Remote Immersive Audio Production: State of the Art Implementation, Challenges, and Improvements

Stefano Giacomelli [1,*], Carlo Centofanti [1], José Santos [2], Mauro Galbiati [3], Tiziano Salvi [4], Fabio Graziosi [1], Claudia Rinaldi [5]

[1] *DISIM - University of L'Aquila, Italy*
[2] *Ghent University - imec, IDLab, Department of Information Technology, Belgium*
[3] *Independent, Italy*
[4] *Suonovivo Tecnologia, Almé, Bergamo, Italy*
[5] *CNIT - Interuniversity Consortium for Telecommunications, UdR L'Aquila, Italy*

***Abstract*—This paper addresses the increasing interest of industries in remote and distributed audio production applications. Specifically, it provides a detailed overview of the current state of the art in this domain, describing a case-study analysis of an immersive audio content production ecosystem, streamed over a distributed network infrastructure. We analyze performance, reliability and latency of algorithms, communication protocols and systems (with multiple geographical distributions of nodes), exploring the potential of leveraging cutting-edge technologies to better meet the requirements of both remote audio content producers and customers. Our proposal reaches good numerical results (in terms of latency and reliability) and is provided with a versatile and optimized set of audio tools capable of transferring real-time native immersive audio recordings, over the web, to distant geographical nodes where customers in turn, can listen and rapidly ask for contents processing with equal flexibility towards their own listening system.**

***Index Terms*—Internet of Sounds (IoS), Spatial Audio, Ambisonics, Immersive Audio, *low-latency* audio, Communication Networks**

I. Introduction

Imagine an audio event where the performance, production, and audience are spread across three different geographical locations. In one room, musicians perform, in another, the production team operates, and in a third, both musicians and the audience are present. This scenario demands exceptional coordination and communication, particularly to ensure that the production team has real-time awareness of the musicians' performance and the audience's experience.

This scenario exemplifies various applications envisioned in the field of the Internet of Musical Things (IoMusT) [1]. Research papers and experiments focusing on Networked Music Performances (NMPs) have been presented, such as [2]–[4]. These papers discuss and sometimes evaluate possible networking architectures, [5], also in terms of Key Performance Indicators (KPIs) satisfaction [2].

Latency and reliability are not the only requirements in remote production. For the service to be of high quality, the remote production team must have precise knowledge of what is happening in the audience's room. Ensuring an immersive experience requires efficient implementation of spatial audio recording, rendering, and transmitting across all scenarios, guaranteeing that production team and intermediate control teams can accurately monitor and adjust the audio environment, ensuring the best results for both performers and the audience.

Efficient implementation of spatial audio recording and rendering in remote production scenarios has the potential to revolutionize how audio events are produced and experienced, setting new standards for quality and immersion. It may also have several significant economic consequences as: *(i)* cost efficiency: avoiding the costs associated with traveling to the event location and consequently saving time, allowing the management of multiple events in different locations; *(ii)* enhancement of client satisfaction: real-time adjustments ensure the audio experience meets client expectations; *(iii)* competitive advantage: providers who can offer such advanced remote services might gain a competitive edge in the market, attracting more clients who value flexibility and precision; *(iv)* economies of scale: by standardizing remote service processes and tools, companies can achieve economies of scale, reducing the cost per event serviced; *(v)* new business models: this technology could give rise to new business models, such as subscription-based remote audio adjustment services or on-demand expert consultations.

This paper highlights the current State of the Art (SoA) in the audio industry for delivering a remote and networked production scenario that guarantees an immersive and coherent experience for all users (i.e., musicians, producers, audience). This computational simulation is similar and related to the one proposed as a *demo* for the current conference, entitled "Demo Proposal: Remote Immersive Audio Production". We provide an overall evaluation of computational and resource requirements, as well as networking impairments, and present possible solutions that leverage cutting-edge communication technologies.

II. Scenario

As hinted in the introduction, the scenario envisioned addresses the need for multimedia service professionals to provide immediate contents feedback to clients. The integration of spatial audio is driven directly by customer requirements, as the production is commissioned with the explicit goal of guaranteeing immersive experiences. It is worth noticing

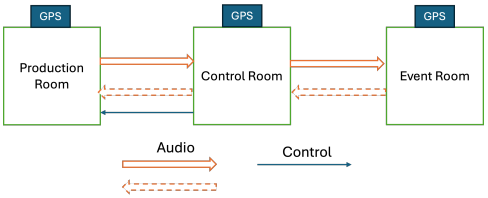

Fig. 1. Blocks Scheme of Our Scenario.

that a first assumption is made by assuming the absence of musicians/performers in the so-called *Event Room* (where the event is reproduced) and the reason for this choice is two-fold. First of all, this scientific paper is related to a demo that is going to take place during the conference, which has practical requirements to be met, secondly we want to focus on how to achieve the highest quality of this technical service by enabling the remote control team with a spatial audio service, trying to guarantee the highest quality of acoustic immersivity, maintaining in background the well known problems of performers low latency necessities imposed in a common NMP scenario.

Typically, audio services are developed in insulated and soundproofed locations, equipped with ad-hoc configured high quality audio gear. The resulting *master* mix is meticulously tailored relying on customers requirements, but is not extensively tested in the actual event environment, until the very day of the event. Consequently, when clients listen to the final results, they may be dissatisfied because of its different sounding from what they have heard in the production stage. In such cases, the production team may not be physically present at the event location and eventual late feedback may require additional re-mixing solutions (which in turn require further hearing-tests phases).

Therefore, this scenario envisions a solution where the service provider (in a *Control Room*) can remotely listen to (or at least in a reliable approximation) what and how the customer is hearing the content in his/her location, so as to make appropriate adjustments to achieve the highest possible level of acoustic immersivity and perceptual satisfaction. This scenario becomes even more complex if we consider that the actual music performance is taking place live in a third, different location (the *Production Room*). The block scheme of this networking context, along with the types of signal exchanges, is depicted in Figure 1.

For our case-study, the *Production Room* is situated in a recording booth at TriangoloLab (Bergamo, Italy). Here, an audio engineer plays back pre-mixed stereo soundscapes while an electronic musician performs live, using a synthesizers ensemble. The performance is captured using two different spatial sound microphone arrays. Our goal is to transmit this immersive electronic production to a *Control Room* in Erlangen (Germany) where an immersive audio loudspeaker system (with required transcoding platforms) is set up. Raw capture signals are transmitted over the network using the Dante *low-latency* protocol [6] and related enabled-devices. Two users in the Control Room can monitor and adjust the soundscapes playback and the synthesis performance in real-time (via ad-hoc Graphical User Interfaces (GUIs), a phisical digital mixer end 2x panner/enconder for soundscape signals). Simultaneously, the controlled immersive scenario is transcoded for another location (the *Event Room*, in this case in the same Erlangen building) where it is reproduced through a resident loudspeakers array (whose disposition at the moment of writing is unknown), a typical scenario where producers propose something in the studio that may sound different at the customer's premises due to variations in device quality, acoustic properties, and spatial configurations.

The aim is to minimize the loss of spatial information (immersivity), and thus a *periphonic* microphones setup is placed in the *Event Room* and its audio is sent back to the *Control Room*, in order for the control engineers to listen (natively with respect to the immersive loudspeakers system) to the final output and properly adjust playback parameters.

