Telemersive audio systems for online jamming

Patrick Müller
Institute for Computer Music and Sound Technology
Zurich University of the Arts
patrick.mueller@zhdk.ch

Roman Haefeli
Institute for Computer Music and Sound Technology
Zurich University of the Arts
roman.haefeli@zhdk.ch

Johannes Schütt
Institute for Computer Music and Sound Technology
Zurich University of the Arts
johannes.schuett@zhdk.ch

Matthias Ziegler
Institute for Computer Music and Sound Technology
Zurich University of the Arts
matthias.ziegler@zhdk.ch

***Abstract*—This paper focuses on aesthetic approaches, specific projects and tools for online jamming involving immersive audio. By 'online jamming', we refer to the widespread practice of individual performers located in different spaces collaborating via the Internet in near real-time. By 'immersive audio', we mean the capture, processing and delivery of sound with three-dimensional directional characteristics — specifically, through binaural rendering via headphones. The first part of the paper focuses on the spatial dimension of telematic performance practices, which, compared to the temporal dimension of said practices, is under-conceptualised in the relevant literature. The second part highlights some perceptual requirements that need to be considered when musicians play together and interact with each other via binaural rendering. We then analyse a number of tools, technical applications and artistic projects that can be subsumed under an expanded understanding of online jamming with immersive audio, two of which were developed in our own research unit. The analysis namely draws on how these projects aesthetically deal with spatial categories and how they respond to the perceptual challenges. The paper concludes by measuring the latencies of spatial audio plug-ins and streaming utilities. The results show that the potential of such telemersive approaches is promising and that such tools can be designed to address latency issues, which are critical for online jamming.**

***Index Terms*—telematic performance, online jamming, telemersion, immersive audio, Ambisonics**

I. Introduction

Today, a wide range of bi- and multidirectional streaming tools are available for live interaction between remote musicians over the Internet for telematic performances, also known as networked (music) performances [1]. In addition to pioneering utilities such as JackTrip [2], LoLa [3], UltraGrid [4], Soundjack [5], Jamulus [6], TPF-Tools [7] or Artsmesh [8], a large number of applications have emerged, particularly in connection with or in the aftermath of the Covid pandemic, such as SonoBus[1], Qacktrip[2], Quaxtrip[3], Maxtrip[4], FarPlay[5], Elk live[6], Koord[7], OVBOX[8] or Digital Stage[9]. The list could easily go on, and different tools each have specific features. They are offered as standalone devices, web applications or VST plug-ins for digital audio workstations, focus more on audio or video streaming or differ in terms of network architecture, internet quality of service flexibility, number of channels available, audio codecs or latency characteristics, to name but a few. This allows users the undeniable advantage of being able to choose a specific tool to suit their individual needs, whether for online teaching, online rehearsals, online jamming or for demanding projects such as simultaneous performances where stages and audiences are connected in (near) real-time. In telematic performance practice, as well as in the engineering of streaming tools that facilitate such a practice, the time domain is well understood and explored. It builds on one of the main specificities of this artistic format: the propagation time of audiovisual signals between distributed locations. While there are a number of systematic treatments of such latency effects from technical and aesthetic perspectives (e.g. [9] [10] [11]), there are no comparable conceptualisations for the spatial domain. This is even more surprising as the coupling of the acoustic characteristics of virtually linked spaces is another key feature that distinguishes telematic performance from other artistic formats. This lack is confirmed by the conventional performance practice of networked performances in real-time in front of a live audience: in order to minimise crosstalk effects in the bidirectional arrangement of the stages involved, instruments or other sound sources are usually captured by microphones placed in the near field (see e.g. [12] [13] [14]), largely erasing the spatial characteristics of the room in which the performance takes place. On the other hand, the signals received from the remote stage are mixed on the local stage, merging the local and remote sound sources into a coherent mix adapted to the local environment [15]. Finally, the inevitable delayed feedback from the remote location is often masked by artificial reverberation added to the local mix [16]. Thus, the spatial characteristics of the remote space(s), as well as the spatial effects of the intermediate space (the Internet), are nullified, ”creating an ambiguous space that confuses rather than elucidates the interaction between the dispersed ensemble” [17] [18].

[1] https://sonobus. net

[2] https://msp.ucsd.edu/tools/quacktrip/

[3] https://github.com/damonholzborn/Quaxtrip

[4] https://github.com/dispersionlab/maxtrip

[5] https://farplay.io

[6] https://www.elk.audio

[7] https://koord.live

[8] https://github.com/gisogrimm/ovbox

[9] https://digital-stage.org/startseite

A. *Typology of spatial configurations*

Compared to latency issues, there seems to be no systematic conceptualisation of spatial configurations in telematic performances. In the temporal domain, Carôt and Werner [5] have presented a now widely used categorisation based on aesthetic aspects, namely regarding musical interaction styles as a consequence of latency effects. The resulting approaches are useful in the design, development and application of technical devices for rehearsing and performing music in telematics. Accordingly, this paper presents a classification of approaches in the spatial domain, derived from a number of projects that consider spatial acoustics as an aesthetic dimension. In addition to a **Space Ignoring Approach (SIA)**, which does not consider the spatial configuration on a conceptual level, we propose a classification into the following strategies:

Single Perspective Approach (SPA): The primary aim of SPA is to create a consistent acoustic scene, integrating both local and remote players as well as the acoustic consequences of latency and feedback phenomena, according to the characteristics of the local stage. The perspective of the latter is thus the sole guiding principle, and the acoustic situation at the different locations may differ considerably – a difference that is not made aesthetically productive. The conventional practice of close-miking, feedback-masking and rendering through a Public Address System, as mentioned earlier, exemplifies this approach when concert stages for respective audiences are concerned. Another example is rendering through headphones in an online jam, where individual musicians are in separate locations and each musician may monitor a discrete personalised mix using techniques like panning, reverberation or spatial audio. Each location thus has an individual acoustic characteristic without being informed about the others.

Virtual Space Approach (VSA): In VSA, a shared virtual acoustic space is created so that performers and the audience at different venues experience an identical, or at least similar, acoustic environment, merging the spatial properties into a synthetic, overarching virtual space. This can be achieved through an installative approach [19], concert performance [20], headphone rendering [21] or by extracting properties of the intermediary space, such as propagation time, and using them to create a virtual acoustic space [22] [23]. The main goal of this approach is to ensure that all performers at all locations inhabit the same acoustic environment and interact according to its properties [24].

Dynamic Space Approach (DSA): In the aforementioned approaches, spatial characteristics are understood as a static category. In contrast, DSA dynamically transforms the acoustic properties of the involved physical or imaginary spaces over time. Listeners (who might also be performers) are guided sequentially through different spaces [25]. They can virtually visit the various rooms of involved performers [26], experience the transformation between acoustic properties of concert venues [27], witness the evolution from a point source rendering over stereo to an Ambisonics sound scene [28] and more.

Coupled Room Approach (CRA): Whereas the approaches mentioned up to this point are based on an understanding of a specific (real or imagined) acoustic space, CRA adopts a synthetic strategy. It actively blends the acoustic properties of the participating spaces into each other, making the influence of the remote space's specificities perceivable at the local site and vice versa [29] [30]. This can be compared to acoustic volumes coupled through an aperture in non-mediated real-world contexts.

Spatial Discontinuity Approach (SDA): SDA actively reveals the rupture lines between the connected acoustic rooms. It does not preserve ”the fidelity in relation to their equivalent real-world sensory modalities” [31]. Instead, it avoids constructing a (real or conceptual) continuity between the involved spaces, leading to a form of conflict or dissonance in the spatial domain by actively making the rupture lines perceivable in the interplay of the acoustic spaces involved [32] [33] [34].

Spatial audio, defined as the discipline of capturing, processing and delivering sound with three-dimensional directional characteristics [35] – which we will refer to as 'immersive audio' in the following – opens up a wide field of exploration regarding the interplay of different room acoustics. This paper focuses on aesthetic approaches, specific projects and tools for online jamming that include immersive audio features. By 'online jamming', we refer to the widespread practice where individual performers are located in separate rooms and are brought together via the Internet — a form of (video) conferencing for the performing arts. The goal of the strategies observed and proposed here is to transform the telematic performance practice to one which might be called 'telemersive': portemanteau consisting of 'telematic' and 'immersive' [36].

II. Online jamming with immersive audio

With very few exceptions, publicly available streaming tools suitable for online jamming in the field of telematic performances lack functionalities in the area of immersive audio. This is particularly remarkable given the widespread adoption of spatial audio in the consumer sector in recent years, namely in web content, gaming and home electronics [37]. A number of videoconferencing utilities have adopted the potential of immersive audio, e.g. Clubhouse[10], Gathertown[11] or respective audio features in Microsoft Teams[12]. Binaural rendering is now widely accessible and could therefore facilitate applications in telematic performances, bringing together performers from different locations who listen to each other through headphones.

However, there are reasons that complicate the introduction of such technologies, primarily falling under two categories: signal path latency and user-friendliness. Tools designed to facilitate online jamming aim to optimise their latency characteristics. The use of immersive audio can counteract this, as the corresponding processing of the recorded and streamed sounds

[10] https://www.clubhouse.com

[11] https://www.gather.town

[12] https://support.microsoft.com/en-gb/office/spatial-audio-in-microsoft-teams-meetings-547b5f81-1825-4ee1-a1cf-f02e12db4fdb

requires additional time [38] [39]. Additionally, incorporating features for auralisation demands extra functionalities, reducing user-friendliness. Furthermore, the application of immersive audio in such contexts almost invariably requires headphones, as few users have suitable loudspeaker systems at home (and even if they do, playing could be significantly disrupted by feedback effects). However, many musicians feel uncomfortable when listening through headphones while playing [40] [41].

However, these practical considerations are challenged by using conventional streaming tools in telematics, where from the player's perspective, other musicians often overlap in the overall sound image — a phenomenon that simple stereo panning can only partially mitigate. Moreover, individual players may apply reverb or other processing to their own mix or transmit an already reverberated signal, leading to mutual uncertainty about the acoustic environments of the other players. This significantly complicates playing together. Incorporating immersive audio into tools for online jamming therefore involves balancing a wide range of possibilities for designing individual and collective spatial environments that binaural rendering offers with considerations of usability, accessibility and the imperative for low latency.

