Immersive Audio and Network Music Performance: Practical Applications and User Testing

Patrick Cairns

Doctor of Philosophy

University of York

School of Physics, Engineering and Technology

February 2025

Abstract

Network Music Performance (NMP) technology allows remote musicians to perform together in real–time over Wide Area Networks (WANs) using low–latency audio streaming methods. Musicians hear and interact with one another, experiencing musical timing as if performing together live. The extension of NMP systems with immersive audio (Immersive Network Music Performance, INMP) allows for simulation of hearing networked performance as if cohabiting a physical 3–Dimensional (3–D) acoustic environment with remote performers. This provides an enhancement of traditional NMP systems, and provides Audio–Visual (AV) coherence in navigation of 3–D immersive display and interface using technologies such as Virtual and Extended Reality (VR/XR).

This thesis proposes that low–latency INMP systems are viable as a practical remote performance workflow, and that immersive audio will improve the User Experience (UX). A series of studies contribute documentation of low–latency INMP system designs which have been implemented in practical use–case contexts, practice–oriented user–testing of these designs, and investigation of the experience of immersive audio.

The first three studies contribute case–studies of INMP systems in the use–cases of remote education, remote virtual production, and remote concert. Results demonstrate the viability of these solutions for supporting remote synchronous performance and positive performance experience with minimal prior training. UX investigation demonstrates the potential for experience improvement and deterioration associated with immersive audio, and highlights relevant practical challenges, informing recommendations for INMP experience optimisation. Novel demonstration of performer engagement with video and VR/XR systems is provided, and limitations of visual display and interface methods are highlighted.

A final study contributes testing of the influence of latency and acoustic parameters in INMP for vocal performance. Results provide novel demonstration of variation in INMP experience quality for different Reverb Times. Here findings inform recommendations for INMP experience optimisation.

Acknowledgments

Gratitude is expressed to:

Prof. Helena Daffern and Prof. Gavin Kearney for supervising this project and supporting research ‘in the wild’.

The various technical magicians at the AudioLab for camaraderie, technical expertise and support, for being part of bringing INMP from theory to practice, and for helping me unwind with some TTRPG.

The various musicians who participated in various events and studies, including Jamie Cullum, Wolph, Ex Corde Vocal Ensemble and various others.

The various industry partners involved in related work, including Bonza Music, the BBC, and Opera North.

Various relevant funding, including the AHRC XR Stories project (AH/S002839/1), EPSRC IAA MINERVA project (EP/R51181X/1), EPSRC IAA SAFFIRE project (EP/X525856/1) and a University of York funded studentship.

This thesis is dedicated to my family, who’s support and encouragement this work would not be possible without.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. All sources are acknowledged as references. This work has not previously been presented for an award at this, or any other, University. Parts of this research have been presented in the following publications:

- P. Cairns, A. Hunt, J. Cooper, *et al.*, "Recording Music in the Metaverse: A case study of XR BBC Maida Vale Recording Studios," in *Proceedings of the AES 2022 International Conference on Audio for Virtual and Augmented Reality*, Redmond, Washington, USA: Audio Engineering Society, Aug. 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21842

- P. Cairns, A. Hunt, D. Johnston, *et al.*, "Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, Jun. 2023, ISSN: 15494950. DOI: 10.17743/jaes.2022.0086. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22139 (visited on 02/25/2025)

- P. Cairns, H. Daffern, and G. Kearney, "A DAW-based approach to immersive audio system evaluation in Network Music Performance contexts," in *Proceedings of the 2023 AES International Conference on Spatial and Immersive Audio*, Huddersfield, United Kingdom: Audio Engineering Society, Aug. 2023. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22186

- P. Cairns, T. Rudzki, J. Cooper, *et al.*, "Singer and Audience Evaluations of a Networked Immersive Audio Concert," *Journal of the Audio Engineering Society*, vol. 72, no. 7/8, pp. 467–478, Jul. 2024, ISSN: 15494950. DOI: 10.17743/jaes.2022.0145. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22644 (visited on 02/25/2025)

- G. Kearney, H. Daffern, P. Cairns, *et al.*, "Measuring the Acoustical Properties of the BBC Maida Vale Recording Studios for Virtual Reality," en, *Acoustics*, vol. 4, no. 3, pp. 783–799, Sep. 2022, ISSN: 2624-599X. DOI: 10.3390/acoustics4030047. [Online]. Available: https://www.mdpi.com/2624-599X/4/3/47 (visited on 07/03/2024)

Contents

Abstract **2**

Acknowledgments **3**

Declaration **4**

Contents **5**

List of Figures **12**

List of Tables **19**

List of Acronyms **22**

1 Introduction **27**
1.1 Statement of Hypothesis 29
1.2 Objectives 29
1.3 Novel Contributions 30
1.4 Thesis Outline 30
1.5 Summary 32

2 Review of Network Music Performance **33**
2.1 Preface 33
2.2 A Brief History of Network Music Performance 33
2.3 Latency Challenges in Network Music Performance 34
2.4 NMP System and Latency Overview 36
2.4.1 Air Propagation, Sound Acquisition and Monitoring 37
2.4.2 Capture and Playback Processes 38
2.4.3 Audio Processing and Application Buffering 38
2.4.4 Network Transport 39
2.5 Low–Latency Audio Streaming Methods for NMP 41
2.5.1 UDP Audio Streaming 41
2.5.2 Jitter Buffering and Packet Sequencing 41
2.5.3 Packet Loss Concealment 42
2.5.4 Clock Synchronisation 43
2.5.5 Session Management 44

2.5.6 Network Architecture . . . 45
2.5.7 Audio Compression . . . 46
2.6 NMP Visual Display . . . 47
2.7 Relevant NMP System Software . . . 47
2.7.1 Audio Routing: Jack Audio Connection Kit . . . 48
2.7.2 Audio Processing Application: Reaper . . . 48
2.7.3 NMP Application: JackTrip . . . 49
2.7.4 NMP Application: SoundJack . . . 49
2.7.5 NMP Application: SonoBus . . . 50
2.8 Evaluation of NMP . . . 50
2.8.1 Synchrony Metrics . . . 51
2.8.2 Quality of Experience Metrics . . . 53
2.8.3 Immersive Experience Modelling and Presence Questionnaires . . . 54
2.8.4 Semi–Structured Interviews . . . 56
2.8.5 Gaze and Motion Capture . . . 57
2.9 Previous Study of Network Music Performance . . . 57
2.9.1 Effect of Latency on Performance Synchrony . . . 57
2.9.2 Effect of Latency on Quality of Experience . . . 58
2.9.3 Effect of Latency on Immersion and Presence . . . 58
2.9.4 Latency–Coping Strategies and Metronomes . . . 58
2.9.5 Effect of Performance Timbre and Dynamics . . . 59
2.9.6 Effect of Performance Material . . . 59
2.9.7 Effect of Prior Practice . . . 60
2.9.8 Evaluation of Visual Display . . . 60
2.10 Musical Performance Practice for NMP . . . 60
2.10.1 Performance Practice in NMP Research . . . 60
2.10.2 Performance Practice in Practical NMP Applications . . . 61
2.11 Summary . . . 62

3 Review of Immersive Audio for Network Music Performance 64
3.1 Preface . . . 64
3.2 A Brief Overview of INMP . . . 64
3.3 Fundamentals of Immersive Audio . . . 65
3.3.1 Sound in 3–D Space . . . 66
3.3.2 Sound in 3–D Acoustic Spaces . . . 67
3.3.3 Acoustic IR Measurement and Simulation . . . 69
3.3.4 Auralisation . . . 70
3.3.5 Interactive Auralisation . . . 74
3.4 Binaural Stereo . . . 75
3.4.1 Duplex Theory and the Head Related Transfer Function . . . 76
3.4.2 Binaural Recording and HRTF Measurement . . . 76
3.4.3 Binaural Auralisation . . . 77
3.5 Ambisonics . . . 79
3.5.1 Spherical Harmonic Representation of Spatial Sound . . . 79
3.5.2 Ambisonic Format, Recording and SIR Measurement . . . 80

3.5.3 Ambisonic Auralisation . . . 83
3.5.4 Loudspeaker Decoding . . . 84
3.5.5 Binaural Decoding . . . 85
3.6 INMP Audio System Design . . . 86
3.6.1 Spatial Audio Microphone Systems . . . 86
3.6.2 Free–Field Auralisation . . . 87
3.6.3 Acoustic Auralisation . . . 88
3.6.4 Algorithmic Auralisation . . . 89
3.6.5 Loudspeaker Rendering . . . 89
3.6.6 Binaural Rendering . . . 90
3.6.7 Streaming Format and Network Architecture . . . 92
3.7 Evaluation of INMP . . . 93
3.7.1 Spatialisation Metrics . . . 94
3.7.2 Acoustic Metrics . . . 94
3.7.3 Comparative Audio System Evaluation . . . 95
3.7.4 Semi–Structured Interviews . . . 95
3.8 Previous Study of INMP . . . 95
3.8.1 Comparison to Mono and Stereo Audio . . . 96
3.8.2 Effect of Virtual Acoustics . . . 96
3.8.3 Effect of Spatialisation . . . 97
3.9 Summary . . . 97

4 Evaluation of an INMP System for Remote Music Education 99
4.1 Preface . . . 99
4.2 An INMP System for Remote Music Education . . . 100
4.3 Study Overview . . . 102
4.3.1 Motivation . . . 102
4.3.2 Objectives . . . 103
4.3.3 Hypothesis . . . 103
4.3.4 Statement of Ethics . . . 103
4.4 System Overview . . . 104
4.4.1 Hardware and Physical Routing . . . 104
4.4.2 Software Overview . . . 106
4.4.3 Audio Networking . . . 106
4.4.4 Audio Rendering . . . 107
4.5 BRIR Measurement . . . 108
4.5.1 Measurement Method . . . 109
4.5.2 BRIR Processing . . . 110
4.5.3 Acoustic Parameters . . . 110
4.6 Evaluation Protocol . . . 111
4.6.1 Protocol Overview . . . 111
4.6.2 Latency Conditions . . . 112
4.6.3 Participants . . . 113
4.6.4 Performance Task . . . 113
4.6.5 Questionnaire . . . 114

4.6.6 Exit Interview . . . 114
4.6.7 Study Limitations . . . 115
4.7 Results . . . 115
4.7.1 Experience Ratings . . . 115
4.7.2 Exit Interview Response . . . 119
4.8 Discussion . . . 121
4.8.1 Overall System Evaluation . . . 121
4.8.2 Experience of Different Audio Rendering Methods . . . 121
4.8.3 Experience of Video Display . . . 123
4.8.4 Experience of Different Participants . . . 123
4.8.5 Experience of Different Performance Material . . . 124
4.9 Hypothesis Revisited . . . 124
4.10 Summary . . . 126

5 Evaluation of an XR INMP System for Remote Virtual Production 127
5.1 Preface . . . 127
5.2 An XR INMP System for Remote Virtual Production . . . 129
5.3 Study Overview . . . 130
5.3.1 Motivation . . . 130
5.3.2 Objectives . . . 132
5.3.3 Hypothesis . . . 132
5.3.4 Statement of Ethics . . . 132
5.4 System Overview . . . 133
5.4.1 Hardware Overview . . . 134
5.4.2 Software Overview . . . 134
5.4.3 Audio Networking . . . 135
5.4.4 Tracking Metadata Networking . . . 136
5.4.5 XR Application . . . 136
5.4.6 Immersive Audio Rendering . . . 138
5.5 SIR Measurement . . . 141
5.6 Evaluation Method . . . 142
5.6.1 Protocol Overview . . . 142
5.6.2 Participants . . . 143
5.6.3 Latency Conditions . . . 143
5.6.4 Performance Task . . . 144
5.6.5 Questionnaire . . . 144
5.6.6 Synchrony Analysis . . . 145
5.6.7 Exit Interview . . . 145
5.6.8 Study Limitations . . . 146
5.7 Results . . . 146
5.7.1 QoE Rating . . . 146
5.7.2 Synchrony Analysis . . . 151
5.7.3 Exit Interview Response . . . 153
5.8 Discussion . . . 154
5.8.1 Overall System Evaluation . . . 154

5.8.2 Experience of Different Latency Conditions . . . 156
5.8.3 Experience of Different Audio Rendering Methods . . . 156
5.8.4 Experience of Visual Display and Interface . . . 158
5.8.5 Experience of Different Participants . . . 159
5.9 Hypothesis Revisited . . . 159
5.10 Summary . . . 160

6 Evaluation of an INMP System for Remote Concert Performance 162
6.1 Preface . . . 162
6.2 An INMP System for Remote Concert Performance . . . 163
6.3 Study Overview . . . 164
6.3.1 Motivation . . . 165
6.3.2 Research Question and Objectives . . . 166
6.3.3 Statement of Ethics . . . 166
6.4 System Overview . . . 167
6.4.1 Hardware Configuration . . . 167
6.4.2 Software Overview . . . 169
6.4.3 Audio Networking . . . 170
6.4.4 Immersive Audio Rendering . . . 170
6.5 SIR Measurements . . . 173
6.6 Part I: Evaluation Method . . . 175
6.6.1 Protocol Overview . . . 175
6.6.2 Participants . . . 176
6.6.3 Performance Material . . . 176
6.6.4 Latency Conditions . . . 177
6.6.5 Performer Questionnaire . . . 177
6.6.6 Audience Questionnaire . . . 178
6.6.7 Performer Semi–Structured Interview . . . 179
6.6.8 Informal Testing of Loudspeaker Monitoring . . . 179
6.6.9 Study Limitations . . . 180
6.7 Part I: Results . . . 181
6.7.1 Audience Rating Results . . . 181
6.7.2 Performer Rating Results . . . 182
6.7.3 Performer Interview Response and Feedback . . . 183
6.8 Part I: Discussion . . . 185
6.8.1 Overall Concert Experience . . . 185
6.8.2 Experience of Immersive Audio Rendering . . . 187
6.8.3 Experience of Visual Display and Interface . . . 189
6.9 Part II: Evaluation Method . . . 189
6.9.1 Protocol Overview . . . 190
6.9.2 Participants . . . 190
6.9.3 Listening Material . . . 190
6.9.4 Rating Questionnaire . . . 191
6.9.5 Study Limitations . . . 192
6.10 Part II: Results . . . 193

6.11 Part II: Discussion . . . 194
6.12 Research Question Revisited . . . 195
6.13 Summary . . . 196

7 Evaluation of the effect of Reverb Time on vocal performance using an INMP system 197
7.1 Preface . . . 197
7.2 An INMP System for Remote Vocal Performance . . . 198
7.3 Study Overview . . . 198
7.3.1 Motivations . . . 199
7.3.2 Objectives . . . 200
7.3.3 Hypothesis . . . 200
7.3.4 Statement of Ethics . . . 200
7.4 System Overview . . . 201
7.4.1 Hardware and Physical Routing . . . 201
7.4.2 Audio Networking . . . 201
7.4.3 Latency Measurement . . . 203
7.4.4 Immersive Audio Rendering . . . 204
7.4.5 Test Protocol GUI and Control System . . . 206
7.5 SIR Measurement and Processing . . . 209
7.5.1 SIR Measurement . . . 209
7.5.2 SIR Processing . . . 209
7.6 Evaluation Method . . . 214
7.6.1 Protocol Overview . . . 214
7.6.2 Participants . . . 214
7.6.3 Acoustic Conditions . . . 217
7.6.4 Latency Conditions . . . 217
7.6.5 Performance Task . . . 218
7.6.6 Experience Rating Questionnaire . . . 218
7.6.7 Synchrony Metrics . . . 219
7.6.8 Onset Detection Method: MM–BOP . . . 220
7.6.9 Statistical Analysis Method . . . 221
7.6.10 Sample Size Calculation . . . 223
7.6.11 Visual Contact . . . 224
7.6.12 Study Limitations . . . 224
7.7 Results . . . 224
7.7.1 Data Collection Errors . . . 224
7.7.2 Normality Tests . . . 225
7.7.3 Experience Rating: Q1, Perception of Delay . . . 227
7.7.4 Experience Rating: Q2, Perception of Synchrony . . . 231
7.7.5 Experience Rating: Q3, Perception of Clarity . . . 235
7.7.6 Experience Rating: Q4, Perception of Reverb Level . . . 238
7.7.7 Experience Rating: Q5, Perception of Reverberation Length . . . 241
7.7.8 Within–Parts Synchrony: Tempo Slope . . . 244
7.7.9 Between–Parts Synchrony: Precision . . . 247

7.7.10 Between–Parts Synchrony: Asynchrony . . . 248
7.7.11 Between Parts Synchrony: Lag . . . 251
7.8 Discussion . . . 252
7.8.1 Effect of Latency . . . 252
7.8.2 Effect of Acoustics . . . 253
7.9 Hypothesis Revisited . . . 255
7.10 Summary . . . 256

8 Conclusion **257**
8.1 Preface . . . 257
8.2 Summary of Work and Contributions . . . 257
8.2.1 Chapter 1: Introduction . . . 257
8.2.2 Chapter 2: Review of Network Music Performance . . . 258
8.2.3 Chapter 3: Review of Immersive Audio for Network Music Performance . . . 258
8.2.4 Chapter 4: Evaluation of an INMP System for Remote Music Education . . . 258
8.2.5 Chapter 5: Evaluation of an XR INMP system for remote virtual production . . . 258
8.2.6 Chapter 6: Evaluation of an INMP System for Remote Concert Performance . . . 259
8.2.7 Chapter 7: Evaluation of the effect of Reverb Time using an INMP system for remote singing practice . . . 259
8.3 Re–statement of Hypothesis . . . 260
8.4 Further Work . . . 261
8.4.1 Further INMP Evaluation . . . 261
8.4.2 Standardising INMP Evaluation . . . 263
8.4.3 Interactive Auralisation . . . 263
8.4.4 Further System Development . . . 264
8.4.5 Visual Display . . . 264
8.4.6 Industry Uptake . . . 265
8.5 Closing Remarks . . . 266

Appendices **267**

A Digital Appendices Overview **268**
A.1 Chapter 4 Supplementary Material . . . 268
A.2 Chapter 5 Supplementary Material . . . 268
A.3 Chapter 6 Supplementary Material . . . 268
A.4 Chapter 7 Supplementary Material . . . 269

References **295**

List of Figures

2.1 A flow chart illustrating the dependency of musical timing on audio delay. Where delay is negligible musical interactions are perceived and responded to with acceptable synchrony. Where delay exceeds recommended thresholds (20–30 ms) musical interactions are perceived as occurring late, and can result in recursive late responses unless interrupted. . 36
2.2 Illustration of Network Music Performance (NMP) signal flow from a performer at endpoint A, over a Wide Area Network (WAN) connection to a discrete performer at endpoint B. System components which contribute significantly to audio throughput delay between the two remote performers are highlighted. 37
2.3 Audio signal path for an NMP endpoint audio interface and computing device. Inclusion of audio processing and audio routing applications is considered. Internal buffering is highlighted, as are associated applications. 39
2.4 Simple network overview between NMP endpoints. 40
2.5 NMP jitter buffer illustration. Frames of audio data ($p[n]$, where n is the frame sequence index) are written to, and read from, a ring buffer. Variable distance between read and write pointers allows frames of audio data to be written at a variable rate, and read at a constant rate. 42
2.6 Illustration of Forward Error Correction (FEC) with a single frame redundancy. $p[n]$ describes a frame of audio data, where n is the frame sequence index. The third packet to be received is missing, however packet loss is avoided by retrieving the missing frame ($p[3]$) from redundancy transport in the second received packet. 43
2.7 Illustration of peer–to–peer network architecture. 45
2.8 Illustration of client–server network architecture. 46
2.9 Illustration of Inter–Onset–Interval (IOI) and Onset Time Difference (OTD) synchrony metrics. 51
2.10 Example illustration of immersive experience and presence modelling factors, from [125]. 56

3.1 Spherical coordinate notation for the point $p(r,\theta,\phi)$ 67
3.2 Simple depiction of the reflection of sound in a room, showing direct sound path, and different orders of reflection. 68
3.3 Simple depiction of Impulse Response (IR) representation of room acoustics, highlighting reverb components of direct sound, Early Reflection (ER)s, and diffuse reverb. 69

3.4 Illustration of non–uniform partitioning in convolution, reproduced with reference to [156]. $x(n)$, $h(n)$ and $y(n)$ are the input signal, IR and output respectively. N is the frame size of the first partition. 71
3.5 Basic comb filter. $x(n)$ is the input, $y(n)$ is the output, M is the delay, and g is the gain. 73
3.6 Basic all–pass filter. $x(n)$ is the input, $y(n)$ is the output, M is the delay, and g is the gain. 73
3.7 Simple illustration of the Image Source Method (ISM) for room acoustics. The source and listener position are shown with direct sound signal path. A boundary, first–order reflection and image–source are detailed. This figure was created with reference to illustration in [162]. 74
3.8 Illustration of x, y and z positional axis, and illustration of yaw, pitch and roll rotations, describing rotation around z, y and x axis respectively. . . . 75
3.9 Example head and torso simulator used in binaural recording, namely a Gras Acoustics Kemar model. 78
3.10 MH Acoustics EM32 Eigenmike, here illustrated in–use for Spatial Impulse Response (SIR) measurement. 81
3.11 Spherical harmonics up to 3rd order. m and i are the degree and index of each spherical harmonic, as detailed in subsection 3.5.1. 82
3.12 Illustration of a 2–Dimensional (2–D) Ambisonic loudspeaker array setup, namely the 8–channel 3rd Order Ambisonic (3OA) array from the SADIE II database. The sweet spot, and an example of off–centre listening position are noted. 85
3.13 Breakdown of factors of audio quality considered in relevant standards, where immersive audio rating items considered in Immersive Network Music Performance (INMP) evaluation are included, from [180]. 93

4.1 Named artist delivering a remote music education masterclass to groups of students at multiple remote locations. 101
4.2 Network view of the system deployment, detailing Local Area Network (LAN) connection between the three endpoints, analogue lines to the recording studio console, and noting the performer associated with each endpoint. 104
4.3 Illustration of hardware configuration at each endpoint, detailing relevant hardware and physical routing, audio and video capture and playback channels, and the LAN connection. 105
4.4 One–Way Trip (OWT) network latency using JackTrip as configured in the INMP system, estimated from loopback test signal sampled over 0.5 hours. 106
4.5 Binaural auralisation signal flow, illustrating auralisation processes, routing, and relevant audio applications. 108
4.6 Binaural Room Impulse Response (BRIR) measurement in Beehive Studios. 109
4.7 BRIR measurement configuration, illustrating KU100 receiver placement and distance, and left, centre and right loudspeaker placement and spacing. 110

4.8 BRIR describing Beehive Studios, detailing truncated BRIR used in auralisation, including diffuse reverb BRIR. 111
4.9 The drummer participant as situated using the INMP system and as captured by the video system. 113
4.10 Box plot illustrating rating of agreement for questionnaire item 1: *'We can play together as though in the same room'*. 115
4.11 Box plot illustrating rating of agreement for questionnaire item 2: *'I feel like I am in the same space as the other performers.'* 116
4.12 Box plot illustrating rating of agreement for questionnaire item 3: *'I have a sense of where other performers are located in the space'*. 116
4.13 Box plot illustrating rating of agreement for questionnaire item 4: *'The other performers sound as thought they are playing from where I see them on screen.'* 117
4.14 Box plot illustrating rating of agreement for questionnaire item 5: *'I am enjoying playing in this acoustic.'* 117
4.15 Box plot illustrating rating of agreement for questionnaire item 6: *'I am conscious of audio delay between performers.'* 118
4.16 Box plot illustrating rating of agreement for questionnaire item 7: *'The quality of audio is of a professional standard.'* 118
4.17 Box plot illustrating rating of agreement for questionnaire item 8: *'I engaged with video display during performance.'*. 119

5.1 Members of the professional rock band *Wolph* captured physically at each endpoint, and represented virtually as avatars using the INMP system. . . 129
5.2 Hardware configuration for reproduction of INMP in this study. Each endpoint is configured with the same basic hardware setup. Signal flow is detailed for both audio and Open Sound Control (OSC) tracking metadata. 133
5.3 Physical space of Maida Vale Studio 4 (MV4) (left) and virtual environment model of MV4 (right). 137
5.4 Floor plan of MV4, detailing performer positions PA03, PA11, PA15 and PA23. 137
5.5 Pass–through display in the Extended Reality (XR) application, here showing passthrough view of the drummer. 138
5.6 Immersive audio rendering signal flow for the INMP system, providing an overview of audio routing and processing. Relevant audio applications and audio plugins are detailed. 140
5.7 Illustration of direct sound removal and truncation of SIR to the onset of the first reflection. The original SIR omnidirectional channel from 3OA measurements (above) and processed SIR (below) are shown. The red line on the plot of the original SIR details the peak of direct sound, and the blue line details the calculated first reflection time. 142
5.8 Box plot for questionnaire item 1, rating agreement with the statement *'We can play together as though in the same room'*. 147
5.9 Box plot for questionnaire item 2, rating agreement with the statement *'I feel like I am in the same space as the other performers'*. 147

5.10 Box plot for questionnaire item 3, rating agreement with the statement *'I have a sense of where other musicians are located in the space'*. 148
5.11 Box plot for questionnaire item 4, rating agreement with the statement *'The other players sound as though they are playing where I see their avatar'*. 148
5.12 Box plot for questionnaire item 5, rating agreement with the statement *'I enjoyed the virtual acoustics during performance'*. 149
5.13 Box plot for questionnaire item 6, rating agreement with the statement *'The audio quality was of a professional standard'*. 149
5.14 Box plot for questionnaire item 7, rating agreement with the statement *'I am conscious of audio delay between performers'*. 150
5.15 Box plot for questionnaire item 8, rating agreement with the statement *'I engaged with the virtual environment and avatars during performance'*. . 150
5.16 Box plot for questionnaire item 9, rating agreement with the statement *'I engaged with pass–through display during performance'*. 151
5.17 Example of tempo analysis of recorded kick drum for the excerpt of the musical performance task. Onset detection is detailed (above, showing the audio recording with onset times indicated by red vertical markers) and associated tempo is detailed (below) including note of starting tempo (intercept) and tempo slope (gradient). 152
5.18 Box plot of measured tempo slope. 152

6.1 *Ex Corde Vocal Ensemble* providing networked concert performance using the INMP system. Here Phipp's Hall is shown. 164
6.2 Hardware setup at Phipp's Hall, University of Huddersfield, and the AudioLab studios, University of York. Audio signal flow and streaming channels are detailed. 168
6.3 Estimated OWT network latency between Phipp's Hall, University of Huddersfield, and the AudioLab studios, University of York. 170
6.4 Floor plan of Phipp's Hall detailing staging for the concert event. The four stage–right positions are the remote singers, each represented by a discrete *Genelec* loudspeaker, and the four stage–right positions are the physically present singers. 171
6.5 Immersive audio processing for the INMP concert system detailing relevant audio processing applications and audio routing. 172
6.6 Illustration of the truncation of the beginning of SIR to the onset of direct sound. Ambisonic Channel Number (ACN) 0 from the AmbiX format 3OA signal is shown, illustrating the original SIR (above) and SIR truncated to the onset of direct sound (below). The vertical line on the upper plot indicates the onset of the direct sound. 174

6.7 Illustration of the truncation of the beginning of SIR to the onset of diffuse reverberation, or *'mixing time'*. ACN 0 from the AmbiX format 3OA signal is shown, illustrating the SIR which has been truncated to the onset of direct sound (above) and SIR truncated to the onset of diffuse reverberation (below). The vertical line on the upper plot indicates the onset of diffuse reverberation. 174
6.8 Acoustic measurement positions for Maida Vale Studio 2 (MV2). The source position (*PA02*) and receiver position (*OA11*) for the SIR measurement used in this study are illustrated. 175
6.9 Experience rating results from the live audience in Phipp's Hall (N=17). . 181
6.10 Experience rating results from the online headphone audience in York (N=4). 182
6.11 Singer experience rating results. N = 7 for the *live rehearsal* and *network rehearsal* sessions, and N = 6 for the *concert* session. 183
6.12 Rating results from performance quality evaluation. Box plots are given for each of the four discrete groups, with each group relating to a discrete combination of performance material and method. Individual ratings are shown to the right of the box plot for each discrete group (N = 36) 193

7.1 Hardware setup deployed for testing across the LAN at the AudioLab studios. Audio analogue routing, head–tracking, JackTrip audio streams and test protocol control application User Datagram Protocol (UDP) signal flow are detailed. 202
7.2 Measured OWT latency from local audio capture to remote audio playback for analogue and JackTrip connections between endpoints. Here measurement is performed with immersive audio rendering bypassed. . . 203
7.3 Immersive audio processing for the INMP test protocol deployment, detailing relevant applications and signal flow. 205
7.4 The participant Graphic User Interface (GUI) for the MATLAB INMP evaluation framework. 207
7.5 Flowchart illustrating progression through the test protocol using the INMP evaluation framework. 208
7.6 SIR measurement positions in the GuildHall, York, detailing source position S4 and receiver position R1. 211
7.7 Illustration of the truncation of the SIR to the direct sound. ACN0 for the original SIR is shown with the vertical line indicating the determined direct sound peak (above), and ACN0 for the truncated SIR is shown (below).212
7.8 Illustration of SIR decay envelope adjustment from approximately 2.3 s to 1.5 s Reverb Time (RT). ACN0 is shown, and the vertical line indicates the first reflection time, onward from which decay envelope adjustment is applied so as not to attenuate direct sound. 212
7.9 Illustration of SIR direct sound removal. ACN0 is shown. The vertical lines indicates the peak of direct sound and the first reflection time to which the SIR is truncated. 213

7.10 RT (T30) for the original SIR (RT2.3), and SIR with decay envelope adjustment applied to approximate RT of 1.5 s (RT1.5) and 0.5 s (RT0.5).A, C and L refer to A–weighted, C–weighted and Linear frequency weighting respectively. 213
7.11 Histogram of participant experience level. 215
7.12 Histogram of participant age. 215
7.13 Histogram of participant gender. 216
7.14 Histogram of participant ethnicity. 216
7.15 Detected onsets for the raw output of the Music Information Retrieval (MIR) Toolbox MIRONSETS function (above, indicating 116 onsets including 84 false positives) and the output with the MM–BOP function implemented (below, indicating 32 onsets with no false positives or negatives). 221
7.16 Example of onset detection using the Marker Metadata–Based Onset Picking (MM–BOP) method in synchrony analysis. 222
7.17 Example of *Tempo Slope* calculation in synchrony analysis. 222
7.18 Example of Mean tempo for participant pairs by group, illustrating variation in participant choice of tempo. 225
7.19 Box plot across RT and latency factors for participant rating of questionnaire item 1, addressing perceived audio delay between performers. . . . 227
7.20 Box plot across the factor of latency for participant rating of questionnaire item 1, addressing perceived audio delay between performers. 229
7.21 Box plot across the factor of RT for participant rating of questionnaire item 1, addressing perceived audio delay between performers. 230
7.22 Box plot across RT and latency factors for participant rating of questionnaire item 2, addressing perceived synchrony between performers. 231
7.23 Box plot across the factor of latency for participant rating of questionnaire item 2, addressing perceived synchrony between performers. 233
7.24 Box plot across the factor of RT for participant rating of questionnaire item 2, addressing perceived synchrony between performers. 234
7.25 Box plot across RT and latency factors for participant rating of questionnaire item 3, addressing perceived clarity in monitoring of auralised performance. 235
7.26 Box plot across the factor of RT for participant rating of questionnaire item 3, addressing perceived clarity in monitoring of auralised performance.237
7.27 Box plot across RT and latency factors for participant rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance. 238
7.28 Box plot across the factor of RT for participant rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance. 240
7.29 Box plot across RT and latency factors for participant rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance. 241

7.30 Box plot across the factor of RT for participant rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance. 243
7.31 Box plot across RT and latency factors for Tempo Slope. 244
7.32 Box plot across the factor of latency for Tempo Slope. 246
7.33 Box plot across RT and latency factors for Precision. 247
7.34 Box plot across RT and latency factors for Asynchrony. 248
7.35 Box plot across the factor of RT for Asynchrony. 250
7.36 Box plot across RT and latency factors for Lag. 251
7.37 Clarity (C50) for the original SIR (RT2.3), and SIR with decay envelope adjustment applied to approximate RT of 1.5 s (RT1.5) and 0.5 s (RT0.5). A, C and L refer to A–weighted, C–weighted and Linear frequency weighting respectively. 254

List of Tables

2.1 A note of common frame sizes and associated sound card throughput delay at a sample rate of 48 kHz. Two frames are assumed in capture and playback buffering, as is typical of pro–audio sound cards and drivers. It is notable that 256–sample frames or smaller are required to avoid excessively high audio delay in audio interface throughput. 38
2.2 Rating factors relating to measurement of presence. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of Quality of Experience (QoE) items. 54
2.3 Rating factors relating to measurement of interactivity. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items. 54
2.4 Rating factors relating to measurement of immersion. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items. 55
2.5 Rating factors relating to measurement of plausibility. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items. 55

4.1 A–weighted acoustic parameters describing the Teaching Room, Listening Room, Studio Live Room and Beehive Studios. 111

5.1 Upload and Download Bandwidth (BW) requirements for each endpoint measured in Mbps. N Ch. Up and N Ch. Down denotes the total number of upload and download channels respectively for each endpoint, where 16–bit, 48 kHz uncompressed audio is considered (for a per–channel BW cost of 768 kbps). Up BW and Down BW denote upload and download total BW requirements for each discrete endpoint. The instrument associated with each endpoint is also noted. 135
5.2 Mean and standard deviation of A–weighted acoustic parameters from acoustic measurement of the physical space of MV4. 141

6.1 RT (T30) of MV2 and Phipp's Hall. 173
6.2 Headphone models used by participants in completion of the evaluation. . 191
6.3 Results of the Friedman test conducted across the four pieces of listening material for each of the questionnaire items. For each rating item the degrees of freedom (df), number of measurements (N), Chi–square statistic (Chi–sq), and p–value (p) are given. 194

7.1 Shapiro–Wilk normality test results for *Tempo Slope* by group. Significant results (i.e., *not normal* distribution) is highlighted in bold. 226
7.2 Shapiro–Wilk normality test results by group for Precision. Significant results (i.e., *not normal* distribution) is highlighted in bold. 226
7.3 Shapiro–Wilk normality test results by group for Aysnchrony. Significant results (i.e., *not normal* distribution) is highlighted in bold. 226
7.4 Shapiro–Wilk normality test results by group for Lag. Significant results (i.e., *not normal* distribution) is highlighted in bold. 226
7.5 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 1, addressing perception of delay. Significant differences are highlighted in bold. . . 228
7.6 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of latency for rating of questionnaire item 1, addressing perception of delay. Significant differences are highlighted in bold. . . 229
7.7 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 1, addressing perception of delay. 230
7.8 Wilcoxon rank sum test significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 1, addressing perception of delay. 230
7.9 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 2, addressing perception of synchrony. Significant differences are highlighted in bold. 232
7.10 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of latency for rating of questionnaire item 2, addressing perception of synchrony. Significant differences are highlighted in bold. 233
7.11 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 2, addressing perception of synchrony. 234
7.12 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 3, addressing perception of clarity in monitoring of auralised performance. Significant differences are highlighted in bold. 236
7.13 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 3, addressing perceived clarity in monitoring of auralised performance. 237
7.14 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance. Significant differences are highlighted in bold. 239
7.15 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance. . . 240

7.16 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance. Significant differences are highlighted in bold. 242
7.17 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance. . 243
7.18 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for Tempo Slope. 245
7.19 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of latency for Tempo Slope. Significant differences are highlighted in bold. 246
7.20 Wilcoxon rank sum significance values (p) from pairwise comparison across the factor of latency for Tempo Slope. Significant differences are highlighted in bold. 246
7.21 Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for Asynchrony. Significant differences are highlighted in bold. 249
7.22 Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for Asynchrony. 250
7.23 Wilcoxon rank sum significance values (p) from pairwise comparison across the factor of RT for Asynchrony. 250

List of Acronyms

2–D 2–Dimensional

2OA 2nd Order Ambisonic

3–D 3–Dimensional

3DoF 3 Degrees of Freedom

3OA 3rd Order Ambisonic

5G 5 GHz

6DoF 6 Degrees of Freedom

ACN Ambisonic Channel Number

ADAT Alesis Digital Audio Tape

ADC Analogue–to–Digital Conversion

AES Audio Engineering Society

AllRAD All Round Ambisonic Decoding

AmbiX Ambisonics Exchangable

ANOVA Analysis of Variance

API Application Programming Interface

ARQ Automatic Repeat Request

AV Audio–Visual

AVB Audio Video Bridging

BAQ Basic Audio Quality

BBC British Broadcasting Corporation

BEM Boundary Element Model

BPM Beats Per Minute

BRIR Binaural Room Impulse Response

BW Bandwidth

CCRMA Center for Computer Research in Music and Acoustics

CLI Command Line Interface

CT Centre Time

DAC Digital–to–Analogue Conversion

DAW Digital Audio Workstation

dB Decibel

DI Direct Injection

DRR Direct–to–Reverberant Ratio

DSP Digital Signal Processing

EDT Early Decay Time

EQ Equalisation

ER Early Reflection

FDN Feedback Delay Network

FEC Forward Error Correction

FFT Fast Fourier Transform

FOA First Order Ambisonic

FOH Front of House

GPS Global Positioning System

GUI Graphic User Interface

HISS Huddersfield Immersive Sound System

HMD Head Mounted Display

HOA Higher Order Ambisonic

HRTF Head Related Transfer Function

ILD Inter–aural Level Difference

IEM In Ear Monitor

INMP Immersive Network Music Performance

IOI Inter–Onset–Interval

IP Internet Protocol

IQR Inter–Quartile Range

IR Impulse Response

ISM Image Source Method

ISO International Standards Organisation

ISP Internet Service Provider

ITD Inter–aural Time Difference

JACK Jack Audio Connection Kit

JND Just–Noticable Difference

LAN Local Area Network

LCR Left–Centre–Right

LTI Linear Time–Invariant

LUFS Loudness Units Full–Scale

MADI Multichannel Audio Digital Interface

Mag–LS Magnitude–Least Squares

MIR Music Information Retrieval

MINERVA Musical Interactions in Networked Experiences using Real–time Virtual Audio

MM–BOP Marker Metadata–Based Onset Picking

MV2 Maida Vale Studio 2

MV4 Maida Vale Studio 4

NMP Network Music Performance

NTP Network Time Protocol

OS Operating System

OSC Open Sound Control

OTD Onset Time Difference

OWT One–Way Trip

PCM Pulse Code Modulation

PLC Packet Loss Concealment

PTP Precision Time Protocol

QoE Quality of Experience

QoL Quality of Life

QoS Quality of Service

RAN Radio Access Network

RT Reverb Time

RTP Real Time Protocol

RTT Round–Trip–Time

SAFFIRE Scalable Audio Frameworks For Immersive Real–time Experiences

SDK Software Development Kit

SIP Session Initiation Protocol

SIR Spatial Impulse Response

SN3D Schmitt Semi–Normalised 3–D

SPL Sound Pressure Level

TCP Transmission Control Protocol

UDP User Datagram Protocol

UK United Kingdom

USB Universal Serial Bus

VBAP Vector–Based Amplitude Panning

VIIVA Vocal Interactions in a Virtual Acoustic

VLE Virtual Learning Environment

VoIP Voice over Internet Protocol

VR Virtual Reality

VST Virtual Studio Technology

WAN Wide Area Network

WAP Wireless Access Point

WLAN Wireless Local Area Network

XR Extended Reality

Chapter 1

Introduction

Ted: *Now what?*
Bill: *I dunno. Philosophize with him!*

Theodore 'Ted' Logan and Bill S. Preston Esq., Bill and Ted's Excellent Adventure (1989)

Many projects generally require groups of individuals to cohabit and interact in a physical space in order to conduct work or engage in a process or experience. Where it is impossible or impractical to congregate in a physical space, online workflows provide a solution by which the project activity can take place remotely. Such remote workflows make project activity possible, or more convenient, by circumventing or alleviating logistical concerns (and expenses) of on–site workflows, such as space hire, travel and accommodation, scheduling and equipment transport.

The case of online systems and remote workflows for musical performance contexts is described as NMP [6]. NMP systems allow remotely–located musicians to connect over the Internet and participate in musical performance. This is achieved by sharing audio between WAN endpoints using low–latency streaming methods, such that performers can hear and interact with one another in real–time. Deployments over recent years have demonstrated that typical shared consumer and academic Internet networks are capable of supporting high quality NMP [7]–[9]. Here, telecommunication audio delay is low enough[1] that performers have negligible experience of lag and can play in time with comparable proficiency to live performance contexts. As such, NMP technology has become of practical interest to the music industry, providing the basis of remote performance workflows.

Modern research explores the extension of low–latency NMP streaming solutions with audio systems which provide networked performers a virtual acoustic simulation of playing together in a live performance space. In live performance contexts performers experience sound in 3–Dimensional (3–D) space, including directional properties of sound sources, and reverberant properties of acoustic spaces [10], [11].

[1] Empirical study estimates a 30 ms system throughput latency threshold as the point at which latency begins to impair musical performance [6].

Audio systems which simulate the experience of sound in 3–D space are described as *immersive audio* systems. In previous literature the term INMP is used to reference immersive audio NMP systems and is adopted in this thesis [7], [8], [12], [13].

Interest in INMP systems is motivated by the potential for use in remote workflows for modern music industry contexts which involve immersive technologies. Indeed immersive delivery contexts are becoming increasingly popular [14]. Practical examples where INMP solutions have been proposed include virtual performance experiences in using traditional video display [15], [16], game, Virtual Reality (VR) and XR environments [14], [17]–[20], immersive venues with surround–sound systems [20]–[23], and other immersive media formats such as 360°video and live–streaming [24]. In these applications immersive audio is required to allow aural navigation of a 3–D environment, often providing audio–visual coherence with navigable 3–D visual display and interface, and improving sensation of *presence* in the immersive experience [25].

Despite numerous proposals for the practical application of INMP, prior work in the field is limited. Previous practical INMP deployments in literature only address high–latency contexts, where it is acknowledged that excessive lag makes achieving high quality performance difficult for popular rhythmic music styles[2] [18], [21]–[23]. These implementations present a significantly different context from INMP solutions proposed for practical music industry use–cases, where low–latency operation is required [14]. In this regard the validity of proposals for low–latency INMP solutions in practical use–cases remains untested. Study of musical performance quality and perceptual experience for INMP is also limited, and does not address proposed practical INMP use–cases. As such a limited understanding of INMP in–practice can be derived from prior work.

The research detailed in this thesis addresses the proposed application of INMP solutions in practical use–case contexts, and works towards an understanding of INMP which can be leveraged to optimise INMP experience and performance quality and inform best–practice. Immersive audio systems, namely *Ambisonic* [26], [27] and *binaural stereo* [10] methods, were adapted to the context of INMP and implemented as extensions to low–latency NMP streaming solutions. Evaluation of musical performance quality and perceptual experience using implemented systems tests the viability of INMP systems in practical use–case contexts and provides practice–oriented study of networked musical interactions using immersive audio systems.

[2]In these cases compositions which do not require strict musical timing are performed, for example ambient electronic music.

1.1 Statement of Hypothesis

Underpinning the work presented in this thesis is the premise that INMP systems can be effective and practical remote workflow solutions. The core hypothesis of the research detailed in this thesis is therefore given as:

- *High quality musical performance can be achieved using immersive audio Network Music Performance systems designed for practical applications.*

The key terminology here is defined as follows:

- *'High quality musical performance'* is defined as identifiable through physical performance metrics which do not deviate significantly from expectations of reference values from live performance workflows, and through perceptual rating metrics which describe a positive performance experience.
- *'Practical applications'* here identifies that relevant INMP systems have been deployed in the field for the relevant musical performance application[3], and are capable of supporting synchronous performance of popular rhythmic musical styles[4] by performers with no prior experience of networked performance[5].

1.2 Objectives

In order to test the core hypothesis of this thesis, the following objectives were devised:

- To implement novel proof–of–concept INMP system designs for practical use–case contexts.
- To determine through evaluation whether or not high quality musical performance and perceptual experience can be achieved using these systems.
- To contribute to a practice–oriented understanding of INMP and variation in INMP quality.
- To identify parameters of immersive audio rendering which can be controlled to optimise musical performance and perceptual experience for INMP in practice.

[3] I.e, the INMP system works in real world deployment.

[4] I.e., performers can play pop music as is industry–relevant

[5] I.e., significant training periods to develop networked musical interaction strategies should not be required.

1.3 Novel Contributions

In achieving the presented objectives, and addressing the core hypothesis of this thesis, several novel contributions are provided:

- Practical implementation and evaluation of INMP for a music tuition use–case.
- Practical implementation and evaluation of INMP for an XR virtual production use–case.
- Practical implementation and evaluation of INMP for a virtual concert use–case.
- Evaluation of virtual acoustic control for INMP vocal performance.
- Other peripheral contributions: Technical detail on INMP system design. A novel INMP evaluation system. MATLAB functions for synchrony analysis and IR processing.

1.4 Thesis Outline

Presentation of research and findings in this thesis is structured as follows. In Chapter 2 a review of NMP is presented. Detail is provided on low–latency NMP streaming methods, state–of–the–art NMP solutions, evaluation approaches, and findings of empirical research[6]. Chapter 3 presents a review of immersive audio for INMP. The fundamentals of spatial sound are surmised. Detail is provided on Ambisonic and binaural immersive audio methods, as well as virtual acoustic methods, focusing here on convolution–based approaches. Implementation of INMP systems in prior work is detailed, including a review of related evaluation approaches and finding. These review chapters provide the reader with background information relevant to the work in this thesis.

Chapters 4, 5 and 6 detail a series of evaluations designed to test the viability of proposed INMP use–cases and explore potential variation in the quality of INMP relating to immersive audio. In Chapter 4 a music tuition use–case is addressed. A virtual stereo–based binaural system was implemented with video and screen display. The system was deployed in collaboration with Bonza Music[7] for a music lesson between the artist *Jamie Cullum*[8] and students at two schools. Following this three music tutors participated in evaluation comparing auralisation of different virtual acoustic spaces. Results demonstrate the viability of this INMP use–case and suggest a preference for environmentally coherent virtual acoustics. Performer engagement with video systems in INMP is novelly demonstrated, however, limitations of video systems are also highlighted with respect to Audio–Visual (AV) asynchrony. The potential for virtual acoustic parameters to influence the experience of INMP is suggested.

[6] Empirical research in NMP evaluation focuses on the effect of latency on performance quality and perceptual experience [6].

[7] `https://bonzamusic.com/`

[8] `www.jamiecullum.com`

A virtual production use–case is presented in Chapter 5. A virtual Ambisonic–based binaural system was implemented with XR display. The system was deployed in collaboration with the British Broadcasting Corporation (BBC) as a digital twin of MV4 and used by the rock back *Wolph*[9] to record a single. The band participated in a following study in which immersive audio rendering is evaluated relative to a mono baseline at different latency levels. Results validate the INMP use–case and suggest perceptual benefits to use of immersive audio. Limitations of XR display and interface are highlighted, and the potential for experience of INMP to vary with monitoring context is suggested.

In Chapter 6 a virtual concert use–case is presented in collaboration with the BBC Singers. An INMP system was implemented using Ambisonic methods for loudspeaker playback and a stereo–enhanced virtual Ambisonic–based binaural approach for headphone playback. The system was deployed for virtual concert performance by *Ex Corde Vocal Ensemble*[10] to an audience in an immersive live venue, and a remote headphone audience with video display. Singers participated in comparative study of live and INMP performance, providing evaluation of performance quality in live conditions and for the concert performance. Audience members for the concert participated in evaluation of the quality of the concert experience. Following this a listening test was conducted, comparing live and INMP performance quality using audio recordings captured throughout the concert production.

Results show that a high quality concert experience was delivered, and suggest that comparable quality performance to live conditions can be achieved with INMP systems in this use–case. Challenges relating to lack of visual contact between singers are highlighted, as are challenges in loudspeaker rendering of immersive audio in concert venues. Practical benefits to immersive audio monitoring are also highlighted. The potential for virtual acoustic parameters to influence the INMP experience is suggested.

Following indication of virtual acoustic preferences in use–case evaluations Chapter 7 presents a controlled study, evaluating INMP under different virtual acoustic and latency conditions in a vocal performance context. A decay envelope adjustment method is applied to Ambisonic IR to control virtual acoustics rendered by a virtual Ambisonic–based binaural system. Results demonstrate that the perceived quality of performance conditions varies with virtual acoustic parameters, including perception of ability to play in time and perception of delay. These results demonstrate that virtual acoustics may be controlled to optimise the quality of INMP.

In Chapter 8 the conclusion to this thesis is given. The core hypothesis is revisited and discussed relative to key findings and the achievement of research objectives. The wider implications of the practical validation of INMP solutions presented in this thesis are considered, and discussion of areas of interest for future work is provided. The thesis is then concluded with closing remarks.

[9] `www.wolphband.com`

[10] `www.excordevocal.com`

1.5 Summary

The topic of INMP has been introduced, as have the motivations for research in this field. The central hypothesis, research objectives and novel contributions associated with the work presented in this thesis have been surmised. This thesis now presents this research as summarised in the thesis outline.

Chapter 2

Review of Network Music Performance

> We are stuck with technology when all we really want is just stuff that works. How do you recognize something that is still technology? A good clue is if it comes with a manual.
>
> *Douglas Adams, The Salmon of Doubt: Hitchhiking the Galaxy One Last Time (2002)*

2.1 Preface

INMP systems are implemented as extensions to existing NMP solutions, and practice builds upon prior work in the field. This chapter provides a review of relevant NMP technology, research and practice. NMP is first introduced and defined as a unique telecommunications technology. The challenge of latency in NMP is summarised and discussed with relevance to NMP practice and system design. A technical description is then provided of practical NMP methods and relevant networking concepts, and an overview is given for popular modern NMP solutions which are relevant to this thesis. Evaluation methods for NMP are then reviewed, providing detail on findings of prior evaluation with relevance to NMP practice.

2.2 A Brief History of Network Music Performance

The case of musical performance between endpoints on digital computer networks can be traced back at least as far as the 1970's, where artists began to experiment with electro–acoustic music performance on LAN [28]. NMP [6], however, more specifically refers to a set of low–latency audio streaming technologies designed to facilitate musical performance between WAN endpoints.

Whereas previous systems for musical performance over WAN consider experimental forms of musical interaction (for example various forms of sonic control and high latency streaming systems [28]), NMP systems are capable of allowing remote performers to achieve synchronous performance through natural musical interactions. Here each performer in an NMP session is able to hear and to interact musically with the networked ensemble, with negligible experience of delay, and perceiving musical timing as if co–located. In this regard NMP technology provides a practical remote alternative to typical applications of musical performance which traditionally require performers to cohabit and interact synchronously in a physical space.

NMP technology and practice emerged in the early 2000s, building on developments such as the deployment of high–speed Internet infrastructure and the widespread adoption of Voice over Internet Protocol (VoIP) telecommunications solutions. Early examples of NMP involved remote musical performance and virtual concert distribution, predominantly using high–speed managed WAN connections [29]–[34] as early public Internet was considered incapable of supporting low–latency NMP [30].

At the time of this thesis NMP technology and practice has developed significantly. The speed, bandwidth and infrastructure available via public and other shared Internet connections has improved significantly, and been demonstrated as capable of supporting low–latency NMP over national distances [7]–[9]. Standard technical NMP designs have emerged from early work and a range of software [6], [35] and hardware solutions are available. Hardware solutions include Linux–based solutions (for example the SoundJack Fast Music Box [36], or Elk Live system [37]) and pro–AV hardware (for example MVTP systems [38]. Experimental pro–LAN solutions for NMP (namely Dante) have also successfully been deployed over managed WAN connections [39], [40].

These modern NMP solutions are used in a range of virtual concert productions. In addition to a wide range of high–latency examples (such as [35]), prolific low–latency examples are also documented in literature (such as [15]). Remote recording and production [41] is also an application of NMP (notably allowing real–time musical interaction between endpoints, thereby differentiating from high–latency pro–WAN solutions for remote mixing and recording such as AudioMovers[1] or Source Elements[2]). Further applications of NMP include use as an access–enabling technology for remote music education [42], [43], and musical performance for social [44] and well–being [45] contexts. Amateur NMP is widespread, and is supported by active user communities, for example through online forums and user groups, and through online performance communities and ensembles.

2.3 Latency Challenges in Network Music Performance

NMP systems may be differentiated from other telecommunications technology primarily with respect to handling of the challenge of audio latency. This challenge is well acknowledged in VoIP systems [46]. Here delay between endpoints offsets the time at which speech interactions are heard.

[1] `https://audiomovers.com/`

[2] `https://www.source-elements.com/`

This can result in speech response times which exceed those typical of natural conversation. This impairs the experience of speech interactions by disrupting the natural flow of conversation [47], [48]. It is typically considered that audio delay below 100 ms is required to achieve acceptable flow in conversation for teleconferencing [49] (though more recent study would suggest teleconferencing audio latency in the range of 30–70 ms may still result in impaired conversation response times [50]). Teleconferencing solutions (designed with consideration of speech audio latency constraints) typically do not meet the audio latency requirements of NMP.

Recent literature describing use of teleconferencing solutions in remote musical performance workflows highlight latency as a major issue [51], [52]. Indeed extensive study of latency constraints in NMP has understood that a system throughput delay below 20–30 ms between performers is required to avoid disruption of musical interactions [6], far smaller than the 100 ms threshold considered in VoIP systems.

The reason for stringent audio latency constraints in NMP is with regard to perception of musical timing in performance. For groups of musicians to perform synchronously (i.e.'in time' with one another) it is required that a common understanding of musical timing is established. In this case each musician hears the musical actions of others. In hearing sequences of musical actions, each performer derives an expectation of periodicity in performance (i.e. musical timing), thereby allowing synchronous musical interaction [53], [54]. Musicians are capable of synchronising timing with a high degree of precision (typically in the order of tens of milliseconds) [53]. Professional musicians are capable of achieving synchrony with an accuracy of up to 0.005–0.02 times the period of one beat [53]. For reference this is 2.5–10 ms standard deviation of precision at 120 BPM. This is, however, dependent on performance context. For example asynchronies of several tens of milliseconds are typical of classical and vocal performance [55], [56].

In the case of networked musical interactions, audio delay between performers offsets the time at which musical actions are heard at remote endpoints. This offset may cause musical actions to be perceived as occurring outwith acceptable synchronisation tolerances (i.e. performed 'late'). Here the perceived timing of musical interactions becomes ambiguous, and synchronous performance becomes disrupted. Empirical study of NMP demonstrates that for audio delay between performers in excess of 30 ms the recursive perception of musical actions as being performed late can lead to recursive adaptation of musical timing (Figure 2.1). This typically leads to tempo deceleration, and potential collapse of performance synchrony [57], [58]. NMP system design therefore focuses on minimisation of audio latency between performers in order to negate or minimise associated disruption of musical synchrony, and provide acceptable performance conditions for professional musicians [6].

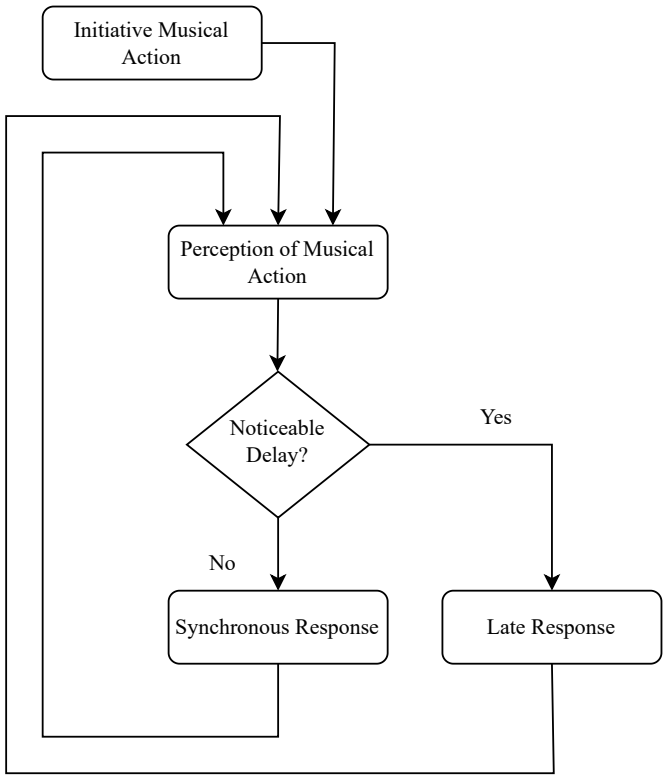


Figure 2.1: A flow chart illustrating the dependency of musical timing on audio delay. Where delay is negligible musical interactions are perceived and responded to with acceptable synchrony. Where delay exceeds recommended thresholds (20–30 ms) musical interactions are perceived as occurring late, and can result in recursive late responses unless interrupted.

2.4 NMP System and Latency Overview

A basic NMP system endpoint (Figure 2.2) consists of audio capture and playback devices (used in sound acquisition and monitoring), a computing device (used to host audio processing and low–latency audio streaming software) and a network connection with Internet access (used to access the Internet backbone over which audio data is transport between NMP endpoints). Management of audio delay between NMP musicians requires consideration of latency contributions across the full NMP system and signal path [6], [12]. This includes air propagation in sound acquisition and monitoring, audio interface capture and playback processes, application processes on computing devices, and network transport between NMP endpoints. The latency between NMP musicians is typically described in terms of One–Way Trip (OWT) from a sound source (i.e. musical instrument) at one endpoint to a receiver (i.e. listener's ears) at another endpoint [6]. The following subsections provide detail on processes which contribute latency to the OWT latency of an NMP system.

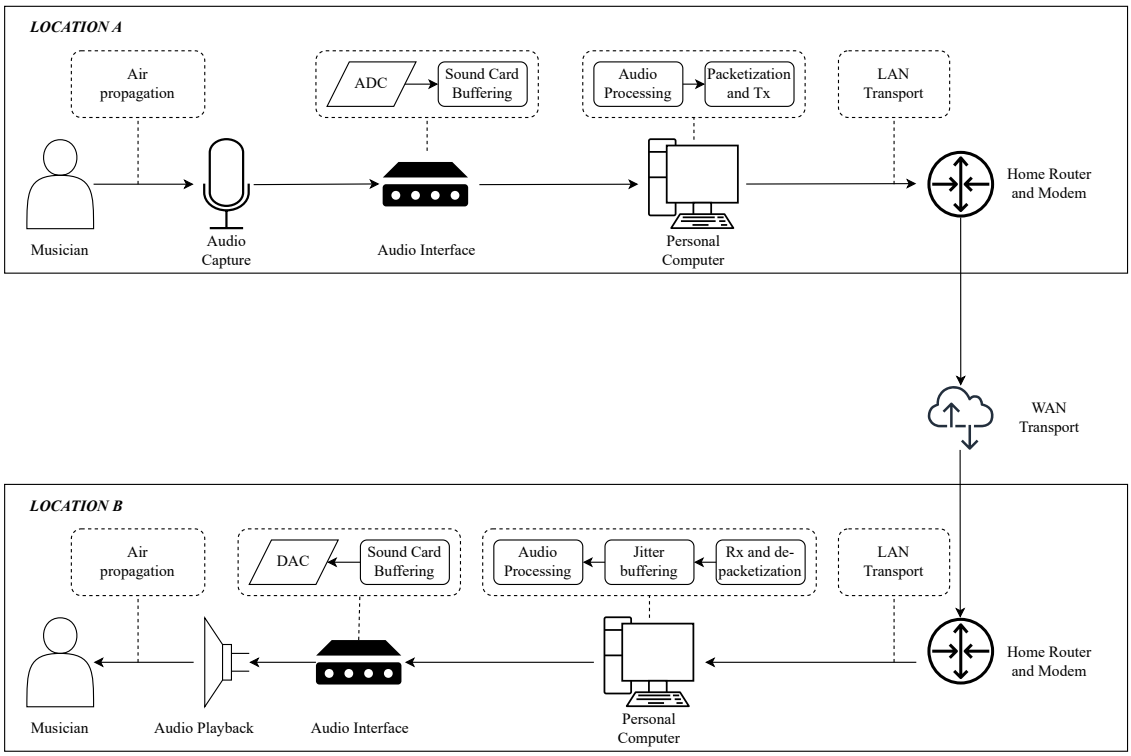


Figure 2.2: Illustration of NMP signal flow from a performer at endpoint A, over a WAN connection to a discrete performer at endpoint B. System components which contribute significantly to audio throughput delay between the two remote performers are highlighted.

2.4.1 Air Propagation, Sound Acquisition and Monitoring

Sound propagates from sources as a wave and traverses free air as the compression and decompression of air particles (i.e. sound pressure). The speed of sound in air at room temperature is given as approximately $c = 344$ m/s at room temperature [59]. 1 m of distance therefore takes approximately 3 ms for sound to traverse. In order to minimise air propagation delay in NMP systems excessive distance can be avoided between sound sources and microphones used in audio capture, or between listeners and loudspeakers used in audio playback.

Close microphone placement, or Direct Injection (DI) for electronic instruments, is most practical here with respect to minimisation of air propagation delay in audio capture. Headphone playback, with loudspeakers mounted directly by the ear, is most practical with respect to minimisation of air propagation delay in monitoring [6]. Use of headphones, particularly In Ear Monitor (IEM), also usefully avoids signal spill from monitoring into audio capture, and associated potential for feedback in monitoring between endpoints [9].

Frame Size (samples)	Frame Size (ms)	Throughput Delay (ms)
32	0.667	1.333
48	1	2
64	1.333	2.667
96	2	4
120	2.5	5
128	2.667	5.333
240	5	10
256	5.333	10.667
480	10	20
512	10.667	21.333

Table 2.1: A note of common frame sizes and associated sound card throughput delay at a sample rate of 48 kHz. Two frames are assumed in capture and playback buffering, as is typical of pro–audio sound cards and drivers. It is notable that 256–sample frames or smaller are required to avoid excessively high audio delay in audio interface throughput.

2.4.2 Capture and Playback Processes

The digitisation of sound to audio Analogue–to–Digital Conversion (ADC), and vice–versa (Digital–to–Analogue Conversion (DAC)), and handling of audio, is performed by audio interface hardware and driver software [60]. Though ADC and DAC incur negligible levels of delay, delay is incurred in capture and playback buffering [6], [61]. Low–latency throughput here typically requires a dedicated audio interface, as common Operating System (OS) default audio drivers and ADC/DAC hardware on common computing devices typically require large frames and internal buffer queuing to operate without audio dropouts. Audio interface capture and playback buffering delay (Table 2.1) is configurable via sample rate and frame size used. Typically pro–audio sound cards and drivers consider a single frame of audio for capture buffering, and another single frame for playback buffering [62]. Performance can, however, vary across equipment [63].

2.4.3 Audio Processing and Application Buffering

Digital Signal Processing (DSP) [64] may be applied to audio data either before or after network transport, typically in the form of audio effects for NMP contexts. Though some NMP software includes limited audio effects (for example JackTrip[3] or SonoBus[4]), generalised use of audio plugins practically requires that a routing application is used to route audio between hardware capture and playback buffers, an audio processing application (i.e. plugin host), and an NMP application (Figure 2.3).

[3] `https://jacktrip.github.io/jacktrip/`
[4] `https://sonobus.net/`

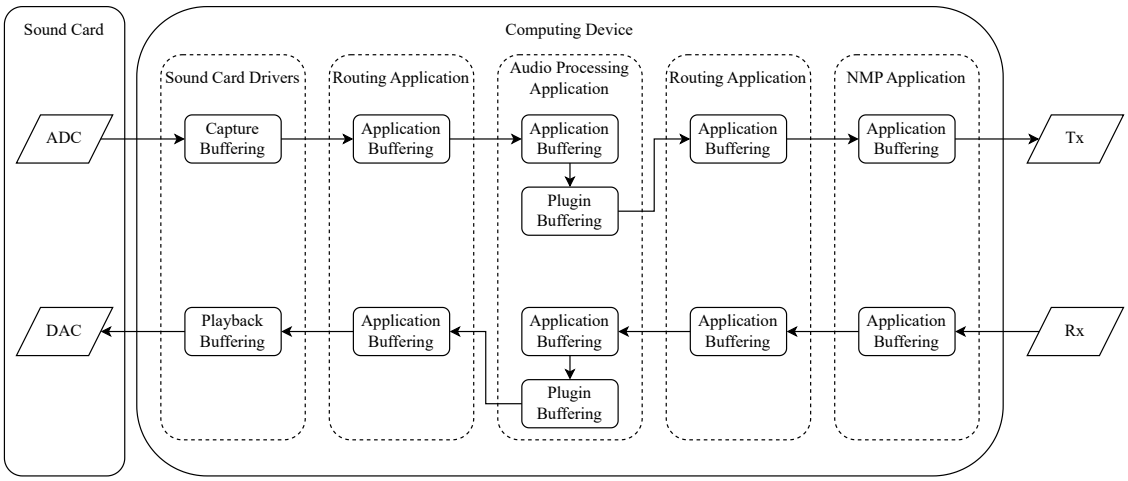


Figure 2.3: Audio signal path for an NMP endpoint audio interface and computing device. Inclusion of audio processing and audio routing applications is considered. Internal buffering is highlighted, as are associated applications.

In audio routing applications audio delay may be incurred in application internal buffering [6]. Minimisation of delay here can be again acknowledged as dependent on specified sample rate and frame size (Table 2.1). Audio routing application buffer queue size should also be minimised where configurable, and routing applications which require significant internal buffer queuing should be avoided.

In audio processing applications audio delay may be incurred both in internal application buffering and by internal buffering in audio plugins themselves [12]. The audio plugin host application buffering delay is once again dependent on configured frame size and sample rate, and any configurable internal buffer queuing. Audio plugin delay is dependent on internal buffer sizes and plugin functionality, which varies on a case–by–case basis (being dependent on the specific plugin implementation) [12], [65]. Minimisation of audio delay here is generally dependent on audio plugin choice and configuration.

2.4.4 Network Transport

The sharing of audio data between NMP endpoints is performed using low–latency audio streaming software hosted on a computing device at each endpoint. This software is responsible for processes associated with sending and receiving audio data. The audio data itself is transported between endpoints via the connecting network. This includes the LAN between the computing device and Internet gateway at each endpoint, and the WAN connection between Internet gateways (Figure 2.4). Audio data transport over a LAN [66] will typically incur negligible levels of latency provided a wired network connection is used (though Quality of Service (QoS) policy [66], may prioritise other network traffic, for example, traffic associated with pro–LAN systems such as Dante, raising NMP LAN transport latency).

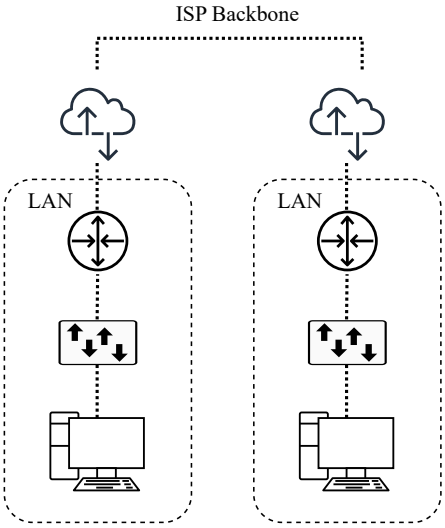


Figure 2.4: Simple network overview between NMP endpoints.

Though wireless NMP networks have been demonstrated as viable using modern wireless technologies, for example Radio Access Network (RAN) and 5 GHz (5G) wireless systems [67], [68], it can be acknowledged that typical wireless connections encountered in practice cannot be expected to provide this performance [69], and can rather be expected to under–perform wired LAN connections [70], providing higher latency and less reliable data transport.

Audio data transport over WAN will typically incur a significant level of delay as audio data transits the Internet Service Provider (ISP) backbone between endpoints [6]. Notably WAN transport delay will also vary over time (i.e. jitter [71]). WAN transport delay and jitter can be recognised as dependent upon distance between NMP endpoints (though geographic distance does not necessarily equate to network distance) [7], [62]. Though previous studies have demonstrated that sub–30 ms audio throughput can be achieved for NMP across national distances [8], [9], international [8] and trans–Atlantic [35] deployments using similar systems result in high audio delay between NMP endpoints due to excessive WAN transport latency.

Considering use of typical shared networks in practical NMP, WAN connection between endpoints is an unmanaged network component and WAN latency is therefore not typically possible to control via network QoS policy. It is possible to configure WAN connections with managed QoS, however, the expense of this is typically beyond the scope of most practical musical performance productions where NMP offers a remote workflow solution. As such, minimisation of latency in WAN transport is dependent on the implementation of low–latency audio streaming methods in NMP software. The internal processes associated with NMP software will incur a significant amount of throughput latency themselves, which can be considered as an application buffer (Figure 2.3).

2.5 Low–Latency Audio Streaming Methods for NMP

Data traverses digital networks following Internet Protocol (IP) [72] networking methods, where frames of data may be transferred from one device to another using a specified structure (i.e. IP packet), and where a specified addressing logic (i.e., IP address[5]) is used to identify discrete devices in data transfer. At the core of low–latency NMP streaming is a pseudo–standard approach to low–latency and reliable audio data transport over IP networks in UDP [74] payloads [6]. This core functionality is supported by session management protocols and clock synchronisation methods, which may vary across implementation. Additional features may also be included, such as optional audio compression and packet loss concealment. The following subsections provide technical detail on the low–latency NMP audio streaming systems.

2.5.1 UDP Audio Streaming

UDP [74] describes transport protocol used in the transport of data across IP networks, implemented on top of IP packets. UDP is defined as a connectionless protocol, indicating that there is no bi–directional communication between send and receive sockets [66]. UDP transport therefore minimises network delay, as data only transits the network path between endpoints once. As a 'best effort' protocol (i.e., there is no acknowledgement of packet receipt or re–transmission function between send and receive sockets native to UDP) UDP provides unreliable data transport [62]. UDP frames may arrive at receive sockets with varying latency as a result of network jitter (out–of–order packets) or may not arrive at all (i.e. packet loss). The core functionality of low–latency NMP software is the streaming of audio data in UDP payloads (in order to provide low latency WAN transport), where methods of improving stream reliability for real–time media are implemented on top of UDP by the NMP application [6].

2.5.2 Jitter Buffering and Packet Sequencing

NMP systems require that audio data is both written and read at constant rates at each endpoint in order to maintain continuous audio streams between endpoints. Network jitter in WAN transport of UDP packets results in packets arriving at a variable rate at each endpoint. In order to allow for audio data received at a variable rate to be read at a constant rate, an intermediate ring buffer is used (i.e., a jitter buffer [71]). Here data is read at a constant rate from a static read pointer, while data is input to the jitter buffer using a variable write pointer [6], [75] (Figure 2.5). In NMP jitter buffers must be sized larger than WAN jitter in order to ensure continuous audio throughput [6]. The process of jitter buffering incurs an application latency with the NMP software, equal to the offset between read and write pointers [75].

[5]Though IPv6 [73] is the current IP standard, IPv4 [72] is still in common usage.

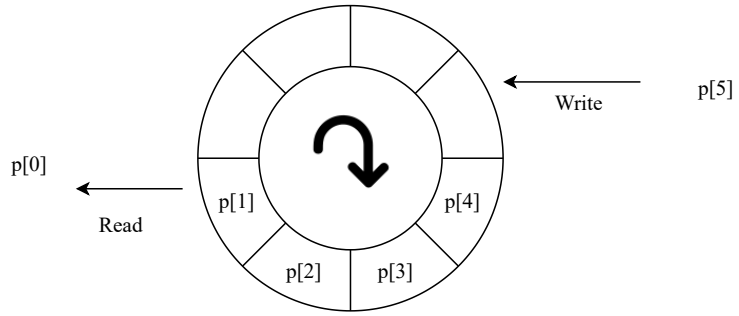


Figure 2.5: NMP jitter buffer illustration. Frames of audio data ($p[n]$, where n is the frame sequence index) are written to, and read from, a ring buffer. Variable distance between read and write pointers allows frames of audio data to be written at a variable rate, and read at a constant rate.

Handling of over–run (i.e. buffer overflow, where frames to be written do not have available write space) and under–run conditions (i.e. buffer starve, where no frames are available to read) may also be implemented [6], [15], [75]. Under–run conditions may be handled by simply introducing a wait state until frames are available to read, or alternatively synthesised audio may be used to fill the gap. For instance , JackTrip [75] allows options to loop through previous frames while in an under–run wait state. Over–run conditions are handled by resetting the variable write pointer position (i.e. discarding oldest frames in the jitter buffer) such that frames in the jitter buffer are available to write to again [75].

Out–of–order packet–sorting is typically implemented in jitter–buffer writing processes in NMP applications. Sequential frames of audio data are made identifiable through the use of sequence indices, used to sort packets by informing write positions for received frames. These may be included in custom headers implemented inside UDP payloads [75]. Some NMP systems implement Real Time Protocol (RTP) [76] inside of UDP payloads, where sequence index headers are included [6], [32].

