Locally Adapted Immersive Environments for Distributed Music Performances in Mixed Reality

Andrea F. Genovese[1], Marta Gospodarek[1], Zack Nguyen[2], Robert Pahle[2], Agnieszka Roginska[1]

New York University, [1]*Music and Audio Research Lab*, [2]*Research Technology*, New York, US.
genovese@nyu.edu, gospodarek@nyu.edu, zack.nguyen@nyu.edu, pahle@nyu.edu, roginska@nyu.edu

***Abstract*—The integration of network-based music performance with immersive media technology opens new compelling avenues for live collaborative multimedia concerts and exhibitions. This paper presents a development workflow for creating cohesive immersive environments for multimedia network music performances, where an exhibition node is connected to a remote ensemble. These environments rely on the usage of local characterization data, such as room acoustic measurements and digital twin assets, to render a mix of audio and motion-capture streams, capable of merging remote and local performers into a single shared cohesive display streamed to eXtended Reality devices. Spatialization and auralization techniques are used to add realism to the auditory elements and ground them in the acoustics of the local exhibition node shared by musicians on stage and audiences. The resulting "hybrid" display combines virtual-reality and mixed-reality principles to create a specific type of novel interactive concert experience.**

***Index Terms*—network music, interactive displays, multimedia, motion capture, augmented concerts, XR, auralization, immersive music**

I. Introduction

In recent years, the fields of network-based music performance (or distributed music) and immersive displays have been often merged to create novel applications and types of entertainment experiences. These applications are currently available not only to academics and researchers but also to consumers, artists, and independent developers [1]–[5]. By leveraging existing tools, such as Virtual Reality (VR) and Mixed Reality (MR) immersive devices, 5G network technology [6], efficient audio exchange software, and immersive audio rendering technology for source spatialization and auralization - the general public is closer than ever to the possibility of enjoying new types of live collaborative artistic displays relating to music and multimedia arts. Moreover, these novel paradigms are interesting not only for developing new forms of audience entertainment experience but also for enhancing the range of creative tools available to artists, enabling more effective and immersive network-based performances.

While most devices and infrastructure services are largely developed by the technology industry for the consumer market, academic research has assumed the role of exploring the design of innovative musical interaction models and discussing the technical and human boundaries that define a new relationship between artists and a new generation of technology. For example, immersive audio rendering embedded in a multimedia display has the potential to make the digital performance experience more analogous to its real-world counterpart, enhancing the engagement of the user's senses. However, that aspect should be integrated to work efficiently with the overall network structure, keeping an eye on engineering costs and signal latency issues that may hinder the ability to play music. It is the trade-offs between these elements that need further exploration in order to provide guidelines that can optimize and maximize the space of artistic freedom and quality of experience against technical limitations and specific challenges pertinent to the production goals.

In the present paper, the authors' objective is to describe a novel framework for developing multimedia immersive distributed performances that are tested using the *Holodeck* research platform, a result of an inter-lab collaboration across *New York University*, powered by the *Corelink* data exchange protocol. In practice, this discussed technical architecture involves "augmentations" such as the usage of motion capture and avatar rendering, spatial audio reproduction, auralization methods, and custom-made data exchange protocols. A preliminary evaluation is provided but more rigorous assessment and extensive latency measurements will be conducted in future works. This framework is of relevance for usage in the Internet of Sounds [7] community, Networked Immersive Audio [8], and Musical Metaverse [9] due to its linkage of a novel research infrastructure[1] with XR and immersive audio.

The framework is explored through the development of a "hybrid" experience that fuses VR visual elements with MR auditory elements to create an immersive and realistic navigable environment experienced through VR headsets. This experience is constructed over two connected nodes, an exhibition node where the audience and local musicians are present, and a remote node where an ensemble is captured and transmitted. The experience relies on the usage of auralization techniques for merging the acoustics of the remote audio streams with the local acoustics inherent to the exhibition space, which matches the character of locally produced sound. Furthermore, a digital twin of the concert space is used as a visual environment for rendering and arranging the avatars of performers, grounding the visual element to the sensorial

The work described in this paper has been partially funded by the NSF MRI Award #1626098

[1]This infrastructure is described extensively in a companion publication submitted by the authors to the same conference: *Holodeck: A Research Framework for Distributed Multimedia Concert Performances*

expectations set forth by the local acoustic space. In this setup, the performers use a "leader-follower" unidirectional interaction paradigm [10] allowing the Corelink server to synchronize streams without concern for latency issues. The paper portrays the elements necessary for the development and implementation of this design and a discussion of its limitations.

II. Background and Framework

A. Literature on Immersive NMPs

Network Music Performances (NMPs) involve performers collaborating from different locations using telecommunication technology. The two main challenges regard signal latency and audio fidelity. Over the past two decades, NMPs have gained popularity due to advancements in high-speed academic networks with reduced latency and the growth of network-based communication tools with personal devices. Today, specialized software enable remote collaborative interactions among artists, such as distributed concerts. For a general audience, digital music collaboration over the internet is, however, still an inadequate substitute for real-life performance settings [11], failing to provide effective key social "connectedness" aspects, such as feeling present with others similar to live situations [12], [13]. It is therefore of interest to this field to explore the potential impact of integrative immersive technologies in both the auditory and visual realms as a way to enhance the quality of experience and effective engagement with the NMP paradigm for live performance, education or rehearsal purposes.

Setting up an NMP network involves balancing latency and audio quality. Latency, which cannot be fully eliminated, is influenced by geographic distance, network infrastructure, and bandwidth. Audio "quality" in NMPs can be further subdivided into audio-fidelity quality, and audio-display quality. Digital interventions on quality improvements usually add additional latency to the pipeline, therefore their usage needs to be motivated by specific application goals or design properties, and enabled by a large enough "latency budget" that is determined by the needs of the artistic experience at hand and the quality of a network connection. Whenever viable, the quality of an auditory display in an NMP can be addressed and improved with addition of processing effects that augment the audio presentation into an interactive immersive environment through spatialization and auralization effects [14].

For these purposes, spatial audio and auralization technology aim to provide a "realistic" auditory virtual environment which simulates a real-world acoustic setting [15]. A spatial audio display is capable of providing the auditory perception of a three-dimensional sound stage, which allows the positioning of sound sources in space through binaural or soundfield signal processing techniques [16]. Spatial audio has been explored in NMPs through binaural headphone rendering [17], [18] or loudspeakers [19]. Additionally, digital auralization techniques allow the overlay of the acoustic character of a shared virtual space to an incoming (non-reverberant) sound stream [17]. This can be achieved using either priorly measured acoustic data or simulated sound reflection patterns and reverberation decay, potentially coherent with a real/virtual visual display. The combination of spatialization and auralization techniques is a powerful mix that can allow performers to perceive sound as "externalized", or "real" [20], thus augmenting the plausibility and quality of a virtual presentation.

