playground to analyze the audio streams captured from the headset's microphone.

In the next section, we detail the design and implementation of the two proposed playgrounds for the "Musical Metaverse".

[4]https://github.com/networked-aframe/networked-aframe/blob/master/src/components/networked-audio-source.js, accessed: 2023-07-12

III. Design

We developed two immersive VR environments designed as playgrounds for creating and testing multi-user interactive MM experiences. The first playground was designed to facilitate musical creation through real-time sound synthesis and binaural rendering. It allows the creation of virtual instruments whose ownership and control can be shared among all connected users. Conversely, the second playground was designed to explore audio analysis and music information retrieval in the context of the MM. Here, the sound stream captured by the headset's microphone is represented in the space through several sound-reactive three-dimensional entities, and features extracted from the sound stream are also presented to the users in textual form.

A. Environments and interactions

Despite their different goals, the two playgrounds were developed with a similar structure in terms of visual appearance and interaction techniques.

Immersive environment: The environments are presented as flat night landscapes. This was chosen to provide a relaxed and neutral environment.

Embodiment: Users are represented with a 3D avatar. We developed a simple embodiment where the users are presented with stylized heads and two virtual hands representing the hand-held tracked controllers. In the second playground, we added to the avatar a simple mouth composed of two audio-reactive lips.

Interactions: Within the two playgrounds, users can move through the environment with 6-DoF, interact with the environments with gestures, and between them through speech. We implemented a system for navigation and one for selection. Users can move in the 3D worlds by rotating their heads and changing direction and speed using the joystick/touchpad placed on the left-hand-held controller. However, we limited the movement to the horizontal plane. For interacting with

virtual objects, we implemented a *gaze-and-commit* type of interaction[5]. This is a type of indirect manipulation suited for both selection and manipulation of distant objects [30]. This interaction technique proved to be intuitive because it closely resembles the mechanics of the "point and click" paradigm used on desktop computers and is already used in several VR applications. A yellow cursor follows the users' head movements. Therefore, users can select virtual objects just by looking at them. The trigger button on the right controller (usually placed on the back side) is used to confirm the selection.

In the next section, we will detail the implementation of the two playgrounds.

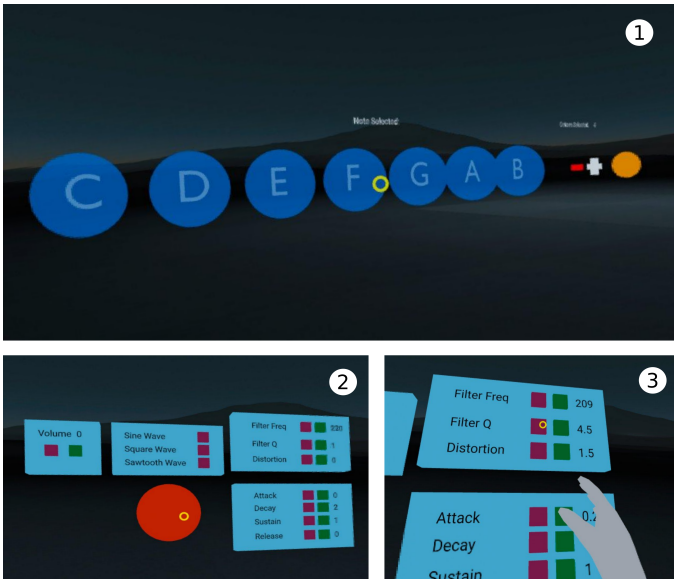

Fig. 2. First Playground: (1) the 3D interface used for selecting the base note, (2) a sound generator, and (3) a detail of the control panels.

IV. Implementation

The source code of the two developed playgrounds is freely accessible online[6].

A. First Playground: shared sound generators

When users enter the virtual world, they will witness a three-dimensional interface placed at the center (see Figure 2 (1)). Users can select one or more musical notes to build their own sound generators through a series of three-dimensional spheres, with the corresponding musical note written inside them, representing a C major scale. With the other two three-dimensional elements -a plus and a minus– users can increase and decrease the octave of the selected notes. Alongside those elements, another 3D element can be used to clear the selected notes. The interface can be used simultaneously by all connected users.

Then, the user can dynamically create a sound generator by clicking on an arbitrary point on the terrain. The sound generators are polyphonic synthesizers that can be controlled in real-time. The sound generators will produce the note that was previously selected.