2.5.3 Packet Loss Concealment

In order to handle the case of packet loss, NMP applications may implement Packet Loss Concealment (PLC) methods [6]. In NMP WAN transport delay does not allow for low–latency re–transmission of missing packets [30]. This form of Automatic Repeat Request (ARQ) requires data to transport the network path between endpoints 3 times for a single re–transmission (original packet sending, packet receipt acknowledgement update messages from the receiver to sender, re–transmission), typically incurring significant latency in WAN transit. PLC must therefore be implemented in a connection–less manner (i.e., implemented and resourced at receive sockets). NMP achieves this through provision of redundant audio–streams, such that where one instance of a sent frame of audio does not arrive at a receive socket, the relevant frame may be retrieved from the redundant stream.

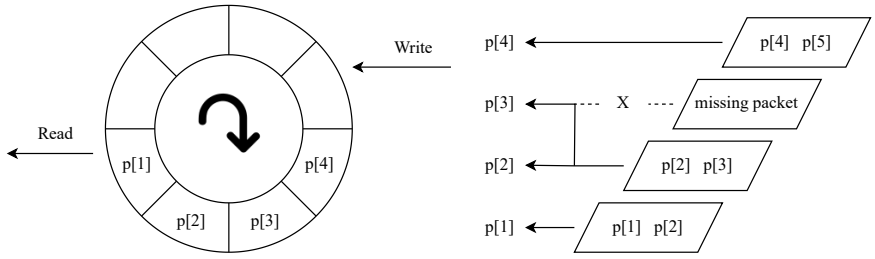


Figure 2.6: Illustration of FEC with a single frame redundancy. $p[n]$ describes a frame of audio data, where n is the frame sequence index. The third packet to be received is missing, however packet loss is avoided by retrieving the missing frame ($p[3]$) from redundancy transport in the second received packet.

FEC methods [77] (Figure 2.6) are popular in NMP systems [6], where redundant frames of audio are transmitted with each frame in a stream such that for a redundant factor of K frames, frames $N-1$ to $N-K$ are sent as redundancy in the same packet as frame N. Implementations of FEC in NMP software include bespoke implementation [75] as well as the Recovery Journal approach associated with RTP [32]. FEC PLC methods, sending additional redundant data, increase the bandwidth cost of NMP streams. Considering a typical NMP stream, a single channel of uncompressed Pulse Code Modulation (PCM) audio data of 16–bit resolution and sample rate of 48 kHz (i.e., pro–audio quality) is associated with a data rate of 768 kbps [62]. IP, UDP and any custom NMP application packet headers and associated metadata will incur additional overhead. The 'rule of thumb' 1 Mbps per channel of audio may prove more useful in practice. For each redundant frame of audio sent per packet, bandwidth requirement will be increased accordingly.

2.5.4 Clock Synchronisation

Due to inconsistencies in precise local clock frequency of devices across endpoints clock drift can occur, where discrete endpoints do not agree on the time elapsed (and associated number frames of audio) over a period of absolute time. With respect to NMP this leads to periodic under–run and over–run conditions in jitter buffers[6] as clock drift causes audio streams to skew [15], [75]. As NMP systems are deployed over a WAN rather than LAN, pro–LAN approaches to clock distribution are not possible. Here LAN–based transport solutions such as Audio Video Bridging (AVB) [78], AES 67 [79] or Dante use layer 3 clock distribution protocols, namely various versions of Precision Time Protocol (PTP) [80] which requires high–speed, low–latency transport between endpoints (i.e., LAN performance).

[6]Where one clock is faster than another, the case will periodically occur where $N+1$ rather than an expected N frames of audio are received. Where this process is recurrent over–run conditions will be periodically incurred. Conversely the endpoint with the faster clock, when receiving data from the endpoint with the slower clock will eventually achieve the case of $N-1$ frames of audio data where N are expected, where recurrence results in periodic under–run conditions being incurred.

Network Time Protocol (NTP) [81] is used in some cases to synchronise NMP endpoint clocks [6], [82]. Here each endpoint periodically receives an update over the WAN from a reference clock, allowing clock drift calculation, and relevant clock rate correction. A reference clock may be similarly derived from RTP [76] timestamp headers [6], [32]. NTP–based synchronisation over WAN, however, can tolerate several milliseconds of drift with performance being dependent on network quality [6].

Global Positioning System (GPS) clock distribution has also been used in several cases to provide a common reference clock across NMP endpoints [6], [9], [33], [39], which is capable of delivering synchronisation within a few microseconds. In practice, however, GPS performance may vary, particularly using makeshift devices in urban environments (where signal interference is possible) [63]. This approach notably requires specialist equipment (GPS modules) which may not be readily available in practical musical performance contexts.

In NMP practice, synchronisation between remote endpoints may be achieved simply through under–run and over–run handling [15], [75]. Where over–run conditions occur, it can be concluded that the local clock is running slow relative to the relevant remote clock. Where over–run condition handling re–sets read and write pointer positions in jitter buffers, the local system is effectively 'reading ahead', counteracting local clock lag relative to a remote endpoint. Where under run–conditions occur, it can be concluded that the local clock is running fast relative to the relevant remote clock. Where under–run condition handling resets read and write pointer positions (zero–filling empty buffers), the system is effectively 'waiting N frames', counteracting local clock lead relative to a remote endpoint [15]. This approach effectively operates as a block–resampling scheme. Notably this approach to real–time synchronisation between endpoints will cause periodic audio streaming artifacts as over/under–run conditions are periodically encountered [15], [39]. Despite this, the method is employed in typical modern NMP solutions (for example JackTrip [75] or SoundJack [15]), requiring no specialist technology, and where performance does not depend on network quality. Prior research has demonstrated that software which utilises such over/under–run handling is not likely to incur deterioration of audio quality relative to reference audio [83], and such software is used with success in professional music performance contexts [15].

An alternative approach to clock synchronisation across endpoints is the use of high–precision reference clocks at each endpoint. Such hardware can provide clocking specified with < 0.01 PPM drift[7] negating the concern of clock drift in practice for any sensible performance duration. Such equipment is specialist and expensive, however typical in AV contexts.

2.5.5 Session Management

As NMP software streams audio data via UDP it is necessary to provide endpoints with the metadata required to de–packetise and process received packets of audio data (i.e., frame size, sample rate, resolution, N channels, sequence number, time stamp).

[7] I.e., below 1 ms drift per day.

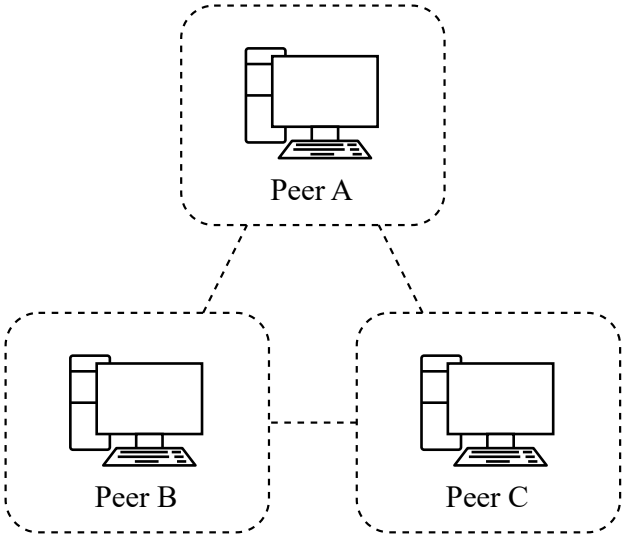


Figure 2.7: Illustration of peer–to–peer network architecture.

This session management metadata may be implemented as headers in custom NMP packet structures [75] or through implementation of protocols containing required headers such as RTP [6], [32], [34], [62], [84]. Alternatively session management may be handled through reliable network protocols such as Transmission Control Protocol (TCP) [85] and Session Initiation Protocol (SIP) [86] where transport of control data is ensured through packet receipt acknowledgement and re–transmission processes between endpoints [6].

2.5.6 Network Architecture

WAN transport delay is dependent on the distance of the network path between endpoints. This network path is defined with respect to the network architecture. peer–to–peer architectures (Figure 2.7) define a shortest–distance direct connection between any endpoint pair in a deployment. This minimises latency, as minimal network distance, and associated network latency, is considered in data transport. As such this architecture is the default mode of operation for most NMP solutions which aim to provide low–latency performance [6]. In peer–to–peer deployments, each peer is required to provide a discrete stream to each other peer. For large ensembles uncompressed audio streaming may therefore result in a high bandwidth requirement. Required upload bandwidth overhead may be calculated as approximately $BW = N_{channels} \times N_{peers} \times 768$ (considering a sample rate of 48 kHz and associated data rate of 768 kbps) and download is simply the sum of channels received from peers.

Client–server architectures (Figure 2.8) consider each NMP endpoint as a client, and provide client–client transport via an intermediate server endpoint. Relative to peer–to–peer connection paths, additional network distance is traveled in client–server architectures, thus incurring additional latency [6]. Here, however, bandwidth usage is greatly reduced, improving the scalability of NMP deployments.

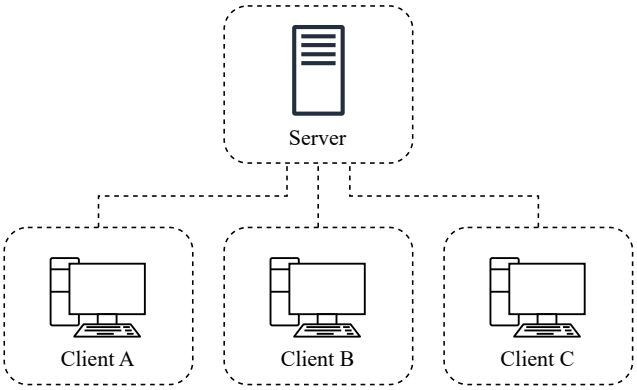


Figure 2.8: Illustration of client–server network architecture.

Here each NMP endpoint only needs to upload its own input channels to the server, and download a mixed monitoring feed of all other endpoints, such that bandwidth requirements do not scale with number of performers in an NMP ensemble. Audio processing may also be deployed to servers, providing audio rendering services for NMP clients without any client processing load [87], [88]. In practice, server–based NMP is supported by several modern NMP solutions, including JackTrip [87] and SoundJack [89].

2.5.7 Audio Compression

As a channel of 48 kHz, 16–bit audio requires 768 kbps of bandwidth [62], it can be observed that using peer–to–peer networking as generally required for low–latency NMP, bandwidth requirements can become a concern for endpoints encountered in practice [8]. The use of audio compression codecs allows for significant bandwidth reduction in NMP. Specifically the Opus codec [90], [91] is utilised in several popular NMP software solutions such as SoundJack [15], [89] and SonoBus. Usefully the Opus codec also handles spatial audio formats [92], [93]. Additionally FEC PLC functionality is also available native to Opus [15]. Opus compression is perceptually transparent for most listeners at 64 kbps in headphone reproduction [93], providing significant bandwidth reduction. Particularly in the context of NMP, Opus incurs minimal additional latency [94]. The codec can handle frames as small as 64 samples, allowing for operation with small frame sizes required to minimise latency in NMP. As blocking delay associated with a 64 sample frame of audio is incurred in the audio interface, the only additional blocking delay incurred by the Opus codec is negligible latency in encoding and decoding, and a look–ahead latency, which may be as small as 32 sample (half the minimum frame size, 0.667 ms at a sample rate of 48 kHz). As such, Opus may be implemented in NMP while incurring less than a millisecond of additional latency. Alternatively the Fraunhofer Ultra Low Delay (ULD) codec [95] offers comparable performance, and has seen use in NMP systems [62].

2.6 NMP Visual Display

NMP systems may provide visual display and interface in addition to core audio functionality. Video systems have seen a high degree of usage in NMP [6]. NMP video–streaming solutions typically look to transport compressed high–definition video such that bandwidth requirements are manageable. Various codecs are noted in literature, traditionally various versions of MPEG [30], [96], [97] and MJPEG [29] formats. Other codecs such as H.264 [98] may also be used [89]. Latency incurred in video compression, however, can result in end–end video latency in the order of hundreds of milliseconds, resulting in AV asynchrony with low–latency audio data [89]. Conversely, systems which stream uncompressed video (for example, LoLa [99]) require extreme bandwidth provisions [35], and are generally not practically possible to deploy outside academic networks. Though several publications address the development of video streaming systems for NMP, particularly UDP transport of video in RTP payloads [33], [96], and low–latency compression methods [89], it is common in modern NMP literature to simply find teleconferencing solutions implemented parallel to NMP audio systems [19], [35]. In several cases NMP hardware which includes low–latency video coding is used to stream video between sites [38].

In recent years NMP research has begun exploring the potential for virtual representation of remote performers using VR/XR systems [19], [45], [100], [101]. Here it is considered that metadata required to remotely control avatars in visual display is low–bandwidth, less sensitive to packet loss, and low–latency in comparison to video. This is dependent on method, however. For example experiments with volumetric capture and playback in XR systems have reported delay between performers of almost 0.5 s [101]. Motion–photon delay (describing time difference between an action and playback of that action on visual display) has been reported for motion–tracking based XR NMP systems with latency of 74 ms, whereas teleconferencing video delay here was 144 ms [19]. In VR/XR visual display and interface for NMP, avatar animation metadata is typically streamed between endpoints as OSC [102] messages in UDP payloads [13], [19].

2.7 Relevant NMP System Software

Presented in the following subsections is further detail on several specific software solutions relevant to study of NMP presented in this thesis, namely audio routing, rendering, and NMP streaming solutions. These solutions have been identified as providing low–latency NMP streaming functionality as outlined in section 2.5 and additional routing and rendering utility as outlined in section 2.4, as required for practical remote performance workflows.

2.7.1 Audio Routing: Jack Audio Connection Kit

Various solutions for audio routing between applications exist that may be used in NMP deployments. This includes proprietary OS–specific solutions (for example Blackhole[8] or AudioMovers OmniBus[9]) solutions integrated into DAWs such as Reaper[10], and open–source cross–platform solutions, such as Jack Audio Connection Kit (JACK)[11]. Various legacy solutions also exist (for example SoundFlower[12] or ReWire[13]). For the work described in this thesis, JACK was selected as the audio routing application used.

JACK is a sound server Application Programming Interface (API), used for managing interface between audio applications and the sound card. JACK is open–source, free–to–use and cross–platform, providing a generalisable solution for audio engineers in NMP contexts. In NMP practice, JACK is used as a daemon application via Command Line Interface (CLI), or via control GUI such as QjackCtl[14]. For CLI operation, JACK systems may be configured as shell scripts.

With respect to NMP latency considerations, JACK optimally adds no additional latency on top of existing application and sound–card buffers. Industry–standard Digital Audio Workstation (DAW) applications (for example ProTools and Reaper) and the majority of sound card and ASIO drivers will interface with JACK, as will a wide range of audio applications. Particularly in the context of NMP, various popular NMP software solutions will also interface with JACK, including JackTrip, SoundJack, and SonoBus.

2.7.2 Audio Processing Application: Reaper

In the context of NMP, audio rendering applications are, in practice, required to operate as plugin hosts. Various cross–platform solutions exist here including usage of DAWs such as Reaper [103] and Ardour[15]. Open–source plugin–host applications are also available (such as Carla[16]) or real–time media processing and programming applications such as Max MSP[17]. For the work described in this thesis Reaper was used to host audio processing. Reaper is a proprietary DAW software and is a common audio production solution in industry practice. With respect to the context of audio processing for NMP, Reaper provides various standard audio processing, routing and recording utilities in addition to hosting of audio plugins. Particularly considering immersive applications, Reaper is capable of handling high channel counts which may be required by immersive audio formats, and supports Virtual Studio Technology (VST) audio plugins which are typical in immersive audio effects.

[8] https://github.com/ExistentialAudio/BlackHole

[9] https://audiomovers.com/omnibus/

[10] Specifically ReaRoute [103].

[11] https://jackaudio.org/

[12] https://github.com/mattingalls/Soundflower

[13] https://archive.steinberg.help/cubase_pro/v11/en/cubase_nuendo/topics/rewire/rewire_c.html

[14] https://qjackctl.sourceforge.io/

[15] https://github.com/Ardour

[16] https://kx.studio/Applications:Carla

[17] https://cycling74.com/products/max

Specific to usage of Reaper in NMP systems with JACK, the two applications may be configured to allow any number of virtual inputs and outputs. Additionally, the two applications will allow Reaper to operate without any additional application buffering latency[18].

2.7.3 NMP Application: JackTrip

JackTrip[19] [75] is an open source, free–to–use, and cross–platform NMP software developed over the course of two decades by various contributors, with origins in early NMP research from the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University [31]. The software operation follows the low–latency streaming outline presented in section 2.5, where JackTrip streams uncompressed audio, provides FEC PLC, and synchronises endpoints through under/over–run condition handling (i.e., block resampling) [75]. JackTrip supports both peer–to–peer [75] and client–server operation [87].

Control of JackTrip is achieved via a desktop application GUI in recent versions. Current and legacy versions may be also controlled via CLI operation, allowing configuration and launching of JackTrip deployments from shell scripts. In addition to the JackTrip desktop application, microcontroller and Pi 4–based installations of the application have been accomplished [104], [105]. JackTrip imposes no limit on the number of channels of audio streamed (notably a practical feature for multichannel formats and large ensemble sizes at endpoints). With respect to interoperability, JACK may be used to route audio to and from JackTrip. Notably, this software is used widely in recent and current NMP research, and performance has been validated through significant documentation in literature. Examples include [7], [8], [35], [63], [83], [106]. JackTrip is also used successfully in a variety of concert performances (for example [35]). The software is also supported by an active user community[20], and provision of virtual machines and a modified JackTrip version for server–based NMP has been commercialised as a service for musicians[21].

2.7.4 NMP Application: SoundJack

SoundJack[22] is a proprietary, free–to–use and cross–platform NMP software developed by Alexander Carôt [89]. Similar to JackTrip, the low–latency UDP streaming approach outlined in section 2.5 is used, streaming uncompressed audio, and providing endpoint synchronisation through over/under–run handling (i.e., block resampling) [89]. Though SoundJack has previously employed Fraunhofer ULD compression [62], the modern version provides an Opus codec option for streaming, which includes FEC PLC utility. SoundJack may be routed to and from via JACK, however SoundJack imposes a channel limit of 16 channels on each endpoint.

[18] Though it should be noted that internal plugin buffering latency [65] may still be incurred by hosted plugins.

[19] `https://jacktrip.github.io/jacktrip/`

[20] `https://groups.google.com/g/jacktrip-users`

[21] `https://www.jacktrip.com/`

[22] `https://www.soundjack.eu/`

Both peer–to–peer and client–server operation is supported. Notably SoundJack consists of a console application back–end controlled via a browser GUI front–end. This web interface allows for other browser–based utility, such as user profiles and text chat. In addition to deployment on typical OS, SoundJack has also been deployed using Pi 4–based systems [36]. The solution is supported by active developer engagement and active online community[23], including professional musicians and online choirs[24]. SoundJack is well detailed in NMP literature [15], [36], [62], [89] providing validation of the performance of the NMP solution, and has been used with success in various concert performances (for example [15]).

2.7.5 NMP Application: SonoBus

SonoBus[25] is an open source[26], free–to–use and cross–platform NMP software developed by Sonosaurus using the JUCE[27] framework. SonoBus streams audio following the common low–latency method outlined in section 2.5 (UDP streaming with over/under–run handling). SonoBus utilises the Audio Over OSC (AoO) protocol[28], which itself is an implementation of RTP [76] streamed as OSC messages inside UDP payloads. SonoBus is implemented as a desktop application and as a VST, allowing use of the NMP solution in–DAW without the requirement for additional routing software, and providing a familiar interface for audio engineers. SonoBus provides significant additional utility internally, including metronome, audio recording and playback, audio effects, and input/output routing. SonoBus may stream either uncompressed or Opus–compressed audio.

Notably, SonoBus allows for connections to be made between peers via a signaling server (used in session configuration), negating the need for any practical network setup such as port forwarding. Though SonoBus has not seen the attention given to JackTrip and SoundJack in literature, and as such has not been robustly validated, some literature does report on the performance of SonoBus [13], [45], [83] providing some degree of validation of the practical success of SonoBus as a low–latency NMP solution.

2.8 Evaluation of NMP

With respect to evaluation of NMP, two common approaches have emerged from prior research which effectively assess the quality of NMP experiences and the quality of performed music, namely synchrony analysis and quality of experience evaluation [6]. Synchrony analysis provides evaluation of the objective musical timing in performance, derived from signal analysis of an audio recording. QoE rating provides evaluation of the perceived quality of the NMP experience, focusing on perceived performance conditions and musical quality.

[23] https://www.soundjack.eu/support/
[24] https://www.ianhowellcountertenor.com/soundjack-real-time-online-music
[25] https://sonobus.net/
[26] https://github.com/sonosaurus/sonobus
[27] https://github.com/juce-framework/JUCE
[28] https://github.com/essej/aoo

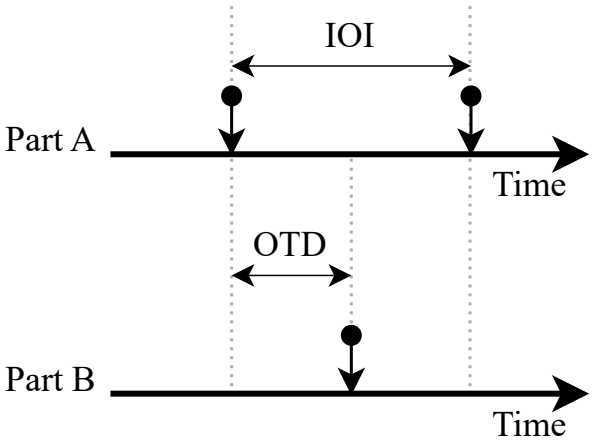


Figure 2.9: Illustration of IOI and OTD synchrony metrics.

Both synchrony analysis and QoE rating have been demonstrated in prior research as capable of identifying differences in evaluation of NMP. In addition, these pseudo–standard evaluation methods, other forms of evaluation which have seen less validation are used in NMP study. This includes extensive immersive experience modelling and presence questionnaires, semi–structured interviews, and metrics derived from motion capture. The following subsections provide detail on evaluation methods for NMP.

2.8.1 Synchrony Metrics

Performance synchrony [55] describes the timing of musical performance, where metrics are derived from onset times of musical events. Though the perceived onset time of a musical event is most accurately described as a probability window, practical signal analysis methods used in onset detection [107] software will define an onset as a single point in time within each musical event. Within–parts synchrony metrics are derived from the IOI [108], describing the time difference between the onset times of sequential musical events within a part (Figure 2.9).

Synchrony metrics derived from IOIs may be described with respect to time (i.e., raw IOIs), or with respect to tempo. The conversion from IOI to tempo, $\bar{\delta}(n)$, is defined mathematically in Beats Per Minute (BPM) as:

$$\bar{\delta}(n) = BD_n * (60/IOI_n) \tag{2.1}$$

Here BD is the Beat Division[29] and N is the number of IOIs in a performance recording, such that the IOI index is described by $n = 0, 1, 2...N$.

[29] Where $FullBeat = 1$, $HalfBeat = 0.5$ etc.

IOI–based metrics of synchrony from prior NMP research include:

- *Pacing*, defined simply as the mean IOI across a performance [6], [16], [109]. This may also be expressed as *Mean Tempo* through the inclusion of tempo conversion as described in Equation 2.1 [19], [110], [111].
- *Regularity*, defined as the Coefficient of Variability (CV, ratio of standard deviation to mean) of IOIs across a performance [6], [16], [109].
- *Tempo Jitter*, defined as the variance of residuals (sum of squared difference between individual tempo samples and linear regression of tempo across a performance) [57].
- *Asymmetry*, ambiguously defined with respect to tempo consistency, described by variation in IOIs [19].
- *Tempo Slope*, describing the gradient of the linear regression of tempo across a performance, often referred to as *Mean Tempo Slope* [8], [16], [19], [57], [58], [108], [110], [112]–[115].

IOIs–based metrics are also in some cases used to compare performance between parts, expressed in terms of IOI–based metrics within each part. Examples here are ratios of tempo–based metrics, such as *Mean Tempi Ratio* [8], [111], [116], describing the ratio of mean tempo from one part to another (such that perfect tempo synchronisation would result in a ratio of $1 : 1$). Another example is tempo difference, simply describing the difference of the tempo of a part from the tempo of a reference part [116].

Between–parts synchrony is derived from OTD, describing the time difference between musical events which are intended to be concurrent [6] (Figure 2.9). OTD may be described either as *absolute OTD* (time difference between onsets of any two discrete parts) or *relative OTD* (time offset of each onset from the mean onset time across parts) [55]. Relative OTD is more practical when describing large ensembles, providing a common reference point across all parts. Absolute OTD, however, may be considered more perceptually relevant, describing actual physical time intervals heard in performance. Metrics derived from absolute OTD are prevalent in NMP synchrony study.

A variety of OTD–based metrics are used in prior NMP research, including:

- Simple expression to raw absolute OTDs between parts, described as *Precision* [56] or *Inter–Subject Time Difference* (ISD) [9] depending on literature.
- The mean of absolute OTDs across ensemble performance, described as *Coordination* [109] or *Mean Precision* [8] depending on literature.
- *Asynchrony*, describing Root Mean Square (RMS) of standard deviations of OTDs across parts in a performance (simply the standard deviation between parts for duet cases) [8], [55].
- *Imprecision*, originally defined as the RMS of IOI variation from reference IOIs across an ensemble [110]. In other literature, however, this metric is reported as the standard deviation of OTDs [6], [16], or is ambiguously defined as variability of OTDs [19].

- Metrics describing the tendency of performers to perform ahead or behind one another (i.e., to lead or follow other performers). *Asymmetry* is defined in some literature as the mean OTD expressed relative to a designated reference part [6], [113], [115]. In other literature the nomenclature *Mean Lead* [110]. In other literature the nomenclature *Tendency to Lead/Lag* [8] is defined as the median OTD relative to a designated reference part.

A lack of consistency in nomenclature, and minor variation of discrete metric definitions can be observed generally in prior NMP synchrony studies. Despite this, a pseudo–standard can be recognised in the consistency with which synchrony metrics describe the averaging or variation of IOIs and OTDs between parts in terms of statistical measures.

2.8.2 Quality of Experience Metrics

QoE evaluation of NMP describes the perceived quality of aspects of the NMP experience through response to QoE items on rating scales. The rating scales used across prior NMP research exhibit a high degree of variation. Discrete scales may express positive or negative rating through agreement with statements[30], through absolute quality rating of bespoke items[31], or through rating of the magnitude of a particular feature[32].

Despite the lack of coherence in metric definition across QoE evaluation of NMP, several items can be identified as common across various studies, and demonstrated in prior study as capable of identifying variation in NMP experience. This pseudo–standard set of QoE rating items includes:

- *Rating of audio delay between performers*, including perceived amount of audio delay [8], [114], [118], the perceived influence of audio delay on performance conditions [9] and perceived performance quality [113], and the perceived degree to which performers perceive audio delay as manageable [8], [9], [35], [108], [113], [119].

- *Rating of performance quality*, including absolute perceived quality of musical performance [16], [109], [118], [120], musical performance ability [16], [113], or perceived performance conditions [8], [19], [115] perceived quality of musical interactions [8], [35], [96], [113], [116], [118], [121], or perception of performance synchrony [114], [118], [122].

- *Rating of perceived overall experience quality* [110], including willingness to participate in NMP experiences [8], [123], enjoyment [8].

- *Rating of perceived audio quality* [35], [118] including BAQ [83].

[30] i.e., strongly disagree–strongly agree
[31] i.e., bad–excellent, as per Basic Audio Quality (BAQ) [117]
[32] i.e., none–lots

Metric	Definition
Auditory co–presence	*"The auditory sensation of being 'together' with your co–performer in a similar way as you experience in traditional co–located rehearsal or performance..."* (with reference to [126]) [16].
Social presence	In terms of agreement with the statement *"While playing I felt like if I was in the same room together with the other musicians"* [121].
Social presence	In terms of agreement with the statement *"I felt like I was in the same room with my remote counterpart"* [19].
Involvement and flow	Described with respect to a subset of rating items addressing presence, attention and engagement in networked interactions [113].

Table 2.2: Rating factors relating to measurement of presence. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items.

Metric	Definition
Connection	Defined with respect to rating of agreement with the statement *"I felt connected with the other musicians."* on a semantic scale [121].
Involvement and connectedness	Measured in terms of agreement with a series of relevant statements [9].

Table 2.3: Rating factors relating to measurement of interactivity. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items.

In addition to these pseudo–standard rating items, other less common rating items are found in literature. Examples include rating of emotional response to musical performance [118], [120], rating of perceived creativity [122], rating of visual display and contact [19], [114] and rating of the usefulness of audio monitoring [35][33].

2.8.3 Immersive Experience Modelling and Presence Questionnaires

In some prior work NMP is considered as an immersive experience, and evaluated using extensive questionnaires which are typical in the study of presence [124] and immersion [125] (Figure 2.10). Such questionnaires aim to provide comprehensive modelling of immersive experiences through rating of subsets of items designed to reflect different aspects of the experience. Examples of relevant rating items which have been adapted to the context of NMP consider immersive experience modelling factors relating to interactivity (Table 2.3), presence (Table 2.2), immersion (Table 2.4), and plausibility (Table 2.5).

[33]This rating item is most specifically relevant to hybrid performance contexts, where musicians perform with both remote NMP musicians, and physically present musicians.

Metric	Definition
Quality of immersion	Described in terms of agreement with a series of relevant statements [9].
Quality of immersion	Described in terms of semantic rating of a series of relevant statements. Defined as *"related to the presentation of the stimuli and can be defined as the set of valid actions supported by the mediating environment"* (with reference to [127]) [113].
Immersion	Described via rating of agreement with the statement *"I felt being involved in the virtual performance space"* [19].
Interface awareness and quality	Described with respect to a subset of rating items addressing attention and distraction factors [113].

Table 2.4: Rating factors relating to measurement of immersion. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items.

Metric	Definition
Naturalness	Defined with respect to rating of agreement with the statement *"I played in a natural way"* on a semantic scale [65].
Naturalness	Defined via rating of agreement with the statement *"my musical interactions in the virtual performance space felt natural"* [19].
Auditory cohesion	Defined as *"The sensation that the sound you create and the sound you hear plausibly fit together in your current physical space..."* [16].
Coherence	Measured in terms of agreement with a subset of relevant statements [9].
Coherence	Defined in terms of *"perceived realness" and "predictability"* [113]..
Consistency	Defined via rating of agreement with the statement *"The virtual performance space was very similar to reality"* [19].

Table 2.5: Rating factors relating to measurement of plausibility. Definitions from relevant literature are detailed, or summarised where defined with respect to subsets of QoE items.

This approach to NMP system evaluation is a relatively recent addition to evaluation methods, and as such has not achieved the pseudo–validation achieved through decades of experimentation with certain QoE rating items. Nonetheless, certain immersive experience modelling and presence rating aspects (namely rating of presence, immersion, and interactivity elements) have been demonstrated as capable of detecting variation in NMP experiences.

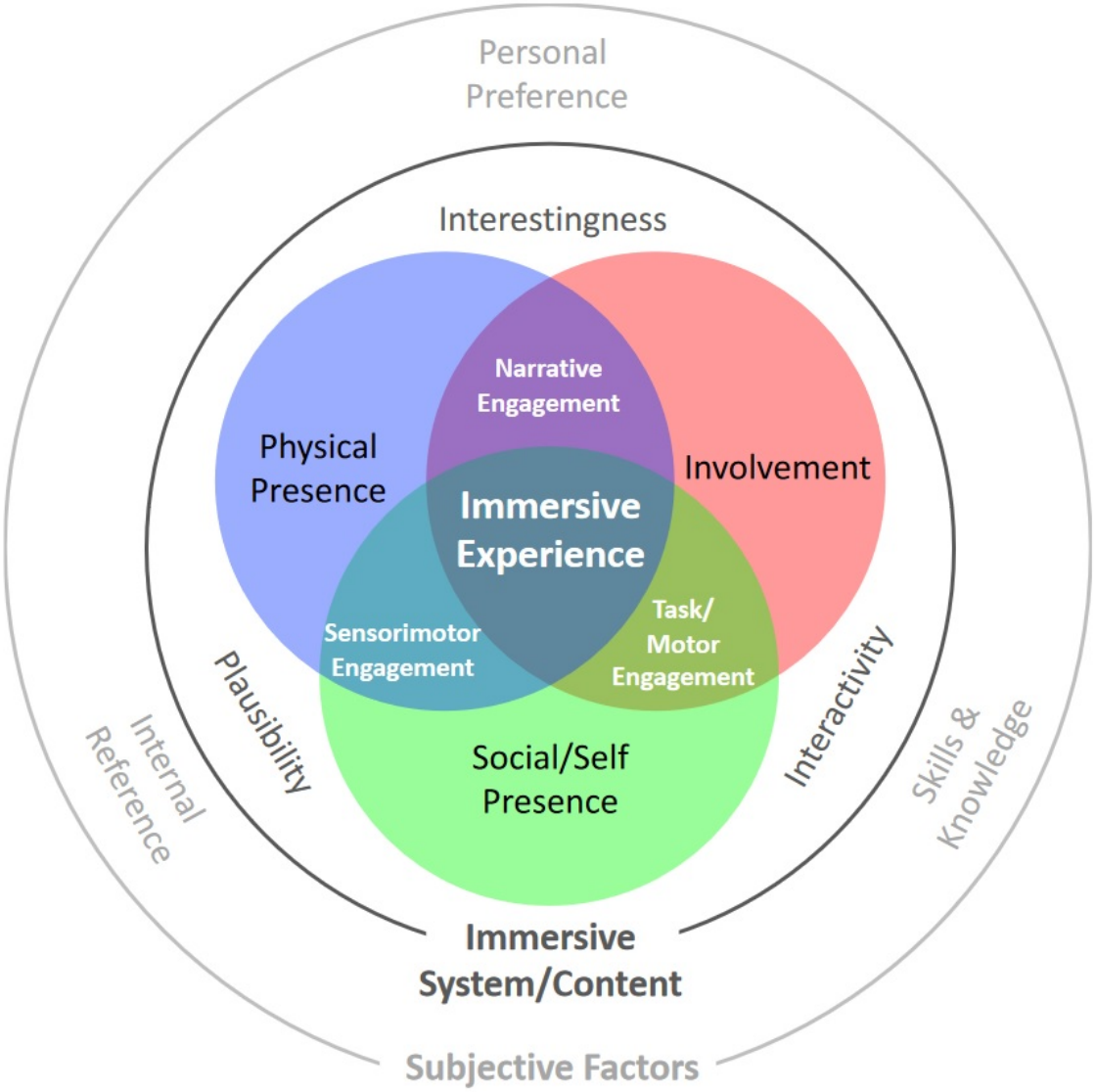


Figure 2.10: Example illustration of immersive experience and presence modelling factors, from [125].

2.8.4 Semi–Structured Interviews

The usage of semi–structured interviews in NMP [15], [43], [44], [128] allows for wider discussion around the performance experience. Here, unlike QoE evaluation (which is investigator–directed), interviewees may provide comment on the experience which may be overlooked in QoE protocols, or may require more nuanced feedback than can be communicated via a rating scale. Prompts for semi–structured interviews are generally dependent highly on the evaluation context. In NMP settings interview prompts may invite comment on challenges found in NMP, perception of performance using NMP systems relative to live conditions, and ability to interface with NMP technology [43], [128]. Alternatively comment may be invited which is relevant to visual display and visual contact between musicians [128].

No pseudo–standard interview prompts are identifiable in literature as demonstrably capable of identifying variation in NMP experience, however, this is likely primarily due to the free–form nature of this evaluation approach, rather than inability to elicit statement of discernment from interviewees.

2.8.5 Gaze and Motion Capture

In the case of NMP systems which use visual display, related evaluation may assess visual contact in performance. Evaluation methods here may monitor physical parameters associated with visual contact between performers using motion–capture technology. Measurement of performer gaze direction has been used in prior study to evaluate time spent looking at visual display [120], [122]. Alternatively head movement distance is used in prior work, considered as a method of monitoring the extent of head movements used in communication of visual cues in performance [122].

2.9 Previous Study of Network Music Performance

Previous study of NMP provides evaluation with respect to previously discussed metrics (predominantly synchrony and QoE). This work largely addresses the effect of latency [6], where a high degree of consistency in the findings of prior work allows some empirical understanding of the context. Other work expands on latency study, providing evaluation of other aspects of the NMP experience. This includes evaluation of the effect of timbre [108], [114] and dynamics [111] of instrumentation, the effect of performance material [108], [119], and the influence of visual display [13], [19], [114], [120], [122]. Cumulatively, these prior evaluations inform best practice in practical applications of NMP systems. The following subsections provide a review of previous study of NMP.

2.9.1 Effect of Latency on Performance Synchrony

Early NMP latency experiments focused on tapping studies [53] where participants attempted to synchronise tapping with no visual contact between participants. In these studies it was identified that tempo deceleration was likely to occur when the OWT audio delay between performers exceeded approximately 30 ms [57], [58], [110]. Further work expands on this on these tapping studies through evaluation of instrumental performance, where similar tempo deceleration and a similar latency threshold was found [8], [109], [119], including in evaluation conducted with video display [113].

Studies also address synchrony between parts, noting deterioration of relevant synchrony measures at audio delays in excess of 30 ms [8], [109], [113]. In some cases a higher latency threshold of approximately 50 ms is suggested for synchrony deterioration [110], [116], however the 30 ms threshold is commonly accepted in literature [6]. Below the 30 ms latency threshold synchronous performance is demonstrated in studies which look to validate NMP deployments[34] [8], [9].

[34] It can be considered that no significant difference in synchrony from a live or zero–latency reference is suitable for validating the ability of an NMP deployment to support synchronous musical performance.

At latencies below approximately 10 ms, it has been noted that there is a tendency for musicians to naturally accelerate tempo [9], [57], [58], though no other impairment of synchrony is noted here.

2.9.2 Effect of Latency on Quality of Experience

Similar to performance synchrony, evaluation of QoE in prior NMP research also identifies a deterioration of evaluation metrics (see subsection 2.8.2) with excessive latency. Rating of metrics relating to perception of delay have been shown to deteriorate with audio delay in excess of 33 ms [113] 36 m [108] and 40 ms [114]. Here the latency threshold at which ratings deteriorate appears to approximately follow the 30 ms latency threshold suggested by synchrony study.

Ratings of metrics relating to perceived performance quality and conditions, ratings have been shown to deteriorate or be rated poorly at, and in excess of, 40 ms [114], 50 ms [116], 56 ms [109], and 81 ms [35], with high–quality rating achieved at and below 20–26 ms [109], [110], [114], [116]. Though reports of deterioration of rating due to latency are less consistent here, all values associated with experience impairment are notably in excess of the 30 ms latency threshold suggested by synchrony studies. Metrics relating to overall enjoyment, such as willingness to participate [123] or overall judgment of the experience [110], have also been suggested as deteriorating as latency increases.

2.9.3 Effect of Latency on Immersion and Presence

Study of presence and immersion under varying latency is limited in NMP literature, though a small number of studies exist. Deterioration in rating of factors of immersion, plausibility and interface awareness have been noted for latency at 33 ms relative to 28 ms latency [113] and above, whereas other study notes deterioration of factors relating to presence at 40 ms and above relative to 20 ms [16]. In other research high rating of presence, immersion and probability factors have been recorded when evaluating systems deployed with audio delays of 8 ms [9] and 21 ms [19]. Though data here is limited, results again appear to conform to the expectation for NMP experience deterioration above 30 ms.

2.9.4 Latency–Coping Strategies and Metronomes

Prior work considers that impairment of NMP relating to latency may also be detectable with respect to performance strategies used to counteract the effect of latency [129]. The *leader–follower* strategy [110], [112], [128] may be described as the case where one performer ignores the performance of others, and instead attempts to maintain steady performance timing (the *leader*), such that other performers (the *followers*) may perform synchronously to this part (effectively acting as a tempo reference for other performers).

The *leader–follower* strategy may be characterised by synchrony metrics which describe the tendency of performers to play ahead or behind other performers (as detailed in subsection 2.8.1).

Here, *follower* parts lag behind a *leader* part. Where the *leader–follower* latency–coping strategy is used, it is inferred that networked performers are experiencing synchrony impairment due to latency. As such, synchrony analysis of lag which describes a *leader–follower* strategy is considered to infer that networked performers are experiencing synchrony impairment due to latency.

Alternatively to synchrony analysis, the *leader–follower* strategy may be identified through self–report by musicians in discussion or QoE rating. Here performers may indicate the degree to which they feel the *leader–follower* strategy was used.

Evaluation of leader–follower strategies in NMP latency study has suggested an increase in perceived adoption of the strategy at 40 ms and above (relative to 20 ms and below) [114], and greater tendency to lead in synchrony measurement above 33 ms (relative to 27 ms and below) [8]. In other work, evaluation of performance at different latency levels is provided with and without a metronome to stabilise performance (acting as a reference for performer timing similar to the leader in leader–follower strategies). Usage of both acoustic and visual metronomes was found to stabilise performance synchrony at 55 ms latency [115].

Notably in neither the case of metronome or leader–follower approaches to timing stabilisation is the musical interactivity the same as sub–30 ms conditions [129]. In each case performers are primarily interacting with the musical timing reference, rather than establishing synchronous performance solely through musical interactions between performers[35]. An alternative method to latency–coping which is popular in high–latency performance is simply to perform avante–garde or ambient material where strict timing is not required (i.e., the *"latency accepting approach"* [129].

2.9.5 Effect of Performance Timbre and Dynamics

A small number of studies evaluate the influence of dynamic and spectral qualities of instrumentation on synchrony and QoE in NMP. Notably here one study identifies varying synchrony results for woodwind and string instruments [109]. In another study it is suggested that synchrony is more sensitive to latency in performance with fast attack times, rather than slow [111], and in performance with bright and noisy timbre [108], [114].

2.9.6 Effect of Performance Material

Aspects of performance material have been evaluated in some NMP studies. Prior evaluations here suggest that performance synchrony and QoE will be more sensitive to the effect of latency when performing material at a high tempo [108], [115], [119], and where the rhythmic complexity of performance material is high [119].

[35]Though it can be acknowledged that performing to a click track is typical in many modern music performance contexts.

2.9.7 Effect of Prior Practice

Throughout NMP literature it is considered that prior practice may influence the ability to achieve synchronous performance in NMP, particularly with respect to performance at latency above 30 ms. Prior work indeed suggests informally that performance quality will improve with practice [6], [35], [116]. More comprehensive evaluation has demonstrated that throughout NMP trials the synchrony of delivered performance is indeed likely to improve as performers become more experienced with the novel context [16].

2.9.8 Evaluation of Visual Display

For NMP systems with visual display and interface, evaluation of the use of visual display, and visual contact in performance, is considered in prior work. In some literature it is reported that performers will typically not use video display in practice due to excessive video latency and AV asynchrony [19], [97], [100]. In other work, however, video display is reported as useful with respect to visual performance cues and communication between performers, supported by self–report from musicians [122], [128], and monitoring of gaze in evaluation [120], [122].

In an investigatory study of systems which use VR and XR display, it is reported that avatar representation is important to the quality of interaction and communication of visual performance cues between performers, particularly hand–tracking and facial expression [13], [128]. Robust comparative study of video display and avatar representation in projector–based XR display [19] suggests that XR display is preferable to video display, being rated as generally providing improved perceived performance conditions, and less likely to cause impairment of performance due to visual inaccuracies and glitches. Additionally XR display is associated with improved rating of presence, immersion, plausibility and interaction features. In another study video display is reported as preferable to simple VR avatar display in some regards, primarily with respect to the ability to communicate via facial expression, and the plausibility of visual interactions [128].

2.10 Musical Performance Practice for NMP

The following subsections provide a non–exhaustive summary of historical and modern performance practice in NMP research, and in practical application of NMP systems over public WANs.

2.10.1 Performance Practice in NMP Research

Early research in NMP focuses on tapping studies, where simple clapping rhythms are performed [57], [58], [110]. These performance tasks are in continued use in modern NMP research [19]. This form of performance practice is designed for research purposes. A simple performance task minimises performance proficiency requirements of participation in studies, and simplifies the case of onset detection associated with synchrony analysis.

This performance practice is prominent in work which studies latency tolerance in NMP. Simple tapping performance practice can also be found in non–melodic instrumental performance, for example tapping simple rhythmic patterns between a cello and violin [111] used to evaluate the influence of attack times in NMP.

Further NMP research expands on tapping studies with performance practice which is more representative of practical musical performance scenarios (i.e., classical and popular rhythmic and melodic musical styles). This includes simple compositions created for the purpose of NMP evaluation. Examples include performance between percussion and bass parts, used to evaluate the impact of rhythmic factors in NMP [94], and simple vocal performance melodies, used to validate NMP synchrony over practical WAN deployment [8]. Other instrumental performance practice for NMP includes the performance of well–known compositions, for example classical or pop music. Classical performance practice includes performance between woodwind (clarinet) parts or string parts (violin and viola), used to asses delay tolerance in performance [109]. In other work classical piano duet performance is used to asses delay tolerance [116].

Operatic performance between vocal parts and piano accompaniment have been delivered in evaluation of NMP latency thresholds [120]. Chamber music groups have also been explored in one study which evaluates presence in NMP (here with performance between percussion, strings and woodwind) [113].

In other work pop music is performed between guitar, keyboard and clarinet parts in a study addressing the impact of timbral features of instrumentation in NMP [108]. Further similar work which investigates the impact of timbral features includes performance of folk, rock, funk and classical styles between string, percussion, woodwind and keys instruments [114]. In other work which aims to address the QoE of NMP rock/pop style performance between guitar, bass, piano, keyboard and drums is delivered [9], [115].

2.10.2 Performance Practice in Practical NMP Applications

Beyond performance practice designed for academic research, a range of NMP systems have been deployed in practical applications. Here performance practice is more ecologically representative of performance contexts in the music industry. Early developments in NMP technology experimented in live–streaming of concert performance across a range of musical styles. Examples here include streaming of swing jazz [30], and symphonic performance [33]. These early developments progressed to supporting early NMP concert events, involving musicians interacting across multiple endpoints. Examples here include concert performance of post–modernist compositions [29], and piano and accordion performance [33], [96].

A wide variety of NMP concert events have since been delivered. Where events were not restricted to LANs, core networks, and private network connections over high–speed academic internet, however, latency was a notable issue. Performance practice here developed from early networked performance [28] where performance material is designed so as not to require rhythmic timing, thus negating the limitations of latency on performance of popular rhythmic musical styles. Examples here include forms of interactive sonic installation [21], ambient and experimental performance material and improvisation [22].

This performance practice continues in modern NMP concert practice [18], [20], [23], [104]. Alternatively, other NMP concert practice includes the delivery of synchronous rhythmic styles, for example classical music [35], where latency–coping strategies (subsection 2.9.4) are implemented to tolerate high latency conditions. In modern NMP concert practice low–latency deployment over public internet connections has been achieved. Here performance practice expands to include pop music styles which can be supported by low–latency deployment [15]. Pro–LAN audio industry, namely Dante, have also experimented with low–latency NMP, again exploring pop music styles [40].

Modern low–latency NMP practice has expanded beyond specialist concert events to include community–based performance practices. Included here are examples such as community choir activity supported by the SoundJack NMP streaming application [36], [130]. Popular NMP software solutions are in routine use by a large community of users for a range of performance practices. This activity is supported through user forums, user groups, and other community resources for software such as SoundJack[36][37], JackTrip[38][39] and SonoBus[40]. NMP systems are also trialled for a range of instruments and musical styles in music tuition applications [42], [43].

2.11 Summary

The review of NMP provided allows the identification of conditions under which synchronous music performance of popular musical styles is achievable remotely. Various novel remote performance systems exist, for example use of high–latency teleconferencing solutions and latency–coping strategies such as leader–follower or ambient performance, or use of networked sequencers or other forms of sonic control. These systems, however, provide an experience which deviates significantly from the natural musical interactions required for synchronous performance of popular music in live contexts. Low–latency NMP streaming methods, however, are capable of allowing such natural synchronous musical interactions remotely across national distances. As such these methods are identified as providing a viable solution for remote musical performance workflows which aim to be practical remote alternatives to musical performance applications which typically require musicians to co–habit a physical location.

A common approach to low–latency audio streaming is identifiable here, and an overview of relevant methods and practice has been provided, as well as identification and review of specific software solutions which are of particular relevance to the work described in this thesis. The performance of these solutions is identifiable as having undergone validation in prior work, and the suitability of these systems for extension with audio processing functionality is outlined, demonstrating the utility of these solutions in a NMP research and development capacity, as well as in practical application of remote performance.

[36] https://www.soundjack.eu/
[37] https://www.ianhowellcountertenor.com/additional-resources
[38] https://groups.google.com/g/jacktrip-users
[39] https://www.jacktrip.com/
[40] https://groups.google.com/g/sonobus-users

Though a lack of standardisation is sorely apparent in prior evaluation of NMP, a high degree of repeatability can be found in evaluation approaches and findings. A pseudo–standard evaluation protocol can be derived from prior evaluation of synchrony and QoE which is suitable for use in study of NMP despite a lack of formal evaluation method validation in the field. In addition to pseudo–standard items, presence and immersion questionnaires also appear to yield evaluation items which will identify experience differences in studies.

From prior study of latency, strict audio OWT delay guidelines can be noted (that a practical remote performance workflow which aims to allow synchronous performance and natural musical interactivity between performers should achieve latency below 30 ms). Other NMP evaluation, though sparser and less robust than latency study, provides suggestion of expectations which are useful in guiding further evaluation. It is noted that there is potential for synchrony and QoE results to vary across timbral, dynamic, instrumentation, visual display, performance material and prior NMP experience level participant groupings.

Chapter 3

Review of Immersive Audio for Network Music Performance

...these go to eleven

Nigel Tufnel,
This is Spinal Tap (1984)

3.1 Preface

INMP systems extend NMP technology with immersive audio capture, processing and playback functionality. This chapter provides a review of INMP technology, research and practice. INMP systems are introduced, briefly summarising practice and system design in prior work and current state–of–the–art. A review of the fundamentals of spatial sound is given, followed by an overview of spatial audio auralisation. Detail is given on specific spatial audio methods which are of relevance to INMP practice and system design, namely Ambisonic and Binaural Stereo approaches. A detailed review is provided of INMP audio system implementation in prior work. A review of INMP system evaluation is then given, including evaluation methods and findings of prior studies.

3.2 A Brief Overview of INMP

Low–latency capture, streaming and playback of traditional mono and stereo audio is sufficient to facilitate synchronous remote performance using NMP systems. It is, however, well acknowledged that this experience is substantially different from the experience of musical interactions between performers who cohabit a physical location. A long–standing interest of NMP system development is therefore the extension of core audio capture, streaming and playback functionality with system components designed to make the experience of networked musical interactions more like that of live musical interactions.

Included here is the integration of immersive audio technology, such that networked performance may be heard as if together with remote performers in a 3–D acoustic space [131]. The terminology INMP is used to refer to NMP systems which include immersive audio functionality [7], [8], [12], [13] and is adopted in this thesis.

Historic INMP design considers spatial microphone systems, used to reproduce the experience of listening in the room with networked performers over remote loudspeaker arrays [29], [31], [33], [34], [96], and spatial panning methods, used to auralise networked performers in remote listening environments over loudspeaker arrays [21], [22]. Here immersive audio systems extend traditional NMP, enhancing video display in systems for NMP musician monitoring and distribution of networked performance as virtual concert to audiences listening in remote venues.

Modern developments in INMP expand on prior work, focusing on immersive audio acoustic simulation and DSP, and interactive rendering of immersive audio for headphone playback [65], [131], [132], such that networked performance may be heard as if together with performers in a simulated 3–D acoustic environment. In addition to enhancement of video display, inclusion of immersive audio rendering is considered as a requirement of usage alongside interactive immersive visual media such as 360° video, VR and XR display systems. Here interactive immersive audio rendering provides AV coherence in navigation of real or virtual environments [25]. Such modern INMP designs are considered as solutions for remote performance workflows in various practical applications, particularly where immersive systems are considered. Examples include virtual concert performance, virtual production, and virtual learning and therapeutic applications [131].

Prior practical implementation of INMP over WAN (largely systems for musician monitoring and virtual concert distribution), however, describes high–latency contexts [18], [20]–[23], [104], [133]–[135]. Here performance of electroacoustic, ambient and experimental performance material is considered as a means of coping with impairment of networked musical interactions associated with high latency deployment[1]. It can be considered, however, that practical implementation of INMP in proposed applications should allow for unimpaired performance of music of popular rhythmic styles, and thus requires design and deployment of low–latency (i.e., sub–30 ms) INMP systems. The design of low–latency INMP systems is the subject of experimental LAN–based systems integrating VR and XR technology [13], [19], [20], [136], [137], the topic of theoretical system design and estimated performance [65], [121], [132], and estimation of performance from simulated networks [70]. Practical WAN deployment of low–latency INMP is, however, extremely limited [7], [8], and does not address INMP system design or performance in practical proposed applications.

3.3 Fundamentals of Immersive Audio

The development of low–latency immersive audio rendering systems for INMP considers various approaches to delivering to remote listeners a simulation or reproduction of the experience of hearing remote performers in a 3–D acoustic space.

[1] I.e., in excess of 30 ms as discussed in Chapter 2.

As such, understanding of the design and operation of INMP systems requires an understanding of the modelling of sound in 3–D acoustic space. The following subsections provide a review of the basics of sound in 3–D acoustic spaces and relevant acoustic simulation methods.

3.3.1 Sound in 3–D Space

Sound propagates through 3–D space as the compression and rarefaction (i.e., longitudinal wave) of air particles (i.e., sound pressure). The radiation of sound in 3–D space may be described with respect to the amplitude, frequency and directionality of sound pressure [10]. The amplitude of sound (i.e., intensity) is described by Sound Pressure Level (SPL), measured in Decibel (dB), and defined in terms of the ratio of measured sound pressure to a reference pressure level:

$$L_P = 20\log_{10}(p/p_0) \tag{3.1}$$

where p is the measured SPL and p_0 a reference pressure level, and where a reference pressure of 20 µPa is given in relevant standards [138]. The frequency of sound, f is defined with respect to the speed of sound, c, (as defined in subsection 2.4.1), and wavelength of longitudinal waves of sound pressure, λ:

$$f = c/\lambda \tag{3.2}$$

Description of the direction of sound sources is generally given as the location of a sound source relative to a reference location in space[2], expressed in spherical coordinate notation. Here the relative direction of a sound source relative to the reference position is described with respect to the radial distance, r, azimuth, θ (describing rotation around the z–axis, with 0°defined as 'forward', and counter–clockwise rotation defined as positive[3]), and elevation, ϕ (describing rotation around the y–axis, with 0°defined as level with the horizontal plane, and counter–clockwise rotation again considered positive) (Figure 3.1). Azimuth and elevation may be derived (including radians–degrees conversion) from relative cartesian description of sound source direction as:

$$\theta = (180/\pi)\tan^{-1}(y/x) \tag{3.3}$$

$$\phi = (180/\pi)\tan^{-1}(z/\sqrt{x^2+y^2}) \tag{3.4}$$

[2]Typically the listener position in immersive audio contexts.

[3]Though it can be noted that in practice many spatial audio software implementations found in audio production consider clockwise rotation as positive, defying conventions [139].

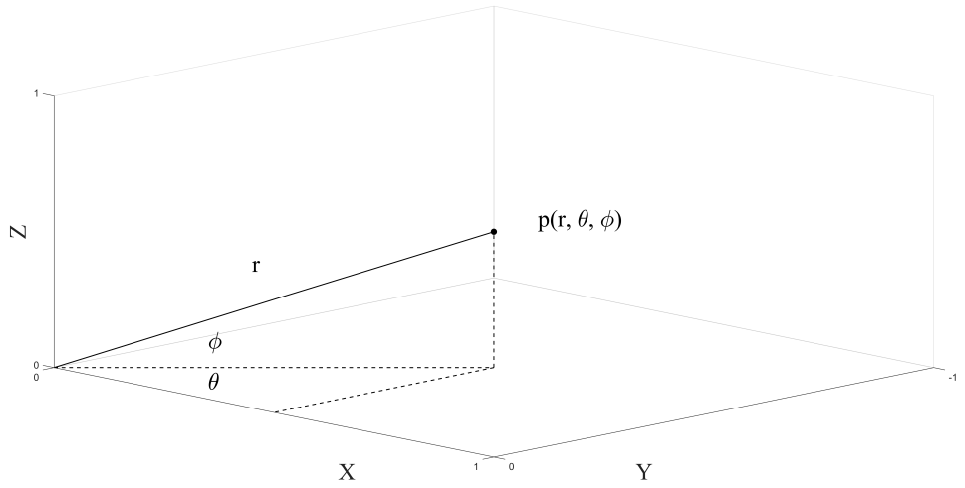


Figure 3.1: Spherical coordinate notation for the point $p(r,\theta,\phi)$.

In the free field the radial distance, r, of a sound source is related to the SPL of relevant sound waves at a point in space as detailed in the equation:

$$L_{P1} = L_{P0} - 20\log_{10}(r_1/r_0) \tag{3.5}$$

such that a halving or doubling of radial distance is associated with an increase or decrease of 6 dB respectively. Here L_{P1} is the SPL at radial distance r_1, and L_{P0} is a measured SPL at radial distance r_0.

3.3.2 Sound in 3–D Acoustic Spaces

As sound propagates through 3–D spaces which include physical boundaries (i.e., rooms, halls and performance venues) sound waves will collide with boundaries, causing sound waves to reflect and reverberate throughout the space. The reverberation of sound in a space [59] is described with respect to the directional qualities of reflections and rate of decay of sound reverberating in a space. This includes frequency–dependent characteristics of reflections (a result of spectral absorption of sound pressure by boundaries and air absorption) and the interference of reflections at a reference position in a space (i.e., room modes). The propagation of sound in a reverberant acoustic environment (Figure 3.2) is generally described component–wise in terms of direct sound, ERs and diffuse reverberation (Figure 3.3). Direct sound describes the signal traveling through the direct in–air path between a sound source and a receiver. Discrete reflections of sound are typically described in terms of reflection order[4]. In some literature ERs are defined as the the 1^{st}–6^{th} order reflections [140].

[4]Where a reflection of order N has reflected from N boundaries in traversal of a sound wave from a source to a receiver.

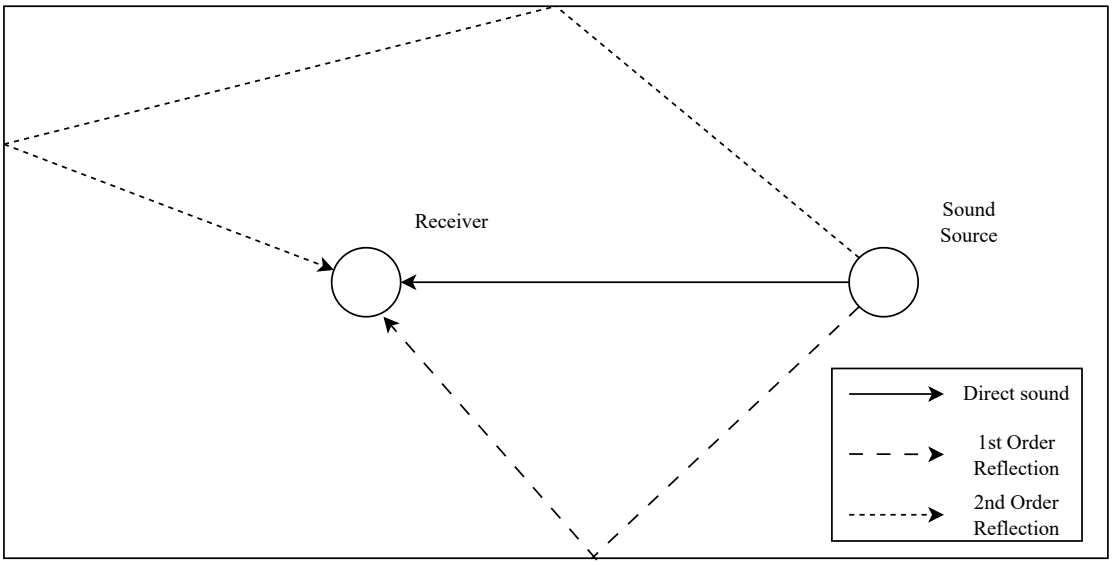


Figure 3.2: Simple depiction of the reflection of sound in a room, showing direct sound path, and different orders of reflection.

Alternatively, early reflections may be defined via arbitrary time values expressed relative to direct sound, for example from the first reflection to 50–100 ms[5] after direct sound [142]. ERs may also simply be defined as all reflections from 1st order up to the beginning of the diffuse component of reverberations. Diffuse reverberation is defined as reflections where the decay characteristics have become approximately even throughout a space[6]. The start of the diffuse reverb component of reflections can be arbitrarily defined with respect to time values used to define the end of ERs. In practice, it is common for the beginning of diffuse reverberation components to be be estimated from room dimensions, where room volume is typically used to define a *'mixing time'* between ERs and diffuse [143].

The relative ratio of direct sound to reverberant energy (Direct–to–Reverberant Ratio (DRR)) for any discrete source and receiver position is defined as [144]:

$$DRR = 10\log_{10}\frac{\sum_{n_{peak}-0.5n_{window}}^{n_{peak}+0.5n_{window}} h^2(n)}{\sum_{n_{peak}+0.5n_{window}}^{\infty} h^2(n)} \tag{3.6}$$

where $h^2(n)$ is the square Impulse Response (IR), n_{peak} is the time of the peak of direct sound, and n_{window} is a time window around the peak of direct sound[7]. DRR varies approximately linearly with distance of sound sources [145], [146].

[5]Choice of arbitrary time value here is typically dependent on room size (as in larger rooms lower order reflections are associated with larger path distances), or may be taken from room acoustic parameters, such as values of 50 ms or 80 ms as per relevant standards [141].

[6]I.e., the order and density of reflections is high enough that reverberant qualities of a space have reached an approximate averaging [59].

[7]A window size of 2.5 ms is noted in related literature [144].

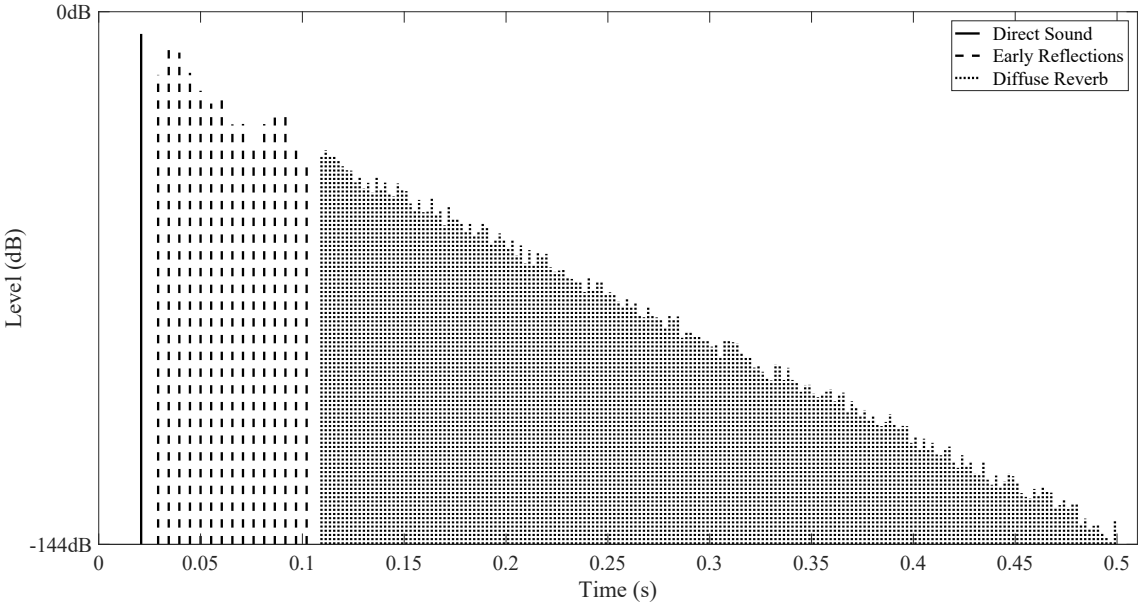


Figure 3.3: Simple depiction of IR representation of room acoustics, highlighting reverb components of direct sound, ERs, and diffuse reverb.

3.3.3 Acoustic IR Measurement and Simulation

The reverberation of sound in acoustic spaces is generally represented as an IR. With respect to DSP an IR provides a 'black box' description of the response of a Linear Time–Invariant (LTI) system, here the difference caused by room acoustics between a signal at a sound source and a signal at a receiver. The IR of a LTI system, $h(n)$ is defined in terms of system input, $x(n)$ and output, $y(n)$:

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(n)h(n-k) \tag{3.7}$$

In room acoustics, an IR is used as a standardised acoustic measurement [141], describing the acoustic response of a room. A discrete IR describes the room response for a discrete sound source and discrete receiver position, and datasets of IR measurements may describe source–receiver positions throughout a space. Acoustic IR measurement [141], [147], [148] may be performed using a range of techniques. The standard and most typical approach uses a loudspeaker to playback a sine–sweep room excitation signal [148], [149], which is recorded using a microphone at a receiver position. Acoustic IR measurement may then be extracted from the audio recording of the excitation signal through convolution with an inverse filter which is appropriately designed for the excitation signal [149]. Acoustic IR may also be simulated, typically using geometric and wave–based acoustic approaches [150], [151].

This can be achieved using practical software solutions such as CATT[8], EASE[9] and Treble[10], or solutions such as MCRoomSim[11] [152]. Measured or simulated acoustic IR may be used to provide a standard set of acoustic parameters considered to characterise the acoustic qualities of the measured or simulated space. The standard set of metrics [141] provide various measures of energy decay and power ratios of IR components. Key acoustic parameters include:

- RT, describing the decay of sound reverberating in a space in terms of the time taken for sound to fall 60 dB below direct sound levels[12].
- Early Decay Time (EDT), describing the time taken for sound energy to decay by 60 dB, based on extension of the linear regression of the decay curve between 0 dB and -10 dB.
- Centre Time (CT), describing the time taken for energy in an IR to decay by half the total energy in the IR.
- Clarity (C_t) describing the ratio of energy in the IR from before to after a specified time, t[13], measured in dB.
- Definition (D_{50}), describing ratio of energy before to after 50 ms from direct sound as a percentage ratio.

Acoustic paramaters may be reported by frequency band, or as a single value derived from weighting [153] of frequency bands [141].

3.3.4 Auralisation

Simulation of spatial sound and room acoustics in INMP allows for the auralisation of audio, such that networked performers may be heard as if listening together with sound sources in a simulated 3–D acoustic environment [150]. Välimäki et al provide categorisation of artificial reverberation methods used in digital simulation of acoustics [154]:

- Delay networks: where networks of delay lines and filters with feedback connections are used to simulate reverberation.
- Convolution algorithms: where convolution with IR is used to simulate reverberation.
- Computational acoustics: where reverberation is simulated through mathematical modelling of room acoustics.

[8] `https://www.catt.se/`

[9] `https://www.afmg.eu/en/ease`

[10] `https://www.treble.tech/`

[11] `https://github.com/Andronicus1000/MCRoomSim`

[12] Most typically the decay envelope will be extended to this point based on the estimated envelope over 20 or 30 dB reduction from direct sound (T_{20} and T_{30} respectively).

[13] It is typical here to use the value $t = 50$ *ms* when considering speech sources and applications, and consider $t = 80$ *ms* for musical instrument sources and performance applications.

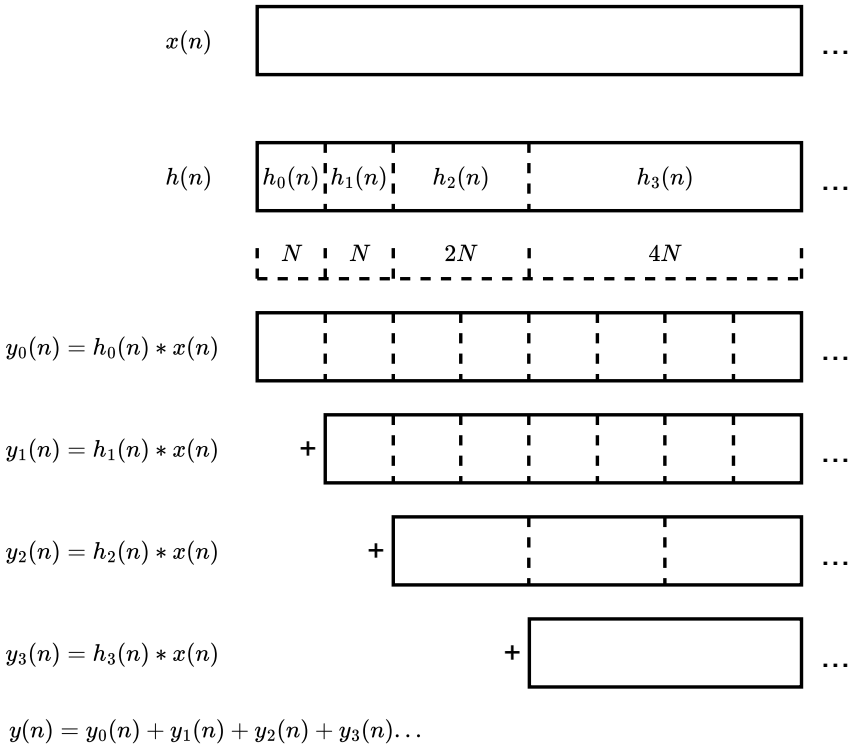


Figure 3.4: Illustration of non–uniform partitioning in convolution, reproduced with reference to [156]. $x(n)$, $h(n)$ and $y(n)$ are the input signal, IR and output respectively. N is the frame size of the first partition.

- Virtual analogue modelling: where reverberation is simulated by simulating the response of mechanical or electrical hardware devices traditionally used to apply reverberation.

With respect to real–time immersive audio production tools used in INMP [65], [132], delay networks, convolution, and specifically the Image–Source Method (ISM) of computational acoustics are relevant. A range of other acoustic simulation methods exist [151], [154], [155], such as geometric acoustics, wave–based models, and particle based and volumetric acoustic modelling. These methods may be found in acoustic simulation software (subsection 3.3.3). These methods are, however, not relevant to INMP practice (given the lack of real–time production solutions in use), and, therefore, not relevant to this thesis, and are not discussed here.

For auralisation of a measured or simulated acoustic response, rendering is achieved through convolution of audio with an IR. In evaluation of artificial reverberation IR convolution may be considered as reference quality [151]. For real–time rendering, partitioned Fast Fourier Transform (FFT)–based convolution methods [157] are used, typically overlap–add algorithms [64]. The state–of–the–art in real–time convolution in the context of INMP is non–uniform partitioned convolution [156].

Here the IR is partitioned into blocks of increasing size ($N, 2N, 4N, 8N...$, where N is the frame size of the first partition) for convolution with input audio. Each partitioned IR frame is convolved with the input audio and summed at the convolution output at the relevant partition time (Figure 3.4). In this manner it is possible to allow for small first partition frame sizes, while alleviating the processing load relative to equal–partition algorithms by computing overlaps at larger frame sizes.

Small first partition frame sizes allow for zero–latency operation with equally small input audio frame sizes typically considered in NMP (section 2.4), such that no additional processing latency is incurred in convolution processes (i.e., no internal buffering delay in audio plugin implementation). Implementation of this algorithm includes multi–channel convolution plugins which provide real–time convolution with SIR with negligible latency[14] [65], [132].

Feedback Delay Network (FDN) reverberation methods are implemented as a feedback matrix of delay lines, comb filters, and all–pass filters. Here each delayed output ('tap') approximates an acoustic reflection, and the cumulative output of delay networks can be used to approximate the acoustic response of a room. A basic comb filter (Figure 3.5) is a signal processing block which sums signal with the input signal via a recursive delay line, causing constructive and destructive interference in the output signal (i.e., comb effect). The recursive delay line is controlled via the delay time, M, and gain applied to the delay line, g. The comb filter is characterised by the transfer function, $H(z)$, [155]:

$$H(z) = \frac{z^{-M}}{1 - gz^{-M}} \tag{3.8}$$

A basic all–pass filter (Figure 3.6) is a signal processing block which amends the feedback comb filter with a feed–forward comb filter path to be summed into the comb filter output. Interference between the two comb filters results in an output with flat frequency response (i.e., 'colourless') while allowing for phase adjustment. The all–pass filter is characterised by the transfer function, $H(z)$, [155]:

$$H(z) = \frac{z^{-M} - g}{1 - gz^{-M}} \tag{3.9}$$

Here g is the filter gain, and M delay time of the delay line. Delay line reverberation designs may also implement low–pass filtering on delay lines to simulate the effect of air absorption [158].