The remainder of this paper is organised as follows: the following sections explain some of the perceptual principles that have to be respected when it comes to telematic performances with immersive audio and binaural rendering over headphones. Although the following considerations are based on some findings in the field of binaural auralisation in general, they take practices in the field of telematic performances with immersive audio as their starting point and therefore do not claim to be valid beyond this limited focus. Chapter III then presents a meta-study where a series of tools and projects are presented and analyzed, including network topologies for the streaming of audio and spatial metadata. Chapter IV adds some latency measurements for spatial audio plugins (SAPs) and for streaming utilities used in the analysed tools and projects.

A. Perceptual considerations

Early experiments with streaming binaural signals date back to the 1930s at Bell Telephone Laboratories. A dummy head named Oscar was equipped with two microphones, allowing listeners in another room to experience striking spatial acoustic effects, feeling "acoustically transported to the location of the microphones, no matter at what distance from them he might actually be." [42] The concept of binaural telephony and teleconferencing has intermittently been explored in various contexts [43] [44] [45]. The early history of telematic performances and installations is rich with reflections on the spatial dimensions – examples include Dieter Schnebel's *Drei-Klang* from 1971 [46], Bill Fontana's intercontinental Sound Bridges from the 1980s [47] and the pioneering "first professional-quality audio streaming in a public musical event" by the SoundWIRE group at Stanford University in 2000, which involved recording, streaming and rendering of first-order Ambisonics [48].

Nevertheless, the situation changes when considering network music performances involving headphones. Once again, the emphasis on time in telematic research appeared to disregard spatial dimensions, notably in early clapping-experiments on latency effects conducted in anechoic environments with dry signals through headphones (e.g. [49]). It has been observed that latencies manipulated under such controlled laboratory conditions "would in real-life be smeared by multiple acoustical reflections" and that reverberation would "almost certainly mask the effect in more natural situations by cushioning sharp-edged signal arrivals" [50]. The results of such studies can therefore be seen as critical in retrospect, as they ignore central perceptual aspects.

Nonetheless, the impact of spatial considerations on interactions among geographically dispersed musicians remains ambiguous. Carot, Werner and Fischinger [51] found no discernible effects in their study involving percussionists; indeed, the latter appeared distracted rather than aided by artificially introduced reverberation. However, their study does not specify the type of reverberation used. It is known that certain of its benefits, such as externalisation, diminish if applied incorrectly: for example, adding diotic reverberation without interaural time or level differences does not enhance externalisation [52]. Moreover, there could be perceptual impairments if virtual acoustics and the listener's perception of the environment do not align, a phenomenon known as the "room divergence effect". It is plausible that the aforementioned study gave little consideration to such sensitivities, as these experiments are often conducted in small rooms with mono signals treated with synthetic reverb – conditions similar to those of online jams.

B. Effects of reverberation

In contrast, latency effect studies focusing on perceptual aspects clearly demonstrate the positive impacts of reverberation. Gang et al. [16] introduced an "audio latency masking" technique using reverberation, creating a virtual scenario where connected musicians perform as if in electronically linked "reverberant music halls." Objective measurements comparing dry and reverberant acoustics indicated significant improvements in performance synchrony with reverberation. Subjective evaluations showed even more pronounced benefits; musicians perceived less audio latency and their involvement was higher, a crucial aspect for collaborative music practices. Farner et al. [53] conducted hand-clapping experiments across three acoustic environments: real reverberant, virtual anechoic and virtual reverberant. Carefully planned reverberation, including virtual reconstruction of real room acoustics using binaural room impulse responses (BRIRs), resulted in similar performance indicators between real and virtual reverberant settings, both significantly better than in virtual anechoic conditions. Schuett's [24] findings confirm, through negative evidence, that ensemble performance exhibits greater resilience when delays are spatially manipulated (e.g., drummers separated by increasing distances and absent eye-contact in a real-world setting) compared to electronically manipulated delays (e.g., time shifts in dry signals over headphones). The hypothesis that aural cues such as reverberation could explain this phenomenon, however, was not supported when artificial reverberation was added to the dry signals: this did not affect the interaction between musicians. Schuett does not dismiss the potential of reverberation outright but suggests that the artificial nature of the reverberation may have prevented players from perceiving a shared auditory space essential for cohesive musical interaction. He therefore proposes to transport some of the reflection paths over the internet separate from the dry signal or the use of convolution reverb for future studies.

In summary, the outcomes of these studies appear to hinge significantly on the manner in which reverberation is applied rather than simply its application in the first place. In addition to influencing objective measures such as precision and asymmetry, subjective factors like engagement and cohesion also prove pivotal. It is widely acknowledged that reverberation, as a component of spatial auditory perception, yields beneficial effects on perception and interaction within online jamming environments.

C. Effects of binaural audification

The aforementioned configuration using BRIRs [16] represents an initial step towards immersive, spatial audio. In addition to acoustics effects mentioned in the previous section, binaural rendering enables more precise placement of sound sources within the acoustic panorama compared to stereo panning, facilitating clearer separation in the perception of remote players [21] [41]. A series of studies demonstrates that binaural auralisation is beneficial in music performance or listening contexts [54] [55]. Comanducci [40] conducted preliminary tests with binaural audio in online jamming, yielding positive results for auditory perception, though subsequent tests used loudspeaker arrays due to the obtrusiveness of headphones for the players. While the respective reports and documentation do not systematically compare binaural to stereo rendering, they offer promising qualitative assessments from the participants.

Tomasetti and Turchet [56] show the superiority of binaural rendering with head-tracking over stereo in a simulated online jamming environment. Here, the signals of individual players were spatialised using 5th order Ambisonics and binaurally rendered with a head-tracking system, ensuring that fellow players remained ‘world-fixed’ in the auditory scene. A qualitative survey found that the binaural plus head-tracking system received higher evaluations than stereo across several dimensions, including localisation of and connection to the virtual musicians, distinction of one’s own contribution, immersion, realism of the acoustic scene, sound quality and social presence. The study’s limited informative value regarding telematic performances stems from its simulated setting: an individual musician played along with pre-recorded songs in different genres, avoiding latency effects or interaction behavior typical of telematic performances. Further in-depth studies are warranted, as this study only placed sources in two dimensions, did not apply any room simulation techniques and used generic Head-Related Transfer Functions (HRTFs) for the binaural rendering.

Nevertheless, an important observation hints at the user’s behaviour in terms of embodied experience: the results which showed that the participants moved their heads to a greater extent in binaural condition suggest a stronger involvement, granting players the possibility of taking a first-person perspective in their performative behaviour, as opposed to the third-person perspective associated with stereo [56].

D. Externalisation

The binaural rendering of immersive audio is, therefore, a promising approach for enhancing the immersive quality of online jamming practices. However, achieving an accurate representation of sound sources in binaural rendering through headphones requires careful consideration of several factors, particularly externalisation and localisation issues. Binaural Ambisonics in particular often lacks externalisation, which is the perception of a sound source positioned at a certain distance outside the head [57]. This inside-the-head locatedness is a primary challenge for binaural rendering over headphones. Individual HRTFs, which map the physical properties of the user’s body and its influence on sound propagation, would be necessary for an ideal representation. However, producing individual HRTFs is complex and costly, making them impractical for the average user in online jamming contexts.

However, there are strategies to compensate for the lack of individualised HRTFs. In addition to contextual information such as visual cues, two types of cues improve the ability to resolve the aforementioned ambiguities: well-controlled environmental cues derived from models of room acoustics and dynamic cues correlated with head motion of the listener [41]. Research indicates that artificial reverberation can effectively enhance the perception of externalised sound images in headphone listening [44] [58]. To synthesise accurate acoustic images, the ratio of direct to reflected energy may also play a role in enhancing externalisation, with no significant difference observed between early-reflection and full-reverberation conditions for externalisation [59]. Studies suggest that a minimal representation of a reverberant acoustic field is sufficient for many contexts: for instance, no significant difference was found when using reverberation algorithms in 1st or 3rd order Ambisonics [60]. Simulating early reflections and higher-order ambisonic (HOA) reverberation can be computationally intensive, suggesting a potential trade-off to reduce processing latency. Nevertheless, early reflections are crucial for creating a realistic auditory scene, enhancing parameters such as envelopment and spaciousness, thereby augmenting the sense of presence. The same applies to the directivity patterns of sound sources, especially in environments where the sources and/or the listening positions are moving, allowing for a perspective in 6 degrees of freedom (6DoF).

E. Localisation

Localisation, the ability of the human auditory system to determine the direction (azimuth, elevation, distance) from which a sound originates, is another critical factor in binaural rendering. Several strategies can be employed to reduce localisation ambiguities in binaural rendering. In the Ambisonics domain, increasing its encoding order from 1st to 3rd significantly enhances lateralisation in the perception of spatialised sound sources [61] [62]. However, further increases in Ambisonics order do not appear to notably improve localisation [63], at least when generic HRTFs are used [64]. Again, this aspect warrants consideration in the context of CPU usage. While individually-tailored HRTFs also enhance lateralisation, data suggests that most listeners can discern directional information adequately using generic HRTFs, especially for azimuth [65], which is more crucial than elevation in online jamming.

The limitations of generic HRTFs in localisation and externalisation can be partly mitigated through dynamic binaural rendering, which incorporates natural head movements of listeners [59]. Self-initiated movements are particularly effective in this regard [44] [66]. As demonstrated in the above mentioned study by Tomasetti et al. [56], musicians appear to exhibit greater movement in binaural rendering compared to stereo, thereby enhancing localisation and externalisation further. However, these benefits are contingent upon the sound sources being 'world-fixed', or stable relative to the environment. Therefore, head-tracking enabled systems are essential, significantly improving the externalisation and localisation of virtual sound sources [67] [68]. In contrast, the absence of head-tracking represents a compromise, diminishing the accuracy of localisation and externalisation cues in binaural rendering. Nonetheless, 'head-fixed' static binaural techniques may still offer advantages over stereo, including enhanced source intelligibility, reduced masking effects between sources and a more spatially immersive experience.