III. Requirements

The scenario presented so far, does impose: *(i)* sufficient *throughput*, i.e. all the audio (and control) signals must be properly transmitted; *(ii)* an aggregated (audio and network) *latency* below 40ms, which is the maximum admissible value as per Dante protocol; *(iii) jitter* below 0.25ms: it refers to the deviation in packet arrival times, high jitter could cause timing issues, leading to automatic playback muting (as handled by Dante ecosystem); *(iv)* near-zero *packet loss*, otherwise could results in audio dropouts, unacceptable for live sound and recording environments. Dante moreover relies on a User Datagram Protocol (UDP) based protocol, thus not considering any recovery routine for packet delays or loss handling. *(v) spatial information preservation*: to fully realize the potential of the remote production scenario, it is crucial to ensure the minimum loss of aural impression (i.e. listener's sense of immersion and presence). Therefore, we have implemented two different spatial audio encodings for the demo. During the demo, we plan to distribute questionnaires to users, to gather feedback about qualities and drawbacks of each solution. This feedback will help us integrating our assessments with subjective and perceptual judgments.

IV. Audio Signals

This subsection details configurations of acoustic devices in each of the locations of Figures 2 and 5. All devices and software use a sample rate of 48KHz, 24-bit depth, and 128-samples buffers (exceptions will be detailed when necessary).

A. Audio Infrastructure Set-Up

The *Production Room* is set up with 4x analog audio synthesizers (mono audio sources) which receive real-time synthesis parameters modulation via Musical Instrument Digital Interface (MIDI) messages [7], [8]. Each synthesizer is wired to a dedicated amplifier (with its speaker), placed at the corner of an ideal quadrilateral perimeter, at different heights. This kind of displacement is completely arbitrary and subject to changes based on *Production Room* availability:

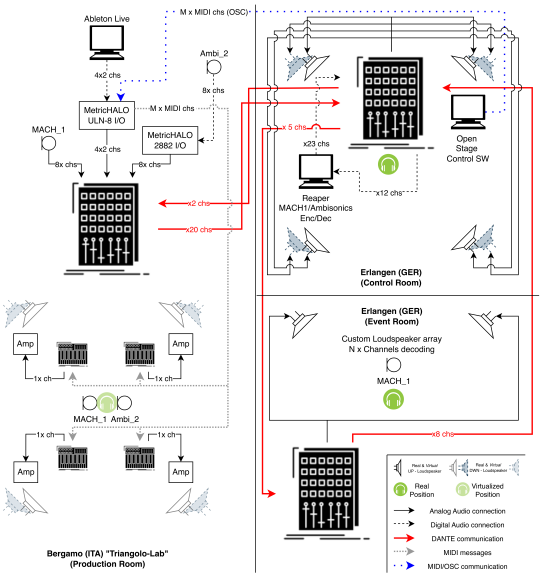

Fig. 2. Our Infrastructure Proposal

as an alternative, they can be replaced by virtual instruments (Virtual Sound Technology (VST)) that can be connected (via audio I/O) to the same amplifiers (hybrid D/A set-up).

Two multi-channel microphones, specifically designed for immersive audio shooting, are placed at the center of the virtual space (designed by the audio sources): an AudioTechnica BP3600 [9] and a VoyageAudio SpatialMic [10]. The BP3600 is a native microphonic near-coincident cubic array, which consists of 8x 3.5" (8.8 cm) hyper-cardioid condenser capsules (SNR: 71dB-A), equidistantly placed from the center core-body: each capsule is 15cm away from the adjacent-corner one (angle between the two: 70.53°). This set-up could be fully replaced with an-hoc Neumann KM184 series 180 [11] custom array (cardioid polar patterns).

The SpatialMic is a 2^{nd} order Ambisonics microphonic array with 8x condenser capsules, manufactured with standard USB/ADAT digital output and a +5V auxiliary power port (SNR: 72dB-A), which is connected to a MetricHalo 2882 audio I/O [12] with on-board Digital Signal Processor (DSP) processors, working as an independent signal converter (ADAT-to-Analog). These microphones are dedicated solely to the musical scenario real-time shooting, for a total of 16x audio channels: 8x of native-*periphony* [13] and 8x of Ambisonics A-Format [14].

The channels pool is extended up to 20x by means of 2x stereo pre-mixed stem tracks, containing live-taking materials of different concrete audio scenarios (*anthropic/cityscape* and *pristine/natural*) played on an Ableton Live session, which handles MIDI-tracks looping, virtual instruments, and the Open Sound Control (OSC)-to-MIDI messages (received from the web) too. All raw audio channels are sent to the analog inputs of a Dante-enabled Behringer X32 Rack Mixer [15], [16] which routes them in data-packets using proprietary *low-latency* Dante protocol [17], [18] (discussed in Section V) to an external Ethernet switch. The switch is driven by a Raspberry Pi-4 with a Linux VM (Raspbian) handling a Precision Time Protocol (PTP) (PTPv2) open-source timecode acquisition [19] from a wired Global Positioning System (GPS) antenna. Finally, another machine runs the Dante Domain Manager software, used to couple and monitor all Dante-interconnected devices, while the Dante-enabling hardware (or the Dante virtual soundcard) ensures that packets are encoded with related timestamps (as per Dante specifications) and sent via a router, over the web.

Our *Control Room* setup includes the same communication node infrastructure: GPS, switch, and router. We acquire signals from the Production Room and send them to one of the two dedicated Dante ports of the audio mixer (for testing purposes an Allen & Heath SQ6). Here, the mixer acts as a digital interface, routing microphone and stem signals to a control PC, which hosts a dedicated REAPER session, for real-time channel encoding (described below in Section IV-B) in MACH1 [20], binaural [21] (for monitoring), and the required format for the Event Room loudspeakers (via 2^{nd} order Ambisonics transcoding [22], [23]). The 8x MACH1 and the 8x B-format-to-MACH1 encoded signals are sent back (via Dante virtual soundcard or Universal Serial Bus (USB) / digital audio I/O available on the digital desk) to the Control Room mixer, and then to the Yamaha HS-7 cubic [24] periphony array, for playback. It has to be stressed once again that the choice of using the periphony array is due to the need of producers to have a precise idea of what is going on in the Event Room context. The same could be exploited either through binaural or virtual soundfield rendering, and we chose to implement both.

Additionally, the mixing engineer and a "creative" user can interact separately with the music and the environmental audio sources: the mixing engineer can spatialize and adjust output gains through a native MACH1 panner [25], while the "creative" can access to custom GUI (via QR code scanning with a smart device), built with Open Stage Control [26] (see Figure 3), with 2x 2D-knobs dedicated to dynamic mixing and panning of audio scenarios, and 3x faders for synthesizers interaction (rhythm, spatial depth and poliphony). User gestures are encoded in OSC messages [27] and sent in real-time via the same communication node infrastructure, back to the Production Room. There, they are acquired, monitored, and converted to MIDI before being applied to the live audio synthesizers.

