Investigation of Server–Based Spatial Audio for Metaverse Concert Distribution

Patrick Cairns
AudioLab, School of PET
University of York
York, United Kingdom
patrick.cairns@york.ac.uk

Helena Daffern
AudioLab, School of PET
University of York
York, United Kingdom
helena.daffern@york.ac.uk

Gavin Kearney
AudioLab, School of PET
University of York
York, United Kingdom
gavin.kearney@york.ac.uk

***Abstract*—Modern music delivery contexts include the real–time distribution of musical performances as immersive media ('metaverse concerts') to online audiences. This paper details an investigatory deployment of a novel server–based spatial audio rendering system for this context, considering distribution to practical wireless Virtual Reality (VR) audience endpoint. Ambisonic auralisation is provided with 6 Degrees of Freedom (6DoF) using a Spatial Impulse Response (SIR) interpolation approach which is efficient for large–scale distribution. Binaural audio is distributed to a mobile device at the audience endpoint using low–latency Network Music Performance (NMP) streaming software. Improvement in spatial audio rendering is achieved relative to that which is possible on consumer HMDs. Latency measurements suggest acceptable motion–sound latency is not achieved for the mobile device, however may be achievable in other distribution contexts, or with optimised systems.**

***Index Terms*—Networked Immersive Audio, Ambisonics, Binaural, Network Music Performance, Virtual Reality, Metaverse, Wireless Networks, 5G.**

I. Introduction

The term 'metaverse concert' broadly describes the distribution of musical performance to audiences as immersive media [1], [2]. One such case is the real–time distribution of musical performance to online audiences in virtual environments presented using game, VR and Extended Reality (XR) systems. This can include distribution of live performance from a broadcast studio, as is of particular interest to modern music industry[1]. Modern research also considers metaverse concert distribution of NMP, where remote musicians perform together over computer networks using low–latency audio streaming solutions [2], [5]. In these contexts the metaverse concert experience is presented to audience members in a navigable and interactive 3D environment. Spatial audio rendering is therefore required, to provide Audio–Visual (AV) coherence in navigation of the virtual environment [2].

The spatial audio rendering capabilities of consumer Head Mounted Displays (HMDs) used for VR and XR systems is, however, limited, and may not provide an optimal audio rendering solution for the distribution of musical performance content. Server–based spatial audio rendering [6], [7] is an alternative which removes processing load from client devices. Access to High Performance Computing (HPC) resource for spatial audio rendering [8], [9] may also be provided. This can allow for complex spatial audio rendering beyond that which is possible on consumer HMDs.

As part of ongoing investigation of the potential for server–based spatial audio in practical metaverse concert distribution scenarios a system was deployed which is the subject of ongoing development. A server endpoint located at the AudioLab, University of York provides 6DoF Ambisonic auralisation using an SIR interpolation approach which is efficient for large–scale distribution. A received NMP stream is auralised by the server for distribution to a typical VR audience client endpoint located in Glasgow, Scotland. Rendered binaural audio is streamed from the server to a mobile device at the client endpoint using low–latency NMP [10] streaming software over a consumer mobile network. Preliminary latency measurements suggest that acceptable motion–sound latency is not achieved in distribution to a mobile device, however may be achievable with optimised hardware and wireless networks. Acceptable motion–sound latency for wired client endpoints with desktop/laptop computing appears practically viable.

II. Related Work

A. Immersive Network Music Performance Systems

NMP systems use low–latency User Datagram Protocol (UDP) audio streaming methods, capable of supporting musical interaction between endpoints with negligible experience of delay[2] [10]. Recent research in the field of Immersive NMP (INMP)[3] explores server and edge computing solutions for spatial audio rendering [7], [12], [13]. Practical implementations provide virtual Ambisonic–based binaural rendering with 3DoF in response to webcam [12] or mounted [13] head–tracking solutions at client endpoints.

In INMP system design a motion–sound latency limit of 50 ms is considered for acceptable performance in spatial audio rendering [14]. Server and edge INMP solutions are capable of achieving spatial audio rendering and distribution with acceptable motion–sound latency over wireless networks.

[1]Several examples of mainstream metaverse concerts are given in [3] and [4]

[2]Empirical research estimates that musical interactivity will become impaired by latency in excess of 20–30 ms throughput [10].

[3]INMP refers to a subset of NMP systems which utilise immersive technology, and where spatial audio rendering is provided [11].