Head-tracking systems are computationally efficient in scene-based audio, as Ambisonic signals can be rotated using simple frequency-independent matrix multiplication. In contrast, dynamic binaural rendering of object-based audio typically requires fading or switching of filters (HRIRs or Binaural Room Impulse Responses [BRIRs]) [69] [70], which is computationally expensive and not yet explored in existing online jamming tools. When working with Spatial Impulse Responses [SIR], these are usually summed into a single complete Ambisonic scene before applying rotation for head-tracking [71] [72].

The complexity of the human auditory system can therefore quickly lead to a similar complexity on the part of the technical apparatus when it comes to simulating auditory scenes. Thus, tools for telematic performances, including immersive audio, have to mitigate a trade-off between the goal of perceptual quality on the one hand and practical elements such as usability, CPU usage, latency or bandwidth limitations on the other. In the following chapter, a number of tools and approaches are presented with their respective aesthetic dimensions.

III. Online jamming tools with immersive audio

In this chapter, a series of systems employed for online jamming including immersive audio, primarily utilising open-source software, is presented. Some of these systems (introduced and referenced below) are publicly accessible (*digital-stage-web, VRR, OVBOX, Tpf-jam-tool*), while others are documented sufficiently to enable reconstruction (*VIIVA-NMP audio system, dispersion.spat, IRENE, Zone*). Two approaches (*Zone* and *Tpf-jam-tool*) have been developed and are utilised within our own research group at the Institute for Computer Music and Sound Technology of the Zurich University of the Arts. The tools are evaluated based on their use of specific room simulation utilities, whether they employ head-tracking in 3 or 6 Degrees of Freedom, their aesthetic approach to spatial configuration (as discussed in Chapter I-A), their coordination of spatial meta data between locations (see Figure 1), the host system, the streaming protocol used, whether there are video functionalities and whether it is publicly available. An overview is provided in Table I. The Ambisonics order is given if applicable, and the main focus is highlighted in bold where the spatial configuration and the spatial coordination is concerned.

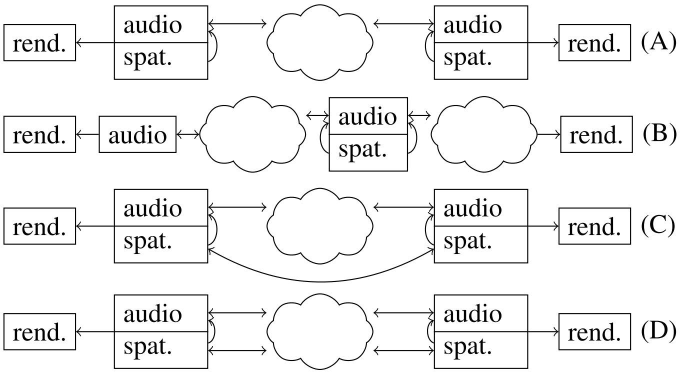

Fig. 1. Signal flows of audio and spatial meta data (spat.) between locations and their rendering (rend.): no coordination of room parameters (A); room simulation on central server (B); coordination by prior agreement (C); or through OSC in real-time (D).

Besides the streaming of audio, which relies on various protocols, and the consideration of perceptual issues when processing immersive audio, as outlined in the previous chapter, tools for online jamming must address how to coordinate spatial parameters across different locations. The following presentation of the tools and projects is organised according to such coordination logics employed. Unless absent (A), these logics are implemented via a central server (B), through the prior exchange of room parameters (C) or through their real-time exchange using the corresponding streaming utility (D). An overview of the signal flow for spatialisation is provided in Figure 1.

A. No coordination of room parameters between locations

Tools in this category focus on the sonic environment of the local performers without accessing or coordinating with spatial information from other locations. Apart from *digital-stage-web*, which is limited to this configuration, most of the other tools of this overview also allow such an approach.

TABLE I

Synopsis of tools for online jamming with immersive audio: spatial audio processing (positioning of sound sources [Pos], room simulation through convolution [IR], feedback delay networks [FdN], ray-tracing [RT]), integration of head-tracking (3 or 6 degrees of freedom), spatial configuration (according to Chapter I-A, spatial coordination (according to Figure 1), host system, streaming protocol, inclusion of video, public availability; in bold: main focus

Tool	Spatial audio processing				Head-tracking	Spatial configuration				Spatial coordination				Host	Protocol	Video	Public
	Pos	IR	FdN	RT		SPA	VSA	DSA	CRA	A.	B.	C.	D.				
digital-stage-web	x				-	**x**				**x**				Browser	WebRTC	x	x
VRR	1st		1st		-		**x**				**x**			Pure Data	AOO		x
dispersion.spat	3rd	x	3rd		3DoF	x	x	**x**			**x**		x	MaxMSP	JackTrip		
VIIVA-NMP (1)	1st	1st			-	x	**x**					**x**		Reaper	JackTrip		
VIIVA-NMP (2)	3rd	3rd			3DoF	x	**x**					**x**		Reaper	SonoBus	x	
OVBOX			1st	x	3DoF		**x**						**x**	generic	zita-njbridge		x
IRENE	3rd	0th	3rd	x	6DoF	x	x		**x**				**x**	MaxMSP	(Dante)	x	
Zone	3rd		3rd		6DoF	x	x	**x**	x				**x**	MaxMSP	TTkit	x	
Tpf-jam-tool	3rd		3rd		3DoF	**x**	x	x		**x**			x	Reaper	Tpf-client	x	x

Although they have the ability to exchange spatial information between locations, immersive audio functionalities can also be applied at each location independently. However, due to their focus on a shared virtual space, in *VRR*, *OVBOX* and in the hub mode of *dispersion.spat*, the spatial coordinates at each location are forced to be identical. Aesthetically, tools in this category represent the technological equivalent of the Single Perspective Approach (SPA, see Chapter I-A).

Digital-stage-web:[13] Digital Space provides a set of tools for online jamming, initiated in the aftermath of the COVID-19 pandemic: *digital-space-pc* is a standalone device for desktop computers and has no immersive audio capabilities; *digital-space-ovbox* is identical to the *OVBOX* device described below; *digital-stage-web* is a browser-based web application based on the WebRTC streaming protocol, which includes a room editor that allows the local positioning of all sound sources in a sound scene. In this application, no further specifications can be applied to modulate the room characteristics.

As a browser application, the spatial audio algorithms executed directly in the browser are limited, and it uses compressed audio formats to limit the bitrate at the cost of introducing additional latency [73]. Consequently, it does not (and does not aspire to) meet the latency standards for online jamming and is therefore more akin to videoconferencing utilities that include immersive audio, as mentioned above in Chapter II.

B. Room simulation on a central server

An obvious approach to providing uniform spatialisation for all locations is to perform the respective processing on a central server. To do this, the signals from all locations must be sent to such a server, at which point they are distributed back in a star topology. The additional signal paths, along with the unpacking and repacking process on the server, add latency, so this approach may conflict with latency-sensitive situations. In addition, the local performer's signal also has to pass through the corresponding signal path, so it is heard with a delay in the overall scene. Alternatively, the local performer can suppress the returning signal, but then they are no longer part of the acoustic scene. *VRR* and *dispersion.spat* in its hub mode use this strategy. Aesthetically, it corresponds to the Virtual Space Approach (VSA). In the case of *dispersion.spat*, options are offered where the spatial and positional characteristics can be changed dynamically, approaching the Dynamic Space Approach (DSA).

Virtual Rehearsal Room (VRR):[14] The audio-only VRR is a standalone application written in Pure Data and based on the Audio Over OSC (AOO)[15] streaming protocol, both of which are open source. As the title suggests, distributed players gather in an environment that models the acoustics of a virtual concert hall, based on IEM's experience with Ambisonics technology. By default, the virtual room is located on a central server, where one of the participants, acting as a 'conductor', places the members of the ensemble in an auditory scene in terms of azimuth and distance and applies reverberation to the overall scene. The individual musician then receives a binaural mix from the central server, either with or without the local player included. There is no head-tracking functionality.

As *VRR* delivers a standalone device, it is relatively easy to use for players in an online jamming context, and the options for virtual acoustics are displayed in a clearly organised interface. The usability nevertheless comes at the price of considerably high latency. The local player is either excluded from the auditory scene, or the musician has to train to hear themselves with delay of latency. An interesting option is to use the latency as first reflection delay, so that it is less disturbing. Additionally, the quality of the sound scene might be compromised due to the lack of a functionality that keeps it world-fixed. A peer-to-peer version where the *VRR* application does the mix for each musician locally is mentioned in the

[13] https://digital-stage.org/browser-version

[14] https://vrr.iem.sh

[15] https://github.com/Spacechild1/aoo

documentation but not elaborated on. VRR does not seem to be operational at present.

dispersion.spat [26] [74]: The utility is based on a similar idea to *VRR*, in that the individual players are connected to a central server where the virtual acoustics are facilitated. The JackTrip protocol is used as the streaming protocol, wrapped in a MaxMSP environment named Maxtrip. Maxtrip can optionally be combined with a network utility, disperf[16], which brokers the JackTrip connections, thus circumventing port forwarding and firewall issues. For room simulation, the Spat software suite developed by the Institut de recherche et coordination acoustique/musique (Ircam)[17] is used in its MaxMSP version, combined with some elements of the Ambisonics Externals for MaxMSP from the Institute for Computer Music and Sound Technology (ICST) of the Zurich University of the Arts[18].

Spat5.oper, as the central unit of Spat, allows differentiated manipulation of the individual sound sources, derived from Ircam's research on perceptual properties. In addition to three-dimensional panning and directivity parameters, it includes parameters such as source presence, warmth and brilliance, room presence, running reverberation and envelopment. Each source can be linked to a room simulation where early reflections, cluster reflections and diffuse reverberation are designed. Instead of choosing contingent source and room parameters, *dispersion.spat* bases them on the analysis of Room Impulse Response (IR) measurements, which are then used in Spat5.oper's reverberation enginge to mimic the acoustic qualities of local and remote rooms. In addition, it proposes an option to interpolate between different room characteristics derived from IRs, so that it is possible to zoom into the different acoustic characteristics of the connected rooms. Finally, a feature for applying geometric trajectories to the individual sources is integrated, technically based on the Ambimonitor of ICST's Ambisonics Externals for MaxMSP and aesthetically based on the developer's previous work with Pauline Oliveros and her EIS system [75].