On the topic of *immersion* and communication quality, visual contact is crucial for fostering a sense of *presence* for musicians and can sometimes be more impactful than acoustic cues for mutual understanding of expressive intentions [21]. Most commercially available tools for distributed performance support video streaming to enhance connection, but video streams typically require higher bandwidth and have greater compression latency than audio. This leads to higher overall transmission latency, sometimes in the order of hundreds of milliseconds, causing out-of-sync rendering. Musicians often disregard the video feed for tempo synchrony, though in some cases, additional latency can be added to the audio buffer to re-synchronize with the video stream when artistically viable [4], [10]. An alternative to video in NMPs is motion-capture (mo-cap) data, which transmits small data loads representing three-dimensional point coordinates of human movement. This method uses fixed tracking camera systems and special suits with trackers [22]. The data is live-streamed and interpreted to recreate digital avatars through a graphics engine, minimizing video delay and allowing smooth visual interactions between participants [23]. This improves interactive visual experience and musical collaboration. Achieving virtual *copresence*, or the feeling of "being together" in a telematic space, involves placing avatars within a shared virtual space, potentially paired with a cohesive shared acoustic environment [13], [24].

This modality has made its way into distributed performance studies paired with virtual immersive environments designed for head-mounted displays (HMDs) devices used for virtual, mixed, and augmented reality. The use of mo-cap streaming does in fact facilitate the connection of real and virtual performers with audiences, in shared virtual spaces. One such example is Coretet, designed for co-located live performances, which allows performers to play sounds from a virtual stringed instrument while in VR [25], or novel interactive interfaces [26]. Other examples explored the use of shared visual spaces for real and virtual sound [27], with coherent acoustic environments [28], [29]. Virtual displays are not necessarily tied to HMDs, Hupke et al. [18] proposed a system called IRENE that connects remote musicians in a shared acoustic space. Instead of using HMDs, which might prove uncomfortable for performance, the virtual meeting space is projected on a wall. Moreover, mo-cap data applies to diverse kinds of arts, and it has been explored for network dance performances [30], [31] as well as combined music and dance connections in the work covered in the later chapters of this paper.

B. Acoustics simulation

For a truly immersive experience, it is important to achieve acoustic cohesion between the reproduced sound of remote

and local performers. Since the reference character is given by locally emitted sound, the remote streams need to be processed at the receiving node with acoustic simulations calibrated to the same exhibition space. In order to achieve that, a virtual acoustic system needs to be implemented.

Every virtual acoustic system consists of three core modules: source modeling, environment modeling, and listener modeling. In the context of NMPs, where the goal is to seamlessly integrate local and remote performers within a single local environment, these modules aim to replicate the characteristics of reality, including the properties of the remote source, the local physical space, and the listener experiencing the performance through headphones. The more accurately each module mirrors reality, the more immersive and convincing the user experience becomes. Additionally, these modeling components should be dynamically responsive, updating in real-time to reflect changes in the sound scene, especially when the listener is in motion, such as in 6DoF dynamic experiences.

Source modeling techniques typically involve modeling the source radiation pattern. Dynamic directivity enables different sound spectra depending on the source's rotation relative to the listener. In most implementations, sound sources are modeled as omnidirectional point sources, which is sufficient for applications with static listeners. However, when the listener is in motion (in 6DoF systems) achieving higher realism requires more detailed rendering. Radiation patterns vary in complexity depending on the source type [32], [33].

The aim of room modeling is to simulate how sound propagates within the local acoustic space. A sound wave generated by the source interacts with its surroundings, reflecting and diffracting when it encounters boundaries. Consequently, the sound wave reaches the listener not as a single event but as a series of reflections. Initially, these reflections arrive at distinct time intervals, but as their energy decreases and echo density increases, they create a diffuse reverberation. The properties of a space can be characterized by a room impulse response (RIR) recorded within that space. The RIR is typically divided into three sections: direct sound, early reflections, and late reverberation. The direct sound segment is usually modeled separately to account for varying source and receiver positions. The amount and type of early reflections depend on the room's shape and the dispersion and absorption coefficients of surface materials. Simulating early reflections is particularly challenging, as their pattern changes based on the source and receiver positions within the room. Accurate rendering of early reflections requires detailed measurements or high-cost calculations, which are often impractical. Therefore, various approximation methods are explored to simplify rendering without sacrificing perceived spatialization quality. Research indicates that in some contexts, maintaining a consistent early reflections pattern may suffice for achieving high auditory plausibility [34]. Late reverberation is generally considered diffuse, with individual reflections not differentiated and uniformly distributed around the listener, making the late reverberation time-frequency envelope independent of position in the room [35].

There are three primary approaches to modeling room acoustics for virtual environments: physical models, convolution with pre-measured RIRs, and algorithms such as delay networks. Physical models encompass geometrical acoustic (GA) methods [36], which aim to compute various propagation paths based on an initial model of the space, receiver, and source. In GA, sound is assumed to propagate as rays. Practical implementations of this approach require a model of the space and the absorption coefficients of surface materials. Using this information, the RIR for a given position of the remote source and local listener is computed and can be updated in real-time.

In the convolution approach, the incoming audio stream is convolved with the RIR measured in the local space. Types of RIRs used include omnidirectional RIRs, binaural RIRs (BRIRs), and spatial RIRs (SRIRs). This technique typically divides the reference IR into two or three segments: direct sound and reverberation, or direct sound, early reflections, and late reverberation. Direct sound is simulated separately to ensure proper sound positioning through convolution with HRIRs. Reverberation usually remains unchanged from the reference IR. Since the spatio-temporal structure of the early reflections segment is challenging to simulate, various simplification methods are used. For example, maintaining the temporal structure and using convolution with HRIRs to obtain a single spatial pattern of reflections independent of room position [37], or modifying the measured early reflections section of the RIR to account for different source and listener positions.