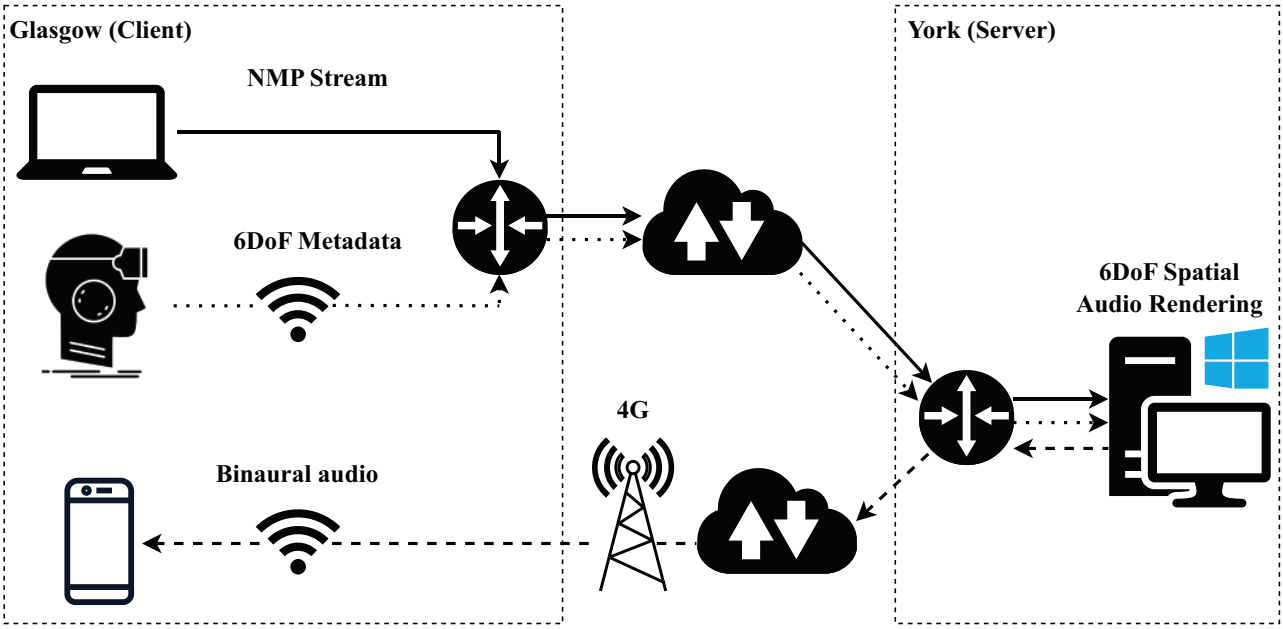

Fig. 1. System overview, showing hardware, network connections, and signal flow of the server–based spatial audio system deployed between Glasgow and York.

This is achieved using low–latency NMP streaming solutions[4]. Indeed previous practical deployment has reported 44ms Round Trip Time (RTT) latency measurement for audio between a 5G wireless client and an edge endpoint [13]. Estimations from network simulations have suggested that sub–30 ms One Way Trip (OWT) latency between clients and edge and server spatial audio rendering services for INMP is achievable on wireless 5G networks [7].

The investigatory system deployment detailed in this paper expands upon this prior work through the inclusion of 6DoF rendering. Server–based spatial audio INMP solutions are also repurposed to the case of practical audience distribution, considering distribution to typical consumer mobile and HMD devices.

B. 6DoF Ambisonic Auralisation

6DoF Ambisonic auralisation provides rendering of spatial audio and virtual acoustics allowing both rotational and positional listener navigation [2]. This can be achieved using sets of SIR measurements [16]. Here each SIR represents the sound field response around a discrete listener position for a fixed source position in a 3D acoustic environment. Auralisation with 6DoF may be computed by interpolating through SIR convolution outputs in response to listener tracking data. An SIR set from an acoustic survey of BBC Maida Vale Studio 4 (MV4)[5] [16] is utilised in the server–based system detailed in this paper (Figure 2).

The server–based system presented in this paper implements an SIR interpolation approach adapted from previous work by De Muynke et al [17]. This prior work evaluates perceived source stability for different SIR grid spacing and interpolation approaches. It is concluded that distance–based weighting of 3 Nearest Neighbour (3NN) SIR achieved good results with 1m SIR grid spacing. 3NN SIR weighting calculation is given as:

$$w_i = \frac{\frac{1}{d_i}}{\frac{1}{d_i} + \frac{1}{d_j} + \frac{1}{d_k}} \tag{1}$$

Where w_i and d_i are the SIR weighting and distance from listener position respectively for NN SIR_i.

III. System Overview

The server–based spatial audio system (Figure 1) was deployed between three endpoints (an NMP endpoint, a VR audience endpoint, and a server endpoint). These three endpoints are designed to be representative of networked audio components of metaverse concert distribution systems (i.e., a channel from NMP concert performance, a server providing spatial audio rendering services, and a VR client to which rendered audio is distributed). The three endpoints are situated across two physical locations, namely the AudioLab, University of York, and a researcher home in Glasgow, Scotland.

The deployment uses no specialised networks, using the existing networks at each site, in order for investigation to consider practical distribution scenarios. The audience endpoint specifically utilises consumer–grade hardware in order to similarly represent a practical scenario.