One of the most aesthetically intriguing aspects of *dispersion.spat* is the idea that room values can be manipulated in real-time, allowing flexible changes in room qualities, e.g. the interpolation between acoustic rooms measured by impulse responses. Additionally, it allows the control of trajectories of individual sound sources. Analysing the IRs and using them in an algorithmic reverberation engine is computationally efficient and allows flexible handling. Like that of *VRR*, *dispersion.spat*'s star-shaped streaming topology does not seem to be optimised in terms of latency, as the signals are auralised on a central server located at the same site as the JackTrip server in hub mode. However, an alternative approach is proposed using *dispersion.spat* instances at each player location, where the room parameters are then coordinated via the exchange of OSC messages between instances. *dispersion.spat* is currently not publicly available.

C. Coordination of room parameters by prior agreement

If coordination between the room parameters at the different location is desired and if the auralisation is not computed on a central server, the relevant parameters must be exchanged between individual locations. If the room simulation in the Ambisonics domain is implemented with an algorithmic simulation, these algorithms can be exchanged in real-time, e.g. by exchanging OSC messages between the locations (see Section III-D below). However, this is not possible when using convolution techniques, where the sound signal is convolved with IRs measured in each of the players' environments (or in any other environment). In this case, the IRs must be swapped and implemented prior to the performance. This is the approach taken by the *VIIVA-NMP audio system*. Aesthetically, this corresponds to the Virtual Space Approach (VSA).

VIIVA-NMP Audio System [76] [77] [71] [78]: The system is based on a parent application called "Vocal Interaction in an Immersive Virtual Acoustic (VIIVA)", which allows the user to participate in a pre-recorded group singing experience by taking on the role of a missing singer in the group. Immersive audio rendering provides a performance scene from the perspective of each singer. For a series of telematic performances of spatially distributed vocal duos from home networks, the system was adapted for Network Music Performance (NMP) using the JackTrip protocol (*VIIVA-NMP (1)*). Auralisation was based on spatial room impulse responses (SRIRs) recorded with a 1st order Ambisonics microphone. Three different rooms were captured, and SRIR recordings were arranged to correspond to a virtual placement of the singers in each room. The singers' direct sounds were individually convolved with the respective SRIRs (using MCFX-convolver[19]) at each location. The performers' locations in the virtual acoustic scene were manually specified using the configuration files, then binaurally rendered with ambiX binaural decoder[20]. No additional sound processing or head-tracking was applied (although the latter was technically implemented). A series of performances were run with the three room characteristics, achieving one-way latencies between 23 and 55 ms [71].

In a further development, the system was extended for an XR application involving a four-piece rock band (*VIIVA-NMP (2)*). Here, the auralisation model was significantly extended: a larger number of SRIRs, now captured with a 3rd order Ambisonics microphone, were required to cover the positions in the virtual environment. Three parallel spatialisation processes were applied: 1. placement of the direct sources in the virtual scene with ambiX encoder; 2. convolving early reflections of each source with (the first part of) individual SRIRs; 3.

[16] https://github.com/dispersionlab/disperf

[17] https://forum.ircam.fr/projects/detail/spat/

[18] https://www.zhdk.ch/forschung/icst/software-downloads-5379/downloads-ambisonics-externals-for-maxmsp-5381

[19] https://github.com/kronihias/mcfx

[20] SIR files may be manually specified to define the performer locations in the virtual acoustic scene and are loaded into the MCFX-convolver plugin using configuration files.

convolving the late reflections of the summed signal of all sources with (the diffuse reverb part of) a single SRIR (using X-MCFX-convolver[21]). All scenes are then summed into a single complete scene so that rotation for head-tracking can be applied (using IEM's SceneRotator[22]), avoiding the need to interpolate between discrete HRTFs for each discrete source. Binaural decoding was achieved using the ambiX binaural decoder and HRTFs from the SADIE II database[23]. Reasonable one-way latencies of around 23 to 26 ms were achieved with the use of the SonoBus streaming software [77].

Among the utilities reviewed, VIIVA-NMP proposes the most advanced system using room impulse responses in the context of online jamming. To mitigate the high computational burden of using multiple SRIRs, splitting them into early reflection and diffuse reverberation is a viable option and should not impose perceptual limitations. At the same time, measuring SRIRs is a complex task that will not be a standard for online jamming outside of academic institutions. It is, of course, possible to use existing SRIR libraries such as Open AIR[24] or the SADIE II database, but even then, the selection and subsequent splitting of the IRs and the application of correct delay lines would be a challenge when applied to a concrete ensemble of players. Furthermore, scaling options might be limited as they lead to an increase in the necessary SRIRs and the computational burden.

D. Real-time coordination of room parameters using OSC

Unlike the use of IRs, which require prior exchange, algorithmically generated room parameters open up the possibility to be exchanged in real-time. In the present cases, this is done using OSC messages. The streaming applications therefore contain an OSC bridge, which allows OSC data to be streamed in addition to audio data. Aesthetically, besides a Virtual Space Approach (VSA), a Dynamic Space Approach (DSA) or a Coupled Room Approach (CRA) is possible in some cases.

OVBOX [21]: The *ORLANDOviols Consort Box (OVBOX)* is another open source, audio-only application with a similar conceptual approach to *VRR*: distributed musicians gathering in a shared virtual acoustic environment. It relies on zita-njbridge[25] as the streaming protocol, and to optimise latency, the application is implemented in C++ on a Linux mini-computer (Raspberry Pi; *OVBOX RP*), although there is also a standalone device for desktop computers (*OVBOX DT*). The central piece is a configuration server, implemented on the local hardware, which coordinates the virtual room acoustics that are computed on the local machine and coordinated with the other locations via OSC messages. The acoustic simulation is based on a geometric acoustic simulation tool, TASCAR[26], which places each individual source in a 3D space with the dimensions azimuth and distance, surrounded by a shoebox room model whose dimensions can be manipulated by the user in a browser interface. Reverberation is simulated using a feedback delay network, but operates in the 1st order ambisonics domain, with rotation operations on each reflection filter. The application includes optional head-tracking. It is publicly available, either from the author's personal website[27] or from *Digital Stage*[28].

The *OVBOX* has chosen an approach that does not rely primarily on spherical harmonics, but on a ”direct” rendering from the virtual source position to the binaural. The optimisation of the hardware through the construction of an embedded system makes it possible to achieve ultra-low latencies that are not comparable with the other systems mentioned in this review. At the same time, compared to some of the other utilities, it is limited to the Virtual Space Approach (VSA) and does not allow more experimental strategies. Although the room parameters are exchanged via OSC messages, a modification in real-time is not applicable (and not intended). In addition, the fact that players have to assemble their own Raspberry Pi may deter users; those opting for the desktop version will find a reduction in the benefits of ultra-low latency.

IRENE [72]: As the extended version of the *VIIVA-NMP audio system*, the *Immersive Room ExtensioN Environment (IRENE)* is an XR application, and similar to *VIIVA-NMP (2)*, it also works with avatars as visual representations of the players, although it only recreates a one-to-one situation. The immersive audio part relies heavily on the IEM plug-in suite: directivity patterns are calculated for the individual sources (DirectivityShaper), then discrete position-dependent early reflections are generated as a function of room size, wall properties and listening position (RoomEncoder). In parallel, an additional acoustic simulator (MCRoomSim[29]) based on a ray-tracing method [79] pre-simulates diffuse reverberation that is convoluted with the direct sounds. All signals are then summed to an ambisonic scene, rotated according to head movement (SceneRotator) and decoded binaurally (BinauralDecoder). In this environment, the performers – one on each side – are tracked and can move in 6 degrees of freedom (6DoF). The system, which is implemented in MaxMSP and exchanges room properties through OSC data, has only been tested on a local area network using Dante[30], where it achieved one-way-latencies of 13 to 30 ms, depending on the buffersize. Nevertheless, it is proposed to be implemented on wide area networks as well.

IRENE is particularly interesting from an aesthetic point of view, as it implements a Coupled Room Approach (CRA), where the characteristics of the rooms involved are blended together as if they were coupled through an aperture. Another important feature is the focus on directivity patterns, which

[21] http://www.angelofarina.it/X-MCFX.htm

[22] https://plugins.iem.at/

[23] https://www.york.ac.uk/sadie-project/database.html

[24] https://www.openair.hosted.york.ac.uk/?page-id=2

[25] https://github.com/digital-stage/zita-njbridge

[26] https://github.com/gisogrimm/tascar

[27] https://github.com/gisogrimm/ovbox

[28] https://digital-stage.org/proben-und-auffuehren/per-ov-box

[29] https://github.com/Andronicus1000/MCRoomSim

[30] https://www.getdante.com

plays an important role in 6DoF reproduction, taking into account the acoustic behaviour of moving sound sources. However, extending the one-to-one situation to a multi-player jamming situation (which is not the intention of *IRENE*) creates obstacles: each source has to be encoded and processed with an individual DirectivityShaper and RoomEncoder, which leads not only to complex handling and coordination problems between the locations, but also to a considerable computational load. Depending on the hardware, this can quickly lead to the need for large buffer sizes and thus high latency.

Zone [25]: As the extended version of *VIIVA-NMP (2)*, *Zone* is an XR application. Like *IRENE*, it is designed for a telemersive one-to-one encounter. And like *dispersion.spat*, it utilises Ircam's Spat5 application, but with one instance at each location. Telemersive Toolkit, a MaxMSP wrapper developed by our research group [36], was used as the streaming device. It allows for the streaming of audio and video data (based on the UltraGrid protocol), of OSC and motion capture data and of other control data (via the integration of Open Stage Control[31]). In the performance, a visitor – together with their dialogue partner from the remote location – wanders through various virtual interior and exterior spaces. The respective scene also corresponds to an acoustic environment, which is played back binaurally via headphones and changes over the course of the piece. All spatial audio processing is done with Spat5.