In the algorithmic approach to room modeling, the local space's reverberation is approximated using a system of delay lines. In typical rooms, reflections accumulate until the mixing time, establishing diffuse reverberation. The late reverberation in a room is a diffuse sound field independent of the source and listener positions. Since individual reflections are no longer noticeable, reverberation can be approximated using delay lines. Feedback Delay Networks (FDNs) are designed with parallel delay lines connected recursively through a feedback matrix [38]. A set of multi-band absorptive filters is connected with delay lines to control the frequency-dependent reverberation time. The advantages of FDN algorithmic reverb include simple design, low computational complexity, and high-quality reverberation that can be easily tuned to match the real room's reverberation characteristics.

The final crucial aspect is listener modeling using digital filters for binaural sound, such as HRTFs. This allows for accurate dynamic source positioning for headphone playback during the listener's movement in the exhibition space. The early reflections and late reverberation segments are often transformed or captured in the Ambisonics domain which allows to preserve the spatial positioning of all of the sound components. The sound captured in the Ambisonics domain is then decoded into binaural for proper playback on headphones. While remote sources can also be reproduced locally through loudspeakers, this reproduction method is less flexible and more challenging to set up than playback on headphones.

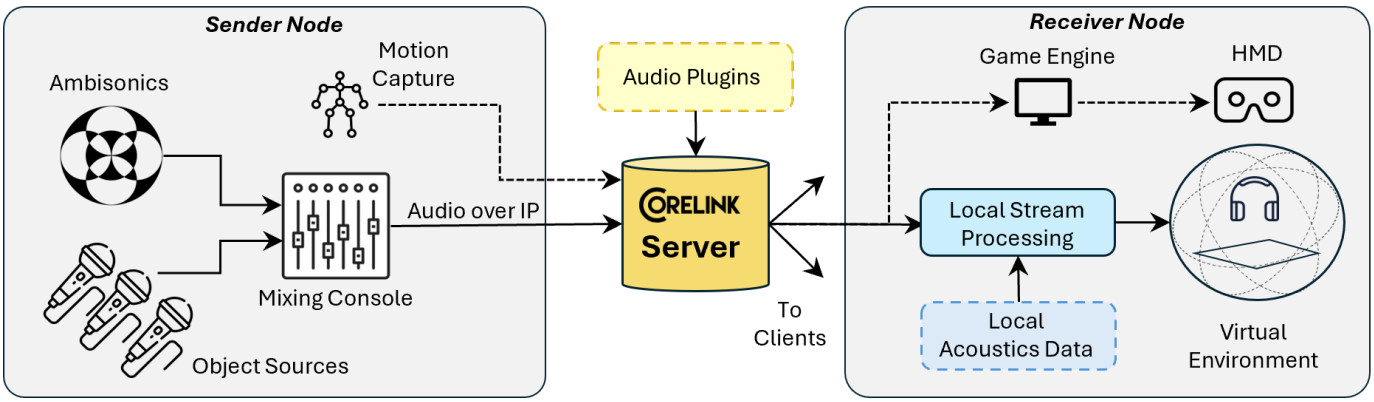


Fig. 1: Example setup of a Corelink sender and transmitter node for sending and rendering immersive audio formats.

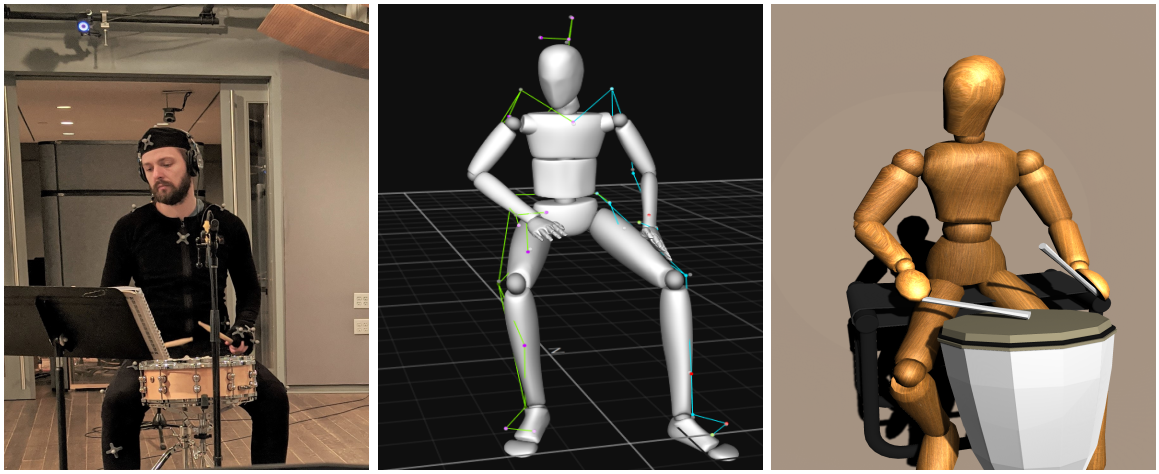
C. The Holodeck and Corelink Framework

The work described in this paper has been developed as part of the *Holodeck* project and facilitated by its network routing tool, *Corelink*. Holodeck is an ambitious collaborative project initiated at New York University (NYU) with support from the National Science Foundation. This project aims to develop an immersive research environment that seamlessly integrates augmented reality (AR), mixed reality (MR), and virtual reality (VR) simulations, transitioning from conventional projection systems to advanced immersive technologies. The Holodeck framework focuses on creating a versatile infrastructure equipped with sophisticated tools designed for national and international research, academic exploration, and creative outputs. Each connected laboratory room within the Holodeck network is outfitted with various sensors, such as cameras and motion tracking rigs for capturing video and body motion data, as well as high-quality microphones for audio capture. This setup enables the transmission of data through a dedicated server, Corelink, and allows for the rendering of requested data tailored to the local configurations of each room. By facilitating the interaction between local and remote participants, the Holodeck enhances collaborative research efforts across multiple disciplines, including computer science, digital media, and educational technology.

Corelink[2], a crucial component of the Holodeck framework. Corelink's architecture supports the specification of workspace and stream types, allowing for the multiplexing and demultiplexing of numerous data streams of various data types within a single internet application. This capability is especially advantageous in scenarios involving virtual performances that necessitate the integration of multiple modalities, such as audio, video, and motion-capture data. The versatility of Corelink extends to its ability to manage complex data flow requirements, ensuring that each data stream maintains its integrity and quality throughout the transmission process. This is achieved through its platform-neutral and data-agnostic design, which accommodates a wide range of data formats and protocols. In virtual performances, where precision and timing are crucial, Corelink's robust streaming capabilities facilitate the synchronization of different modalities, creating a harmonious and engaging performance environment. This flexibility not only enhances the technical execution of virtual performances but also opens up new possibilities for creative expression and collaborative innovation in the realm of digital arts and multimedia.