B. Spatialization Techniques

Both MACH1, a proprietary implementation of Spatial PCM Sampling (SPS), and Ambisonics spatialization techniques leverage the same approach, for which any captured or played-back signal, can be considered as recorded by a microphone

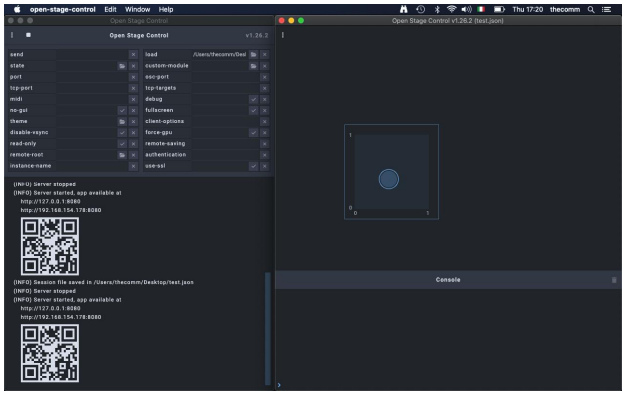

Fig. 3. Open Stage Control Interface Example.

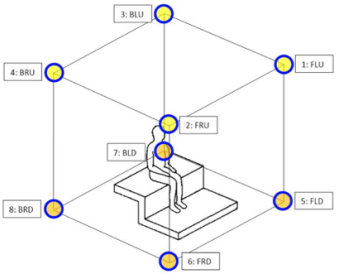

Fig. 4. MACH1 Octahedron Disposition.

positioned in a specific location, with a given direction, and with a certain frequency-dependent directivity (*polar*) pattern [28]. A *virtual* microphone signal y_v, can be synthesized by convolving the signals captured from multiple real capsules x_m, performing a spatial sampling of the resulting soundfield:

$$y_v(t) = \sum_{m=1}^{M} x_m(t) \otimes h_{m,v}(t) \tag{1}$$

The approach used to obtain the convolution filtering coefficients $h_{m,v}$ for any v microphone does not rely on theoretical models, and simply aims to minimize the deviation of the measured polar pattern from the ideal one. Denoting $c_{m,d}$ the matrix of the measured impulse responses from D directions and M microphones, for each of these and at any given frequency, the virtual microphone should produce a "nominal" gain p_d:

$$\sum_{m=1}^{M} c_{m,d}(t) \otimes h_m(t) \rightarrow p_d \cong Q_d \cdot \delta \tag{2}$$

obtained applying a directional gain Q_d to a delayed unit-amplitude *Dirac's delta* function. This computation is simplified in the frequency domain via the Fast Fourier Transform (FFT) algorithm on the N-point impulse responses c, h, and p. Resulting complex spectra are noted C, H, and P turning previous equations into:

$$\sum_{m=1}^{M} C_{m,d,k}(t) \cdot H_m(t) \rightarrow P_d \tag{3}$$

$$[H_k]_{M \times V} = \frac{[P]_{D \times V}}{[C_k]_{D \times M}} \tag{4}$$

a system solvable applying an error minimization algorithm (i.e.: *least squares* method) and a frequency dependent regularization coefficient, as presented in [29].

2^{nd} order Ambisonics [30] represents the spatial information of a sound field using signals that correspond to virtual microphones possessing complex polar patterns (*spherical harmonics* functions), mathematically expressed as an orthonormal basis of *Legendre functions* and a set of normalization coefficients: related expressions can be examined in [31]. Its recording format is named "A-format". An intermediate representation (the B-format) can be obtained exploiting directivity of the real capsules and ad-hoc "decoders" for the target loudspeaker setup [32].

The SPS instead, is a multi-channel stream (also P-format) in a 2D spherical-coordinate space, analogous to the traditional Pulse Code Modulation (PCM) waveform representation (in the time domain). An SPS track comprises signals recorded with coincident directive microphones, arranged to uniformly cover a sphere surface. These signals encode spatial information based on amplitude, rather than phase, slightly differing from Ambisonics for exploiting a non-perfectly orthonormal transformation basis. By knowing the azimuth Az_m and elevation El_m of the recording capsules (up to 32x), any mono signal can be encoded as an n-channels P-format directional signal. The angle between the incoming sound and each microphone's axis is computed using the Haversine formula:

$$\phi_m = 2 \arcsin \left[\sin^2 \left(\frac{El_m - El_{in}}{2} \right) + \cos(El_m) \cdot \cos(El_{in}) \cdot \sin^2 \left(\frac{Az_m - Az_{in}}{2} \right) \right]^{\frac{1}{2}} \tag{5}$$

and substituted to obtained the n^{th} channel gain:

$$Q_n(\phi) = [0.5 + 0.5 \cos(\phi)]^4 \tag{6}$$

To drive our cubic periphony (SPS8 – MACH1) represented in Figure 4, with n-channels SPS signals, we use a pre-defined "decoding matrix" [20] of nx8 Finite Impulse Response (FIR) filters (F), whose spatial coordinates are listed in Table IV-B. The nx SPS signals are thus convolved with the F filters, to produce speakers feeds vector.

The Ambisonics A-Format 8-channel recording stream is encoded in B-Format using the ad-hoc "Voyage Audio A-B format converter" plug-in [33]. The B-Format signals are then converted to the AmbiX interchange format [34] and spatialized on the cubic periphony using the All-Round Ambisonics Panning and Decoding (AllRAP/AllRAD) technique [23], [35]. The AllRAD technique renders virtual loudspeaker locations and related output gains through *t-designed* approximations of a sphere (*t*=3, for an *octahedron*), optimally truncating the spherical harmonics re-synthesis summation to the desired

TABLE I
MACH1 OCTAHEDRON FEATURES

Mic #	Azimuth (°)	Elevation (°)
1	45	35.26
2	315	35.26
3	135	35.26
4	225	35.26
5	45	-35.26
6	315	-35.26
7	135	-35.26
8	225	-35.26

sound field approximation. It relies on an energy-compensated Vector-based Amplitude Panning (VBAP) mechanism [36]–[38] to simulate the *phantom source* signals for the cubic loudspeakers' real coordinates (provided in `.json` format). The resulting scaling coefficients are simply the composition of the *t*-discretized Ambisonics and VBAP gain coefficients. Engineers in the Control and Production Room can monitor the AllRAD and/or MACH1-SPS spatialization outcomes via headphones binauralization, achieved through real-time Head-related Transfer Functions (HRTF) convolution [21].

Native MACH1 microphone signals are directly routed to the cubic periphony, while the environmental stereo stems pass through a proprietary MACH1-Panner plugin [25], enabling 3D independent panning and amplitude scaling for each scene.

In the Event Room, transcoding DSP for MACH1-SPS streams is performed through 2nd order Ambisonics B-Format conversion using the "SPARTA AmbiENC" plugin [22], which exploits azimuth and elevation parameters, as detailed in Table IV-B. Both MACH1-to-B-Format and native Ambisonics B-Format streams can ultimately be rendered to the location's loudspeaker array using the AllRAD technique.

V. Network Architecture

This section presents the proposed network architecture for remote audio transmission, highlighting key components and their interactions across three primary sites (Figure 6) to ensure high-quality, low-latency audio transmission and real-time control messages between geographically distributed locations.