As performance and audience move through different acoustic spaces in 6DoF, perceptive aspects like early reflections and directivity patterns of the sound sources are of utmost importance to ensure that the situation would not be perceived as unnatural. In contrast to IEM's RoomEncoder, which takes into account the spatial relationship between the sound source and the listener, this relationship in this case had to be additionally calculated using a transformation matrix, computed inside MaxMSP. The advantage of this method is that it can be easily scaled to multiple locations and performers. In contrast to *IRENE*, all sound sources can be summarised in a single sound scene here, which increases clarity and computational efficiency.

Tpf-jam-tool[32]: The tool has been designed with the intention of packaging our research group's experience with immersive audio and its powerful potential in telemersive performance contexts into an easy-to-use tool for online jamming. The local and remote sound sources are locally encoded into an Ambisonics scene using the AmbiEncoder plugin of the ICST-Ambisonics-Plugin-Suite[33]. Ambisonic reverberation is added to the scene using the FdnReverb plugin from the IEM suite, its SceneRotator is used for head-tracking and finally the binaural rendering is done using the IEM's BinauralDecoder. The tool is implemented in Reaper DAW as the host system from which the streaming utility – the Tpf-client based on the AOO protocol[34] – can be launched and controlled, and it is possible to select pre-configured templates for ensembles of 2 to 16 players, including routing between the different software. These functionalities are implemented using Reaper's scripting options, namely Lua.[35] In its standard version, Reaper allows for the independent design of the acoustic scene at each location. However, since the AmbiEncoder and FdnReverb plugins can both send and receive OSC messages, the Tpf-client includes an OSC bridge that allows relevant data to be exchanged between the locations. In addition, the positioning of the players in the Ambisonics scene can be modified at each location to give the central perspective of the local player, using the transformation matrix mentioned above in the description of *Zone*.

TABLE II
LATENCY MEASUREMENTS OF STREAMING UTILITIES, IN MILLISECONDS FOR DIFFERENT BUFFER SIZES (BS)

Utility	Mode	64BS	128BS	256BS	512BS
JackTrip 2.3.0	P2P	5.33	6.66	21.33	42.66
	Hub	13.33	26.66	53.33	106.66
SonoBus 1.7.2	P2P	n/a	13.33	18.66	53.33
UltraGrid 1.9	Standard	109.33	85.33	85.33	96.00
	Ultra	10.66	18.66	21.33	42.66
Tpf-client 2.0-b15	P2P	10.66	12.00	16.00	25.33

Tpf-jam-tool takes a modular approach, relying on different plug-in suites for spatial audio processing according to their sound quality and latency behaviour. With the integration of the DAW Reaper and the streaming utility Tpf-client, it is specifically streamlined for online jamming with different ensemble sizes. However, its open and modular architecture allows it to be extended for more individual applications. On a perceptual level, the room simulation is limited to distance encoding (in the AmbiEncoder plugin) and an Ambisonics FDN reverb, so no early reflections or directivity patterns are computed.

The panorama of tools and projects collected in this chapter impressively demonstrates the variety of solutions proposed for the trade-off between perceptual aspects of binaural hearing, technical workflows and aesthetic decisions. The different models may provide orientation for the design of further projects and tools in the field of telemersive performance.

IV. LATENCY MEASUREMENTS

Due to different motivations of the project authors and tool designers, ultra-low latency is not a primary goal in all of the above approaches. Some focus on the detailed design of tonal and spatial sound parameters, knowing full well that high and immersive sound quality can also lead to greater latency tolerance. In other projects, the synchronisation of audio and video components is more important than low audio latency. Issues of usability or hardware limitations, especially CPU consumption, are other aspects that can counteract latency

[31] http://openstagecontrol.ammd.net/

[32] https://github.com/jschuett/tpf-jam-reaper-tool

[33] https://github.com/schweizerweb/icst-ambisonics-plugins/wiki

[34] https://github.com/zhdk/tpf-client

[35] https://www.reaper.fm/sdk/reascript/reascript.php

reduction. Nevertheless, good latency performance remains an important goal in the design of tools that combine telematic performance practices with immersive audio. From a practical point of view, it can be said that low latency expands artistic possibilities: reducing an existing threshold is in some cases not possible or only possible with great effort, while artificially adding latency, on the other hand, is always an option. The specificity of the telematic medium also invites practitioners to accept latency as an artistic resource [5] [80] [33].

Latency measurements therefore provide important clues when it comes to building tools to facilitate telematic performances with immersive audio. Referring to a recent study by Turchett and Tomasetti [39], the total audio latency path Γ from a musician acting as a sender to a musician acting as a receiver is composed as follows:

$$\begin{aligned}\Gamma = {} & \lambda_{\text{ADC}} + \lambda_{\text{audio-buffer}} + \lambda_{\text{packetization}} \\ & + \lambda_{\text{network}} + \lambda_{\text{jitter-buffer}} + \lambda_{\text{depacketization}} \\ & + \lambda_{\text{spatial-audio}} + \lambda_{\text{DAC}}\end{aligned}$$

The latency for the conversion of analogue to digital signals, λ_{ADC} and λ_{DAC}, is highly dependent on the hardware used. The behaviour of the latter can be influenced by building an integrated system – e.g. using a Linux mini-computer as in *OVBOX* – or by optimising desktop computers by giving real-time priority to the signal processing threads of audio software, as proposed for the desktop version of *OVBOX* on Linux[36]. However, this is beyond the scope of this study. Another highly variable parameter is the latency of the network, λ_{network}. Depending on its quality of service, it also affects the size of the jitter buffer, $\lambda_{\text{jitter-buffer}}$, used to compensate for network jitter. Finally, the audio buffer, $\lambda_{\text{audio-buffer}}$, due to the acquisition of the system's digital signal, depends on the performance of the hardware used as well as the computational requirements imposed by the processing of the spatial audio tools, $\lambda_{\text{spatial-audio}}$. The settings for both jitter and audio buffer have a major impact on the latency behaviour.

To enable comparability, two aspects were therefore moved centre stage and measured: firstly, the latency of the streaming utility $[\lambda_{\text{packetization}} + \lambda_{\text{network}} + \lambda_{\text{jitter-buffer}} + \lambda_{\text{depacketization}}]$ as a function of $\lambda_{\text{audio-buffer}}$ and with optimised λ_{network} and minimised jitter. Secondly, the latency of $\lambda_{\text{spatial-audio}}$ was measured, which in turn is divided into the aspects of encoding, room simulation, scene rotation and binaural decoding:

$$\begin{aligned}\lambda_{\text{spatial-audio}} = {} & \lambda_{\text{encoding}} + \lambda_{\text{roomsim}} \\ & + \lambda_{\text{scene-rotation}} + \lambda_{\text{binaural-decoding}}\end{aligned}$$

The methodology of the measurements is explained in the following two subsections; some of the results can be found in Table II and III. It should be noted that the measurements do not reflect the use of the tools in real-world contexts, where many factors such as sound card hardware, clock synchronisation, buffer size, internet quality of service, etc. add additional latency. Nevertheless, experience shows that it is helpful to look at the discrete elements of the data flow and use comparisons of these discrete elements to identify relevant latency factors and to optimise overall behaviour.

TABLE III
LATENCY MEASUREMENTS OF SPATIAL AUDIO PLUG-INS IN SAMPLES
*ONE MEASUREMENT FOR ALL PARAMETERS IN THE CASE OF SPAT5

Parameter	Suite	Utility	Subparam	samples
λ_{encoder}	IEM	MultiEncoder		0
		DirectivityShaper		0
	ambiX	ambiX encoder		0
	ICST	AmbiEncoder		0
	Sparta	ambiENC		64
λ_{roomsim}	IEM	RoomEncoder		396
		FdnReverb		0
	ambiX	MCFX	York-tof	37
			York+tof	566
			Maida Vale	152
	Sparta	ambiRoomSim		352
		matrixconv	York-tof	4133
	IRCAM	Spat5		48*
$\lambda_{\text{rotation}}$	IEM	SceneRotator		0
	ambiX	ambiX rotator		0
	Sparta	rotator		64
$\lambda_{\text{decoding}}$	IEM	BinauralDecoder		126
	ambiX	binaural	Icosaheder	4
			SADIE	92
		MCFX	ViveCinema	12
	DearVR	AmbiMicro		24
	Sparta	ambiBIN		1536

A. Latency measurements of streaming utilities

Table II gives an overview of the measured roundtrip times (RTT) for the streaming utilities used in the systems presented in Chapter III. In the test design, two computers – a MacBook M3 running MacOS 12.6.3 and a Linux laptop running Ubuntu 24.04 – were the clients on each side. To minimise network latency and jitter, we connected them on a local network, including the server on the local infrastructure where applicable. We used the Jack Audio Connection Kit[37] for routing, and round-trip latency was measured using the Jack-delay utility[38]. A sample rate of 48kHz was selected. Jitter buffers were set as low as possible ('receive buffer' 2 in Tpf-client; 'queue buffer length' 2 in JackTrip, 'minimum jitbuffer' in SonoBus). JackTrip was run alternately in peer-to-peer and hub mode. UltraGrid was run alternately with its default parameters and with the 'low-latency-audio ultra' option. SonoBus in its MacOS version does not support Jack; therefore, a second Linux laptop running Ubuntu 22.04 was used for testing in this case. In this setup, no clean audio transmission could be achieved with the buffer size of 64 samples. We did not measure *OVBOX*' zita-njbridge or *VRR*'s AOO, as it builds on an old version.

In all tests uncompressed audio (PCM) was streamed. However, most utilities allow the use of audio codecs at the cost of additional latency. The Opus codec is considered to be one of the most efficient [81], and it may be indicated in the case of bandwidth limitations. In addition, the Opus codec

[36] https://github.com/gisogrimm/ovbox/wiki/LinuxLatency

[37] https://jackaudio.org/

[38] http://kokkinizita.linuxaudio.org/linuxaudio/

natively includes forward error correction, which may improve audio quality in some environments.