An example setup for an Holodeck experience enabled by Corelink is shown in Fig. 1. In this scenario, a sender node captures audio in different formats, for example using soundfield multichannel audio (e.g. Ambisonics) for the local ambiance and using spot directional microphones for individual sources. The combination of these audio streams (which may or may not be pre-processed by a local machine), is then sent over the Corelink protocol along Motion-capture data captured concurrently. Thanks to time-stamping and sample rate headers, the server can synchronize the streams and send them over as a package to one or more receiver nodes running a Corelink decoder application. The visual environment is compatible with both two-dimensional displays (screens, projectors) and interactive virtual displays like HMDs. The server has the ability to process the data arbitrarily for any passing stream, meaning that potentially sound effects or data corrections can be added through plugins. The plugins (e.g. artificial reverb for audio, or skeleton data interpolation for motion-capture) are driven by metadata parameters specified by either the transmitter or receiver node and sent along the stream, and can be specified differently for any node subscribing to the stream. For lower-latency processing, sound effects can also be processed at the decoder side, using a local machine present in the receiver node. In the portrayed example, the received streams are optionally spatialized at the receiver node using the local orientation and positioning data (retrieved by the HMD) for accurate 3D sound placement. Auralization can occur by using locally captured acoustic data to characterize the sound and merge it cohesively with the reproduction space and any locally produced sound for a full mixed-reality experience.

[2]Corelink is an open-source project available to the public, the authors encourage participation in the code repository: *Corelink repository link: https://dev.hsrn.nyu.edu/corelink.*



(a) Tracked performer in studio (b) Raw mo-cap points and skeleton (c) Rigged and rendered avatar

Fig. 2: Capture, skeleton, and rendered avatar of a drum performer

III. Experience Design

A platform such as the Holodeck allows novel interaction paradigms to be studied for both technical and artistic feasibility, as well as behavioural evaluation studies linked with distributed immersive NMP performance. To explore the potential of the platform, the authors experimented with multimedia streams and immersive reproduction devices through the design of conceptual augmented music performance experiences. This process involved the enacting of test scenarios that gradually enhanced and integrated the usage of the Corelink tools within NMPs. This section dives deeper into the motion-capture aspect of these experiences, providing a summary of integration and production guidelines for "live" or pre-captured performances [22].

In this case-study discussion, the authors cover the creation of a "hybrid" virtual- and mixed-reality experience, compatible with 6-degrees-of-freedom virtual environments. In this setup, a live motion-tracked performer plays along a remote "leader" node, which is either live or pre-recorded. The transmitted multimedia streams are rendered at the receiving node using an audiovisual characterization pipeline that adapts the transmitted content to the local listening environment and displays it to co-located audiences. Through this work, the authors explored viable production methods for delivering an immersive multimedia experience to a co-located audience wearing HMDs. The process allows for the tailoring of the acoustic character to the reproduction space and the display of locally rendered avatars. The authors emphasize that this implementation is here described as a generalizable model. For more technical implementation details, please refer to [28] and [22]

A. Motion Capture overview

One of the core types of data supported by the Corelink routing protocol, and interesting for NMPs, is motion capture (mo-cap) data. This data is usually obtained from commercial tracking hardware and companion tracking software. The tracking multi-camera system functions by observing - within an area range - the position of several infrared reflective tracker objects, organized according to arrangement schemes that signal different body parts to the system. Using this data the tracking software can infer the position, pose, and rotation of a digital human skeleton representation. This "skeleton data" - or in the case of objects "rigid-body" data - is retrieved by the Corelink sender application from the tracking software output port and sent to the central node server for potential processing (e.g. automatic data cleaning or smoothing). The mo-cap stream, plus any annexed stream, is then retrieved at the requesting client node, which routes the data from its receiving port to a 3D graphics game engine, using a manufacturer-made plugin interpreter [39]. At this stage, the data must be annexed to pre-rigged avatar graphics objects to correctly assign each point-data skeleton information to the related body part of the digital character.

The main challenge in recording motion capture and audio simultaneously is managing infrared reflections (IR). Motion capture systems rely on IR light reflected by markers to track motion accurately. However, shiny surfaces and equipment can also reflect IR light, causing issues such as calibration failures or the appearance of unintended markers (artifacts). Compromises are often needed in microphone placement and instrument handling to minimize these light reflections and ensure accurate data capture. Additionally, musicians' body movements and instrument placement need to be restricted, as their natural motions might cause data occlusions. The subtlety of natural performance movements could be hindered by the placement of markers on specific body parts, necessitating a choice between compromising musical execution or reducing tracking resolution (such as by excluding the tracking of fingers). When artifacts are present, the skeleton fitting loses geometrical reference points, resulting in avatars that are prone to disjointed limbs and jittery motions. Usually, this requires manual editing interventions, but a more modern solution to

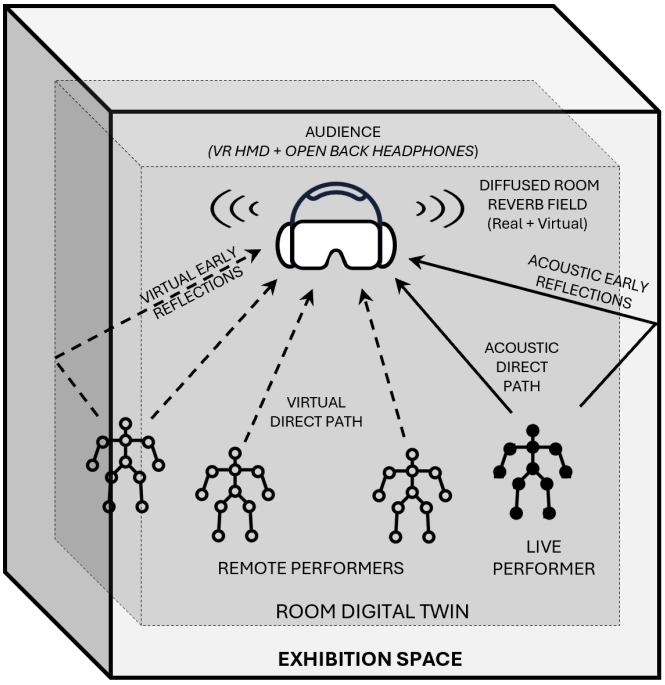


Fig. 3: Conceptual overview of a hybrid cohesive setup. The visual element is displayed over a VR HMD showing a digital twin of the real space, while the auditory element is received in MR and grounded in local acoustics through interactive auralization

this problem, efficient for live streaming, is the application of automatic interpolation or denoising algorithms that process the mo-cap buffers with pre-trained correction models [40]. Fig. 2 illustrates an overview of an avatar creation process from capture to final rendering. For more information on the capture of large ensembles please refer to [22].