To maintain precise audio and controls synchronization across all connected locations, the Production Room employs GPS timecode synchronization Coordinated Universal Time (UTC)-based. This ensures that all audio and control signals are accurately time flagged, exploiting the Raspberry PI-4 network adapter as *grandmaster* clock. Each room Ethernet switch instead work as *boundary* clock used to minimize static and dynamic communication fluctuations, maintaining the integrity of the packets trasmission and avoiding latency. For secure and reliable communication, the Production Room uses 2x Virtual Private Networks (VPNs). VPN 12 connects the Production Room with the Control Room in Erlangen, ensuring that audio and control signals are transmitted securely over the internet. Additionally, VPN 13 is exploited to establish a secure link with the Event Room, also in Erlangen, facilitating the direct transmission of audio data. These VPN tunnels encrypt the data, guaranteeing secure and reliable communication channels.

In the Control Room, VPN 12 connects towards the Production Room, for secure audio and control signal transmission, while VPN 23 ensure audio pre-processed playback transmission towards the Event Room. This setup decouples the *periphony* architecture from our specific implementation, making it adaptable to various remote production scenarios, ensuring flexibility for different geographical and technical requirements.

The Event Room, is where the final audio is rendered and played back. GPS timecode synchronization ensures that playback is accurately timed with the Production and Control Rooms. An internal network switch manages incoming audio data packets, while VPN 23 ensures secure reception of binaural recordings and transmission back to the Control Room. This setup not only guarantees secure and reliable communication but also decouples the resident architecture from our specific immersive implementation. This flexibility allows the architecture to be applied to different geographical and technical requirements and scenarios.

Dante: The Dante protocol, developed by Audinate, is a proprietary *network layer* audio and transport protocol for transmitting high-quality, low-latency digital audio over standard Ethernet networks. It is widely used in professional audio fields like live sound, broadcasting, recording studios, and sound installation systems due to its flexibility, scalability, and ease of use. It encapsulate audio data as payload in an IP packet: Dante integrates hardware, control software, and the transport protocol itself. Audio sources must be either Dante-enabled or connected to a device that supports Dante.

Dante operates on a switched Ethernet network of 100 Mbps or higher, with a recommended Gigabit Ethernet backbone. It utilizes the VoIP DiffServ category to prioritize audio and control data, allowing it to share a Local Area Network (LAN) with other types of data, without requiring dedicated bandwidth allocation. With Fast Ethernet, a single Dante connection supports up to 48x channels of bi-directional 24bit, 48KHz audio signals. On Gigabit Ethernet, it supports 512x channels. The actual channel limit varies with respect to audio stream quality: higher quality reduces the number of channels, and lower quality increases it. Dante can operate over any Ethernet topology. The network's latency is configured by the system administrator using Dante control software and is based on the network size. For live sound, Audinate advises setting the latency to no more than 1ms. If a packet arrives before the latency period ends, the device will delays it slightly to ensure all nodes stay synchronized. Dante control software enables to allocate network bandwidth for *time-critical* paths, like to playback loudspeakers. However, lower latency requires more bandwidth, limiting channel capacity. Dante networks support up to 10 hops and is compatible with AES67. The proprietary protocol ensures precise synchronization using PTP, offering redundancy with primary and secondary network paths, and employing Quality of Service (QoS) to prioritize audio traffic.

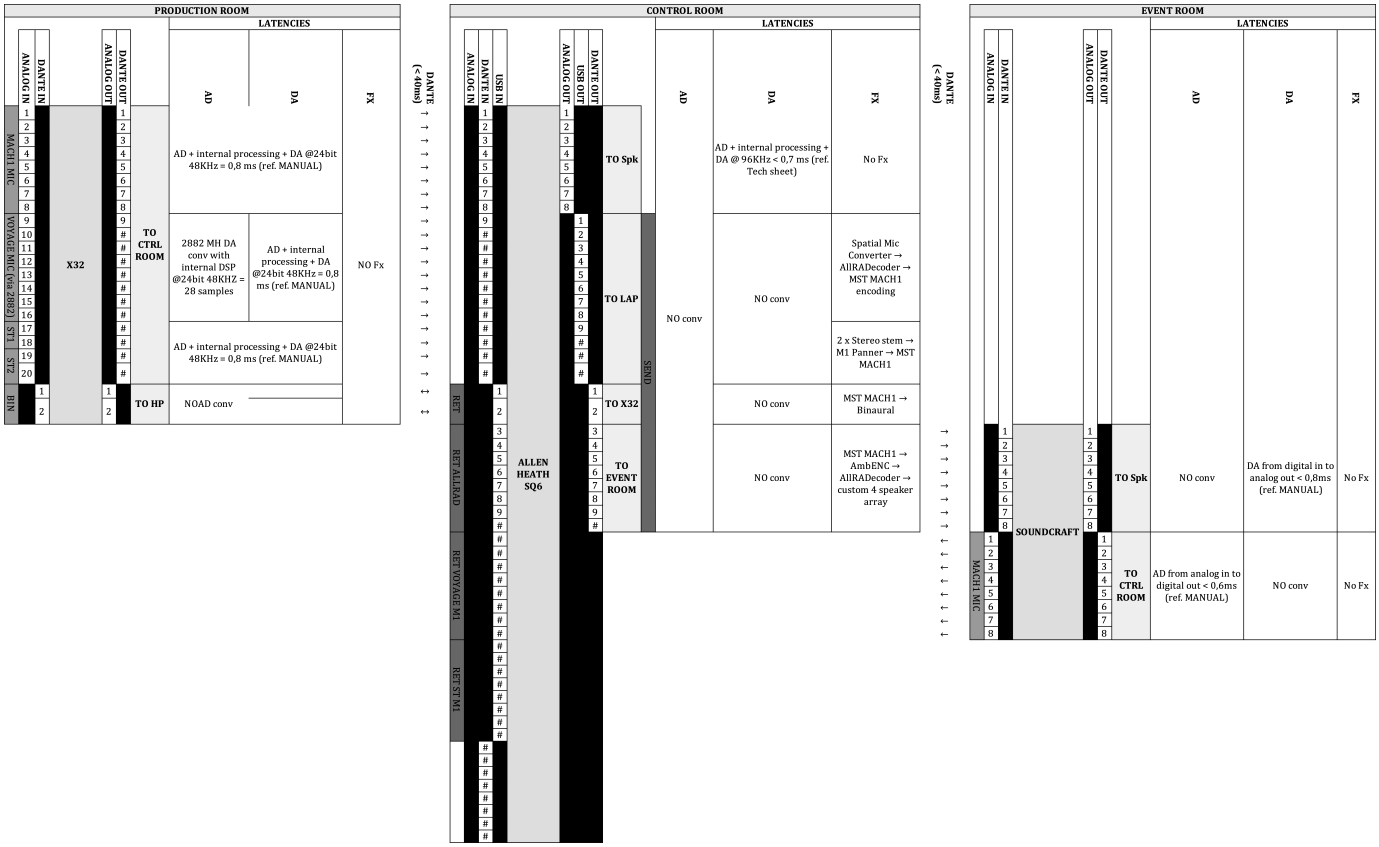

Fig. 5. Audio Connections.