As the results show, JackTrip in peer-to-peer mode performs best at low buffer sizes (64 and 128 samples), while Tpf-client performs best at higher buffer sizes (256 and 512 samples). JackTrip's hub mode adds a considerable amount of latency, although its performance is still impressive at low buffer sizes. However, network topologies with a central server can have a major impact on latency, as signals from all sites have to pass through this central node. UltraGrid needs to be run in 'low latency audio' mode to be useful in latency sensitive environments, but this is at the expense of lower reliability.

B. Latency measurements of spatial audio plugins

Table III gives an overview of the latencies induced by the spatial audio plug-ins (SAPs) used in the systems presented in Chapter III or mentioned in the relevant literature. The measurements repeat a similar study by Tomasetti, Farina and Turchet [38] and extend it to include some of the research desiderata proposed there. One of these is the impact of SAP processing latencies on different machines and with different host programs. Another is to measure latencies with different IRs in the room simulations and with different filter matrices on the SAPs dealing with binaural decoding.

The test design used a MacBook M1 running macOS 13.4. We chose Reaper as the host software, with the exception of Spat5, which was run on MaxMSP. Alternatively, we also tested the SAPs on MaxMSP and did not find any divergence to the measurements with Reaper. Where applicable, measurements were made with 3rd order Ambisonics signals. To measure the processing latencies, we used an impulse as the audio signal (Dirac's Delta), recorded the output of the signal separately for each plugin and analysed the latency in samples. We tested the SAPs sequentially with buffer sizes of 64, 128, 256 and 512 samples, and found no differences in their latency behaviour (with the exception of some of the Sparta plug-ins, which do not work with buffer sizes below 128 samples).

A summary of the results is presented in the next section, following the categorisation presented at the beginning of Chapter IV.

$\lambda_{\textbf{encoding}}$: With the exception of the Sparta encoder, which has a latency of 64 samples, all Ambisonic encoders operate with zero latency. In the case of IEM's DirectivityShaper, the Ambisonics encoding represents the directivity and orientation of the sound source, resulting in an encoded multichannel stream that is usually referred to as O-format [82]. In addition, ICST's AmbiEncoder includes some elements of room simulation, namely distance encoding.

$\lambda_{\textbf{roomsim}}$: Of the room simulation plugins, only IEM's FdnReverb does not add any latency, which is hardly surprising, as it is a pure reverb plugin in the Ambisonics domain. The algorithmic SAPs such as IEM's RoomEncoder or Sparta's ambiRoomSim, on the other hand, introduce significant latency in a dimension that is relevant in the NMP context.

It is important to note that, as far as convolution-based SAPs are concerned and compared with the benchmark study by Tomasetti, Farina and Turchet [38], a new situation has arisen with regard to Matthias Kronlachner's ambiX and MCFX plugins. With their new release in version v0.3.0 (ambiX) and v0.6.3 (MCFX), published at the end of 2023, they adopt an approach proposed by Angelo Farina in his X-MCFX plugin fork[39]: instead of hard-coding the internal buffer at 512, the new versions now allow to work in "zero latency" mode, where the plugin's buffer is aligned with the buffer of the host system. Both amiX and MCFX SAPs now include this functionality.

However, these SAPs do not work without latency – this is due to the IRs used for reverberation, as the pre-ringing of the direct sound is highly dependent on the length of the IRs. Our measurements are therefore consistent with the comparative study [38], where X-MCFX reached a latency of 37 samples with an IR from the OpenAIR IR library[40], a dimension now also measured with the new version of the MCFX-plugin. However, we removed the time-of-flight (530 samples) of the IR from the Openair library before convolution (OpenAIR-tof); without this removal, the latency is much higher (OpenAIR+tof). Applying the SIR measured at the Maide Vale studios, as suggested by the *VIIVA-NMP (2)*, the induced latency is again significantly higher, even after truncating the time-of-flight (130 samples) and only using the first part of the SRIR. The Sparta room simulation plugins seem not to be suitable for these contexts as their latencies are high, notably for convolution, which is CPU intense and only works with high buffer sizes.

$\lambda_{\textbf{scene-rotation}}$: With the exception of the Sparta scene rotator, which has a latency of 64 samples, all scene rotators in the Ambisonics domain work with zero latency.

$\lambda_{\textbf{binaural-decoding}}$: Regarding the algorithmic binaural decoders, the high latency of the Sparta decoder is noteworthy, while dearVR's AmbiMicro plugin[41] shows the best performance when compared to a higher value for IEM's binaural decoder. The same situation as used in the room simulation is repeated for the convolution-based decoders, depending on the filter matrices (HRIRs) taken from the ambiX binaural decoder pre-coded settings (Icosaeder), from the SADIE database (as proposed by *VIIVA-NMP*) or as used in [38] (ViveCinema-Ambix2Bin.wav[42]).

In the case of tools and projects that use IRCAM's Spat5 (*dispersion.spat, Zone*), it is ineffective to split up the subprocesses of the spatial audio processing, as the corresponding algorithms largely summarise them. Spat5 is optimised especially for live contexts [83], and this is reflected in our measurements in an extraordinarily favourable latency behaviour.

The latency measurements of the SAPs largely confirm the findings of the comparative study [38], and they seem

[39] http://www.angelofarina.it/X-MCFX.htm

[40] https://www.openair.hosted.york.ac.uk/?page%20id=502

[41] https://www.dear-reality.com/products/dearvr-ambi-micro

[42] http://www.angelofarina.it/Public/Xvolver/Filter-Matrices/Ambix-2-Binaural/

independent from hardware (MacBook or Dell Alienware), operating systems (Apple M1 or Windows 10) or host systems (Reaper, MaxMSP or Plogue Bidule[43]). In summary, Ambisonic processing through channel gains (i.e. encoding, rotation) can typically be computed with zero latency. Real-time geometric acoustics are computationally more expensive in comparison and require processing buffers, which add latency. In contrast, Ambisonic convolution can be computed with zero additional processing latency; however, additional latency may be part of the impulse response used in convolution, and the measurements show that the time-of-flight and length of IRs must be carefully considered. Finally, the use of Spat5 as a spatial audio engine for telemersive performances seems a viable option, given its low processing latency, if there is a willingness to work within the MaxMSP environment.

V. Conclusion

This paper analyses a series of projects and tools for telematic performances, including immersive audio, in terms of space and time domains. In order to mitigate important aspects of this format such as perceptual quality, ease of use, CPU consumption or latency behaviour, these projects and tools follow different paths in terms of network topologies, choice of streaming protocols, coordination of metadata for spatialisation and utilities for spatial audio processing with reference to algorithmic or convolution-based methodologies. In addition, the examples show clear choices in terms of aesthetic aspects that include space as a causal factor in telemersive performance practices. Some of the findings are summarised in Table I. They can be helpful for the creation of further tools and projects in this area, and indeed, they have been for our own exploration into developing appropriate tools and protocols in XR contexts (*Zone*) or for online jamming purposes (*Tpf-jam-tool*).

The analysis also includes latency measurements. The results are summarised in Table IV. For the compilation, a distinction was made, where possible, between the measurement of the SAPs and the measurement of the streaming utility used in the individual tools and projects, as measured in Chapter IV. This is not possible for *digital-space-web* and *OVBOX* due to their integrated systems, so only their total round trip latency through the whole system is given here. As the measurements for *digital-web* and *OVBOX* in its Raspberry Pi version would require a different methodology from the one proposed in this study, we refer to the information provided by the developers. These are therefore not directly comparable with the other measurements and are shown in brackets in Table IV. The information might still be useful as a guide. In the case of projects using Spat5, only the latency of the entire spatial audio processing is given. For the streaming utilities, we have chosen the variant that is central to the project in question. For the sake of comparability, we have chosen the values for a block size of 126 samples, as smaller block sizes are not realistic for most hardware environments in the context of online jamming.

The SAP values refer to one-way latencies. From a perceptual perspective and in a bidirectional setting, the audio signal has to be auralised on each side before it reaches the ear of the respective performers. The values therefore have to be doubled for the round trip time (RTT). The total latency results from the sum of twice the value of the SAP latency (except for *dispersion.spat*, where this process only takes place once on the central server) and the value of the streaming utility measurement.

As the values demonstrate, the use of immersive audio in online jamming practices is within the realms of possibility, even in latency-sensitive environments. However, the measured values are largely theoretical for the time being; the specific hardware and software components, block size settings, internal routing and other hardware latencies, such as analogue-to-digital conversion, will play a significant role. When measuring the streaming utilities, the propagation time of the internet was ignored and jitter was kept extremely low; both do not correspond to real conditions and are a source of considerable latencies in telematic practice. These are even more pronounced in network topologies that choose an approach with a central server. Even if the comparison of the various tools and projects under real-world conditions remains a desideratum, the measurements do provide some indications of what needs to be taken into account.

Besides placing sound sources in the virtual acoustic scene, two fundamentally different approaches are chosen for the room simulation: geometric acoustics or the use of impulse responses. Perceptually, the latter produces more natural results, but the collection and handling of IR data is complex and quickly becomes computationally expensive when scaled. To overcome this difficulty, tools that incorporate convolution techniques usually distinguish between direct reflections and diffuse reverberation. *VIIVA-NMP* uses SRIRs for both, convolving the former with the individual sound sources and the latter with the whole sound scene only in favour of computational efficiency. *IRENE* combines algorithmic early reflections with diffuse room responses derived from a ray-tracing room model. *dispersion.spat* takes the opposite approach: it feeds the room simulation engine, based on Fdn reverb, with parameters derived from IR measurements of actual performance spaces. Finally, the authors of *VIIVA-NMP (2)* propose another hybrid approach with convoluted SRIRs for early reflection and feedback delay networks for diffuse reverberation as a research desideratum.

Working with SRIRs also leads to difficulties with movement, whether of the listener's head or of moving sound sources, in which case interpolations between different SRIRs would be necessary. Their absence may be perceptually negligible in static sound scenes, but tools that allow 6DoF have other affordances and should include the audio object's directivity pattern so that sound intensity and spectral changes may be simulated convincingly upon variation in its orientation relative to the position of the listener [84]. Approaches presented in *IRENE*, *VIIVA-NMP (2)* and *Zone* transcend the idea of online jamming towards acoustic practices in the Metaverse.