As shown in Figure 2 (2), a sound generator is composed of a sphere that controls its state: when blue, the generator is off; when turned on, the color will change to red. Near the sphere, there are four three-dimensional panels composed of different buttons. They control different parameters of the sound generators (see Figure 2 (3)). The purple buttons are used to decrease the corresponding value, while the green buttons are used to decrease them. Some parameters that can be controlled are a) the general amplitude; b) a band-pass filter (main frequency, roll-off, and Q); c) the amount of distortion; d) an ADSR envelope. In addition, the users can select different types of sound waves (i.e., square, sawtooth, and sinewave). Once created, any user active in the virtual scene can control the parameters of the sound generators that have been placed in the environment. The sound generators are based on a template that merges 3D objects created with A-Frame components, and the sound synthesis was developed with a Tone.js chain of effects. Differently from speech (that uses the media stream function from WebRTC), the synthesizers are controlled with messages that are broadcasted between peers to minimize latency and bandwidth requests.

To spatialize the audio produced by the generators, we employed binaural synthesis. This method is preferred for three-dimensional environments and is intended for experiencing spatial sound through headphones [31]. Conveniently, Tone.js provides a useful wrapper of the PannerNode of the WAA[7]. When a sound generator is created, it defines the point where the sound will be emitted. Then, in the same position, a PannerNode is also created. A listener allows each user to experience the audio in three dimensions by moving inside the environment. Each sound source is then convolved using a generalized Head-Related transfer function (HRTF), which was also provided by the WAA[8].

B. Second Playground: multi-user audio reactive environment

In this playground, the experience revolves around the features extracted from the sound acquired by the microphones integrated into the HMDs worn by each user. These features are visualized both as three-dimensional virtual objects, as well as displayed in textual form as numbers and text.

We selected three main features of the audio signal. The first is the signal's amplitude, expressed as Root Mean Square (RMS); the second is its harmonic characteristics, calculated using the Harmonic Pitch Class Profiles (HPCP). These two features were implemented with the respective functions available in Essentia.js. Specifically, the HPCP was calculated

[5]https://aframe.io/docs/1.4.0/introduction/interactions-and-controllers.html#gaze-based-interactions-with-cursor-component, accessed: 2023-07-12

[6]https://github.com/CIMIL/MusicalMetaversePlaygrounds

[7]https://tonejs.github.io/docs/14.7.58/Panner3D, accessed: 2023-07-20

[8]https://developer.mozilla.org/en-US/docs/Web/API/PannerNode, accessed: 2023-07-18

through the "TonalExtractor"[9]. The third feature was the frequency data of the audio stream. These were extracted using the Tone.js implementation of the Fast Fourier Transform (FFT).

We developed a custom pipeline to integrate the real-time sound analysis within NAF and its adapter (see Figure 3). At first, the audio stream from the microphone is accessed through the getUserMedia() function[10]. To process this signal in real-time, we used the AudioWorklets of the Web Audio API[11]. With this interface, we created two personalized nodes dedicated to audio processing on a separate thread to ensure low latency. Inside both nodes, we implemented a custom processor that is automatically called by the browser while accessing simultaneously the audio stream captured from the microphone. One processor executes the functions for the RMS and HPCP, while the other executes the function for the FFT. Since these processes are executed in two separate threads, we also implemented a two-way pipeline for transmitting the analysis results to NAF, in order to use them in the scene.

When entering the second Playground users will witness a series of animated vertical bars representing the data derived from the FFT. In front of them, each user sees a small panel designed as a personal display, where the results of the HPCP and RMS analysis are presented as text. Both displays are shown in Figure 4. In this playground, each avatar possesses a stylized animated mouth composed of two lips set in motion according to the RMS values derived from the microphone signal (Figure 5). In addition, by using the *gaze-and-commit* interaction, users can generate unlimited 3D objects. These objects are used as additional visualizers of the RMS. They are designed to resemble a simple vertical loudness meter: the height of the virtual object increases as the loudness increases. To avoid confusion, users can control only the objects they have created, as shown in Figure 6.

V. Evaluation

Following a user-centered design approach [32], we involved prospective users in the early stages of the development process of our applications. The goal of the evaluation was to gather preliminary feedback about the user experience and the functionalities of the system, understand their limitations, and identify possible artistic usage of the two Playgrounds. Additionally, we wanted to understand how to improve our systems and gather potential implications for the MM at large.