[4]Lists of modern NMP software solutions can be found in [10] and [15].
[5]https://doi.org/10.5281/zenodo.10020866

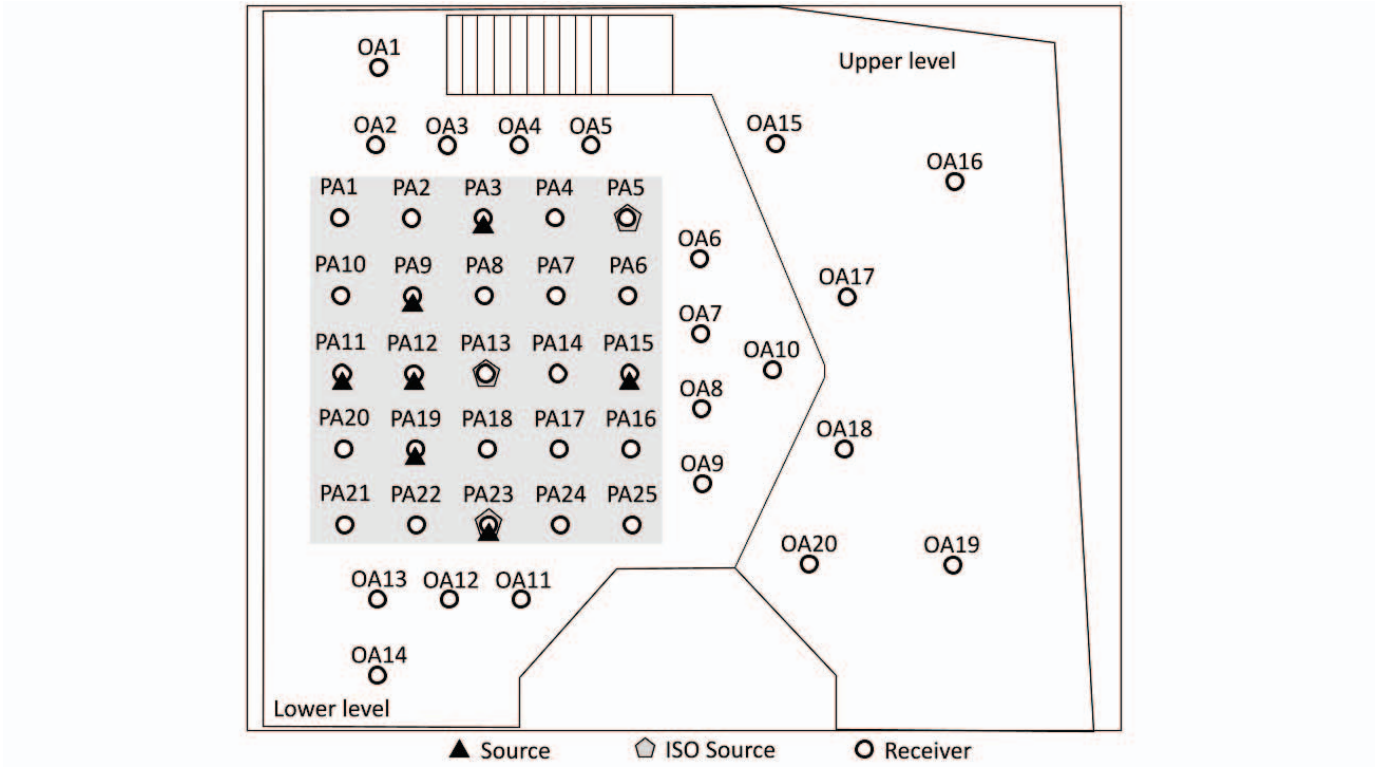

Fig. 2. A floorplan of MV4. SIR measurement positions are shown as detailed in [16]. The fixed source position PA3 is shown, as is the 6DoF listener area described by positions PA1–PA25.

A. NMP Endpoint

The NMP endpoint was located in Glasgow, Scotland. Here a laptop computer (*i7-11800H 2,3 GHz, 32 GB RAM, Win10*) with audio interface (*Focusrite Scarlet 2i2g3, 48 kHz, 24–bit, 128 sample buffer*) is connected (*wired Cat5e*) to the home network internet connection (*BT Smart Hub router, 9.17 Mbps upload/46.1 Mbps download*). Audio playback from Reaper[6] is used to represent a channel of audio capture from NMP concert. JackTrip [18] is used to stream this audio to the server via the consumer home internet connection (*two channels, jitter buffer length of four frames, and Forward Error Correction, FEC, redundancy of two frames*). Jack Audio Connection Kit (JACK)[7] is used to route audio between hardware buffers, Reaper, and JackTrip.

B. Server Endpoint

The server endpoint was located at the AudioLab, University of York. A standard managed desktop computer (*i7-6700K, 16 GB RAM, Win11*) with audio interface (*Focusrite Scarlet 2i2g2, 48 kHz, 24–bit, 128 sample buffer*) and a wired connection to the shared university network (*913.8 Mbps upload/818.5Mpbs download*) served as a pseudo–server device. A JackTrip stream is received from the NMP endpoint, and routed to Reaper, where audio rendering processes are hosted as detailed in section IV.

Open Sound Control (OSC) [19] data is received from the HMD at the audience endpoint which controls 6DoF rendering in Reaper. SonoBus[8] is used to provide an Opus–compressed [20] (*64 kbps*) stream of rendered binaural audio to a mobile device at the audience endpoint over a consumer mobile wireless network.

C. Audience Endpoint

The VR audience endpoint was situated at the same location as the NMP endpoint in Glasgow, Scotland. A HMD (*Oculus Quest 2*) runs a custom Unity application. 6DoF rendering metadata is derived from avatar transforms and streamed to the server as OSC messages (*ExtOSC package*[9]). This data is streamed via the wireless 5G connection of the same home network used by the NMP endpoint (subsection III-A).