In regards to the audio capture, it is necessary to minimize the capture of acoustic reflections by placing the microphone as close as possible to the source. This facilitates the capture of a clean signal, maintaining flexibility for the potential addition of artificial reverb effects or spatial audio rendering. To balance these two sources of noise, highly directional microphones are best kept hidden or covered from the tracking cameras (e.g. with matte coverings) while being placed near the sound source. The audio and motion-capture platforms are also separate asynchronous systems. The mo-cap sample rates are significantly slower than audio sample rates, and the equipment operates on different hardware clocks. Corelink can align the streams using timestamp metadata. For further correction, an audiovisual slate can be recorded to determine the cross-modal latency and manually refine the adjustment.

B. Hybrid Mixed-Reality Experience

A small-scale case study for exploring the integration of motion capture streams in immersive NMP environments was designed in the form of a hybrid mixed-reality concert experience, targeted to one or more audiences wearing HMDs and co-located with part of the ensemble. This type of distributed performance is referred to as "hybrid" due to the combination of a virtual-reality (VR) visual display, and a mixed-reality (MR) auditory display, as shown in Fig. 3. A first implementation using pre-recorded ensemble capture data instead of live streams was presented in [28][3]. Overall, the experience is based on a "leader-follower" NMP interaction style between a remote ensemble and one or more performers co-located with the audience at the receiving node. In this example, the ensemble was composed of four African percussion instruments (Djembe) playing a piece consisting of four voices, in which one member was local and three remotes. The single-member audience consisted of various academic experts who assisted to the performance in turns.

The high-level system design is illustrated as a flexible prototype framework pipeline, fully shown in Fig. 4. The multimedia streams are collected by a local machine and fed to a game engine (e.g. *Unity 3D*) in charge of rendering mo-cap data into digital avatars and processing audio streams with spatialization and auralization effects, responding interactively to the audience's position and orientation in the exhibition space. The concert is experienced by the audience through a tethered VR device displaying a six-degrees-of-freedom virtual environment mixed with the real sound of the local performers.

To establish cohesiveness between auditory and visual realms, the VR environment relies on a pre-built digital twin of the local exhibition space. The twin was realized as a navigable scene asset by carefully measuring the space to create a one-to-one digital copy of the real room, with precise dimensions and boundary placement and approximate imitation of its materials and furniture[4]. The grounding of the experience in the local space characteristic is dictated by the presence of live acoustic sound emitted from the local performers, co-located with the audiences, that generates reverberation and reflected sound that is inherent to said space. Thus, a digital twin provides coherent visuals that can elicit an improved "realism" or plausibility to the holistic experience. There is, however, no necessary limitation to the type of HMD that can be used for this setup. While mixed- or augmented-reality devices would make it easier to place the virtual content in a passthrough-type visual stage without the need for a digital twin, a VR display would allow the local performer to be also motion-tracked and rendered as an avatar along the remote performers. Fig. 4 shows the workflow process for a tethered VR setup.

In regards to the audio layer, it is desirable to achieve acoustic cohesion between the remote and the local performers. Since the reference character is given by the local emitted sound, the remote streams need to be processed at the receiving node with acoustic simulations calibrated to the same exhibition space. One way to acquire the necessary auralization data is to record in situ BRIRs [41] using source-receiver positions corresponding to the arrangement of the remote avatars in the virtual stage of the room's digital twin

[3]A video recording of this implementation of this experience can be found at this link: https://youtu.be/-0VqIn1pTA0?si=gLGJ75lfkUvYAXFq

[4]Digital twin created by the Future Reality Lab at NYU

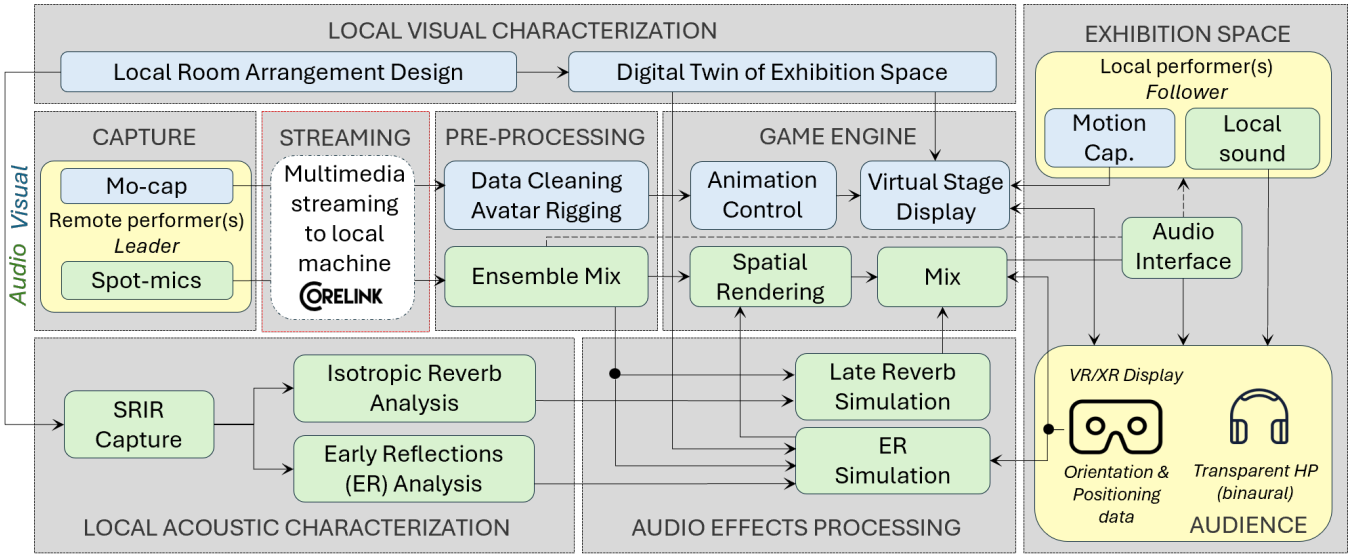


Fig. 4: Design process for a hybrid remote/co-located performance displayed to a locally present audience wearing (in this case) a VR HMD. The remote performers act as "leaders" in the musical interaction, and may be either pre-recorded or live.