The adopted methodology is based on a think-aloud protocol [33] where users interact with the system while verbally describing its functions and commenting about its usability. Subsequently, we applied an inductive thematic analysis on the collected data [34].

[9]https://essentia.upf.edu/reference/std_TonalExtractor.html, accessed: 2023-07-21

[10]https://developer.mozilla.org/en-US/docs/Web/API/MediaDevices/getUserMedia, accessed: 2023-07-21

[11]https://developer.mozilla.org/en-US/docs/Web/API/AudioWorklet, accessed: 2023-07-21

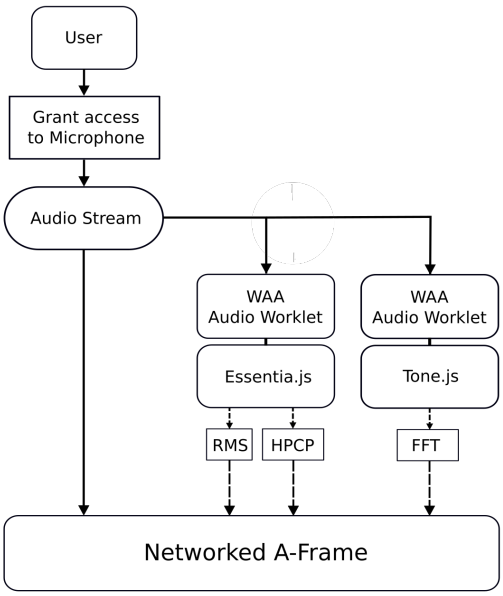

Fig. 3. Second Playground: A schematic representation of how the audio stream is managed with the AudioWorklet from the Web Audio API.

Eight practitioners with experience in both VR and music were invited to test the two playgrounds (8 males, aged between 23 and 31, mean = 25.12, standard deviation = 2.7). During the evaluation, each participant was paired with one of the authors to act as a partner. They both wore an Oculus (now Meta) Quest 1 and a pair of stereo headphones. The experimenter and the participants were always in the same building, connected to the same Wi-Fi network, but placed in separate rooms with an area ranging from 10 to 50 meters. For each playground, there was a 15-minute session where participants had to explore the functionalities of the environments. They could engage with the experimenter through voice chat and ask questions in case of issues or problems. Before starting the evaluation session, the experimenter provided a five-minute briefing about the system. Subsequently, a URL was provided to each participant to access the experience.

A. User Study

We report the main themes derived from the thematic analysis conducted on the participants' comments. We refer to the eight participants as A, B, C, D, E, F, G, and H.

1) First Playground:

- **Spatial audio needs visual aids**: The eight participants experienced difficulty in monitoring the state of each sound generator from a distance (D: *"When I was moving around, I was not sure which was on and which off, as well I was not sure in which configuration they were at the moment"*). According to the participants, since the sound of each generator is spatialized, the only way to experience the sound produced is by standing close to them. To alleviate this, they suggested including more clear and elaborated forms of visual feedback to com-

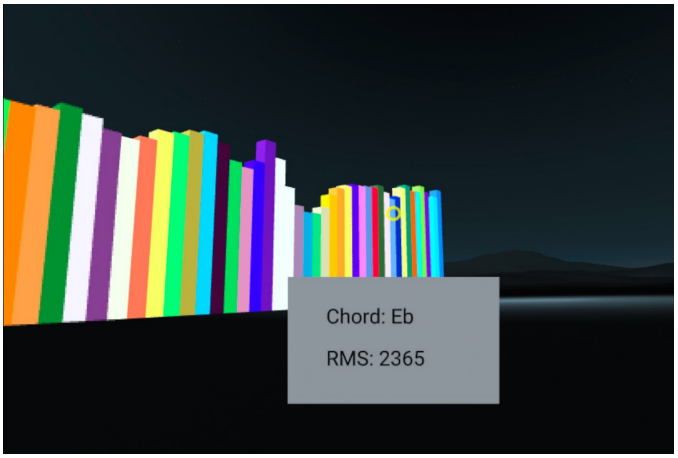

Fig. 4. Second Playground: On the foreground, the personal display placed in front of each user, with the calculated HPCP and RMS values presented in textual form. In the background the FFT data visualization is made visible.

municate the state of each sound generator despite their location in space (B: *"I would have expected much more visual markers and feedback regarding the synthesizers"*). However, they all appreciated the spatialization of the sound produced by the different generators.