A mobile device (*Google Pixel 6, Android 14, 48kHz, 256 sample buffer*) received an Opus–compressed binaural stream from the server over a wireless mobile network (*4G coverage, 3 Mbps upload/20 Mbps download*) using the NMP software SonoBus. Headphone (*Beyerdynamic DT990 Pro*) output from the mobile phone (notably here, a low–grade consumer USB-c–3.5 mm converter with internal Digital to Analogue Converter, DAC) is used for audio playback.

[6]https://www.reaper.fm/
[7]https://jackaudio.org/
[8]https://www.sonobus.net/
[9]https://github.com/Iam1337/extOSC

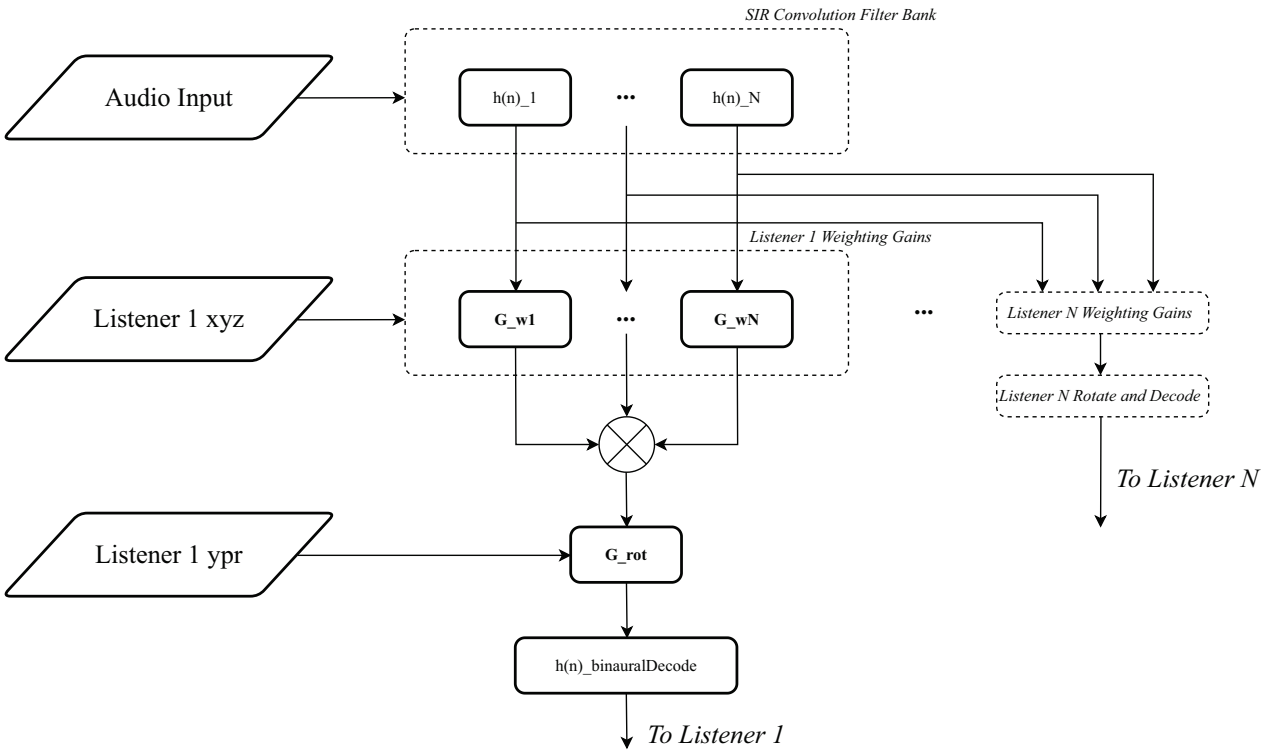

Fig. 3. Audio rendering signal flow for server–based 6DoF auralisation. Audio input may be a received NMP stream or local capture. The audio source is convolved with SIR describing each measurement position in a dataset ($h(n)_1$ to $h(n)_N$ for a set of N measurement positions). SIR weighting gains for each listener (G_{w1} to G_{wN}) are derived for each SIR convolution output (as defined in Equation 1) in response to positional data (xyz) from each listener. Per–listener 3DoF is then computed normally as the Ambisonic rotation matrix G_{rot} in response to listener rotation data (yaw, pitch, roll, ypr). Binaural decoding is then achieved through convolution with decoding filters $h(n)_{binauralDecode}$.

Notably a mobile device was used for audio distribution and playback here. Game development tools used in the creation of VR applications do not include readily available NMP streaming functionality. NMP streaming software (i.e., SonoBus) is, however, readily available for mobile devices. Mobile devices are also a typical consumer wireless endpoint which is suitable for usage with 6DoF movement, and were therefore considered as a solution for low–latency NMP distribution to typical VR audience endpoints.