(see Section II-B). However, for a more interactive experience compatible with listener-movement tracking technology (3-degrees-of-freedom, or 6-degrees-of-freedom), a more flexible solution is to instead record a single *Spatial Room Impulse Response* (SRIR) through a multichannel soundfield microphone. The acoustic response of an SRIR can then be used to analyse and extract the late reverb decay and early reflection patterns (time, level, and direction of arrival), separating them from the direct sound signal. A possible method to achieve this separation is by using the spatial decomposition method [42]. The assumptions of an isotropic late reverb allows the creation of a position- and rotation-independent reverb bus where the simulation is fed with the unprocessed remote mix. For simulating early reflections, a more sophisticated model (e.g. GA model based on the digital twin geometry) can be used to dynamically modify the recorded early reflection patterns according to the virtual listener-sources positional relationship in the room supported by the HMD tracking and accounting for the virtual stage arrangement. Further adjustments in level ratios between the mixed parts can be applied to account for the distance between the listener and virtual sources.

Separate from the auralization pipeline, the incoming received buffers can be brought into the game engine display and associated with the desired spatial arrangement. A spatial audio rendering plugin can thus be used to process the received streams with *Head-Related Transfer Functions* (HRTFs) to create a binaural version with localizable "object" sources. Finally, the direct spatialized sound is aligned and mixed with the output of the early reflections and the late reverb simulation blocks (accounting for onset delays). The resulting concert experience is that of a dynamic immersive spatial display, with a high degree of plausibility due to its cohesiveness with the visual environment and the locally produced sound.

The leader-follower interaction paradigm [10] has the inherent property of not providing for a two-way interaction. While this has the disadvantage of reducing the performer's engagement with each other, there is a benefit towards the accurate delivery of tempo-critical music to an audience. This paradigm does in fact allow for the different remote media types to be aligned in sync through either prior editing of the pre-recorded material, or delay manipulation for asynchronous stream alignment. Due to the one-way stream connection, live performers are thus capable of delivering their part consistently, without concern for signal latency provided there is absent or negligible jitter. Being this an "audience-first" design, the live performer is not necessarily the target of the immersive experience, although there are no particular technical limitations in replicating the setup for the performer using an additional rendering machine and HMD. However, this is not suggested due to potential impediments in performance mechanics and potential sensitivity to self-delay and acoustic incoherence. From the point of view of an audience, the only potential perceived delay may lie in the avatar rendering of the local musician being off-sync with its live acoustic sound. If the visual rendering is handled by a local machine, and a tethered HMD is used in combination, the graphical rendering delay with respect to sound would only depend on the local system's graphics computation speed without being affected by any transmission latency. This usually amounts to a negligible overall latency, within established JND levels for discrimination of cross-modal asynchrony [43].

Latency Considerations: The work described in this paper has not yet been fully evaluated in regards to observable objective latencies. For this publication, informal assessment by the authors has identified areas of further study. For the visual part, the capture latency would depend on the number of full body "skeletons" as each one would be packetized into a Corelink stream in interleaved manner. Lowering the

capture sample rate can allow the available network bandwidth to handle more skeletons at the cost of visual smoothness. On the rendering side, the latency is defined by the complexity of the visual environment and its interaction with the local machine graphic capabilities. In regards to transmission, the Corelink latency is subject to standard LAN or WAN-related latency and jitter that equally affect alternative systems.

Audio-wise, the latency of this system would be affected by the spatial audio elements, in addition to the standard latency stages inherent to capture, encoding, transmission and decoding. The application of auralization and spatialization filters comes with inherent added latency. Modern spatial audio renderers that are commercially available for game-engines are capable of running binaural rendering close to real-time thanks to short HRIR filters and HOA-domain interpolation. The latency induced by the application of reverb does, however, depend on the desired accuracy towards the ground truth. In fact, a locally measured reverb is usually applied through signal convolution, which would add increasing delay according to the number of filter taps, or duration of sound decay. Common methods to address excessive latency in reverb processing involve the approximation of the real local reverb to an efficient synthetic format, compatible with systems such as a feedback delay network [38]. Additionally, motion-to-sound latency may occur, especially if the tracker technology is wireless.

IV. Discussion

The proposed workflow has been piloted using a musical interaction approach that is not fully interactive, as in a real-life "analog" performance. This limitation can be challenged by studying the application of efficient and practical signal processing techniques capable of reducing the computational load on the rendering machines and allow the total latency to reach usable levels. Further improvements can be investigated by testing other methods of interactive auralization and customizable spatialization (e.g. by using individual HRTFs) and polling changes in reported quality of experience metrics. At the same time, the introductions of approximations in rendering accuracy for these particular displays can be investigated by leveraging the limits of the human auditory system, which does not always necessitate accurate spatialization or auralization to achieve plausible externalization of the binaural audio signal. Related to that, further correlations can be drawn between accuracy needs and musical density or background noise.

This setup is also limited by the need of locally captured SRIRs and availability of a Digital Twin asset. Geometrical acoustic approaches such as raytracing techniques can eliminate the need of capturing SRIRs and directly synthesize the acoustics using the visual room representation. This method is difficult to run in real time but it could potentially provide very accurate acoustic simulations, provided that the underlying visual model is of high-fidelity towards the actual target space of performance. Obtaining a room digital twin is not trivial, this would usually be obtained via manual graphics designer intervention to create the visual assets for a game engine. However, recent photogrammetry techniques have been exploring the automated reconstruction of a digital room from photos and scans [44]. The accuracy of these methods is limited and it may not provide important information for the local acoustics such as the materials of surfaces and obstructive frames present in the room.

Deeper investigations are also required for a potential extension of the system to provide audiovisual interaction capabilities directly to multiple performers. Some promising work has been done for real-time interactive immersive displays in 3DOF modality [45] showing that first order ambisonics is a viable option for an ensemble of interacting singers. A 3DOF auditory display is sufficient for stationary musicians who do not navigate the virtual space environment, and recent work suggests that this format is well-received and desirable for musicians [46]. Regarding the visual environment, there are stronger challenges to face in regards to the material obstruction that HMD can cause to an artists range of movement and vision of self-motion that can be sensitive to small delays between action and visual rendering. Using projectors may be a more viable option for performers [18] although there is a higher cost in equipment and flexibility.