With respect to INMP systems, a range of modern delay network designs are available [17], [25], [159], including implementations as real–time plugins [160], [161]. Usefully such delay network solutions allow for parametric control over simulated reverberation through control of filter gains and delays, and do not require measured or simulated IR resources as in convolution methods.

[14] Though it can be noted that delay may be incurred through convolution with *'dead air'* (i.e., zeros) captured in IR [65]. IR used in real–time convolution should therefore generally be truncated to direct sound onset to remove any delay encoded in IR this way.

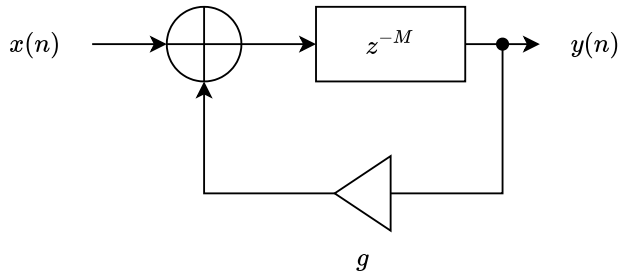


Figure 3.5: Basic comb filter. $x(n)$ is the input, $y(n)$ is the output, M is the delay, and g is the gain.

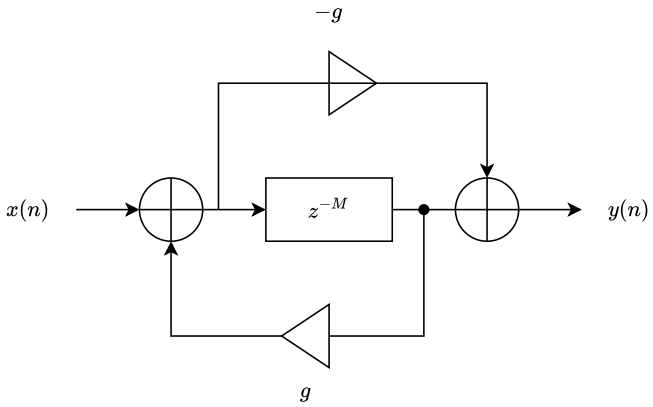


Figure 3.6: Basic all–pass filter. $x(n)$ is the input, $y(n)$ is the output, M is the delay, and g is the gain.

The ISM [154] is a computational acoustics method where a discrete boundary reflection in an acoustic space is modelled as a discrete mirror image sound source of the signal path incoming to the boundary (i.e., an 'image source'). Higher order boundary reflections may then be expressed as further image sources. The acoustic response of a space at a listener position may, therefore, be expressed as a sum of image sources arriving at a listener position. With respect to INMP the ISM (Figure 3.7) is implemented in several real–time plugins, including SPARTA[15] and IEM[16] room simulation plugins. It can be noted, however, that the ISM becomes more computationally demanding as the number of reflections and signal paths modelled increases [162]. With respect to INMP, ISM room simulation plugins typically incur additional delay due to internal process buffering [65].

[15] `https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/`
[16] `https://plugins.iem.at/docs/plugindescriptions/`

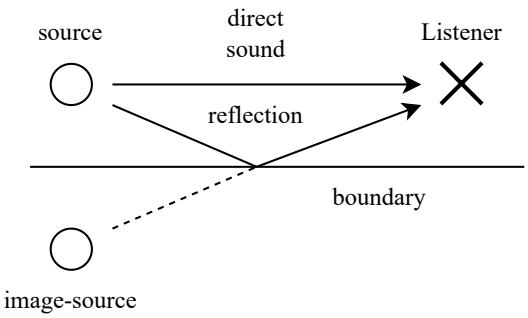


Figure 3.7: Simple illustration of the ISM for room acoustics. The source and listener position are shown with direct sound signal path. A boundary, first–order reflection and image–source are detailed. This figure was created with reference to illustration in [162].

3.3.5 Interactive Auralisation

Auralisation in INMP systems considers varying degrees of interactivity in immersive audio rendering. The most simple case is auralisation of a fixed perspective, where DSP is not responsive to listener movements. Alternative to fixed–perspective rendering, immersive audio may be rendered with 3 Degrees of Freedom (3DoF) (i.e., such that immersive audio DSP is responsive to listener head rotation). Listener head rotations here are described as yaw, pitch and roll, describing rotations around z, y and x 3–D spatial axes respectively (Figure 3.8). Here 3DoF head–tracking systems are required. A range of proprietary solutions exist, including wireless designs such as the Waves Nx Head Tracker[17] or wired USB solutions such as the Supperware Head Tracker[18]. Open source designs also exist, such as the nvsonic head tracker[19]. Alternatively experimental head–tracking solutions, such as webcam video–based solutions [134], [163].

Interactive auralisation may be extended to provide immersive audio rendering which is interactive with 6 Degrees of Freedom (6DoF) (such that immersive audio rendering is responsive to both listener head rotations and listener positional movement, typically described in terms of 3–D spatial x, y and z axis, Figure 3.8). Here tracking systems are required which describe both head rotations and positional movement. In practice this tracking most typically comes in the form of on–board motion tracking from Head Mounted Display (HMD) devices used in VR and XR systems [18], [20], [137], and from bespoke motion capture solutions which are typical of virtual production systems [20], [136], [137], such as solutions provided by professional industry such as Optitrack [20] and Vicon[21].

[17] `https://www.waves.com/hardware/nx-head-tracker`
[18] `https://supperware.co.uk/headtracker-overview`
[19] `https://github.com/trsonic/nvsonic-head-tracker`
[20] `https://optitrack.com/software/motive/`
[21] `https://www.vicon.com/`

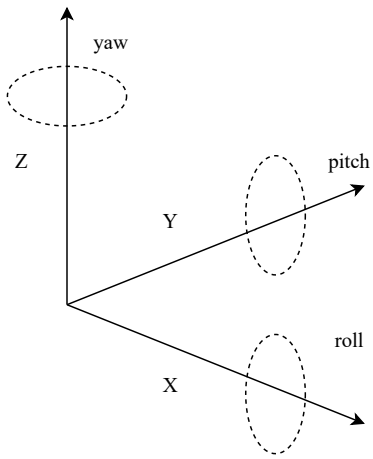


Figure 3.8: Illustration of x, y and z positional axis, and illustration of yaw, pitch and roll rotations, describing rotation around z, y and x axis respectively.

For implementation of interactive immersive audio in INMP, it is typical to consider network transport of tracking metadata associated with 3DoF and 6DoF movement using OSC messages [102]. OSC describes a custom protocol, typically implemented inside UDP payloads, and designed for network transport of parameter control metadata in networked audio systems. For implementation with game engines used in VR and XR system development, various OSC tools exist, such as OSCJack[22] or ExtOSC[23] for Unity, or native OSC plugins in Unreal Engine[24]. Specific to INMP contexts, where real–time interaction between endpoints requires real–time synchronisation, it is required that streams of tracking data in OSC message payloads do not drift due to clock skew between devices. This can occur where tracking data originates from a discrete device from the device hosting NMP software (for example, when taking tracking data is captured by and received from HMDs or motion capture systems). For immersive audio plugins implemented in INMP systems design [65], [132] synchronisation may be approximated by resampling to the most recently received OSC message at a set refresh rate.

3.4 Binaural Stereo

With respect to immersive audio rendering in INMP, headphone–based approaches are considered most practical. Here the requirement of specialist hardware considered in multi–channel loudspeaker array approaches is circumvented, making system implementation logistically simpler [132]. Additionally air propagation delay in musician monitoring of networked performance is reduced to negligible levels, minimising overall system latency in INMP systems.

[22] `https://github.com/keijiro/OscJack`

[23] `https://github.com/Iam1337/extOSC`

[24] `https://dev.epicgames.com/documentation/en-us/unreal-engine/osc-plugin-overview?application_version=4.27`

Finally, monitoring with headphones, rather than loudspeaker monitors, is widely understood to minimise signal spill in sound production [9]. Systems which consider headphone rendering of immersive audio are described in *Binaural Stereo* methods [10]. Binaural Stereo approaches to immersive audio consider rendering of spatial sound in terms of modelling of the human hearing system. The following subsections provide detail on Binaural Stereo methods.

3.4.1 Duplex Theory and the Head Related Transfer Function

Binaural systems consider that reproduction of spatial sound over headphones, where speakers are effectively mounted at the entrance to the ear canal, may be achieved by reproducing the signals from a sound source as it propagates to the ear canal entrance for each ear in the human hearing system. Binaural modelling of the human hearing system may be described with respect to *Duplex Theory* [164] and the *Head Related Transfer Function (HRTF)* [165]. Duplex Theory describes human hearing with respect to inter–aural differences (i.e., differences between signal from a sound source arriving at the ear canal entrance for each ear). HRTFs describe the signal from a sound source arriving at the ear canal entrance for each ear as an acoustic IR, where inter–aural differences are detailed in the measured IR.

Inter–aural differences are surmised in terms of directionally–dependent Inter–aural Time Difference (ITD), Inter–aural Level Difference (ILD), and spectral filtering of sound described in HRTFs. ITD describes the difference in the time at which a signal from a sound source arrives at the ear canal entrance for each ear, resulting from the path distance difference between a sound source and each ear. ITD is the dominant azimuthal localisation cue for low–frequency sounds [166]. ILD describes the difference in amplitude of a signal from a sound source arriving at the ear canal entrance for each ear. This amplitude difference is a result of sound absorption and reflection associated with the physical boundaries of the human hearing system. ILD is the dominant localisation cue for high–frequency sounds. ITD and ILD for given sound source directions may be derived from measured HRTFs, or may be estimated based on simplified models of the human hearing system (for example, the spherical head model [167]). In addition to amplitude and time differences, the difference between a signal at a sound source and a signal at the ear canal entrance for each ear has a unique frequency response for any given source direction. This frequency response is described in measured or simulated HRTFs.

3.4.2 Binaural Recording and HRTF Measurement

Binaural methods consider a two–channel representation of spatial sound, with the first and second audio channels containing the sound to be reproduced at the left and right ear canal entrances respectively.

Native microphone capture of binaural audio is achieved using microphones placed at the ear canal entrances for human subjects [168] or may be recorded using binaural dummy head (for example the Neumann KU100[25]) and head and torso simulator (e.g., Gras Acoustics Kemar[26], Figure 3.9).

Measurement of HRTFs considers IR measurement using binaural microphones, where a discrete HRTF measured pair is associated with a discrete source direction, and sets of HRTFs describe measurement for direction sets, typically describing a 3–D sphere around a receiver position. The number of discrete measurement directions in HRTF datasets is not standardised, and varies from sub–100 to over 1000 discrete HRTF directions [169], [170]. HRTFs are typically measured in anechoic or semi–anechoic conditions. Though HRTF measurement has traditionally been considered a complex process, various efficient methods now exist for measurement of HRTFs for human subjects. Measurement of HRTFs may be achieved using 3–D loudspeaker arrays as sources around a listening position [169]–[171]. Other approaches circumvent specialist hardware requirements, for example HRTF measurement using consumer–grade HMD and a loudspeaker [172]. Alternative to measurement, HRTFs may be generated through Boundary Element Model (BEM) simulation of a human head and torso or head and torso simulator [173], or may be sourced from open source datasets, such as the SADIE II binaural dataset[27] [169] or the meta Sound Sphere 2 dataset[28].

3.4.3 Binaural Auralisation

Auralisation of spatial sound for the binaural stereo format is based on convolution with measured HRTFs. In the simplest case, auralisation of a source at a fixed direction for a fixed head position (i.e., head–locked and non–interactive auralisation), y_{left} and y_{right}, can be achieved through convolution of a sound source signal, $x(n)$, with the HRTF pair for the relevant direction (using spherical coordinate notation as outlined in subsection 3.3.1):

$$y_{left}(n) = x(n) * h_{left}(n, \theta, \phi) \tag{3.10}$$

$$y_{right}(n) = x(n) * h_{right}(n, \theta, \phi) \tag{3.11}$$

Convolution with the HRTF, measured in anechoic conditions, is suitable for spatialisation of sound in the free field (i.e., auralising direct sound only). HRTF measurements performed in 3–D acoustic spaces, and which include spatial detail of reverberation in IR, are referred to as a BRIR. Convolution with a BRIR allows for auralisation of the measured environment, considering a discrete source and receiver positioning associated with a discrete BRIR measurement.

[25] www.neumann.com/en-gb/products/microphones/ku-100

[26] www.grasacoustics.com/products/head-torso-simulators-kemar/kemar-non-configured/product/733-45bb

[27] www.york.ac.uk/sadie-project/database.html

[28] https://github.com/facebookresearch/SS2_HRTF

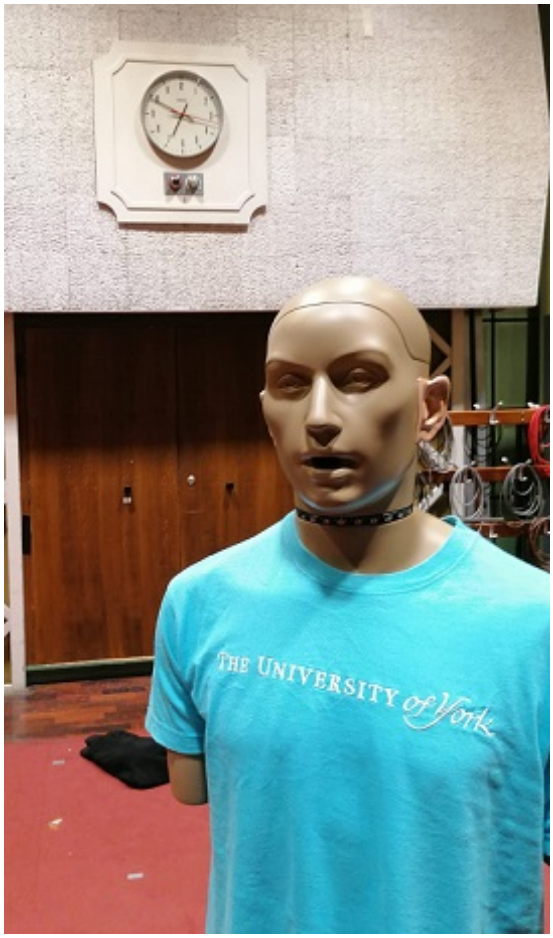


Figure 3.9: Example head and torso simulator used in binaural recording, namely a Gras Acoustics Kemar model.

Interactive auralisation in the binaural format may be achieved by real–time interpolation between HRTF or BRIR measurements. Dense sets of measurements are required for binaural rendering here, therefore requiring interpolation through large numbers of HRTF or BRIR filters. This is acknowledged as a computationally demanding approach for real–time and interactive audio applications, both historically [174], and in modern research [175]. Interpolation through sparse HRTF measurements is considered as a practical approach for real–time applications, however this approach can result in significant error in the output binaural signal [174]–[176]. Modern work here focuses on spherical harmonic interpolation of sparse HRTF sets, working towards the identification of optimal (with respect to signal error minimisation) sparse measurement sets and spherical harmonic orders [175], [176].

Most commonly, virtual loudspeaker approaches to interactive binaural auralisation are considered. Here it is considered that the binaural rendering of loudspeaker arrays may be achieved by convolving loudspeaker feeds with HRTFs for relevant loudspeaker directions, and summing the rendered binaural audio for each loudspeaker.

This allows for immersive audio methods such as Ambisonics [177], Vector–Based Amplitude Panning (VBAP) [178] and other loudspeaker–based 2–D and 3–D panning methods [161] to be rendered to binaural stereo as an end–step auralisation process. Usefully this allows for alternative immersive audio formats to be used as processing formats, such that immersive audio DSP for interactive auralisation may be computed in the alternative format, avoiding HRTF or BRIR interpolation and only requiring a sparse set of static HRTFs to be used in binaural decoding. Prior study has demonstrated that localisation, externalisation [179], [180][29] and reduced front–back reversal[30] are achieved in binaural rendering where individualised HRTFs are used, where 3DoF is provided in rendering, and where ERs are auralised [181].

3.5 Ambisonics

Considering both immersive audio systems for loudspeaker playback and immersive audio processing formats for end–step binaural decoding, an overwhelming prevalence of Ambisonic methods can be observed in practical INMP systems design [65], [132]. Ambisonic methods provide representation of spatial sound through physical modelling of sound pressure in 3–D space, implemented as a multi–channel audio format. The following subsection provide detail on Ambisonic methods.

3.5.1 Spherical Harmonic Representation of Spatial Sound

Ambisonic methods consider a sound field representation of spatial sound, where 3–D sound around a point in space is described in terms of spherical harmonics [27], [182]. Spherical harmonics (Figure 3.11) are a set of directional functions describing spatial sampling of a unit sphere. In Ambisonics spherical harmonics are used to define directionality of spatial sampling of a 3–D field of sound pressure, where sampled directional sound pressure is represented by weighting of spherical harmonic components and where weighting is defined by spherical harmonic coefficients.

Following notation as defined in [139], [182], spherical harmonics are grouped with respect to order, M, where $(M+1)^2$ spherical harmonic components are grouped within a given order. Higher orders therefore correspond with higher numbers of discrete spatial sampling directionalities, providing higher spatial resolution in sound field sampling. Within a given order, discrete spherical harmonic components are denoted with respect to the degree, m, and index, i. Degree denotes subsets of spherical harmonics, where subset m contains spherical harmonic components included in order M, excluding spherical harmonic components included in orders 0 to $M-1$. Within subset m, index, i identifies discrete spherical harmonic components.

[29] A common issue in rendering of binaural audio is for sound sources to be perceived as 'in the head' rather than positioned external to the listener as intended.

[30] Another common issue in binaural audio rendering, where sources rendered in the frontal hemisphere are perceived as localised in the rear hemisphere and vice–versa.

Discrete spherical harmonic coefficients, $Y_m^i(\theta,\phi)$, are calculable [182] for sound pressure sampling in a given direction θ,ϕ:

$$Y_m^i(\theta,\phi) = N_m^{|i|} P_m^{|i|}(\sin\phi) \begin{cases} \cos|i|\theta & \text{where} \quad i \geq 0 \\ \sin|i|\theta & \text{where} \quad i < 0 \end{cases} \tag{3.12}$$

Here $P_m^{|i|}(\sin\phi)$ denotes the associated Legendre functions (where derivation of discrete Legendre polynomials is possible via recursion as detailed in [183]). N denotes a normalisation factor, designed to normalise directional sensitivity of spatial sampling. A range of normalisation schemes exist, however Schmitt Semi–Normalised 3–D (SN3D) is most common, denoted as $N_m^{|i|}$ and calculable as [139]:

$$N_m^{|i|} = \begin{cases} 1 & \text{where} \quad m = 0 \\ \sqrt{2(m-|i|)!/(m+|i|)!} & \text{where} \quad m > 0 \end{cases} \tag{3.13}$$

3.5.2 Ambisonic Format, Recording and SIR Measurement

Ambisonic audio is stored as a multichannel audio signal, where each channel represents a discrete spherical harmonic component (i.e., spatial sampling of a sound field in a discrete direction), and where spherical harmonic components cumulatively describe the sampled 3–D sound field. Channel ordering is typically implemented following the Ambisonics Exchangable (AmbiX) standard[31] [184].

Ambisonic signal may be recorded using Ambisonic microphones [185], [186]. Examples of Ambisonic microphones include the First Order Ambisonic (FOA) Soundfield B–Format microphones[32] (where *'B-Format'* refers to Ambisonic signal derived from FOA microphone recording [26], [185]), and the 2nd Order Ambisonic (2OA) Core Sound OctoMic[33]. For recording of Higher Order Ambisonic (HOA), professional examples include the MH Acoustics EM64 Eigenmike[34] which can provide up to 6th Order Ambisonic capture (Figure 3.10).

IR measurement with Ambisonic microphones allows for measurement of SIR, detailing the acoustic sound field response for the measured source–receiver positions, or where measurement sets describe source–receiver positions throughout a space [187]. Alternatively SIR may be simulated, or sourced from online datasets (for example the Open AIR Ambisonic Impulse Response Library[35]).

[31] It can be noted that legacy formats, namely Furse–Malham, Fu–Ma, channel definitions may be found in practice [182].

[32] `https://www.soundfield.com/`

[33] `https://www.core-sound.com/products/octomic`

[34] `https://eigenmike.com/eigenmike-64`

[35] `https://www.openair.hosted.york.ac.uk/`

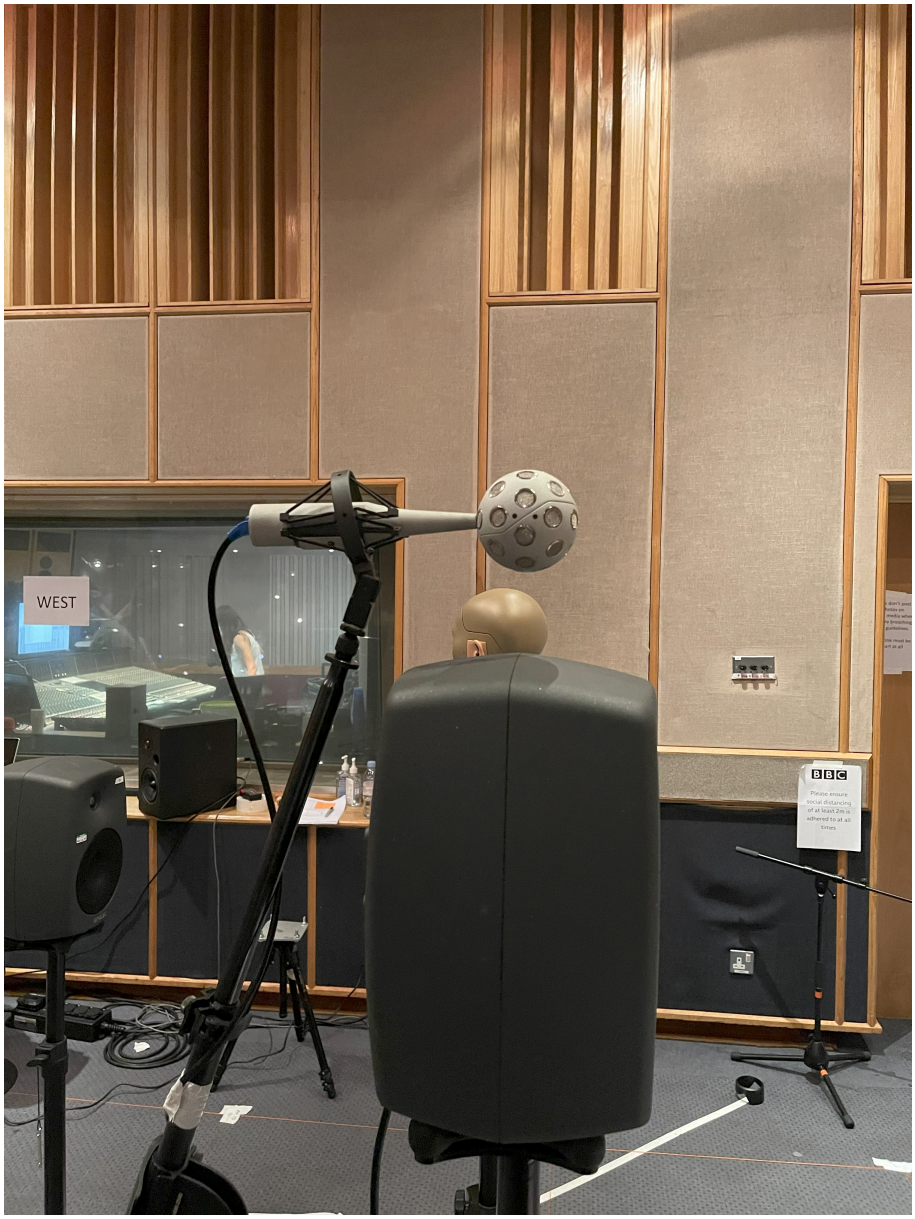


Figure 3.10: MH Acoustics EM32 Eigenmike, here illustrated in–use for SIR measurement.

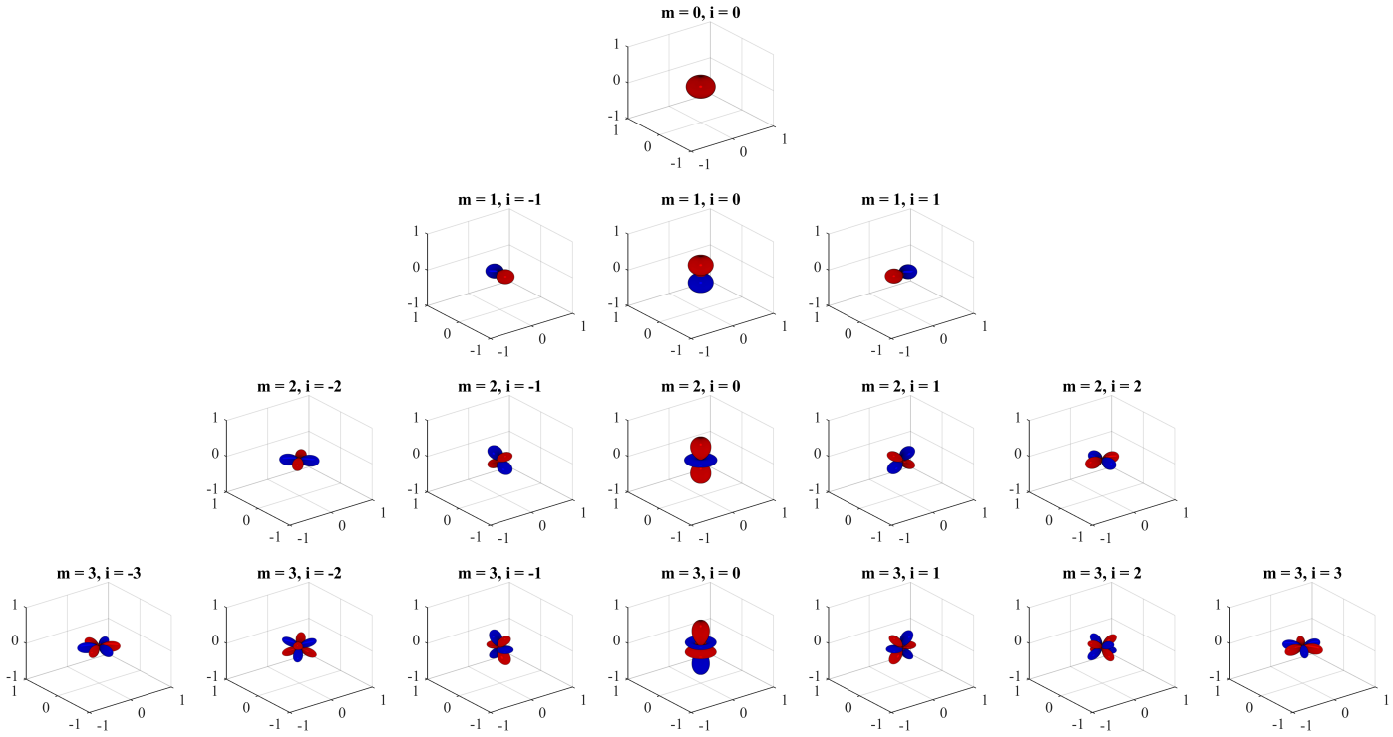


Figure 3.11: Spherical harmonics up to 3rd order. m and i are the degree and index of each spherical harmonic, as detailed in subsection 3.5.1.

3.5.3 Ambisonic Auralisation

Free–field auralisation in the Ambisonic format can be achieved using Ambisonic encoding methods. Here an Ambisonic signal, β, describing auralisation of a source at a set direction, is achieved by applying spherical harmonic coefficients for the relevant direction and order, $Y_m^i(\theta,\phi)$, as a gain [27]:

$$\beta = s(t)Y_m^i(\theta,\phi) \tag{3.14}$$

where $Y_m^i(\theta,\phi)$ are the spherical harmonics as previously defined (Ambisonic channel definitions used in encoding up to 7$^{\text{th}}$ order may also be found noted explicitly online[36]).

Acoustic auralisation of spatial sound in the Ambisonic format can be achieved via convolution with measured or simulated SIRs, providing simulation of a 3–D acoustic sound field for the relevant source and receiver positions in the auralised space [187]. Alternatively a range of parametric [17], [25], [160], [161] implementations exist for rendering ERs and diffuse reverb in the Ambisonic format. Here Ambisonic FDN reverberator designs [159] are implemented, or systems are implemented using FDN designs for diffuse reverb, and allowing precise control of ERs using feed–forward designs [25]. Geometric acoustic rendering is also possible using 3–D image source designs implemented for the Ambisonic format [159].

For interactive auralisation Ambisonic methods usefully allows rendering with 3DoF as a gain operation. Here a rotated Ambisonic signal β_{Rot} may be computed by multiplication of the original Ambisonic signal β with an Ambisonic rotation matrix defined with respect to transforms associated with yaw, $T_z(\theta)$, pitch, $T_x(\phi)$ and roll, $T_y(\gamma)$ angles:

$$\beta_{Rot} = \beta.T_z(\theta).T_x(\phi).T_y(\gamma) \tag{3.15}$$

Derivation of yaw, pitch and roll rotation transforms is detailed in [27], [183], and can be found noted explicitly online up to 21$^{\text{st}}$ order[37]. Ambisonic rendering with 6DoF can be achieved through dynamic parameter control of parametric and image source Ambisonic reverberators. Considering SIR convolution–based approaches, it is also possible to provide 6DoF Ambisonic auralisation by interpolating through discrete measurements in SIR datasets [188]–[190]. Alternatively some systems consider arrays of Ambisonic microphones, where real–time interpolation between live microphone signals allows for 6DoF auralisation [191].

With respect to practical real–time implementation in INMP systems, Ambisonic encoding, rotation, FDN and image source methods are all implemented in popular Ambisonic processing plugins which are capable of operation with minimal additional application processing latency [65], [132].

[36] `http://www.angelofarina.it/Aurora/HOA_explicit_formulas.htm`

[37] `https://iaem.at/ambisonics/xchange/fileformat/docs/spherical-harmonics-rotation`

3.5.4 Loudspeaker Decoding

The sound field described by an Ambisonic signal may be reproduced over 3–D loudspeaker arrays [182]. Decoding of an Ambisonic signal, β to a set of L loudspeakers is achieved as a gain operation, where a matrix of gains (i.e., decoding matrix, D) is used to provide weighting of each of K spherical harmonic components (Ambisonic channels) to the signal $s(t)$ for playback over an array of L loudspeakers. Here the loudspeaker array signal, $s(t)$ is a vector of L individual loudspeaker signals:

$$s(t) = \begin{bmatrix} s_1(t) \\ ... \\ s_L(t) \end{bmatrix} \tag{3.16}$$

The Ambisonic signal, β is a vector of K individual Ambisonic channels:

$$\beta = \begin{bmatrix} \beta_1 \\ ... \\ \beta_K \end{bmatrix} \tag{3.17}$$

For standard Ambisonic loudspeaker layouts the typical loudspeaker decoding method is *mode–matching* [182], [192]. Here the decoding matrix, D is derived from the *re–encoding matrix*, C, defining spherical harmonic weighting for each loudspeaker for discrete loudspeaker directions in spherical coordinate notation:

$$C = \begin{bmatrix} Y_1(\theta_1,\phi_1) & Y_1(\theta_l,\phi_l) & ... & Y_1(\theta_L,\phi_L) \\ Y_k(\theta_1,\phi_1) & Y_k(\theta_l,\phi_l) & ... & Y_k(\theta_L,\phi_L) \\ ... & ... & ... & ... \\ Y_K(\theta_1,\phi_1) & Y_K(\theta_l,\phi_l) & ... & Y_K(\theta_L,\phi_L) \end{bmatrix} \tag{3.18}$$

The decoding matrix, D, is then calculated as the inverse of C where $L = K$:

$$D^{inv} = C^{-1} \tag{3.19}$$

Alternatively, where $L > K$, the pseudo–inverse is used:

$$D^{pinv} = pinv(C) \tag{3.20}$$

Loudspeaker array configurations are generally required to fulfill Ambisonic loudspeaker array regularity requirements to provide accurate sound field reproduction. Requirements of loudspeaker array regularity are detailed in [182], where standard array configurations are described by the *platonic solids* and *t–designs*.

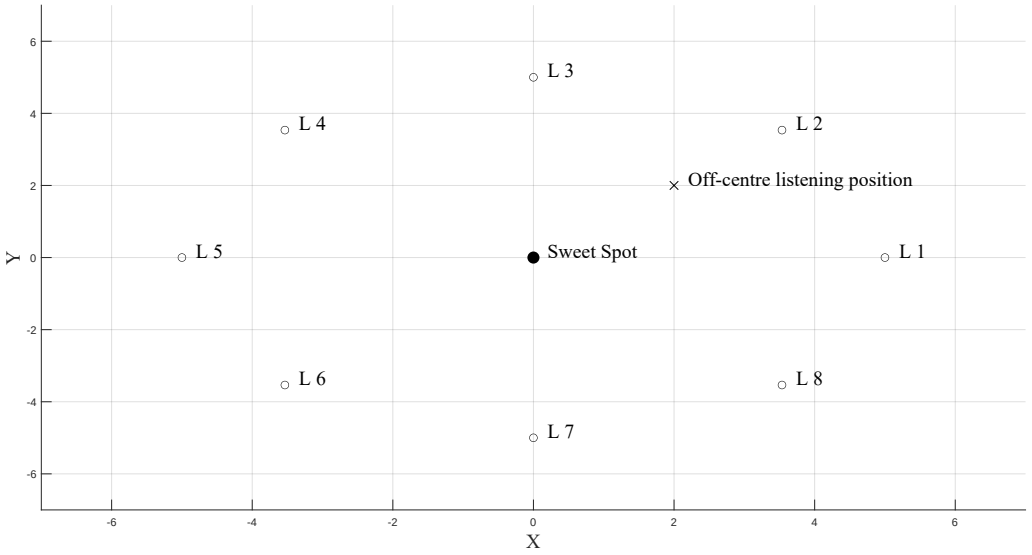


Figure 3.12: Illustration of a 2–D Ambisonic loudspeaker array setup, namely the 8–channel 3OA array from the SADIE II database. The sweet spot, and an example of off–centre listening position are noted.

All Round Ambisonic Decoding (AllRAD) decoding approaches may be used where irregular loudspeaker arrays are considered [193]. Here VBAP methods are used to position phantom sources describing an optimal regular layout to which an Ambisonic signal may be decoded. Though not required in basic Ambisonic loudspeaker decoding, additional weighting may be applied in decoding, designed to optimise aspects of loudspeaker reproduction, such as high frequency response or phase [182]. Here dual–band approaches are practical, which provide basic decoding at low frequencies, and $max - r_E$ weighting for high frequencies, with order–dependent crossover as detailed in [182].

Loudspeaker reproduction of Ambisonics typically considers sound field reconstruction at a *'sweet spot'*, around which the array is centred (Figure 3.12). For lower–order Ambisonics (i.e., FOA and 2OA systems) the experience of immersive audio over Ambisonic loudspeaker arrays deteriorates when listening off–centre (i.e., displaced from the *'sweet spot'*) [194]. For HOA loudspeaker arrays, however, off–centre listening may not deteriorate the experience of sound field reproduction [195], [196], particularly across large arrays used for venues, where a loudspeaker coverage area in which a sizable audience may be situated must be considered [197].

3.5.5 Binaural Decoding

Binaural rendering of Ambisonics has traditionally been achieved using the virtual loudspeaker approach as outlined in subsection 3.4.3. Here each discrete decoded loudspeaker signal is convolved with a HRTF pair for the given loudspeaker direction, and convolution outputs are summed to provide binaural rendering of the loudspeaker array [177].

Alternatively to direct virtual loudspeaker approaches, HRTFs used in virtual loudspeaker rendering may be encoded in the spherical harmonic domain [182]. Practically, this is where a transposed decoding matrix, D^T is multiplied with HRTFs, such that loudspeaker decoding is embedded in the resultant IRs. Notably the number of resultant IRs is equal to the number of Ambisonic channels, rather than number of loudspeakers. Given that $L > K$ is typical of Ambisonic loudspeaker designs, this generally reduces the number of channels of convolution required in binaural decoding, alleviating processing requirements in real–time implementation.

Magnitude–Least Squares (Mag–LS) binaural decoding methods are the focus of modern implementation [198]. Here a set of HRTFs encoded in the spherical harmonic domain are considered. A transposed decoding matrix, to be multiplied with HRTFs to provide binaural decoding filters, is found through least–squares error minimisation of magnitude in reconstructed HRTFs, where HRTF phase is discarded above an Ambisonic order–dependent crossover frequency. With respect to practice Mag–LS binaural decoding demonstrably outperforms virtual loudspeaker approaches [139]. Considering INMP system design, Mag–LS binaural decoding methods are implemented in popular immersive audio plugins, and are capable of operation with minimal additional application processing latency [65], [132].

3.6 INMP Audio System Design

Previous work in the field of INMP provides a practice–oriented overview of applications, and common methods in modern systems. In addition to a range of practical systems for virtual concert applications, previous INMP literature focuses on the design and estimated performance of low–latency WAN systems which are suitable for practical applications. Here particular focus is given to the design of low–latency immersive audio rendering using Ambisonic methods as a processing format, and considering both loudspeaker and binaural playback. The following subsections provide a review of immersive audio methods in INMP system design.

3.6.1 Spatial Audio Microphone Systems

Early developments in INMP and live–streaming focused on spatial sound capture using spatial audio microphone arrays. With respect to live–streaming, Ambisonic microphones [31] and other spatial microphone arrays [33] were used in distribution of concerts to remote audiences listening on 2–D and 3–D loudspeaker arrays, typically matching perspective presented in video display. Similar designs are used in musician monitoring of early INMP trials over internal university networks [33], [34], [96]. Examples also exist of usage of head and torso simulators for native binaural capture and remote playback for musician monitoring for WAN deployment of INMP [29]. Microphone–based approaches to INMP continue to see usage for musician monitoring in modern practice. Ambisonic microphones are considered for musician monitoring using popular NMP solutions such as SoundJack [15].

Alternative spatial audio microphone arrays (such as coincident arrays for the Mach–1[38] format) are also used in modern experimental INMP systems where pro–LAN solutions are deployed over WAN connections [135]. Spherical microphone arrays are also considered in INMP system design for experimental virtual performance and concert infrastructures [20], considering both musician monitoring using IEMs and audience distribution in live–streaming, and in loudspeaker array playback in venues. Here matching of perspective delivered in XR visual display is considered in microphone capture.

Microphone–based INMP design generally requires minimal DSP, achieving spatial sound description through direct capture rather than scene synthesis (typically only format conversion between discrete microphone capsule signals and a relevant immersive audio format, if necessary, implemented with negligible latency native to microphone systems). Practically for INMP this means negligible audio processing latency is associated with the approach, minimising overall system latency. Spatial audio microphone capture is also particularly relevant to modern systems for the capture and production of performance content for immersive media, where INMP has been proposed as a practical remote workflow solution for content creation, and real–time distribution of content as a virtual concert. Examples include use of HOA spherical microphone array capture in 360 °video production [24]. In delivery of performance content as VR media, distribution of multiple spherical microphone arrays allows for simulation of 6DoF movement throughout a space, interpolating through individual spherical microphone array signals in response to listener position [191]. Where such vast arrays are not considered, however, microphone–based capture of spatial audio with a single microphone system provides description of a fixed position in space (and in the case of binaural microphones, fixed head pose also).

3.6.2 Free–Field Auralisation

With respect to immersive audio DSP for networked performance, spatialisation of direct sound (i.e., free–field auralisation) is most typically achieved in the Ambisonic format using encoding methods outlined in subsection 3.5.3. Generally this is achieved using spatial audio plugins implemented in a host application (where Max MSP and Reaper are popular solutions). Ambisonic encoding plugins from the AmbiX plugin suite[39] [199] are considered in various INMP designs and practical deployments [22], [65], [132]. Alternatively solutions such as ICST Ambisonics[40] [132], IEM[41] [121], [132], and SPARTA[42] [132], [160] plugin suites are practically utilised. Further alternative solutions such as Blue Ripple HOA plugins[43], or legacy solutions such as the FaceBook Spatial Workstation[44] [65].

[38] https://www.mach1.tech/
[39] www.github.com/kronihias/ambix
[40] https://github.com/schweizerweb/icst-ambisonics-plugins
[41] https://plugins.iem.at/
[42] https://leomccormack.github.io/sparta-site/
[43] https://www.blueripplesound.com/o3a_core
[44] http://www.angelofarina.it/Public/Spatial_Workstation/

In practice free–field auralisation for INMP is utilised in different remote concert performance systems for both musician monitoring and playback of performance to remote audiences over 3–D loudspeaker arrays. Notably here electroacoustic performance material is considered (making high–latency conditions tolerable), and production of immersive audio considers artistic licence in real–time production, rather than matching of perspective presented in visual display. Free–field auralisation is also implemented in processing for systems for auralising networked performance for video live–streaming with binaural headphone playback [133]. With respect to low–latency INMP design, it can be noted that popular solutions to Ambisonic encoding typically incur negligable additional application processing latency [65], [132].

3.6.3 Acoustic Auralisation

Simulation of 3–D acoustic spaces in INMP audio DSP is again most typically achieved using spatial audio plugins implemented in a host application. As with free–field modelling, acoustic auralisation is most typically achieved in the Ambisonic format, here using a range of techniques. In practical low–latency WAN INMP deployment [7], [8], and in modern systems for musician monitoring and audience distribution of networked performance with XR display [20], [137] convolution with measured Ambisonic SIR is implemented to auralise networked performance. For convolution, use of MCFX Convolver from the AmbiX plugin suite [200] is noted in literature [7]. X–MCFX Convolver[45], a modified version of the vanilla MCFX Convolver is considered as a practical solution for multichannel convolution [65], [132]. Direct convolution with BRIR is also implemented in some cases, where head–locked headphone playback is considered [16]. Both MCFX convolver, and X–MCFX convolver, allow for zero–latency convolution (subsection 3.3.4) using non–uniform partition algorithms (such that no additional latency is incurred internally in plugin processing) [65].

In the majority of implementations, however, auralisation is achieved using parametric systems [132]. Here rendering of direct sound is typically achieved using Ambisonic encoding methods as discussed in subsection 3.6.2. Auralisation of ERs and/or diffuse reverb is achieved using FDN reverberator designs. In practical INMP implementation IEM FDNReverb plugin is utilised here [132]. Notably simplified auralisation systems may be considered here, where modelling of ERs is not considered, instead simply auralising direct sound and diffuse reverb. Ambisonic image source reverberators are also used in INMP system design [65], [132], [136]. Here particularly IEM RoomEncoder and Sparta AmbiRoomSim plugin options are considered. Practical INMP implementations may also use combinations of methods. For example image source plugins may be used to auralise direct sound and early reflections in HOA format, whereas diffuse reverb may be rendered via convolution with 0^{th} order Ambisonic IR [136].

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

Alternative approaches are considered here other than usage of Ambisonic format and production tools. Notably, usage of Spat[46] [161] is considered in several practical INMP implementations and designs [23], [132], [134]. Here direct sound may be spatialised using a range of spatial panning methods, with parametric reverberation using FDN methods [161].

Parametric rendering approaches to acoustic auralisation for INMP generally consider interactive binaural rendering of networked performance. Practical implementation considers head–locked audio used in musician monitoring and audience distribution in live–streaming with video display [23], [132], and rendering with 3DoF using webcam–based tracking systems [134]. As with practical implementation of free–field rendering in virtual concert systems, electroacoustic and ambient performance material is considered here. Theoretical design of low–latency parametric rendering systems for INMP [65], [132] and practical implementation of image source rendering designs for low–latency INMP on LANs [136] primarily considers rendering for musician monitoring with VR and XR systems with 3DoF or 6DoF.

With respect to low–latency INMP systems design [65], [132], it can be noted that convolution–based approaches may be achieved with negligible latency. With respect to parametric approaches, direct sound may typically be rendered with negligible latency. Though latency may be incurred in rendering of reverberation, as direct sound is rendered with negligible latency, any rendering delay in reverb auralisation relative to this effectively acts as a pre–delay. Considering image source reverberation, it can be considered that in practice [65], [132], [136] several milliseconds of delay will be incurred in audio plugin throughput.

3.6.4 Algorithmic Auralisation

In INMP designs for musician monitoring [18] and distribution to audiences using VR and XR systems [18], [20], [137] algorithmic immersive audio rendering is considered. Here immersive audio rendering is implemented in game engines [20], [137] or web development frameworks [18] using immersive audio Software Development Kit (SDK) solutions such as Resonance Audio[47], typically considering distribution to musicians and audience members using HMDs. Though algorithmic rendering solves the issue of providing interactive rendering for large groups, where real–time sound production of individual mixes is impractical, such solutions are acknowledged as incurring high audio processing latency when implemented for INMP musician monitoring, and electroacoustic and ambient material is again performed to manage high–latency conditions in practice [18].

3.6.5 Loudspeaker Rendering

Practical implementation of loudspeaker playback of immersive audio in INMP generally considers the Ambisonic format where methods are detailed, with decoding methods as detailed in subsection 3.5.4. Practical implementation in musician monitoring contexts notes use of the Kronlachner Ambix plugins [22].

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

[47] `https://resonance-audio.github.io/resonance-audio/`

For loudspeaker playback in audience distribution Kronlachner Ambix plugins are also considered [22]. Alternatively bespoke Ambisonic encoding and decoding is implemented directly in Max MSP in other work [21]. In other INMP audience distribution cases, playback of Ambisonic microphone capture is considered, matching video [31] or XR [20] stage display. In practice, playback on a range of loudspeaker array types is considered. Included here are standard surround–sound loudspeaker configurations such as THX 10.2 [34], [96] or Dolby 5.1 [31] and 5.1.4 systems [20]. Alternatively, 3–D Ambisonic loudspeaker arrays may be used [21], [22].

Considering INMP musician monitoring of loudspeaker playback, it can be considered that it is typical for spatial audio loudspeaker arrays to be far–field (for example in literature describing Ambisonic arrays speakers are placed at radial distances above 2 m [195], with larger radial distances considered in practice, particularly for venue systems [197]. Considering air propagation latency in NMP design, it can be noted that loudspeaker array monitoring will result in several milliseconds of additional latency, and may not be suitable for low–latency INMP design. Indeed modern low–latency INMP system development rather focuses on headphone–based systems for musician monitoring [65], [131], [132].

With respect to distribution of networked performance as concerts to remote audiences, it can be noted that the Ambisonic format is convenient, being possible to decode to a range of loudspeaker arrays [197]. For large multichannel arrays, playback of Ambisonics provides effective results for off–centre listening, allowing for an audience to be positioned throughout a coverage area. For FOA systems and sparse 2–D arrays, however, (for example 5.1 systems) it can be noted that perception of immersive audio deteriorates for off–centre listening [195], [196]. It may be considered for such cases that other loudspeaker spatialisation methods may be more appropriate, where Spat [161] allows for several alternative methods (for example VBAP).

With respect to interactivity in immersive audio rendering, it can be noted that loudspeaker rendering in INMP considers reproduction of immersive audio with a fixed positional perspective. All listeners in a loudspeaker coverage area are rendered with the same fixed perspective, where 3DoF in listening is achieved through natural head rotations in the reproduced sound field. In this regard loudspeaker systems are limited in that sound sources may only be positioned external to the enclosure of loudspeaker arrays. Here it can be considered that such systems are most suited to matching of perspective presented in stage displays, as considered in practical INMP implementation.

3.6.6 Binaural Rendering

Binaural rendering in INMP design generally considers interactive immersive audio rendering. Here, the Ambisonic format is typically used as a processing format for auralisation and computation of 3DoF and 6DoF in rendering, typically implemented as audio plugins in a host application [65], [132]. 3DoF is typically computed via Ambisonic rotation matrices, and 6DoF rendered using parametric methods or image source methods.

In practical INMP implementation, interactive binaural rendering is provided with 3DoF in musician monitoring with video display and a webcam head–tracking system, here using parametric rendering methods (namely Spat, implemented in Max MSP) [134]. In LAN–based INMP test system implementation, interactive binaural rendering with 6DoF is provided, using dynamic control of an Ambisonic image source reverberator (namely IEM Room Encoder[48]) implemented in Max MSP, with 6DoF tracking from a motion capture system [136].

In the design of low–latency INMP solutions, several interactive binaural systems are also presented [65], [132]. Here systems consider tracking from mounted head–trackers, HMDs and motion capture systems. Ambisonic encoding is used for free–field rendering, and FDN reverberation, providing simplified 6DoF auralisation. Auralisation through SIR convolution is also considered, rendered with 3DoF using Ambisonic rotation methods. Alternatively image source auralisation is also considered for 6DoF implementation. Notably here auralisation is possible to achieve with minimal additional audio rendering latency, and where Ambisonic auralisation and rotation can be implemented in host applications using immersive audio plugins (again considering primarily the Kronlachner Ambix, IEM, and Sparta plugin suites, and usage of Spat). For convolution, MCFX Convolver and X–MCFX convolver are considered.

Binaural decoding in the design of low–latency INMP systems again considered implementation as plugins. Here both virtual loudspeaker and Mag–LS approaches are considered. Virtual loudspeaker methods may be achieved using the Kronlachner Ambix plugin suites [7], [65], [132], whereas Mag–LS decoding is achieved using IEM and Sparta binaural decoder plugins [65], [132]. Minor audio processing latency is expected to be incurred in binaural decoding from previous literature [65], [132]. It can be considered, however, that binaural decoding of Ambisonics is a multichannel convolution operation, and can be computed with negligable latency using multichannel convolver plugins such as X–MCFX. Low–latency interactive binaural INMP system design largely focuses on systems which are suitable for usage with immersive technologies such as VR and XR [131], considering musician monitoring systems. Practical implementation of binaural rendering systems also include distribution of performance to audiences using HMDs in virtual concert applications [18], [20], [137].

Head–locked binaural rendering is also implemented in practical INMP design, here used in musician monitoring and audience distribution considering acoustic auralisation [23] and free–field auralisation in the Ambisonic format [133]. Here typically concert performance applications are considered, where head–locked binaural playback is suitable for live–stream audiences, enhancing video display where tracking solutions are not expected to be available.

[48] `https://plugins.iem.at/docs/plugindescriptions`

3.6.7 Streaming Format and Network Architecture

Considering several cases of INMP system implementation for musician monitoring with spatial microphone capture [15], [20], [33], [135], and one implementation of immersive audio auralisation with DSP [21], audio is streamed between sites in spatial audio formats, where audio either describes microphone capture, or a scene created from summing mixing. Such format–based multichannel streaming is also practical in audience distribution, where scenes created from summing mixing and spatial microphone capture are also typically considered [20]–[23], [137]. The Ambisonic format can be noted as particularly useful in this case. Considering remote loudspeaker playback, Ambisonics may be decoded for playback both as binaural stereo on headphones, and for arbitrary loudspeaker array configuration [197]. Additionally transport of multichannel immersive audio formats may incur significant bandwidth requirement due to high channel counts. Here it can be noted that the Ambisonic format is compatible with low–latency compression using the Opus codec [92], [93] which is implemented in several modern NMP solutions. The design of low–latency INMP systems, however, typically considers object–based streaming between sites, as peer–to–peer design is required for low–latency operation, and the bandwidth cost of multiple peer streams for multichannel immersive audio formats such as Ambisonics can be considered excessive[49].

With respect to network architecture in INMP design, practical implementation largely considers client–server architectures, where servers are used for rendering of musician monitor mixes [21], [23], [134] and for rendering of audio for audience distribution [20], [23], [133], [137]. Such server–based designs can be considered to simplify audio production, allowing for all audio production to occur at a single site, and making practical operation of INMP systems logistically simpler for audio engineers, and alleviating processing requirements for clients.

Additionally, where particularly large networked ensembles are considered, server–based architectures may significantly reduce bandwidth requirements, where each client only needs to receive a scene–based render created from summing mixing of networked performers, rather than large sets of object–based streams considered in peer–to–peer design. Low–latency INMP design, however, explicitly considers peer–to–peer architectures, as such systems typically achieve lower network latency than client–server architectures.

[49] For example, the bandwidth cost of sound source description as a single channel of audio (48 kHz, 16–bit) in a peer–to–peer system requires approximately 768 kbps of bandwidth per–peer [62]. A 3OA signal, which may describe one or more sound sources, here is approximately 12.288 Mbps of bandwidth requirement per–peer. In this example it becomes bandwidth–efficient to use the 3OA format in the case that there are 17 or more sources at each endpoint. It can also be recognised that in deployment of systems for multiple peers, scaling of 12.288 Mbps per–peer may easily result in bandwidth requirements which are not met by many practical WAN endpoints

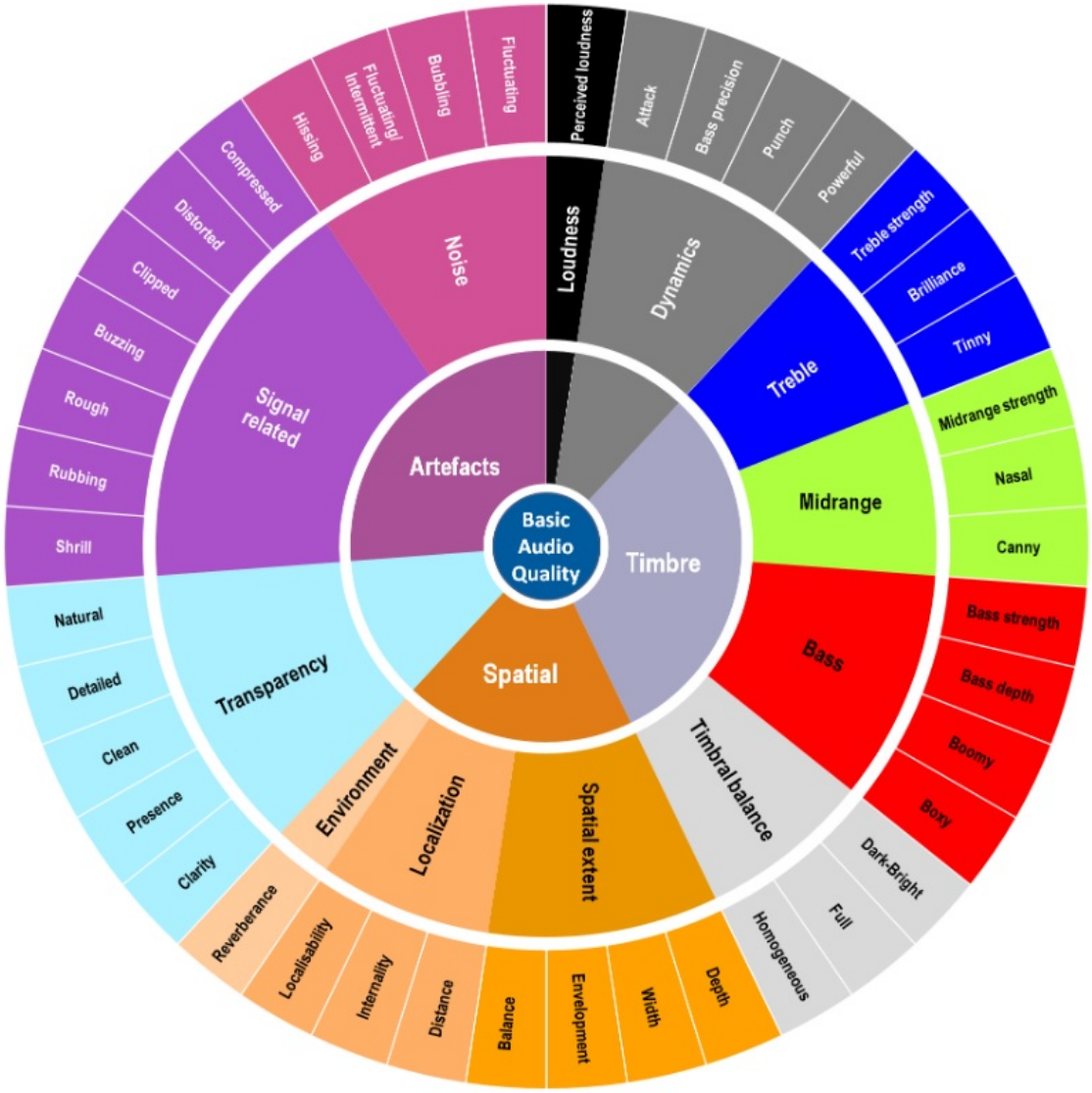


Figure 3.13: Breakdown of factors of audio quality considered in relevant standards, where immersive audio rating items considered in INMP evaluation are included, from [180].

3.7 Evaluation of INMP

The evaluation of INMP systems expands upon evaluation methods considered standard in evaluation of traditional NMP systems, namely synchrony analysis and QoE rating [6]. Rating items relating to presence [124] and immersive experience modelling [125], also common in evaluation of standard NMP systems, are given particular attention in INMP evaluation. Building upon these existing methods, INMP evaluation introduces rating items relating to spatial factors of BAQ [180] and the perception of immersive audio [201], [202] (Figure 3.13). The following subsections provide a review of evaluation methods for INMP systems.

3.7.1 Spatialisation Metrics

In several studies which evaluate immersive audio rendering systems for musician monitoring in INMP contexts rating items are used which describe either the ability localise sound sources, or to distinguish between discrete sound sources. Items which address the ability to localise sound are most relevant to the standard immersive audio rating item of *localisation*, considered an important aspect of evaluation of immersive audio systems [202], and where definition in audio system evaluation is given in relevant standards as *"The degree of precision to which the position and extent of a source or ensemble can be identified"*, with a suggested rating scale of imprecise–precise [180]. Rating of localisation in INMP evaluation includes:

- Rating of the statement *"While playing I was able to localize the other musicians"* on a scale of not accurately–very accurately [121].
- Rating of agreement with the statement *"I was able to localize my musical counterpart"* on a scale of strong disagreement–strong agreement [19].

Items which address the ability to distinguish between discrete sounds[50], again considered important to perception of immersive audio [202], are most relevant to the standard immersive audio rating item of *'spatial clarity'*, where standard definition is given as *"the impression of how clearly different elements in a scene can be spatially distinguished from each other"*, with a suggested rating scale of unclear–clear [180].

Rating of spatial clarity in INMP evaluation includes:

- Rating of the statement *"While playing I could distinguish well my contribution from that of the other musicians"* on a scale of little difference–much difference [121].
- Rating of the statement *"I was able to distinguish between the sounds I initiated and the sounds produced by my musical counterpart"* on a scale of strongly disagree–strongly agree [19].

3.7.2 Acoustic Metrics

Also introduced in studies evaluating immersive audio systems for musician monitoring in INMP are rating items which address perception of auralised acoustics. Here rating is considered relative to quality of interaction with real–time acoustic rendering. This form of rating can be considered related to immersive experience modelling in terms of interaction with system components [125], rather than discrete qualities of perception of acoustics described in immersive audio rating such as envelopment, externalisation, or spaciousness [180], [202].

[50] What audio and music professionals may colloquially refer to as *'separation'*.

Rating of acoustic responsiveness in INMP evaluation includes:

- Rating of agreement (strongly disagree–strongly agree) with the statement *"The virtual acoustic performance space was acoustically responsive to sounds I initiated or performed"* [8].
- Rating of agreement (strongly disagree–strongly agree) with the statement *"The virtual performance space was acoustically responsive to sounds I initiated"* [19].

3.7.3 Comparative Audio System Evaluation

An aspect of evaluation introduced in INMP is the comparative rating of different audio rendering systems. Here studies [16], [110], [119], [121], [136] consider mono or stereo audio as a baseline which is representative of traditional NMP audio systems. In these studies immersive audio systems and baseline systems are evaluated using synchrony analysis as well as QoE, presence and immersion, and immersive audio rating items. Variation in synchrony, or in item rating, between audio systems is considered here as demonstration of improved performance, and of elicitation of perceptual difference. Comparative study is also used in this manner to provide comparison between various immersive audio rendering systems for INMP [19], and in various immersive audio rendering approaches for INMP (i.e., variation of acoustic simulation parameters) [16].

3.7.4 Semi–Structured Interviews

In INMP evaluation, discussion prompts for semi–structured interviews are extended to include items relevant to immersive audio rating. In practice, exit interview discussion follows rating questionnaires and includes discussion of localisation and spatial clarity aspects [13], [121]. With respect to comparative study, interview prompts may directly address perceived preferences in systems under evaluation, and justification for those preferences [121].

3.8 Previous Study of INMP

Prior study of INMP provides system evaluation through previously discussed NMP evaluation metrics. Considering INMP as a solution for immersive applications, particular focus is given to presence and immersive experience rating items, and previously highlighted immersive audio rating items are also included. INMP evaluation focuses on comparative evaluation of immersive audio rendering systems, typically considered relative to baseline audio systems. Another key aspect of prior work is the evaluation of different auralisation methods within INMP systems, here focusing on variation in acoustic simulation complexity, and parametric variation of auralised acoustics. The following subsections provide a review of prior study in the field of INMP.

3.8.1 Comparison to Mono and Stereo Audio

Prior study [121] provides comparative evaluation of the experience of performing with a backing track, considered representative of INMP contexts. Here comparison is provided between baseline stereo conditions and an immersive audio rendering system providing direct sound spatialisation (without reverb) and 3DoF in interactive binaural rendering. Evaluation, provided through QoE, presence and immersion, and immersive audio rating items demonstrates a clear preference for interactive immersive audio rendering in musician monitoring.

In further study [19] similar evaluation is provided, here with the inclusion of video display with mono baseline audio conditions, and projector–based XR visual display for immersive audio conditions, where rendering is provided with 6DoF, and considers both free–field modelling (i.e., direct sound only) and reverberant auralisation. Again a preference for immersive audio rendering is noted across presence, immersion, and immersive audio rating factors. It is considered here, however, that results relate largely to both visual and audio display and interface, and poor rating of baseline conditions may be influenced by usage of video display. Notably baseline conditions appear to result in more stable synchrony than immersive audio monitoring. This, however, may be due to difference in latency between conditions, where baseline conditions are evaluated at a significantly lower latency than immersive audio conditions.

3.8.2 Effect of Virtual Acoustics

Evaluation of virtual acoustics is a topic of particular interest in INMP evaluation. Early INMP research considers the implementation of reverberation effects as having the potential to counteract the impairment of networked musical interactions by latency, supposed to provide smoothing of onsets [58]. In a following study, use of BRIR convolution to provide immersive audio rendering is suggested as improving stability of tempo in performance [110]. Other work, however (notably considering drummers),a finds no variation between baseline conditions and monitoring with *'artificial reverberation'*, instead reporting a preference from musicians for non–reverberant conditions, considered to provide superior clarity in monitoring [119].

Further study here suggests improved synchrony specifically at high (100 ms) latency under reverberant conditions than non–reverberant conditions [203]. In another study considering immersive audio rendering with XR display and 6DoF in interactive binaural rendering, free–field (i.e., direct sound only) and reverberant conditions are compared. Notably conditions where virtual acoustics are included achieve superior rating than free–field auralisation for rating factors relating to presence, immersion and immersive audio perception. No difference, however, is suggested in rating of perceived latency [19].

In addition to comparison to baseline dry mono or stereo audio, previous studies also evaluate variant approaches to virtual acoustics. Prior work comparing head–locked binaural auralisation of different measured acoustic spaces for INMP systems suggests more stable tempi ratio between parts for auralisation of medium and cathedral halls than for auralisation of a studio booth [8].

Other work, considering 6DoF binaural rendering for INMP with XR visual display, shows no significant difference in subjective rating or synchrony analysis between auralisation of a listening room and large hall [19]. Comprehensive work evaluating BRIR auralisation approaches provides comparative rating and synchrony analysis for INMP [16]. Congruent (where remote sources are auralised to sound as if located in the local physical environment) and incongruent (where remote sources are auralised as if heard in a simulated acoustic environment which is different from the local physical environment) are compared, as is symmetric (where each performer is provided auralisation of the same acoustic space) and asymmetric (where each performer is provided auralisation of a discrete acoustic space) auralisation are also compared. Here it is suggested that, though latency dominates experience of presence in networked interactions at high latency (i.e., above 30 ms), congruent auralisation may affect improved sense of presence in networked interactions for low latency conditions.

3.8.3 Effect of Spatialisation

Improvements in rating associated with immersive audio rendering relative to mono or stereo baseline audio rendering included immersive audio rating items associated with spatial clarity and localisation [19], [121]. It is discussed in these studies that spatialisation of sources in immersive audio rendering may improve experience in INMP, making it easier to pay attention to discrete sources, shift attention between sources, and generally direct attention in networked musical interactions (i.e., such that it is easier to interact with discrete parts in performance) [19]. Indeed exit interview responses in other INMP studies report appreciation of the ability to localise discrete sources [13], [121]. Further interview responses report appreciation of source separation in immersive audio rendering relative to limited stereo systems [121].

3.9 Summary

Though prior practical implementation of INMP considers high–latency deployments, review of prior work, particularly modern research which address low–latency INMP system design [7], [8], [65], [132], [136], suitable low–latency immersive audio rendering designs are identifiable. A common approach can be noted in Ambisonic methods used for auralisation, loudspeaker playback, and interactive binaural rendering. Notably typical plugin solutions [7], [65], [132] are considered for implementing immersive audio rendering with minimal additional application processing delay added to the NMP signal path. The Ambisonic format can be noted as particularly useful here, being possible to deploy to arbitrary loudspeaker arrays [197] which may be used in INMP musician monitoring and audience distribution in virtual concert systems, and being compatible with low–latency compression codecs such as Opus which are implemented in modern NMP streaming software where streaming of Ambisonic format audio is considered [92], [93].

Notably, practical low–latency INMP musician monitoring systems, and virtual concert distribution systems, consider implementation of immersive audio rendering in real–time sound production environments such as DAW applications. It can be considered here that though processing methods are derived from modelling of spatial sound, in practice an engineer will have access to audio processing parameters, and will be able to adjust parameters according to their professional opinion and artistic license in real–time sound production of musical content. In this regard it can be acknowledged that INMP audio rendering design may not follow the objective of authentic reproduction or simulation of spatial sound. In practice it is likely that engineers will modify parameters to their own taste.

Practical implementation of INMP in modern applications in virtual concert systems, though considering high–latency contexts, confirms the practicality of INMP system design for use as an enhancement to traditional video display, and as a compatible audio rendering system for use alongside XR display. Evaluation of INMP systems indeed demonstrates that inclusion of immersive audio rendering results in general experience improvements for NMP, typically relating to NMP QoE, presence, and introduced immersive audio experience rating. Here introduced immersive audio rating items contribute further practical evaluation approaches for the context. Notably prior evaluation demonstrates the potential for aspects of immersive audio rendering such as spatial clarity, localisation, and auralisation have the potential to impact the experience of networked interactions, providing practical direction for further investigation.

Chapter 4

Evaluation of an INMP System for Remote Music Education

> Science is not about building a body of known 'facts'. It is a method for asking awkward questions and subjecting them to a reality–check, thus avoiding the human tendency to believe whatever makes us feel good.
>
> *Terry Pratchett,*
> *The Science of Discworld (1999)*

4.1 Preface

This chapter details an evaluation of an INMP system designed for the practical application of remote music education. The practical use of the system in the field is briefly surmised. A following study is then detailed, where the system was reproduced and evaluated by three professional musicians and music tutors. An overview of the study design, objectives and hypothesis is given, and relevant background literature is discussed. Detail is provided on the INMP system reproduction and related acoustic measurements. This is followed by an overview of evaluation methods. Results are then presented and discussed, and summary remarks are given on key findings.

The INMP system presented in this chapter was implemented and operated by a project team from the AudioLab, Univeristy of York. This team were associated with the Musical Interactions in Networked Experiences using Real–time Virtual Audio (MINERVA) project[1], led by Professor Helena Daffern and Professor Gavin Kearney. The contributions of members of this project team to the work described in this chapter and the relevant INMP system are acknowledged here.

[1] `https://audiolab.york.ac.uk/minerva/`

The immersive audio rendering method (section 4.4) was specified by Helena Daffern, Gavin Kearney and Patrick Cairns, in collaboration with Bonza Music[2]. Relevant BRIR measurements (section 4.5) were performed using a MATLAB script and measurement protocol implemented by Gavin Kearney. BRIR measurement of Beehive Studios was performed by Gavin Kearney and Ben Lee. Other BRIR measurements described in this chapter were performed by Patrick Cairns. The INMP system (section 4.4) was implemented by Patrick Cairns with the assistance of Ben Lee and Anthony Hunt. The evaluation detailed in this chapter was conducted by Patrick Cairns with the assistance of the MINERVA project team and technical staff at the AudioLab.

Thanks is expressed to the following individuals for their contribution:

- Gavin Kearney (MINERVA project lead/sound engineer).
- Helena Daffern (MINERVA project lead/production).
- Anthony Hunt (network engineer).
- Ben Lee (network engineer/video operator).
- Jacob Cooper (sound engineer).
- Josh Gregg (sound engineer).
- Jay Harrison (sound engineer).
- Gerardo Acosta Martínez (sound engineer/production).
- Francesca Lorico (production).
- Panos Tsagkarakis (production).
- Andrew Chadwick (AudioLab studio manager).

Thanks is also expressed to technical staff from Bonza Music who assisted as video operators in the field (section 4.2).

4.2 An INMP System for Remote Music Education

Over 2021/22 a collaboration between the AudioLab, University of York, and Bonza Music developed a low–latency INMP system design for the context of remote music education. A prototype system, implemented as an extension to SoundJack NMP streaming software, was deployed over practical WAN connections and used to facilitate a remote music masterclass.

[2] `https://bonzamusic.com/`



Figure 4.1: Named artist delivering a remote music education masterclass to groups of students at multiple remote locations.

Here a named artist at a studio location in the United Kingdom (UK) led the masterclass, providing tuition to and performing with groups of students located at two remote locations in the UK (Figure 4.1). The INMP system provides musicians at each endpoint with an auralisation of the experience of listening to local and remote performers in a simulated 3–D acoustic environment in binaural format, with head–locked playback over headphones, and matching perspective presented in video display of remote performers. The system uses a stereo–based approach to audio processing and streaming, such that conventional stereo audio production workflows and tools may be used, and such that processing and bandwidth requirements of the INMP system are achievable on typical personal computing devices and WAN networks respectively.

A OWT latency of approximately 15–40 ms is reported between endpoints[3] when using the INMP system in the practical masterclass event. Two modes of auralisation were trialled in the masterclass event, reflecting the two primary modes of operation for immersive audio systems for immersive musical experiences [25]:

- *'They are all here'*: This case considers the auralisation of remote performers as if located in the local listening environment at each endpoint. This case is typically considered where remote performers are displayed visually in the local listening environment, for example through video or XR screen display, or using pass–through display in HMD–based XR systems.

- *'We are all there'*: This case considers the auralisation of local and remote performers as if located in a simulated acoustic environment which is not the local listening environment at each endpoint (i.e., a virtual space). This case is typically considered where local and remote performers are displayed visually in a virtual environment (for example using HMD–based VR systems), or in audio–only presentation of a virtual space.

[3]Calculated from network latency reporting in the SoundJack GUI.

4.3 Study Overview

Following the masterclass event a study was undertaken in order to evaluate and further investigate the experience of performance using the practical INMP system design. The system was reproduced over the LAN in the AudioLab and configured to provide equivalent OWT latency between endpoints as reported from practical WAN deployment. Considering evaluation by a small group of expert assessors, three professional musicians and music tutors participated in a study, performing using the system, and providing evaluation through interview feedback and rating of the performance experience. Included in the study is an evaluation of both *'The are all here'* and *'We are all there'* auralisation approaches. Additionally evaluation was conducted where immersive audio rendering was bypassed and traditional mono or stereo audio with no acoustic simulation DSP is considered. This is considered to provide baseline conditions as a reference for evaluation of different immersive audio rendering approaches. The following subsections provide an overview of the motivations for this study, as well as related objectives, hypotheses and statement of ethics.

4.3.1 Motivation

Remote music education workflows are considered as a means of delivering tuition and engaging in group practice where physical access is impossible or impractical (for example, with respect to travel requirements) [42], [44], [204]. This is most typically achieved using standard teleconferencing software [42], [204]. It is, however, widely acknowledged that high latency conditions here significantly impair musical interactivity and requires latency–coping strategies [51], inhibiting the quality of music education experiences [204]. Practical trials using NMP remote education solutions with video display note significant improvement relative to teleconferencing solutions [204], and propose integration of immersive audio rendering to improve the quality of the networked performance experience [15], [205]. Motivated by the potential for improvement of remote music education with INMP systems, this study addresses the viability of a practical novel INMP system design, demonstrated through performance experience evaluation.

Expanding on the motivation of experience improvement, evaluation here also works towards a practice–oriented understanding of performance experiences in the novel context of practical INMP systems[4]. This includes testing of expected improvement of networked performance with immersive audio rendering relative to mono or stereo baseline audio systems [16], [19], [121]. Also included is consideration of video display as is typical in remote music education contexts. Here some literature reports video display to be problematic due to high video latency and AV asynchrony with low–latency audio [15], [19], [97]. Other literature, however, report that video display is useful in communication between remote performers [122], and in delivery of visual performance cues [128].