[43] https://www.plogue.com/products/bidule.html

TABLE IV
TOTAL LATENCY IN ROUND-TRIP TIME (RTT) OF TOOLS FOR ONLINE JAMMING WITH IMMERSIVE AUDIO (ACCORDING TO TABLE II AND TABLE III)

Tool	$\lambda_{\text{encoding}}$	λ_{roomsim}	$\lambda_{\text{rotation}}$	$\lambda_{\text{decoding}}$	SAP	Streaming	RTT
	latency in samples				latency	latency	in ms
digital-stage-web	–						(120 ms)
dispersion.spat	spat5.oper 48 s				48 s / 1 ms	JackTrip 27 ms	28 ms
OVBOX DT	-						(29 ms)
OVBOX RP	-						(10 ms)
VIIVA-NMP (1)	-	MCFX 37 s	-	ambiX-binaural 4 s	41 s / 1 ms	JackTrip 7 ms	9 ms
VIIVA-NMP (2)	ambiX-encoder 0 s	MCFX 152 s	SceneRotator 0 s	ambiX-binaural 92 s	244 s / 5 ms	SonoBus 13 ms	23 ms
IRENE	DirectivityShaper 0 s	RoomEncoder 396 s	SceneRotator 0 s	ambiX-binaural 4 s	400 s / 8 ms	(Dante) -	(16 ms)
Zone	spat5.oper 48 s				48 s / 1 ms	TTkit 19 ms	21 ms
Tpf-jam-tool	AmbiEncoder 0 s	FdnReverb 0 s	SceneRotator 0 s	BinauralDecoder 126 s	126 s / 3 ms	Tpf-client 12 ms	18 ms

Advanced IR-techniques will not be practical for an online jamming tool which should flexibly adapt to different ensemble sizes and players at home networks. For this specific use case, it is preferable that the individual sound sources need not to be simulated individually on individual tracks but can be integrated in a multi-encoder graphic interface. *dispersion.spat, Zone* and *Tpf-jam-tool* offer such a functionality.

Another avenue for future work concerns the involvement of musicians to evaluate the qualities of SAPs from the point of view of musicians' perception. Some studies exist, particularly for binaural decoders [85] [86] [87]. At the same time, a systematic comparison of room simulation SAPs would be difficult to achieve, as they vary widely across multiple parameters or differ in terms of the IRs chosen. Aesthetic decisions in this area are largely left to the artistic experience of the practitioner.

ACKNOWLEDGMENT

The research for this paper is funded by the Swiss National Science Foundation and took place at the Institute for Computer Music and Sound Technology and the Immersive Arts Space of Zurich University of the Arts as part of the Telematic Performance Format research unit and its project Spatial Dis/continuities in Telematic Performances. We thank our diverse collaborators as well as the further members of the research team, besides the authors, Benjamin Burger, Martin Fröhlich, Joel De Giovanni, Eric Larrieux, and Hannah Walter.

REFERENCES

[1] Eric C Lemmon. Telematic Music vs. Networked Music: Distinguishing Between Cybernetic Aspirations and Technological Music-Making. *Journal of Network Music and Arts*, 1(1), 2019.

[2] Juan Pablo Cáceres and Chris Chafe. Jacktrip: Under the hood of an engine for network audio. *Journal of New Music Research*, 39(3):183–187, 2010.

[3] Carlo Drioli, Claudio Allocchio, and Nicola Buso. Networked performances and natural interaction via LOLA Low latency high quality A/V streaming system. In *Proc. 2nd Int. Conf. Inf. Technol. Perform. Arts, Media Access, Entertainment*, pages 240–250, 2013.

[4] Peter Holub, Jiri Matela, Martin Pulec, and Martin Srom. UltraGrid: low-latency high-quality video transmissions on commodity hardware. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1457–1460, 2012.

[5] Alexander Carôt and Christian Werner. Network Music Performance – Problems, Approaches and Perspectives. In *Music in the Global Village*, page 13, Budapest, 2007.

[6] Peter Fischer. Case Study: Performing Band Rehearsals on the Internet With Jamulus.

[7] Roman Haefeli, Johannes Schütt, and Patrick Müller. TPF-Tools. A Multi-Instance Jacktrip Clone. In *Proceedings of the 17th Linux Audio Conference*, 2019.

[8] Alan B. Tyson, Scott Deal, and Kenneth Fields. Artsmesh: An Incremental Development in Telematic Art. 2016.

[9] Gil Weinberg. Interconnected musical networks: Toward a theoretical framework. *Computer Music Journal*, 29(2):23–39, 2005.

[10] Alvaro Barbosa. Displaced SoundScapes. A Survey of Network Systems for Music and Sonic Art Creation. *Leonardo Music Journal*, (13):53–59, 2003.

[11] Roger Mills. *Tele-Improvisation. Intercultural Interaction in the Online Global Music Jam Session*. Springer International Publishing AG, Cham, 2019.

[12] Alvaro Barbosa. *Computer-Supported Cooperative Work for Music Applications*. PhD thesis, Pompeu Fabra University, 2006.

[13] Jonas Braasch. The Telematic Music System: Affordances for a New Instrument to Shape the Music of Tomorrow. *Contemporary Music Review*, 28(4-5):421–432, 2009.

[14] Chris Chafe. Living with Net Lag. In *AES 43rd International Conference*, pages 1–6, 2011.

[15] Jonas Braasch, Chris Chafe, Pauline Oliveros, and Doug Van Nort. Mixing-Console Design Considerations for Telematic Music Applications. In *Audio Engineering Society Convention*, 2009.

[16] Ren Gang, Samarth H Shivaswamy, Stephen Roessner, Akshay Rao, Dave Headlam, and Mark Bocko. Audio Latency Masking in Music. In *Audio Engineering Society Convention*, 2012.

[17] Helen Thorington. Breaking out: The trip back. *Contemporary Music Review*, 24(6):445–458, 2005.

[18] Peter Traub. Sounding the net: Recent sonic works for the internet and computer networks. *Contemporary Music Review*, 24(6):459–481, 2005.

[19] Michael Gurevich, Dónal Donohoe, and Stéphanie Bertet. Ambisonic Spatialization for Networked Music Performance. In *International Conference on Auditory Display*, 2011.

[20] Jonas Braasch, Daniel Valente, and Nils Peters. Sharing Acoustic Spaces over Telepresence using Virtual Microphone Control. *123rd Convension of Audio Engineering Society*, 2007.

[21] Giso Grimm, Angelika Kothe, and Volker Hohmann. Low-delay interactive rendering of virtual acoustic environments with extionsions for distributed low-delay transmission of audio and bio-physical sensor data. *The Journal of the Acoustical Society of America*, 153(3), 2023.

[22] Chris Chafe. Tapping into the Internet as an Acoustical/Musical Medium. *Contemporary Music Review*, 28(4-5):413–420, 2009.

[23] Chris Chafe. I Am Streaming in a Room. *Frontiers in Digital Humanities*, 5(November):1–9, 2018.

[24] Nathan Schuett. *The Effects of Latency on Ensemble Performance*. PhD thesis, 2002.

[25] Benjamin Burger, Joel De Giovanni, Martin Fröhlich, Eric Larrieux, Hannah Walter, and Patrick Müller. The Zone. Technical report, 2024.

[26] Rory Hoy and Doug Van Nort. A technological and methodological ecosystem for dynamic virtual acoustics in telematic performance contexts. In *ACM International Conference Proceeding Series*, pages 169–174, 9 2021.

[27] Jonas Braasch, Nils Peters, Pauline Oliveros, Doug Van Nort, and Chris Chafe. A Spatial Auditory Display for Telematic Music Performances. In Yoiti Suzuki, Douglas Brungart, Yukio Iwaya, Kazuhiro Iida, Densil Cabrera, and Hiroaki Kato, editors, *Application of Spatial Hearing*, number November, pages 436–451. Singapore, 2011.

[28] Matthias Ziegler. Osmosis: Asymmetries in Telematic Performance. *Journal of Network Music and Arts*, 5(1), 2023.

[29] Braxton Boren and Andrea Genovese. Acoustics of Virtually Coupled Performance Spaces. In *Conference on Auditory Display*, 2018.

[30] Henning Piper, Marcel Nophut, Robert Hupke, Stephan Preihs, and Jürgen Peissig. Investigations on Loudspeaker-based Auralization of Immersively Connected Rooms. In *DAGA Jahrestagung für Akustik*, 2020.

[31] Mel Slater. A Note on Presence Terminology. Technical report, 2003.

[32] Franziska Schroeder, Alain Renaud, Pedro Rebelo, and F. Gualda. Addressing The Network: Performative Strategies for Playing Apart. *International Computer Music Conference*, (January), 2007.

[33] Patrick Müller, Matthias Ziegler, and Johannes Schütt. Towards a Telematic Dimension Space. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2019.

[34] Hannah Walter, Robert Torche, Cedric Spindler, Eric Larrieux, and Patrick Muller. kompost*klang*küche*: Telematic WORLDing with Glitch. *Journal of Network Music and Arts*, 6(1), 2024.

[35] Francis Rumsey. *Francis Rumsey, Spatial Audio, Oxford: Focal Press, 2001*. Focal Press, Oxford, 2001.

[36] Martin Fröhlich, Patrick Müller, and Roman Häfeli. Telemersive Toolkit: Exchange Multi Media Streams for Distributed Networked Performance Installations. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, pages 354–360. Association for Computing Machinery (ACM), 6 2024.

[37] Jack Kelly, Wieslaw Woszczyk, and Richard King. Are you there? : A Literature Review of Presence for Immersive Music Reproduction. In *149th Convention Audio Engineering Society*, 2020.

[38] Matteo Tomasetti, Angelo Farina, and Luca Turchet. Latency of spatial audio plugins: a comparative study. In *2023 Immersive and 3D Audio: from Architecture to Automotive, I3DA 2023*, 2023.

[39] Luca Turchet and Matteo Tomasetti. Immersive networked music performance systems: identifying latency factors. In *2023 Immersive and 3D Audio: from Architecture to Automotive, I3DA 2023*, 2023.