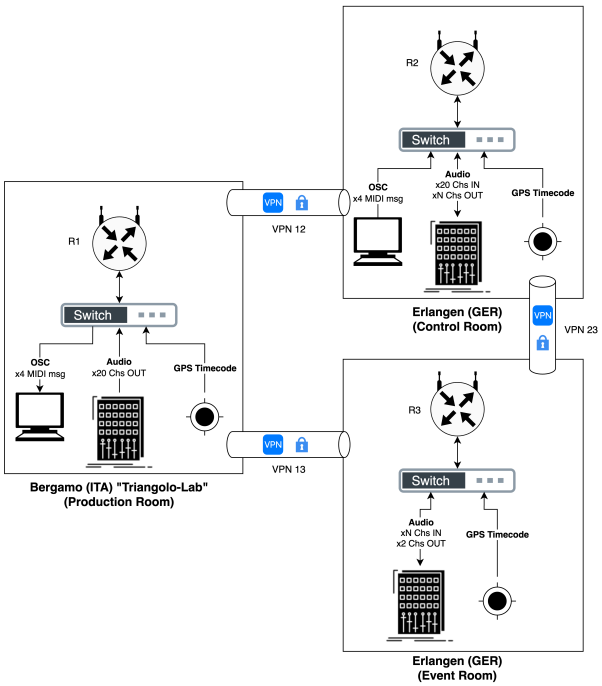

Fig. 6. Network Reference Architecture.

VI. Simulations and Results

Various simulations were conducted to evaluate the feasibility of the proposed scenario. Relying on the maximum number of channels to be transmitted (this happens from the Production Room to the Event Room) and assuming sampling rate as defined in Section IV, the data rate required can be computed as follows:

$$\begin{aligned}\text{Bit rate (per channel)} &= \text{SR} \times \text{Bit Depth} = \\ &= 48000\text{samps/sec} \times 24\text{bits/samps} = 1152000\text{bits/sec} \\ &= 1.152\text{Mbps}\end{aligned} \tag{7}$$

$$\begin{aligned}\text{Tot. Bit rate} &= \text{Bit rate (per channel)} \times \text{Num. of chs} = \\ &= 1.152\text{Mbps/ch} \times 24\text{chs} = 27.648\text{Mbps}\end{aligned} \tag{8}$$

including overhead and additional protocol data (a factor of about 15%):

$$\begin{aligned}\text{Tot. Bandwidth } (+\text{ overhead}) &= \text{Tot. Bit rate} \times 1.15 = \\ &= 27.648\text{Mbps} \times 1.15 = 31.7952\text{Mbps}\end{aligned} \tag{9}$$

with VPN overhead and some margin, the final estimate would be about 45Mbps.

The latency evaluation has been performed separately for the audio part and the network part, thus allowing to take decisions before deploying the final solution and provide

guidance to the reader to better understand our approach and reproduce within their custom use cases.

Concerning the audio, the overall latency contribution of converters and DSP plugins is summarized in Figure 8, where A/D and D/A converters delays are directly supplied by mixer manuals, and DSP delays are estimated thanks to Plugin Delay Compensations (PDCs), as performed by REAPER. For the sake of completeness, Plugin Delay Compensation (PDC) in REAPER is a feature that ensures all audio tracks in a project remain perfectly synchronized, even when some of them are processed by tools introducing latency, i.e. those that perform complex processing tasks, such as: *lookahead* compression, convolution, or linear-phase EQ. Without meticulous compensation, the introduced latency would cause tracks to playback slightly out-of-sync with the rest of the project. We tried to minimize plugins requirement, exploiting the most stripped encoding-decoding chain for the immersive audio reproduction (as required by both producers and clients) and the most versatile ready-to-use solution for the Event loudspeakers array. The *A/D-to-Channel strip-to-D/A* stage, performed on our Behringer X32Rack, causes a latency of 0.8ms on both types of microphonic inputs (MACH1 and Ambisonics). Possibly being generated by a *virtual instrument* (VST), synthesizers MACH1 signals could suffer of a negligible 64 samples pre-buffering (on Ableton live), while Ambisonics signals strictly require both I/O buffering (5.4ms) and plugin's signal encoding, which causes no latency overhead.

Once the signals have been carried to the *Control Room*, they need to be acquired by the Control PC by means of a Dante virtual sound card (which avoids overheading A/D conversions, otherwise a 2.7ms should be accounted for). MACH1 signals are processed in I/O audio chunks by the REAPER Digital Audio Workstation (DAW) (128samples $=$ 2.7ms latency) and traverse the proprietary panning plugin, which simply apply vectors gain coefficients (no latency). Ambisonics B-format signals require re-encoding and AllRAD processing (performed at chunks level, 2.7ms of latency). Finally signals are sent back (via Dante virtual sound card) to the Control Room Allen & Heat SQ6 mixer, which apply an additional D/A delay ($<$ 0.7ms) before audio monitoring on the *periphony*. Instead, special care should be paid to the introduction of *multichannel-to-binaural* spatialization/encoding algorithms, anywhere inside the DSP-net communication process: requiring HRTF convolution, they could add considerable latency (breaking DANTE protocol requirements), especially when including adaptive filters adjustments, related to head's position/tilt tracking. The SPARTA binauralizer [21] alone, which simply performs fixed Head Related Impulse Response (HRIR) convolution [39], actually requires a PDC of 1536samples (32ms at 48KHz). Only an additional D/A delay is applied to the overall DSP network (in the Event Room) which actually affects only the signal amplification stage (post-network communication). On the *Latency* row in Figure 8 we performed the above latency summations (per channel-type, per path), including *network* delays estimated as described below. We examined the practical effects of network latencies

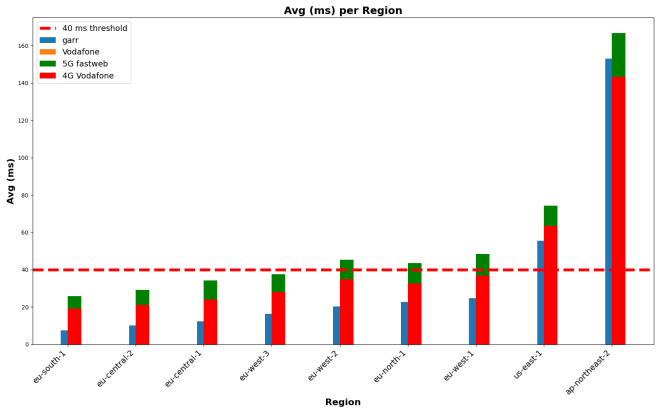

Fig. 7. Average latency from L'Aquila to different AWS cloud regions.

across a selection of server locations provided by Amazon Web Services (AWS), with a particular focus on their implications for user experiences and network reliability. We targeted AWS server destinations in the following regions: Frankfurt (eu-central-1), Ireland (eu-west-1), London (eu-west-2), Milan (eu-south-1), Paris (eu-west-3), Stockholm (eu-north-1), Zurich (eu-central-2), Northern Virginia (us-east-1), and Seoul (ap-northeast-2). To obtain an accurate representation of network latencies, we designed a measurement procedure using a Bash script that employed the Internet Control Message Protocol (ICMP) via the `ping` command. This script was executed to conduct 1000x measurements for each server location, which were systematically repeated over a set duration.

Following the data collection phase, we conducted statistical analyses to compute the average, minimum, and maximum latency values, as well as the standard deviation for each set of measurements. It is important to note that the measurements represent the Round-Trip Time (RTT); however, for our study, we required only the one-way latency. To derive the one-way latency values, we divided each RTT measurement by two.