[4] i.e., for INMP system designs which have been deployed in–application over practical WAN connections, and where performance conditions experienced in–application are reproduced.

4.3.2 Objectives

Central to this study is a practice–oriented testing of the viability of a practical INMP system design for remote music education in terms of experience rating. The following core objective is therefore considered:

> *To evaluate the quality of musical performance experience using the INMP system under evaluation where performance conditions experienced in–application are reproduced.*

Within this practice–oriented evaluation, this study additionally looks to test the potential for experience variation between immersive audio rendering and baseline audio systems, and to test the degree to which performers engage with video display. The following sub–objectives are therefore considered:

- *To evaluate the quality of musical performance experience using the INMP system for immersive audio and baseline mono and stereo audio rendering methods and provide comparison.*
- *To evaluate the degree to which performers engage with video display using the INMP system under evaluation.*

4.3.3 Hypothesis

With respect to the highlighted objectives of this study, the following primary hypothesis is considered:

> *Positive rating of the performance experience will be achieved using the INMP system under evaluation where performance conditions experienced in–application are reproduced.*

With respect to highlighted sub–objectives relating to testing for experience variation with audio rendering methods and relating to testing for performer engagement with video display, the following hypotheses are also considered:

- *Rating of the performance experience using the INMP system under evaluation will vary across immersive audio rendering methods and mono and stereo baseline audio rendering methods.*
- *Positive rating of engagement with video display will be achieved using the INMP system under evaluation.*

4.3.4 Statement of Ethics

Ethical approval for this study was obtained from the University of York School of Physics, Engineering and Technology Ethics Committee, reference *Daffern111021*.

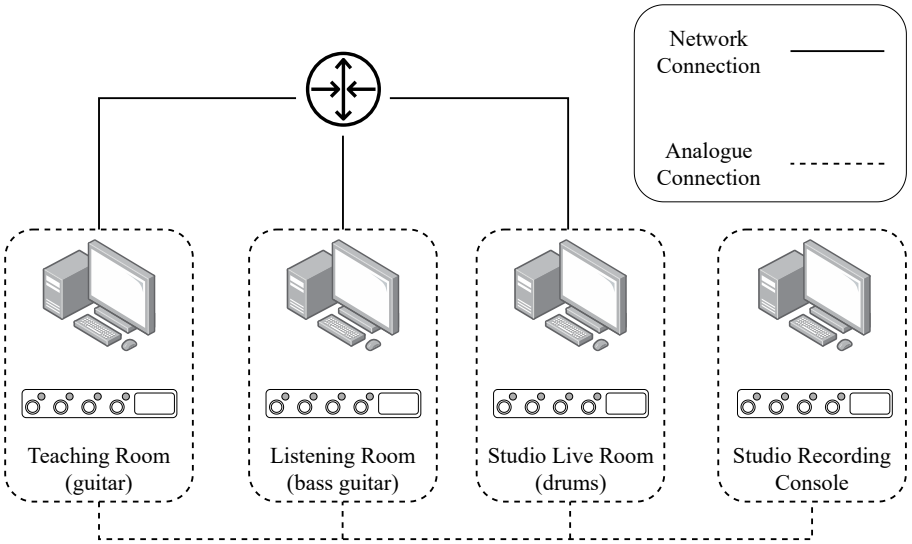


Figure 4.2: Network view of the system deployment, detailing LAN connection between the three endpoints, analogue lines to the recording studio console, and noting the performer associated with each endpoint.

4.4 System Overview

Reproduction of the INMP system design was deployed between three endpoints on the LAN at the AudioLab (Figure 4.2), University of York for use in the study. The INMP system design was implemented with minor modification for the study protocol, and in order to approximate conditions reported in WAN deployment. Each of the three endpoints is associated with a physical room, namely the AudioLab Teaching Room, Listening Room, and Studio Live Room. At each endpoint a performer is presented with an engineer–operated audio system, providing audio capture, low–latency streaming, and immersive audio rendering and monitoring functionality, and an engineer–operated video system, providing teleconferencing based video capture, streaming and screen display. The following subsections provide technical detail on the INMP system implementation.

4.4.1 Hardware and Physical Routing

Each endpoint is configured with the same basic hardware setup (Figure 4.3). Microphone or DI musician capture, and an engineer microphone used for talk–back and private technician communications, are input to a digital stereo mixing console (*Presonus StudioLive 16*[5]). Inputs are mixed, allowing application of traditional dynamic effects and Equalisation (EQ). The stereo mix of local instruments to be streamed to remote endpoints is output from the mixing console main outputs. A mono send to local musician self–monitoring, a mono send to local musician diffuse reverberation, and a mono send to a private technician communications channel, are output from the mixing console auxiliary outputs. This allows for control from the mixing console over self–monitoring direct sound and reverb levels in local musician monitor mixes.

[5] `www.presonus.com/products/studiolive-classic-1602-usb-digital-console-mixer`

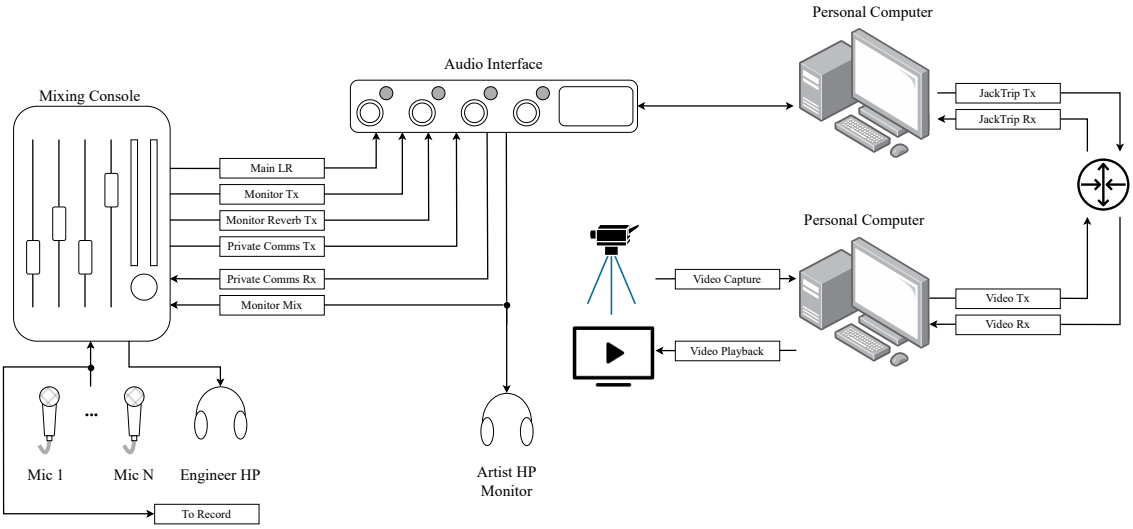


Figure 4.3: Illustration of hardware configuration at each endpoint, detailing relevant hardware and physical routing, audio and video capture and playback channels, and the LAN connection.

Mixing console outputs are connected to the inputs of an audio interface (*Focusrite Scarlet 18i20 g3*[6]*, 48 kHz, 128–sample frame size*) connected to a laptop computer (*(i7-11800H, 2.3 GHz, 32GB RAM, Win10*) which hosts low–latency NMP streaming and immersive audio rendering processes, and which is connected to the LAN. The local musician monitor mix is output from the audio interface to musician headphone monitoring (*Beyerdynamic DT770*), and a duplicate of the musician monitor mix is output from the audio interface to the mixing console allowing the local engineer to also monitor the mix. A summed mix of remote engineer private communications is also output from the audio interface to the mixing console for the local engineer to monitor.

Audio capture at each endpoint is specified by instrument. For the Teaching Room electric guitar (DI) and guitarist vocals (*Neumann u87*) were captured. For the Listening Room electric bass guitar (DI) and bassist vocals (*Neumann u87*) were captured. For the Studio Live Room an acoustic drum kit (*Shure Beta 52a, SM57, Rode M5, Sennheiser e604, AKG C414*) and drummer vocals (*Shure SM58)* were captured. In addition to mic inputs to the local mixing desk at each site a split signal for each channel was sent to the AudioLab Studio recording console via analogue lines for controlled recording (Figure 4.2).

A video system is deployed across endpoints which is asynchronous to the audio system (i.e., no attempt is made to synchronise video and audio, given that video latency will almost certainly exceed audio latency and AV asynchrony is an open issue native to the context of NMP). Here video is captured via a digital video camera, input to a laptop computer device which hosts teleconferencing software used in video streaming, and which is connected to the AudioLab LAN and Internet access. Two widescreen displays are used for playback of video, where each screen displays video capture from a discrete remote site, and screens are placed in front of musicians at each site.

[6] `www.focusrite.com/products/scarlett-18i20-3rd-gen`

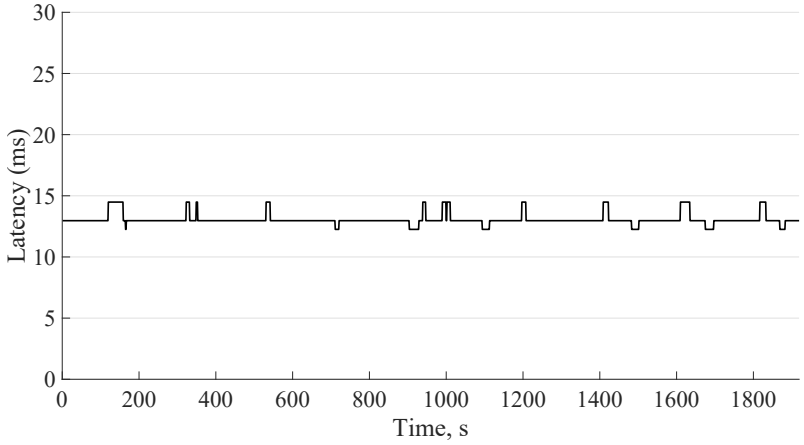


Figure 4.4: OWT network latency using JackTrip as configured in the INMP system, estimated from loopback test signal sampled over 0.5 hours.

4.4.2 Software Overview

At each endpoint two laptop devices are situated, one associated with each of the audio and video system, and hosting relevant software and processes. The video system laptop hosts common teleconferencing software (namely Zoom[7]), used to stream video between endpoints. The audio rendering system laptop at each endpoint hosts several audio applications (Figure 4.5). Audio interface driver software (*Focusrite Control*) provides control over capture and playback buffers and internal input and output routing. The NMP low–latency audio streaming software JackTrip provides network audio transport between endpoints for the INMP system. Immersive audio rendering is implemented as an extension to JackTrip, computed using audio plugins hosted in the DAW software Reaper. The audio routing application JACK is used to route audio between JackTrip, Reaper, and audio interface hardware buffers.

4.4.3 Audio Networking

At each endpoint two instances of JackTrip (specifically Jacktrip version 1.3[8]) are used, each providing low–latency audio transport to and from a discrete remote endpoint. Each JackTrip instance is configured for a jitter buffer size of 4 frames, FEC redundancy of two frames, and four channels of duplex audio (the stereo scene representing local performers at each endpoint, the private technician communications channel, and a spare channel). Estimation of network latency for JackTrip in this configuration was achieved by sending a test signal (namely a 1 Hz pulse train with 1 kHz duty cycle) from one endpoint to a remote endpoint and looping the signal back. OWT network delay between endpoints is then estimated as half the delay between relevant leading edges in the test signal and loopback recording (the Round–Trip–Time (RTT) delay). OWT network latency of $13\ ms \pm 1.5\ ms$ was estimated this way (Figure 4.4).

[7] `www.zoom.com`

[8] `www.github.com/jacktrip/jacktrip/releases`

At each endpoint the stereo mix of local performers and local private technician communications capture is routed to relevant JackTrip send channels for each remote endpoint. The stereo mix for each remote endpoint is routed from relevant JackTrip receive channels to Reaper inputs, and received private communications channels from remote endpoints are summed and routed to audio interface outputs. Transport of local musician audio to remote sites as a stereo scene represents a practical INMP design choice. Here the bandwidth requirement of each additional endpoint is limited to a duplex stereo stream plus an additional private communications channel for any number of performers at each discrete endpoint[9]

Notably the INMP system reproduction utilises JackTrip rather than SoundJack. In–application SoundJack provides useful features such as browser–based GUI control (simplifying operation and suitable for integration of a Virtual Learning Environment (VLE)), video display and Opus compression. JackTrip, however, allows for operation via CLI and configuration from shell script, and as an open–source software allows for greater control over the system. As such, where technicians are familiar with JackTrip, this solution is more practical for the development of evaluation frameworks. JackTrip and SoundJack, however, both follow standard low–latency NMP streaming design, and interchanging of software here is not expected to impact system performance.

4.4.4 Audio Rendering

The INMP system design implements a stereo–based binaural approach. Immersive audio auralisation here provides listeners with the experience of hearing networked performance as if listening over a stereo loudspeaker system in a simulated 3–D acoustic environment. This is achieved using stereo–based BRIR convolution following the virtual loudspeaker approach, and providing head–locked playback over headphones. Auralisation of diffuse reverberation is computed discretely from auralisation of direct sound and ERs. Audio routing (Figure 4.5) is configured internally in both audio interface driver software and DAW, and using JACK to route between applications and interface hardware buffers. BRIR convolution is performed in–DAW, implemented as audio plugins, namely instances of X–MCFX Convolver operating in zero–latency mode.

Stereo audio received from each remote endpoint via JackTrip is summed into a single stereo scene, such that a stereo image of all remote performers is produced. From the summed stereo mix left, right, and a summed mono channel are output. Left and right channels are convolved with a BRIR describing direct sound and ERs for the relevant virtual loudspeaker position. The mono sum from remote endpoints is summed with the mono diffuse reverb send from the local mixing console (controlling the amount of local instrument signal to diffuse reverberation), and the summed signal is input to BRIR convolution describing diffuse reverberation. Outputs from convolutions are then summed producing a single binaural scene describing auralised remote performers, and auralised diffuse reverberation for local performers. Using audio interface driver software (i.e., bypassing application buffering), the mono local musician self–monitoring signal received from the mixing console, describing local musician direct sound, is summed with the binaural scene, such that self–monitoring latency for local musicians is minimised.

[9] Within reason for the context of music tuition.

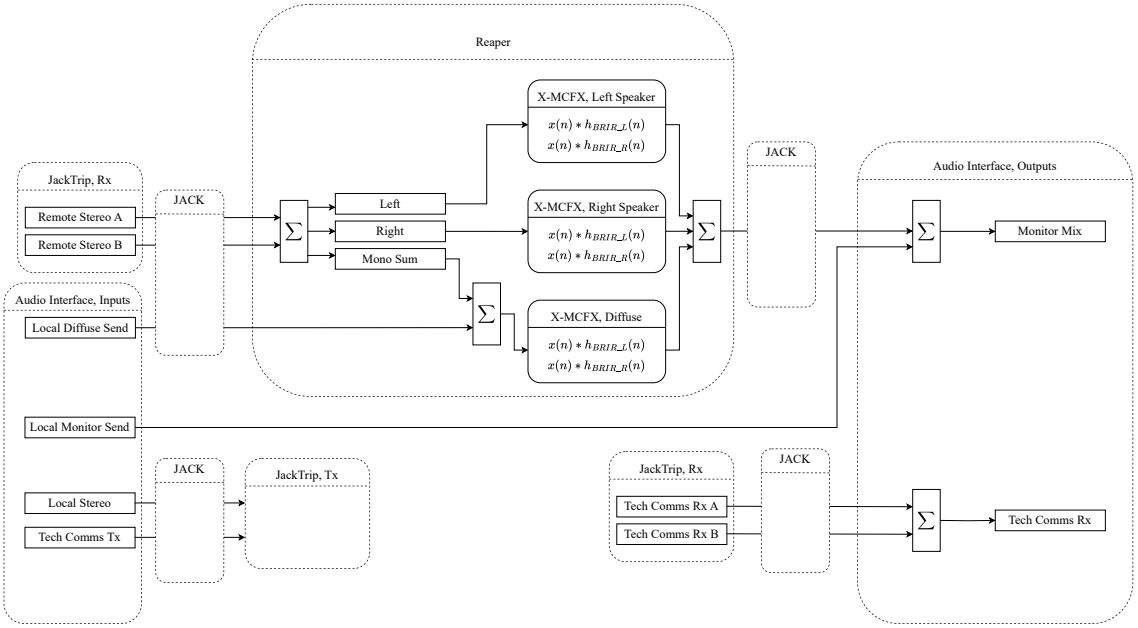


Figure 4.5: Binaural auralisation signal flow, illustrating auralisation processes, routing, and relevant audio applications.

The resultant stereo audio channels, containing binaural stereo auralisation of remote performers and local performer diffuse reverb, summed with mono self–monitoring signal, is output to provide the musician monitor mix at each endpoint.

The stereo–based approach here simplifies immersive audio rendering, such that for any number of local or remote performers only 6 channels of convolution are required, limiting processing requirements. Seperation of diffuse auralisation from direct sound and ER auralisation allows for resource–sharing in processing (i.e., all diffuse reverb is computed from a single BRIR convolution) and minimises the length of filters required in direct sound and ER auralisation.

4.5 BRIR Measurement

BRIR resource used in auralisation were captured in four discrete 3–D acoustic environments. This includes:

- The three local environments used as endpoints in the study (namely the Teaching Room, Listening Room and Studio Live Room at the AudioLab, University of York). These measurements provide resource used in *'They are all here'* auralisation methods.
- A medium hall at Beehive Studios, London (Figure 4.6). Measurement of this space provides resource used in *'We are all there'* auralisation methods.

The following subsections provide technical detail on BRIR measurement and processing, and report the acoustic parameters of relevant measured spaces.



Figure 4.6: BRIR measurement in Beehive Studios.

4.5.1 Measurement Method

BRIR measurement was performed using the sine sweep method [148] as in acoustic measurement standards [141]. A MATLAB script created by Gavin Kearney was used to generate a test signal and inverse filter, parse test signal recordings, and extract BRIR. For all BRIR measurements a Genelec 8040 A[10] loudspeaker was used to output the test signal, and test signal was recorded using a Neumann KU–100 Binaural Dummy head system[11]. Measurement considers a Left–Centre–Right (LCR) loudspeaker configuration (Figure 4.7), allowing for an optional centre channel. The centre channel is not used in this study.

For the Teaching Room, Listening Room and Studio Live Room measurements loudspeaker height was measured at 1.4 m from floor to tweeter, receiver height was measured at 1.2 m from floor to ear canal entrance, loudspeaker level was calibrated to 66–70 dBA (measured at 1 m), noise floor ranged from 40–44 dBA, rooms were unoccupied, and a 0.5 s sweep length was used in BRIR measurement. For Beehive Studio measurements, loudspeaker height was measured at 1.5 m from floor to tweeter, receiver height was measured at 1.2 m from floor to ear canal entrance, loudspeaker level was calibrated to 85 dBC (measured at 1 m), and the space was unoccupied. Noise floor and sweep length were not noted for these measurements.

[10] `www.genelec.com/previous-models/8040a`

[11] `www.neumann.com/en-gb/products/microphones/ku-100`

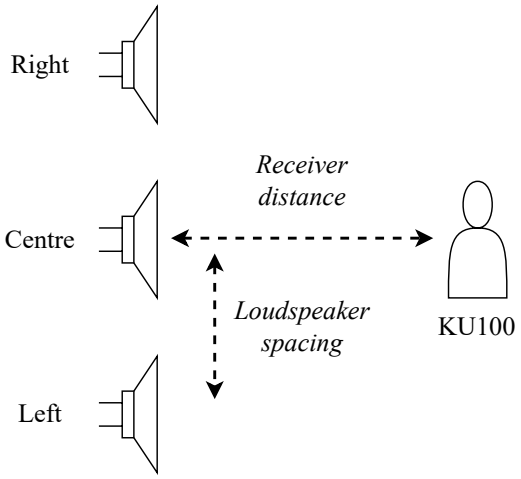


Figure 4.7: BRIR measurement configuration, illustrating KU100 receiver placement and distance, and left, centre and right loudspeaker placement and spacing.

Loudspeaker spacing and receiver distance from the centre loudspeaker position in measurement (Figure 4.7) was not noted and varied across measured environments depending on available space and reasonable positioning of video display screens and performers. In addition to LCR loudspeaker positions, a separate BRIR measurement was performed for each space. Here loudspeaker and receiver positioning was considered arbitrary. This measurement was taken to be processed, removing direct sound and ERs to produce a BRIR containing only diffuse reverberation.

4.5.2 BRIR Processing

The MATLAB script used in BRIR measurement performs post–processing on acoustic measurements. The start of each of the LCR loudspeaker BRIR is truncated to 4 ms before the peak of direct sound, adding a 2 ms half–Hanning window fade–in (removing latency associated from convolution with 'dead air' in IR for real–time auralisation). Following this each BRIR here was truncated to 1 s length. For diffuse BRIR measurement the start of the BRIR is truncated to 10 ms after the peak of direct sound and a half–Hanning window fade–in is applied from 10 ms after the peak of direct sound to 30 ms after the peak of direct sound. This is considered to remove direct sound and ERs leaving only diffuse reverberation (Figure 4.8).

4.5.3 Acoustic Parameters

Acoustic parameters (Table 4.1) for the Teaching Room, Listening Room and Studio Live Room were estimated from reference IR provided by the studio manager, Andrew Chadwick. These measurements were performed using a Genelec 8040A loudspeaker, and Earthworks M–30 omnidirectional microphone[12].

[12] `www.earthworksaudio.com/measurement-microphones/m30/`

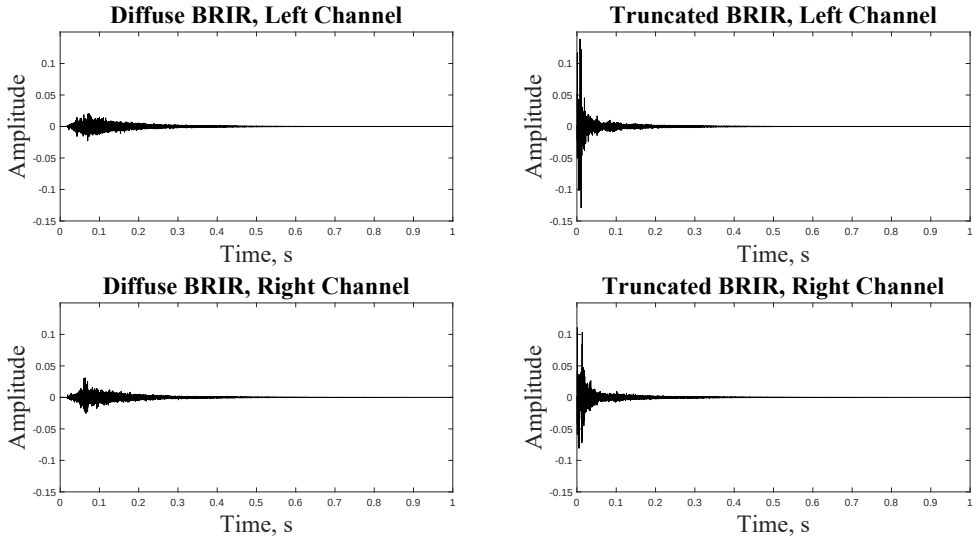


Figure 4.8: BRIR describing Beehive Studios, detailing truncated BRIR used in auralisation, including diffuse reverb BRIR.

No reference IR measurement was performed for Beehive Studios. Here acoustic parameters are estimated from an IR where the left and right channels of the centre loudspeaker BRIR measurement are summed. As such these measurements are considered a coarse estimation.

4.6 Evaluation Protocol

The following subsections provide detail on the protocol used to evaluate the INMP system.

4.6.1 Protocol Overview

Evaluation of the INMP system is conducted through standard NMP performance task and QoE rating protocol design. Here three participants (one located at each endpoint performing a discrete instrument) were required to attempt a musical performance task multiple times using the INMP system, providing rating of the performance experience via questionnaire response after each attempt of the performance task.

Room	EDT (s)	T30 (s)	C80 (dB)	D80 (%)
Studio Live Room	0.1524	0.1894	30.5	99.91
Teaching Room	0.1165	0.3217	25	99.68
Listening Room	0.1014	0.1494	40.2	99.99
Beehive Hall	0.225	0.931	12.1	94.14

Table 4.1: A–weighted acoustic parameters describing the Teaching Room, Listening Room, Studio Live Room and Beehive Studios.

Rating of the performance experience tests the core hypothesis of this study, where positive rating of the experience is considered to infer the viability of the practical INMP design. In order to test for variation of performance experience with audio rendering method, the performance task is attempted for different audio rendering configurations, namely:

- *Mono*: Where BRIR rendering is bypassed, and where the stereo monitor mix for each networked musician is summed to a mono channel, and dual–mono playback is provided over monitoring headphones.
- *Stereo*: Where BRIR rendering is bypassed, and where the stereo mix for each networked musician is sent directly to monitoring headphone playback.
- *They are all here*: Where BRIR used in auralisation describe the local listening environment (namely the Studio Booth, Listening Room and Teaching Room, with each environment associated with a discrete musician), and auralised binaural audio is sent to monitoring headphone playback.
- *We are all there*: Where BRIR used in auralisation for each networked performer describe Beehive Studios, and auralised binaural audio is sent to monitoring headphone playback.

The performance task is attempted three times for each audio rendering method in a randomised order. Testing of performer engagement with video display is addressed via a discrete rating item included in the questionnaire. This protocol was completed for two pieces of performance material, namely:

- Cissy Strut, by The Meters.
- Kashmir, by Led Zeppelin.

Following completion of the protocol for both pieces of performance material participants discussed the experience in a semi–structured exit interview.

4.6.2 Latency Conditions

In order to reflect performance conditions experienced in practical use the INMP system is configured such that latency which approximates the 21–41 ms which is reported in–application. The system was configured to provide OWT network latency of approximately 13 ms as detailed in subsection 4.4.3. Summing this with audio interface throughput latency of 5.3 ms (two blocks of 128 samples at 48 kHz) and 2 ms of latency from mixing console throughput, OWT system throughput[13] is estimated as approximately 20 ms.

[13]Throughput latency for convolution is considered negligable.



Figure 4.9: The drummer participant as situated using the INMP system and as captured by the video system.

4.6.3 Participants

This study was completed by three expert assessors, each performing a discrete instrument (namely electric guitar, electric bass guitar and drums). All participants are professional musicians and music tutors. Though all participants had prior experience in remote music performance using teleconferencing solutions, none had prior experience with low–latency NMP systems, however, were allowed a practice period using the INMP system prior to the study protocol starting.

4.6.4 Performance Task

For each attempt of the performance task the three musicians were required to perform the relevant performance material together using the INMP system (Figure 4.9). The exact content of each performance task varied under instruction from the lead engineer (delivered via engineer microphone input). Participants were instructed to begin performance from different parts of the performance material, instructed when to end performance, and instructed to improvise on occasion.

4.6.5 Questionnaire

The questionnaire consisted of the following items, each rated in terms of agreement (7–point scale, strong disagree (1)–strong agree (7) with a neutral mid–point (4):

1. We can play together as though in the same room.
2. I feel like I am in the same space as the other performers.
3. I have a sense of where the other performers are located in the space.
4. The other performers sound as though they are playing from where I see them on screen.
5. I am enjoying playing in this acoustic.
6. I am conscious of audio delay between performers.
7. The quality of audio is of a professional standard.
8. I engaged with video display during performance.

Item 1 and item 6 consider standard NMP QoE rating items, relating to perceived performance conditions, and perceived latency conditions respectively [6]. Item 2 provides rating of perceived presence [124], [125]. Item 3 relates to localisation, however, does not explicitly pertain to aural localisation. Item 4 specifically provides rating of aural localisation [180]. Item 5 introduces rating of acoustics, here in terms of the influence of acoustics on enjoyment of the NMP QoE rating concept of perceived performance conditions. Item 7 provides an absolute rating of BAQ, as considered in both audio system rating [180] and standard NMP QoE rating [6]. Item 8 addresses performer engagement with video display directly, which may be considered relevant to immersive experience modelling with relevance to system interaction factors [125]. Performer rating of questionnaire items was recorded via a Qualtrics[14] survey interface using a tablet device.

4.6.6 Exit Interview

In addition to questionnaire item rating the three musicians participated in a brief semi–structured exit interview. Interview prompts here addressed:

- Prior experience of remote music performance.
- Participant reaction to the performance experience using the INMP system.
- Comparison of the experience of performing using the INMP system to prior participant performance experiences.
- Participant perception of different audio rendering methods.
- Perceived performance conditions.
- Perception of video display.

[14] `www.qualtrics.com/`

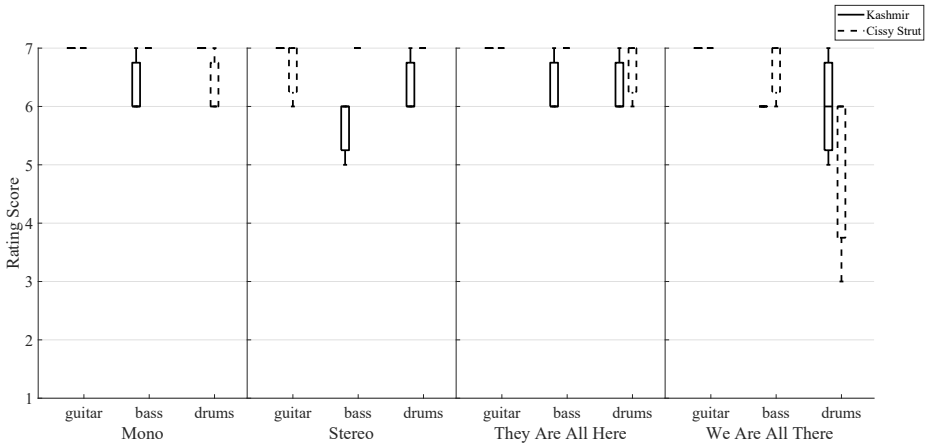


Figure 4.10: Box plot illustrating rating of agreement for questionnaire item 1: *'We can play together as though in the same room'*.

4.6.7 Study Limitations

Several limitations of the protocol can be acknowledged. Firstly, a sample size of only three participants is considered. Statistical hypothesis testing is not considered reliable with this sample size. The exact content of the performance task also varies across the protocol, presenting a minor lapse in study control. Various aspects of monitoring context also vary across the three participants. This includes the individual participant, instrument performed, monitor mix and video display content, local physical environment, and virtual loudspeaker and listening positioning. This monitoring context variation is representative of a practical use–case scenario, however can reasonably be expected to influence varition in experience ratings across monitoring contexts.

4.7 Results

The following subsection reports the results of the INMP system evaluation.

4.7.1 Experience Ratings

Item 1, relating to perceived performance conditions (Figure 4.10) is rated positively with ratings of slight–strong agreement in all but one case, namely drummer rating of the *'We are all there'* audio conditions for the performance material *Cissy Strut*. Here ratings range from slight disagreement–agreement. Notably the drummer provides lower ratings for the *'We are all there'* audio rendering method than for other methods evaluated.

Guitar and drum performers both provide rating of item 2 (Figure 4.11), addressing perceived presence, which is generally positive, ranging from slight–strong agreement. The bass performer, however, provides rating here ranging from slight disagreement–agreement for both stereo and *'We are all there'* audio rendering methods.

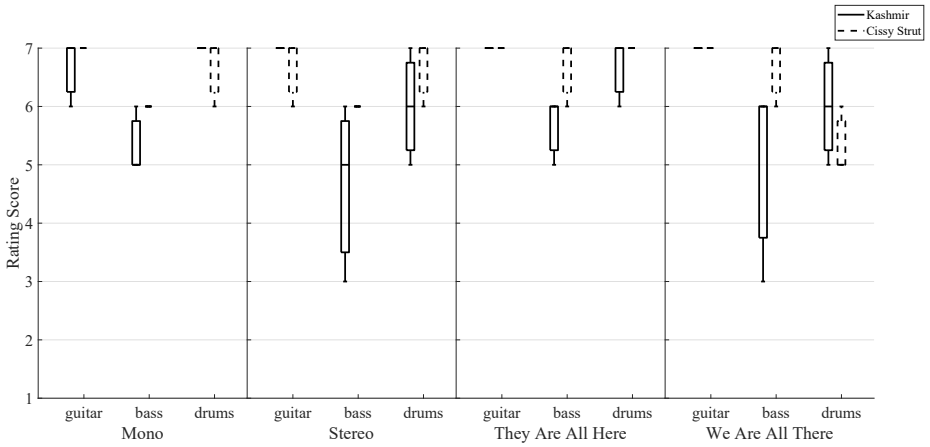


Figure 4.11: Box plot illustrating rating of agreement for questionnaire item 2: *'I feel like I am in the same space as the other performers.'*

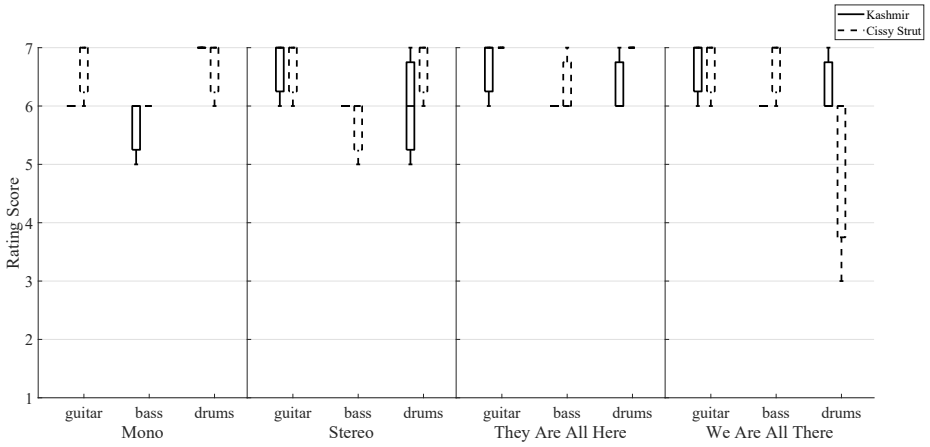


Figure 4.12: Box plot illustrating rating of agreement for questionnaire item 3: *'I have a sense of where other performers are located in the space'*.

Rating of item 3 (Figure 4.12), describing localisation (without specifying aural localisation), is again generally positive, with the majority of ratings ranging from slight–strong agreement. The one exception here is rating of *'We are all there'* audio methods for *Cissy Strut* performance material from the drum performer. Ratings here range from slight disagreement–agreement.

Specifically aural localisation is addressed in questionnaire item 4 (Figure 4.13), where rating varies. The guitarist generally provides positive rating of slight–strong agreement, however, rates stereo conditions between slight disagreement–agreement for performance of *'Cissy Strut'*, and rating of neutrality–slight agreement for performance of *'Kashmir'* under the *'We are all there'* audio rendering method. The bass performer also generally provides positive rating between slight–strong agreement. The exception here is rating of performance of *'Cissy Strut'* under stereo audio rendering methods, where rating between slight disagreement–agreement is given. The drum performer here consistently provides positive rating from slight–strong agreement.

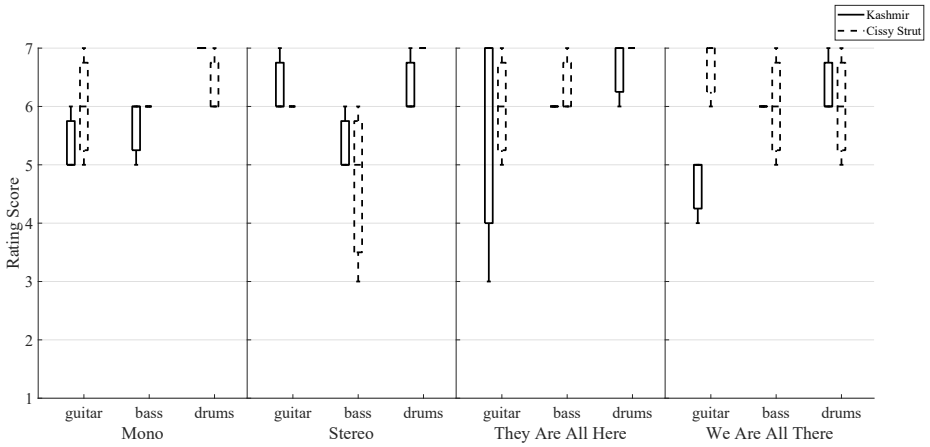


Figure 4.13: Box plot illustrating rating of agreement for questionnaire item 4: *'The other performers sound as thought they are playing from where I see them on screen.'*

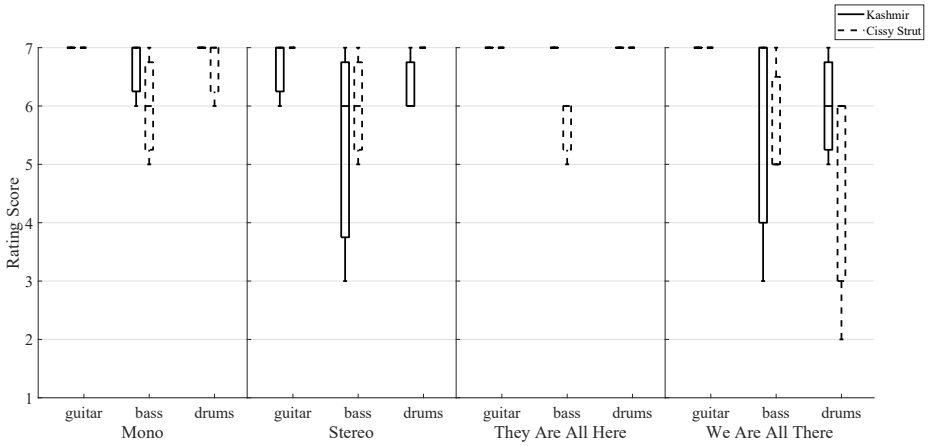


Figure 4.14: Box plot illustrating rating of agreement for questionnaire item 5: *'I am enjoying playing in this acoustic.'*

Rating of questionnaire item 5 (Figure 4.14), addressing perceived enjoyment of rendered acoustics varies by participant. The guitarist provides ratings of strong agreement for all performance material and audio rendering methods. The bassist generally provides positive ratings from slight–strong agreement, however, notably rates stereo and *'We are all there'* audio rendering methods from slight disagreement–strong agreement for performance of *Kashmir*. The drummer again generally provides positive ratings between slight–strong agreement, where, here, the notable exception is ratings of *'We are all there'* audio rendering methods for performance of *'Cissy Strut'*.

Rating again varies across participants for rating of questionnaire item 6 (Figure 4.15), which addresses perception of latency. The guitarist here provides negative ratings, providing consistent ratings from slight–strong agreement across performance material and audio rendering methods. The bassist provides ratings from neutrality–disagreement for stereo and *'They are all here'* audio rendering methods and ratings from disagreement–strong disagreement for the *'We are all there'* audio rendering method.

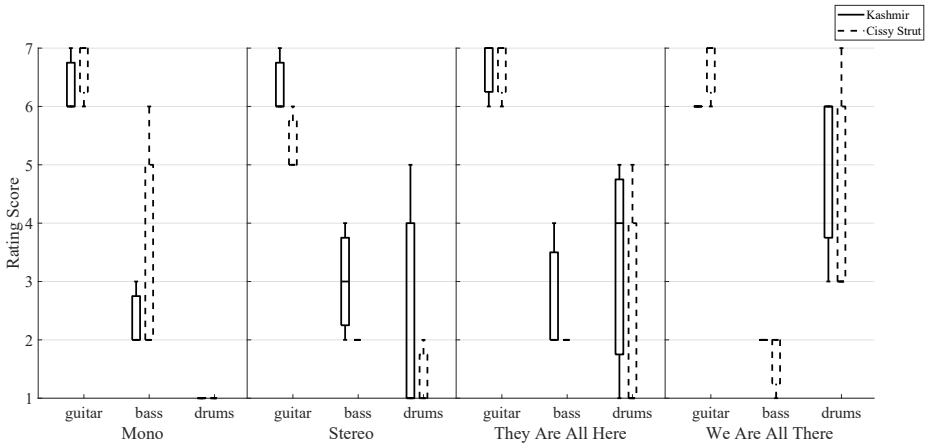


Figure 4.15: Box plot illustrating rating of agreement for questionnaire item 6: *'I am conscious of audio delay between performers.'*

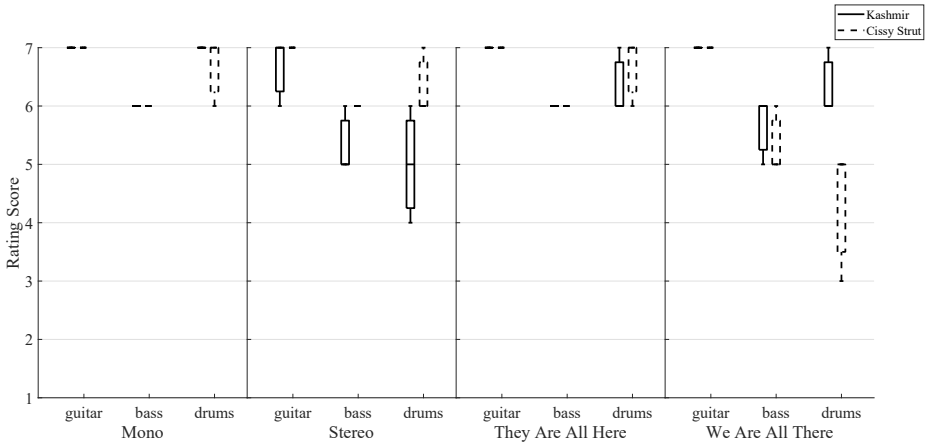


Figure 4.16: Box plot illustrating rating of agreement for questionnaire item 7: *'The quality of audio is of a professional standard.'*

For mono audio methods the bassist rates the performance of *Kashmir* from slight disagreement–disagreement, and the performance of *Cissy Strut* from disagreement–agreement. The drummer provides consistent ratings of strong disagreement for mono conditions, strong disagreement–slight agreement for stereo and *'They are all here'* conditions, and rating from slight disagreement–strong agreement for *'We are all there'* audio rendering conditions (notably this audio rendering method is rated poorer than alternative audio rendering methods).

For item 7 (Figure 4.16), addressing perceived BAQ, the guitarist provides consistent positive rating between agreement–strong agreement. The bassist also provides consistent positive rating, here ranging from slight agreement–agreement. The drummer generally provides positive rating between agreement–strong agreement, however, notably for performance of *Cissy Strut* provide lower rating for the *'They are all here'* audio rendering method, here ranging from slight disagreement–slight agreement.

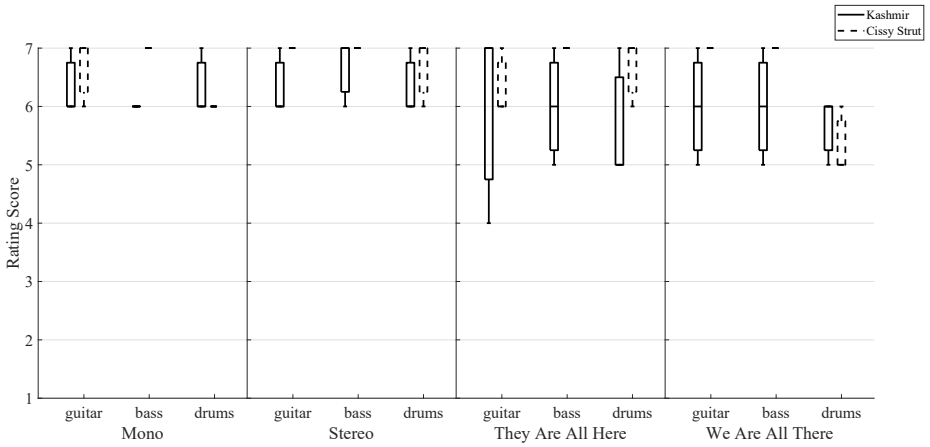


Figure 4.17: Box plot illustrating rating of agreement for questionnaire item 8: *'I engaged with video display during performance.'*.

Rating of item 8, addressing performer engagement with video display (Figure 4.17) is again generally positive. For mono and stereo audio methods all participants provide rating from agreement–strong agreement. For *'They are all here'* audio methods rating generally ranges from slight–strong agreement, with the exception of the guitarist, who provides rating from neutrality–strong agreement for performance of *Kashmir*. For *'We are all there'* audio methods ratings range from slight–strong agreement.

4.7.2 Exit Interview Response

Each of the participants reported previous experience using teleconferencing solutions for remote music performance and tuition. In discussion participants described their previous experience, where all participants identified the issue of latency, impairment of musical interactivity, and requirement for latency coping strategies. The bassist commented that to play with remote performers *'you just need to play out of time'* (acting as a backing track for remote performers). The drummer noted that teaching *'was a bit of a nightmare...either I couldn't hear them or they couldn't hear me...there was nothing where you could perform together'*. The guitarist noted that they *'never play with anyone'* and that instead they provide backing tracks for students, noting *'they play and I listen to them, because it just doesn't work...so much delay on it'*.

Participants responded positively when asked about their initial impression performing using the INMP system. The drummer comments *'I wasn't expecting it to be as good as it was. I just didn't think it could happen'*. When discussing what previous experience performance using the INMP system was perceived as most similar to, the drummer and bassist considered the experience most similar to live performance using IEMs. Participants also reported perception of the experience varying with practice. The bassist commented *'It changed for me throughout the day. (To start) it was like a (studio) session, where everyone was in booths... by the end of it, it had become more interactive I think.'*

The drummer provided similar comment, stating *'the beginning it felt like a recording session... it did feel like it could be like doing a rehearsal... as you got more used to it we were interacting more... you could get to the stage where it could feel like a gig'*.

When asked about what variations participants noticed between performances across the study, participants highlighted different aspects of audio rendering. Firstly the guitarist notes *'some of the reverbs were longer and some were shorter.'* In following discussion of reverberation, the drummer comments on preferences in rendered acoustics, noting that for *'We are all there'* audio methods *'there was a delay on the hi–hats then, it felt like it was really dragging... when it was drier it felt better for me personally.'*. Also relating to *'We are all there'* audio rendering methods, the bassist comment *'the more ambient settings I was losing the front end of my note on my bass... I was losing that direct attack... I was really overplaying all the time on those ones... but they felt vibey–er. It would be great if I could get myself on the dry setting and (remote performers) on the ambient setting'*. The drummer provides comment in agreement here stating *'you could play better together (in dry acoustics), but it didn't have the same ambiance.'* The guitarist provides following comment *'I like the reverb, if you play without it just sounds really dry to me'*.

The bassist identifies an audio rendering condition between dry audio and *'We are all there'* methods, noting *'there was a couple (that felt in—-between)... not as much of the vibey ambient sound, but did have enough detail on the bass.'* Participants also commented on spatialisation in audio rendering. Here the bassist stated *'there was one where it panned really far left and right and then I lost the spatial positioning of (remote performers)'*. Variation in perceived spatialisation was noted by multiple participants.

When prompted to consider experience of presence, the bassist notes variation in experience with time spent performing, stating *'I felt more like I was in the room by the end of the day than I did at the beginning... I've got into it.'* The drummer comments on differences in experience of presence relating to audio rendering method and perceived difficulty of the performance task. Here the drummer notes that for *'We are all there'* conditions *'it was harder to play so (I was) more concentrating on keeping in time as opposed to just relaxing, once we did that it felt much more like we were in the same room.'*

Participants reported some variation in experience when discussing ability to interact musically and perform synchronously using the INMP system. Participants agreed that they were able to interact with one another, and were not adopting leader–follower strategies (associated with high latency conditions). When asked about the degree to which performers perceived they had to concentrate to play in time, the bassist comments *'not at all for me.'* The drummer, however, considers variation associated with audio rendering methods, and time spent using the INMP system, stating *'I did a bit... especially early on in the day but as we went on I didn't as much, but when it was more ambient, I could hear the reverbs, then I had to concentrate more.'*

Participants provided comment on video display throughout the exit interview. All participants particularly noted an awareness of AV asynchrony. The bassist considered that awareness of video delay and AV asynchrony depended on instrumentation, commenting *'it wasn't throwing me at all day watching (guitar) play – nature of guitar isn't it, you are sometimes anticipating a move so your hands are there before you played it, but (drums) at first was really (noticeably out of time).'*

The guitarist noted a focus on audio rather than video, stating that when performing a clapping exercise during the practice period using the INMP system *'I just listened to the one, in my left ear.'* The drummer considers that they engaged with video display, but ignored visual cues of hand–clapping during this exercise, commenting *'I didn't look at the hands clapping, but I would look (at video display).'* The bassist notes that *'threw me at first and then didn't at all by the end of the day'*, reporting a variation in experience of video display with time spent using the INMP system.

4.8 Discussion

In the following subsections the results of evaluation are discussed with respect to overall evaluation of the INMP system, the experience of immersive audio, the experience of video display, the experience of different participants, and the experience of different performance material.

4.8.1 Overall System Evaluation

For rating items 1–4 and rating items 7 and 8, positive ratings are generally provided, with negative and neutral ratings being outliers. This implies that a general high–quality experience is provided with respect to performance quality, presence, localisation, BAQ, and video display engagement.

For rating of enjoyment of performance acoustics (item 5) it is highlighted that *'We are all there'* auralisation methods may result in some experience impairment. The INMP system is otherwise rated generally positively for this item, indicating a high–quality experience with respect to enjoyment of performance acoustics.

Rating of perceived latency varies highly and contains ratings which indicate high awareness of latency, and therefore impairment of performance experience quality. Despite this, in exit interview discussion performers indicate no associated impairment in ability to perform in time, no requirement for latency–coping strategies, and consider performance using the INMP system comparable to live monitoring contexts such as studio and stage monitoring systems. It is possible that rating of perceived audio delay highlights the participants' awareness of NMP systems and the issue of latency, rather than poor system performance using the INMP system under evaluation.

Considering the core objective of validating the practical INMP system under evaluation, results generally indicate that a high quality performance experience is indeed achieved using this system. Negative or neutral ratings of *'We are all there'* rendering methods, and variation in rating between audio rendering methods, does, however, indicate cases where high quality performance experience is not achieved.

4.8.2 Experience of Different Audio Rendering Methods

In exit interview discussion, participants indicate an awareness of different audio rendering approaches in musician monitoring using the INMP system.

Notably, for rating items 1, 5, 6 and 7 (relating to perceived performance quality, enjoyment of acoustics, experience of latency and BAQ) the drummer provides lower ratings for *'We are all there'* audio rendering methods than other methods, including negative responses to performance experience rating. The drummer also specifically identifies a preference for 'dry' acoustics in exit interview comment, considering greater performance difficulty for *'We are all there'* audio rendering methods. The bassist provides lower rating of *'We are all there'* audio rendering methods, including negative experience rating, for item 5 (enjoyment of performance acoustics). In exit interview discussion the bassist reports greater difficulty in performance here, noting poor clarity under *'We are all there'* audio rendering methods.

The RT of *'We are all there'* audio rendering methods (0.93 s) is substantially higher than the RT of *'They are here'* conditions (0.15–0.32 s) or dry mono or stereo baseline conditions. C80 also varies between *'We are all there'* (12.1 dB) and *'They are all here'* (25–40.2 dB) conditions. Relatively high RT and low C80 may explain negative ratings of *'We are all there'* audio rendering methods. Indeed prior study of reverberation effects in NMP reports that dry conditions are preferred over reverberant conditions due to improved clarity for drummers [119]. Considering ratings of perceived delay (item 5), the drummer provides ratings which imply delay is more noticeable in *'We are all there'* audio rendering conditions. This is contrary to the findings previous literature which suggests that reverberation will mask perception of delay [203].

With respect to localisation, participants note in exit interview discussion that the positioning of sources appeared to change throughout the study protocol. It can be acknowledged that the stereo image presented in baseline stereo and binaural auralisation is not coherent. The pan positions are scaled to match the width of a virtual loudspeaker array in binaural auralisation, however, in baseline stereo conditions they are presented directly over headphones. Similar issues with stereo image consistency have been reported in other work [121]). Despite awareness of localisation variation, overt discernment between audio rendering methods is not evident in relevant rating items (items 3 and 4). It may be the case that ventriloquism[15] [206] associated with video display impacts ratings here.

In this study discernment between immersive audio and baseline mono and stereo audio rendering approaches is not generally overtly evident across experience rating items. Work by Genovese, however, provides similar evaluation of audio rendering methods for INMP with video display [16]. Here it is suggested that *'They are all here'* audio rendering methods are preferable. Indeed in rating of enjoyment of performance acoustics (item 5) in this study, it can be noted that *'They are all here'* audio rendering methods achieve highest median rating. Considering the objective of evaluating the potential for INMP experience quality to vary with audio rendering method, rating results from this study do not identify overt improvement in performance experience between baseline and immersive audio rendering methods. It is, however, identified that INMP experience quality may vary with auralised space, suggesting that the quality of INMP experience may be related to acoustic parameters of auralised spaces.

[15] A perceptual phenomenon where sound sources appear localised to match visual stimuli.

4.8.3 Experience of Video Display

Ratings of questionnaire item 8, addressing participant engagement with video display, report almost unanimous agreement from all participants. Though some previous literature reports that video display may be ignored in NMP experiences [15], [97], [100], results from this study are coherent with literature which reports performer engagement with video systems [122], [128].

It may be the case that experience of video display varies with aspects of the performance experience using INMP systems. Though all performers noted AV asynchrony in video display due to high video latency (as is reported in other study of INMP systems [19]) one participant in this study did consider experience of video display to improve with time spent using the INMP system. Indeed prior work has suggested that the experience of NMP will improve with the amount of practice using the relevant NMP system [6], [16], [35], [116]. It was also considered in exit interview that experience of AV asynchrony was dependent on visibility of visual performance cues. For example one performer noted AV asynchrony being more noticeable for drums than for guitar, and another noted that their attention focused on audio when conducting a clapping exercise.

Considering the study objective of testing performer engagement with video display using the INMP system, it is clear from results that performers do indeed engage with video display. With respect to best–practice recommendations for INMP, it is suggested that optimisation of experience of video display in INMP should consider visibility of visual performance cues, and should afford performers time to practice using the INMP system.

4.8.4 Experience of Different Participants

For the majority of questionnaire items (with the exception of item 8) some degree of variation in rating can be observed in the response provided by discrete participants. In some cases variation is minor, for example in items 1 or 3, where only drummer rating of *'We are all there'* audio rendering methods deviates from other ratings. In other cases rating variation is more pronounced, for example in rating of items 2, 4, 5, 6 and 7, where multiple cases of rating differences between participants may be highlighted. As discussed briefly in subsection 4.6.7, differences between participants include the individual and various aspects of the monitoring context (for example instrumentation, monitoring and video display content, local physical environment, virtual loudspeaker and video display and capture configuration). Though factors which influence rating variation between participants may be ambiguous here, some key factors of monitoring context may be considered.

Prior work has shown that experience of NMP may vary with dynamic qualities of performance, where fast attack times are associated with a more acute experience of latency than slow attack times [111]. Literature also reports variation of NMP experience with timbral qualities, where brighter and noisier timbre is associated with a more acute experience of latency [108], [114].

Indeed, prior study has reported variation in synchrony and quality of NMP under different latency conditions for string and woodwind instruments [109]. In this study, several cases can be noted where ratings from the drummer vary across audio rendering methods but ratings from the bassist and guitarist do not. Examples include questionnaire items 1, 3, 6 and 7. It can be considered that both the dynamic and timbral qualities of drums are distinct from guitar and bass guitar. Considering this, it is possible rating variation between participants is associated with instrumentation.

It can also be considered that the monitoring context varies across participants. This includes aspects of immersive audio auralisation (i.e., auralised room for *'They are here'* auralisation methods, virtual loudspeaker configuration, and monitoring mix content) and aspects of visual display (i.e., video display placement and content). Particularly of note the drummer has high SPL in the room from the kit which may spill into headphone monitoring. Guitar and bass, however, are captured via DI, and hear only the monitor mix over headphones. Considering this it is possible that rating variation between participants is associated with monitoring context. It is also possible that rating variation reflects individual perception and/or preferences. For example with respect to rating item 6, relating to perception of latency, each participant provides different ratings despite all participants experiencing the same latency conditions.

Though the reason for rating variation between participants is unclear, it is evident that the experience of performance using the INMP system under evaluation does vary across participants. In acknowledgment of this, it can be considered a practical recommendation that individualised control of audio rendering is provided in INMP systems, such that audio rendering system design may cater to various preferences and monitoring contexts.

4.8.5 Experience of Different Performance Material

Though comparison of performance material is not a focus of this study, the study protocol was completed for multiple pieces of performance material. There are some cases in experience rating, where rating varies between performance material, for example in drummer rating of *'We are all there'* audio rendering methods for items 3 and 7. Rating of the performance experience generally does not appear to vary with performance material, however, and no observable pattern emerges in results here. It can, however, be noted that prior work demonstrates variation in NMP experience with aspects of performance material, such as rhythmic complexity [119], and more extreme variation of performance material may elicit observable variation in experience rating.

4.9 Hypothesis Revisited

At the beginning of this chapter the primary objective was presented:

> *To evaluate the quality of musical performance experience using the INMP system under evaluation where performance conditions experienced in–application are reproduced.*

This objective was accomplished by conducting QoE evaluation of an INMP system which was used in practical WAN deployment, and where evaluation conditions approximate the conditions experienced in practical WAN deployment.

The primary objective was accompanied by the sub–objectives:

- *To evaluate the quality of musical performance experience using the INMP system for immersive audio and baseline mono and stereo audio rendering methods and provide comparison.*
- *To evaluate the degree to which performers engage with video display using the INMP system under evaluation.*

The first sub–objective was achieved by conducting QoE evaluation for different audio rendering conditions, including immersive audio rendering, and aforementioned mono and stereo baselines. The second sub–objective was achieved through the inclusion of rating of video display engagement in QoE evaluation.

In completion of these objectives, the related primary hypothesis was presented:

> *Positive rating of the performance experience will be achieved using the INMP system under evaluation where performance conditions experienced in–application are reproduced.*

As discussed in subsection 4.8.1, a generally positive experience was indeed achievable for all participants (though the specific conditions under which this is the case for each performer did vary). Participant feedback also indicated that synchronous performance was possible. The primary hypothesis is therefore accepted. It is, however, acknowledged that the conditions under which this hypothesis is accepted may vary with the performer in question.

The relevant secondary hypotheses were presented:

- *Rating of the performance experience using the INMP system under evaluation will vary across immersive audio rendering methods and mono and stereo baseline audio rendering methods.*
- *Positive rating of engagement with video display will be achieved using the INMP system under evaluation.*

As discussed in subsection 4.8.2, rating, and feedback regarding the performance experience, did indeed vary for different audio rendering methods. The first of the secondary hypotheses is therefore accepted. It can be acknowledged, from participant rating and feedback, that audio rendering preferences were different for different performers. The sample size in this study is, however, too small for any statistical modelling of preferences in audio rendering. As discussed in subsection 4.8.3, participants unanimously provided positive rating of engagement with video display. The second of the secondary hypotheses is therefore accepted.

4.10 Summary

The INMP system evaluation presented novelly demonstrates the ability of a practical INMP system design from remote music education applications to provide a high–quality performance experience. This contributes a practical validation of this form of remote music education solution. Within this evaluation, this study novelly identifies the potential for INMP experience to vary both with auralised acoustics, and with monitoring context. This contributes to an understanding of INMP practice, from which practical recommendation can be derived. Particularly it is prudent in INMP system design and operation to allow for individualised control of immersive audio rendering (to cater to variant monitoring contexts), particularly with respect to parametric control of rendered acoustics, such that auralisation of acoustics may be optimised for different monitoring contexts, and cases which may result in negative rating of performance experiences may be avoided.

Notably results relating to perception of reverberation contribute the suggestion that in the case of low–latency INMP assumption of improved experience of latency under reverberant conditions, as suggested in previous study, may not be accurate, and in fact the inverse may be true for acoustics with high RT and low C80. Additionally this study provides a novel evaluation of performer engagement with video display in INMP. Here results contribute a demonstration that performers do indeed engage with video display, at least in the case of low–latency INMP, and contribute a suggestion that this video display engagement may be dependent on visibility of visual performance cues, and practice using the INMP systems under consideration.

Chapter 5

Evaluation of an XR INMP System for Remote Virtual Production

> Who wants the machine to show them an alternate reality?
>
> *Professor Hubert J. Farnsworth, Futurama S04E03 (1999)*

5.1 Preface

This chapter describes the evaluation of an XR INMP system design for the practical application of remote virtual production. Here the virtual production case is simulation of a recording studio environment to facilitate networked performance. This is motivated by the desire to facilitate remote performance in the wider field of remote music recording, which may be used alongside existing pro–audio remote recording solutions. Detail is briefly provided on practical usage of the system in–application. A following study is then detailed, where the system is reproduced for evaluation by a professional rock band. An overview of the study design is provided, describing the motivation, objectives and hypothesis, including a brief summary of relevant background literature. Technical description of the INMP system reproduction and related acoustic measurements is given, followed by a description of evaluation methods. Results of evaluation are then presented and discussed, followed by a summary of key findings and conclusions of this study.

The INMP system presented in this chapter was implemented and operated by a project team from the AudioLab, Univeristy of York. This team were associated with the MINERVA project[1], led by Professor Helena Daffern and Professor Gavin Kearney. The contributions of members of this project team to the work described in this chapter and the relevant INMP system are acknowledged here.

Acoustic measurement of BBC Maida Vale Studios used in the INMP system (section 5.5) was conducted by the authors of the related publication:

[1] `https://audiolab.york.ac.uk/minerva/`

- G. Kearney, H. Daffern, P. Cairns, *et al.*, "Measuring the Acoustical Properties of the BBC Maida Vale Recording Studios for Virtual Reality," en, *Acoustics*, vol. 4, no. 3, pp. 783–799, Sep. 2022, ISSN: 2624-599X. DOI: 10.3390/acoustics4030047. [Online]. Available: https://www.mdpi.com/2624-599X/4/3/47 (visited on 07/03/2024)

Measurement was performed using MATLAB scripts and a measurement protocol implemented by Gavin Kearney. Additional physical measurement of the space was performed by Jacob Cooper.

The INMP system (section 5.4) was implemented in collaboration with the BBC[2] by the authors of the related publication:

- P. Cairns, A. Hunt, J. Cooper, *et al.*, "Recording Music in the Metaverse: A case study of XR BBC Maida Vale Recording Studios," in *Proceedings of the AES 2022 International Conference on Audio for Virtual and Augmented Reality*, Redmond, Washington, USA: Audio Engineering Society, Aug. 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21842

Audio rendering and networking was implemented by Patrick Cairns with the assistance of Ben Lee and Anthony Hunt. The XR unity application was implemented by Dr. Dan Johnston with the assistance of Patrick Cairns. The virtual model of MV4 was implemented by Jacob Cooper.

Evaluation of the INMP system was conducted by Patrick Cairns with the assistance of the authors of the related publication:

- P. Cairns, A. Hunt, D. Johnston, *et al.*, "Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, Jun. 2023, ISSN: 15494950. DOI: 10.17743/jaes.2022.0086. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22139 (visited on 02/25/2025)

Thanks is expressed to the following individuals for their contribution:

- Gavin Kearney (MINERVA project lead/sound engineer/acoustic measurement).
- Helena Daffern (MINERVA project lead/production).
- Anthony Hunt (network engineer/sound engineer).
- Ben Lee (network engineer/sound engineer).
- Dan Johnston (XR developer).
- Jacob Cooper (acoustic measurement/3–D modelling).
- Gerardo Acosta Martínez (production).
- Andrew Chadwick (AudioLab studio manager).

Thanks is also expressed to the sound engineers from BBC Maida Vale Studios who assisted in the field.

[2] www.bbc.co.uk/



Figure 5.1: Members of the professional rock band *Wolph* captured physically at each endpoint, and represented virtually as avatars using the INMP system.

5.2 An XR INMP System for Remote Virtual Production

Across 2021/22 a collaborative project between the AudioLab, University of York, and the BBC developed a low–latency INMP system design with XR visual display and interface. A prototype of this system, implemented as an extension to the SonoBus NMP streaming software, was deployed over the LAN at BBC Maida Vale Studios and used to perform in the recording of a live version of the single *'Shoulders of Giants'* by the professional rock band *Wolph* (Figure 5.1). Each performer uses a HMD and is able to see other performers represented as avatars in a VR reproduction[3] of BBC MV4, while being able to see their physical instrument using pass–through XR display. Auralisation provides simulation of hearing the networked performance as if located in the 3–D acoustic environment of MV4 as presented visually, and rendered with 3DoF. The system uses a convolution–based virtual Ambisonic approach to interactive rendering of binaural audio utilising SIR resources from a comprehensive acoustic survey of Maida Vale Studios[4] [5]. An object–based[5] peer–to–peer streaming approach is used.

In practical application of networked performance, the INMP system was deployed across LAN endpoints, reporting a latency of approximately 20 ms OWT between performers. In WAN deployment of the system by research technicians over consumer ISP backbones, however, deployment between Glasgow, York and Rochdale was associated with informally reported latency ranging from 25–40 ms, which is coherent with previous practical INMP system latency measurements across the UK from related prior work [7], [8].

[3] Often referred to as a *'digital twin'*.

[4] `https://zenodo.org/records/10020866`

[5] An audio *'object'* is here defined as audio describing a discrete sound source.

5.3 Study Overview

Following practical virtual performance for recording using the XR INMP system, the band *Wolph* participated in a study designed to test the viability of the novel remote performance solution, and further investigate the novel performance experience provided. The INMP system was reproduced and deployed over the LAN at the AudioLab, University of York, configured to recreate performance conditions experienced in practical WAN deployment. The members of the band *'Wolph'* performed using the system reproduction and participated in evaluation of the INMP system. Here evaluation is achieved through exit interview feedback, experience rating, and tempo analysis of performance, where unimpaired synchrony and positive experience rating is considered to infer high performance experience quality, and the viability of the system. This study includes evaluation of immersive audio rendering implemented in the INMP system relative to a baseline mono monitoring system. Additionally evaluation is repeated across three latency levels considered representative of WAN deployment across national ranges. In the following subsections an overview is provided of the motivations for this study, as well as related objectives, hypotheses and statement of ethics.

5.3.1 Motivation

Remote music recording workflows allow for the capture and production of musical media where physical facilities access and the gathering of musicians and engineers is impractical or impossible. Practical examples here include production scenarios, where unfeasible travel is otherwise required by artists or engineers, or education scenarios, where remote workflows allow for collaborative production tuition to be delivered to remote students for whom it is impractical to congregate physically [207]. With respect to the state–of–the–art in modern industry practice, a range of pro–WAN recording and production solutions are available, for example solutions from AudioMovers[6] and Source Elements[7]. Traditional pro–LAN solutions such as Dante[8] have also been employed in experimental WAN deployment [39], [40], [135]. Such remote workflow solutions are in common usage in industry practice, for example Abbey Road Studio reports usage of AudioMovers solutions on over 90% of sessions[9].

The capabilities of modern pro–WAN remote recording and production workflows is, however, limited with respect to real–time interactivity. Here remote synchronous DAW playback monitoring and overdubbing is possible, however, group performance and associated recording, as is typical in traditional in–person recording workflows, is not supported [41], [207]. NMP workflows for remote recording and production are proposed as a method for overcoming the limitations of current pro–WAN solutions and allowing for real–time performance and capture across multiple remote endpoints. Prior practical implementation of such systems has been achieved using NMP streaming solutions [41] or experimental WAN deployment of pro–LAN solutions [135].

[6] `https://audiomovers.com/`
[7] `https://www.source-elements.com/`
[8] `https://www.getdante.com/`
[9] `https://www.abbeyroad.com/audiomovers`

Remote recording workflows using INMP system designs, where immersive audio rendering is included, are proposed for remote collaboration and virtual production of immersive media [131]. INMP system design is, however, more generally relevant to systems which integrate interactive VR and XR visual display and interface systems. Here systems consider display of virtual avatars and environments in HMD–based VR systems [18], [20], [45], [100], [128], [137] and XR display of virtual environments and remote performer avatars using screen and projector–based systems [19], [20], [136], [137]. Volumetric capture and display of performers has also been explored, however, is considered impractical due to the high latency incurred in visual capture and playback [101]. Such VR and XR NMP systems simulate for networked performers the visual experience of cohabiting a real or virtual 3–D environment, providing a novel interaction modality for remote performance workflows [131]. Here immersive audio rendering is considered a requirement for provision of AV coherence in user navigation of real or virtual 3–D environments presented in visual display and interface [25].

Usefully, in the context of NMP workflows, avatar representation of remote performers offers a visual display and interface method which can circumvent the limitations of traditional video systems. For video systems in NMP AV asynchrony is widely considered problematic in networked performance experiences [15], [19], [97]. High video latency here is incurred in video compression codecs required to operate within the constraints of typical WAN bandwidths. Motion tracking and avatar animation systems provide an alternative to video display in NMP. Here motion tracking metadata streaming can be achieved with low latency, low bandwidth overhead, and additionally may benefit from being more robust to experience degradation due to packet loss and other visual display and interface artifacts [13], [19], [136].

Motivated by the potential for XR INMP systems to facilitate networked interactions in remote recording and production workflows this chapter addresses validation of a novel XR INMP system design for the practical application of virtual production. Though this system simulates a recording studio space, and is intended to facilitate performance for virtual production of musical content, it is acknowledged that synchronous remote recording is beyond the scope of this study. Rather the XR INMP virtual production system is intended to facilitate remote musical collaboration, and may be used alongside previously mentioned existing pro–audio remote recording solutions.

Related evaluation is further motivated by the desire to contribute towards a practice–oriented understanding of the experience of the novel XR INMP context. This includes performer experience of novel HMD–based XR pass–through display. Also included is a testing for experience improvement associated with immersive audio rendering relative to mono baseline conditions, as described in related INMP system evaluation [19], [121]. The XR INMP system described in this chapter builds upon prior work in the field of virtual performance from the research output of the AudioLab, University of York. Particularly here the XR INMP system builds upon prior system development for single–user VR singing simulation [208], and system development for acoustic simulation of virtual performance spaces in practical INMP deployments [7], [8].

5.3.2 Objectives

The focus of this study is a practice–oriented testing of the viability of a practical XR INMP system design for the application of remote virtual production, considered in terms of quality of musical performance achieved using the system. The central objective of this study is therefore:

> *To evaluate the quality of musical performance experience using the INMP system, where performance conditions are representative of WAN deployment.*

Secondary to the core focus, this study also explores potential for experience variations associated with immersive audio rendering using the practical INMP design, considered through comparative evaluation of immersive audio and baseline mono audio rendering conditions. Additionally, this study tests performer engagement with novel XR visual display in the context of INMP. As such the following sub–objectives were defined:

- *To evaluate the quality of musical performance experience using the INMP system for immersive audio and baseline mono audio rendering methods.*
- *To evaluate the degree to which performers engage with XR visual display using the INMP system.*

5.3.3 Hypothesis

With respect to the primary objective of this study, the following hypothesis is considered:

> *Positive rating of the performance experience, and unimpaired performance synchrony, will be achieved using the INMP system under evaluation where performance conditions are representative of WAN deployment.*

With respect to sub–objectives of this study regarding comparison evaluation of immersive audio rendering relative to a mono baseline audio rendering system, and performer engagement with XR visual display, the following hypotheses are also considered:

- *Rating of the performance experience and performance synchrony will vary between baseline mono audio rendering conditions and immersive audio rendering conditions using the INMP system.*
- *Positive rating of performer engagement with XR visual display will be achieved using the INMP system.*

5.3.4 Statement of Ethics

Ethical approval for this study was obtained from the University of York School of Physics, Engineering and Technology Ethics Committee, reference *Daffern300621*.