[40] Luca Comanducci. Intelligent Networked Music Performance Experiences. In *Springer Briefs in Applied Sciences and Technology*, pages 119–130. Springer Science, 2023.

[41] Valentin Bauer, Dimitri Soudoplatoff, Leonard Menon, and Amandine Pras. Binaural Headphone Monitoring to Enhance Musicians' Immersion in Performance. In *Advances in Fundamental and Applied Research on Spatial Audio*. IntechOpen, 10 2022.

[42] Harvey Fletcher. An Acoustic Illusion Telephonically Achieved. *Bell Laboratories Record*, 11(10), 1933.

[43] Seong Hoon Kang and Sung Han Kim. Realistic audio teleconferencing using binaural and auralization techniques. *ETRI Journal*, 18(1):41–50, 1996.

[44] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Heli Nironen, and Sampo Vesa. Techniques and applications of wearable augmented reality audio. In *AES 114th Convention*, 2003.

[45] Jussi Rämö and Vesa Välimäki. Digital augmented reality audio headset. *Journal of Electrical and Computer Engineering*, 2012.

[46] Gisela Nauck. *Dieter Schnebel. Lesegänge durch Leben und Werk*. Schott, Mainz, 2001.

[47] Stefan Drees. Bill Fontanas urban sound sculptures und die Idee der Relokalisierung von Klängen. In Stefan Drees, Andreas Jacob, and Stefan Orgass, editors, *Musik – Transfer – Kultur. Festschrift für Horst Webe*, pages 459–474. Olms, Hildesheim, 2009.

[48] Chris Chafe, Scott Wilson, Randal Leistikow, Dave Chisholm, and Gary Scavone. A simplified aproach to high quality music and sound over IP. In *COST G-6 Conference on Digital Audio Effects*, 2000.

[49] Chris Chafe, Michael Gurevich, Grace Leslie, and Sean Tyan. Effect of Time Delay on Ensemble Accuracy. In *Proceedings of the International Symposium on Musical Acoustics*, volume 2004, pages 3–6, 2004.

[50] Chris Chafe and Michael Gurevich. Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry. In *Proc. of the 117th Convention of the Audio Eng. Soc.*, 2004.

[51] Alexander Carôt, Christian Werner, and Timo Fischinger. Towards a Comprehensive Cognitive Analysis of Delay-Influenced Rhythmical Interaction. In *International Computer Music Conference*, 2009.

[52] Jasmina Catic, Sébastien Santurette, and Torsten Dau. The role of reverberation-related binaural cues in the externalization of speech. *The Journal of the Acoustical Society of America*, 138(2):1154–1167, 8 2015.

[53] Snorre Farner, Audun Solvang, Asbjørn Saebø, and Peter Svensson. Ensemble hand-clapping experiments under the influence of delay and various acoustic environments. In *Audio Engineering Society Convention*, pages 1–19, San Francisco, 2006.

[54] Jeanette Tamplin, Ben Loveridge, Ken Clarke, Yunhan Li, and David J Berlowitz. Development and feasibility testing of an online virtual reality platform for delivering therapeutic group singing interventions for people living with spinal cord injury. *Journal of Telemedicine and Telecare*, 26(6):365–375, 7 2020.

[55] Yui Ueno, Mitsunori Mizumachi, and Toshiharu Horiuchi. Comparison of subjective characteristics between binaural rendering and stereo width control method. In *Proceedings of International Conference on Technology and Social Science*, volume 2020, 2020.

[56] Matteo Tomasetti and Luca Turchet. Playing with Others Using Headphones: Musicians Prefer Binaural Audio with Head Tracking over Stereo. *IEEE Transactions on Human-Machine Systems*, 53(3):501–511, 6 2023.

[57] Daniel Rudrich and Matthias Frank. Improving Externalization in Ambisonic Binaural Decoding. In *DAGA 2019 Fortschritte der Akustik*, 2019.

[58] Thibaud Leclère, Mathieu Lavandier, and Fabien Perrin. On the externalization of sound sources with headphones without reference to a real source. *The Journal of the Acoustical Society of America*, 146(4):2309–2320, 10 2019.

[59] Durand Begault and Elizabeth M Wenzel. 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*, 49, 2001.

[60] Lorenzo Picinali, David Poirier-Quinot, Alexander Wallin, and Yuli Levtov. Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context. Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context. In *142nd Convention Audio Engineering Society*, 2017.

[61] Ville Pulkki and Toni Hirvonen. Localization of virtual sources in multichannel audio reproduction. *IEEE Transactions on Speech and Audio Processing*, 13(1):105–119, 1 2005.

[62] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and Olivier Warusfel. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acustica united with Acustica*, 99(4):642–657, 7 2013.

[63] Thirsa Huisman, Axel Ahrens, and Ewen MacDonald. Ambisonics Sound Source Localization With Varying Amount of Visual Information in Virtual Reality. *Frontiers in Virtual Reality*, 2, 10 2021.

[64] Lewis Thresh, Calum Armstrong, and Gavin Kearney. A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeaker Rendering. In *143rd Convention of the Audio Engineering Society*, 2017.

[65] Elizabeth M. Wenzel, Frederic L. Wightman, and Doris J. Kistler. Localization with non-individualized virtual acoustic display cues. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 351–359. Association for Computing Machinery, 1991.

[66] Etienne Hendrickx, Mathieu Paquier, Vincent Koehl, and Julian Palacino. Ventriloquism effect with sound stimuli varying in both azimuth and elevation. *Journal of the Acoustical Society of America*, 138(6):3686–3697, 2015.

[67] Tomasz Rudzki, Ignacio Gomez-Lanzaco, Jessica Stubbs, Jan Skoglund, Damian T. Murphy, and Gavin Kearney. Auditory localization in low-

bitrate compressed Ambisonic scenes. *Applied Sciences (Switzerland)*, 9(13), 7 2019.

[68] Frederic L Wightman and Doris J Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustic Society of America*, 105(5):2841–2853, 1999.

[69] Veronique; Pernaux Jean-Marie Jot, Jean-Marc; Larcher. A Comparative Study of 3-D Audio Encoding and Rendering Techniques. In *16th International Conference: Spatial Sound Reproduction*, 1999.

[70] Didier Pinchon and Philip E. Hoggan. Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes. *Journal of Physics A: Mathematical and Theoretical*, 40(7):1597–1610, 2 2007.

[71] Patrick James Cairns. *VIIVA-NMP Audio System: The design of a low latency and naturally interactive Ambisonic audio system for Immersive Network Music Performance*. PhD thesis, University of York, York, 2021.

[72] Robert Hupke, Stephan Preihs Id, and Jürgen Peissig. Immersive Room Extension Environment for Networked Music Performance. In *AES 153rd Convention*, 2022.

[73] Matteo Sacchetto, Paolo Gastaldi, Chris Chafe, Cristina Rottondi, and Antonio Servetti. Web-Based Networked Music Performances via WebRTC: A Low-Latency PCM Audio Solution. *AES: Journal of the Audio Engineering Society*, 70(11):938–950, 11 2022.

[74] Michael Palumbo, Doug Van Nort, and Rory Hoy. Disperf: A Platform for Telematic Music Concert Production. In *International Computer Music Conference*, 2020.

[75] Pauline Oliveros and Panaiotis. The Expanded Instrument System (EIS). In *International Computer Music Conference*, pages 404–407, 1991.

[76] Patrick Cairns, Helena Daffern, and Gavin Kearney. Parametric Evaluation of Ensemble Vocal Performance Using an Immersive Network Music Performance Audio System. *Journal of the Audio Engineering Society*, 10(10):1–10, 2021.

[77] Patrick Cairns, Anthony Hunt, Daniel Johnston, Jacob Cooper, Ben Lee, Helena Daffern, and Gavin Kearney. Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios. *AES: Journal of the Audio Engineering Society*, 71(6):313–325, 6 2023.

[78] Gavin Kearney, Helena Daffern, Lewis Thresh, Haroom Omodudu, Calum Armstrong, and Jude BRERETON Audiolab. Design of an Interactive Virtual Reality System for Ensemble Singing. In *Proceedings of the Interactive Audio Systems Symposium*, 2016.

[79] Andrew Wabnitz, Nicolas Epain, Craig Jin, and André Van Schaik. Room acoustics simulation for multichannel microphone arrays. In *International Symposium on Room Acoustics*, 2010.

[80] Franziska Schroeder. Dramaturgy as a model for geographically displaced collaborations: Views from within and views from without. *Contemporary Music Review*, 28(4-5):377–385, 2009.

[81] Ben Lee, Tomasz Rudzki, Jan Skoglund, and Gavin Kearney. Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality. *AES: Journal of the Audio Engineering Society*, 71(4):145–154, 2023.

[82] Dylan Menzies. W-panning and O-format, tools for object spatialisation. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[83] Thibaut Carpentier, Markus Noisternig, and Olivier Warusfel. Twenty Years of Ircam Spat: Looking Back, Looking Forward. In *41st International Computer Music Conference*, pages 270–277, 2015.

[84] Jean-Marc Jot, Rémi Audfray, Mark Hertensteiner, and Brian Schmidt. Rendering Spatial Sound for Interoperable Experiences in the Audio Metaverse. In *International Conference on Immersive and 3D Audio*, 2021.

[85] Gregory Reardon, Andrea Genovese, Marta Gospodarek, Juan Simon Calle, Gabriel Zalles, Marta Olko, Christal Jerez, Patrick Flanagan, and Agnieszka Roginska. Evaluation of Binaural Renderers: A Methodology Convention e-Brief 168 Evaluation of Binaural Renderers: A Methodology. In *143rd Convention of the Audio Engineering Society*, 2017.

[86] Gregory Reardon, Andrea Genovese, Gabriel Zalles, Patrick Flanagan, and Agnieszka Roginska. Evaluation of Binaural Renderers: Externalization, Front/Back and Up/Down Confusions. In *144th Convention of the Audio Engineering Society*, 2018.

[87] Gregory Reardon, Andrea Genovese, Gabriel Zalles, Patrick Flanagan, and Agnieszka Roginska. Evaluation of Binaural Renderers: Localization. In *144th Convention of the Audio Engineering Society*, 2018.