In Figure 7 we present the evaluated one-way latency to various AWS cloud locations. For this analysis, we applied a threshold value of 40ms, which aligns with the latency requirements specified by the Dante protocol. It is important to note that the overhead induced by VPN usage in terms of latency is negligible compared to the measured one-way latency values. Finally the jitter problem has been solved by exploitation of GPS synchronization.

A. Results Discussion

For our scenario (Figure 8), we present the cumulative sums of both DSP and network components, acting on each audio signal path type. We included all possible combinations of DSP chains, excluding the VoyageAudio (Ambisonics) on-board acquisition DSP (far below 1ms) and the MetricHalo 2882 A/D-DSP chain (28samples $=$ 0.5ms at 48KHz).

The audio/DSP latency generated inside the Production Room ranges between 3.5ms (for pre-recorded soundscapes D/A) and 6.2ms for both MACH1 and Ambisonics recording.

<table>
<tr><th colspan="4">Production Room</th><th rowspan="2">L'Aquila --> Frankfurt Network (Min, MAX, Avg)</th><th colspan="3">Control Room</th><th rowspan="2">Erlangen --> Erlangen Network (Min, MAX, Avg)</th><th rowspan="2">Event Room</th></tr>
<tr><th>Source</th><th>Recording</th><th>A/D conversion</th><th>DAW I/O buffer</th><th>DSP plugins</th><th>DAW I/O buffer</th><th>DSP plugins</th><th>D/A conversion</th><th>D/A conversion</th></tr>
<tr><td rowspan="2">Music Synth</td><td>MACH1 BP3600 (VST or Analog)</td><td rowspan="3">0.8ms</td><td rowspan="2">2.7ms (x2)</td><td>None: 0ms</td><td>4G (Vodafone): 20.99ms - 37.12ms - 24.03ms</td><td rowspan="3">2.7ms (x2)</td><td>None: 0ms</td><td><0.7ms</td><td rowspan="3">Local: 4.01ms - 20.62ms - 6.17ms</td><td rowspan="3">0.8ms</td></tr>
<tr><td>VoyageAudio Ambisonics Mic</td><td>Ambisonics A --> B (Ambix): 0ms</td><td>5G (Fastweb): 29.12ms - 55.01ms - 32.07ms</td><td>AmbiEnc --> AllRAD: 2.7ms</td><td rowspan="2">None: 0ms</td></tr>
<tr><td>Soundscapes</td><td>Pre-recorded audio files mix</td><td>2.7ms</td><td>None: 0ms</td><td>Optical Net (Vodafone): 13.95ms - 19.10ms - 15.33ms</td><td>Panner: 0ms</td></tr>
<tr><td rowspan="16">Latency</td><td colspan="4" rowspan="16">[3.5, 6.2] ms</td><td>Optical Net (GARR): 11.18ms - 18.29ms - 12.28ms</td><td colspan="3"></td><td rowspan="16">+ Local(Avg) + D/A (Event)</td><td></td></tr>
<tr><td>Min(PR latency) + 4G(Avg) + Min(CR latency)</td><td colspan="3">32,93</td><td>39,9</td></tr>
<tr><td>Min(PR latency) + 4G(Avg) + Max(CR latency)</td><td colspan="3">36,33</td><td>43,3</td></tr>
<tr><td>Min(PR latency) + 5G(Avg) + Min(CR latency)</td><td colspan="3">40,97</td><td>47,94</td></tr>
<tr><td>Min(PR latency) + 5G(Avg) + Max(CR latency)</td><td colspan="3">44,37</td><td>51,34</td></tr>
<tr><td>Min(PR latency) + ONVodafone(Avg) + Min(CR latency)</td><td colspan="3">24,23</td><td>31,2</td></tr>
<tr><td>Min(PR latency) + ONVodafone(Avg) + Max(CR latency)</td><td colspan="3">27,63</td><td>34,6</td></tr>
<tr><td>Min(PR latency) + ONGarr(Avg) + Min(CR latency)</td><td colspan="3">21,18</td><td>28,15</td></tr>
<tr><td>Min(PR latency) + ONGarr(Avg) + Max(CR latency)</td><td colspan="3">24,58</td><td>31,55</td></tr>
<tr><td>MAX(PR latency) + 4G(Avg) + Min(CR latency)</td><td colspan="3">35,63</td><td>42,6</td></tr>
<tr><td>MAX(PR latency) + 4G(Avg) + Max(CR latency)</td><td colspan="3">39,03</td><td>46</td></tr>
<tr><td>MAX(PR latency) + 5G(Avg) + Min(CR latency)</td><td colspan="3">43,67</td><td>50,64</td></tr>
<tr><td>MAX(PR latency) + 5G(Avg) + Max(CR latency)</td><td colspan="3">47,07</td><td>54,04</td></tr>
<tr><td>MAX(PR latency) + ONVodafone(Avg) + Min(CR latency)</td><td colspan="3">26,93</td><td>33,9</td></tr>
<tr><td>MAX(PR latency) + ONVodafone(Avg) + Max(CR latency)</td><td colspan="3">30,33</td><td>37,3</td></tr>
<tr><td>MAX(PR latency) + ONGarr(avg) + Min(CR latency)</td><td colspan="3">23,88</td><td>30,85</td></tr>
<tr><td>MAX(PR latency) + ONGarr(avg) + Max(CR latency)</td><td colspan="3">27,28</td><td>34,25</td></tr>
</table>

Fig. 8. Audio & Networks Latency.

The Ambisonics A-to-B-Format plugin generates no latency overhead, allowing artists and technicians in Bergamo to monitor their real-audio results with a total latency of 6.7ms on the monitor loudspeaker setup.

When traversing the *L'Aquila-to-Frankfurt* network, a variable amount of latency is applied. We performed statistical computations by summing the average values of each network latency with the previously obtained DSP minimum and maximum latency. Our results show that signals are likely to be transmitted efficiently, with almost all latency sums below the 40ms requirement. However, we occasionally encountered insufficient performance with 5G (40 up to 47ms), though more tests are needed to confirm this.

Some results related to 4G (highlighted with a gray background) warrant a technical consideration: our infrastructure is split across three locations (even though two are in the same geographical location), which allows for a 40ms $\times$ 2 Dante connections overhead (latency is evaluated by Dante hardware, *point-to-point*). If we collapse the Control and Event rooms together, summing the *Erlangen* path of the network and performing DSP operations on a single production setup, we might overcome Dante latency requirements due to an increase in computational requirements at that site, particularly with mobile network technologies (4G and 5G). This confirm the need of geographically distributing DSP solutions through advanced communication networks to guarantee customers satisfaction in a remote production framework.

We can conclude that our audio DSP chain offers good versatility in terms of audio monitoring. It supports various configurations such as MACH1-to-Ambisonics, Ambisonics-to-MACH1, MACH1-to-binaural, Ambisonics-to-binaural, and MACH1/Ambisonics-to-*custom array*. It minimizes computational requirements with SoA solutions and allows both technical teams and the audience (or clients) to monitor audio outcomes according to their respective infrastructure availability (loudspeakers).