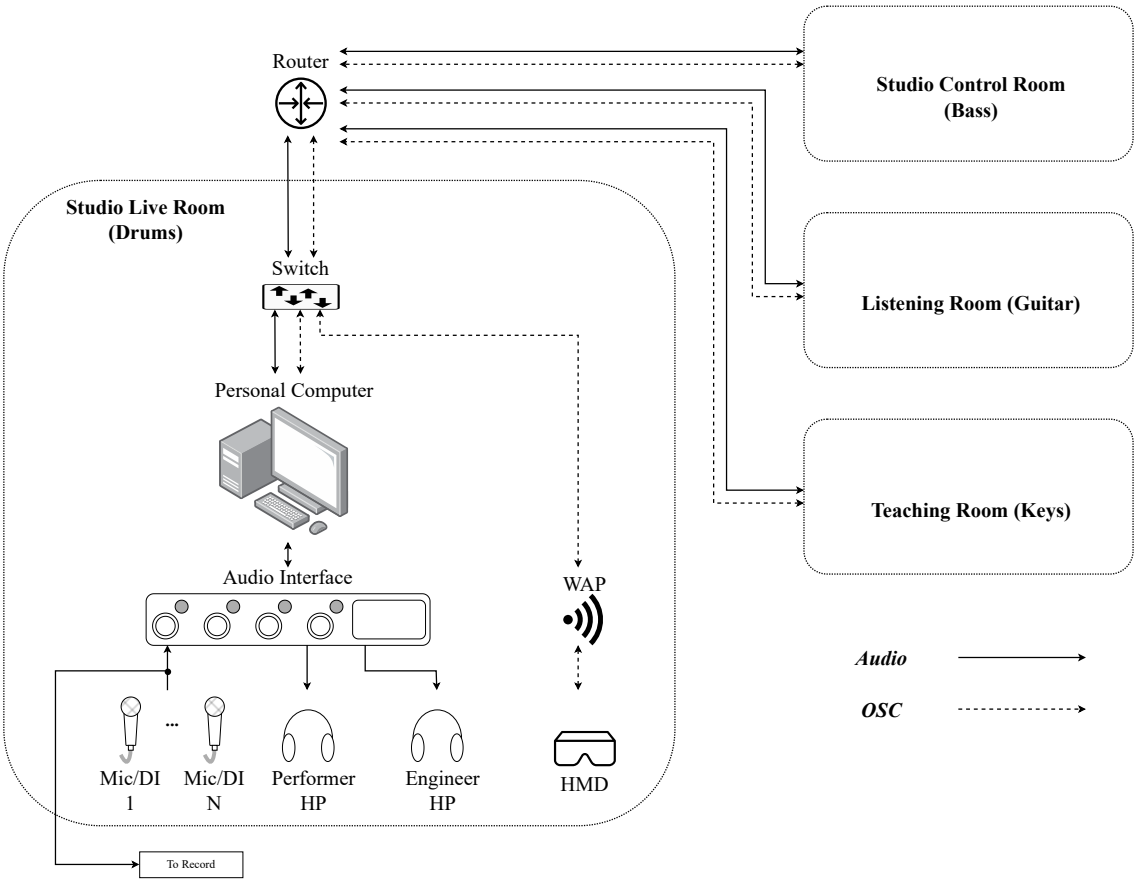


Figure 5.2: Hardware configuration for reproduction of INMP in this study. Each endpoint is configured with the same basic hardware setup. Signal flow is detailed for both audio and OSC tracking metadata.

5.4 System Overview

The XR INMP system was reproduced and deployed across four endpoints on the LAN at the AudioLab, University of York (Figure 5.2). System reproduction included minor modification for use in the study. This includes additionaly artificial delay used to configure latency conditions to the same range recorded in WAN deployment of the INMP system (as discussed in section 5.2). Each discrete endpoint is associated with a discrete physical space and musician, namely the Studio Live Room (*drums*), Studio Control Room (*bass*), Teaching Room (*keys*) and Listening Room (*guitar*) at the AudioLab. At each endpoint the performer is presented with an engineer–operated audio system, providing audio capture, low–latency streaming, immersive audio rendering and audio playback for musician monitoring. Each performer is also presented with an XR system, providing motion tracking and relevant metadata streaming, and visual display and interface via a HMD. The following subsections provide technical detail on the INMP system implementation.

5.4.1 Hardware Overview

Hardware situated at each endpoint follows the same basic configuration. Microphone or DI capture channels for the local performer and an engineer communications microphone channel are input to the audio interface at each endpoint (*Focusrite Scarlet 18i20 g3* and *2i2 g3* models, *48 kHz, 128–sample frame size*). Output from the audio interface at each endpoint is the local performer headphone monitor mix (*Beyerdynamic DT770*).

An additional engineer headphone output is used to monitor the musician monitor mix and private engineer communications. The audio interface at each endpoint is connected to a laptop computer (*i7- 11800H, 2.3 GHz, 32GB RAM, Win10*) which hosts low–latency NMP streaming software and immersive audio rendering processing. The laptop computer at each endpoint is connected to a private LAN.

Audio capture at each endpoint is specified depending on the discrete relevant instrument. In the Studio Live Room drummer vocals (*Shure SM58*) and drums (*Shure Beta 52a, Shure SM57, Rode M5, Sennheiser e604, AKG C414*) are captured via microphone. In the Studio Control Room the bassist's vocals are captured via microphone (*Rode NT2a*) and bass synthesiser is captured via DI. In the Teaching Room the keyboard player and lead vocalist's voice is captured via microphone (*Neumann u87*) and electric keyboard is captured via DI. In the Listening Room the guitarist's voice is captured via microphone (*Neumann u87*) and electric guitar is captured via DI. Audio capture channels at each endpoint are also passed to the Studio Control Room console via analogue lines in order to provide controlled recording of performance across endpoints.

At each endpoint the performer is also presented with the XR system, consisting of a HMD, providing motion tracking and XR display, and tracking networking processes hosted on the local laptop device. The HMD device at each endpoint is connected to the Wireless Local Area Network (WLAN) on the 5GHz band via a Wireless Access Point (WAP) (*ASUS RT–AX 55*).

5.4.2 Software Overview

At each endpoint the INMP system is hosted across the local laptop device and the HMD. The HMD device hosts an XR application developed in Unity. This application provides XR visual display, motion capture with 3DoF interactivity, and streaming of motion capture metadata between the HMD and the laptop device at each endpoint. The laptop device at each endpoint hosts a range of software used in the audio system, and in motion capture metadata streaming. Audio interface driver software (*Focusrite Control*) allows for control over interface capture and playback routing. JackTrip [75] is used as the NMP software which provides low–latency audio streaming between endpoints. The DAW solution Reaper[10] is the audio processing application which hosts plugins used in immersive audio rendering processes. JACK[11] is the audio routing application, used to route audio between audio applications and audio interface capture and playback buffers.

[10] `https://www.reaper.fm/`

[11] `https://jackaudio.org/`

Endpoint	N Ch. Up	Up BW	N Ch. Down	Down BW
Studio Live Room (drums)	30	23.04	10	7.680
Studio Control Room (bass)	9	6.912	10	13.056
Listening Room (guitar)	9	6.912	17	13.056
Teaching Room (keys)	12	9.216	16	12.288

Table 5.1: Upload and Download BW requirements for each endpoint measured in Mbps. N Ch. Up and N Ch. Down denotes the total number of upload and download channels respectively for each endpoint, where 16–bit, 48 kHz uncompressed audio is considered (for a per–channel BW cost of 768 kbps). Up BW and Down BW denote upload and download total BW requirements for each discrete endpoint. The instrument associated with each endpoint is also noted.

5.4.3 Audio Networking

At each endpoint a discrete instance of JackTrip (specifically version 1.3[12]) is used to form a peer–to–peer connection with each other discrete endpoint, providing low–latency audio streaming between endpoints.

JackTrip instances are configured for a jitter buffer size of four frames, and FEC redundancy of two frames. Latency measurement characterising JackTrip in this configuration is described in subsection 4.4.3, where an OWT network latency is estimated as $13\ ms \pm\ 1.5\ ms$ between endpoints. Each endpoint streams to each other endpoint local performer capture channels and an additional technician communications microphone channel. The total number of channels upload or download at each discrete endpoint therefore varies across endpoints, depending on relevant sent or received audio channels. Upload BW requirement ranges from 6.912–23.04 Mbps, and download BW requirement ranges from 7.68–13.056 Mbsp (Table 5.1). At each endpoint received audio channels describing performer voice and instrument capture are routed to Reaper inputs, and received technician communications channels are routed to audio interface hardware outputs.

The reproduced INMP system design notably substitutes the NMP software SonoBus[13] for the NMP software JackTrip. In the application of remote virtual production, SonoBus provides several useful features. Firstly SonoBus can exist as a VST plugin in–DAW, negating requirement for usage of audio routing applications such as JACK, and providing a familiar interface for audio engineers who may not be familiar with NMP technology to work with. SonoBus also provides Opus compression capabilities, allowing for usage in narrow–band networks, particularly considering the object–based streaming approach used and associated BW requirements for uncompressed audio, and provides other general studio utility, such as backing track playback. As discussed in subsection 4.4.3, however, JackTrip can be convenient for usage in system reproduction in evaluation protocols. Here JackTrip allows for CLI operation such that the INMP system may be launched and configured from shell script. Both SonoBus and JackTrip follow similar standard low–latency NMP streaming design, and interchanging of the NMP software is not expected to influence system performance.

[12] `www.github.com/jacktrip/jacktrip/releases`

[13] `www.sonobus.net/`

5.4.4 Tracking Metadata Networking

At each endpoint performer motion tracking is captured via the Unity[14] XR application on the performer HMD. This metadata is streamed over the WLAN from the HMD to a bridge application instance hosted on the local laptop device. Both the Unity XR application and bridge application (also developed in Unity) were developed by Dr. Daniel Johnston, University of York. The local instance of the bridge application is responsible for forwarding local tracking metadata to immersive audio rendering processes hosted on the local laptop device, and to an instance of the bridge application at each remote endpoint for use in avatar animation at remote endpoints. The local instance of the bridge application receives tracking metadata from each remote performer and forwards this to the local performer HMD, where the Unity XR application applies this metadata in animation of relevant remote performer avatars.

The bridge application allows for input of IP address and port of each endpoint used in motion tracking networking (the local HMD, local immersive audio rendering, and remote bridge application instances). By default each instance of the XR application on the HMD receives metadata on port 6000, each instance of the bridge application receives metadata on port 9000, and immersive audio rendering processes at each endpoint receive metadata on port 8000. These port choices are arbitrary. Metadata is transported over networks and between applications on laptop devices as OSC messages using the OSCJack[15] Unity implementation of the OSC protocol [102]. Each bridge application instance passes tracking metadata to immersive audio rendering in relevant OSC format, namely yaw, pitch and roll messages used by IEM SceneRotator VST[16]. In instances of the bridge application and XR application motion tracking metadata is refreshed to most recent received values at frame–rate, such that metadata streams are effectively resampled at these endpoints in network transport. This effectively continuously handles overflow or underflow in metadata receive processes associated with clock drift between endpoints, and serves to minimise clock drift and maintain approximate synchronisation of motion tracking across endpoints in the INMP system.

5.4.5 XR Application

The Unity XR application, hosted on the HMD at each endpoint, provides each performer with display of a virtual environment which simulates the physical environment of MV4. The MV4 environment model (Figure 5.3) was created by Jacob Cooper (University of York) in SketchUp[17] and imported to Unity for use in the XR application.

[14] `www.unity.com/`
[15] `www.github.com/keijiro/OscJack`
[16] `https://plugins.iem.at/docs/osc/`
[17] `www.sketchup.com/`



Figure 5.3: Physical space of MV4 (left) and virtual environment model of MV4 (right).

Each performer is placed at a discrete set position in the virtual reproduction of MV4, and is able to see remote performers represented as avatars in the virtual environment (Figure 5.1). The floor plan of MV4 and performer positions in the virtual space are detailed in Figure 5.4. Positions PA03, PA11, PA15 and PA23 are used for the *bass*, *drum*, *keys*, and *guitar* performers respectively. Only avatar head movement is animated, responding to remote performer motion capture metadata received from the relevant bridge application instance. Each performer's avatar is personalised based on the performers' physical appearance using Ready Player Me[18].

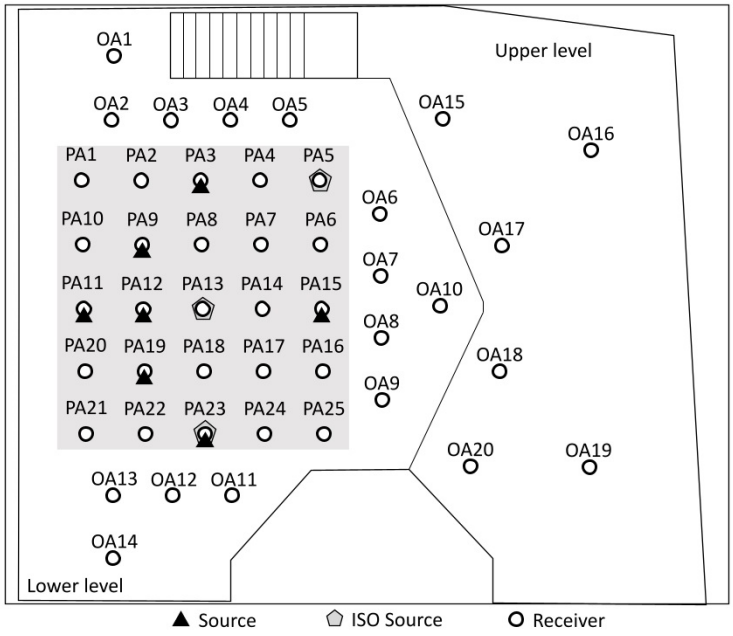


Figure 5.4: Floor plan of MV4, detailing performer positions PA03, PA11, PA15 and PA23.

[18] `https://readyplayer.me/`



Figure 5.5: Pass–through display in the XR application, here showing passthrough view of the drummer.

In addition to virtual environment and avatars in HMD display each performer is able to observe their physical environment using pass–through display implemented with the *Oculus Passthrough API*[19]. Each performer is able to create a rectangular window through which pass–through display is visible, and which may be re–sized and re–positioned (Figure 5.5). Pass–through display here uses the *Quest 2* tracking cameras. These cameras were not designed for the purpose of video display, and therefore provide a low–resolution (i.e., 'grainy') video display, and provide only black and white video with no colour.

5.4.6 Immersive Audio Rendering

The INMP system uses a virtual Ambisonic approach to immersive audio rendering of binaural stereo audio for headphone playback (Figure 5.6). Here each performer is provided with a simulation of hearing networked performance as if listening together with remote performers in MV4, matching the perspective and environment presented in XR display and interface. Immersive audio rendering is performed using a combination of parametric and SIR convolution approaches, hosted as audio plugins in Reaper. Audio routing in immersive audio rendering is configured across audio interface driver software (defining hardware capture and playback buffer routing), internal routing in Reaper (defining routing between audio plugins), and JACK (defining routing between audio applications).

At each endpoint local performer channels are available from audio interface inputs, and remote performer channels are available from JackTrip receives. Here each discrete channel represents a discrete sound source. Auralisation of direct sound for remote performer channels and local performer instrument channels is achieved using Ambisonic encoding methods [27], where the relative direction of each instrument is defined as a static value relative to the relevant listening position.

[19] `https://developers.meta.com/horizon/documentation/unity/unity-passthrough/`

This is implemented in Reaper using an instance of the Encoder VST from the Kronlachner Ambix plugin suite[20] [199] to provide free field auralisation of each sound source as a 3OA signal. These 3OA signals are then summed into a single 3OA bus representing direct sound for all remote performer channels, and local performer instrument channels. The local performer voice channel is treated separately in direct sound auralisation. Here dual mono monitoring of own voice direct sound is achieved using audio interface direct monitoring, where the signal is summed into the musician monitor mix at the audio interface outputs. This allows for own voice to monitored with zero latency.

Auralisation of reverberation for all local and remote performer channels is achieved through convolution with measured 3OA SIR. For vocal channels, and for *guitar*, *bass* and *keys*, each channel is associated with a discrete instance of the X–MCFX convolver plugin, which performs convolution with SIR describing the relative source and receiver positions. For auralisation of the drum kit each part in the kit is auralised as a virtual source positioned across a 2–D vertical triplet of source positions, namely positions PA–09, PA–12 and PA–19, Figure 5.4. These three measurement source positions describe the extremities of a drum kit, with PA–12 located at the kick drum position, and with PA–19 and PA–09 located at the left and right overhead positions respectively. This is achieved using three corresponding instances of X–MCFX convolver[21] performing convolution with measured 3OA SIRs. The 3OA signal output from convolutions is summed into a single 3OA signal representing reverberation components of auralisation.

Each SIR used in auralisation of reverberation is pre–processed to remove direct sound components such that only reverberation is described in SIRs, and where the start of each SIR is truncated to the onset of the first reflection. In order for reverberation onset to be correctly time aligned relative to direct sound onset it is required that a delay is introduced to the Ambisonic signal describing auralised reverberation. This is achieved by allowing for a large first partition in convolution, which also serves to alleviate computational cost of real–time convolution processes. The delays incurred here are based on performer position (Figure 5.4) and trigonometric calculation of path distance between the direct sound path and ground reflection path for each performer, where the speed of sound is used to calculate time offsets. For remote performers an additional 128 samples is incurred in convolution (2.7 ms). This approximates the approximately 1 m path difference between direct sound and ground reflection between any two positions. For a performer's own voice the 3 m path difference between direct sound and ground reflection is approximated incurring an additional 512 samples (10.7 ms at 48 kHz).

The 3OA signal describing auralised direct sound is summed with the 3OA signal describing auralised reverberation to provide a 3OA signal describing the full auralised scene from the perspectice designated to the relevant local performer. An instance of IEM SceneRotator VST[22] is used to provide Ambisonic rotation [27], [183] for 3DoF interactivity in rendering. This plugin instance receives yaw, pitch and roll tracking metadata from the XR system local bridge application instance, received as OSC messages as described in subsection 5.4.4.

[20] `https://github.com/kronihias/ambix`

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

[22] `https://plugins.iem.at/docs/scenerotator/`

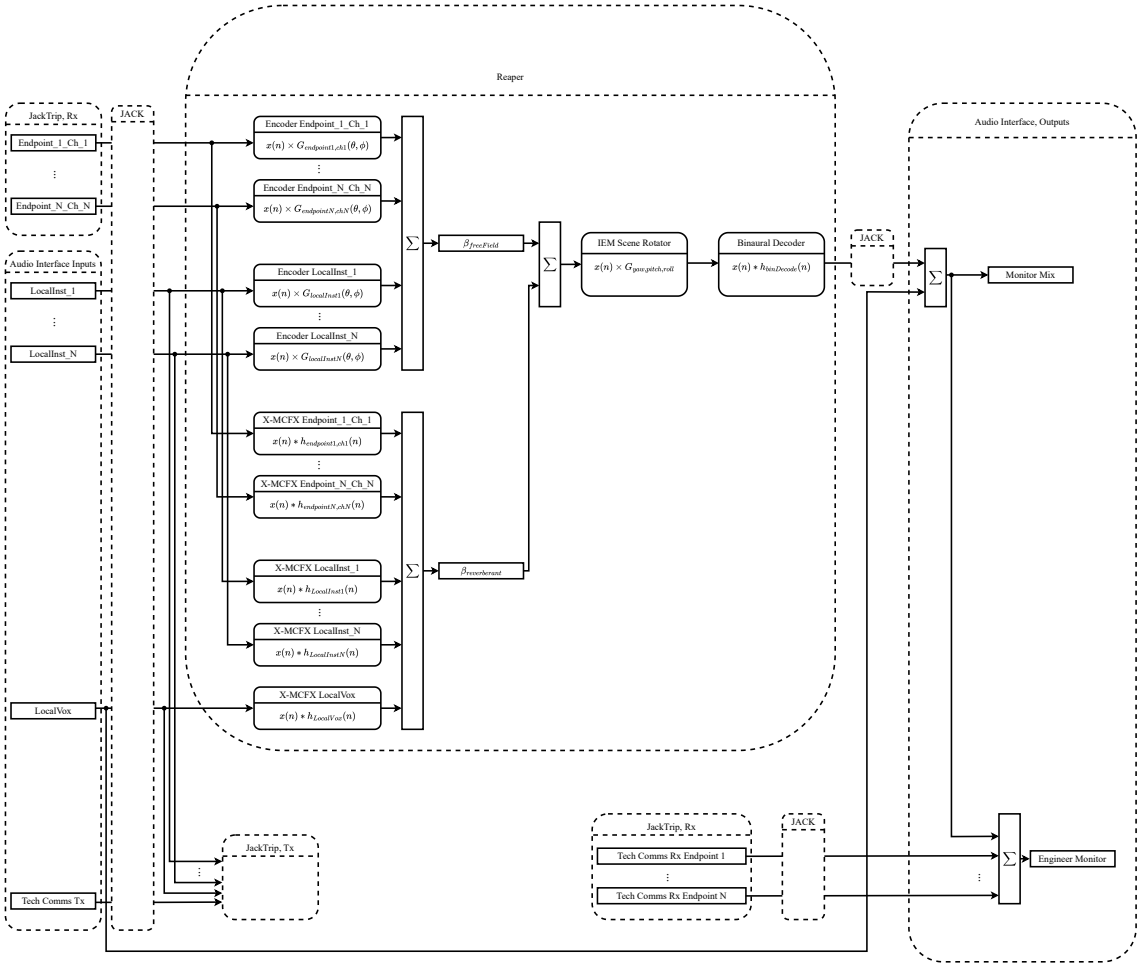


Figure 5.6: Immersive audio rendering signal flow for the INMP system, providing an overview of audio routing and processing. Relevant audio applications and audio plugins are detailed.

The rotated 3OA scene is decoded to binaural stereo format using an instance of Ambix Binaural Decoder from the Kronlachner AmbiX plugin suite [199]. Resources from the SADIE II database[23] [169] are used to perform binaural decoding with Ambix Binaural Decoder, namely HRTF measurements from the Neumann KU100 for virtual loudspeaker binaural decoding [177] for a 26–point Lebedev grid. Binaural stereo auralisation is then output via the audio interface for monitoring by the local performer. Approximately 1 ms latency is incurred in binaural decoding, resulting from HRTF filters not being minimum phase[24].

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

[24] As discussed in [65] the convolution plugins used do not incur processing latency, however, do incur latency as a result of convolution with zeros (or *'dead air'*) before the onset of signal described in IR, which are, in practice, included in HRTF measurements to preserve the natural attack of signal and the ITD.

	RT (T30), s	EDT, s	CT, s	C80, dB	D80, %
mean	0.3726	0.3646	0.0237	13.7945	95.6864
std. dev	0.0101	0.0584	0.0060	1.9654	1.5723

Table 5.2: Mean and standard deviation of A–weighted acoustic parameters from acoustic measurement of the physical space of MV4.

Implementation of immersive audio rendering in–DAW allows for adjustment of direct and reverberation levels according to engineer taste, and according to feedback from performers in sound–check. Considering this it can be noted that DRR deviates from authentic reproduction of the experience of listening in MV4. This approach alternatively allows modification which may be more appropriate for listener preferences in binaural monitoring of Ambisonic auralisation [144]. Additionally it is possible to apply typical production effects to audio monitoring, for example, dynamic processing and EQ of audio capture.

5.5 SIR Measurement

SIR used in auralisation are sourced from a dataset of acoustic measurements[25] performed in the physical space of MV4 [5]. SIR measurements used in the INMP system were performed using the sine sweep method [148], where an overlapping–sweep approach is used for simultaneous measurement of multiple source positions [5], where 20 s overlapping sweeps were used. *Genelec 8040 A* loudspeakers[26] were used as sources in measurement, with loudspeakers calibrated to 85 dBC at 1 m distance.

Receiver positions were measured using an *MH Acoustics em32 Eigenmike*[27] from which 3OA SIR used in the INMP system were derived. A receiver height of 1.6 m was measured for positions PA03, PA11 and PA15, whereas a receiver height of 1.2 m was used for the seated drummer position at PA11 (Figure 5.4). Source heights of 1.5 m (measured from floor to loudspeaker tweeter) were used with the exception of source measurement for positions PA09, PA11 and PA19 (the seated drummer and overhead positions respectively), where a source height of 1.2 m was set. For PA12 (describing the kick drum position) the loudspeaker was placed at floor level.

International Standards Organisation (ISO)–3382 [141] standard acoustic measurement of the space captured with an omni–directional reference microphone (*AKG CK-77*) are available from the online dataset. Acoustic parameters (Table 5.2) were estimated from averaging of the 18 related IR measurements using the MATLAB Acoustic Parameters toolbox from Alex Southern[28] [209]. SIR used in auralisation represent approximation of a hyper–cardiod directivity pattern, achieved through weighting of directional coincident source measurements [5].

[25] `https://zenodo.org/records/10020866`
[26] `https://www.genelec.com/previous-models/8040a`
[27] `https://eigenmike.com/eigenmike-em32`
[28] `https://doi.org/10.5281/zenodo.6505345`

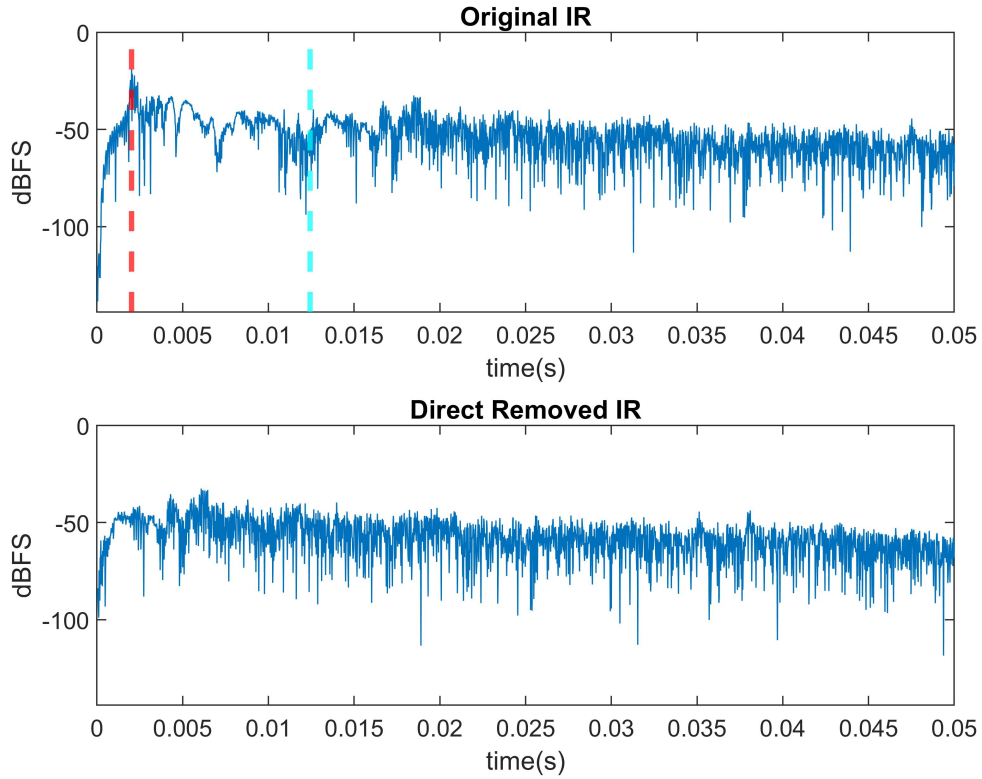


Figure 5.7: Illustration of direct sound removal and truncation of SIR to the onset of the first reflection. The original SIR omnidirectional channel from 3OA measurements (above) and processed SIR (below) are shown. The red line on the plot of the original SIR details the peak of direct sound, and the blue line details the calculated first reflection time.

For each SIR the direct sound component is removed and SIR are truncated to the onset of the first reflection (Figure 5.7). Calculation of the time of first reflection relative to direct sound is here achieved by calculating the in–air time difference between direct sound path from source to receiver, and the shortest reflection path, here assumed to be the ground reflection path. A 2 ms linear fade is applied to the truncated SIR.

5.6 Evaluation Method

The following subsections provide detail on the protocol and methods used to evaluate the INMP system.

5.6.1 Protocol Overview

The XR INMP system is evaluated following a standard NMP performance task protocol, where evaluation is achieved through QoE rating and synchrony analysis. The four participants, members of the rock band *Wolph*, each at a discrete endpoint, used the INMP system to attempt a performance task multiple times.

For each attempt of the performance task audio is recorded for synchrony analysis, and following each attempt of the performance task participants provide rating of the performance experience via a questionnaire response. QoE rating and synchrony analysis results test the core hypothesis of this study, where positive rating and unimpaired synchrony is considered to infer viability for the XR INMP design in–application. Performers also participated in a semi–structured exit interview following completion of the study protocol in order to provide further feedback on the performance experience.

To address the sub–hypothesis that performance experience will vary with audio rendering method, the performance task is attempted for two audio rendering methods:

- *3OA* audio rendering methods, where the virtual Ambisonic system implemented in the practical INMP design is used in musician monitoring.
- *Mono* audio rendering methods, where the virtual Ambisonic system implemented in the INMP system is bypassed and mono monitoring of networked performance is delivered to participants with no reverberation.

With respect to providing evaluation which reflects conditions reported for practical WAN deployment, the performance task is attempted for three latency levels estimated from previous reporting of INMP deployment across endpoints in the UK [7], [8]. OWT system throughput latency levels of 19 ms, 24 ms, and 29 ms are evaluated. For each discrete combination of the three latency levels and two audio rendering methods (6 groups) the performance task is attempted three times for a total of 18 performances, attempted in a random order. Following completion of the test protocol, participants provided feedback on the performance experience in a semi–structured exit interview. Testing of participant engagement with XR pass–through display and the VR reproduction of MV4 is tested through rating of discrete QoE items.

5.6.2 Participants

This study was completed by four expert assessors, namely members of the professional rock band *Wolph*. The band consists of a lead vocalist and electric keyboard player (*keys*), an electric guitar player and backing vocalist (*guitar*), a bass synthesiser player and backing vocalist (*bass*) and a drummer (*drums*). All participants had 1 day of prior experience performing using the INMP system, and had prior experience of performing for music recording in the physical space of MV4 (such that each performer had a controlled reference for comparison of live performance and INMP).

5.6.3 Latency Conditions

Latency levels in evaluation were configured such that evaluation conditions are representative of practical WAN deployment of the XR INMP system. In related prior work [8] a similar INMP audio system was deployed over consumer ISP backbones across Europe, where OWT latency estimates range between 23–55 ms. For UK national deployments OWT latency estimates range between 23–33 ms.

Additionally, informal testing of the XR INMP system design over WAN between researcher endpoints reports sub–30 ms OWT latency. Latency levels evaluated in this study address the range of 19–29 ms OWT. This is considered to be representative of practical WAN deployment across a national range. OWT system latency is estimated here from summing measured network latency of 13 ms (subsection 5.4.3) with audio interface throughput latency of 5.3 ms (48 kHz, 128–sample frame size), and audio processing latency of approximately 1 ms, summing to a baseline OWT audio latency of 19 ms. This value was incremented for 24 ms and 29 ms latency levels using audio plugin instances to add artificial delay to system throughput.

5.6.4 Performance Task

The performance task is a live studio recording performance of the band's single *Shoulders of Giants*. The content of each attempt of the performance task varied to a degree. In order to limit the time taken by the study protocol, and any associated fatigue, performers were not required to complete the performance material for each attempt, and were instructed to end the performance early for some attempts of the performance task. Performers were also instructed to improvise for some attempts of the performance task. Prior to the study protocol performers participated in a sound check and were allowed time performing using the system in order to familiarise themselves with the system. Prior to each attempt of the performance task, participants engaged in a brief clapping exercise together, to allow for any relevant adjustment to performance conditions.

5.6.5 Questionnaire

The questionnaire required that level of agreement was rated for each of the following statements using a 7–point rating scale ranging from strong disagreement (1)–strong agreement (7) via a neutral point (4):

1. We can play together as though in the same room.
2. I feel like I am in the same space as other performers.
3. I have a sense of where other performers are located in the space.
4. Other players sound as though they are playing where I see their avatar.
5. I enjoyed the virtual acoustics during performance.
6. The audio quality was of a professional standard.
7. I am conscious of audio delay between performers.
8. I engaged with the virtual environment and avatars.
9. I engaged with the pass–through display during performance.

Several rating items address standard NMP QoE rating items [6], [114], [118], including rating of perceived performance quality (item 1), BAQ [180] (item 6) perceived latency conditions (item 7). The immersive audio rating item of localisation [180] is addressed generally for the XR INMP experience (item 3) and specific to auditory localisation (item 4). Perceived presence [124], [125] is measured in rating of item 3. Item 5 introduces rating of simulated acoustics, here in terms of enjoyment of performance experience. Rating items 8 and 9 specifically address performer engagement with virtual environment and avatars and pass–through display respectively. These items may also be considered related to low–level interactivity aspects of immersive experience modelling [125]. Each participant provides their response to this rating questionnaire via a Qualtrics[29] survey using a tablet device.

5.6.6 Synchrony Analysis

To provide characterisation of synchrony an excerpt of the performance material was selected for tempo analysis of the drum performance from the kick drum recording. The excerpt selected was from the introduction of the performance. This excerpt was selected as this excerpt was included in each attempt of the performance task, and as the excerpt of the performance material was intended to be performed with steady tempo. The kick drum is selected to characterise tempo of performance as the impulse–based signal allows for error–free and precise algorithmic onset detection.

Evaluation of kick drum tempo is achieved with the MIR Toolbox[30] [210] for MATLAB, which provides onset detection functionality. Performance tempo is derived from onset times as described in subsection 2.8.1. Analysis of tempo variation is conducted with respect to *Tempo Slope*, describing the linear best fit of tempo samples, and characterising the stability of tempo across the audio recording. This allows testing for significant tempo deceleration, which is indicatory of impairment of musical interactivity due to latency [211].

5.6.7 Exit Interview

Following completion of the test protocol the four performers of the band *Wolph* participated in a semi–structured exit interview discussion. Here interview prompts address:

- Participant reaction to the experience of performing using the INMP system.
- Comparison of performance using the INMP system to the experience of performance in the physical space of MV4.
- Participant perception of different audio rendering methods.
- Perceived performance conditions.
- Perception of XRvisual display and interface.

[29] `www.qualtrics.com/`
[30] `https://github.com/olivierlar/mirtoolbox`

- Suggested improvements for the INMP system and performance experience.

These prompts are designed to expand upon concepts addressed in the experience rating questionnaire (subsection 5.6.5), and encourage broader exploratory feedback on the performance experience.

5.6.8 Study Limitations

The exact content of the performance task varies slightly across evaluation, presenting a minor lapse in control of evaluation conditions. Monitoring conditions also varies across participants, with respect to individual participant, instrument performed, audio monitoring and XR visual display content, and auralised listening position in simulation of MV4. This variation in monitoring conditions can be surmised as 'monitoring context'. Whereas this is representative of practical application of the INMP system, variation in monitoring context can be reasonably expected to influence rating across the study. It can be recognised that though audio capture and playback levels may be configured at system deployment these levels were changed during the study. This is a result of engineer production decision and performer requests for monitoring mix adjustments in sound–check. Though this is representative of practice, as desired in this study, it can be acknowledged here that this is also a lapse in control of monitoring level, which may have influenced results.

Motion–photon latency in XR pass–through display and avatar representation of remote performers was not measured for this study. This presents an unknown element and lapse of control in this study. With respect to analysis and interpretation of results it can be acknowledged here that the sample size (four participants) considered in this evaluation is insufficient for any robust statistical analysis.

5.7 Results

The following subsections report the results of evaluation of the INMP system.

5.7.1 QoE Rating

For rating of item 1 (Figure 5.8), addressing perceived performance quality, variation can be observed between participants, and across audio rendering methods within participants. The *keys* performer provides superior rating of *3OA* conditions relative to *mono* baseline conditions. The *drum* performer, however, indicates an inverse preference. Here *mono* conditions are rated as superior to *3OA* audio rendering methods. Both the *bass* and *guitar* performers provide consistently positive rating of both *mono* and *3OA* audio rendering methods. For all participants no overt impact of latency can be observed.

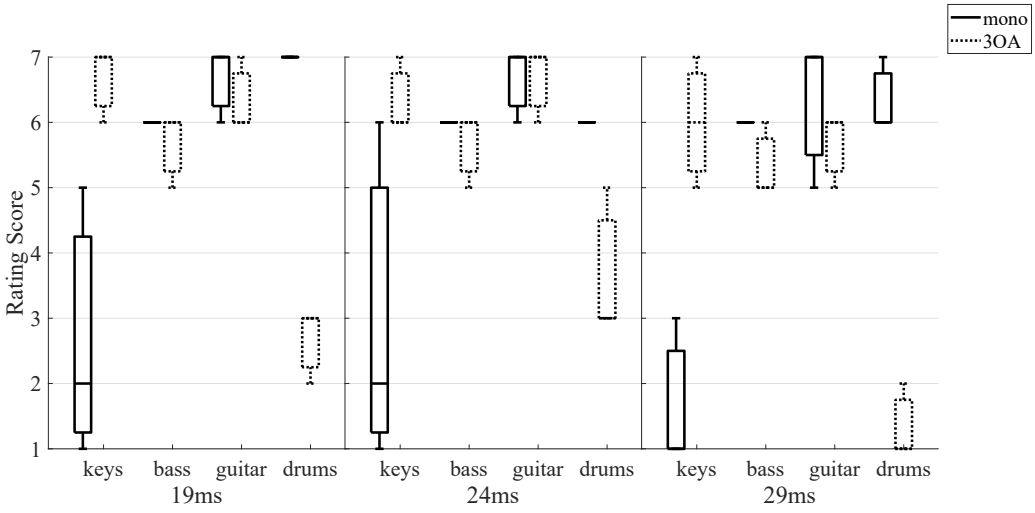


Figure 5.8: Box plot for questionnaire item 1, rating agreement with the statement *'We can play together as though in the same room'*.

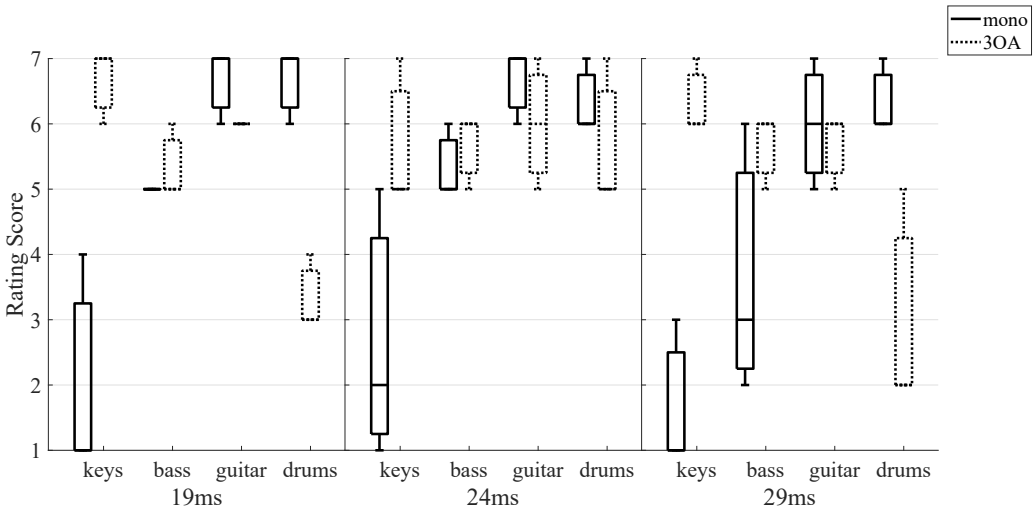


Figure 5.9: Box plot for questionnaire item 2, rating agreement with the statement *'I feel like I am in the same space as the other performers'*.

Rating of item 2 (Figure 5.9), addressing perceived presence, exhibits variation between participants, and across audio rendering method within participant. The *keys* performer consistently rates *3OA* methods positively, whereas *mono* baseline conditions demonstrate a trend for negative rating. The *drums* participant provides inverse response, rating *mono* baseline methods as superior to *3OA* audio rendering methods at 19 ms and 29 ms latency. The *guitar* performer provides consistent positive rating of *mono* and *3OA* methods. The *bass* performer indicates a preference for *3OA* conditions specifically at 29 ms latency.

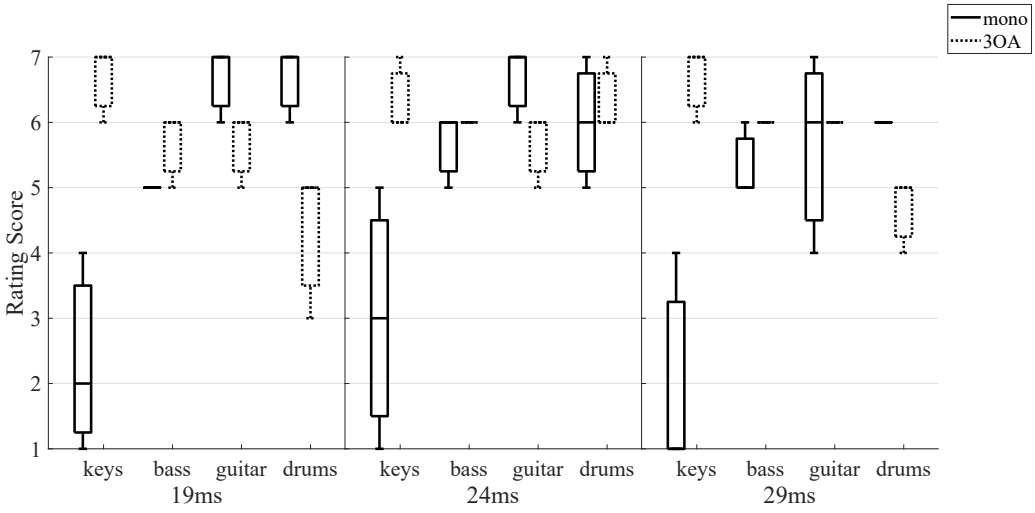


Figure 5.10: Box plot for questionnaire item 3, rating agreement with the statement *'I have a sense of where other musicians are located in the space'*.

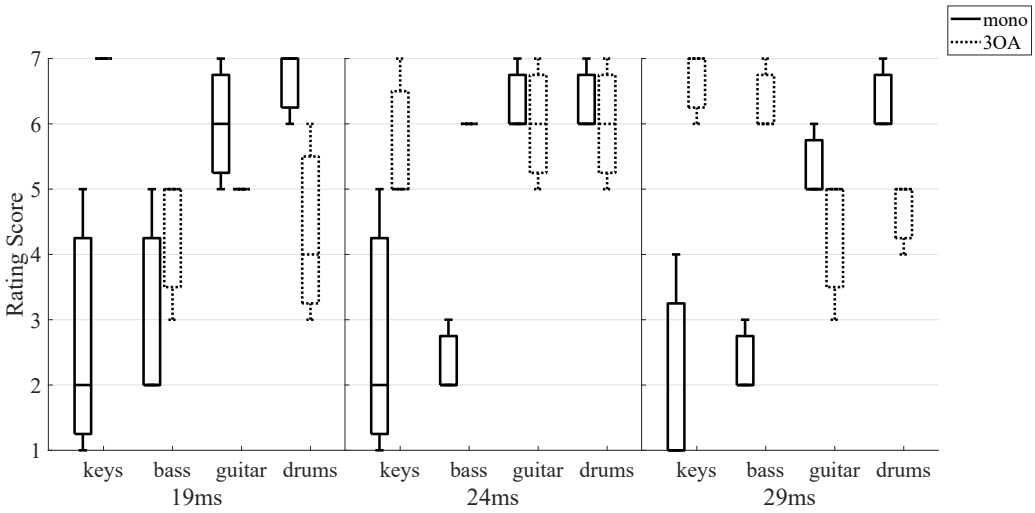


Figure 5.11: Box plot for questionnaire item 4, rating agreement with the statement *'The other players sound as though they are playing where I see their avatar'*.

A variation of rating between participants, and across audio system rating within participants, is observed for questionnaire item 3 (Figure 5.10), addressing localisation. The *keys* performer again consistently rates *3OA* methods as superior to *mono* baseline audio rendering methods. The *drums* performer provides inverse rating, indicating superior perception of *mono* methods relative to *3OA* audio rendering methods. The *bass* and *guitar* performers provide consistent positive rating across all conditions.

Questionnaire item 4 (Figure 5.11), relating specifically to aural localisation, demonstrates rating variation between participants and between audio rendering methods. The *keys* and *bass* performers provide superior rating of *3OA* methods relative to the *mono* baseline. The *drums* performer, however, rates the *mono* baseline as superior to *3OA* methods. The *guitar* performer provides consistently positive rating across all conditions.

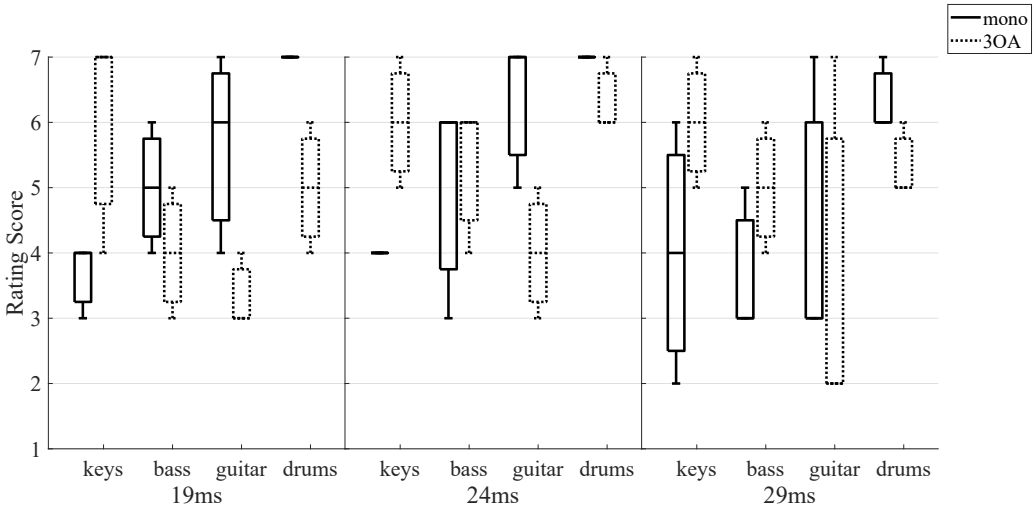


Figure 5.12: Box plot for questionnaire item 5, rating agreement with the statement *'I enjoyed the virtual acoustics during performance'*.

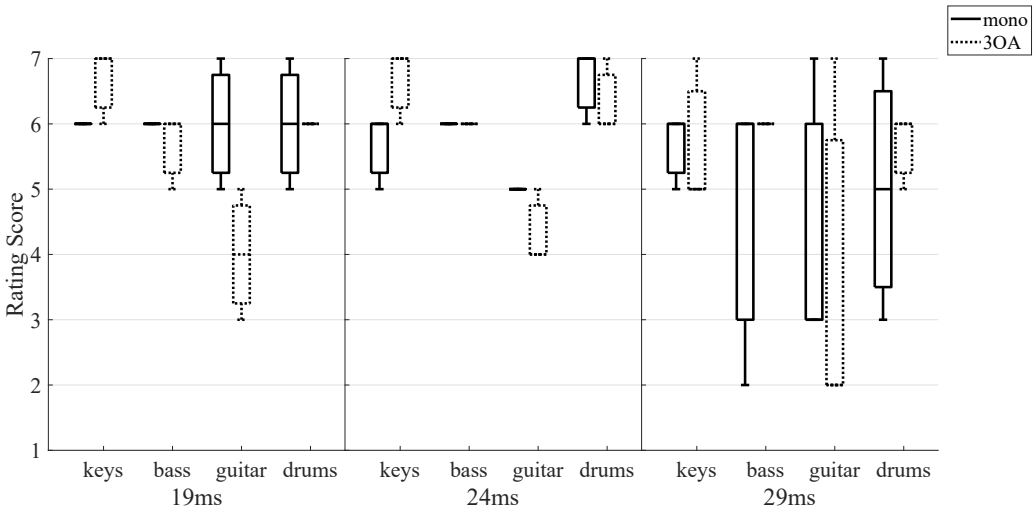


Figure 5.13: Box plot for questionnaire item 6, rating agreement with the statement *'The audio quality was of a professional standard'*.

For item 5 (Figure 5.12, enjoyment of acoustics), variation is observed between participants and across audio rendering method within participants. Here the *keys* performer provides superior rating for *3OA* methods relative to the *mono* baseline. Both the *drums* and *guitar* performers, however, rate the *mono* baseline as superior to *3OA* audio rendering methods. The *bass* performer's ratings are ambiguous here, providing a mix of positive and negative ratings for both the *mono* baseline and *3OA* rendering methods.

Rating of perceived audio quality is given for questionnaire item 6 (Figure 5.13). Generally, a positive response is given across 19 ms and 24 ms latency levels, however, several negative ratings were given at 29 ms latency. With respect to audio rendering methods, the *guitar* performer notably indicates a preference for the *mono* baseline at 19 ms latency.

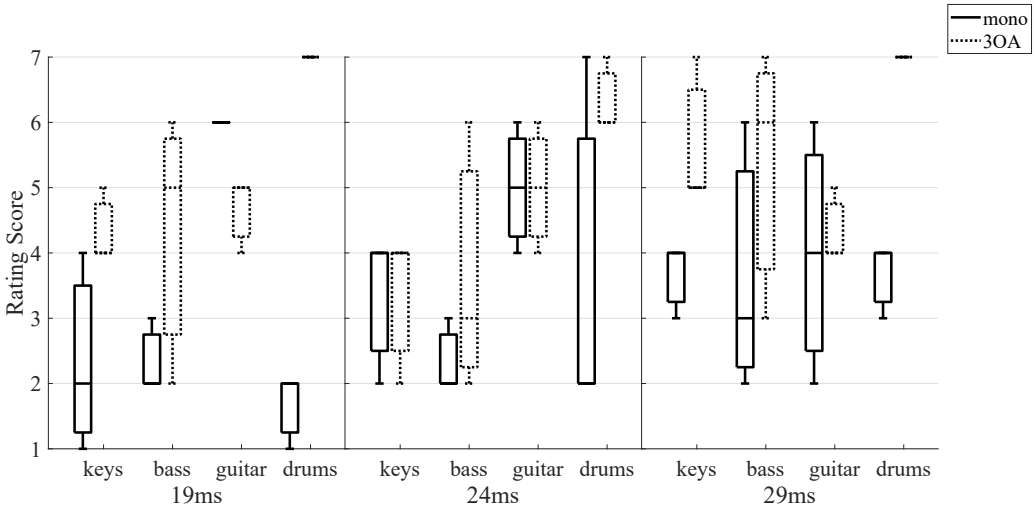


Figure 5.14: Box plot for questionnaire item 7, rating agreement with the statement *'I am conscious of audio delay between performers'*.

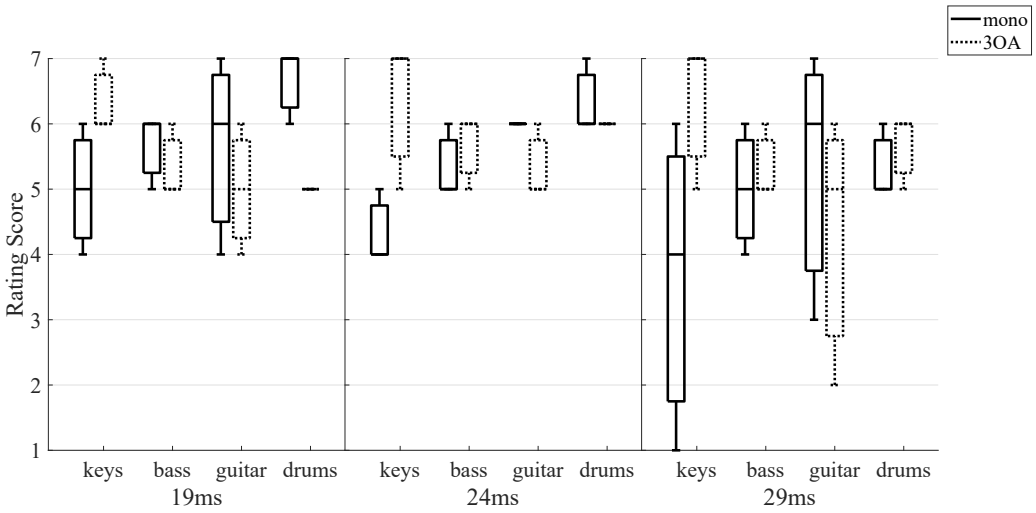


Figure 5.15: Box plot for questionnaire item 8, rating agreement with the statement *'I engaged with the virtual environment and avatars during performance'*.

For item 7, rating perceived audio latency (Figure 5.14), variation is again evident between participants and across audio rendering methods within participants. Notably the *keys*, *bass* and *drums* performers all indicate that they are more aware of audio delay for *3OA* audio rendering methods relative to the *mono* baseline. A general trend is slightly suggested for audio delay to be rated as more noticeable for 29 ms than other latency levels.

Rating of item 8 (Figure 5.15) includes some variation between participants, and across audio rendering methods. Notably here the *keys* performer indicates a preference for *3OA* relateive to the *mono*, baseline. Otherwise it can be highlighted that *3OA* audio rendering methods are generally rated positively by all performers.

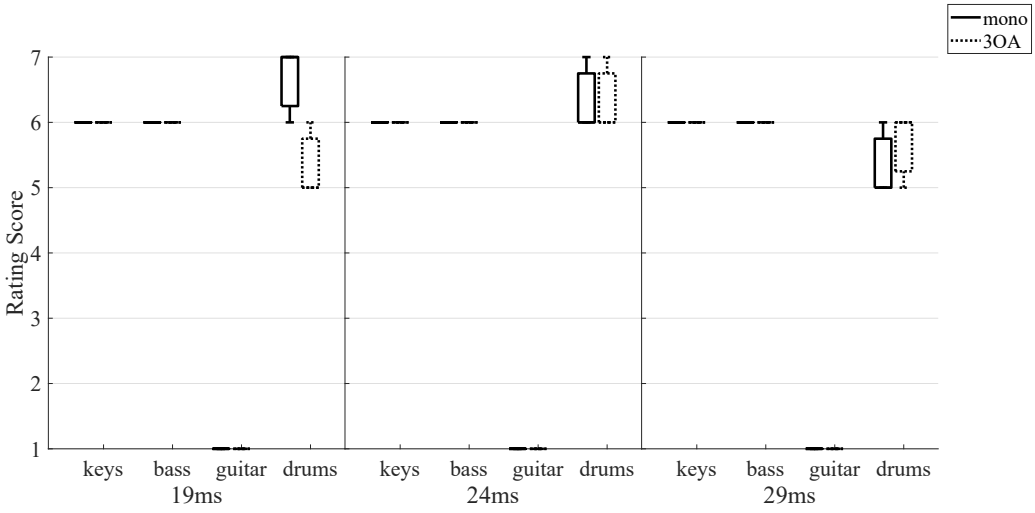


Figure 5.16: Box plot for questionnaire item 9, rating agreement with the statement *'I engaged with pass–through display during performance'*.

Item 9, which addresses engagement with XR pass–through display (Figure 5.16), is rated consistently from agreement–strong agreement by the *keys, bass* and *drums* performers across the study. The *guitar* performer, however, provides consistent rating of strong disagreement across the study.

5.7.2 Synchrony Analysis

Tempo analysis of the recording of the kick drum was performed on an excerpt of each attempt of the performance task as detailed in subsection 5.6.6 (Figure 5.17). Starting tempo across attempts of the performance task varied slightly from 107–116 Beats Per Minute (BPM).

Tempo slope (Figure 5.18) evaluation suggests that tempo is more stable under baseline *mono* audio rendering conditions than under *3OA* audio rendering conditions. Tempo slope also appears to be less stable at 29 ms latency than at lower latency levels. Some level of tempo variation is expected in musical performance and is evident in tempo slope measurements. No excessive tempo deceleration, associated with impairment of networked performance due to excessive latency, was suggested. At 29 ms latency level a trend is suggested for tempo acceleration.

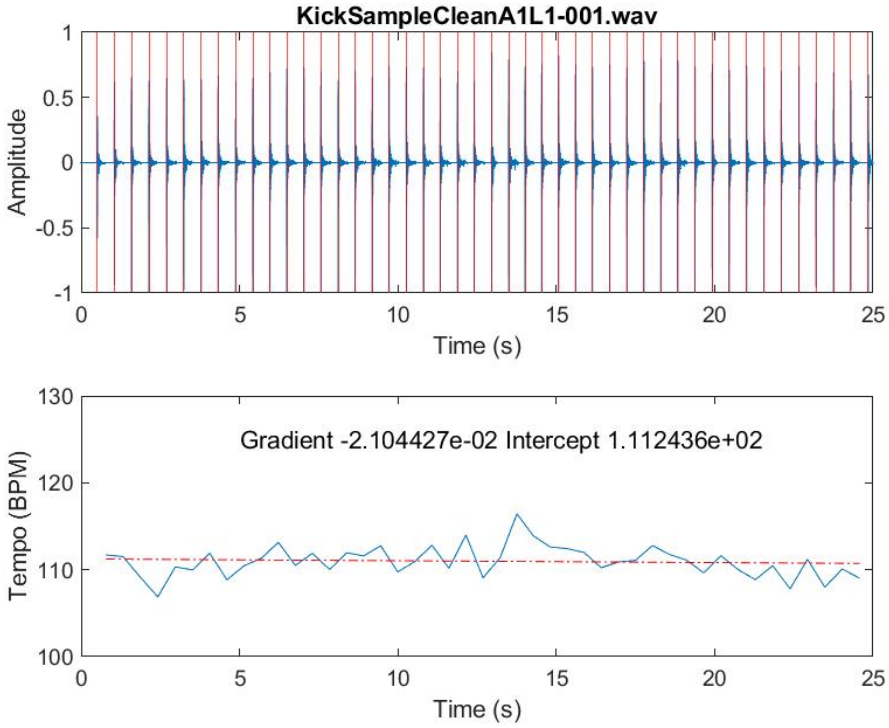


Figure 5.17: Example of tempo analysis of recorded kick drum for the excerpt of the musical performance task. Onset detection is detailed (above, showing the audio recording with onset times indicated by red vertical markers) and associated tempo is detailed (below) including note of starting tempo (intercept) and tempo slope (gradient).

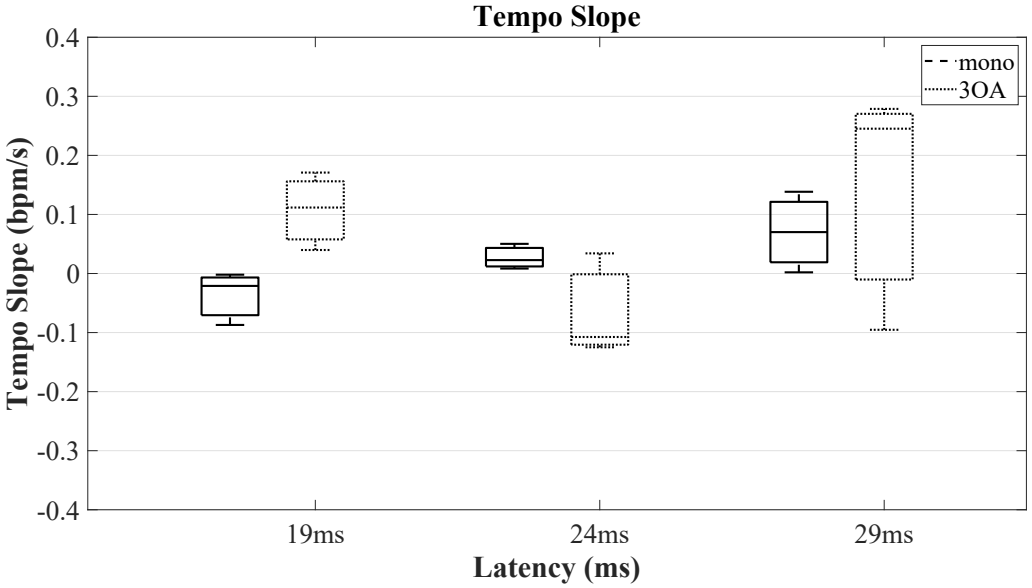


Figure 5.18: Box plot of measured tempo slope.

5.7.3 Exit Interview Response

In discussion participants all indicated some awareness of latency performing using the INMP system. Additionally participants reported awareness of latency variation across the test protocol, and considered the degree to which latency impaired performance conditions to vary across the study.

Participants reported an awareness of audio system variation across the study. The *keys* performer states *'you could tell there were certain takes which were 2D and some which were 3D'*, and indicates a positive response to immersive audio rendering, commenting *'I could hear the drums are set up...like a real drum kit...I think it's really cool.'* The *bass* performer comments on superior localisation under *3OA* conditions, stating *'even (keys), who was off at an angle, went from being, it felt like hard left, kind of behind where their avatar was, to certain takes where it was bang on...it did feel like they were there, the sound was coming from where their actual avatar (was)...sometimes it just felt like a stereo mix.'* The *drum* participant comments on the plausibility of auralisation, and notes that *'it feels like you are in the room listening to something...it has that ambience'*.

Participants also reported improved spatial clarity and source seperation for immersive audio rendering. The *keys* performer comments that *'when you have the separation I know (moog) is over there because I can hear their claps coming from there...if you are a bit out of time you can pull yourself back in, because you can hear if 3 people are in time you can pull back in with them.'* Both the *drum* and *guitar* performers report confusing immersive audio rendering for real sources. The *guitar* performer comments *'when the 3D audio was on, and it all clicked in, and I just instantly turn around to turn it down (a guitar amp), it's quite weird. It's like: 'it's not there is it?'...it sounded like it was on the floor.'* Performers commented that the rendering approach, where electric instruments are localised at performer avatars, was unusual. Participants rather considered that these sources should be localised to virtual amplifier models placed in the virtual environment.

When discussing XR visual display elements of the INMP system performers note several limitations. The *guitar* comments on pass–through video quality, stating *'the cameras on the VR, obviously they are not HD or anything...it makes it a little bit harder'*.

The *bass* performer provides similar feedback, commenting *'you lose that granular detail...playing the moog there's obviously lots of controls...you have to use a bit of memory knowing which is the right button...it's a bit of a challenge'*. The perspective provided in pass–through display is also considered as a limitation. Particularly the *keys* performer considers the 2–dimensional nature of pass–through display as contributing to performance difficulty, and the *drums* performer considered placement of pass–through 'windows' as contributing to performance difficulty, stating *'my window was on the wrong angle'*. The *guitar* performer also notes that as they moved in performance they lost view of their instrument and pass–through display. Here it is commented *'I didn't have a portal on my guitar...it's complete guesswork'*. This was a result of the guitarist moving, such that their guitar had moved away from the static pass–through display window.

Participants also noted limitations in simplified avatar representation, with the *bass* performer noting that *'the other challenge is everyone is severed torsos'*, and notes that visual performance timing cues, such as the drummer striking a drum *'weren't necessarily always available so you had to fall back on what you could hear or pick up on.'*

Despite limitations of avatar representation the *bass* performer reports that *'little bits of visual interaction did positively influence performance.'* Participants note that the shared virtual environment and avatars allowed for interaction which is not typically available when performers are isolated for studio recording. The *drum* performer comments here that *'when I talk to (guitar) I know I'm talking to (guitar)...I've got them engaged...even on a basic level, you know you are looking at each other...it eliminates a large chunk of miscommunications that can happen while tracking live.'* HMD profile was also highlighted by participants. Here the *bass* performer notes that they *'found the headset quite heavy after a while.'*

Considering the wider context of XR musical performance, participants commented that comfort using the INMP system improved with practice. The *keys* participant comments *'The first time...we felt a bit queasy and weird about it, but coming back to it, it did feel a bit more natural'*. With respect to wider improvement of the INMP system, performers expressed several suggestions. Virtual instrument selection was suggested for keyboard instruments. Individual performer control of monitor mixes was also suggested. Participants also expressed a desire for visual effects in the XR experience, and considered expansion of XR visual display and interface to include video elements.

5.8 Discussion

The following subsections present discussion of results of the INMP system evaluation. Included here is discussion of overall system evaluation, and the experience of latency, immersive audio, visual display, and different participants.

5.8.1 Overall System Evaluation

Considering the object of validating the quality of musical performance using the XR INMP system, synchrony analysis (subsection 5.7.2) describes no significant deceleration[31] or tempo instability which is considered indicative of impairment in musical interactivity and performance quality [6]. This demonstrates that synchronous performance is indeed possible using the INMP system under all latency and audio system conditions.

Rating items, considered to provide validation of the quality of performance experience delivered using the INMP system provide mixed results. For the *keys* and *bass* performers, a high quality experience is indicated through generally positive mean rating across latency levels for *3OA* audio rendering methods across rating items addressing performance quality (item 1, Figure 5.8), presence (item 2), localisation (item 3 and item 4,), enjoyment of acoustics (item 5), audio quality (item 6) and visual display engagement (item 8 and item 9). Here it is suggested that for these factors a high quality performance experience was provided, where immersive audio monitoring was delivered. Experience rating for *mono* audio rendering methods, however, includes several negative mean ratings, suggesting that for baseline audio rendering methods a high quality experience was not delivered.

[31] In previous tapping study mean deceleration of approximately 0.2 BPM per second or greater is indicative of impairment of performance quality due to latency [211].

For the *drum* performer positive mean rating is generally provided across latency levels for baseline *mono* audio rendering methods. This includes rating of items addressing the factors of performance quality (item 1), presence (item 2), localisation (item 3 and item 4), enjoyment of acoustics (item 5), audio quality (item 6) and visual display engagement (item 8 and item 9). This suggests that for the *drum* performer a high quality performance experience was delivered for these factors using the INMP system with immersive audio rendering bypassed and providing baseline *mono* monitoring of networked performance. Rating of performance for *3OA* audio rendering methods, however, contains several negative ratings across these factors, suggesting that a high quality experience was not achieved when immersive audio monitoring was provided.

For the *guitar* performer results are more ambiguous. Both *3OA* and baseline *mono* audio rendering methods are generally rated positively across latency factors for items addressing performance quality (item 1), presence (item 2), and localisation (item 3 and item 4). Some negative rating, however, is provided for both *3OA* and *mono* methods in rating of items addressing enjoyment of acoustics (item 5) and audio quality (item 6). It is possible that negative rating of audio quality here may indicate perception of audio artefacts associated with jitter buffer over–run and under–run in NMP streaming software [15], [39]. The *guitar* performer provides generally positive rating of VR elements of visual display and interface (item 8). Experience impairment, however, is clearly indicated in rating of XR pass–through display (item 9). This suggests that the INMP system did not provide a high quality performance experience, and the XR remote performance workflow may be problematic for this participant.

For rating of perceived delay (item 7) all participants indicate some impairment in performance experience related to audio latency for all latency levels and audio rendering methods. Given that unimpaired synchrony and positive rating of perceived performance conditions and quality is provided by all participants, it is considered that this does not reflect an impairment in ability to perform, but more directly an impairment of perceived performance experience quality.

With respect to the central objective of validating the INMP design as a remote performance solution mixed conclusions are drawn. From synchrony analysis and rating of performance quality it is evident that the XR INMP system provides a viable remote workflow for achieving synchronous musical performance.

The quality of performance experience achieved using the INMP system is, however, suggested as conditionally sufficient. For the *keys* and *bass* performer a high quality experience may be delivered where immersive audio monitoring is provided. For the *drums* performer a high quality experience may be delivered where baseline mono audio conditions are provided. For the *guitar* performer results suggest it is not possible to validate a high quality performance experience, where particularly experience of novel XR pass–through display is considered detrimental to the performance experience. Though audio delay was not considered to impair performance quality (as is expected below 30 ms [6]) awareness of latency was highlighted and is suggested as negatively impacting the performance experience.

5.8.2 Experience of Different Latency Conditions

From exit interview discussion it is evident that all participants were aware of latency between performers, and reported perceived variation of latency conditions across the study protocol. Indeed, synchrony analysis does suggest that tempo may be less stable for the 29 ms latency level. Though synchrony is less stable, tempo deceleration and instability associated with impairment of performance due to latency was not apparent in results, nor was any impairment in rating of perceived performance conditions.

Notably for rating item 6, relating to BAQ, some negative rating is provided from the *bass*, *guitar* and *drum* participants at 29 ms latency. This may suggest some deterioration in perceived audio quality relating to latency conditions. QoE rating across other items does not, however, indicate any clear pattern in rating relating to latency level. Notably this includes rating item 7 which directly addresses perception of audio delay. It may be the case that the 5 ms increments and 10 ms latency range in this study are too small to elicit a variation in rating. Study of NMP QoE rating variation in other work [114] would suggest that 40 ms latency variation is associated with variation in rating. Other work suggests that it is typical for other parts to adopt leader–follower strategies to follow drummers [211], and that this may stabilise performance synchrony, and mediate impairment of performance experience associated with latency.

5.8.3 Experience of Different Audio Rendering Methods

With respect to the objective of providing comparative evaluation of immersive audio monitoring relative to baseline mono methods, synchrony analysis suggests no difference in the ability to achieve synchronous performance. A high degree of discernment in audio rendering method is, however, suggested in experience rating and explicitly stated in exit interview response. From related literature describing evaluation of immersive audio monitoring systems for musical performance [121] and for projector–based XR performance [19] it is expected that *3OA* will achieve superior rating than mono or stereo baseline conditions with respect to perceived performance conditions (item 1), presence (item 2), and localisation (item 3 and 4).

The *keys* performer conforms to these expectations across rating items 1–4, indicating a preference for immersive audio monitoring relatively to baseline *mono* methods. The *bass* performer conforms to expectations in rating of auditory localisation (item 4), providing superior ratings of immersive audio monitoring relative to *mono* methods. The *guitar* performer also provides positive rating of immersive audio monitoring across rating items 1–4, however, does not overtly indicate any preference for immersive audio rendering relative to *mono* baseline methods in rating. These three performers all provide coherent responses in exit interview discussion, indicating appreciation of immersive audio monitoring with respect to localisation and presence factors.

The *drums* performer, however, deviates from expectations. Here a preference for baseline *mono* monitoring relative to immersive audio monitoring is indicated across rating items 1–4. Notably this includes localisation. It may be the case that high rating of localisation for item 3 may not specifically relate to aural localisation, and that high rating of *mono* audio rendering in aural localisation (item 4) may therefore be the result of ventriloquism [206] influenced by visual display and interface.

Though the *drums* performer expresses some appreciation of immersive audio rendering in their exit interview response, it can be noted that these comments relate to the plausibility [125] of auralisation rather than performance quality, localisation and presence factors addressed in this study. Preference for *mono* monitoring over *3OA* auralisation is consistent with findings of prior work which report a preference of drummers for dry baseline monitoring over reverberant monitoring conditions [119]. Here the preference is attributed to improved clarity in non–reverberant baseline conditions. Indeed, with respect to variation in monitoring context between participants, it can be recognised that the *keys*, *bass* and *guitar* performers are captured via DI, and the level of sound in the room is well controlled for monitoring. For the *drum* performer, however, a high SPL from the kit may interfere with immersive audio monitoring, and a preference for the clarity of a dry mono mix may be considered reasonable.

In exit interview discussions participants notably comment on the perception of improved spatial clarity [180][32] and plausibility [125] of auralisation for immersive audio monitoring. These factors were suggested as improving the perceived quality of performance experience, and improving ease of synchronous performance, however, were not addressed directly in the rating questionnaire. Indeed subsequent study of projector–based XR INMP has included factors relating to plausibility and spatial clarity, and demonstrates improvement in INMP experience associated with immersive audio monitoring relative to mono baseline monitoring methods for these factors [19].

With respect to enjoyment of acoustics (item 5) discernment between immersive audio and baseline *mono* monitoring methods is somewhat ambiguous, though rating from the *keys* and *drums* performers do suggest some preference for *3OA* and *mono* monitoring methods respectively. In exit interview discussions participants note a desire for individual control of monitor mixes and consider that this will improve enjoyment of monitoring and performance experiences. It is considered that ambiguity in the rating of enjoyment of acoustics in this study may reflect the individual preferences of performers. Achieving a high quality rating of enjoyment of monitoring may be best achieved by affording participants the control required to adjust monitoring mixes to suit individual preferences. Indeed several practical high–latency INMP designs include some form of GUI for this purpose, allowing control of immersive audio mixing [22], and allowing for choice between immersive audio and traditional mono or stereo monitoring methods [104].

For rating of awareness of latency (item 7) it is notable that the *keys*, *bass* and *drum* performers all provide rating which suggests audio delay between performers is more noticeable for immersive audio monitoring than for baseline *mono* conditions. This would suggest that the monitoring method influences the perception of audio delay in the latency–sensitive context of INMP, where immersive audio rendering may influence a more acute perception of delay. Given that the additional latency incurred by immersive audio rendering (approximately 1 ms) is well below any Just–Noticable Difference (JND) [212], however, the reason for this trend in rating is unclear.

[32] More typically described as *'source separation'* [19].

It can be concluded that the experience of performance using the XR INMP system is indeed influenced by audio rendering method. It is suggested that for some participants immersive audio monitoring resulted in experience improvement with respect to perception of performance conditions, localisation, presence, spatial clarity, and plausibility of auralisation. Indeed confidence is given in this conclusion by coherent findings in subsequent relevant studies [19], [121]. It is demonstrated in this study, however, that experience of immersive audio monitoring may be conditional, and that participants may prefer traditional monitoring methods. Particularly, it can be considered that findings relating to evaluation of melodic instrument performance with an immersive audio backing track [121] or a tapping task using an XR INMP system [19] describe a significantly different monitoring context to the practical remote performance scenario addressed in this study, and that assumption of improved experience of immersive audio rendering in ecological use–cases may not be sensible.