A. Future Work

Beyond the technical improvements of the spatial audio setup, the authors intend to follow up the conceptual framework and initial implementations with a more comprehensive study on the user experience and the latency of the various system components. The paradigm of distributed network music performance is very particular and calls for a different type of evaluation, more centred on the assessment and impact of *co-presence* rather than *telepresence* [47]–[49]. Furthermore, it needs to be taken into account that, from a performer's perspective, the cognitive load of a music performance task may in itself influence the state of "immersion", for example through engagement with a motor activity that interacts with the effects of a media rendering system [29], [50]. In practice, the creation of a formalized methodology may entail a variety of approaches to collect appropriate metrics from audiences and artists according to the taxonomy of the scenario at hand (e.g. according to remote or co-located audiences, musical hierarchies, supported reproduction system, etc.). As a starting point, user studies for both performers and audiences will be conducted by adapting questionnaire templates from existing validated sources in telematic media, such as the System Usability Scale [51], the User Experience Questionnaire [52], and other recent NMP-specific formats [53]–[56].

Latencies for both audio and visual modes will be thoroughly assessed and analyzed in a future stage, including a breakdown of their components, as anticipated in section III-B. To achieve a real "live" interaction for the proposed hybrid system, where the two nodes communicate bi-directionally, significant latency reduction need to be achieved. In practice, some of the spatial audio components may be a hinder due to their added latency, and a balance must be struck between

auralization realism and signal latency. For musical purposes, care must be put in to maintain the signal latency within the 20ms threshold [57].

Future work around Holodeck and Corelink will focus on exploring their boundaries, developing new frameworks and development pipelines, establishing evaluation methods and assessment scales, relating the platform improvement to the creation of improved artistic musical practices. Corelink's future work include improvements for the jitter-management and stream synchronization features, as well as cross-node clock synchronization[5]. In the long term, the holy grail of NMP and immersive media engineering is to democratize existing tools and move towards the possibility of transferring the Holodeck experience to a mobile ecosystem, available at a low cost to the public. A further expansion of this paradigm may thus leverage wearable or IoT sensor functionality [58] to flexibly and dynamically characterize the local acoustic and visual environments and accordingly drive the rendering of a shared shared virtual space for a potential "Musical Metaverse" [9].

V. Conclusions

This paper has outlined a comprehensive framework for integrating NMPs with immersive media technology, providing an innovative approach to one-way live collaborative multimedia concerts and exhibitions. The development workflow presented leverages local characterization data, such as room acoustic measurements and digital twin assets, to render a mix of audio and motion-capture streams that seamlessly merge remote and local performers into a single cohesive display. This integration is achieved through advanced spatialization and auralization techniques, enhancing the realism and immersion of the auditory elements within the local exhibition space. The resulting hybrid display combines VR and MR principles, offering a novel and interactive concert experience that bridges the gap between remote and in-person performances. This work is of interest for applying higher degrees of auditory realism to spatial audio display in Musical XR, enhancing the set of tools for the Networked Immersive Audio and Musical Metaverse community.

The future work on the Holodeck and Corelink platforms will focus on pushing the boundaries of these immersive environments and exploring their implications for artistic musical practices. As the system and exchange protocol evolve, there will be opportunities to test various variables against system latency and subjective evaluations. This exploration will address open questions related to interactive virtual displays in NMPs, considering factors such as multimedia combinations, hierarchical musical organizations, rendering asymmetries, and application purposes. The Holodeck framework's adaptability and reconfigurability facilitate experimental research in internet-based communication and interactive media, driving the proliferation of research-oriented concert events and promoting the development of novel evaluation methodologies and artistic techniques.

[5]Please refer to the authors' other submitted publication for more details on Corelink: *"Holodeck: A Research Framework for Distributed Multimedia Concert Performances"*

Acknowledgment

The authors would like to acknowledge all the people who made these experimental works possible, in particular the Corelink development team, the performers, the dancers, and the NYU Holodeck Consortium, in particular, NYU's *Future Reality Lab* and *NYU MAGNET* who lent the required equipment and spaces.

References

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

[2] M. Bosi, A. Servetti, C. Chafe, and C. Rottondi, "Experiencing remote classical music performance over long distance: A Jacktrip concert between two continents during the pandemic," *Journal of the Audio Engineering Society*, vol. 69, no. 12, pp. 934–945, 2021.

[3] J.-P. Cáceres and C. Chafe, "Jacktrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010. [Online]. Available: https://doi.org/10.1080/09298215.2010.481361

[4] C. Drioli, C. Allocchio, and N. Buso, "Networked performances and natural interaction via LOLA: Low latency high quality A/V streaming system," in *Information Technologies for Performing Arts, Media Access, and Entertainment*, P. Nesi and R. Santucci, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 240–250.

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

[6] 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 Journal of the Communications Society*, 2024.

[7] L. Turchet, M. Lagrange, C. Rottondi, 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*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[8] L. Turchet and M. Tomasetti, "Immersive networked music performance systems: identifying latency factors," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2023, pp. 1–6.

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

[10] A. Carôt and C. Werner, "Fundamentals and principles of musical telepresence," *Journal of Science and Technology of the Arts*, vol. 1, no. 1, pp. 26–37, 2009.

[11] K. E. Onderdijk, F. Acar, and E. Van Dyck, "Impact of lockdown measures on joint music making: playing online and physically together," *Frontiers in Psychology*, vol. 12, no. 5, 2021.

[12] K. E. Onderdijk, D. Swarbrick, B. Van Kerrebroeck, M. Mantei, J. K. Vuoskoski, P.-J. Maes, and M. Leman, "Livestream experiments: the role of COVID-19, agency, presence, and social context in facilitating social connectedness," *Frontiers in psychology*, vol. 12, 2021.

[13] V. Y. Oviedo, K. A. Johnson, M. Huberth, and W. O. Brimijoin, "Social connectedness in spatial audio calling contexts," *Computers in Human Behavior Reports*, vol. 15, p. 100451, 2024.

[14] A. Roginska and P. Geluso, *Immersive Sound*. Focal Press, 2017.

[15] S. Tatlow, "Authenticity in sound design for virtual reality," in *History as Fantasy in Music, Sound, Image, and Media*, 1st ed., J. Cook, A. Kolassa, A. Robinson, and A. Whittaker, Eds. New York, USA: Routledge, 2024, ch. 8, pp. 161–183.

[16] R. Wilson, "Aesthetic and technical strategies for networked music performance," *AI and Society*, vol. 38, no. 5, pp. 1871–1884, 2023.