- **Ownership of sound generators**: Three participants (B, C, H) found confusing that any connected user could modify the sound generators, even if they weren't the creators. Even if these participants appreciated such a possibility from the perspective of collaboration (H: *"I'm not used to this type of control that is shared among different musicians, but it might open a lot of possibilities"*), they highlighted some critical issues. According to these users, such type of collaborative activity (i.e., changing together the parameters of a sound generator) is dependent on the type of community involved and on how much a person will trust the other users (C: *"Basically, you can go anywhere and build your castle, but this might bother someone"*). They highlighted that it should be important to set some boundaries. Furthermore, they suggested the idea of implementing a mechanism to determine whether to permit other users to share with them the ownership of the generators they have created. Moreover, they suggest the idea of adding a mechanism to decide whether to allow or not to share the ownership of the generators they have created. This will help users to preserve their setups and sonic compositions (C: *"It will be really awful if someone comes and disrupts my work"*).

2) Second Playground:

- **Audio-visual latency**: According to six participants (A, C, D, F, H, G), there was a noticeable latency between the sound input (processed as RMS values) and the movement of respective virtual entities used for visualization. Subjects reported some inconsistency in the movement (A: *"I think there some sort of a lag"*, C: *"Seems that the cube was starting and finishing to move later compared to my speech"*). Within all, three added that they perceived a slight latency regarding the movement of the avatars' mouth (also mapped to the RMS values), but it was not considered harmful for the experience (D: *"I felt the lips were not precise, but it didn't bother me"*).

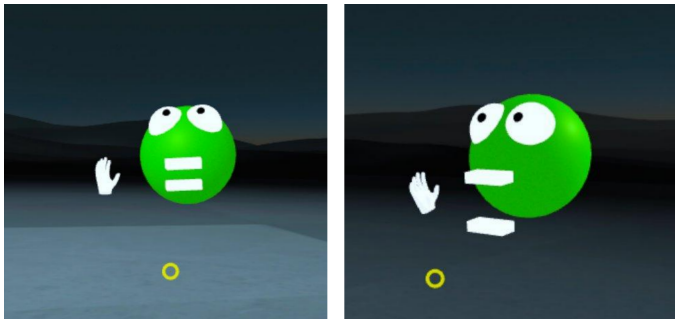

Fig. 5. Second Playground: The image shows the animated mouth of the avatar controlled by the RMS value.

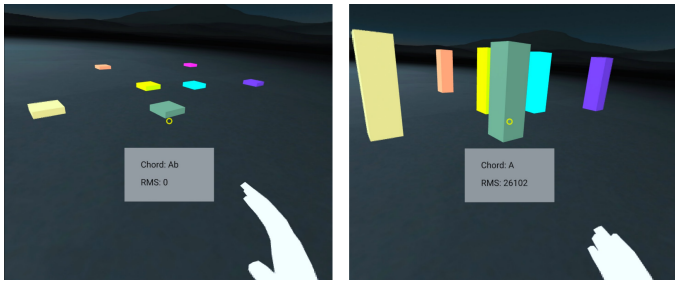

Fig. 6. Second Playground: The virtual entities used to visualize the amplitude of a user's voice. The two figures show the behavior of these virtual entities when sound is detected (on the right), and when sound is not detected (on the left).

- **Audio feature extraction and visualization**: While all participants appreciated the presence of audio-reactive three-dimensional objects, four (B, E, F, G) pointed out that to convey information, these elements should provide more details (B: *"It is not really clear what these virtual bars mean"*). For example, participant A suggested that the FFT visualization should include more information about the different frequencies analyzed. Participants F and G highlighted that virtual entities generated by the users should be used to visualize not only the amplitude of a signal but also other characteristics of the sound (G: *"I would like to see the pitch represented apart from volume"*).

3) General usability issues:

- **Enter a WebXR experience**: All participants emphasized that the sequence of actions needed for accessing the immersive experience within the browser inside the HMD is not intuitive. Initially, they must manually input the URL into the browser, and grant several permissions (i.e., allowing the browser to access the microphone and the HMD to enter the immersive session), a series of steps that they found to be less than seamless. Three

participants (B, D, F) noted that differently from accessing web pages on smartphones, with HMDs it is not possible to scan a QR code. Moreover, typing the URL with the tracked controllers was perceived as a tedious and imprecise activity.