5.8.4 Experience of Visual Display and Interface

Rating of performer engagement with the virtual environment and avatars is generally positive with few exceptions. This suggests that participants do indeed engage with these aspects of visual display and interface using the XR INMP system. Participant feedback in the exit interview corroborates this, however, highlights limitations here particularly with respect to avatar representation. Indeed the limitations of avatar representation for delivery of visual performance cues is suggested in another study of VR NMP [128]. Successive research provides further exploration of this aspect of the XR INMP system [13]. This work indicates confirmation that the INMP experience and performer engagement with visual display and interface is here affected by avatar representation, and suggests that inclusion of full upper–torso and facial tracking provides significant improvement.

Rating of participant engagement with pass–through display provides clear demonstration of the usefulness of novel XR visual display in INMP. The exception here is the *guitar* performer, who provides clear indication of difficulty with XR pass–through display. Particularly the profile of HMDs and the perspective provided in pass–through display can be acknowledged as suitable for instruments such as keyboards and drums which can easily fit in pass–through display, however, are impractical for instruments which are not laid out on front of performers, for example, the guitar. Alternative to HMD–based XR, other research considers screen or projector systems for XR display of avatars and virtual environments [19], [137]. It is considered that this solution may provide improved experience of XR display and interface where instruments such as guitar are considered, and achieve improved performer engagement in comparison to HMD–based approaches.

5.8.5 Experience of Different Participants

Across the evaluation of quality of performance experience, immersive audio rendering, and engagement with visual display, a variation in experience across individual participants is demonstrated. In the case of visual display engagement rating variation is clearly relevant to instrumentation, where XR visual display and interface was highlighted as problematic when the instrument did not suitably fit the pass–through display. Rating variation of performance experience and audio rendering method, however, is less clearly attributed. In prior work it is demonstrated that the timbral [108], [114] and dynamic qualities [111] of instrumentation are likely to affect the experience of NMP. Indeed, other work directly demonstrates variation of synchrony across instrumentation in NMP [109]. In this regard it is considered that rating variation between participants may relate to instrumentation. Given the variation in monitoring context between discrete participants in this study, however, rating variation could reasonably be impacted by the numerous factors (as outlined in subsection 5.6.8) which differ across the unique monitoring context provided for each discrete participant. Alternatively rating variation may be attributed to individual preferences of participants.

5.9 Hypothesis Revisited

At the beginning of this chapter the primary objective was presented:

> *To evaluate the quality of musical performance experience using the INMP system, where performance conditions are representative of WAN deployment.*

This objective was achieved by conducting QoE and synchrony evaluation of the INMP system, where performance conditions approximate conditions experienced over WAN deployment in informal testing, and expected from related prior work.

The primary objective was accompanied by two sub–objectives, namely:

- *To evaluate the quality of musical performance experience using the INMP system for immersive audio and baseline mono audio rendering methods.*
- *To evaluate the degree to which performers engage with XR visual display using the INMP system.*

The first of these sub–objectives is achieved by conducting evaluation for both immersive audio rendering, and for mono baseline audio rendering. The second of these sub–objectives is achieved by including evaluation of performer engagement with XR display in QoE rating and exit interview discussion.

The primary objective is summarised in the primary hypothesis:

> *Positive rating of the performance experience, and unimpaired performance synchrony, will be achieved using the INMP system under evaluation where performance conditions are representative of WAN deployment.*

As discussed in subsection 5.8.1, stable synchrony was achievable under all evaluated conditions. Positive rating of QoE was achieved for the *bass*, *drums* and *keys* performers (although the specific conditions under which positive QoE was achieved varied with the performer). The *guitar* performer, however, provides clear rating of experience impairment across evaluation, associated with XR display. The primary hypothesis is therefore only partially accepted, given that positive experience rating was not achieved for the *guitar* performer.

Related to the sub–objectives, the secondary hypotheses were presented:

- *Rating of the performance experience and performance synchrony will vary between baseline mono audio rendering conditions and immersive audio rendering conditions using the INMP system.*
- *Positive rating of performer engagement with XR visual display will be achieved using the INMP system.*

Though synchrony does not vary with audio rendering method, QoE rating did vary between immersive audio rendering and the mono baseline (as discussed in subsection 5.8.3). The first of these secondary hypothesis is therefore accepted. It can, however, be acknowledged that this is with limitations. Namely that the individual preferences in audio rendering was different across performers, and that the sample size in this study is insufficient for any statistical modelling of the impact of audio rendering method.

As discussed in subsection 5.8.4, all participants other than the *guitar* performer indicate unanimous engagement with the XR display and interrface. The *guitar* performer here provides negative rating of engagement with XR display and interface. The second of the secondary hypotheses conditionally accepted, where the acceptance of this hypothesis depends on the participant.

5.10 Summary

The evaluation of the XR INMP system presented in this chapter contributes a practice–oriented validation of the novel system as a remote performance solution, here considering a practical system design from the context of remote virtual production. Though it is demonstrated that synchronous performance is possible under conditions which reflect practical WAN deployment and an ecologically representative performance scenario, it is novelly demonstrated here that the INMP experience will vary across the practical monitoring contexts presented to participants.

This evaluation of immersive audio rendering contributes to an understanding of INMP practice. It is novelly suggested that immersive audio monitoring in XR INMP may contribute to an improved performance experience with respect to immersive audio rating, NMP QoE, and presence–modelling factors. The assumption of improved experience quality with immersive audio monitoring is, however, novelly demonstrated as impractical, where it is instead suggested that traditional mono or stereo monitoring systems may be preferable for particular monitoring contexts. From this the practical recommendation of individual monitor mix control for performers in INMP can be made, allowing for monitoring of networked performance to be adjusted for different monitoring contexts.

A preliminary investigation of performer engagement with XR display and interface in INMP is also contributed, here evidencing that performers did indeed engage with virtual environment and avatar components. Engagement with XR pass–through display, however, is novelly demonstrated as dependent on instrumentation. Summarily, this study contributes a preliminary validation of XR INMP as a practical remote performance workflow, and contributes to a practice–oriented understanding of the experience of immersive audio and XR systems in this novel context.

Chapter 6

Evaluation of an INMP System for Remote Concert Performance

Just remember to sing– S. I. N. G.

Special Agent Gracie Hart, Miss Congeniality (2000)

6.1 Preface

This chapter details the evaluation of an INMP system for the application of a remote concert. Practical use of the system for a remote concert event is briefly summarised. A two–part study is then detailed, including an overview of the relevant research question, motivation and background material. The first study tests the viability of the INMP remote concert workflow through evaluation of the concert event. The performance experience is evaluated by the professional vocal ensemble delivering concert performance, and the audience experience is evaluated by audio professionals attending the concert event. The second study evaluates the quality of networked performance recorded from the concert production, evaluated relative to reference recordings of a live performance. A technical description of the INMP system and related acoustic measurements are given. Evaluation, methodology and results are presented and discussed. The chapter then closes with a summary of key findings and conclusions.

The INMP system presented in this chapter was implemented and operated by a project team at the AudioLab associated with the Scalable Audio Frameworks For Immersive Real–time Experiences (SAFFIRE) project[1] led by Professor Helena Daffern and Professor Gavin Kearney. The contributions of members of this project team to the work described in this chapter and the relevant INMP system are acknowledged here.

[1] `https://audiolab.york.ac.uk/saffire/`

The INMP system was implemented in collaboration with the BBC Singers[2] by the authors of the related publication:

- P. Cairns, T. Rudzki, J. Cooper, *et al.*, "Singer and Audience Evaluations of a Networked Immersive Audio Concert," *Journal of the Audio Engineering Society*, vol. 72, no. 7/8, pp. 467–478, Jul. 2024, ISSN: 15494950. DOI: 10.17743/jaes.2022.0145. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22644 (visited on 02/25/2025)

Acoustic measurements of MV2 were performed by the MINERVA project team as outlined in section 5.1. The immersive audio rendering system was implemented by Gavin Kearney and Tomasz Rudzki with the assistance of Patrick Cairns. Acoustic measurement of Phipp's Hall was performed by Tomasz Rudzki. Audio networking was implemented by Patrick Cairns with the assistance of Andrew Chadwick. The evaluation detailed in this chapter was conducted by Patrick Cairns.

Thanks is expressed to the following individuals for their contribution:

- Gavin Kearney (SAFFIRE project lead/sound engineer/acoustic measurement).
- Helena Daffern (SAFFIRE project lead/production).
- Tomasz Rudzki (sound engineer/acoustic measurement)
- Andrew Chadwick (sound engineer/network engineer/AudioLab studio manager).
- Jacob Cooper (sound engineer/video operator).
- Gerardo Acosta Martínez (sound engineer/video operator).
- Anthony Hunt (production).
- Kim Steele (production).

Thanks is also expressed to *Ex Corde Vocal Ensemble*[3] for performing in the concert event.

6.2 An INMP System for Remote Concert Performance

In 2023 a low–latency INMP system for remote concert performance was designed and implemented as part of a collaboration between the BBC Singers and the AudioLab, University of York.

[2] www.bbc.co.uk/singers
[3] www.excordevocal.com/



Figure 6.1: *Ex Corde Vocal Ensemble* providing networked concert performance using the INMP system. Here Phipp's Hall is shown.

A prototype system, implemented as an extension to JackTrip, was deployed over the WAN between the AudioLab, University of York, and Phipp's Hall, University of Huddersfield. Here the system was used by a vocal octet, *Ex Corde Vocal Ensemble*, with four singers located at each endpoint, to provide networked concert performance as a pre–event for the 2023 Audio Engineering Society (AES) International Conference on Spatial and Immersive Audio. The concert performance was auralised in the acoustics of BBC MV2, and delivered to a live audience in Phipp's Hall and a live–stream audience at the AudioLab.

The system uses an Ambisonics–based approach to immersive audio rendering. Singer headphone monitor mixes provide a stereo–enhanced binaural auralisation of performing together in MV2. The live–stream audience at the AudioLab also monitors the performance auralised in the acoustics of MV2 over headphones, here with a video display of the concert. The live audience in Phipp's Hall experience the local performers live on stage, with remote performers auralised as if on stage alongside the local performers, and with the acoustics of MV2 auralised over a multichannel loudspeaker array.

6.3 Study Overview

Adjunct to the concert event, a two–part study was conducted in order to test the viability of the novel INMP remote concert workflow and further investigate the experience of the practical INMP use–case. The eight members of *Ex Corde Vocal Ensemble* participated in evaluation of the performance experience, providing experience rating and expert feedback across rehearsal and concert performance. Audience members from the live audience in Phipp's Hall and from the live–stream audience at the AudioLab participated in the evaluation of the concert experience, providing experience rating following the concert event.

Following the concert event a listening test was conducted, where participants provided performance quality evaluation of the recorded performance using the INMP concert system relative to the live performance reference recordings. In the following subsections an overview is provided of the motivations for the two–part study, as well as related research objectives and statement of ethics.

6.3.1 Motivation

Remote concert performance workflows provide a means by which remote performers may collaborate for concert performance and deliver concert performance to remote audiences without the requirement for cohabitation of a physical space. Such systems are considered as a means of allowing for collaboration between artists for whom collaboration opportunities are limited by travel requirements [35], and as means of extending the reach and accessibility of concert events for both artists and audiences [20]. The traditional practical use–case of immersive audio in this context is simulating for audiences the experience of listening to concert performance as if together with remote musicians in a physical acoustic space. This includes cases where concert performance from artists at one or more remote endpoints are displayed in venues to live audiences using immersive video or screen–based XR display, and with immersive audio playback over loudspeaker systems [16], [31], [34], [104], [197]. The case of immersive audio production for binaural live–streaming of networked performance is also relevant [133].

Unlike immersive audio playback and streaming systems, where latency is not critical, INMP systems for remote concert support real–time interactivity between endpoints. Here immersive audio rendering has various practical use–cases. INMP is considered to enhance musician monitoring for concert performance in live–streaming of networked performance [23]. INMP systems are proposed as supporting presence in concert experiences where one or more remote musicians are simulated for a live audience as if performing alongside physically present musicians [15], [131]. For VR and XR INMP systems [18], [20] immersive audio rendering provides auralisation and AV coherence in navigation of real or virtual 3–D concert environments [25] shared by networked artists and audiences.

Practical INMP remote concert workflows include VR concerts in virtual environments [18], [20], XR concerts in venues with live audiences and performers [20], live–streaming of networked performance to online audiences with video display [23], [133] and concert systems where novel immersive display and artistic license in immersive audio production delivers novel, interactive concert experiences across multiple physical venues [18], [21], [22]. Practical concert content, however, is limited to performance material without strict synchrony requirements (i.e., ambient, electro–acoustic and experimental performance material), considered as a strategy for tolerating high latency conditions in networked performance.

The case of low–latency INMP, which is suitable for the performance of rhythmic musical styles, however, remains unaddressed. Rather than being limited to novel performance styles low–latency INMP provides a suitable remote workflow for popular music styles which are most relevant to industry practice. Motivated by the potential for remote workflows to improve the accessibility and reach of artist collaboration and concert experiences, and to support novel immersive concert experiences, this study tests the viability of a practical low–latency INMP concert system. Given the novelty of the low–latency INMP concert context, this study investigates the experience of performers and audience members, motivated by a desire to contribute to a practice–oriented understanding of the novel performance context presented which may be utilised to optimise the experience and quality of remote concert productions.

6.3.2 Research Question and Objectives

Central to this study is the research question:

> *Can a low–latency INMP system provide a practical remote concert performance workflow?*

To address this research question the INMP concert system and practical concert event outlined in section 6.2 was used as a case study. The viability of the INMP concert system was tested in a two–part study. In the first part the viability of the remote concert workflow is considered with respect to the experience of performers and audiences, where a high quality experience is considered indicative of workflow viability, and where experience rating provides an investigation of the practical experience of participants in the novel INMP concert event. Here the relevant objective can be surmised:

> *To evaluate the performer and audience experience of INMP concert performance.*

In the second part the viability of the remote concert workflow is considered with respect to the quality of musical performance achieved using the INMP concert system, judged via audio recordings in a listening test. Here high quality rating of musical performance quality, and comparable rating of musical performance quality using the INMP concert system to live reference performance, is considered as demonstration of workflow viability with respect to the ability to support high–quality concert performance. Here the relevant objective can be surmised:

> *To evaluate the quality of musical performance from audio recording for live performance and for performance using the INMP concert system.*

6.3.3 Statement of Ethics

Ethical approval for this study was obtained from the University of York School of Physics, Engineering and Technology Ethics Committee, reference codes *Cairns20230728* and *Cairns20122023* for the first and second studies respectively.

6.4 System Overview

The INMP concert system was deployed over a WAN connection between two endpoints, namely the studios at the AudioLab, University of York, and Phipp's Hall, University of Huddersfield. At the AudioLab four singers were situated in the Studio Live Room, and in Huddersfield four singers were located in a stage area in Phipp's Hall (Figure 6.1). At each endpoint singers are presented with an engineer–operated audio system providing microphone capture, low–latency streaming, immersive audio rendering, and headphone monitoring of networked performance. No visual display and interface is provided between remote performers. At Phipp's Hall an audience is located in the hall with the singers, experiencing four singers physically on stage, and experiencing the four remote singers auralised as if alongside the local singers in the stage area over a multichannel loudspeaker array. No visual display of remote singers is provided for this audience. In the AudioLab an audience is located in the Teaching Room, listening to an auralisation of the networked performance over headphones, and viewing a video display of the networked concert in a live–streaming context. In the auralisation for singer monitor mixes and audiences the networked performance is auralised in the acoustics of MV2. In addition to the INMP concert system itself, an extended performance capture system was used at Phipp's Hall. This includes a spatial microphone array and 360° video from the centre of the audience area and traditional video capture from the rear of the audience area.

In the following subsections a technical description of the INMP system is presented.

6.4.1 Hardware Configuration

The system hardware configuration at each endpoint has the same core configuration, where the difference in system components between endpoints is primarily the audience auralisation and playback system (Figure 6.2). At each endpoint singer microphone capture (*DPA 4464/4464-F*) and capture of an engineer private communications microphone (*Shure SM58*) is input to a microphone pre–amplifier (*RME OctaMic II*[4] and *Focusrite Scarlet OctoPre*[5] for the Huddersfield and York endpoints respectively). Audio rendering processes are hosted on a laptop or personal computer device at each endpoint. Local audio capture at each endpoint is passed from pre–amplifiers over Alesis Digital Audio Tape (ADAT) to audio interface inputs (*RME UFX III*[6] for each endpoint) for the audio rendering device. NMP streaming is hosted on a discrete device laptop/personal computer device at each endpoint. Here local audio captured is passed from pre–amplifier line outputs to the networking device audio interface inputs (*RME UCX II*[7]). Recieved remote performer audio channels are output from the networking device audio interface line outputs and passed to the audio rendering device audio interface line inputs.

[4] `https://rme-audio.de/octamic-ii.html`
[5] `https://focusrite.com/products/scarlett-octopre`
[6] `https://rme-audio.de/fireface-ufx-3.html`
[7] `https://rme-audio.de/fireface-ucx-ii.html`

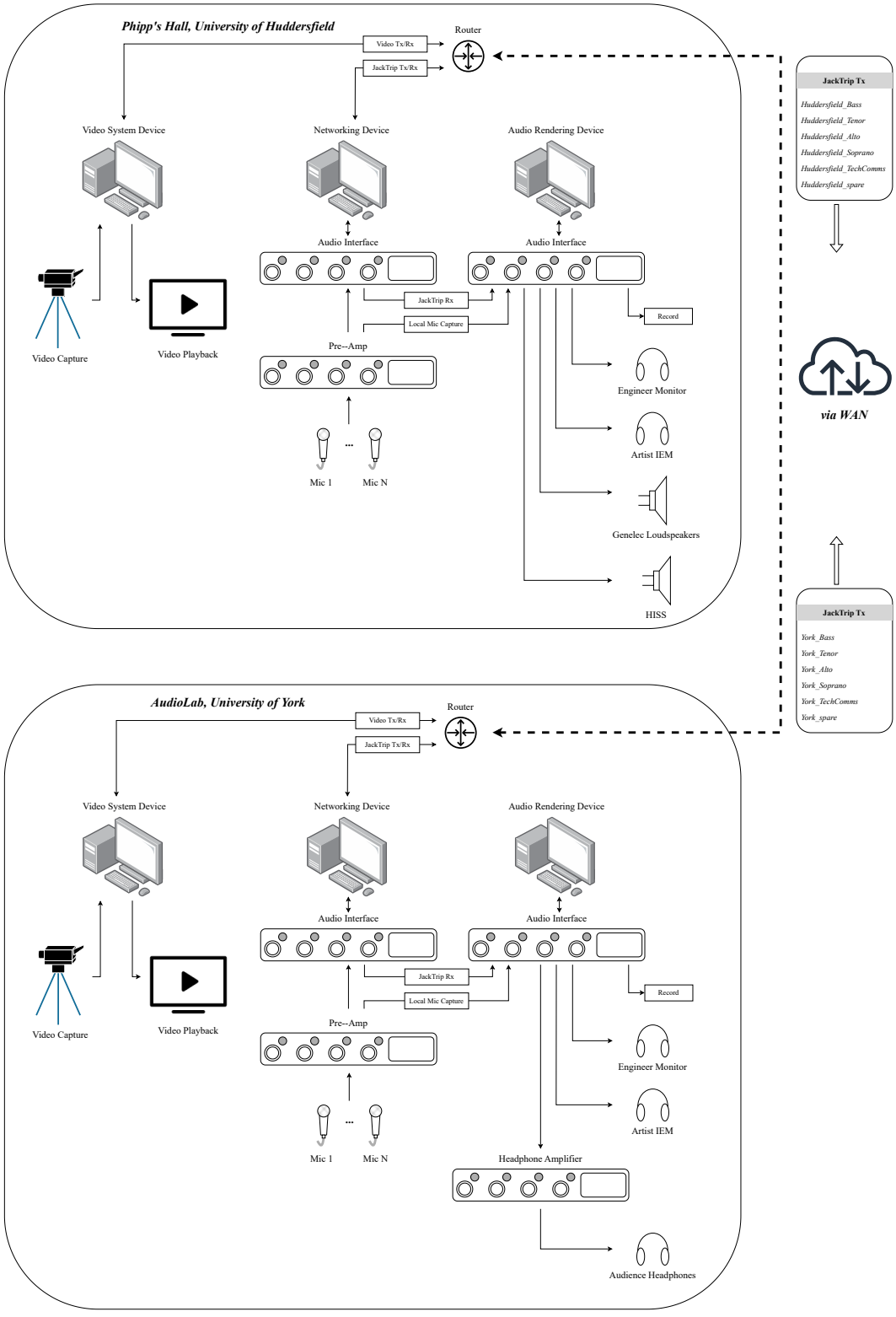


Figure 6.2: Hardware setup at Phipp's Hall, University of Huddersfield, and the AudioLab studios, University of York. Audio signal flow and streaming channels are detailed.

At each endpoint immersive audio monitor mixes are output from the audio rendering device interface to singer headphones (*Beyerdynamic DT 990* open–back headphones and *Shure SE215* IEMs at York and Huddersfield respectively). Engineer headphone monitoring (*Beyerdynamic DT 770*) of networked performance and engineer private communications is also output from the audio rendering device at each endpoint. At Phipp's Hall the Front of House (FOH) audio system used for audience playback consists of a multi–channel loudspeaker array, namely the Huddersfield Immersive Sound System (HISS)[8] and four *Genelec 8040 A* loudspeakers. The HISS is used for playback of auralised acoustics in the Ambisonic format, output from the audio rendering device interface via Multichannel Audio Digital Interface (MADI). Each discrete Genelec loudspeaker is used for playback of the direct signal from a discrete remote performer, output from the audio rendering device interface line outputs. A sample rate of 48 kHz and frame size of 128 samples is used across the audio system.

A video system was also present at each site. A digital video camera is used to capture video of the performers, and a discrete laptop device at each endpoint is used to stream video between endpoints using a common teleconferencing software, and to display video capture from each endpoint on screen display for the audience located in the AudioLab Teaching Room. In addition to the INMP concert system an extended performance capture was configured at Phipp's Hall. This includes Ambisonic recording of the concert performance, using the *MH Acoustics em32 Eigenmike*[9], 360° video recording (*Insta 360 ONE X*[10]) and traditional video camera recording.

6.4.2 Software Overview

The audio networking device at each endpoint hosts an instance of the NMP software JackTrip [75] to provide audio streaming functionality between endpoints. RME TotalMix[11] audio interface control software provides control over interface input and output routing. JACK[12] is used as the audio routing application, passing audio between JackTrip and the audio interface. The video system laptop at each endpoint hosts the teleconferencing software Zoom[13], used to stream video of performers between endpoints. The audio rendering device at each endpoint hosts rendering processes across the audio interface control and mixing software, RME TotalMix, and a DAW application, namely Reaper[14], used to host audio plugins used in immersive audio DSP.

[8] `https://www.thehiss.org/`
[9] `https://eigenmike.com/eigenmike-em32`
[10] `https://www.insta360.com/product/insta360-one`
[11] `https://rme-audio.de/totalmix-fx.html`
[12] `https://jackaudio.org/`
[13] `https://www.zoom.com/`
[14] `https://www.reaper.fm/`

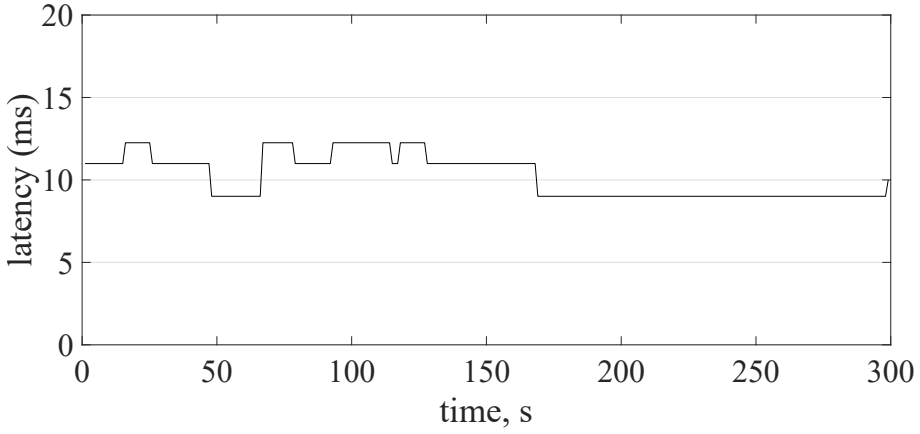


Figure 6.3: Estimated OWT network latency between Phipp's Hall, University of Huddersfield, and the AudioLab studios, University of York.

6.4.3 Audio Networking

Audio is streamed between sites via a peer–to–peer JackTrip (specifically version 1.3[15]) connection. The JackTrip audio stream between the two endpoints is configured for an uncompressed stream of 16–bit audio (48 kHz, 128–sample frame size), with a jitter buffer length of two frames, and FEC redundancy of two frames. A duplex six channel stream was established between the two endpoints, with an associated symmetrical upload and download BW estimate of 4.608 Mbps. The six audio channels streamed from one endpoint to the other are the four local musicians' microphone capture channels, the technician private communications channel, and a spare channel (Figure 6.2). JACK is used to route audio between JackTrip and audio interface capture and playback buffers. Network latency was measured with the INMP deployed between two endpoints using a loop–back 1 Hz pulse train with 100 Hz duty cycle test signal to measure the RTT over 5 minutes. RTT is then halved to provide an estimate of OWT (Figure 6.3). 11 $ms \pm 3\ ms$ network OWT audio delay was estimated from RTT measurement. The peer–to–peer JackTrip stream between endpoints was deployed over the existing academic Internet access at each endpoint, where BW is sufficiently over–provisioned (in the order of hundereds of Mbps upload and download).

6.4.4 Immersive Audio Rendering

The INMP concert system renders immersive audio for multiple playback systems, namely headphone playback for artist monitoring, headphone playback for the live–stream audience situated in the AudioLab, York, and loudspeaker playback for the live audience situated in Phipp's Hall, Huddersfield. Immersive audio processing follows an Ambisonics–based approach, allowing for auralisation of the acoustics of MV2 in a manner which is suitable for use in both headphone and loudspeaker playback systems [197]. Ambisonic auralisation is extended with stereo and multi–channel loudspeaker methods which are specific to playback system.

[15] `https://github.com/jacktrip/jacktrip/releases`

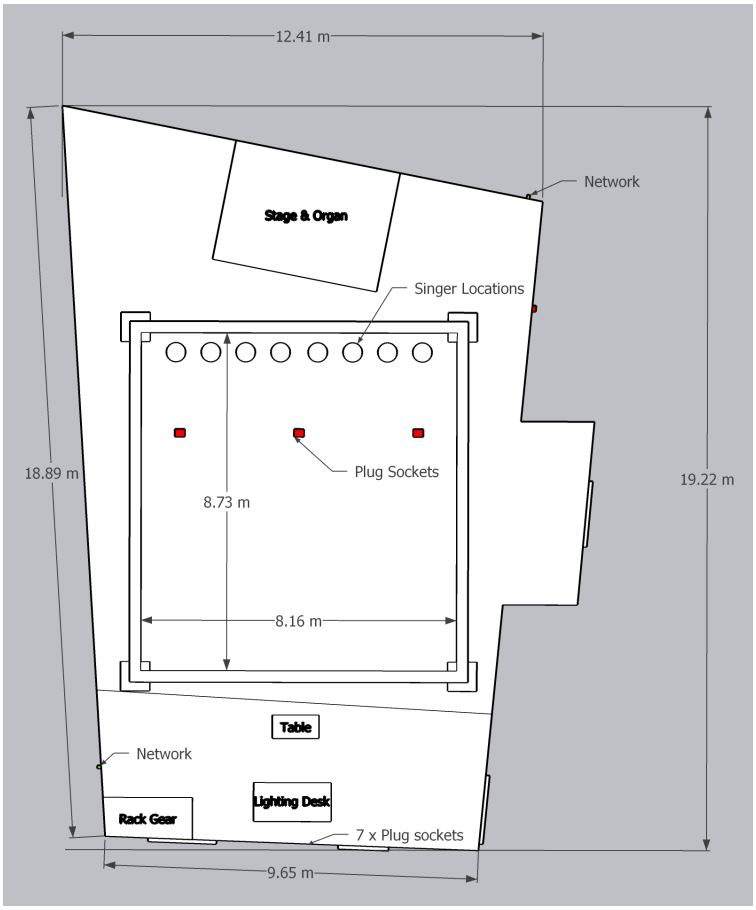


Figure 6.4: Floor plan of Phipp's Hall detailing staging for the concert event. The four stage–right positions are the remote singers, each represented by a discrete *Genelec* loudspeaker, and the four stage–right positions are the physically present singers.

For performer monitoring, direct sound is rendered using audio interface direct monitoring software (RME TotalMix) to minimise system latency. At each endpoint each of the four local performers is rendered a mix of local and remote performer direct sound, where sources are spatialised (following singer staging described in Figure 6.4) with traditional stereo panning, and where traditional live–sound production (dynamic effects and EQ) can be applied (Figure 6.5).

For auralisation of diffuse reverberation, discrete sends from local and remote performer's channels are summed and input to Reaper, which hosts an instance of MCFX convolver [200]. Here the input is convolved with a filter describing diffuse reverb 3OA SIR auralisation and Mag–LS [198] binaural decoding. This filter was rendered by pre–convolving the relevant SIR and binaural decoding filters in the spherical harmonic domain, where filters were generated using custom decoding software by Tomasz Rudzki. The output binaural signal describing diffuse reverberation of MV2 is routed from DAW output to RME TotalMix, where the diffuse reverb return is summed into the direct sound monitor mix for each of the local performers.

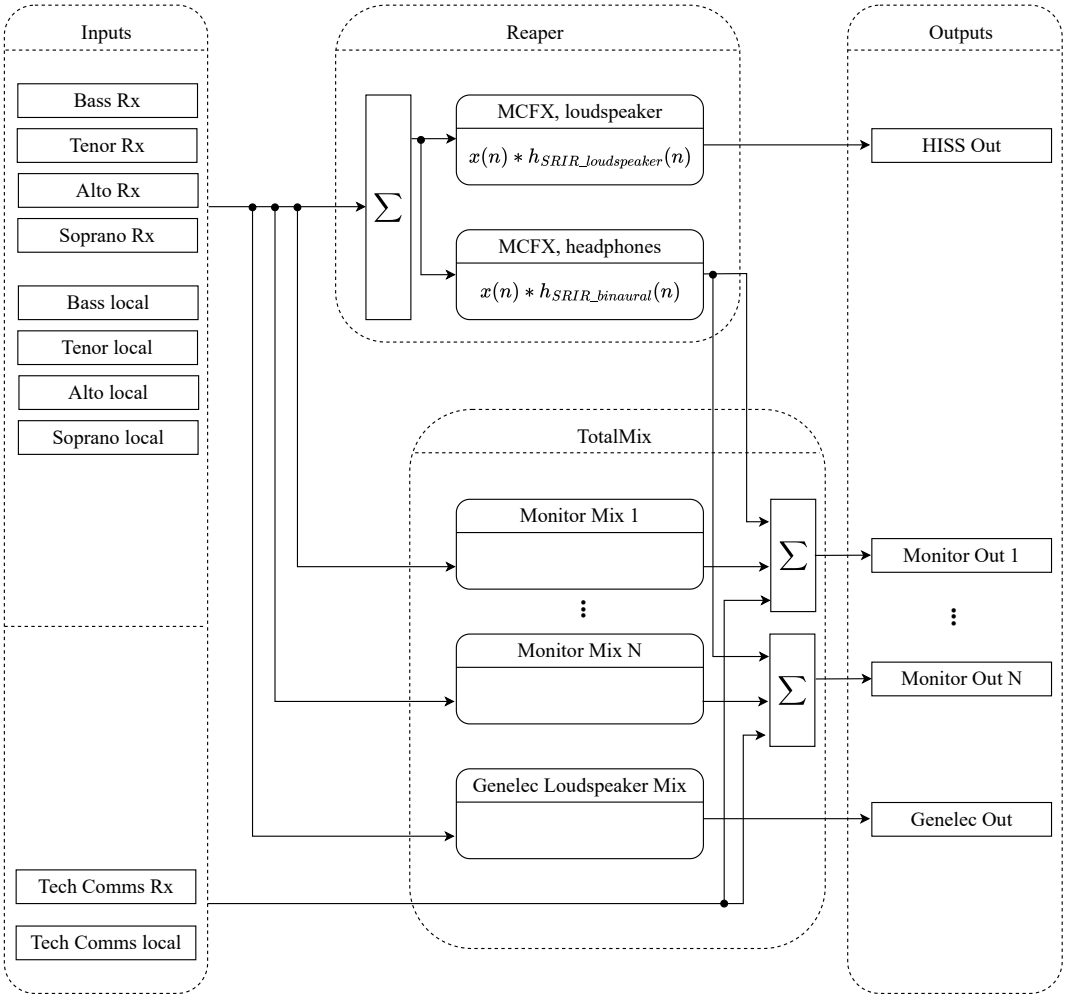


Figure 6.5: Immersive audio processing for the INMP concert system detailing relevant audio processing applications and audio routing.

In order to correctly time–align diffuse reverberation onset relative to the direct sound signal a large first partition is used to incur processing delay in convolution, while alleviating the computational load of convolution. For the live–stream audience in the AudioLab, York, a similar stereo–enhanced binaural auralisation of networked performance in the acoustics of MV2 is rendered. Here each listener in this audience receives the same mix, where stereo signal output from the audio interface is input to a headphone amplifier to be distributed for audience playback. Head–locked auralisation is rendered for all headphone playback using the INMP system.

Immersive audio rendering for the live audience in Phipp's Hall is also performed across RME TotalMix and Reaper. The four remote performer signals are mixed in TotalMix and the direct sound for each of the four performers is output to a discrete loudspeaker located on–stage, auralising the direct signal from each remote singer as a monophonic source. Local and remote performer channels are summed into a mono signal and input to Reaper which hosts a relevant instance of MCFX convolver.

Frequency (Hz)	125	250	500	1000	2000	4000	8000
Maida Vale Studio 2	1.12	1.03	1.15	1.24	1.32	1.12	1.00
Phipps Hall	0.97	0.93	0.84	0.80	0.79	0.73	0.57

Table 6.1: RT (T30) of MV2 and Phipp's Hall.

The input is convolved with a filter describing diffuse reverb 3OA SIR convolution, and Ambisonic loudspeaker decoding to the HISS array using AllRAD [193] methods. This filter was rendered through spherical harmonic domain pre–convolution of SIR and loudspeaker decoding filters generated from custom decoding software and provided by Tomasz Rudzki. The output signal describing diffuse reverberation is played back over the HISS array. The live audience here hears direct sound from local performers acoustically in the room, direct sound from each remote performer played back over a discrete loudspeaker in the stage area, and auralised diffuse reverb of MV2 played back over the HISS array.

6.5 SIR Measurements

The 3OA SIR used in the auralisation are resourced from a dataset describing acoustic measurement of the physical space of MV2 [5]. Relevant SIR measurement from this dataset were performed using sine sweep methods [187], where a 20 s sweep length was used. *Genelec 8040 A* loudspeakers[16] were used in measurement at a height of 1.5 m from floor to tweeter, and calibrated to 85 dBC at 1 m distance. The *MH Acoustics em32 Eigenmike*[17] was used as the spatial microphone array receiver from which 3OA SIR are derived [187], positioned at a height of 1.7 m. The measurement dataset for MV2 includes ISO–3382 stand acoustic measurements [141]. Additionally IR measurements for Phipp's hall were measured with the INMP concert system deployed from which acoustic parameters were estimated. Estimates of RT from these acoustic measurements (Table 6.1) demonstrate a slightly higher RT for MV2 than Phipp's Hall, which may be considered more suited to concert listening [213].

As the SIR measurement was to be processed to remove direct sound and early reflections such that only diffuse reverberation is described, specific source and receiver positions for the measurement are arbitrary. The SIR used describe source position PA02, using a directional measurement with the speaker pointing away from the audience area (Figure 6.8), and receiver position OA11. The *'mixing time'* used to estimate the onset of diffuse reverberation was calculated as the square root of the room volume of MV2 [143], giving an approximated diffuse reverb onset time of 44 ms. The 3OA SIR was first truncated to 1 ms prior to the peak of direct sound with a 1 ms linear fade–in (Figure 6.6). Following this the SIR was then truncated to the mixing time relative to the onset of direct sound, with a 1 ms linear fade–in (Figure 6.7).

[16] `https://www.genelec.com/previous-models/8040a`
[17] `https://eigenmike.com/eigenmike-em32`

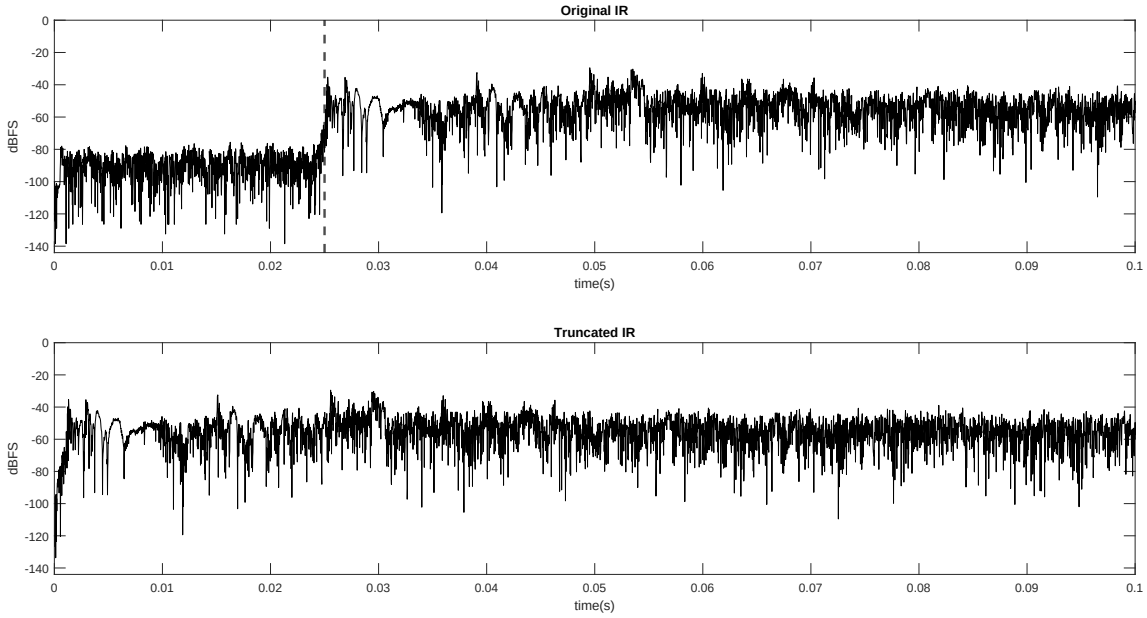


Figure 6.6: Illustration of the truncation of the beginning of SIR to the onset of direct sound. ACN 0 from the AmbiX format 3OA signal is shown, illustrating the original SIR (above) and SIR truncated to the onset of direct sound (below). The vertical line on the upper plot indicates the onset of the direct sound.

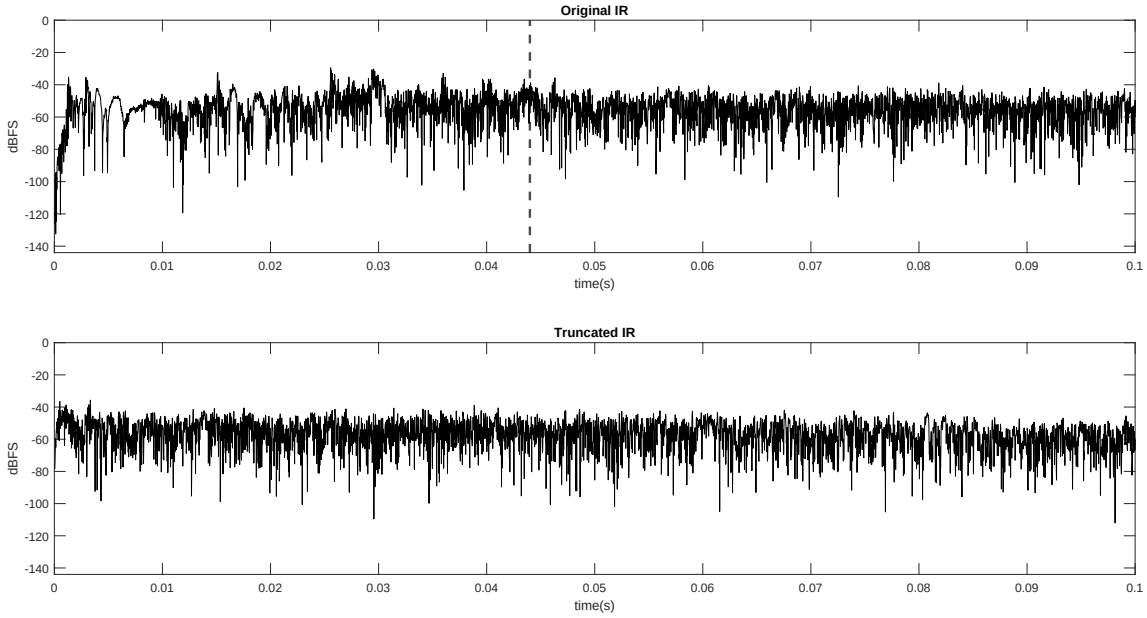


Figure 6.7: Illustration of the truncation of the beginning of SIR to the onset of diffuse reverberation, or *'mixing time'*. ACN 0 from the AmbiX format 3OA signal is shown, illustrating the SIR which has been truncated to the onset of direct sound (above) and SIR truncated to the onset of diffuse reverberation (below). The vertical line on the upper plot indicates the onset of diffuse reverberation.

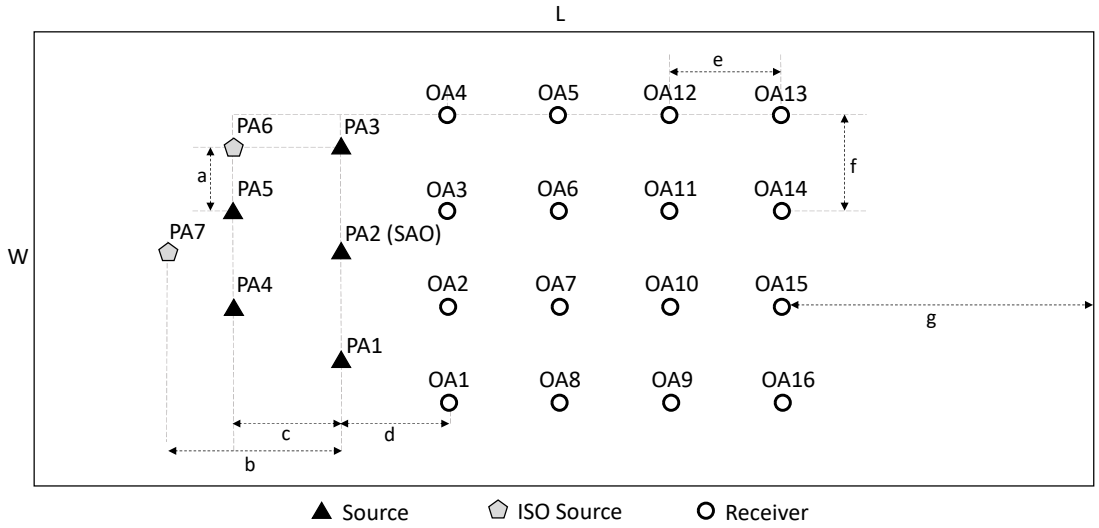


Figure 6.8: Acoustic measurement positions for MV2. The source position (*PA02*) and receiver position (*OA11*) for the SIR measurement used in this study are illustrated.

6.6 Part I: Evaluation Method

The following subsections detail the evaluation protocol used in the first part of the INMP system evaluation.

6.6.1 Protocol Overview

This study follows the common NMP QoE evaluation approach [6]. Here the quality of performer and audience concert experience is judged through post–session response to rating items and exit interview discussion [15], [20], [35]. For performer evaluation the singers of *Ex Corde Vocal Ensemble* were invited to provide ratings of the performance experience after three discrete sessions:

- *Live rehearsal:* where performers gathered in a physical space (namely the Teaching Room, a classroom at the AudioLab, University of York) and rehearsed the concert performance material acoustically while cohabiting a physical space. This is considered to provide reference quality rating of relevant items.
- *Network rehearsal:* where performers used the INMP as deployed for the concert experience to rehearse the concert performance material. This occurred one day after the *live rehearsal*.
- *Concert:* where performers deliver the actual concert performance to audiences in York and Huddersfield. This occurred one day after the *network rehearsal*.

Additionally performers provide comment throughout the concert production, and provide interview discussion. Post–concert, the audience are invited to provide rating of the concert experience and additional comment.

6.6.2 Participants

For performer evaluation the eight members of *Ex Corde Vocal Ensemble* participated in the study. This vocal octet consists of four male and four female singers (two of each of bass, tenor, alto and soprano parts). Participant age ranged from 26–49 years, and they also reported professional or semi–professional levels of singing expertise. Four performers had previous experience in remote performance using teleconferencing solutions, one had experience teaching music using teleconferencing solutions (but no ensemble performance experience here), and three participants had no prior remote performance experience. No participant had prior experience with NMP technology. All eight participants engaged in interview discussion. Seven of the eight participants responded to rating questionnaires following the *live rehearsal* and *network rehearsal* sessions, and six responded to rating questionnaires following the *concert* session.

For audience evaluation, 21 individuals participated in the study. 17 participants were from the live audience in Phipp's Hall, and four participants were from the live–stream audience at the AudioLab. 15 participants were male, five were female and one did not provide gender information.

6.6.3 Performance Material

The performance material delivered in the INMP concert consisted of a program of 16th and 17th Century vocal pieces, namely:

- O La Oche Bonecchо (*Orlando di Lasso*)
- Hear My Prayer (*Henry Purcell*)
- Ave Maria (*Franz Schubert*)
- Ave Regis Caelorum (*Marc-Antoine Charpentier*)
- Dixit Dominum (*George Frideric Handel*)
- Laudate Dominum Omnes Gentes (*Tomás Luis de Victoria*)
- O Magnum Mysterium (*Morten Lauridsen*)
- Crucifixus (*Antonio Lotti*)
- If Ye Love Me (*Thomas Tallis*)
- Tutto Lo Di (*Orlando di Lasso*)

The sequence of performance material was delivered as listed during the *concert* session. During the *live rehearsal* and *network rehearsal* the content of performance was not controlled, and performers rehearsed freely. The members of *Ex Corde Vocal Ensemble* considered performance to be *'democratic'*, indicating that performers responded to one another throughout delivery of performance material without any distinct leader part or parts.

6.6.4 Latency Conditions

Latency conditions were estimated through summing the measured OWT network delay between endpoints (approximately 11 ms, subsection 6.4.3) with the audio interface capture and playback buffering associated with the network devices at each endpoint (5.3 ms). As direct sound is monitored via audio interface direct monitoring software, which incurs negligible (sub–millisecond) latency, a system throughput of approximately 16 ms OWT is therefore estimated between endpoints. Notably, audience listening in Huddersfield is over loudspeakers, where position–dependent additional latency is incurred between loudspeakers and listeners in air–propagation of sound. For performer monitoring of own voice and of local performers over headphones sub–millisecond latency is incurred through audio interface direct monitoring software.

6.6.5 Performer Questionnaire

Performers provided evaluation of the performance experience by rating their agreement with the following items on a seven–point rating scale raging from strong disagreement (1)–strong agreement (7) via a neutral point (4):

1. I was able to perform in time with the other performers who are in the room with me.
2. I was able to perform in time with other performers who I am connected to remotely.
3. I was able to deliver a high quality musical performance.
4. Audio delay was not noticeable between myself and other performers who I am in the room with.
5. Audio delay was not noticeable between myself and other performers who I am connected to remotely.
6. The audio quality was of a professional standard.
7. It sounded like I was in a different acoustic space than the studio/concert hall I was physically in.
8. It sounded like I was performing in the same acoustic space as the rest of the ensemble.
9. I enjoyed the performance experience.
10. The virtual acoustics (the simulated *'room reverb'*) influenced my performance.

Rating items 1–6 address standard NMP QoE rating items including perceived synchrony, performance quality [8], [9], [19], [20], [35], [109], [113], [114], [118], [120], [121], perceived delay [8], [9], [19], [35], [114], [118] and audio quality [35], [118], [180]. Given that performers monitor both local and remote singers, relevant rating items address perception of local and remote performer interactions discretely. Overall experience quality [8], [20], [110] is also addressed in their rating of item 9, here with respect to enjoyment.

Rating of items 7 and 8 address rating of presence [9], [19], [20], [113], [121] here with respect to the concepts of *transportation* and *co–presence* [125] respectively. Given that other relevant studies report acoustics having the potential to impact the experience of NMP [8], [16], [19], [203], item 10 of this study addresses the degree to which performers felt acoustics affected their performance. Participants responded to the rating questionnaire after the *live rehearsal*, *network rehearsal* and *concert* sessions via a Qualtrics[18] survey. Only rating items 1, 3 and 9 were considered relevant and addressed for the *live rehearsal* session, where the INMP system was not used. The Qualtrics survey provided participants the additional opportunity to provide open comment on the performance experience.

6.6.6 Audience Questionnaire

Evaluation of the concert experience of the audience members is achieved by rating their agreement with the following items on a seven–point rating scale raging from strong disagreement (1)–strong agreement (7) via a neutral point (4):

1. The local and remote musicians were performing in time with one another.
2. The musicians delivered a high quality musical performance.
3. The audio quality was of a professional standard.
4. It sounded like the performers were in a different virtual acoustic space than the concert hall.
5. It felt like I was in the virtual acoustic space with the performers.
6. I enjoyed the concert experience.

[18] `https://www.qualtrics.com/`

The audience questionnaire is created by extending NMP QoE and INMP presence rating factors (as outlined in subsection 6.6.5) to the audience perspective as in other practical NMP concert evaluation [20], [35]. Items 1–3 address perceived performance synchrony, performance quality and audio quality. Items 4 and 5 address presence in terms of *transportation* and *co–presence*. Item 6 addresses overall concert experience quality in terms of enjoyment. Participants responded to the rating questionnaire after the *concert* session via a Qualtrics survey. The questionnaire additionally allows participants an opportunity to provide open comment on the concert experience with relevance to each rating item.

6.6.7 Performer Semi–Structured Interview

The performers of *Ex Corde Vocal Ensemble* engaged in discussion of the performance experience throughout the *live rehearsal*, *network rehearsal* and *concert* sessions. Additionally participants engaged in an interview discussion regarding the experience of networked performance after a sound–check on the day of the *concert* session. Interview prompts here encouraged discussion and feedback around the following topics:

- Initial reactions to using the INMP concert system.
- Experience of latency.
- Ability to perform synchronously.
- Experience of immersive audio monitoring.
- Visual contact.
- Communication in performance.
- Potential applications of remote performance workflows.

6.6.8 Informal Testing of Loudspeaker Monitoring

During a sound check prior to the *concert* session performers in Phipp's hall removed their IEMs and briefly attempted performance monitoring local singers acoustically, and remote singers over the FOH loudspeaker array. Though this was a brief informal test, discussion with performers here is relevant to the case of INMP concerts and is reported.

6.6.9 Study Limitations

It can be acknowledged that monitoring level was not controlled across this study. For live sources (i.e., local performers), monitoring level for each listener will vary with the dynamic performance delivered by the relevant live source and the relevant listening position. In headphone monitoring of networked performance for performers and the live–stream audience at the AudioLab, and in loudspeaker playback for the live audience in Phipp's Hall, microphone capture and discrete source monitoring level was adjusted throughout all sessions in response to performer request and the production preferences of the relevant audio engineers. Lack of control over monitoring level may be reasonably expected to influence rating of the concert experience, however, this may also be considered a native aspect of the ecologically representative context of this study.

Various aspects of listening context can be acknowledged as varying across participants in this study. Each performer is presented with a discrete monitoring mix content and visual perspective. The physical environment for York and Huddersfield singers can also be acknowledged as varying, with the York singers situated in a studio live room, and the Huddersfield singers situated in a stage area in Phipp's Hall before a live audience. For audience members it can be particularly noted that the live–stream audience in the AudioLab and the live audience in Phipp's Hall are presented particularly distinct concert experiences, with the Audiolab audience experiencing screen display and headphone playback in a screening environment, and the Phipp's Hall audience experiencing live musicians and loudspeaker playback in a venue environment. Here it can be reasonably assumed that it is inappropriate to group data from these two audience groups. Within the live audience in Phipp's Hall each audience member perceives the concert from a discrete listening position. Cumulative differences in listening context across participants may be reasonably expected to have the potential to influence results in experience rating. This presents a lapse of control in experience rating, however, is again a native factor of the ecologically representative scenario considered in this case study.

Notably for headphone monitoring a combination of stereo and Ambisonics–based binaural approaches are used. Though this is representative of a practically manageable real–time sound production system for INMP, the findings of this study can be acknowledged as of limited generalisability with respect to immersive audio monitoring systems, where Ambisonic or binaural stereo methods of free–field auralisation may be used.

Though this case study describes WAN deployment of the INMP concert system, where some packet loss may be reasonably expected, no monitoring of packet transmission reliability was taken in this study. As such, streaming artifacts, which may influence experience rating for performers and audience members, was not controlled or measured. The latency and quality of video display presented to the live–stream audience in the AudioLab was not measured or controlled, and can, again, be reasonably expected to influence the rating of the audience experience here.

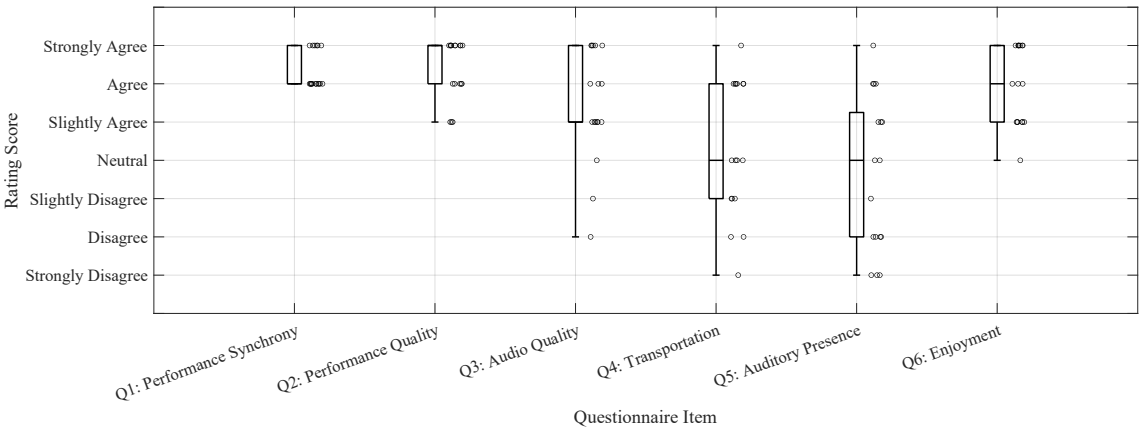


Figure 6.9: Experience rating results from the live audience in Phipp's Hall (N=17).

6.7 Part I: Results

The following subsections report the results of the first part of the INMP system evaluation.

6.7.1 Audience Rating Results

For ratings of items 1 and 2, addressing perceived performance synchrony and quality, Inter–Quartile Range (IQR) of ratings ranges between slight and strong agreement (Figure 6.9, Figure 6.10). Participants also provide relevant comment in questionnaire responses from the audience in Phipp's Hall. One participant perceived superior dynamic control in delivery from live singers relative to remote singer reproduction. Another participant perceived the delivery from live singers to be more expressive than remote singers due to visual contact and perception of singer motion. Audio quality rating (item 3) IQR ranges from slight–strong agreement, however, one neutral and two negative ratings are given from the audience in Phipp's Hall. Two participants from this audience provide relevant comment, indicating audio quality impairment due to streaming artifacts in remote performer loudspeaker playback. One participant comments that *'slight clicks broke the immersion'*.

Nine participants from the audience in Phipp's hall provided comment on loudspeaker reproduction of remote singers. Participants demonstrate varying perception of timbral aspects of reproduction. One participant comments *'timbral balance was excellent between real and remote singers.'* One participant considered that there was *'a slight spectral difference'* between live singers and reproduction of remote singers, whereas another participant considered the difference to be *'a distinct timbral change.'* Close microphone techniques used in singer audio capture was considered by one participant to result in noticeable differences between live singers and remote singer loudspeaker reproduction due to proximity effect. Other participants considered that the naturalness of loudspeaker reproduction of voice was not transparent nor natural. One participant considered loudspeaker reproduction of remote singers to be perceived with localisation *'blur'* and larger apparent source width relative to live singers.

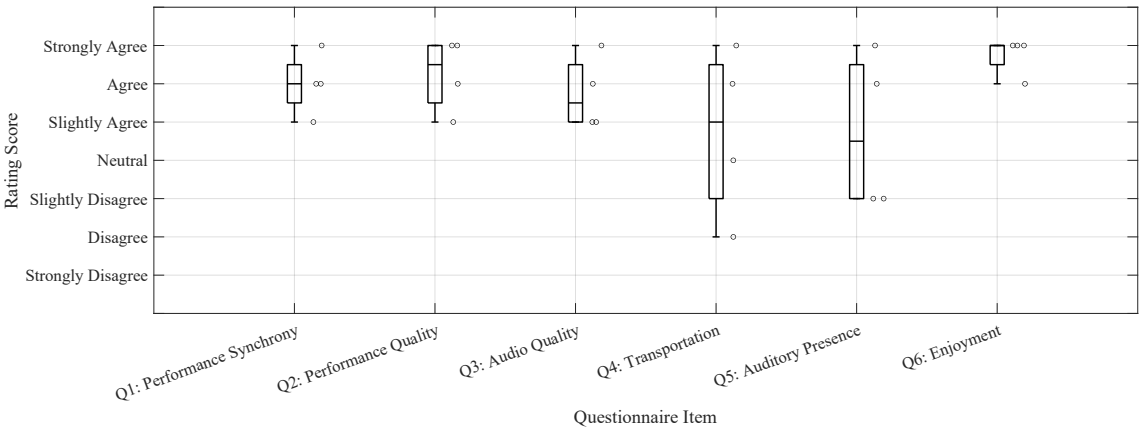


Figure 6.10: Experience rating results from the online headphone audience in York (N=4).

Another participant, however, provided differing comment, considering that loudspeaker reproduction of remote singers was *'surprisingly similar in localisation, distance, reverberation, timbre (and) sense of presence.*

The rating of the presence factor of transportation (item 4) demonstrates IQR ranging from slight disagreement to strong agreement, and with ratings ranging the entire rating scale. Rating of co–presence (item 5) also ranges the entire rating scale, with IQR ranging from disagreement to agreement for the Phipp's Hall audience and IQR ranging from slight disagreement to strong agreement for the headphone audience in York. With respect to perception of immersive audio rendering, one participant from Phipp's Hall comments on lack of awareness of playback of the simulated acoustics of MV2, stating that they *'couldn't tell that (the) HISS system was playing anything.'* Another participant from Phipp's Hall suggests that awareness of virtual acoustics was affected by relative loudness to live singers, stating that *'virtual acoustics were masked for the on–site performers...direct sound was way louder'*. One participant from Phipp's Hall expressed a desire for auralisation of different spaces, suggesting that *'reverberant rooms or churches'* may be a more suitable acoustic for delivery of the performance material.

For overall enjoyment of the concert experience (item 6) IQR for the Phipp's Hall audience ranges from slight–strong agreement, and between agreement and strong agreement for the headphone audience at the AudioLab. Here one participant from Phipp's Hall comments on loudspeaker reproduction of remote singers, stating that they *'enjoyed female vocals through the speakers (more) than the male vocals'*, which may be considered as relevant to the vocal range of parts in the octet.

6.7.2 Performer Rating Results

The rating of NMP QoE factors is generally positive for all sessions (Figure 6.11). For rating of item 1 and 2, (perceived performance synchrony), ratings are clustered from slight–strong agreement for *live rehearsal*, *network rehearsal* and *concert* session. For rating of item 3 (performance quality) IQR of rating ranges from neutral–agreement for the *live rehearsal* session, however, is generally positive, clustered between slight and strong agreement, for the *network rehearsal* and *concert* sessions.

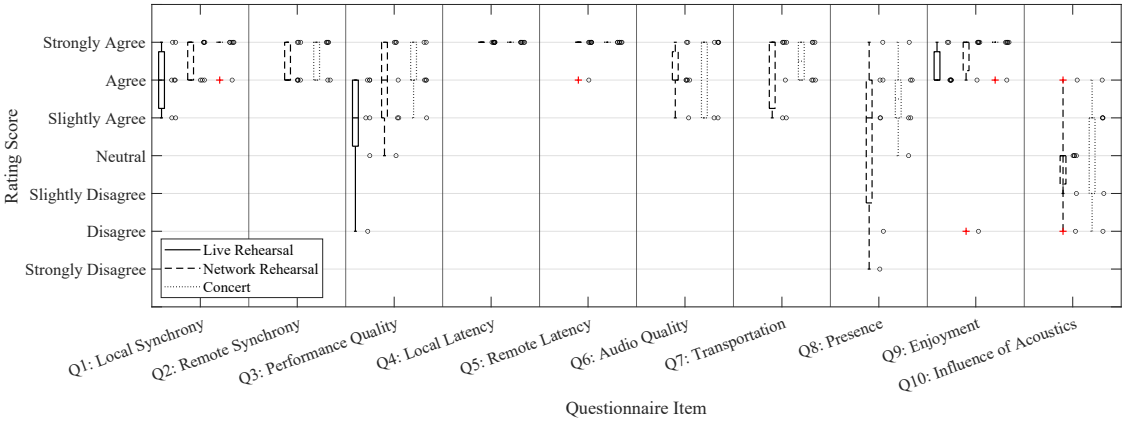


Figure 6.11: Singer experience rating results. N = 7 for the *live rehearsal* and *network rehearsal* sessions, and N = 6 for the *concert* session.

The rating of latency with local (item 4) and remote (item 5) performers is consistent with medians of strong agreement for *network rehearsal* and *concert* sessions. The rating of audio quality (item 6) is clustered between slight and strong agreement for the *network rehearsal* and *concert* sessions. For rating of enjoyment of the performance experience (item 9) ratings are clustered between agreement and strong agreement for *live rehearsal*, *network rehearsal* and *concert* sessions.

The rating of presence factors shows some variation. For item 7, addressing presence in terms of transportation, ratings are generally positive for the *network rehearsal* and *concert* sessions, with IQR ranging from slight–strong agreement, and from agreement–strong agreement respectively. Rating of presence in terms of co–presence, however, ranges the full rating scale for the *network rehearsal* session, however, is generally positive for the *concert* session with IQR ranging from slight agreement to agreement. The rating of the degree to which acoustics were considered as impacting performance quality is ambiguous, with ratings ranging from disagreement–agreement for both the *network rehearsal* and *concert* sessions.

6.7.3 Performer Interview Response and Feedback

With respect to the initial reaction of the performers to using the INMP concert system, all participants responded positively, with some indicating a preference for INMP relative to prior experiences in remote performance using typical teleconferencing software. One participant comments that they had *'a fantastic time'* and another comments that they *'really enjoyed the experience.'* When questioned regarding perception of delay, participants unanimously agreed that no audio latency was perceived either in monitoring of local performers who are physically present, or in monitoring of remote performers.

Considering perceived synchrony performers generally commented that they were able to perform in time with ease. For networked performance one participant notes that *'as well as strict time it felt possible to be spontaneously flexible.'* Another participant notes that due to *'high concert energy'* the ensemble delivered the performance *'at a different tempo than rehearsed'*, considering that remote performers were *'very quick to match this tempo and adapt even with no visuals.'* One performer notes that audio artifacts made performance difficult during rehearsal, however, it can be acknowledged that this comment relates to faulty IEM hardware (specifically a loose connection) which was corrected prior to the concert experience. One participant considered that familiarity with the rest of the ensemble assisted in performance, stating that *'this would be difficult to do with seven strangers.'* One participant notes that during the *live rehearsal* session there were *'errors due to sight–reading.'*

In discussion of the challenges experienced, participants note the lack of visual contact as the most prolific challenge, however, the ensemble was quick to develop performance strategies to rely on aural cues such as breathing in vocal performance. One participant noted that lack of visual contact made communication between the ensemble more difficult, stating that the absence of visual contact with a conductor introduced *'a few complications'* such as *'agreeing how to end pieces.'* Participants in York noted that the lack of visual contact with the audience in Phipp's Hall was a significant divergence from traditional concert performance experiences.

Discussion of performer monitoring addresses several practical considerations of concert performance. Headphone monitoring of performance was considered to be an irregular case for vocal ensemble concert, and participants noted that the unfamiliar monitoring context may have impacted their performance. One performer particularly noted that headphone monitoring impacted presence factors, stating *'it felt very artificial.'* This participant also noted that they found monitoring of their own voice via IEMs disorientating. Another participant noted that for monitoring of local performers they focused on sound in the room (heard through open–back headphones) rather than headphone monitoring.

With respect to monitoring level, participants noted a lack of confidence that monitor mixes were representative of acoustic levels delivered to the audience in Phipp's Hall. Here one participant comments that *'when you are in the room together you adjust your dynamics to blend with the whole group, when you have a different mix you are not quite sure who has been turned up and who has been turned down in the headphones.'* When asked if participants desired individual control of levels in headphone monitoring all responded affirmatively, with one participant noting that the distributed nature of the concert system added some difficulty in communicating with engineers to express monitoring needs. One participant, however, noted the benefits of authenticity in immersive audio auralisation, stating *'adjusting our own individual balances is pretty useful for what we want to hear but its not very useful for the overall effect...no—one in the group can balance what they are singing as they would in an acoustic...a lot of what we do musically is acoustically balance between the parts.'*

Participants note benefits in spatialisation and clarity in immersive audio monitoring. One participant comments that *'particularly Sopranos and Altos...along the front we can often not hear other voice parts...it was quite nice to absolutely know we can hear every part.'*

Performers at both the AudioLab and Phipp's Hall noted an awareness between the acoustics of their physical location and the auralised acoustics of MV2 in monitoring. One participant expressed an enjoyment of the experience of performing in auralised acoustics. Another participant provides comment that for monitoring of aurlisation over headphones they *'found that the closer proximity of the sound of the other singers to my ears and head made it really fun to explore the music as I could easily choose which voices to tune into and feed off of. The acoustic phenomenon of constructive interference when two tones in the same harmonic series are played together allowed for more fine tuning to make the chords ring in my ears/head.'* One participant, however, commented that during rehearsal they were unaware of acoustic rendering, and perceived their monitor mix as *'in their head'*, and considered monitoring over headphones more *'precise'* than live room sound.

In Huddersfield participants experienced some audio artifacts associated with a loose connection in IEM hardware during rehearsal. This was corrected prior to the concert event, however, is considered to have impaired monitoring audio quality to a degree. In York participants noted some streaming artifacts (i.e., *'clicks'*), however, considered artifacts tolerable and not to affect performance. Mixed comment was given on the brief experience of participants in Phipp's Hall monitoring over the FOH array rather than IEMs. Notably the Bass and Tenor participants considered this preferrable to IEMs. The Bass participant states that it was *'liberating'* relative to headphone monitoring, and felt comparable to live co–located performance.

The Soparano and Alto parts, however, provided contrary statements, noting a preference for the clarity and spatialisation of headphone monitoring. Participants also provided discussion of potential further practical scenarios for INMP workflows as practicing professional musicians. Here participants considered the access–enabling potential for remote performance workflows in education and rehearsal contexts. Participants also considered the logistical benefits of remote workflows, noting reduced disruption to lifestyle due to circumvention of travel requirements, reduced cost (considered to improve project viability), and reduced environmental impact associated with travel.

6.8 Part I: Discussion

In the following subsections discussion is given of the results of the first part of the INMP system evaluation. This includes the overall concert experience, and the experience of immersive audio and visual display.

6.8.1 Overall Concert Experience

With respect to the overall concert experience it can be noted that both audience and performer evaluation infers a high level of enjoyment was achieved, and that performance synchrony and overall performance quality was experienced positively. Indeed coherent comment reflecting an overall positive concert experience is provided from both performers and audience members. This is coherent with prior research in low–latency NMP concert workflows using traditional (i.e., stereo) audio systems, where a high quality concert experience is reported [15].

Particularly considering performer evaluation, it can additionally be noted that *network rehearsal* and *concert* sessions achieve equivalent or superior rating for performance synchrony, quality and enjoyment to the *live rehearsal* session, considered as reference conditions. Indeed performer ratings and feedback demonstrates no awareness of audio latency. In prior NMP research it has been well–demonstrated that where low–latency (sub–30 ms) is achieved performers will not experience any impairment of performance conditions [6]. As such comparable ratings of *live rehearsal* and networked sessions here can be expected. Notably ratings of performance quality appear to improve for networked sessions relative to the *live rehearsal* session. It is considered here that perceived performance quality is likely to improve with time spent rehearsing, and that as performers progress through sessions the time spent rehearsing increases.

Positive responses in both audience and performer experience ratings indicates that professional audio quality was delivered using the INMP system. Several audience participants from Phipp's Hall, however, do identify some streaming artifacts in audio playback, and indeed some neutral and negative rating of audio quality is given by a minority of this group. Performers also identify the presence of streaming artifacts in discussion, however consider this tolerable, and provided rating of agreement that audio quality was of a professional standard. Indeed prior work evaluating the impact of streaming artifacts on NMP QoE reports such artifacts tolerable by performers, and is not associated with any clear experience impairment [35], [118]. Though it may be concluded generally that professional quality audio was delivered by the INMP concert system, it is highlighted that the presence of streaming artifacts did present a noticeable difference to traditional live concert experiences for some participants.

Streaming artifacts can result from packet loss in audio transport, however are most typically associated with jitter buffer over/under–run handling [75] used to maintain throughput and approximate synchronisation across endpoints in low–latency NMP streaming software [15], [39].

Practical approaches to minimisation of streaming artifacts require synchronisation of clocks across endpoints. Here GPS systems [39], [115], [135] are implemented in several systems to achieve synchronisation, however, performance here may vary in signal–dense [115] or highly shielded environments. Alternatively, in pro–AV contexts, high–precision reference clocks may be deployed as clock masters at each endpoint, such that clock skew across the length of typical sessions is minimised to negligible levels. NTP servers are considered in many cases, though may tolerate some significant drift between endpoints [6]. Otherwise over–provisioning of network BW and jitter buffer size is effective in minimising audio artifacts due to packet loss.

With respect to presence factors, performers provide positive rating here for the *concert* session. Rating of presence varies for performers in the *network rehearsal* session, however, this is likely due to changes in sound production and issues with faulty equipment during this session, and not considered a significant concern with respect to rating of the INMP system. Notably audience rating of presence factors, however, provides mixed results.

From participant comments a variation in experience can be identified relating to perception of the FOH audio system (discussed further in subsection 6.8.2). Participant comment also identifies visual factors as relevant to experience of presence factors in the concert experience (discussed further in subsection 6.8.3).

Regarding the research objective of validating the INMP system as a remote concert workflow, results demonstrate that a high quality concert experience was indeed achieved for performers and audiences, and that for performers comparable concert performance quality to live reference conditions was indeed supported. The INMP concert system is therefore demonstrated as providing a high quality concert experience. With respect to wider investigation of the INMP concert experience, results indicate that streaming artifacts may impair experience of audio quality, and that audience participant experience of presence factors may not be optimised both with respect to immersive audio rendering and with respect to visual display and interface.

6.8.2 Experience of Immersive Audio Rendering

With respect to performer experience of immersive audio monitoring, some degree of appreciation and usefulness is indicated. Here participants highlighted improved spatial clarity relating to spatialisation of sources [180] in immersive audio monitoring relative to live performance contexts, where acoustic monitoring and singer positioning may provide sub–optimal monitoring conditions. Indeed subsequent research reports high rating of immersive audio monitoring for INMP contexts with respect to localisation and spatial clarity[19] factors [19], [121].

Considering the practical scenario presented by the case study, it is highlighted that singers do find headphone monitoring uncomfortable and unfamiliar. Indeed prior research in NMP concert performance highlights similar issues [35]. With respect to monitoring preferences, it can be highlighted that when loudspeaker monitoring was briefly and informally trialled, the Bass and Tenor performers, positioned nearest centre–stage preferred loudspeaker monitoring, whereas the Alto and Soprano performers, positioned nearest the wings, preferred headphone monitoring. Preferences in loudspeaker monitoring may relate to distance from source, where the amount of additional air–propagation latency and the natural spatial clarity associated with a discrete listening position may predict the suitability of loudspeaker monitoring.

Singers also highlighted the practical issue of monitoring level in immersive audio monitoring influencing dynamic control in performance. Indeed in both live [11] and virtual [214] performance scenarios acoustics are demonstrated as influencing aspects of delivered performance and perceived performance conditions. Though it is reasonable to assume similar expectations in INMP, participant ratings are unclear as to whether or not rendered acoustics influenced performance. Here, however, immersive audio auralisation is proposed as useful with respect to plausible auralisation.