A separate discussion is warranted regarding the latency issues concerning OSC/MIDI communication. These protocols transfer control signals sampled at a lower rate, usually one-third of that of audio, necessitating a longer delay time for effective application in the sound synthesis ecosystem. This well known issue in the IoMusT research field requires prior awareness of both the amount and type of control delegated to the user/performer, as well as careful integration within the flow of the musical performance to avoid disrupting its formal structure.

VII. Solutions beyond State of the Art

A. Alternatives for the Audio Infrastructure set-up

Despite the Dante spread over the music industry domain being the most performing low-latency all-in-one solution, at least one open-source alternative exists, i.e.: SonoBus [40].

SonoBus is an application for high-quality and low-latency P2P audio streaming, between devices over the internet or a local network, with multi-user, multi-platform and fine-grained settings support. Audio encoding quality can be selected from uncompressed PCM (16, 24, or 32 bit) up to compressed bit-rates (16-256KBps, per channel) using the low-latency standardized Opus codec (IETF - RFC6716, [41]). It runs on MS Windows, MacOS and GNU/Linux, as well as in mobile platforms like IOS and Android, and AudioUnit (AU)/VST/AAX plugins (for in DAW DSP).

As reported in its official User Guide, with 64 samples buffering seems to ensure stable audio communications over wired and Wi-Fi networks (with PCM data), via direct UDP connections to a specific host IP. In its User Interface (UI), detailed statistics of audio network and communications are reported (jitter, latency and packet-loss).

SonoBus does not use any encryption in data communication. All audio is sent P2P directly between users and the connection server is only exploited for users inside a *public* or *private group*, so that they can find each other relying on the session name and an automated networking routine. Audio streams networking success relies only on the overall quality of the connections between users audio gears and DSP chains.

It allows to setup custom (multi)-channel based or high-level (stereo, 5.1 or 7.1) tracks grouping, before online streaming. Online sessions recording is supported too. Deeper tests are

required with our audio networking setup to confirm its viability, but its features are unquestionably promising.

B. Access Networks for Low-Latency Audio Production

This subsection examines how various access network technologies, such as mobile networks (4G and 5G), wireless access networks, and fixed-line solutions like Fiber to the X (FTTx), can support low-latency audio production. Each network type has distinct advantages and limitations affecting optimization in different environments.

1) Wireless and Mobile Access Networks: Wireless Access Networks use radio or microwaves for data transmission, including Wi-Fi for local connectivity, and 4G and 5G for wide-area coverage. Technologies like WiMAX and Low-Power Wide-Area Network (LPWAN) extend coverage and support low-power IoT devices. Wireless networks offer flexibility but tend to be less reliable than wired networks, resulting in higher latency, inconsistent jitter, and potential interference issues.

4G Networks: 4G has improved mobile internet speed and reduced latency compared to previous technologies. It supports remote access to cloud-based audio tools, but with latencies of 50-100 ms, it can limit live high-fidelity audio mixing and complex multi-channel setups. Advanced buffering and predictive algorithms can help but may introduce perceptible delays.

5G Networks: 5G offers lower latency, increased reliability, and higher data rates. With potential latency as low as 1 ms for industrial IoT applications, 5G is crucial for live audio applications requiring precise timing and synchronization. It supports real-time remote audio mixing, higher device density, and network slicing for tailored media production needs. 5G increased bandwidth also handles higher-resolution audio and more channels simultaneously.

Non Stand-Alone (NSA) vs. Stand Alone (SA) Architectures: Most 5G implementations use NSA architecture, relying on 4G LTE infrastructure for control functions while 5gnr handles user data. This allows quick deployment and cost savings but can limit 5G low-latency potential due to bottlenecks at the 4G core. In contrast, SA 5G uses a fully 5G core, minimizing data travel distance and significantly reducing latency, enhancing real-time processing capabilities.

Future Net: 6G networks will build on the SA model with advanced edge computing, further reducing latency, improving resilience, and increasing bandwidth. These advancements will support even more complex and latency-sensitive audio production and interactive technologies.

2) Fixed Access Networks: Fixed Access Networks provide permanent, high-capacity connections to end-users through wired infrastructure. These networks today typically utilize technologies such as xDSL (Digital Subscriber Line), which transmits data over existing copper telephone lines, and FTTx (Fiber to the x), which employs optical fiber for superior speed and reliability. They are essential for delivering consistent internet, video on demand, multimedia services in general, and telephone services to homes and businesses, offering higher bandwidth and stability compared to wireless alternatives.

The "x" in FTTx stands for different possible termination points of the fiber optic cable, such as in Fiber to the Home (FTTH), Fiber to the Building (FTTB), Fiber to the Curb (FTTC), Fiber to the Node (FTTN). Each variation provides a balance between deployment cost and performance, with closer fiber terminations generally offering better speed and reliability.

C. Towards a Cloud-Native Infrastructure for Low-Latency Audio Production Protocols

In the upcoming years, the audio domain will inevitably see the virtualization of many, if not all, components that are currently hardware-based due to demanding performance factors such as latency, bandwidth, jitter, and technological advancements. The cloud continuum [42], with its orchestration capabilities and the ability to move services closer to the end user, plays a crucial role in this transition. Technologies such as custom Kubernetes schedulers enable efficient resource management and dynamic service deployment, bringing processing power closer to the user to reduce latency [43] [44]. Despite the advancements in cloud-native infrastructures, several challenges remain. Optimizing network performance and reliability is critical, and technologies like Software-Defined Networking (SDN) and Virtual Private Cloudss (VPCs) can provide dynamic resource allocation and efficient traffic management, ensuring flexible and scalable infrastructures. Embracing microservice-based architectures alongside popular container orchestration platforms will allow modularization and efficient management of audio production applications

VIII. Conclusions and future works

In this paper, we explored the design and implementation of a remote production scenario using state-of-the-art technological solutions. We focused on providing producers with a realistic impression of what is happening in both the Production Room and the Event Room by minimizing latency and optimizing DSP/Network reliability. This approach enables producers to have a precise understanding of the audio modifications needed to improve the QoS.

The solution presented leverages advanced technologies like the Dante protocol over standard IP networks with guaranteed QoS and precise GPS synchronization. Unlike dedicated networks, our approach is designed to operate effectively on shared infrastructures, making it suitable for a broader range of applications. While recent commercial and research efforts in remote production focus on more accessible NMP streaming approaches, our solution provides superior performance, particularly in scenarios where low latency and high reliability are critical, justifying its use in specialized contexts.

Quality of Experience (QoE) assessments will be conducted during the *demo* proposal. We believe that any spatialization algorithm requires supplementing its reliability with subjective judgments provided by listeners. The same applies to the quality and effectiveness of the controls in the GUI interfaced with the Control Room teams.

Our audio-networked infrastructure exceeded results previously achieved in [45], overcoming challenges related to distance range coverage, algorithmic complexity, and spatial and channel distribution of signals. This was combined with more thorough testing of multiple networking (mobile) technologies.

ACKNOWLEDGEMENTS

The authors would like to thank Audinate for their support and provision of a non-expiring license of Dante Domain Manager and Audio-Technica and SISME srl for free loan of BP3600 microphones.
This work is partially supported by the European Union through H2020-MSCA-RISE OPTIMIST project (GA: 872866), NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem (ECS00000041 - VITALITY - CUP E13C22001060006, and partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART" - CUP E83C22004640001). José Santos is funded by the Research Foundation Flanders (FWO), grant number 1299323N.