B. Post-task interview

The trial tests were followed by an unstructured interview where participants were asked to reflect on their experience and identify possible usages of the two playgrounds. We grouped the emergent themes as follows:

- **The use of voice increases engagement and collaboration**: All participants commented positively on the presence of a real-time voice chat system (D: *"it really enhanced the interactive aspect"*). The possibility of talking with other connected users was to keep participants motivated and interested in the experience (A: *"Listen to another person really helped me to move forward in the experience"*). Moreover, when voice communication was paired with an action (i.e., creating and modifying the parameters of a sound generator) voice communication reinforced the idea of collaboration (D: *"If I am asked of an opinion, I can quickly act and reply"*; E: *"playing can become a collaborative game"*). In addition, all participants highlighted that spatialized voice contributed to creating a feeling of immersion and realism.
- **Social presence**: All participants commented that the avatars' animated hands and mouths made the experience more interesting and engaging. The animated mouth – even in its simplicity– was considered an important element for visually discriminating which user was talking (C: *"I will be able to understand who is talking, even if I don't know the voice in advance"*). Similarly, three participants (A, B, G) suggested that the 3D elements that reacted to sound in real-time should be used as a visual indicator for social activity and spatial presence. These audio-reactive virtual objects can also be employed more explicitly to represent the loudness of a musical performance or of a group of people talking. This can potentially allow every connected user to develop a better awareness of the level of noise present in the virtual environment (B: *"When I enter the virtual world, I want to look if other people are present...if I see something moving according to the voice of users I can understand if people are talking even from far away"*).

Next, we present the main usages and applications envisioned by participants for the first and second playgrounds.

- **Spatial compositions**: Six participants (A, B, D, E, F, G) regarded the first Playground as a comprehensive tool for creating and experiencing interactive music through the spatial arrangement of sound sources. While the method used in the current prototype to create sound generators was considered somewhat limiting (each generator is bound to a single note), it made these participants think about the possibility of composing and playing musical melodies that develop and unfold in space. Participants B and E suggested that a composer might create "*sound trails*" or "*music journeys*", such as a virtual composition made of musical phrases that can be listened to by the audience moving through them at different speeds. Participants A and F said that placing infinite sound generators in space could allow the creation of "virtual ambient compositions" that can evolve and be modified for an extended period of time.
- **Voice group training and performance**: Five participants (A, B, C, D, G) envisioned the second Playground as a valuable tool for voice training (B: *"You might use the visual representation to understand how your voice is performing"*, C: *"I can see it perfect for teacher-student interactions"*); and for promoting activities such as group singing and virtual choirs (A: *"I imagine a situation like an alto and a soprano, like a choir"*, G: *"different singers can rehearse together"*).

VI. Conclusion and Future Work

In this paper, we presented two playgrounds for exploring technologies and user experiences in the context of the "Musical Metaverse". Regarding the technologies used, our work showed how web frameworks based on the WebXR (i.e., A-Frame, and NAF, used for designing the virtual environments) and Web Audio (i.e., Tone.js and Essentia.js, used for managing the audio processing) can be successfully integrated to create audio-first shared virtual experiences [35]. Moreover, these technologies enable immersive experiences that can be accessed directly from web browsers compatible with commercially available standalone HMDs. However, these technologies (in terms of hardware and software) have their own limitations.

The first group of limitations pertains to the equipment used for the development and test phase, such as the Oculus Quest and Meta Quest 2, but also extends to other standalone headsets. As already mentioned, these HMDs come with integrated omnidirectional microphones. These microphones are placed on the bottom part of the headset, where the user's mouth will be positioned. They also come with integrated functions for dynamic compression and noise cancellation. Such an arrangement is designed specifically to isolate human voice from the environment and to avoid potential conflicts with the sound coming from the speakers embedded in the HMDs. While this design is very suitable for speech, it is unsuitable for musical purposes, such as recording singing voices. Similarly, these HMDs do not come with headphones but only with speakers. Therefore, to experience binaural audio, developers, and users must use their own headphones that must be connected to the headset. These are important elements to consider when designing multi-user musical virtual environments.