[19] Often referred to as 'source separation.'

The practical case is presented that plausible auralisation allows remote performers to hear the performance as if in the room together, and allows for tuning of performance in the same manner as live scenarios. With respect to practical remote concert workflows, it is considered that allowing performers individual control of levels in immersive audio monitoring is a viable solution considered useful by participants here, such that performers may adjust to meet their own performance needs.

Each performer participant notably had a reference experience of performing in the acoustics of their local physical environment in addition to experience of performing with acoustic simulation in immersive audio monitoring. It can be highlighted that performers provided positive rating of presence factors for the *concert* session, and that achieving this may be influenced by the reference acoustic experience of the local physical environment (i.e., awareness of difference, as confirmed in relevant discussion).

Audience participants from Phipp's Hall, however, provide varying rating of presence factors, and indeed a lack of awareness of simulated reverb playback over the HISS was highlighted in participant comments. The RT difference between Phipp's Hall and MV2 exceeds the 5% difference suggested in relevant standards [141] above 250 Hz (Table 6.1). It may be the case that audience participants, with no reference experience of performance without immersive audio playback, did not notice the virtual acoustics due to this lack of reference. Lack of awareness of virtual acoustics may have influenced rating of presence factors. Indeed presence rating items in the audience questionnaire relate specifically to perception of virtual acoustics, rather than wider experience of presence in the remote concert event. It is considered that affecting presence in simulated acoustics here may be improved by raising diffuse reverb level in FOH playback over the HISS. Indeed it is demonstrated in prior work that reverberation level may influence perception of both *envelopment* [215] and *externalisation* [216]. Additionally, it can be considered that the superposition of the acoustics of Phipp's Hall on playback of the simulated acoustics of MV2 may have influenced perception of acoustics and related presence factors.

The mixed rating of presence factors by the Phipp's Hall audience participants may also relate to extensive comment on perception of immersive audio loudspeaker playback. Monophonic loudspeaker playback of remote singer direct sound is implemented in the INMP concert system design as a method of achieving improved localisation relative to Ambisonic loudspeaker arrays and other surround–sound loudspeaker arrays for position–independent listening across an audience area [194].

Whereas some audience participants perceived this loudspeaker reproduction as similar to an experience of live singers, several note impairment of the naturalness and plausibility [125] of loudspeaker reproduction of the remote singers. Particularly timbral, dynamic, and localisation factors are noted in comments on perceived differences. Notably some participants propose the use of filters to improve the naturalness of loudspeaker reproduction of voice as a practical system improvement. Indeed the naturalness of loudspeaker reproduction of voice is understood in other research to have an impact on presence factors in telecommunications contexts [217].

Position–independent listening of loudspeaker reproduction of direct sound and diffuse reverb can be acknowledged as resulting in varying DRR across the audience in Phipp's Hall. Listening preferences in DRR for musical content are identified in other work [144]. It may therefore be possible that DRR, and other factors of position–independent listening influence differences in audience responses to immersive audio rendering.

With respect to the research objective of further investigation of the INMP concert experience, results and participant feedback novelly demonstrates the potential for remote performance experience improvement with immersive audio rendering, with respect to practical performance factors such as clarity in monitoring, or supporting performance tuning for delivery in simulated or remote acoustic environments. Limitations of immersive audio monitoring are, however, also highlighted, particularly with respect to singer discomfort with headphone monitoring. Concerning the audience results it is particularly highlighted that loudspeaker audience playback in INMP is likely to provide mixed results, where the audience listening experience is likely to depend on listening position and perception of plausibility and naturalness in loudspeaker reproduction.

6.8.3 Experience of Visual Display and Interface

Though visual factors of the INMP concert experience are not addressed in the audio–centric performer and audience experience rating questionnaires, participants from both groups provide significant comments on visual elements. For performers, visual factors of performance context were highlighted both with respect to communication between performers, and communication between performers and audience. With respect to inter–performer communication, participants note a lack of visual contact as the biggest challenge of the experience. Indeed, visual cues in ensemble singing have been highlighted as important to both inter–performer communication [122] and objective quality of synchrony in prior studies [56]. Though lack of visual contact was considered the primary challenge of using the INMP concert system, participants do, however, note that adaptation of performance to rely on aural cues was possible and effective.

Presence factors of the performance experience, namely co–presence and communication between remote performers and audiences were considered by both groups to be limited by the lack of visual display and interface. Remote performers indicated an experience of being isolated from the audience, which may have impacted their delivery of the performance. Comments from the audience provides coherent indication that live singer performance is perceived as more expressive than remote performers. Relating to the research objective of further investigating the novel INMP concert experience, results clearly demonstrate that lack of visual display and interface limits performance interaction and audience experience, and inclusion of visual display and interface may improve the INMP experience with respect to presence and communication factors.

6.9 Part II: Evaluation Method

The following subsections provide detail on the evaluation protocol for the second part of the INMP system evaluation.

6.9.1 Protocol Overview

Evaluation of the quality of musical performance using the INMP concert system is achieved via a novel listening test approach. Here listeners are presented with audio recordings of excerpts of musical performance achieved using the INMP concert system (namely performance of *Crucifixus* and *Hear my Prayer*, recorded during the *network rehearsal* session). Listeners provide ratings of the perceived quality of the recorded musical performance for each excerpt via a questionnaire of rating items adapted from NMP QoE rating methods. For each of the INMP excerpts audio recording of musical performance of the same material in live performance conditions is also presented and rated, and considered to provide a reference measurement.

The listening test was created as a Qualtrics[20] survey. Participants are requested to provide demographic information, and are required to confirm use of headphones. The survey interface presents a sequence of audio recordings. For each audio recording participants are directed to listen and provide ratings of perceived performance quality via the survey GUI. Participants provide three repeated ratings of each of the four audio recordings in the listening test, for a total of 12 ratings per participant. Audio recordings are presented to all participants in the same randomised order. Participants are free to listen to each audio recording any number of times, and are free to move between discrete audio recordings and are able to modify relevant ratings throughout the listening test.

6.9.2 Participants

12 individuals participated in this study. Nine participants were male and three participants were female, and participant age ranged from 23–76 years. Two participants had no prior experience in music performance evaluation. Six participants reported previous participation in 1–5 prior evaluations. Four participants had significant prior assessor experience, reporting participation in more than 20 prior music performance evaluations. Each participant confirmed that they were monitoring the listening test audio over headphone playback. The discrete model of headphone used by each participant is reported in Table 6.2.

6.9.3 Listening Material

Listening material for performance of *Crucifixus* and *Hear my Prayer* for both live performance and performance using the INMP concert system was resourced from a dataset associated with the concert production[21]. For each of the two pieces of performance material an excerpt was selected. Audio recordings of the relevant excerpts were taken from audio recording of *Ex Corde Vocal Ensemble* for live performance, and for performance using the INMP system. For each excerpt audio is prepared in Reaper. The mono audio recording for each discrete singer is first normalised to an arbitrary choice of -23 Loudness Units Full–Scale (LUFS). Each of the eight sources are then panned

[20] `https://www.qualtrics.com/`

[21] `https://zenodo.org/records/10402916`

Headphone Brand and Model	Frequency
Beyerdynamic DT 770 Pro	2
Beyerdynamic DT 990	2
Beyerdynamic DT 235	1
Sennheiser HD 650	1
Sennheiser HD 598	1
Sony MDR 7506	1
Bowers & Wilkins P9 signature headphone	1
Ollo Audio S4X	1
JBL (no model noted)	1
Apple (no model noted)	1

Table 6.2: Headphone models used by participants in completion of the evaluation.

in stereo to a discrete pan angle. Bass, Tenor, Alto and Soprano parts were panned respectively 10°, 30°, 50°and 70°to the left for the York quartet, and to the right for the Huddersfield quartet, following performer physical placement in the INMP concert event (Figure 6.4). The stereo master for each piece of listening material was then normalised to -23 LUFS in compliance with loudness recommendations for streaming of audio[22] [218]. The stereo master for listening the test material was then exported as a .wav file (48 kHz, 16–bit).

6.9.4 Rating Questionnaire

Participants provided an evaluation of the quality of musical performance for each piece of listening test material by providing ratings on a 5–point scale (Terrible, Poor, Fair, Good and Excellent) for each of the following rating items:

1. Enjoyment of listening.
2. Overall quality of musical performance.
3. Musical timing of ensemble performance.
4. Dynamic balance of ensemble performance.
5. Pitch tuning of ensemble performance.

Rating item 1 considers enjoyment of the experience of listening to networked performance. This item is adapted from NMP QoE rating items which considers overall enjoyment of networked performance from either a performer [8] or audience [20] perspective, and aims to address overall experience of listening to networked performance.

[22]Streaming loudness recommendations were selected, as this was considered the most relevant case to INMP concert, where no recommendations or standards exist.

Rating item 2 is adapted from the NMP QoE rating items which address the overall quality of networked performance, rating the performance experience [35], [109], [110], [118], [120], or rating from listening of networked performance [20], [35]. Rating item 3 is also adapted from previous NMP QoE rating items, here addressing perceived quality of performance synchrony [114], [118], [122].

NMP performance quality evaluation has traditionally focussed on synchrony [6]. Rating items 4 and 5 expand on this by addressing performance dynamics and tuning respectively. These rating items are derived from the wiser field of musical performance evaluation, for example, evaluation of live [11] or virtual [214], [219] performance acoustics. Indeed participant responses to the first part of this study indicates that dynamic quality of performance may vary between live performance and INMP methods. The terrible–excellent quality rating scale, as can be found common across some prior NMP evaluation [19], [20], [35], [110], [118], is adopted in the interest of cohesion and repeatability across research.

6.9.5 Study Limitations

The audio level of the listening experience was uncontrolled in this experiment. Given that loudness is well understood as affecting perception of audio this lapse in control may have influenced ratings of performance quality.

Differences can also be acknowledged in the hardware across the study, namely in headphones used for listening. Though suitable quality open or closed back headphones are noted by the majority of participants, the headphone model is unknown in two cases. Though headphone model differences may influence differences in listening experience across participants this is not considered likely to impact rating items in this study.

Audio recordings of performance for INMP and live reference performance methods are notably from discrete performances, and thus natural variation between performances may be considered a lapse in control. Additionally, the degree of practice of performance material at the time of recording varies for INMP and live performance, where more practice time is associated with INMP methods (where recording occurs a day after live recording). Though variation between performances is inevitable, *Ex Corde Vocal Ensemble* are a professional vocal group, and all performers report semi–professional and professional proficiency (subsection 6.6.2), and can therefore be expected to deliver high–quality performance with some consistency in live performance conditions.

Though building upon prior NMP evaluation the listening test approach to performance quality evaluation here is novel in the INMP context, the suitability of this method as a means of evaluation has not been formally validated (though can be found in some prior NMP concert evaluation [35]). Additionally, though a reference performance is provided for each excerpt, no low anchor is implemented in the listening test, such that only one ‘goalpost’ for performance quality is defined (though it can be recognised that ‘anchor–quality’ performance may be difficult to define). As such there is some limitation to the confidence in results of the second part of this study.

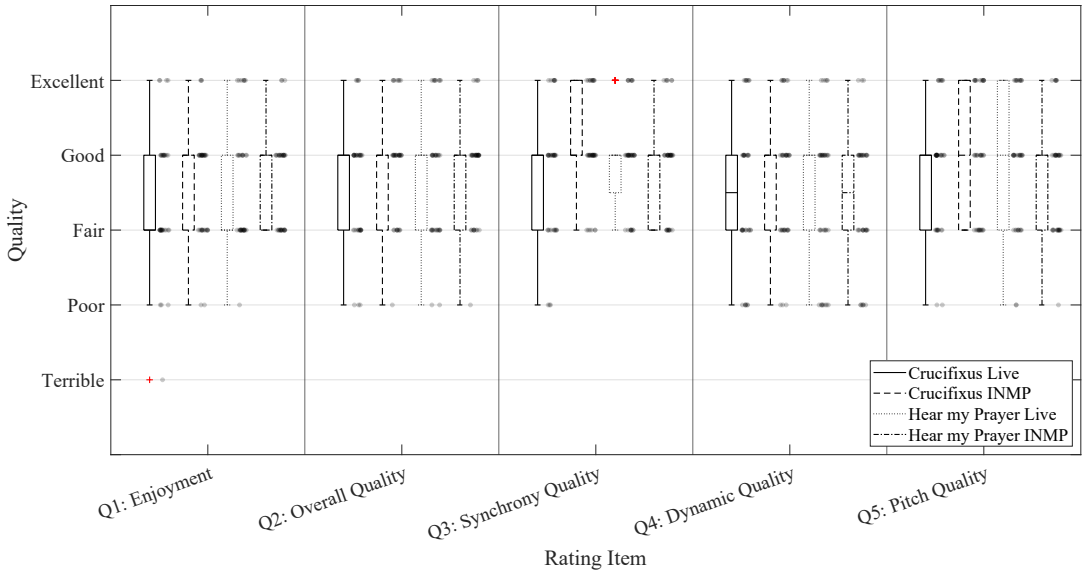


Figure 6.12: Rating results from performance quality evaluation. Box plots are given for each of the four discrete groups, with each group relating to a discrete combination of performance material and method. Individual ratings are shown to the right of the box plot for each discrete group (N = 36)

6.10 Part II: Results

The five–point rating scale used in this study may be considered too coarse for robust assumption of normality in data. As such, non–parametric statistical testing was used for analysis of rating results.

Statistical analysis was conducted using MATLAB[23]. For each of the five questionnaire items, a Friedman Analysis of Variance (ANOVA) was used to test for significant difference in rating across the four pieces of listening material (*Crucifixus Live, Crucifixus INMP, Hear My Prayer Live, and Hear My Prayer INMP*, as detailed in subsection 6.9.3). Each piece of listening material is characterised by a specific piece of performance material (*Crucifixus* and *Hear My Prayer*) and specific performance conditions (Live or INMP). Testing for difference across listening material for each questionnaire item, therefore, tests for differences associated with the factors of performance material and performance conditions.

Results of this analysis (Table 6.3) show that for each of the questionnaire items no significant difference is found across the four piece of listening material (tested at $p < 0.05$). It can, therefore, be concluded that there is no significant difference in performance quality rating associated with the factors of performance material and performance conditions. Follow–up pairwise comparisons were, therefore, not necessary.

[23] `https://uk.mathworks.com/products/matlab.html`

IQR (Figure 6.12) ranges from fair–good for all groups in rating of enjoyment of listening (Q1), overall quality of performance (Q2) and dynamic balance in performance (Q4). For rating of performance synchrony (Q3) IQR generally ranges from fair–good with the exception of performance of *Crucifixus* using INMP methods, where IQR ranges from good–excellent. For rating of pitch tuning in ensemble performance (Q5), IQR ranges from fair–good for live performance of *Crucifixus*, and performance of *Hear my Prayer* using INMP methods. For rating of performance of *Crucifixus* using INMP methods, and for live performance of *Hear my Prayer*, IQR ranges from fair–excellent. Medians for all material across all rating items range from fair–good.

6.11 Part II: Discussion

Statistical analysis of performance quality rating factors demonstrates no difference in rating of performance quality between live performance and performance using the INMP concert system for either of the tested pieces of performance material. This was observed to be the case across all rating items. The hypothesis that musical performance quality will vary between live performance and performance using the INMP system is therefore rejected in this case.

Considering the research objective to test the viability of the INMP system through comparative rating of INMP and live performance quality, results suggest the viability of the INMP concert system with respect to supporting high quality concert performance. It can also be acknowledged that ratings of performance quality were generally positive for both live and INMP workflows. These findings are coherent with rating of performance quality from performers and audience members in the first part of this study (as discussed in subsection 6.8.1). Indeed, from prior investigation of latency in NMP, performance quality is not expected to be impaired for the low–latency (sub–30 ms [6]) conditions achieved by the practical INMP concert system deployment.

The generalisability of these findings can, however, be acknowledged as limited with respect to practical INMP concert experiences, given the lack of previous validation of the listening test evaluation approach, and given the substantial differences in listening conditions between those presented in the listening test and practical listening conditions in INMP concert experiences.

Rating item	df	N	Chi-sq	p
Q1: Enjoyment	3	36	1.15	0.7644
Q2: Overall Quality	3	36	2.85	0.4156
Q3: Synchrony Quality	3	36	3.38	0.3373
Q4: Dynamic Quality	3	36	4.37	0.224
Q5: Pitch Quality	3	36	3.29	0.3488

Table 6.3: Results of the Friedman test conducted across the four pieces of listening material for each of the questionnaire items. For each rating item the degrees of freedom (df), number of measurements (N), Chi–square statistic (Chi–sq), and p–value (p) are given.

It can be highlighted that ratings typically range from fair–good across evaluation items. This indicates that the singers achieved performance which was rated positively both for INMP and live performance. Notably, however, no performance was rated with means which correlate to a low anchor rating (terrible) or high anchor rating (excellent). Indeed, as discussed in subsection 6.9.1, a live performance recording by the same musicians as INMP is used as a reference in evaluation, and no low anchor was implemented (as discussed in subsection 6.9.5). The limitation can, therefore, be acknowledged, that the specific ratings given in this study are not generalisable beyond this study. What is rated to be 'good' quality in this study cannot necessarily be assumed of higher quality than a rating of 'fair' in another study. The inherent limitation here is the lack of absolute definition of 'terrible' and 'excellent' quality in performance evaluation (i.e., low and high anchor), which may be used as common points of reference across wider research. Definition and validation of low and high anchors for discrete performance material is, however, beyond the scope of this study, and accepted as a limitation.

6.12 Research Question Revisited

At the beginning of this chapter the research question is presented:

> *Can a low–latency INMP system provide a practical remote concert performance workflow?*

This research question was addressed in a two–part study. In the first part of the study, the objective was defined:

- *To evaluate the performer and audience experience of INMP concert performance.*

This objective was achieved through the performer and audience evaluation attached to the concert event. As discussed in section 6.8, results of evaluation indicate that performance quality which is equivalent to live performance quality was achievable, and positive rating of the concert performance was generally achieved for both performers and audience members. The first part of this study, therefore, provides an answer to the research question by demonstrating that the INMP system can indeed provide a practical remote concert performance workflow. Several limitations of the workflow, particularly with respect to immersive audio rendering and visual display, were, however, also identified. This provides further detail regarding the research question, identifying limits on the practical use of the INMP concert system.

In the second part of this study the objective was defined:

- *To evaluate the quality of musical performance from audio recording for live performance and for performance using the INMP concert system.*

Evaluation of performance quality was achieved via a listening test, rating the quality of performance using the INMP system, and live performance of the same performance material. As discussed in section 6.11, results of evaluation demonstrate positive performance quality rating associated with listening material recorded using the INMP system.

Additionally no significant difference is found between rating of INMP performance quality and live performance quality. The second part of this study therefore addresses the research question by demonstrating that performance quality which is achieved under live performance conditions was achieved using the INMP concert system. The INMP system is, therefore, reasonable to consider as a practical solution for facilitating concert performance.

6.13 Summary

The first part of this study contributes a practice–oriented evaluation of the novel context of INMP concert. Results of this practical user–testing suggest the viability of the INMP system as a remote concert workflow through demonstration of a high quality concert experience for performers, and for both the live–stream and live audiences. These findings are consolidated in the second part of this study, where results demonstrate comparable rating of musical performance quality for performance using the INMP system and reference–quality live performance. With respect to the research question associated with this study (subsection 6.3.2), it is concluded that low–latency INMP systems can indeed provide a practical remote concert performance solution. Further investigation contributed by experience evaluation highlights practical usefulness of immersive audio monitoring in remote concert performance workflows, particularly with respect to timbral and spatial clarity in monitoring, and supporting tuning of performance for delivery in simulated acoustics.

Also contributed in evaluation, however, is the highlighting of several challenges of practical INMP concert scenarios. This includes position–independent listening and acoustic divergence for loudspeaker array reproduction of immersive audio for live audiences, dynamic control in performer immersive audio monitoring, and the impact of visual display absense on presence and communication factors of the remote concert experience. This study also contributes to an online and publicaly available dataset associated with the networked concert event[24]. This dataset includes media recordings from the concert event and rehearsal, recordings of interview discussions with performers, and experience rating results from the audience and performer participants of this study, providing resources for use in further INMP research.

[24] `https://zenodo.org/records/10402916`

Chapter 7

Evaluation of the effect of Reverb Time on vocal performance using an INMP system

Oooh! What does THIS button do?

Dee Dee, Dexter's Laboratory
(1996–2003)

7.1 Preface

This chapter presents an evaluation of the impact of RT using a practical INMP system design for vocal performance. A brief summary of related prior system development and practical WAN deployment is given. A study is then detailed where the system was reproduced over the AudioLab LAN and evaluated by pairs of singers. An overview of the study design is given, including a summary of motivations and relevant literature, and study objectives and hypothesis. The INMP system design and implementation in a novel evaluation framework is detailed, as is related SIR measurement and processing. Evaluation methods are detailed, including description of a novel onset–detection workflow. Results are then presented, followed by relevant discussion. The chapter then closes with a summary of key findings and conclusions.

The evaluation presented in this chapter was conducted by Patrick Cairns. The SIR measurements of York Guildhall used in the INMP system were performed by Patrick Cairns, Anthony Hunt, Josh Gregg and Andrew Chadwick. These measurements were performed as part of the Architexture project[1] using MATLAB scripts and a measurement protocol implemented by Andrew Chadwick.

[1] `https://architexture.york.ac.uk/`

7.2 An INMP System for Remote Vocal Performance

In 2020, an INMP system design for remote vocal performance was deployed between performers' homes over practical consumer ISP WAN backbones [7], [8]. Here an Ambisonics–based approach to auralisation and playback of head–locked binaural audio was used to auralise for singers the experience of playing together in different performance spaces.

Convolution–based auralisation was implemented, using SIR measurements from the Open Air SIR library[2]. Practical OWT latency values reported ranged from approximately 23–33 ms connecting to sites across the UK from Glasgow, Scotland, and up to approximately 55 ms connecting to sites across Europe from Glasgow.

The INMP system design builds upon prior development in simulation of singing experiences at the AudioLab. Here systems for real–time Ambisonic auralisation of vocal performance [220] were adapted to systems which simulate the experience of singing with a pre–recorded choir in a simulated acoustic environment for a single user [208], namely the Vocal Interactions in a Virtual Acoustic (VIIVA) system [221]. This system is used in active projects such as Sing From Your Seat[3] and Virtual Choirs in Care Homes[4]. The low–latency INMP implementation of the VIIVA audio system design [7], [8] was implemented as an extension of JackTrip NMP streaming software [75]. This development works towards the extension of such simulated choir experiences to multiple users and allowing remote access. The continued development of this system detailed in this chapter extends immersive audio rendering from FOA to 3OA methods and includes interactive rendering with 3DoF.

7.3 Study Overview

The INMP system was deployed over the LAN at the AudioLab, University of York, as part of a novel INMP audio system evaluation framework which extends the system with a networked multi–user test protocol control and participant GUI utility. A study was conducted where pairs of singers provide evaluation of the experience of duet vocal performance using the INMP system via experience rating and synchrony analysis. Evaluation is provided for three discrete RTs in auralisation (namely 2.3 s, 1.5 s, and 0.5 s), derived from SIR measurement of the GuildHall, York. In order for evaluation to reflect practice, RT is evaluated for three discrete OWT latency levels (13 ms analogue connection, providing a baseline, and 18 ms and 38 ms JackTrip connections, reflecting OWT latency observed in practical deployment [8]). The following subsections describe the motivations for this study, and detail the study objectives, hypotheses and statement of ethics.

[2] https://www.openair.hosted.york.ac.uk/

[3] https://audiolab.york.ac.uk/sing-from-your-seat/

[4] https://audiolab.york.ac.uk/virtual-choirs/

7.3.1 Motivations

Remote systems for group vocal performance (or *'virtual choirs'*) are explored as a remote workflow solution for accessing choir activity where physical access is impractical or impossible (for example, circumventing travel requirements) [36]. Whereas use of consumer teleconferencing software is most typical here it is well understood that in practice the musical interactivity between remote singers is significantly limited by system latency, where the ability to listen and respond to others and experience of music timing is drastically impaired [51], [130]. NMP solutions provide an alternative where singers are able to interact with one another in real–time and experience performance timing in a similar manner to live performance [36], [51], [130]. In this context INMP systems provide vocal ensembles the experience of hearing networked performance as if together in a simulated acoustic environment. This is considered to enhance the experience of networked vocal performance, making the experience of networked musical interactions more like live performance experiences by including simulation of room acoustics and spatial properties of live listening [8], and considered as a design requirement of remote vocal performance workflows [130].

Reports of the influence of acoustics on networked musical interactions in prior work, however, are limited, and provide varying reporting of perception of acoustics. In one prior study, inclusion of reverberation in monitoring is suggested as improving synchrony in clapping exercises [110]. In another study, dry audio monitoring is reported as preferred, where inclusion of reverberation is reported as impairing clarity in monitoring [119]. One study suggests that reverb (here with an RT of 1.2 ms) will reduce impairment of NMP QoE associated with high OWT latency between clarinet performers [203]. Evaluation of clapping rhythms conducted using an XR INMP monitoring system reports no significant differences in QoE or performance synchrony in comparison of auralisation of a listening room (RT of 0.46 s) and a hall (RT of 2.05 s) [19]. In other work, evaluating instrumental performance using an INMP system with video display, it is suggested that optimisation of presence is achieved through environmentally congruent auralisation (where acoustic parameters are not divergent from expectations of a physical or simulated environment presented visually) [16].

Specific to the case of vocal performance using INMP systems prior practical system development relating to the system described in this chapter [7], [8] suggests that variation in auralisations of a studio booth, medium hall and cathedral may influence both performer experience and synchrony. It is clear from previous INMP evaluation that simulated acoustics are likely to have an impact on INMP experiences, including perception of NMP–specific experience factors such as perception of latency. Indeed this conforms to findings evaluating the influence of live [11] or simulated acoustics [214], [219], [220], [222] on performance, where it is suggested that acoustics may influence perceived performance conditions and physical parameters of performance such as dynamics or synchrony.

Though simulated acoustics are expected to influence the INMP experience, previous literature does not describe relationships between acoustic parameters and INMP QoE and synchrony. In this regard no practical guidelines can be derived which can inform sound production and INMP experience optimisation in practice. With the motivation of INMP experience improvement this study contributes towards a practice–oriented understanding of the influence of acoustic parameters on INMP duet singing which may be used to inform best–practice in INMP productions.

7.3.2 Objectives

With the motivation of contributing towards a practice–oriented understanding of the effect of acoustic parameters (here considered with respect to RT) on INMP duet singing, the primary objective of this study is surmised:

> *To evaluate the influence of RT in duet vocal performance using a practical INMP system design.*

As a practice–oriented evaluation, evaluation of RT is performed for two networked latency levels considered to reflect practical deployment of the INMP system, and baseline analogue latency levels considered to provide a reference in evaluation. The secondary objective of this study can therefore be given:

- *To perform evaluation of the influence of RT in INMP duet vocal performance across different practical latency levels.*

7.3.3 Hypothesis

With relevance to the primary objective of this study, the related hypothesis is defined:

> *INMP QoE and/or synchrony for duet vocal performance will vary for different RT levels.*

With respect to the secondary objective, addressing evaluation across practical latency levels, two secondary hypotheses are defined:

- *INMP QoE and/or synchrony for duet vocal performance will vary for different latency levels.*
- *Interaction between RT and latency levels will affect variation in INMP QoE and/or synchrony for duet vocal performance.*

7.3.4 Statement of Ethics

Ethical approval for this study was obtained from the University of York School of Physics, Engineering and Technology Ethics Committee, reference *Cairns20230207*.

7.4 System Overview

The INMP system for remote vocal performance outlined in section 7.2 was deployed over the LAN at the AudioLab between two endpoints (namely the Studio Live Room and Studio Control Room). Here the system is extended with a novel test framework which provides networked control of the test protocol and a participant GUI. At each endpoint a participant is presented with the INMP system, consisting of microphone audio capture, a head–tracking device, headphone monitoring, and a personal computing device which host audio streaming, audio rendering, and test protocol control and GUI applications. Additionally the researcher is located in the Studio Control Room with a personal computing device providing headphone monitoring and microphone capture and talk–back utility in order to supervise the test protocol. The following subsection provide technical detail on the INMP system implementation.

7.4.1 Hardware and Physical Routing

Each participant endpoint is equipped with the same hardware configuration (Figure 7.1). Performer voice is captured with a *DPA 4464* omnidirectional microphone headset[5]. The microphone signal is routed via a microphone splitter to the local audio interface (*Focusrite Scarlett 18i20 g3*[6]*, 48kHz, 64–sample buffer*) inputs and to the remote audio interface input via the studio tie–lines (providing the analogue connection associated with baseline latency levels). The local audio interface at each endpoint is connected to a laptop device (*i7-11800H, 2.3 GHz, 32GB RAM, Win10*) hosting relevant streaming, rendering and protocol control and GUI software. Monitoring of the auralised performance is delivered via headphone playback (*Beyerdynamic DT990 Pro*[7]). A headphone–mounted head–tracking device (used in 3DoF rendering) is connected via Universal Serial Bus (USB) to the local laptop device. Initially the *nv sonic head tracker* developed by Tomasz Rudzki[8] was used. However, when available the *Supperware Head Tracker 1*[9] was used as this model was more convenient to mount on headphones. No significant performance variation is expected between the two head–tracking devices.

7.4.2 Audio Networking

Two discrete connections are configured between the two endpoints. In the first case local audio capture is passed to the remote endpoint at each site via analogue connection. This connection is used for baseline latency conditions, where no additional latency associated with network transport is incurred. In the second case a duplex audio stream was configured between the two endpoints using JackTrip [75] (specifically JackTrip v1.3[10]). JackTrip was here configured between participants for a jitter buffer size of four frames, FEC redundancy of two frames and one channel of audio (local voice capture).

[5] `https://www.dpamicrophones.com/microphones/headset-and-earset/4466`
[6] `https://focusrite.com/products/scarlett-18i20-3rd-gen`
[7] `https://beyerdynamic.com/p/dt-990-pro`
[8] `https://github.com/trsonic/nvsonic-head-tracker`
[9] `https://supperware.co.uk/headtracker-overview`
[10] `https://github.com/jacktrip/jacktrip/releases`

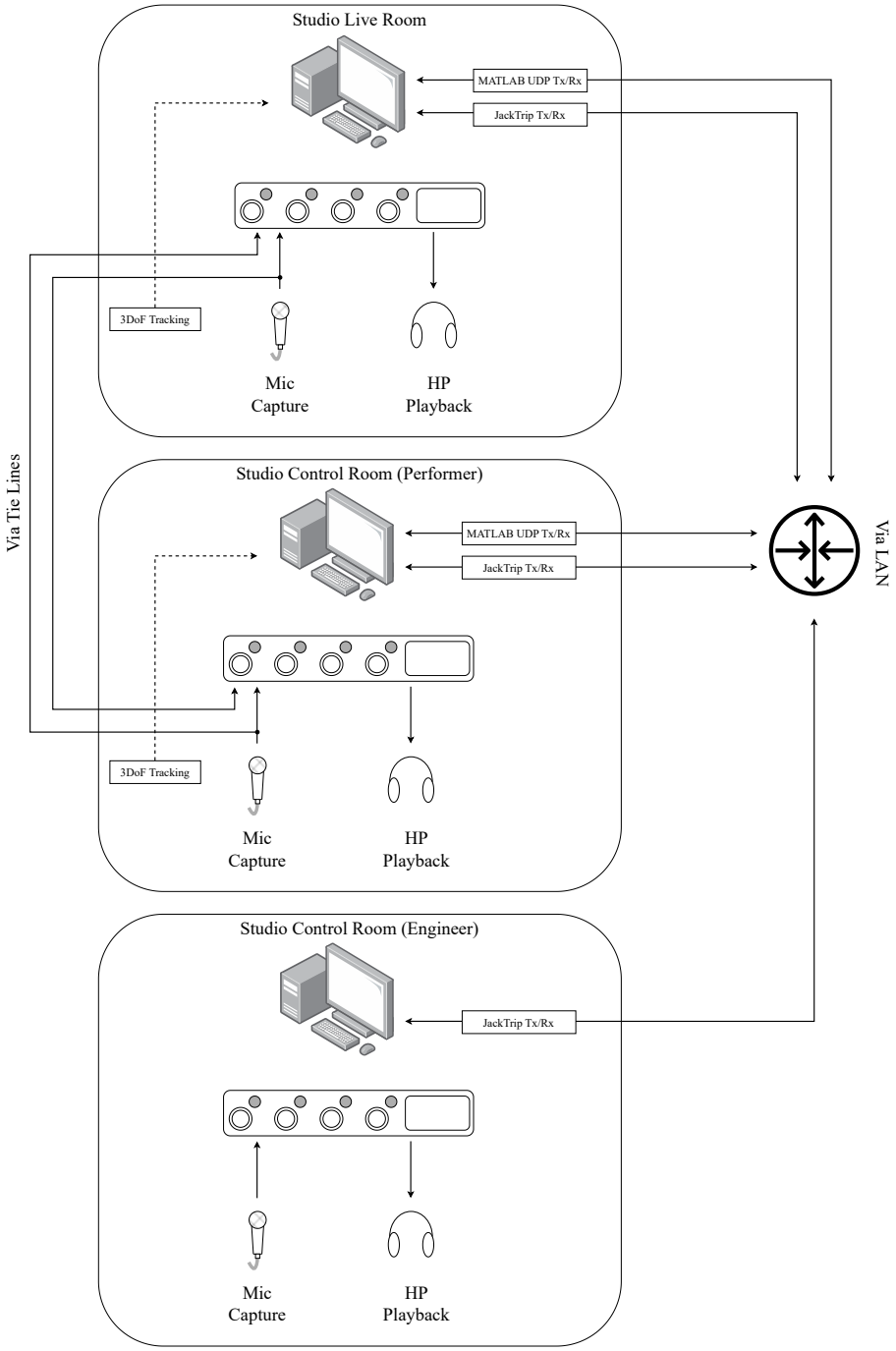


Figure 7.1: Hardware setup deployed for testing across the LAN at the AudioLab studios. Audio analogue routing, head–tracking, JackTrip audio streams and test protocol control application UDP signal flow are detailed.

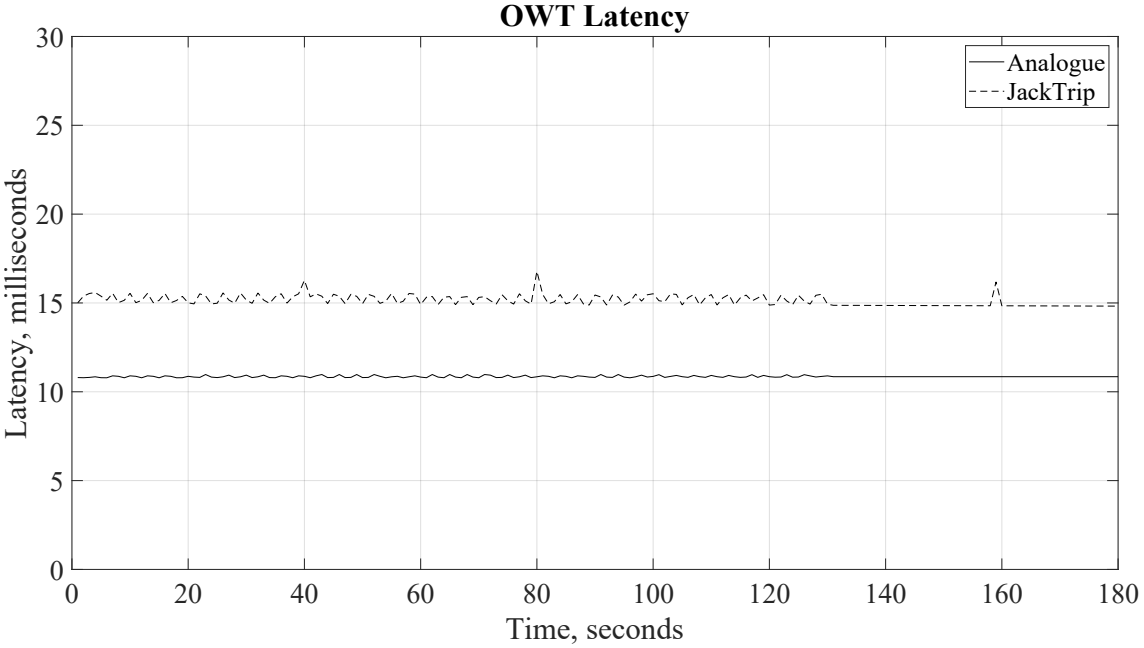


Figure 7.2: Measured OWT latency from local audio capture to remote audio playback for analogue and JackTrip connections between endpoints. Here measurement is performed with immersive audio rendering bypassed.

An additional JackTrip stream is implemented with the same configuration between each endpoint and the researcher laptop device. Here local voice capture at each endpoint is passed to the researcher for monitoring, and each participant endpoint receives the engineer talk–back microphone signal.

7.4.3 Latency Measurement

Latency measurement for the INMP system as deployed was performed by inputting a test signal (1 Hz pulse train with 1 kHz duty cycle) in place of microphone capture at one endpoint, and calculating the time difference for pulse leading edges in the test signal return from relevant outputs at the other endpoint, thereby measuring OWT between performers. A 3–minute test signal measurement was taken to characterise system performance (Figure 7.2).

Direct monitoring latency was measured as sub–millisecond and is considered negligible. Audio rendering throughput latency was measured at a static value of approximately 2.5 ms. The analogue OWT latency between performers was measured as 10.9 ± 0.1 *ms*. Summed with the 2.5 ms audio rendering latency, analogue OWT system throughput is given as 13.4 ± 0.1 *ms*. JackTrip OWT latency between performers was measured as 15.1 ± 1.7 *ms*. Summed with the 2.5 ms rendering latency, JackTrip OWT latency between performers was measured as 17.6 ± 1.7 *ms*. The 2.5 ms of audio rendering latency is associated with throughput in binaural decoding [65] as outlined in subsection 7.4.4.

7.4.4 Immersive Audio Rendering

The INMP system follows a convolution–based virtual Ambisonic approach to interactive auralisation and binaural stereo headphone playback (Figure 7.3). Here each singer of a vocal duet is provided with a simulation of hearing networked performance as if co–located in the simulated acoustic space. The GuildHall, York, was simulated using measured SIRs and processed iterations of this SIR as detailed in section 7.5, where the RT is artificially controlled through modification of damping coefficients. Audio capture channels at each endpoint (analogue input of local and remote performer voice, and JackTrip Rx for remote performer voice) are routed to Reaper[11], the DAW application used to host immersive audio rendering processes as plugins. JACK[12] is used to route audio between hardware capture and playback buffers, and DAW and NMP applications (i.e., Reaper and JackTrip respectively).

Within Reaper audio input channel sends to immersive audio rendering processes are controlled via OSC as outlined in subsection 7.4.5. Additionally an instance of ReaDelay[13] is present on the remote performer JackTrip Rx channel, such that OWT between performers for the network connection may be artificially increased by 20 ms (this may be activated or bypassed via OSC message as outlined in subsection 7.4.5). Inputs from the remote performer may be routed to each of three instances of X–MCFX Convolver[14] configured to provide convolution of mono input with 3OA SIR. Each instance is associated with auralisation of discrete RT conditions (*0.5 s*, *1.5 s* and *2.3 s RT*). SIRs used in auralisation of the remote performer include full direct sound and reverberation information, and describe a source approximately 2 m away in a stage area (as outlined in section 7.5).

Local performer input may similarly be routed to each of three instances of the X–MCFX Convolver providing 3OA SIR convolution. Here the SIRs used in convolution describe measurement for coincident source and receiver positions in the stage area of the Guildhall (again outlined in section 7.5). For 'own voice' auralisation direct sound is rendered as dual mono via the audio interface control software (*Focusrite Control*). This avoids self–monitoring delay, and avoids inaccuracy in direct sound auralisation for coincident source and receiver positions in SIR measurement [5]. Only reverberation is therefore auralised in convolution instances for the local performer, where the direct sound component of relevant SIR is removed (as outlined in section 7.5). Convolution instances in local and remote performer auralisation may be activated or bypassed via OSC control as outlined in subsection 7.4.5, and are configured for zero–latency operation. The output auralised signal from the active convolution for each of the local and remote performer is summed into a single Ambisonic scene. An instance of IEM SceneRotator[15] is used to rotate the Ambsionic signal [183] in response to OSC input from the local performer head–tracking device[16], providing 3DoF interactivity in auralisation.

[11] `https://www.reaper.fm/`
[12] `https://jackaudio.org/`
[13] `https://www.reaper.fm/reaplugs/`
[14] `http://www.angelofarina.it/X-MCFX.htm`
[15] `https://plugins.iem.at/docs/scenerotator/`
[16] `https://plugins.iem.at/docs/osc/`

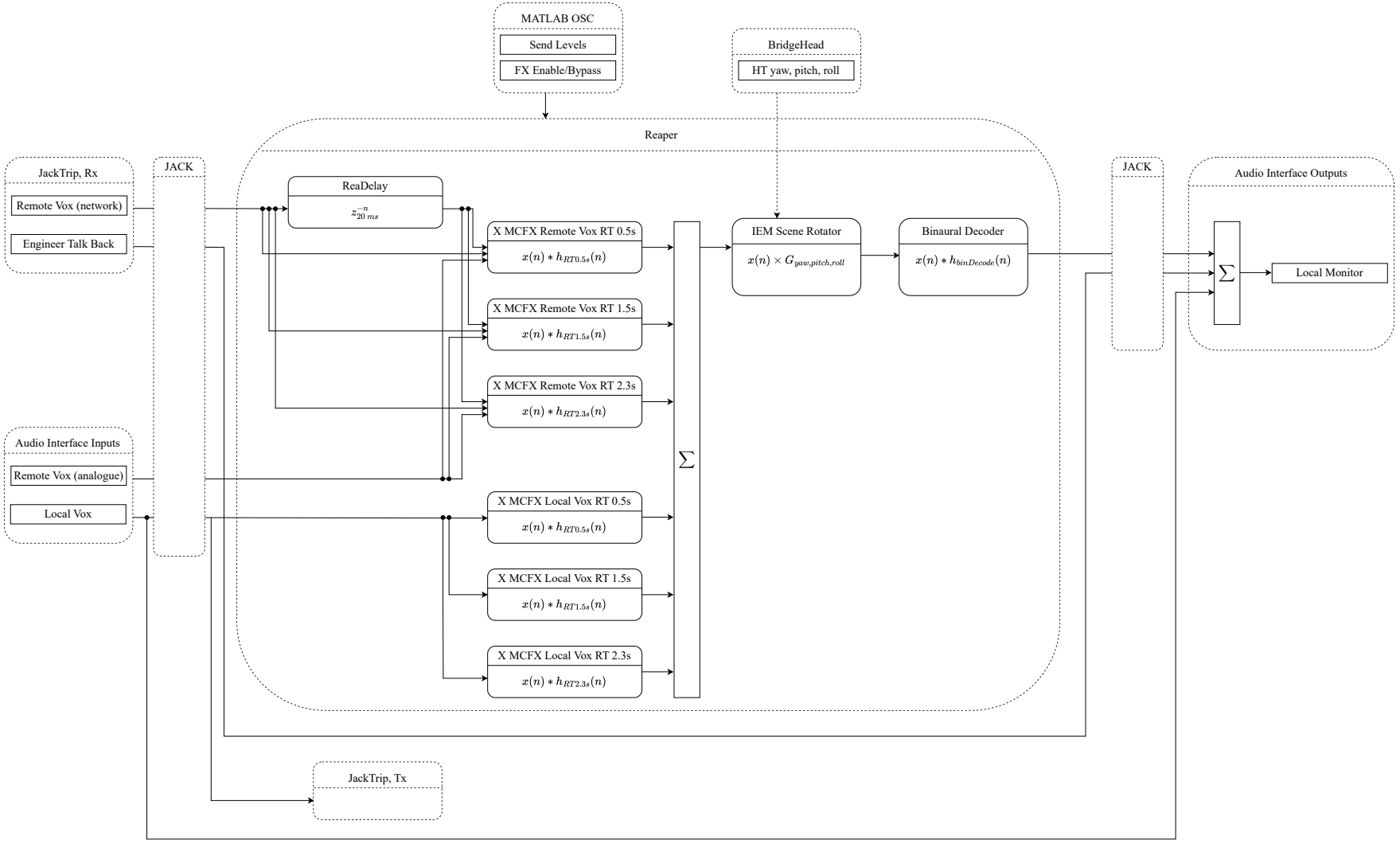


Figure 7.3: Immersive audio processing for the INMP test protocol deployment, detailing relevant applications and signal flow.

The rotated Ambisonic signal is then decoded to binaural stereo using an instance of IEM Binaural Decoder[17] which implements Mag–LS binaural decoding methods [198]. Binaural stereo output from Reaper is routed via JACK to the audio interface outputs, where local performer direct monitoring of the voice is summed into the mix. For the purposes of the test protocol an additional JackTrip receive from a control room endpoint is summed into the monitor output allowing engineer talk–back.

For the local performer's monitoring of their own voice, direct sound monitoring is sub–millisecond, whereas monitoring of auralised reverb through DAW–based auralisation incurs approximately 13 ms latency. The SIR measurement describing the singer's own voice is truncated to the first reflection (Figure 7.9). Thus the offset between direct sound and onset of reverb in monitoring (13 ms) approximates the desired offset between direct sound and first reflection from the measured impulse response (approximately 10 ms).

7.4.5 Test Protocol GUI and Control System

For the purposes of INMP audio system evaluation, a novel test framework was developed using MATLAB[18] which was implemented as an extension to the INMP system. This system was designed to allow control of INMP system configuration across multiple endpoints, such that the controlled progression through test protocols may be supported and supervised by a single research engineer. Additionally the system provides a participant GUI for QoE item rating input (Figure 7.4).

A test protocol sequence and associated set of audio rendering and latency configurations are specified in a spreadsheet. OSC messages [102] for Reaper[19] are used to control INMP system configuration. Here latency configuration is controlled by OSC messages defining send levels from inputs (allowing selection between analogue input and JackTrip Rx), and defining the effects bypass toggle value for an instance of ReaDelay, used to artificially raise the latency of the JackTrip Rx channel by 20 ms. This allows selection between three discrete latency conditions, namely those associated with an analogue baseline connection between endpoints, a JackTrip network connection between endpoints, and a JackTrip network connection between endpoints with delay incremented by 20 ms by the ReaDelay insert.

The audio rendering configuration is controlled by OSC messages defining effects bypass toggle values for relevant X–MCFX convolver inserts. This allows selection between convolutions associated with auralisation using SIRs describing three discrete RT values, namely 0.5 s, 1.5 s and 2.3 s T30, produced from SIR measurement of the GuildHall, York (as detailed in section 7.5). Reaper OSC messages for INMP system configuration are prepared as a spreadsheet. In this manner the MATLAB control application may read a specified test protocol sequence, associated INMP configuration, and required Reaper OSC messages from the prepared spreadsheets. OSC utility for the MATLAB control application is implemented using the sendOSC[20] function.

[17] `https://plugins.iem.at/docs/plugindescriptions`

[18] `https://uk.mathworks.com/products/matlab.html`

[19] `https://www.reaper.fm/sdk/osc/osc.php`

[20] `https://uk.mathworks.com/matlabcentral/fileexchange/31400-send-open-sound-`



Figure 7.4: The participant GUI for the MATLAB INMP evaluation framework.

The MATLAB control application considers progression through a standard performance task–based NMP evaluation protocol. Here the control application at each endpoint configures the INMP system based on the test protocol sequence index, and based on the configuration specified in the spreadsheets. Participants are then prompted to attempt the relevant performance task, and provide rating via a GUI which is saved to a spreadsheet including relevant evaluation conditions.

Participants then indicate that rating is complete via the GUI. This prompts the MATLAB application to send a status notification message via UDP to the remote endpoint, flagging that this participant is ready to progress to the next test protocol sequence index. The system then enters a 'wait' state until a similar status update message is received from the remote endpoint. When this is received the control application increments the index of the test protocol and progresses to repeat this process for the next evaluation condition through to termination at the end of the test protocol (Figure 7.5). In this manner progression through the test protocol is synchronised across the two endpoints.

The test framework is available online as an open resource[21], including the MATLAB control application, Reaper sessions, configuration spreadsheets, and INMP system configuration specified as shell scripts. Guides on the setup and usage of the test framework are also included.

control-osc-messages

[21] https://github.com/SpaceCadetAlba/NMP-Evaluation-Method

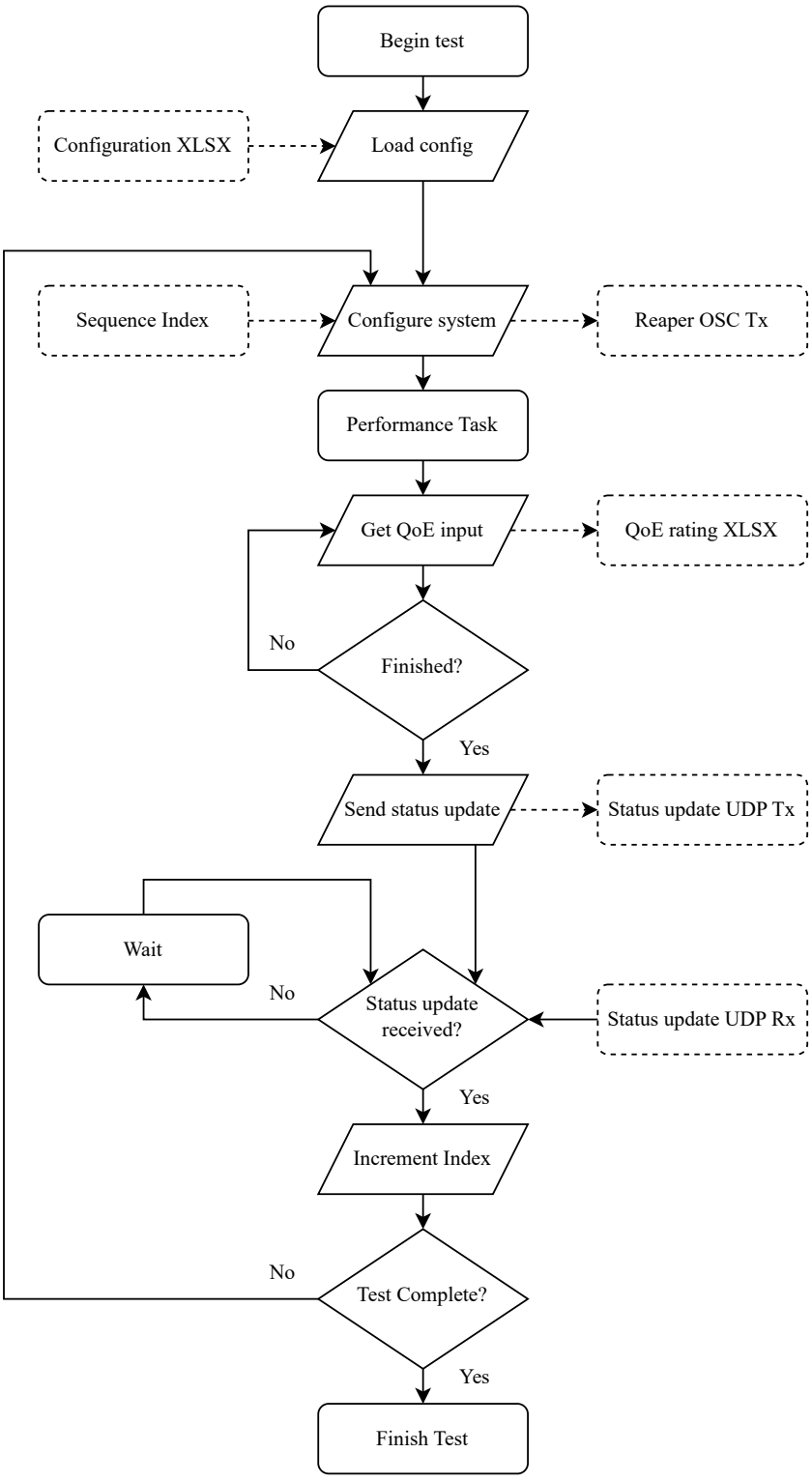


Figure 7.5: Flowchart illustrating progression through the test protocol using the INMP evaluation framework.

7.5 SIR Measurement and Processing

The following subsections detail the measurement and processing of SIR measurements used in immersive audio rendering for the INMP system.

7.5.1 SIR Measurement

The 3OA SIRs used in auralisation were measured in the Guildhall, York. Measurement was performed using the sine sweep method [148], where a *Genelec 8040 A*[22] loudspeaker was used as the source, and the *MH Acoustics em32 Eigenmike*[23] was used as the receiver from which the 3OA signal is derived [187]. A source height of 1.5 m and receiver height of 1.6 m were used. The loudspeaker calibration level was not noted. Two relevant measurements were taken. For the first the source was positioned two meters in front of the receiver position, directed at the receiver position (source position S4 directed at receiver position R1, approximately 2 m apart (Figure 7.6). This measurement is used to auralise listening from position R1 to a singer at position S4. Additionally the coincident source receiver measurement, with both the source and receiver placed at S4, was taken, with the source directed towards position R1. This measurement is used to auralise listening to the singers own voice for position S4.

7.5.2 SIR Processing

In order to avoid incurring additional delay through convolution with zeros (i.e., *'dead air'*) at the beginning of SIR [12], all SIRs were first truncated to the onset of direct sound. Here the SIR beginning is truncated to 2 ms before the peak of the direct sound with a 1 ms linear fade–in (Figure 7.7).

In order to achieve acoustic parameter variation for auralisation in the test protocol, the RT described by each SIR is modified by modifying the decay envelope of the relevant SIR. Decay envelope adjustment is applied using a MATLAB implementation of the method described in [223] applied to the 3OA AmbiX audio format. Here the processed SIR with adjusted decay envelope, $y(t)$ is defined in terms of the unprocessed SIR, $x(t)$ as:

$$y(t) = x(t)e^{-t(\delta_1 - \delta_0)} \tag{7.1}$$

Here the envelope function $e^{-t(\delta_1-\delta_0)}$ describes decay envelope adjustment of $x(t)$ in terms of damping coefficients δ_1 and δ_0, of the adjusted and original IRs respectively. The damping coefficients are expressed in terms of RT:

$$\delta = \ln(10^6)/2RT \tag{7.2}$$

Where the RT for an unprocessed IR is known the enveloping function for any desired RT may be therefore be derived.

[22] `https://www.genelec.com/previous-models/8040a`
[23] `https://eigenmike.com/eigenmike-em32`

In the MATLAB implementation the function required for decay envelope adjustment is defined in terms of the A–Weighted RT (T30) of ACN0 from the original SIR and desired A–Weighted RT. The enveloping is then applied to the AmbiX channels uniformly. The original SIR, with measured RT of approximately 2.3 s was adjusted to render SIRs of 1.5 s and 0.5 s RT (Figure 7.8). Here the decay envelope was specified to be applied from the first reflection onward. Here the first reflection is assumed as the ground reflection, and path difference between direct sound and ground reflection, calculated from source and receiver heights and distance, allows derivation of time difference between direct sound and first reflection.

It can be noted that the original SIRs used are of 16–bit resolution. In decay envelope adjustment an envelope of floats is applied to the 16–bit audio data, which is then written as 24–bit resolution (Figure 7.8). The difference in resolution (and associated dynamic range) between use of 16 bit and 24 bit SIRs is not expected to be perceptually noticeable.

For the SIR used in the auralisation of the singers' own voice (i.e., coincident source and receiver measurement of position S4) the direct sound is removed from the SIR using a custom MATLAB function (Figure 7.9). Removing direct sound serves two purposes. Firstly, the SIR measurement is not truly coincident, with the receiver placed above the source, and thus direct sound auralisation of own voice direct sound through convolution may provide noticeably inaccurate results. Secondly, in INMP contexts direct monitoring of the singer's own voice allows for minimisation in self–monitoring delay.

Acoustic parameters [141] were estimated from ACN0 of each SIR using the MATLAB Acoustic Parameters Toolbox from Alex Southern[24] [209] (Figure 7.10). It can be observed that as the decay envelope is attenuated to a more extreme degree (i.e., adjusted for a shorter reverb time) spectral qualities of the decay envelope are smoothed. Desired RT values are, however, approximated sufficiently (i.e., the measured acoustic parameters of the processed SIR is not considered noticeably different from the specified desired RT). As an alternative to auralisation of rooms with different RT values, the decay envelope adjustment approach is considered as more controlled. Here the arrival time of measured reflections is retained, and the associated frequency response, though smoothed, is retained to a degree. The decay envelope adjustment method, implemented as a modification of damping coefficient, can be considered as an artificial method of approximating acoustic treatment of a real space (for example it will be typical in performance venues to deploy curtains and other acoustic treatment to effectively raise absorption coefficients in a room.

Custom MATLAB scripts associated with the truncation to direct sound, removal of direct sound and decay envelope adjustment are available online as an open source resource[25].

[24] https://doi.org/10.5281/zenodo.6505345

[25] https://github.com/SpaceCadetAlba/IRutilities

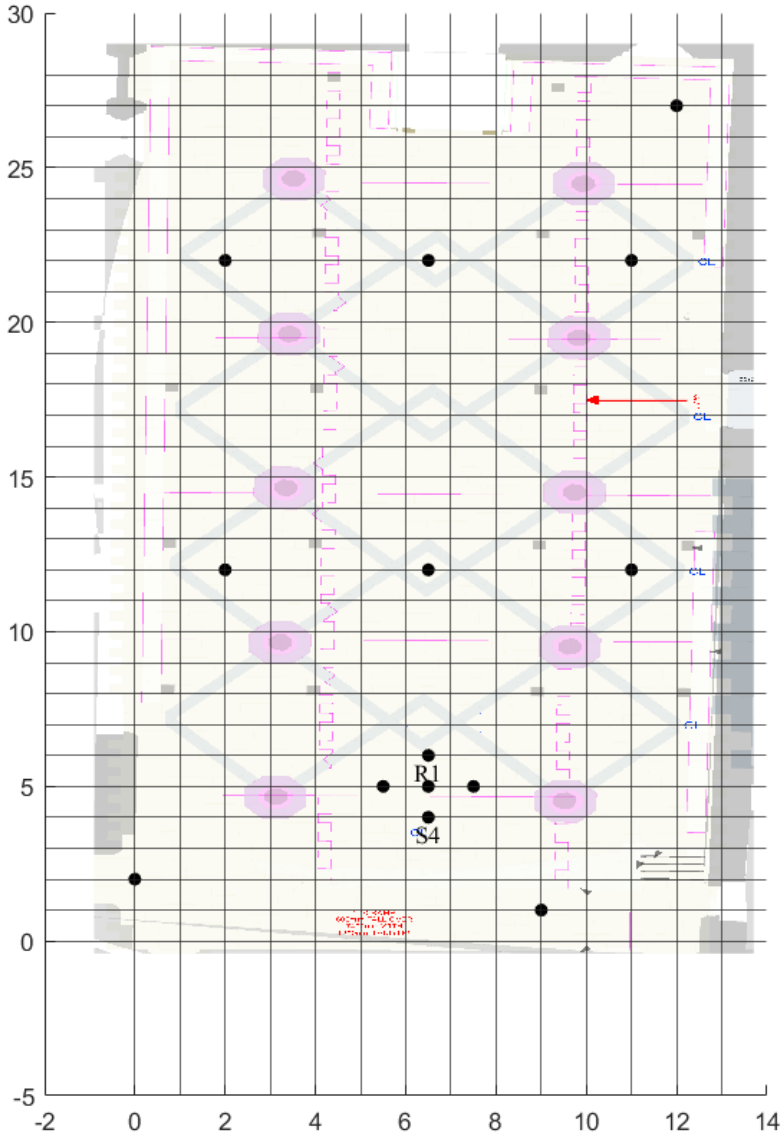


Figure 7.6: SIR measurement positions in the GuildHall, York, detailing source position S4 and receiver position R1.

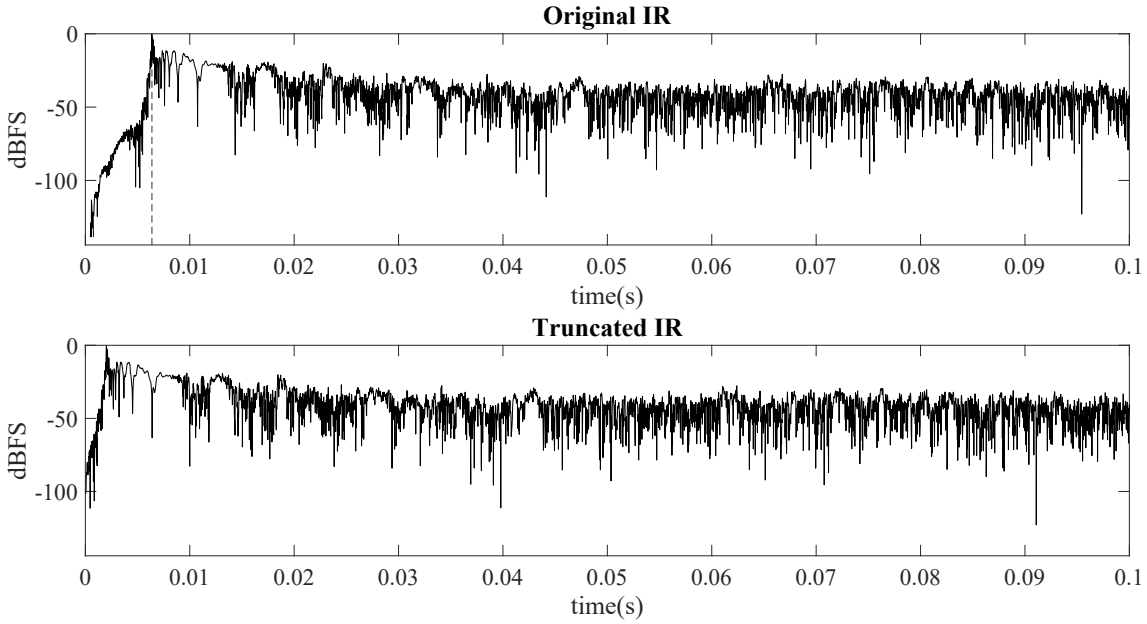


Figure 7.7: Illustration of the truncation of the SIR to the direct sound. ACN0 for the original SIR is shown with the vertical line indicating the determined direct sound peak (above), and ACN0 for the truncated SIR is shown (below).

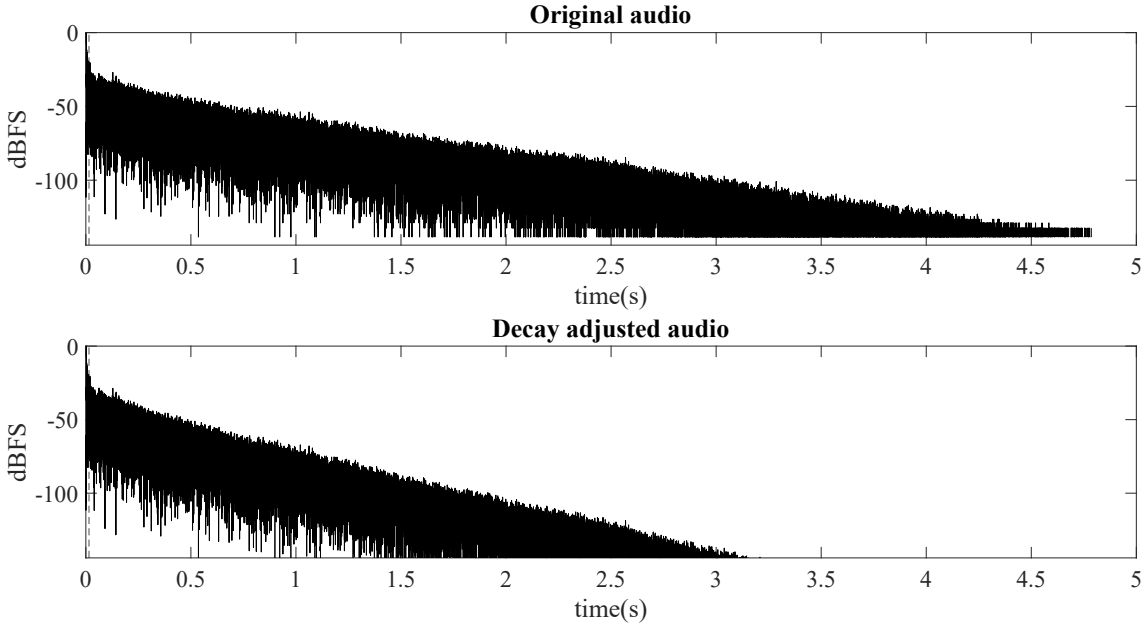


Figure 7.8: Illustration of SIR decay envelope adjustment from approximately 2.3 s to 1.5 s RT. ACN0 is shown, and the vertical line indicates the first reflection time, onward from which decay envelope adjustment is applied so as not to attenuate direct sound.

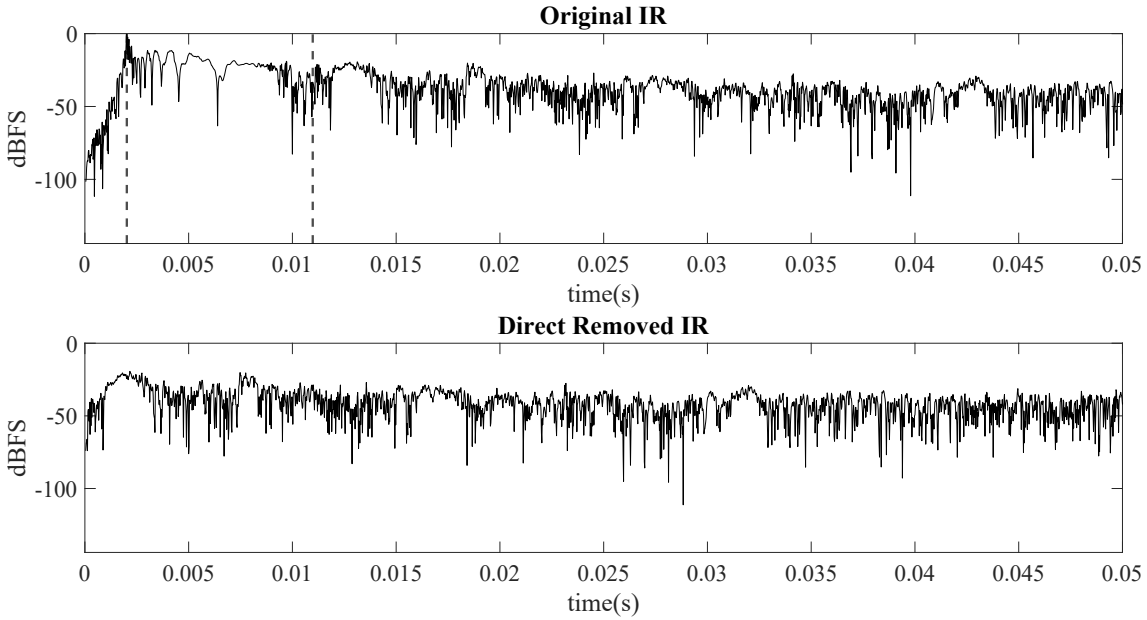


Figure 7.9: Illustration of SIR direct sound removal. ACN0 is shown. The vertical lines indicates the peak of direct sound and the first reflection time to which the SIR is truncated.

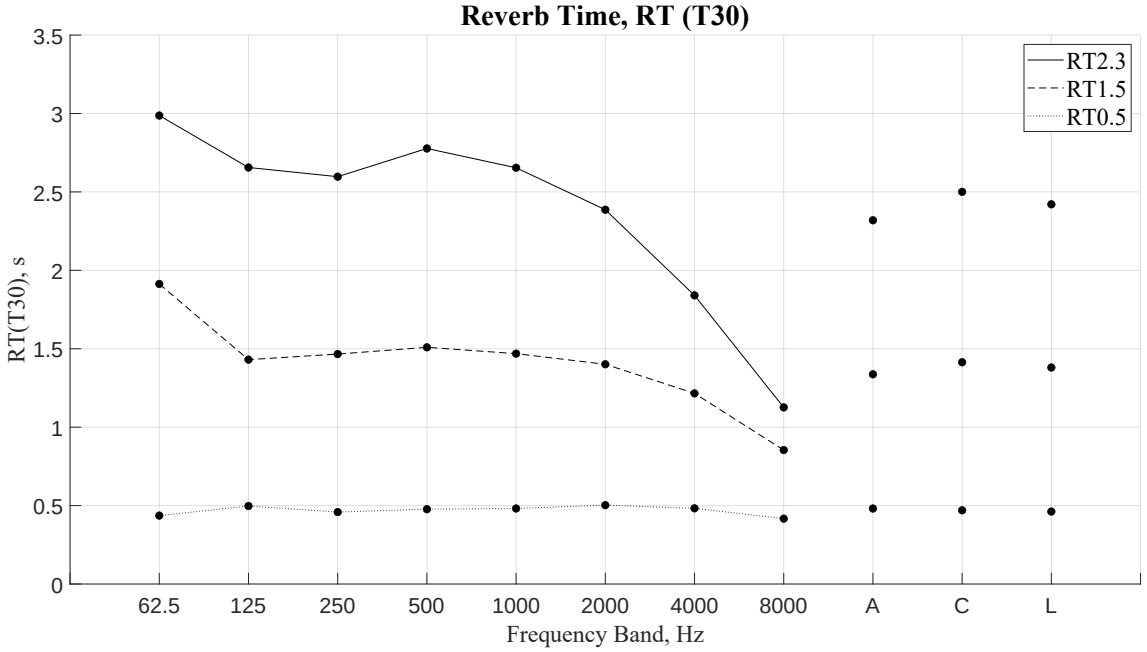


Figure 7.10: RT (T30) for the original SIR (RT2.3), and SIR with decay envelope adjustment applied to approximate RT of 1.5 s (RT1.5) and 0.5 s (RT0.5).A, C and L refer to A–weighted, C–weighted and Linear frequency weighting respectively.

7.6 Evaluation Method

The following subsections provide detail on the protocol and methods used in INMP system evaluation.

7.6.1 Protocol Overview

Evaluation of the impact of RT is conducted using a standard NMP performance task design. Singers participate in duet pairs, with each singer located at a discrete endpoint in a separate room (the studio Live Room and Control Room, Figure 7.1) with no visual contact. The duet pairs use the INMP system to attempt a duet vocal performance task multiple times. For each performance participants provide rating of the performance experience used in QoE evaluation, and audio recording is captured for synchrony analysis. To address the hypothesis that QoE and synchrony will vary with RT the performance task is attempted and evaluated for three discrete RT conditions, namely RT of 0.5 s, 1.5 s and 2.3 s as outlined in subsection 7.6.3.

Considering the objective of practice–oriented evaluation and testing of the hypothesis that latency, or interaction between latency and RT, will impact performance QoE and synchrony, evaluation of RT is repeated for three discrete latency conditions. These are namely 13 ms latency with analogue audio connection between endpoints, 18 ms latency via the JackTrip network connection between endpoints and 38 ms latency via the JackTrip network connection between endpoints with additional added delay, as outlined in subsection 7.6.4.

Participants completed the performance task and related evaluation three times for each discrete combination of the three latency levels and three RT levels (9 discrete groups) for a total of 27 performances. All participants proceed through evaluation following the same randomised order of conditions. Additionally participants were allowed to practice the performance between the two endpoints prior to the performance, where immersive audio rendering is bypassed and dry dual mono monitoring of the JackTrip connection is provided (this is intended to mediate any unfamiliarity factors which may influence performance experiences). The test protocol is monitored by an engineer from a control room endpoint (Figure 7.1).

7.6.2 Participants

Participants were recruited from the general public via outreach with the requirement of involvement in group vocal performance activity. 16 individuals participated in this study. Each participant provided non–compulsory demographic information at the beginning of the test protocol via participant GUI. Self–reported proficiency level ranged from novice–professional (Figure 7.11). Age (Figure 7.12), gender (Figure 7.13) and ethnicity (Figure 7.14) were also recorded for the purposes of demographic monitoring. No participant had prior experience with NMP technology, though some had prior experience singing using teleconferencing solutions.