IV. Audio Rendering

The server–based spatial audio design considered in this paper receives audio input from local capture at the server endpoint, or from a received NMP stream. An input audio source is auralised as a fixed source for multiple listeners. As detailed in subsection II-B the SIR interpolation approach is used. This approach has been adapted to consider scalability for large numbers of listeners, as is required in metaverse concert distribution contexts.

A. Efficient SIR Interpolation

3NN interpolation (as detailed in subsection II-B) is computed as a weighted sum of SIR convolution outputs. It is typical here to consider an instance of SIR interpolation processes per–listener. In this case, for a set of K listeners, $3 * K$ active convolutions are required. Here computational expense of SIR convolution scales directly with number of listeners.

As the SIR auralisation for each listener may be expressed as a weighted sum of SIR outputs, this paper considers a capped finite N convolutions[10] which may be shared for all listeners. Here the per–listener process of 6DoF SIR interpolation is computed as multichannel gains controlling the weighting of all N SIR outputs for each discrete listener[11] (Figure 3). Using this approach the number of SIR convolutions is capped at a known quantity, and the per–listener scaling of computational load becomes multichannel gains rather than multichannel convolutions for SIR interpolation processes. Following SIR interpolation 3DoF rotation is applied in response to HMD rotation metadata. The 6DoF auralisation is then decoded to binaural stereo (where the MagLS appraoch is currently standard [21]) and streamed to the relevant audience client.

[10]Reflecting the N SIR measurement positions described in an SIR dataset.

[11]With 3NN SIR weighted as per Equation 1 and all other SIR weights set to 0.

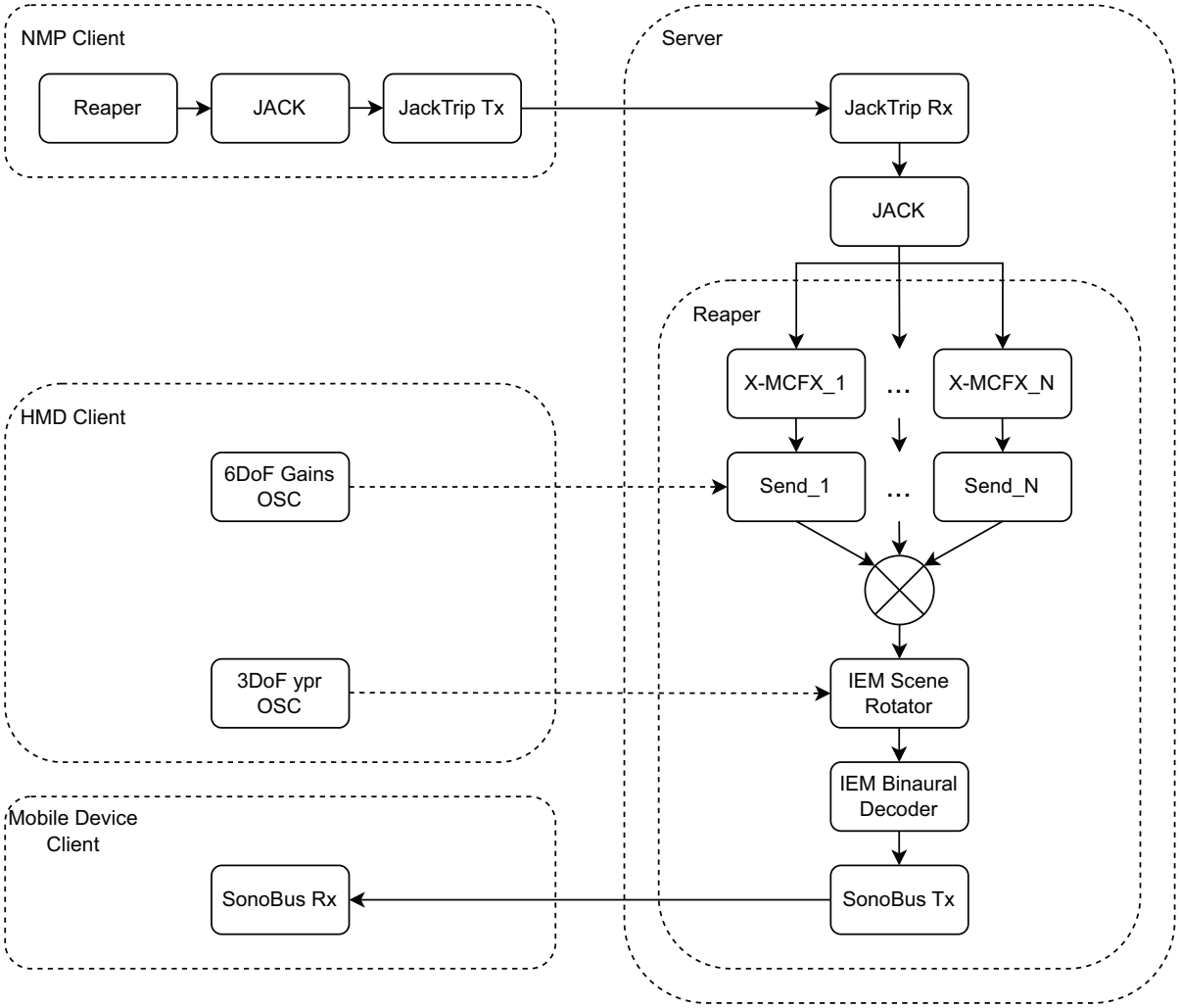

Fig. 4. Signal flow overview of the proof–of–concept implementation. Software solutions used for internal routing, streaming and audio rendering processes are detailed.