REFERENCES

[1] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.
[2] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5g-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, pp. 1–1, 2024.
[3] L. Comanducci, *Intelligent Networked Music Performance Experiences*. Cham: Springer International Publishing, 2023, pp. 119–130.
[4] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.
[5] R. Hoy and D. Van Nort, "A technological and methodological ecosystem for dynamic virtual acoustics in telematic performance contexts," in *Proceedings of the 16th International Audio Mostly Conference*, ser. AM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 169–174. [Online]. Available: https://doi.org/10.1145/3478384.3478425
[6] DanteAudinate. (2024) GetDante.com. [Online]. Available: https://www.getdante.com/
[7] MIDIassociation. (2024) Midi.org official site. [Online]. Available: https://midi.org/
[8] F. Avanzini, V. Faschi, and L. Ludovico, "A web-based midi 2.0 monitor," in *Proceedings of the Sound and Music Computing Conference*. Stockholm, SWE: KTH Royal Institute of Technology, Royal College of Music in Stockholm, 2023, pp. 148–153.
[9] AudioTechnica. (2024) Audiotechnica Immersive Audio Microphone BP3600. [Online]. Available: https://www.audio-technica.com/en-eu/bp3600
[10] VoyageAudio. (2024) Voyageaudio SpatialMic USB. [Online]. Available: https://voyage.audio/spatialmic/
[11] Neumann. (2024) Km 184 (Series 180). [Online]. Available: https://www.neumann.com/en-gb/products/microphones/km-184-series-180/
[12] MetricHalo. (2024) 2882-3D recording interface. [Online]. Available: https://mhsecure.com/metric_halo/index.php?option=com_content&view=article&id=131:2882&catid=36:hardware&Itemid=103
[13] M. J. Gerzon, "Periphony: With-height sound reproduction," *Journal of The Audio Engineering Society*, vol. 21, pp. 2–10, 1973. [Online]. Available: https://api.semanticscholar.org/CorpusID:110210326
[14] A. Farina. (2024) Ambisonics/sps pages. [Online]. Available: http://pcfarina.eng.unipr.it/Ambisonics.htm
[15] DanteAudinate. (2024) Dante-enabled Products list. [Online]. Available: https://www.getdante.com/products/dante-enabled/
[16] Behringer. (2024) X32 Rack Mixer. [Online]. Available: https://www.behringer.com/product.html?modelCode=0604-AAA
[17] J. Sharkey. (2024) wycliffe: a clean room implementation of the dante protocol (github). [Online]. Available: https://github.com/jsharkey/wycliffe
[18] T. Wox. (2024) Inferno - unofficial implementation of Dante protocol. [Online]. Available: https://github.com/teowoz/inferno
[19] TimeBeat. (2024) TimeBeat App. [Online]. Available: https://www.timebeat.app/solutions/software
[20] MACH1 Technologies. (2024) Mach1 encoding. [Online]. Available: https://www.mach1.tech/
[21] L. McCormak. (2024) Sparta Plugin suite: Binauraliser. [Online]. Available: https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#binauraliser
[22] ——. (2024) Sparta Plugin suite: Ambienc. [Online]. Available: https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#ambienc
[23] IEM. (2024) IEM Plugin suite: AllRA Decoder. [Online]. Available: https://plugins.iem.at/docs/plugindescriptions/#allradecoder
[24] Yamaha Corporation. (2024) Series HS loudspeakers. [Online]. Available: https://it.yamaha.com/it/products/proaudio/speakers/hs_series/index.html
[25] MACH1 Technologies. (2024) Mach1-panner. [Online]. Available: https://www.mach1.tech/guides/Mach1-Panner.pdf
[26] OpenStageControl. (2024) Open Stage Control: Libre and modular osc/midi controller. [Online]. Available: https://openstagecontrol.ammd.net/docs/getting-started/introduction/
[27] CCRMA. (2024) Open Sound Control.org. [Online]. Available: https://ccrma.stanford.edu/groups/osc/index.html
[28] A. Farina, A. Capra, L. Chiesi, and L. Scopece, "A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in post production," in *AES 40th International Conference*, 10 2010, pp. 8–10.
[29] A. Farina, A. Amendola, L. Chiesi, A. Capra, and S. Campanini, "Spatial pcm sampling: A new method for sound recording and playback," in *AES 52nd International Conference*, 09 2013, pp. 1–12.
[30] F. Zotter and M. Frank, *Ambisonics : A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, 1st ed., ser. Springer Topics in Signal Processing, 19. Cham: Springer Nature, 2019.
[31] A. Farina. (2024) ACN-N3D formulas for High Order Ambisonics. [Online]. Available: http://www.angelofarina.it/Aurora/HOA_ACN_N3D_formulas.htm
[32] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of The Audio Engineering Society*, vol. 60, pp. 807–820, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:55805919
[33] Voyage AUdio. (2024) Spatial mic converter. [Online]. Available: https://voyage.audio/downloads/#spatial-mic-converter-plugin
[34] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix - a suggested Ambisonics format," in *Proceedings of the Ambisonics Symposium*, 07 2011, pp. 1–11.
[35] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of The Audio Engineering Society*, vol. 60, pp. 807–820, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:55805919
[36] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources, part 1: Stereophonic panning," *J. Audio Eng. Soc.*, vol. 49, pp. 739 – 752, 09 2001.
[37] V. Pulkki, "Localization of amplitude-panned virtual sources ii: Two- and three-dimensional panning," *J. Audio Eng. Soc*, vol. 49, p. 753–767, 09 2001.
[38] ——, "Virtual sound source positioning using vector base amplitude panning," *Journal of The Audio Engineering Society*, vol. 45, pp. 456–466, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:108593534
[39] N. Markus, S. Alois, M. Thomas, and H. Robert, "A 3d ambisonic based binaural sound reproduction system," *Journal of the audio engineering society*, no. 1, June 2003.
[40] SonoBus. (2024) High quality network audio streaming. [Online]. Available: https://www.sonobus.net
[41] IETF. (2024) Opus-codec. [Online]. Available: https://opus-codec.org/
[42] J. Santos, T. Wauters, B. Volckaert, and F. De Turck, "Towards low-latency service delivery in a continuum of virtual resources: State-of-the-art and research directions," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2557–2589, 2021.
[43] C. Centofanti, W. Tiberti, A. Marotta, F. Graziosi, and D. Cassioli, "Taming latency at the edge: A user-aware service placement approach," *Computer Networks*, vol. 247, p. 110444, 2024.
[44] J. Santos, C. Wang, T. Wauters, and F. De Turck, "Diktyo: Network-aware scheduling in container-based clouds," *IEEE Transactions on Network and Service Management*, vol. 20, no. 4, pp. 4461–4477, 2023.
[45] F. Paul, C. Chris, and G. Simon, "Trans-europe express audio: testing 1000 mile low-latency uncompressed audio between edinburgh and berlin using gps-derived word clock, first with jacktrip then with dante." *Journal of the Audio Engineering Society*, no. 605, may 2020.