The second group of limitations is related to NAF and the adapters used for sharing data and audio streams between users. As reported by the developers of NAF[12] –and confirmed

[12]https://github.com/networked-aframe/networked-aframe/wiki/NAF-adapters-comparison#easyrtc-adapter, accessed: 2023-07-21

by our preliminary tests– with the standard *easyrtc* adapter it is difficult to provide a stable experience with more than 4-5 users connected at the same time. In fact, during our testing phase, we experienced some inconsistency when multiple users (from two to five) if connected not only to the same but also to different Wi-Fi networks. Among these issues we report participants' inability to: 1) receive microphone input from others and 2) view the avatars of connected users.

A third group of limitations pertain to the user study. We assessed the system with only eight male participants. The study should be expanded with a larger pool of subjects, especially to capture a more diverse group of users. These problems represent some major impediments to the creation of fully functional shared musical experiences. Future work might include the development of custom adapters for NAF that can support a larger number of connected users without impacting data synchronization and audio quality during streaming.

While the user study highlighted a few usability issues that will be addressed in the next iterations of the two playgrounds, participants provided some interesting insights on future applications of MM environments.

The first playground points towards a more systematic exploration of the creative use of real-time sound synthesis combined with binaural audio in shared virtual experiences. Future work in this direction should involve composers and performers investigating musical practices that use space as a compositional tool [36].

As suggested by participants, the second playground should be further developed to support musical activities based on group singing. Virtual choirs not only have a long tradition [37], but they have been identified as an important model of collaboration in NMP [38], and found to be beneficial for improving well-being [39]. Future work might include understanding the impact of 3D avatars on the user experience and the role of binaural audio in the context of group singing. Additionally, real-time voice analysis should be explored to create interactive visual displays for self and context awareness (i.e., represent the noise level in an environment during a virtual concert).

In conclusion, web technologies such as WebXR and Web Audio might not provide a definitive solution to the main problems affecting networked music systems (i.e., supporting the largest number of concurrent users, network latency, jitter, and packet loss concealment), and multi-user virtual environments (i.e., synchronization of virtual entities, social interactions). However, being open source and widely supported, they offer interesting opportunities for researchers and practitioners to extend their functionalities.

Our work showed that such systems can successfully satisfy the primary function of shared virtual environments, such as synchronous communication. Moreover, we showed that a combination of WebXR and Web Audio API can be used to build usable prototypes of "Musical Metaverse" applications, and exploited to develop new interactions and shared virtual musical instruments.

VII. Acknowledgment

This work has been supported by the Italian Ministry for University and Research under the PRIN program (grant n. 2022CZWWKP). We would like to thank Pierre Folgarait and Eros Ribaga for their help in the early development of the two playgrounds.

References

[1] M. Ball, *The metaverse: and how it will revolutionize everything*. Liveright Publishing, 2022.

[2] Y. K. Dwivedi, L. Hughes, A. M. Baabdullah, S. Ribeiro-Navarrete, M. Giannakis, M. M. Al-Debei, D. Dennehy, B. Metri, D. Buhalis, C. M. Cheung *et al.*, "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 66, p. 102542, 2022.

[3] R. C. Waters and J. W. Barrus, "The rise of shared virtual environments," *Ieee Spectrum*, vol. 34, no. 3, pp. 20–25, 1997.

[4] N. Durlach and M. Slater, "Presence in shared virtual environments and virtual togetherness," *Presence: Teleoperators & Virtual Environments*, vol. 9, no. 2, pp. 214–217, 2000.

[5] B. Loveridge, "Networked music performance in virtual reality: current perspectives," *Journal of Network Music and Arts*, vol. 2, no. 1, p. 2, 2020.

[6] L. Turchet, "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, pp. 1–17, 2023.

[7] B. Loveridge, "An overview of immersive virtual reality music experiences in online platforms," *Journal of Network Music and Arts*, vol. 5, no. 1, p. 5, 2023.

[8] D. Dziwis and H. von Coler, "The entanglement: Volumetric music performances in a virtual metaverse environment," *Journal of Network Music and Arts*, vol. 5, no. 1, p. 3, 2023.

[9] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, "Evaluation of metaverse music performance with bbc maida vale recording studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, 2023.

[10] L. Men and N. Bryan-Kinns, "Lemo: Exploring virtual space for collaborative creativity," in *Proceedings of the 2019 Conference on Creativity and Cognition*, ser. CC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 71–82.