B. Investigatory Implementation

In current investigation a proof–of–concept implementation of the detailed 6DoF auralisation design is deployed. A single NMP source and a single 6DoF listener are considered. The source is auralised for the listener in the acoustics of MV4 using the dataset described in subsection II-B. These SIR are in 3rd Order Ambisonic (3OA) format. The position PA3 was selected for the fixed source location. Positions PA1–PA25 (describing a 5x5 grid with 1m spacing along length and width room axis) were used for 6DoF listener auralisation (Figure 2). Prior to use the 25 SIR used were aligned by truncating the beginning of each SIR to the direct sound.

Audio rendering processes are hosted in Reaper (Figure 4). 25 instances of X-MCFX Convolver[12] were active, each representing one of the 25 receiver positions from the MV4 SIR dataset. SIR weighting gains are computed as send gains from each of the convolution outputs. These weighting gains are received as OSC from the audience client HMD as detailed in section III. The output of the weighted sends is summed, providing the interpolated SIR auralisation.

The rendered scene is rotated using IEM Scene Rotator[13], receiving rotational data from the audience client HMD as OSC. Binaural decoding of the 6DoF auralisation is then decoded to binaural stereo using IEM Binaural Decoder. SonoBus is then used to stream binaural audio to a mobile device at the audience client endpoint as detailed in section III.

[12]https://github.com/JB-Luke/X-MCFX
[13]https://plugins.iem.at/

TABLE I
LATENCY ESTIMATES FROM CURSORY MEASUREMENT.

Endpoint	Measured OWT from NMP Endpoint	Network OWT from Server	Full System RTT from Server
NMP Endpoint	-	24 ms	83 ms
Mobile (4G Mobile Network)	140 ms	51 ms	239 ms
Mobile (5G WLAN)	115 ms	26 ms	189 ms

V. LATENCY MEASUREMENT

A brief pulse test signal was used to provide cursory measurement of latency[14] between points on the network (Table I), taken as the difference between the leading edge of the test signal at input and output endpoints. Prior to each latency measurement music was streamed along the relevant signal path to allow cursory listening for the purpose of ensuring stream stability (that audible packet loss was not occurring). No objective test of stream reliability has been taken at this stage of development. Streaming parameters were fixed as detailed in section III when a stable stream was achieved. A static latency of 21.3 ms was incurred in audio rendering at the server[15].

Network latency between the NMP endpoint and the server was tested. The test signal was output from Reaper at the NMP endpoint, streamed to the server, and looped back. A network RTT of 48 +/- 2 ms was measured. Hardware capture and playback latency for the NMP endpoint was a static 14 ms. A system RTT of 62 ms and OWT of 31 ms was therefore estimated between the NMP endpoint and the server. Adding immersive audio rendering an RTT latency of 83 ms is estimated here.

Latency between the NMP endpoint and audience endpoint mobile device was also measured. The test signal was sent from Reaper at the NMP endpoint and routed to the audience endpoint mobile device audio output. Two network connections were tested for the mobile device, namely the 4G mobile network, and wireless 5G connection to the home network.

When connected to the mobile service provider 4G network a jitter buffer size of 18 ms was required for stable audio. A OWT latency of 140 +/- 5 ms was measured. Subtracting 31 ms (the path between the NMP endpoint and server) gives an estimated OWT of 109 ms between the server and mobile device (and RTT estimate of 218 ms). Adding immersive audio rendering latency results in an estimated RTT of 239 ms between the mobile device and server. With a wireless 5G connection to the home network a jitter buffer size of 8 ms was sufficient for stable audio. A OWT latency of 115 +/- 5 ms was measured between the NMP endpoint and mobile device. Subtracting the 31 ms path between the NMP endpoint and server an OWT and RTT are estimated as 84 ms and 168 ms respectively between the mobile device and server. Adding spatial audio rendering, RTT is estimated as 189 ms.

[14]Breakdowns of system component and process latency for INMP can be found in [10] and [14].

[15]This is incurred through the use of large convolution first partitions required to process a large number of convolutions without audio dropout on this particular device.

The throughput latency of the mobile device was measured by capturing an impulse sound (clap) with a microphone and via the mobile device audio output and measuring the difference. A hardware throughput of 58 ms was measured for the mobile device. Subtracting this hardware delay from OWT estimates between the mobile device and server gives an estimate of network latency between the endpoints. A network OWT delay of 51 ms for the 4G mobile network and 26 ms for wireless 5G connection to the home network is estimated.