[17] P. Cairns, H. Daffern, and G. Kearney, "Parametric evaluation of ensemble vocal performance using an immersive network music performance audio system," *Journal of Audio Engineering Society*, vol. 69, no. 12, pp. 924–933, 2021.

[18] R. Hupke, S. Preihs, and J. Peissig, "Immersive room extension environment for networked music performance," in *153rd Audio Engineering Society Convention*, New York, USA, 2022.

[19] L. Comanducci, *Intelligent Networked Music Performance Experiences*. Springer International Publishing, 2023. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-15374-7_10

[20] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[21] J. W. Davidson, "Visual perception of performance manner in the movements of solo musicians," *Psychology of Music*, vol. 21, no. 2, pp. 103–113, 1993.

[22] C. Bui, A. Genovese, T. Bradley, and A. Roginska, "Multimodal immersive motion capture (MIMiC): A workflow for musical performance," in *Audio Engineering Society Convention 149*. Audio Engineering Society, 2020.

[23] A. Hunt, H. Daffern, and G. Kearney, "Avatar representation in extended reality for immersive networked music performance," in *AES International Conference on Spatial and Immersive Audio*, Huddersfield, UK, 2023. [Online]. Available: https://unity.com/

[24] M. F. Schober, "Virtual environments for creative work in collaborative music-making," *Virtual Reality*, vol. 10, no. 2, pp. 85–94, 2006.

[25] 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.

[26] 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.

[27] M. Gospodarek, A. Genovese, D. Dembeck, C. Brenner, A. Roginska, and K. Perlin, "Sound design and reproduction techniques for co-located narrative VR experiences," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[28] A. Genovese, M. Gospodarek, and A. Roginska, "Mixed Realities: a live collaborative musical performance," in *5th International Conference on Spatial Audio ICSA*, Ilmenau, Germany, 2019.

[29] A. F. Genovese, *Acoustics and Copresence: towards effective auditory virtual environments for distributed music performances.* New York University. Ph.D. Thesis., 2023.

[30] D. Strutt, A. Schlegel, N. Coghlan, C. Debaig, and Y. 'Friendred' Peng, "New telematic technologies for remote creation, rehearsal and performance of choreographic work," *Journal of Embodied Research*, vol. 4, no. 2, 2021. [Online]. Available: https://doi.org/10.16995/jer.82

[31] D. Strutt, "A simple tool for remote real-time dance interaction in virtual spaces, or "dancing in the metaverse"," *Critical Stages*, vol. 2022, no. 25, pp. 1–18, 2022.

[32] B. B. Monson, E. J. Hunter, and B. H. Story, "Horizontal directivity of low- and high-frequency energy in speech and singing," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 433–441, 2012. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4725963

[33] M. Karjalainen, J. Huopaniemi, and V. Välimäki, "Direction-Dependent Physical Modeling of Musical Instruments," in *Int. Congr. on Acoustics (ICA\'95)*, vol. 3, no. 5, 1995, pp. 561–563.

[34] M. Gospodarek, *Acoustic and Perceptual Factors Affecting Plausibility in Sound Design for Audio Augmented Reality Experiences*. New York University. Ph.D. Thesis., 2024.

[35] M. Barron and L. J. Lee, "Energy relations in concert auditoriums. I," *Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 618–628, 1988.

[36] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4926438

[37] C. Pörschmann and S. Wiefling, "Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses," *International Conference on Spatial Audio*, no. December 2016, 2015.

[38] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *Audio Engineering Society Convention 90*, 1991.

[39] (2024) Optitrack unity plugin. [Online]. Available: https://docs.optitrack.com/plugins/optitrack-unity-plugin

[40] D. Holden, "Robust solving of optical motion capture data by denoising," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–12, 2018.

[41] J. Vanasse, A. Genovese, and A. Roginska, "Multichannel impulse response measurements in MATLAB: An update on ScanIR," in *AES International Conference on Immersive and Interactive Audio*, York, UK, 2019.

[42] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, 2013.

[43] K. W. Grant, V. van Wassenhove, and D. Poeppel, "Discrimination of auditory-visual synchrony," in *International Conference on Auditory-Visual Speech Processing*, 2003, pp. 31–35.

[44] B. Denkena, M.-A. Dittrich, S. Stobrawa, and J. Stjepandic, "Automated generation of a digital twin using scan and object detection for data acquisition," *Simulation in Produktion und Logistik*, pp. 49–60, 2019.

[45] P. Cairns, H. Daffern, and G. Kearney, "Parametric evaluation of ensemble vocal performance using an immersive network music performance audio system," *Journal of the Audio Engineering Society*, vol. 69, no. 12, pp. 924–933, 2021.

[46] M. Tomasetti and L. Turchet, "Playing with others using headphones: Musicians prefer binaural audio with head tracking over stereo," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 501–511, 2023.

[47] S. T. Bulu, "Place presence, social presence, co-presence, and satisfaction in virtual worlds," *Computers & Education*, vol. 58, no. 1, pp. 154–161, 2012.

[48] R. E. Kraut, D. Gergle, and S. R. Fussell, "The use of visual information in shared visual spaces: Informing the development of virtual co-presence," in *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, 2002, pp. 31–40.

[49] S. Zhao, "Toward a taxonomy of copresence," *Presence*, vol. 12, no. 5, pp. 445–455, 2003.

[50] H. Lee, "A conceptual model of immersive experience in extended reality," 2020.

[51] J. R. Lewis, "The system usability scale: past, present, and future," *International Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, 2018.

[52] M. Schrepp, A. Hinderks *et al.*, "Design and evaluation of a short version of the user experience questionnaire (ueq-s)," 2017.

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

[54] cairns patrick, rudzki tomasz, cooper jacob, hunt anthony, steele kim, acosta martínez gerardo, chadwick andrew, daffern helena, and kearney gavin, "singer and audience evaluations of a networked immersive audio concert," *journal of the audio engineering society*, vol. 72, pp. 467–478, september 2024.

[55] B. Van Kerrebroeck, G. Caruso, and P.-J. Maes, "A methodological framework for assessing social presence in music interactions in virtual reality," *Frontiers in Psychology*, vol. 12, p. 663725, 2021.

[56] R. Hupke, D. Jan, N. Werner, and J. Peissig, "Latency and quality-of-experience analysis of a networked music performance framework for realistic interaction," in *Audio Engineering Society Convention 152*. Audio Engineering Society, 2022.

[57] 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.

[58] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *Ieee access*, vol. 6, pp. 61 994–62 017, 2018.