[11] R. Hamilton, "Coretet: A Dynamic Virtual Musical Instrument for the Twenty-First Century," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, pp. 1395–1395.

[12] ——, "Collaborative and competitive futures for virtual reality music and sound," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, pp. 1510–1512.

[13] L. Turchet, R. Hamilton, and A. Çamci, "Music in Extended Realities," *IEEE Access*, vol. 9, pp. 15 810–15 832, 2021.

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

[15] L. Turchet, M. Lagrange, R. C., G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The Internet of Sounds: Convergent Trends, Insights and Future Directions," *IEEE Internet of Things Journal*, 2023.

[16] "WebXR Device API," https://immersive-web.github.io/webxr/, accessed: 2023-06-12.

[17] "Web Audio API," https://www.w3.org/TR/webaudio/, accessed: 2023-06-15.

[18] "Networked A-Frame (NAF)," https://github.com/networked-aframe/networked-aframe, accessed: 2023-06-05.

[19] "Tone.js," https://tonejs.github.io/, accessed: 2023-06-08.

[20] A. A. Correya, D. Bogdanov, L. Joglar-Ongay, and X. Serra, "Essentia.js: A javascript library for music and audio analysis on the web," in *Proceedings of the 21st International Society for Music Information Retrieval Conference; 2020 Oct 11-16; Montréal, Canada: ISMIR; 2020. p. 605-12.* International Society for Music Information Retrieval (ISMIR), 2020.

[21] A. A. Correya, J. Marcos Fernández, L. Joglar-Ongay, P. Alonso Jiménez, X. Serra, and D. Bogdanov, "Audio and music analysis on the web using essentia.js," *Transactions of the International Society for Music Information Retrieval. 2021; 4 (1): 167-81.*, 2021.

[22] C. Roberts, G. Wakefield, and M. Wright, "2013: The Web Browser as Synthesizer and Interface," *A NIME Reader: Fifteen Years of New Interfaces for Musical Expression*, pp. 433–450, 2017.
[23] B. Smus, *Web Audio API: advanced sound for games and interactive apps*. O'Reilly Media, Inc., 2013.
[24] L. Turchet and M. De Cet, "A web-based distributed system for integrating mobile music in choral performance," *Personal and Ubiquitous Computing*, pp. 1–14, 2023.
[25] "Howler.js," https://howlerjs.com/, accessed: 2023-06-15.
[26] "A-Frame," https://aframe.io/, accessed: 2023-06-05.
[27] "Threejs," https://threejs.org/, accessed: 2023-06-05.
[28] "WebRTC," https://webrtc.org/, accessed: 2023-06-05.
[29] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, H. Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, X. Serra *et al.*, "Essentia: An audio analysis library for music information retrieval." International Society for Music Information Retrieval (ISMIR), 2013.
[30] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev, *3D User Interfaces: theory and practice*. Addison-Wesley Professional, 2017.
[31] H. Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992.
[32] S. Kujala, "User involvement: a review of the benefits and challenges," *Behaviour & information technology*, vol. 22, no. 1, pp. 1–16, 2003.
[33] J. Nielsen, "Estimating the number of subjects needed for a thinking aloud test," *International journal of human-computer studies*, vol. 41, no. 3, pp. 385–397, 1994.
[34] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.
[35] A. Çamcı and R. Hamilton, "Audio-first VR: New perspectives on musical experiences in virtual environments," *Journal of New Music Research*, vol. 49, no. 1, pp. 1–7, 2020.
[36] A. Basanta, "Extending musical form outwards in space and time: Compositional strategies in sound art and audiovisual installations," *Organised Sound*, vol. 20, no. 2, pp. 171–181, 2015.
[37] C. Bendall, "Defining the virtual choir," *The Choral Journal*, vol. 61, no. 5, pp. 69–77, 2020.
[38] J. Galván and M. Clauhs, "The virtual choir as collaboration," *The Choral Journal*, vol. 61, no. 3, pp. 8–19, 2020.
[39] H. Daffern, D. A. Camlin, H. Egermann, A. J. Gully, G. Kearney, C. Neale, and J. Rees-Jones, "Exploring the potential of virtual reality technology to investigate the health and well being benefits of group singing," *International journal of performance arts and digital media*, vol. 15, no. 1, pp. 1–22, 2019.