VI. DISCUSSION

Estimated system RTT for the NMP endpoint and server only slightly exceed the 50 ms motion–sound threshold. The investigatory deployment is, however, sub–optimal. Audio interface latency (14 ms RTT) can be substantially reduced through hardware improvement[16]. Audio rendering latency (21.3 ms) may also be reduced to negligible levels using HPC systems [7] rather than the standard managed desktop considered in the investigatory deployment. With such optimisations it seems likely that the server–based spatial audio distribution system will perform with acceptable motion–sound latency for this endpoint type. Indeed estimated network RTT between the NMP endpoint and server (48 ms) certainly implies that a sub–50 ms RTT is achievable for laptop and desktop clients with a wired network connection. Such endpoints correspond with audiences using game systems or audiences in immersive venues. This endpoint type is not, however, typically suitable as a consumer VR audience endpoint where 6DoF physical motion is considered.

RTT latency estimates between the server and audience mobile device demonstrate values far in excess of the acceptable 50 ms motion–sound threshold. The investigatory system is, again in this case, sub–optimal. Only 4G mobile coverage of the area for distribution was available over the mobile network. Improved results can be expected over 5G networks [7]. Again server processing latency may be reduced to near–zero by replacing the pseudo–server workstation with HPC resource. Particularly the mobile device hardware buffering incurs a significant amount of latency. This may be improved through the use of USB–c to 3.5 mm converters with low latency internal DAC. With such optimisations it may be possible for wireless mobile devices to achieve motion–sound latency which is acceptably close to the recommended 50 ms RTT. Indeed previous work certainly records sub–30 ms OWT between endpoints using NMP solutions with laptops, workstations, and embedded solutions [13], [22], [23].

[16]For example 5.3 ms audio interface RTT is achieved in [4].

Such optimisations are, however, specialist, and cannot be expected at a typical consumer VR audience endpoint for practical metaverse concert contexts. The investigation in this paper suggests that sub–50 ms motion–sound latency may not be achievable for stable streaming to mobile devices over practical wireless networks. Indeed previous work which evaluates NMP streaming over practical wireless networks suggests that stable streams are unlikely to be achieved with sub–30 ms OWT between two endpoints [24], indicating that sub–50 ms RTT is unlikely also. Though high–latency is expected in practical server–based spatial audio distribution to mobile device endpoints, a high–quality 6DoF auralisation was achieved. It is considered that this may still yield experience improvement relative to on–board HMD rendering despite high motion–sound latency.

VII. Further Work

The system presented in this paper is investigatory at this stage, and is the subject of ongoing development. As such many improvements may be made. Visual rendering should be addressed, as this is understood to be important in virtual performance experiences [25]. Audio rendering methods may be improved, particularly the SIR interpolation approach used, which can contain artefacts [17]. System optimisations as outlined in section VI should be implemented and deployed for further measurements. Reliability (i.e., packet loss) metrics should be included in further measurements. Formal latency testing, rather than cursory estimates, should be undertaken. Comparative evaluation of server–based and HMD–based spatial audio rendering should be conducted. Other distribution contexts should be explored, including distribution to game systems and immersive venues.

VIII. Conclusion

A novel system design for server–based spatial audio using the NMP streaming approach has been presented. 6DoF auralisation is provided, and distribution of metaverse concert content to a practical VR audience endpoint via a wireless mobile device is considered. A practical investigatory deployment of the system over real–world networks is detailed, as is related latency measurements used to suggest performance of the system with respect to motion–sound latency thresholds.

Discussion of wired workstation and laptop endpoints considers that prior systems [12], [13] which provide 3DoF rendering may be extended to 6DoF while maintaining acceptable motion–sound latency. This may be of use in venue distribution contexts, and for distribution to audience endpoints using game systems. Distribution to VR audience endpoints via wireless mobile devices may be possible with an optimised system. It is considered here that practical distribution considering typical consumer hardware is, however, likely to exceed acceptable motion–sound latency. Notably the ‘weak link’ in this distribution path is the mobile device hardware buffering, which incurs significant latency.

The server–based system does, however, allow access to audio rendering capability beyond what is possible on board a HMD. It is considered that this audio rendering improvement may still result in experience improvement despite poor motion–sound latency.

Acknowledgments

This research is supported in part by the UK AHRC XR Stories project, grant no. AH/S002839/1, and in part by a University of York funded PhD studentship.

References

[1] L. Turchet, “Musical Metaverse: vision, opportunities, and challenges,” *Personal and Ubiquitous Computing*, pp. 1–17, 2023, publisher: Springer.

[2] J.-M. Jot, R. Audfray, M. Hertensteiner, and B. Schmidt, “Rendering Spatial Sound for Interoperable Experiences in the Audio Metaverse,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. Bologna, Italy: IEEE, Sep. 2021, pp. 1–15. [Online]. Available: https://ieeexplore.ieee.org/document/9610971/

[3] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, “Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios,” *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, 2023, publisher: Audio Engineering Society.

[4] P. Cairns, T. Rudzki, J. Cooper, A. Hunt, K. Steele, G. Acosta Martínez, A. Chadwick, H. Daffern, and G. Kearney, “Singer and Audience Evaluations of a Networked Immersive Audio Concert,” *Journal of the Audio Engineering society*, vol. 72, pp. 467–478, july 2024.

[5] D. Dziwis, H. Von Coler, and C. Pörschmann, “Orchestra: A Toolbox for Live Music Performances in a Web-Based Metaverse,” *Journal of the Audio Engineering Society*, vol. 71, no. 11, pp. 802–812, Nov. 2023. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22345

[6] C. Rinaldi, F. Franchi, A. Marotta, F. Graziosi, and C. Centofanti, “On the Exploitation of 5G Multi-Access Edge Computing for Spatial Audio in Cultural Heritage Applications,” *IEEE Access*, vol. 9, pp. 155 197–155 206, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9617628/

[7] 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. [Online]. Available: https://ieeexplore.ieee.org/document/10543134/

[8] D. Moore and J. Wakefield, “The Potential of High Performance Computing in Audio Engineering,” in *AES 126th Convention*, Munich, Germany, May 2009.

[9] L. Turchet, F. Vella, and S. L. Fiore, “The potential of high-performance computing for the Internet of Sounds,” in *2023 4th International Symposium on the Internet of Sounds*. Pisa, Italy: IEEE, Oct. 2023, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10335270/

[10] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, “An Overview on Networked Music Performance Technologies,” *IEEE Access*, vol. 4, pp. 8823–8843, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7769205/

[11] P. Cairns, G. Kearney, and H. Daffern, “Immersive Network Music Performance: Design and Practical Deployment of a system for Immersive Vocal Performance,” in *AES 149th Convention*, Online, Oct. 2020. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=20918

[12] R. Hoy and D. Van Nort, “Towards Accessible and Embodied Control of Telematic Sonic Space Through Browser-Based Facial Tracking,” in *2023 4th International Symposium on the Internet of Sounds*. Pisa, Italy: IEEE, Oct. 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10335395/

[13] F. Martusciello, C. Centofanti, C. Rinaldi, and A. Marotta, “Edge-Enabled Spatial Audio Service: Implementation and Performance Analysis on a MEC 5G Infrastructure,” in *2023 4th International Symposium on the Internet of Sounds*. Pisa, Italy: IEEE, Oct. 2023, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/10335480/

[14] L. Turchet and M. Tomasetti, "Immersive networked music performance systems: identifying latency factors," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. Bologna, Italy: IEEE, Sep. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10289169/

[15] M. Bosi, A. Servetti, C. Chafe, and C. Rottondi, "Experiencing Remote Classical Music Performance Over Long Distance: A JackTrip Concert Between Two Continents During the Pandemic," *Journal of the Audio Engineering Society*, vol. 69, no. 12, pp. 934–945, Dec. 2021. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21542

[16] G. Kearney, H. Daffern, P. Cairns, A. Hunt, B. Lee, J. Cooper, P. Tsagkarakis, T. Rudzki, and D. Johnston, "Measuring the Acoustical Properties of the BBC Maida Vale Recording Studios for Virtual Reality," *Acoustics*, vol. 4, no. 3, pp. 783–799, Sep. 2022. [Online]. Available: https://www.mdpi.com/2624-599X/4/3/47

[17] J. De Muynke, D. Poirer-Quinot, and B. F. G. Katz, "{Creating navigable auralisations using RIR convolution: Impact of grid density and panning method on perceived source stability}," in {*AES 2023 International Conference on Spatial and Immersive Audio*}, Huddersfield, United Kingdom, Aug. 2023. [Online]. Available: https://hal.science/hal-04259798

[18] J.-P. Cáceres and C. Chafe, "JackTrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010, publisher: Taylor & Francis.

[19] M. Wright, A. Freed, and A. Momeni, "2003: OpenSound Control: State of the Art 2003," in *A NIME Reader*, A. R. Jensenius and M. J. Lyons, Eds. Cham: Springer International Publishing, 2017, vol. 3, pp. 125–145, series Title: Current Research in Systematic Musicology. [Online]. Available: http://link.springer.com/10.1007/978-3-319-47214-0_9

[20] B. Lee, T. Rudzki, J. Skoglund, and G. Kearney, "Context-Based Evaluation of the Opus Audio Codec for Spatial Audio Content in Virtual Reality," *Journal of the Audio Engineering Society*, vol. 71, no. 4, pp. 145–154, 2023, publisher: Audio Engineering Society.

[21] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *DAGA 2018*, Mar. 2018.

[22] A. Carot, F. Sardis, M. Dohler, S. Saunders, N. Uniyal, and R. Cornock, "Creation of a Hyper-Realistic Remote Music Session with Professional Musicians and Public Audiences Using 5G Commodity Hardware," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. London, UK: IEEE, Jul. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9105995/

[23] L. Turchet and P. Casari, "Latency and Reliability Analysis of a 5G-Enabled Internet of Musical Things System," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 1228–1240, Jan. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10162178/

[24] J. Dürre, N. Werner, S. Hämäläinen, O. Lindfors, J. Koistinen, M. Saarenmaa, and R. Hupke, "In-depth latency and reliability analysis of a networked music performance over public 5G infrastructure," in *Audio Engineering Society Convention 153*. Audio Engineering Society, 2022.

[25] A. Hunt, H. Daffern, and G. Kearney, "Avatar representation in extended reality for immersive network music performance," in *AES 2023 International Conference on Spatial and Immersive Audio*, Huddersfield, Aug 2023, paper 35. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=22183