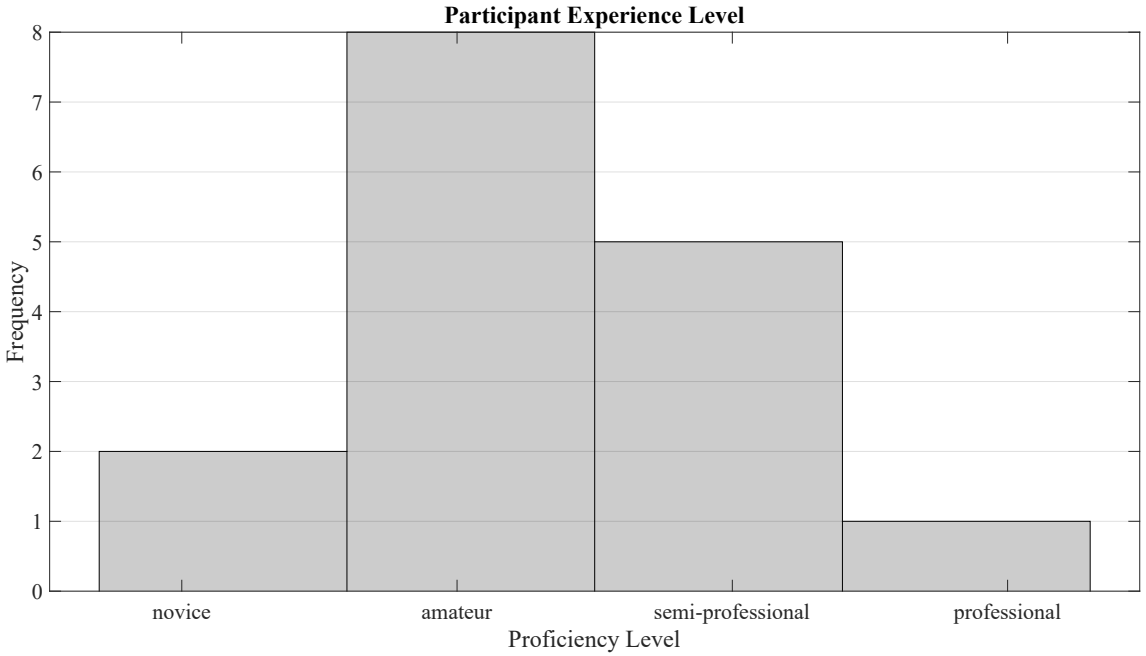


Figure 7.11: Histogram of participant experience level.

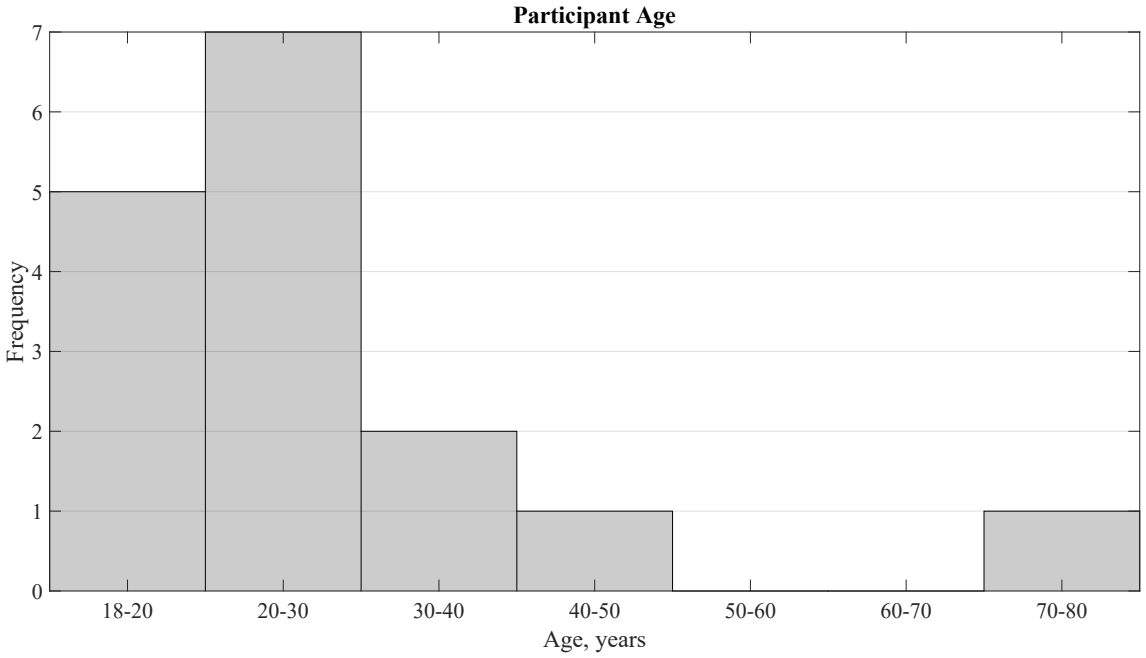


Figure 7.12: Histogram of participant age.

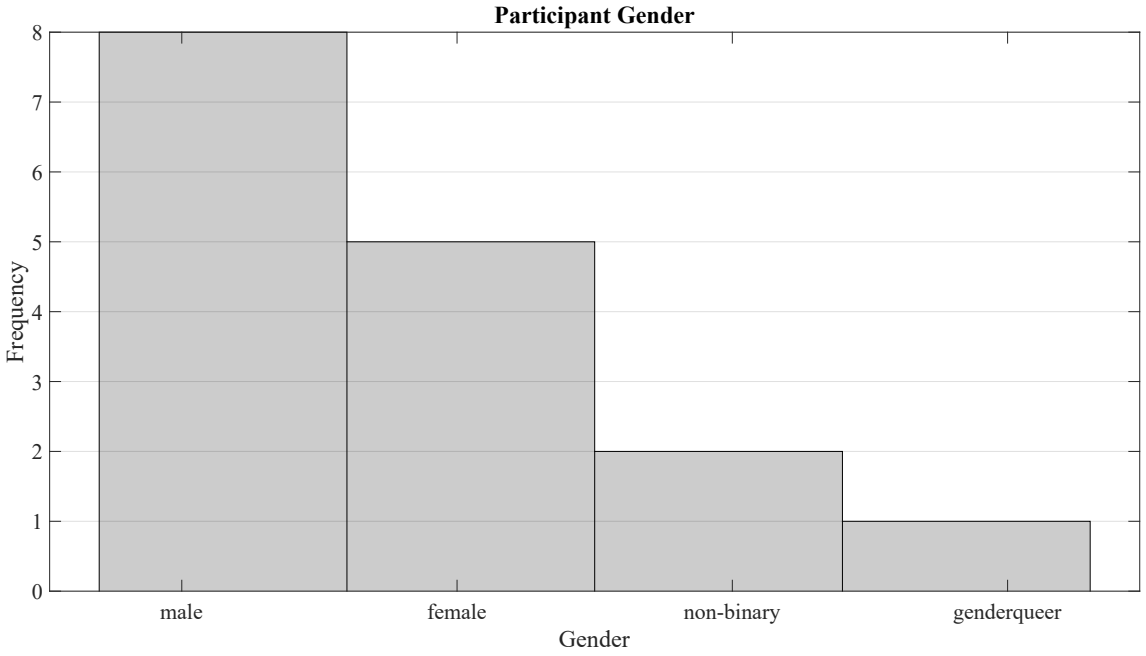


Figure 7.13: Histogram of participant gender.

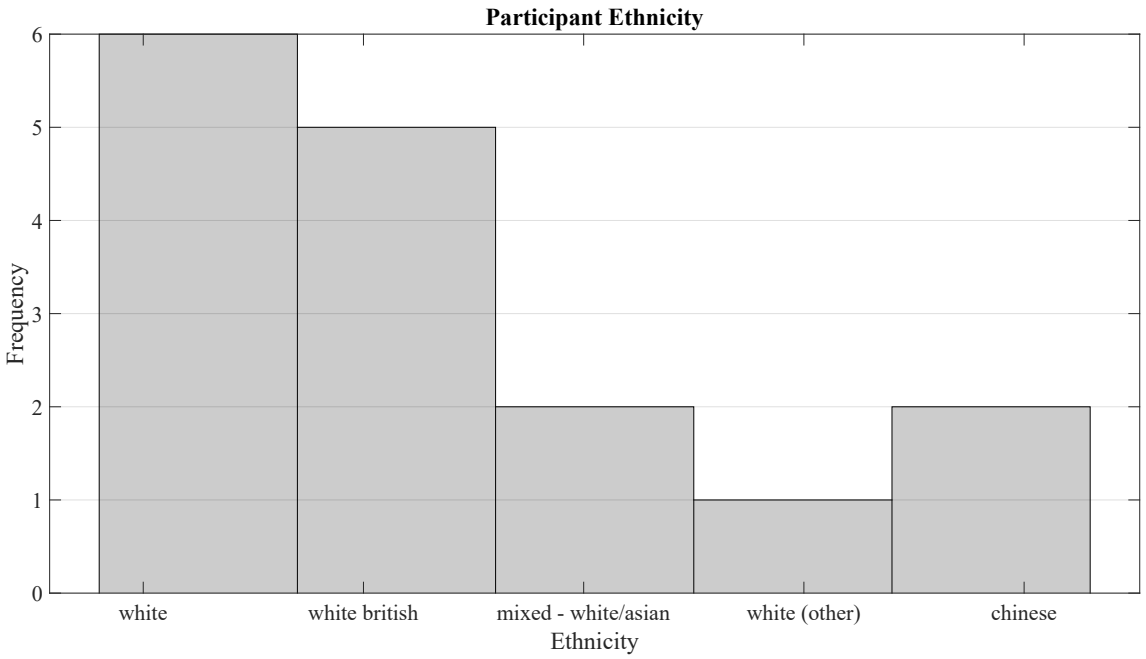


Figure 7.14: Histogram of participant ethnicity.

7.6.3 Acoustic Conditions

The evaluation of three different RT conditions is achieved where the performance task is attempted using the INMP system configured for SIRs describing three discrete RTs. The three different SIRs used in auralisation are as described in section 7.5. These three SIR configurations are associated with RT conditions, namely:

- *RT of 0.5 s*, codified as *RT0.5* for brevity in presentation of results. This RT is considered representative of smaller spaces encountered in practical group singing contexts (for example, a small–medium sized rehearsal space).
- *RT of 1.5 s*, codified as *RT1.5* for brevity in presentation of results. This RT is considered representative of halls typically encountered in group vocal performance (for example concert venues [213]).
- *RT of 2.3 s*, codified as *RT2.3* for brevity in presentation of results. This RT is directly representative of the measured space, namely the GuildHall, York.

Each acoustic condition is defined with by a single RT value for the purpose of brevity in presentation of results. As discussed in section 7.5, however, it is acknowledged that RT varies across the frequency spectrum for each SIR (Figure 7.10).

7.6.4 Latency Conditions

Three different latency conditions are considered in the evaluation. These conditions are defined so as to provide baseline conditions, providing reference measurements in evaluation. Latency conditions are also defined such that evaluation conditions are representative of values reported from practical WAN deployment of a system implementation in prior work [8]. The three discrete latency conditions evaluated are namely:

- *13 ms (analogue)*, codified as *L13* for brevity in presentation of results. This is the OWT audio delay between performers rounded to the nearest millisecond for the analogue audio connection between endpoints as detailed in subsection 7.6.4. This is considered as reference–quality baseline conditions, where no network delay is incurred.
- *18 ms (network)*, codified as *L18* for brevity in presentation of results. This is the OWT audio delay between performers rounded to the nearest millisecond for the JackTrip connection between endpoints as detailed in subsection 7.6.4. This is considered to be representative of best–case OWT latency in practical WAN deployment.
- *38 ms (network)*, codified as *L38* for brevity in presentation of results. Here the OWT of the JackTrip connection is raised by 20 ms as detailed in subsection 7.4.4. This is considered to be representative of the high end of OWT latency experienced in practical WAN deployment over a national range.

7.6.5 Performance Task

The performance task attempted by participant pairs is a duet vocal performance of the melody of *Freres Jacques* in rounds. Here performers sing using only the /*ta*/ phonetic in order to control acoustic features of expression in onsets in performance for synchrony analysis. Each performance contains 32 onsets per–part (giving 31 IOI samples), and 18 coincident onsets between parts (giving 18 OTD samples). Participants were provided with sheet music and example audio for the performance task prior to participating in evaluation.

7.6.6 Experience Rating Questionnaire

For each attempt of the performance task participants provide evaluation of the experience through rating agreements with the following statements using a 7–point rating scale ranging from strong disagreement (1) to strong agreement (7) via a neutral point (4):

1. The audio delay between myself and duet partner is not noticeable.
2. I am able to play in time with my duet partner.
3. I can hear my duet partner's voice clearly.
4. The loudness of the reverberation is good for singing in.
5. The length of the reverberation is good for singing in.

Item 1 addresses the standard NMP QoE item of perception of delay [8], [9], [19], [35], [114], [118], here specifically addressing the degree to which delay was noticeable. Item 2 addresses the standard NMP QoE item of perception of performance quality [8], [9], [19], [20], [35], [109], [113], [114], [118], [120], [121], here specifically addressing the perceived degree to which performers were able to play in time.

Whereas prior INMP studies have addressed perception of reverberation with respect to aspects such as *responsiveness* [8], [19], this study approaches rating of reverberation using items derived from standardised definitions and rating items [180]. Items 4 and 5 address perception of reverberation in terms of level and length, both notably low–level concepts relating to physical parameters of room acoustics. Here perception of *reverberation level* and *reverberation length* is addressed with respect to perceived performance conditions. Item 3 addresses perception of *clarity* with respect to perceived performance conditions. Here definition of clarity can be considered relevant to both the *clean* and *clarity* attribute definitions.

7.6.7 Synchrony Metrics

Synchrony measurement provides an objective physical measure of performance quality. Here metrics are derived from IOI within–parts and OTD between–parts. Onset detection for synchrony analysis was achieved using the MM–BOP method detailed in subsection 7.6.8.

For measurement of within–subject synchrony the standard metric of *Tempo Slope* is used as in empirical work [6], [57], [58], [211] and research [8], [19], [113], [115]. This is defined as the gradient of the linear best fit of tempo samples across a performance. Though some work measures *Tempo Slope* relative to the mean BPM of performance as a percentage[26] [19], [115], in this study *Tempo Slope* is measured in BPM per second, as in empirical work [57], [58]. Reference values for *Tempo Slope* are available from prior work, where mean *Tempo Slope* of deceleration in excess of approximately 0.2–1 BPM per second is considered indicative of synchrony impairment [57], [58], [115], [211].

As detailed in subsection 2.8.1 various definitions exist for between–parts synchrony metrics. These can generally be identified as expressing different ways of averaging of absolute OTDs, averaging signed OTDs, and measuring variation in OTDs. These three characteristics of OTDs are addressed in this study using the metrics of *Precision* [8], [56], [224], *Lag* [8], [56], [224], and *Asynchrony* [8], [55].

Precision is defined as the mean absolute OTD across performance. Reference measurements for live performance are available from prior studies in the range of 50–70 ms [56], [224]. *Asynchrony* is defined as the mean of the standard deviation of OTD magnitude across parts. In this study, where only one OTD set is derived from duet performance, this is simply the standard deviation. Here reference values from live performance range from approximately 30–50 ms [8], [55], [56], [224]. *Lag* is defined as the median signed OTD, where reference measurement for live performance in prior study ranges across $\pm$ 30–40 ms [8], [56], [224]. Lag which exceeds this magnitude is considered in the field of NMP to be indicative of leader–follower latency–coping strategies [6], [129].

[26] I.e., effectively normalising *Tempo Slope* measurement across the range of possible tempos.

7.6.8 Onset Detection Method: MM–BOP

In synchrony analysis temporal metrics describing performance are derived from associated onset times. Onset times are typically measured using onset detection software based on MIR approaches [107][27]. Here an 'onset' is defined with respect to MIR feature parameters derived from audio recording (for example dynamic, spectral, pitch and phase properties), allowing labelling of onset times from analysis of audio in onset detection algorithms. Usefully this provides a controlled definition of onset times across synchrony analysis, and avoids time–consuming tedious manual annotation.

Particularly for the case of vocal performance, however, typical MIR–based onset detection algorithms are well understood to perform[28] poorly with respect to accuracy and repeatability [224], [226]. This poor performance in onset detection can result in errors in synchrony analysis [227]. Though voice–science informed onset detection approaches[29] [8], [224] can provide improved performance, results still contain false negatives and positives. In practice, achieving accurate synchrony from onset detection requires the development of bespoke software and a degree of time–consuming manual post–processing to achieve accurate results.

In order to overcome these challenges a novel onset detection workflow was implemented as a MATLAB wrapper function for the MIR Toolbox[30] [210] onset detection algorithm. Here, manually input marker metadata, describing coarse annotation of onset times, is used to partition audio for onset detection. Within each audio partition the MIRONSETS function returns results of onset detection. The wrapper function then checks for the case of false positives (i.e., more than one onset returned where only a single onset was indicated to exist by marker metadata), and provides onset candidate sorting. The best candidate is defined as the one with the smallest temporal distance from marker metadata, and the return of a single onset is forced. This effectively eliminates false positives. Where no onset is detected (false negative) the script will stop running and the relevant onset index may be noted for manual handling. This eliminates false negatives from onset detection. This method therefore eliminates error in onset detection from synchrony analysis, given manual markers annotated onsets with a sufficient degree of accuracy.

This novel approach, named MM–BOP (Figure 7.15), accepts marker metadata input from marker export data from Reaper[31], and has since been extended by Sarabeth Mullins [228] to allow for input from marker export data from Sonic Visualiser[32].

[27] Statistical models have also been used in some cases [107]. Machine–learning methods are also explored as an approach to onset detection [225].

[28] Performance here is expressed in the metric *'F–measure'*, describing false negatives and false positives in onset detection with respect to a specified error tolerance of temporal accuracy [224].

[29] `https://github.com/SpaceCadetAlba/TIMEX-Lite`

[30] `https://github.com/olivierlar/mirtoolbox`

[31] `https://www.reaper.fm/`

[32] `https://www.sonicvisualiser.org/`

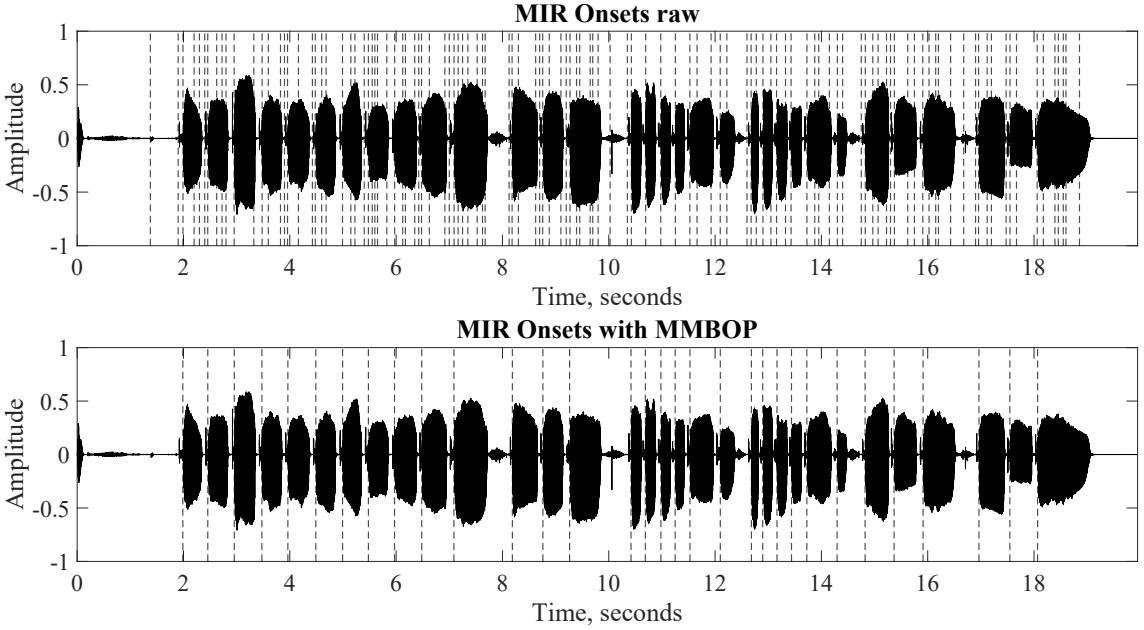


Figure 7.15: Detected onsets for the raw output of the MIR Toolbox MIRONSETS function (above, indicating 116 onsets including 84 false positives) and the output with the MM–BOP function implemented (below, indicating 32 onsets with no false positives or negatives).

The wrapper function itself has been contributed as an open–source resource[33] [229]. As additional utility the script also allows for computation of tempo from onset times given an array of known beat divisions associated with onset marker metadata.

Synchrony analysis for each part was performed using the MM–BOP onset detection approach as a wrapper function for MIRONSETS (Figure 7.16). The function is extended with a custom MATLAB function for tempo analysis (also available in the MM–BOP repository) and calculation of OTDs and between–parts synchrony metrics, returning *Tempo Slope* (Figure 7.17), *Precision*, *Asynchrony* and *Lag* measurements.

7.6.9 Statistical Analysis Method

Given that for ordinal questionnaire rating response normality cannot be assumed, and given that synchrony metrics such as *Tempo Slope* are typically bimodal (i.e., tempo slope is more likely to be slightly above or below zero rather than stable at zero), where non–normality is again likely, and given various standard ANOVA assumptions, such as normality and independence of groups, are violated, it may be reasonable to consider non–parametric approaches (i.e., Friedman ANOVA and Wilcoxon rank–sum pairwise comparisons) are most appropriate with respect to safety against Type I error[34].

[33] `https://github.com/SpaceCadetAlba/MM-BOP`

[34] I.e., false positives in significance results.

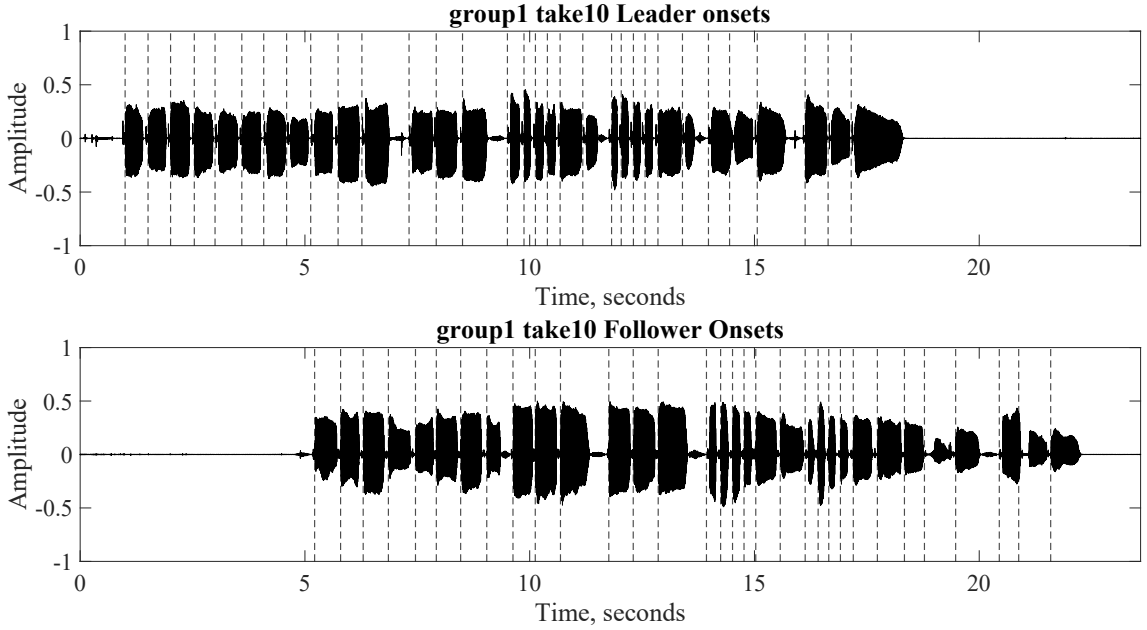


Figure 7.16: Example of onset detection using the MM–BOP method in synchrony analysis.

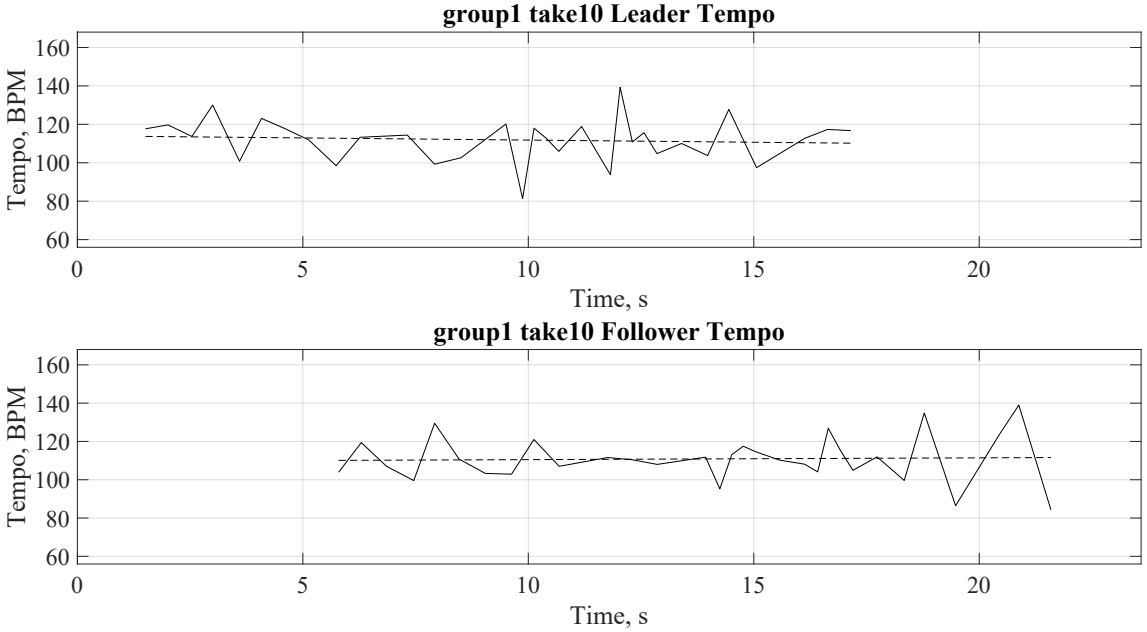


Figure 7.17: Example of *Tempo Slope* calculation in synchrony analysis.

It is, however, most typical to consider ANOVA methods in NMP evaluation, both with respect to empirical [57], [58], [211] and recent [8], [19], [121] literature. It can be recognised that ANOVA methods are extremely robust to assumption violation (i.e., risk of Type I error) [230]. From practical testing of different analysis methods it was indeed found that, where Bonferroni confidence interval adjustment was included in factorial ANOVA and pairwise comparisons, the analysis design is more robust against Type I error than non–parametric alternatives (at least considering MATLAB implementation of factorial ANOVA[35] [231], Friedman ANOVA[36] [232], and Wilcoxon rank sum implementations[37] [233] used in this study). In the interest of protecting against Type I error, and in the interest of coherence with wider NMP research, factorial ANOVA methodology was therefore adopted. Despite the disregard for ANOVA assumptions this is considered as the safest approach. Normality–testing is reported nonetheless, and where significance is borderline non–parametric follow–up is conducted to verify that there is no Type II error[38] in results.

7.6.10 Sample Size Calculation

A–priori sample size calculation was performed using G Power[39] power analysis software. A factorial ANOVA design including main and interaction effects was selected. The two independent variables here are RT and latency, each with three levels. The number of groups here is therefore nine, and degrees of freedom in statistical analysis is four. Large effect sizes are reported in relevant previous work [8], however a medium effect size is considered in power analysis in order to over–provision sample size estimates. A resultant total sample size of 302 was estimated, for an actual study power of 0.95.

Considering nine groups, 34 data points are required per–group to achieve the desired study power. For QoE rating 16 participants each provide 3 ratings for each discrete group (48 samples). This exceeds the required sample size. For within–parts synchrony analysis 14 participants each provide 3 samples for each discrete group for a total of 42 samples, again exceeding required sample size. For between–parts synchrony analysis seven duet pairs each provide three samples per discrete group, for a total of 21 samples per group. This is short of the 34 data points required per group for a medium effect size. For a large effect size, as reported specifically in the case of between–parts synchrony measures in related prior work [8], however, a total sample size of 112 data points is required. This is 13 per discrete group, which between–parts synchrony measurement here exceeds.

[35] `https://uk.mathworks.com/help/stats/anovan.html`

[36] `https://uk.mathworks.com/help/stats/friedman.html`

[37] `https://uk.mathworks.com/help/stats/ranksum.html`

[38] I.e., false negatives in significance results

[39] `https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower`

7.6.11 Visual Contact

Prior work has demonstrated that visual contact between singers will support more precise synchrony than conditions with no visual contact [56], [224]. In NMP literature, however, visual contact via video display is considered potentially problematic due to AV asynchrony between high–latency video system throughput and low–latency audio system throughput [15], [97], [100], [128]. No visual contact is provided in this study, in order to control expected but unknown impact of visual contact via video display in networked musical interactions.

7.6.12 Study Limitations

Though microphone levels were attempted to be limited to a -18 dB peak in audio capture, participants inevitably varied significantly between calibration and actual performance. As such there is a lapse in level control in this study, which may be reasonably expected to introduce noise to results, particularly with respect to rating of clarity.

NMP is considered to improve with practice [6], [35], [116]. It can be acknowledged that all participants completed the evaluation protocol in the same sequence. This may be considered a lapse in control, as it could be reasonably expected for synchrony to improve across the sequence. Participants were allowed a practice period at the beginning of the evaluation protocol to mediate this.

The majority of participants are of *novice* or *amateur* self–reported proficiency level. It can reasonably be expected that singer proficiency level may impact synchrony and experience of performance, and that this presents a lapse in control in evaluation. It is considered that a population of semi–professional and professional singers may provide more consistency in evaluation. Participants delivered performance in evaluation at various tempos (Figure 7.18). Although it was intended that participants establish their own tempo it can be acknowledged that tempo is known to impact perception of latency in NMP [108], [115], [119]. It is therefore acknowledged that tempo variation between groups is a lapse in control in this study. Tempo variation is not extreme, however, and not expected to skew results.

7.7 Results

The following subsections report the results of the INMP system evaluation presented in this chapter.

7.7.1 Data Collection Errors

An error in the protocol control system for group 8 resulted in missing data points in their QoE rating. Here QoE rating data was lost for one take of *RT2.3L38* conditions for one participant. For the other participant in this group one take of each of *RT0.5L18*, *RT2.3L13*, *RT2.3L18* and *RT2.3L38* was lost. These missing data points were replaced with neutral ratings to maintain equal size in groups for analysis. This is not considered likely to skew results. This error caused an index in the test sequence to be skipped.

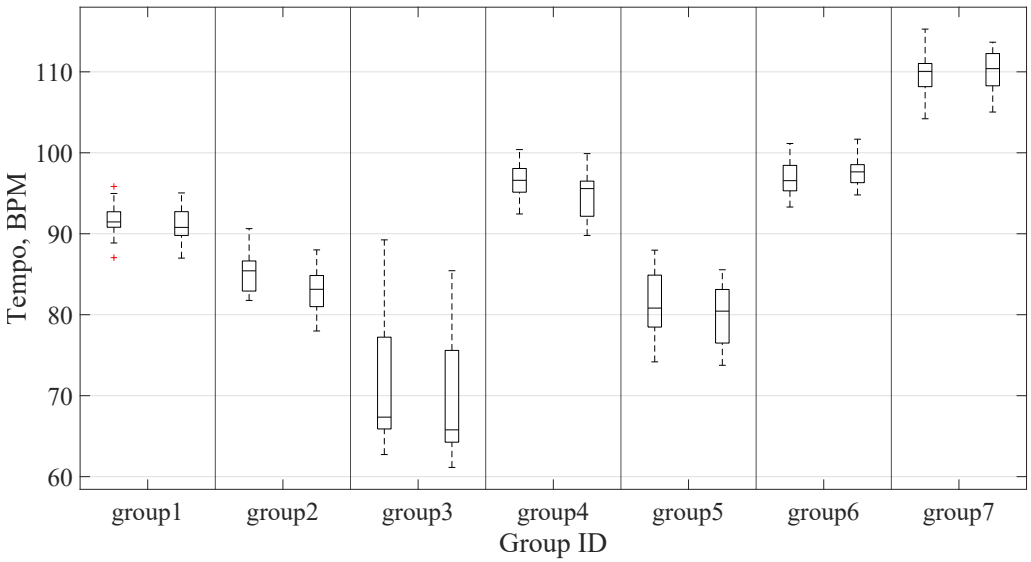


Figure 7.18: Example of Mean tempo for participant pairs by group, illustrating variation in participant choice of tempo.

Though QoE ratings are saved with note of evaluation conditions and could be safely retained, audio recording are not similarly tagged, and thus audio recordings could not be reliably correlated with evaluation conditions. Synchrony data for group 8 was therefore not collected.

Error in performance task delivery by a participant also resulted in missing data. For group 5 the follower part does not perform the last 6 notes of the performance task for one take of *RT0.5L38* conditions. These onsets were therefore removed from synchrony analysis. For group 6 the follower part omits the last 3 notes for one take of *RT0.5L18* conditions. These onsets were therefore removed for synchrony analysis. For one take *RT2.3L18* one participant deviates from the melody and timing of the performance task (i.e., performs the wrong part). In this case synchrony data for this take is discarded and replaced with means from other takes for these conditions to maintain equivalence in group size for analysis. This is not considered likely to skew results, as this only accounts for 64 onset times from a dataset of 12,096 onset times (0.005% of data).

7.7.2 Normality Tests

Ordinal data from questionnaire rating response by definition cannot be normally distributed[40] and therefore was not subjected to normality tests. Shapiro–Wilk normality testing was performed for synchrony metrics. For *Tempo Slope* (Table 7.1) group *RT0.5, L13* was identified as of non–normal distribution. For *Precision* (Table 7.2) groups *RT0.5, L18*, *RT0.5, L38*, *RT1.5, L13* and *RT1.5, L18* were identified as of non–normal distribution. For *Asynchrony* (Table 7.3) detected non–normal distributions were detected for groups *RT0.5, L18*, *RT0.5, L38*, *RT1.5, L13*, *RT1.5, L18*, *RT1.5, L38*, *RT2.3, L13*. Testing of groups for *Lag* (Table 7.4) identifies non–normal distribution for group *RT1.5, L38*.

[40]Normal distribution requires continuous numerical data.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
p	**0.0439**	0.3044	0.4502	0.5050	0.5927	0.4248	0.0841	0.1385	0.5050

Table 7.1: Shapiro–Wilk normality test results for *Tempo Slope* by group. Significant results (i.e., *not normal* distribution) is highlighted in bold.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
p	0.4734	$<$ **0.0001**	**0.0010**	$<$ **0.0001**	**0.0034**	0.1545	0.3022	0.1116	0.1442

Table 7.2: Shapiro–Wilk normality test results by group for Precision. Significant results (i.e., *not normal* distribution) is highlighted in bold.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
p	0.2870	**0.0042**	**0.0185**	**0.0165**	**0.0087**	**0.0191**	$<$ **0.0001**	0.6817	0.5663

Table 7.3: Shapiro–Wilk normality test results by group for Aysnchrony. Significant results (i.e., *not normal* distribution) is highlighted in bold.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
p	0.4249	0.0552	0.0865	0.0539	0.2082	**0.0349**	0.1533	0.4713	0.1161

Table 7.4: Shapiro–Wilk normality test results by group for Lag. Significant results (i.e., *not normal* distribution) is highlighted in bold.

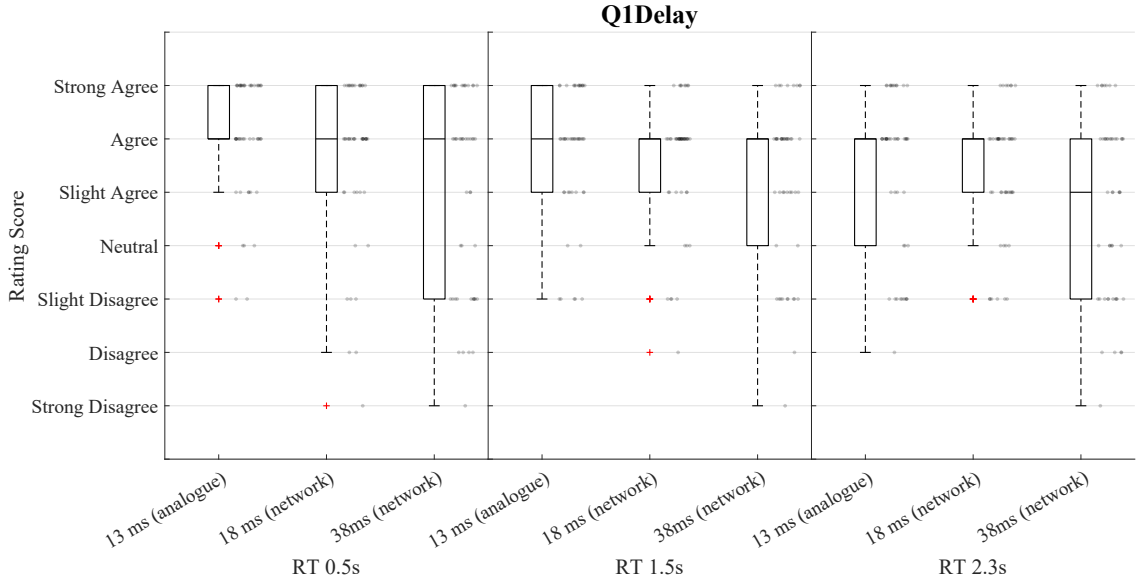


Figure 7.19: Box plot across RT and latency factors for participant rating of questionnaire item 1, addressing perceived audio delay between performers.

7.7.3 Experience Rating: Q1, Perception of Delay

In the rating of questionnaire item 1 (Figure 7.19), addressing perceived delay in the performance experience, ANOVA results report a significant effect of both Latency ($F = 10.6308, p < 0.0001, \eta^2 = 0.0503$) and RT ($F = 3.3851, p = 0.0348, \eta^2 = 0.0160$). No interaction effect of RT and Latency was identified ($F = 0.4092, p = 0.8021$), indicating independency in effects. Bonferroni–adjusted pairwise comparisons (Table 7.5) identify several significant differences.

Bonferroni–adjusted pairwise comparison by the discrete factor of latency (Table 7.6, Figure 7.20) identifies that *13 ms (analogue)* and *18 ms (network)* latency conditions achieve significantly superior rating to *38 ms (network)* latency conditions. Bonferroni–adjusted pairwise comparison by the discrete factor of RT, however, suggests no significant difference (Table 7.7, Figure 7.21). This would indicate that the significant difference reported in the ANOVA results is a Type I error, and should not be accepted. Given the borderline significance results, and given the risk of Type II error associated with ANOVA assumption violation, non–parametric follow–up was also conducted (namely Wilcoxon rank sum pairwise tests). Here *RT2.3* was reported as significantly different from *RT0.5* ($z = 3.5059, p = 0.0005$) and *RT1.5* ($z = 2.2068, p = 0.0273$) (Table 7.8). Relevant ANOVA pairwise comparison is therefore acknowledged as potentially a Type II error. This would suggest that latency is perceived as more noticeable for *2.3 s RT* conditions.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	**0.0217**	1	1	0.0924	0.8817	0.7339	**0.0003**
RT0.5, L18	-	-	1	1	1	1	1	1	0.0734
RT0.5, L38	-	-	-	0.6082	1	1	1	1	1
RT1.5, L13	-	-	-	-	1	1	1	1	**0.0217**
RT1.5, L18	-	-	-	-	-	1	1	1	0.1159
RT1.5, L38	-	-	-	-	-	-	1	1	1
RT2.3, L13	-	-	-	-	-	-	-	1	0.8817
RT2.3, L18	-	-	-	-	-	-	-	-	1
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.5: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 1, addressing perception of delay. Significant differences are highlighted in bold.

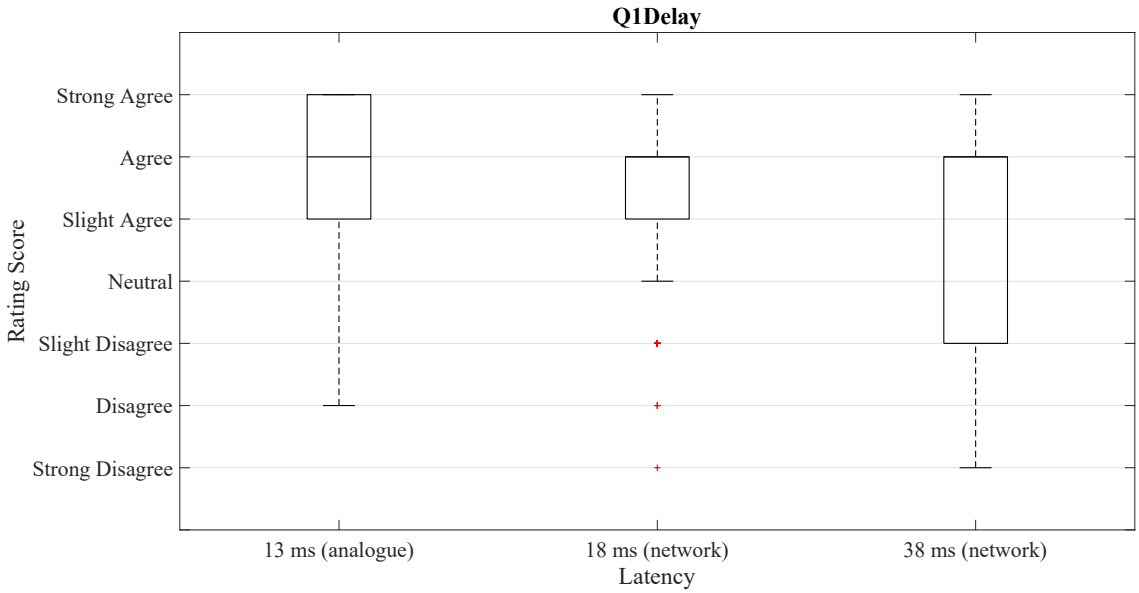


Figure 7.20: Box plot across the factor of latency for participant rating of questionnaire item 1, addressing perceived audio delay between performers.

	L13	L18	L38
L13	-	0.7682	$<$ **0.0001**
L18	-	-	**0.0033**
L38	-	-	-

Table 7.6: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of latency for rating of questionnaire item 1, addressing perception of delay. Significant differences are highlighted in bold.

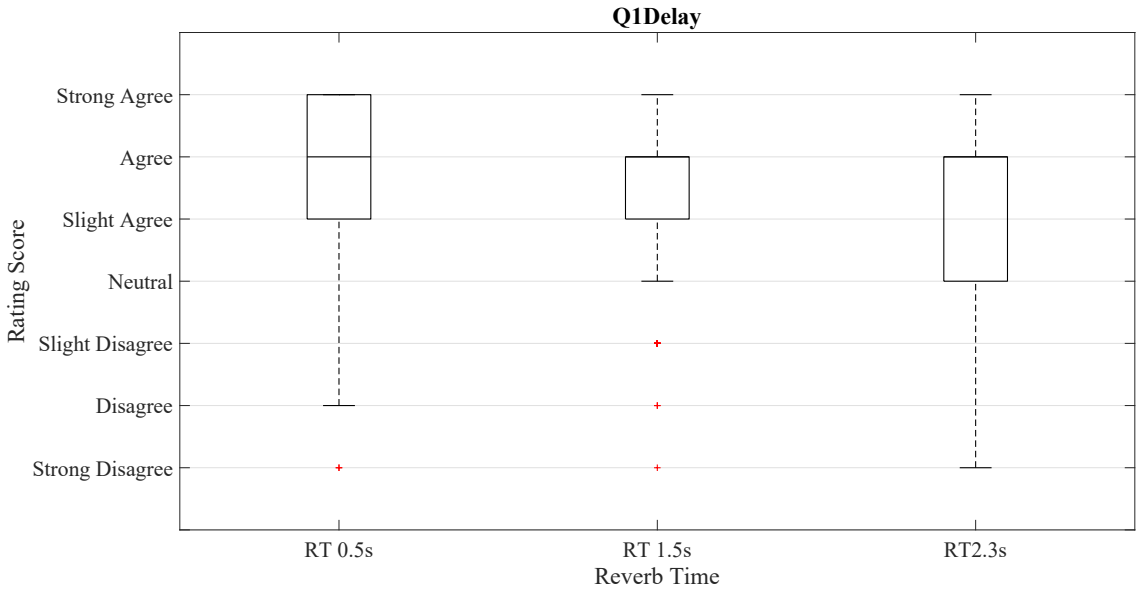


Figure 7.21: Box plot across the factor of RT for participant rating of questionnaire item 1, addressing perceived audio delay between performers.

	RT0.5	RT1.5	RT2.3
RT0.5	-	1	0.0509
RT1.5	-	-	0.1528
RT2.3	-	-	-

Table 7.7: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 1, addressing perception of delay.

	RT0.5	RT1.5	RT2.3
RT0.5	-	0.1046	**0.0005**
RT1.5	-	-	**0.0273**
RT2.3	-	-	-

Table 7.8: Wilcoxon rank sum test significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 1, addressing perception of delay.

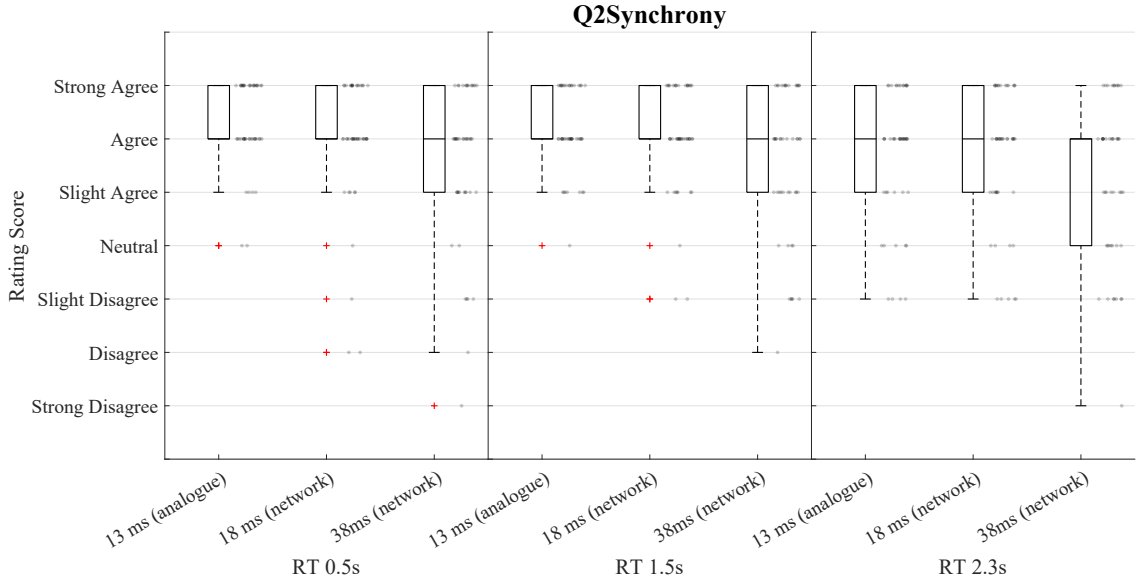


Figure 7.22: Box plot across RT and latency factors for participant rating of questionnaire item 2, addressing perceived synchrony between performers.

7.7.4 Experience Rating: Q2, Perception of Synchrony

For questionnaire item 2 (Figure 7.22), addressing perceived delay in the performance experience, ANOVA of participant rating results reports a significant effect of both Latency ($F = 10.24, p < 0.0001, \eta^2 = 0.0484$) and RT ($F = 5.56, p = 0.0042, \eta^2 = 0.0263$). No significant interaction effect of RT and Latency was identified ($F = 0.19, p = 0.9413$), indicating that effects of Latency and RT are independent. Pairwise comparisons with Bonferroni adjustment (Table 7.9) reports differences between *RT0.5, L13* and both of *RT0.5, L38* and *RT2.3, L38*. For the discrete factor of latency relevant Bonferroni–adjusted pairwise comparisons demonstrate significant differences between *L38* and other latency conditions (Figure 7.23, Table 7.10). Here it is identified that *13 ms (analogue)* and *18 ms (network)* latency conditions achieve superior rating relative to *38 ms (network)* latency conditions.

For the discrete factor of RT, pairwise comparisons with Bonferroni adjustment report significant differences between *RT2.3* and other RT conditions (Table 7.11, Figure 7.24). It is identified here that superior rating is given for RT of 0.5 s and 1.5 s relative to RT of 2.3 s.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	0.1069	1	1	0.0810	0.6211	0.3034	**<0.0001**
RT0.5, L18	-	-	1	1	1	1	1	1	0.0810
RT0.5, L38	-	-	-	0.3877	1	1	1	1	1
RT1.5, L13	-	-	-	-	1	0.3034	1	0.9703	**0.0036**
RT1.5, L18	-	-	-	-	-	1	1	1	**0.0340**
RT1.5, L38	-	-	-	-	-	-	1	1	1
RT2.3, L13	-	-	-	-	-	-	-	1	1
RT2.3, L18	-	-	-	-	-	-	-	-	1
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.9: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 2, addressing perception of synchrony. Significant differences are highlighted in bold.

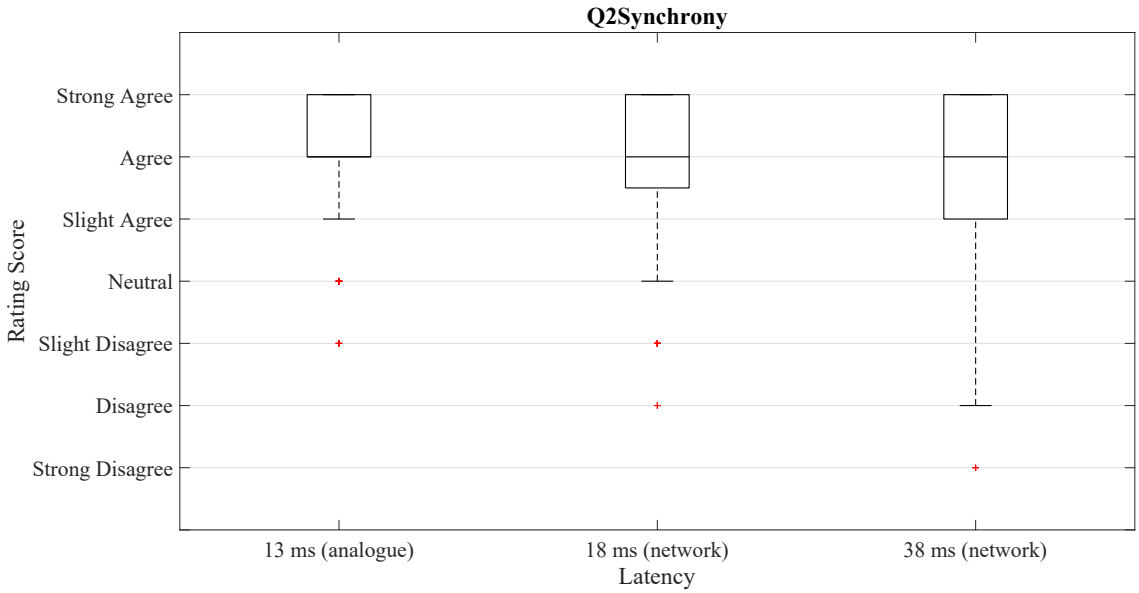


Figure 7.23: Box plot across the factor of latency for participant rating of questionnaire item 2, addressing perceived synchrony between performers.

	L13	L18	L38
L13	-	0.6649	<**0.0001**
L18	-	-	**0.0055**
L38	-	-	-

Table 7.10: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of latency for rating of questionnaire item 2, addressing perception of synchrony. Significant differences are highlighted in bold.

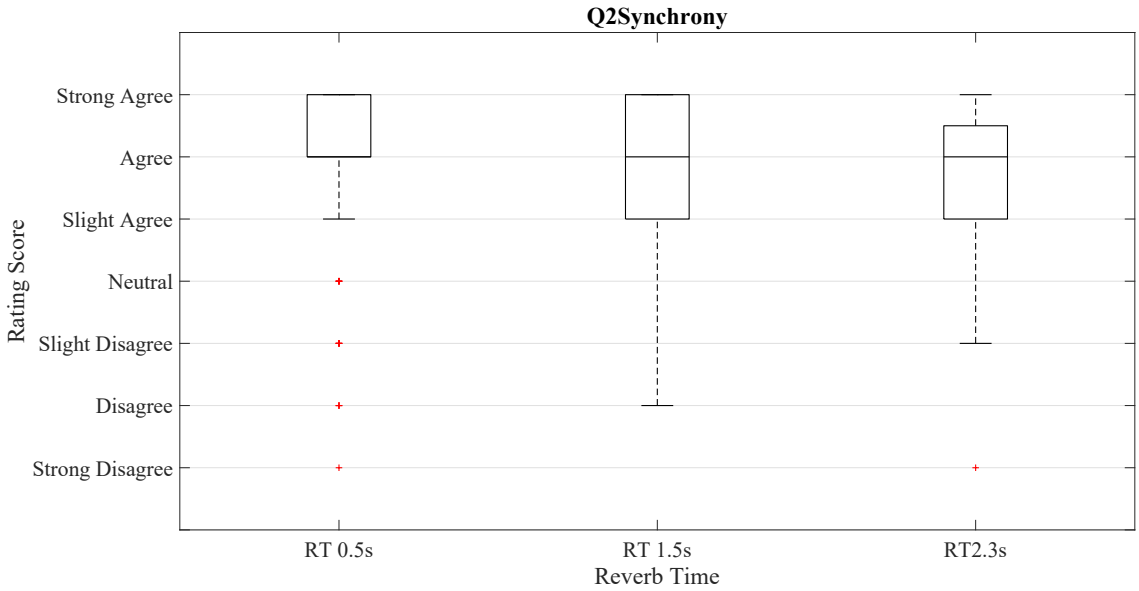


Figure 7.24: Box plot across the factor of RT for participant rating of questionnaire item 2, addressing perceived synchrony between performers.

	RT0.5	RT1.5	RT2.3
RT0.5	-	1	**0.0115**
RT1.5	-	-	**0.0180**
RT2.3	-	-	-

Table 7.11: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 2, addressing perception of synchrony.

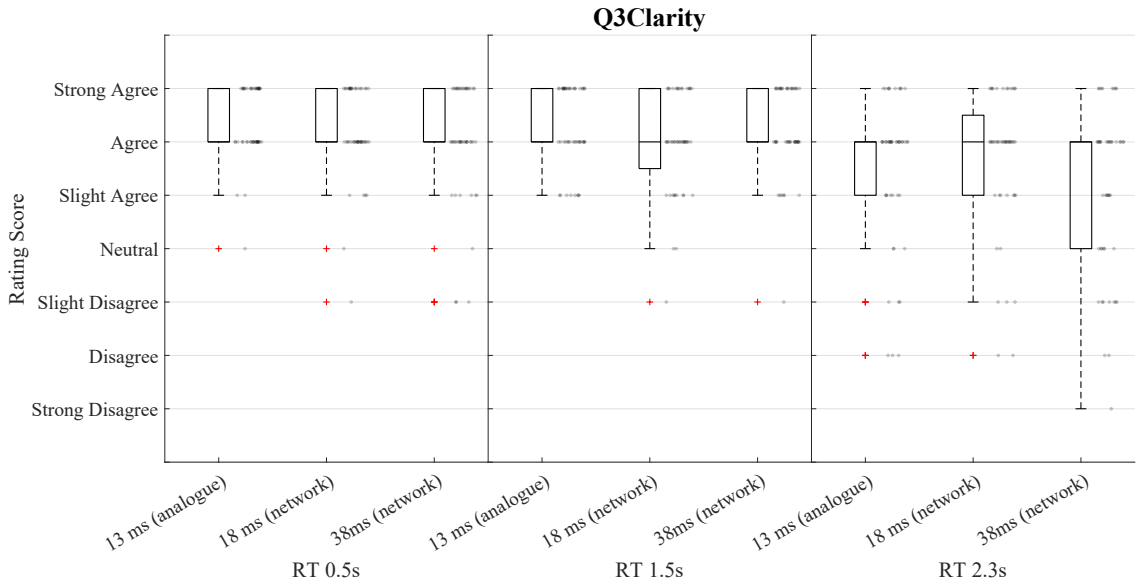


Figure 7.25: Box plot across RT and latency factors for participant rating of questionnaire item 3, addressing perceived clarity in monitoring of auralised performance.

7.7.5 Experience Rating: Q3, Perception of Clarity

With respect to questionnaire item 3 (Figure 7.25), addressing perceived clarity in monitoring of auralised performance, ANOVA of rating results shows a significant effect of RT ($F = 24.43, p < 0.0001, \eta^2 = 0.1155$). No significant effect is shown for latency ($F = 2.53, p = 0.0806, \eta^2 = 0.0120$), and no interaction effect of latency and RT is shown ($F = 1.91, p = 0.1079$). Pairwise comparison with Bonferroni confidence interval adjustment demonstrates a range of significant differences between groups for *RT2.3* and other groups for other RT conditions (Table 7.12). Bonferroni adjusted pairwise comparisons addressing the discrete factor of RT demonstrates significant differences between *RT2.3* and other RT conditions (Figure 7.26, Table 7.13). Here a higher rating of clarity is achieved for RT of 0.5 s and 1.5 s than for RT of 2.3 s.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	1	1	1	1	**0.0033**	0.0541	<**0.0001**
RT0.5, L18	-	-	1	1	1	1	0.0541	0.5388	<**0.0001**
RT0.5, L38	-	-	-	1	1	1	0.6938	1	<**0.0001**
RT1.5, L13	-	-	-	-	1	1	**0.0143**	0.1815	<**0.0001**
RT1.5, L18	-	-	-	-	-	1	1	1	**0.0015**
RT1.5, L38	-	-	-	-	-	-	**0.0392**	0.4152	<**0.0001**
RT2.3, L13	-	-	-	-	-	-	-	1	1
RT2.3, L18	-	-	-	-	-	-	-	-	0.2410
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.12: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 3, addressing perception of clarity in monitoring of auralised performance. Significant differences are highlighted in bold.

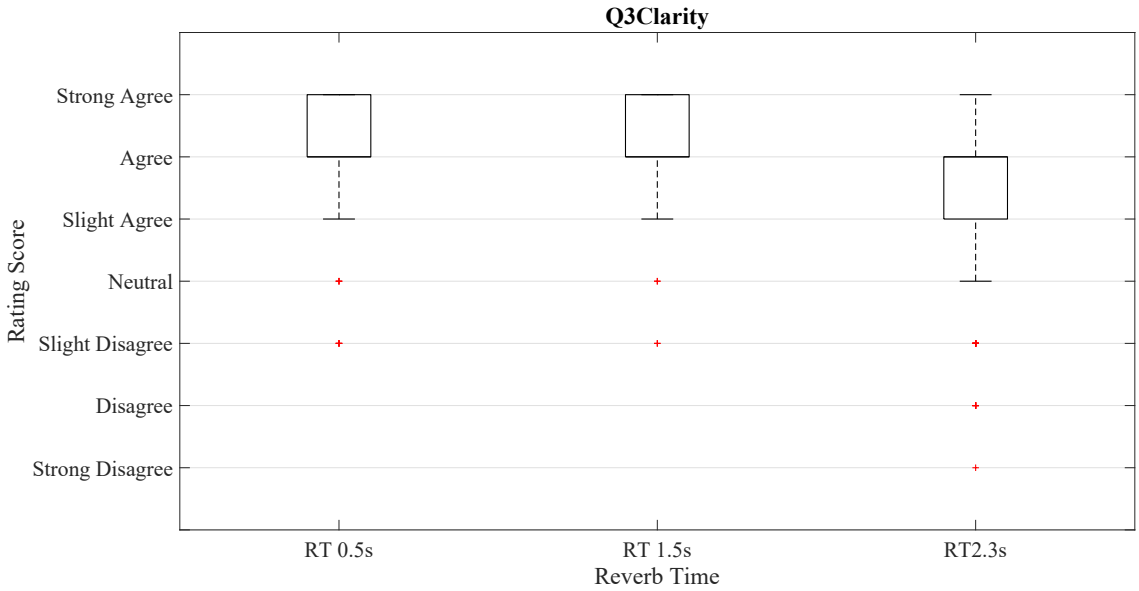


Figure 7.26: Box plot across the factor of RT for participant rating of questionnaire item 3, addressing perceived clarity in monitoring of auralised performance.

	RT0.5	RT1.5	RT2.3
RT0.5	-	1	**<0.0001**
RT1.5	-	-	**<0.0001**
RT2.3	-	-	-

Table 7.13: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 3, addressing perceived clarity in monitoring of auralised performance.

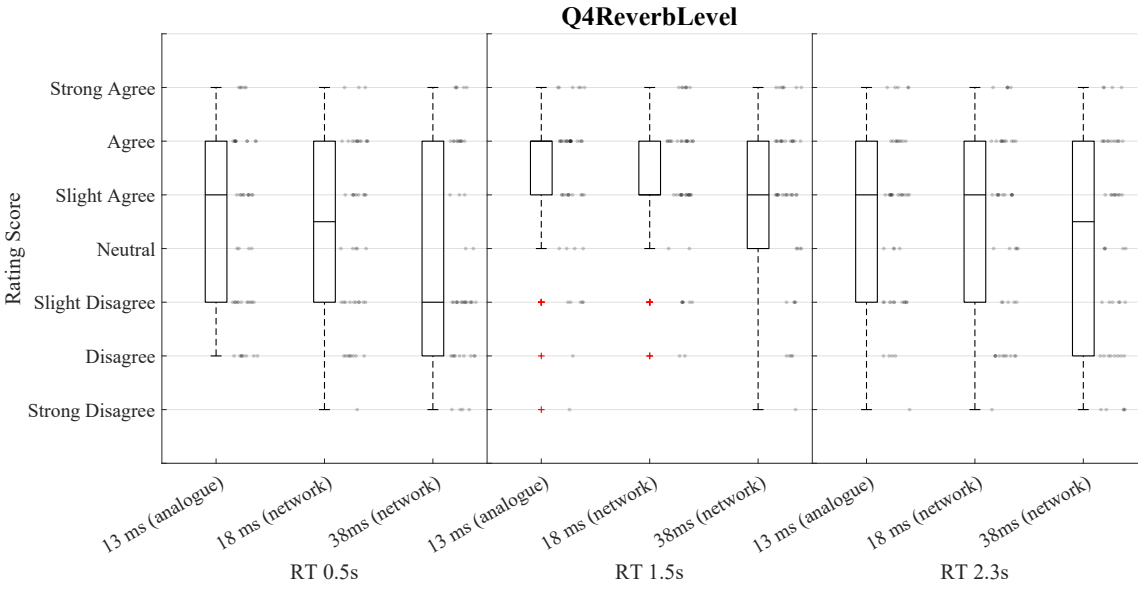


Figure 7.27: Box plot across RT and latency factors for participant rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance.

7.7.6 Experience Rating: Q4, Perception of Reverb Level

In the rating of questionnaire item 4, addressing perception of reverberation level (Figure 7.27), ANOVA results report a significant effect of RT ($F = 14.16, p < 0.0001, \eta^2 = 0.0669$). No significant effect of latency was detected ($F = 2.14, p = 0.1188, \eta^2 = 0.0101$), nor was any interaction effect between RT and latency factors ($F = 0.1, p = 0.9820$). Bonferroni–adjusted pairwise comparisons show significant differences between groups for *RT1.5* and groups for *RT0.5* and *RT2.3* (Table 7.14). Addressing the discrete factor of RT, pairwise comparisons demonstrate significant difference between *RT1.5* and other RT conditions (Figure 7.28, Table 7.15). In this case *1.5 s RT* is rated higher than *0.5 s RT* or *2.3 s RT*.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	1	0.4300	0.5101	1	1	1	1
RT0.5, L18	-	-	1	0.1741	0.2100	1	1	1	1
RT0.5, L38	-	-	-	**0.0021**	**0.0027**	**0.0278**	1	1	1
RT1.5, L13	-	-	-	-	1	1	1	0.6031	**0.0428**
RT1.5, L18	-	-	-	-	-	1	1	0.7109	0.0528
RT1.5, L38	-	-	-	-	-	-	1	1	0.3612
RT2.3, L13	-	-	-	-	-	-	-	1	1
RT2.3, L18	-	-	-	-	-	-	-	-	1
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.14: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance. Significant differences are highlighted in bold.

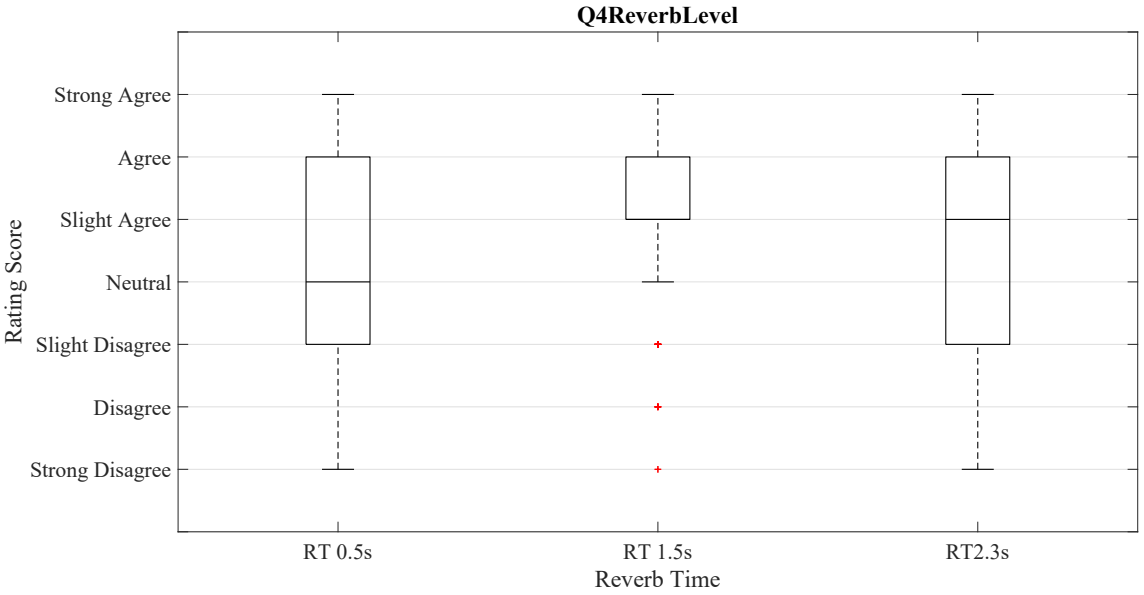


Figure 7.28: Box plot across the factor of RT for participant rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance.

	RT0.5	RT1.5	RT2.3
RT0.5	-	$<$**0.0001**	1
RT1.5	-	-	$<$**0.0001**
RT2.3	-	-	-

Table 7.15: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 4, addressing perception of reverberation level in monitoring of auralised performance.

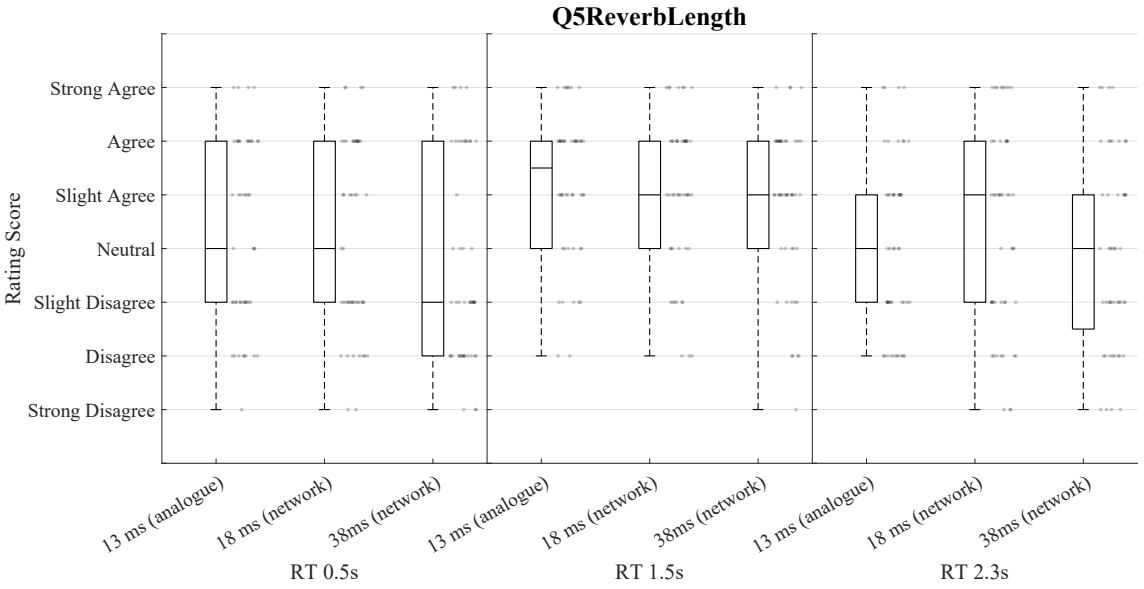


Figure 7.29: Box plot across RT and latency factors for participant rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance.

7.7.7 Experience Rating: Q5, Perception of Reverberation Length

ANOVA of the rating results for questionnaire item 5 (Figure 7.29), addressing perception of reverberation length in monitoring of auralised performance, demonstrates a significant effect of RT ($F = 15.18, p < 0.0001, \eta^2 = 0.0718$). No effect of latency was identified ($F = 2.63, p = 0.0732, \eta^2 = 0.0114$), and no interaction effect of RT and latency was identified ($F = 0.31, p = 0.8697$). Pairwise comparisons with Bonferroni confidence interval adjustment, report significant differences between groups for *RT1.5* and groups for other RT conditions (Table 7.16). Bonferroni–adjusted pairwise comparisons for the discrete factor of RT demonstrate significant differences between *RT1.5* and alternative RT conditions, where *1.5 s RT* is rated as superior to *0.5 s RT* and *2.3 s RT* (Figure 7.30, Table 7.17).

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	1	0.2303	0.8913	1	1	1	1
RT0.5, L18	-	-	1	0.1315	0.5496	1	1	1	1
RT0.5, L38	-	-	-	**0.0004**	**0.0032**	1	1	**0.0318**	1
RT1.5, L13	-	-	-	-	1	1	**0.0484**	0.4648	**0.0052**
RT1.5, L18	-	-	-	-	-	1	0.2303	1	**0.0318**
RT1.5, L38	-	-	-	-	-	-	1	1	1
RT2.3, L13	-	-	-	-	-	-	-	1	1
RT2.3, L18	-	-	-	-	-	-	-	-	0.2303
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.16: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance. Significant differences are highlighted in bold.

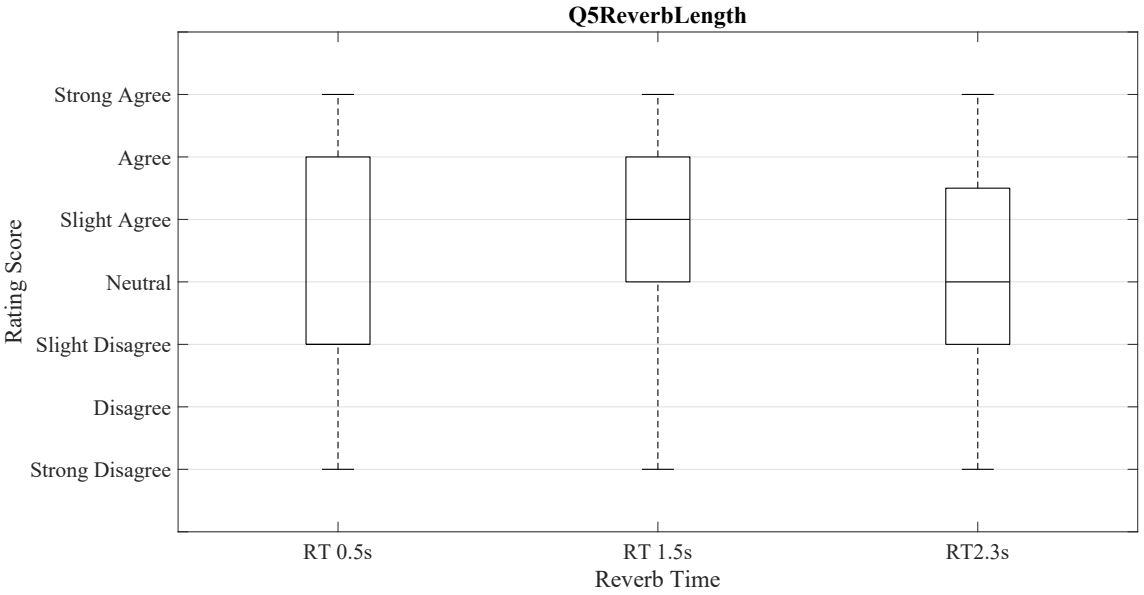


Figure 7.30: Box plot across the factor of RT for participant rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance.

	RT0.5	RT1.5	RT2.3
RT0.5	-	$<$**0.0001**	1
RT1.5	-	-	$<$**0.0001**
RT2.3	-	-	-

Table 7.17: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for rating of questionnaire item 5, addressing perception of reverberation length in monitoring of auralised performance.

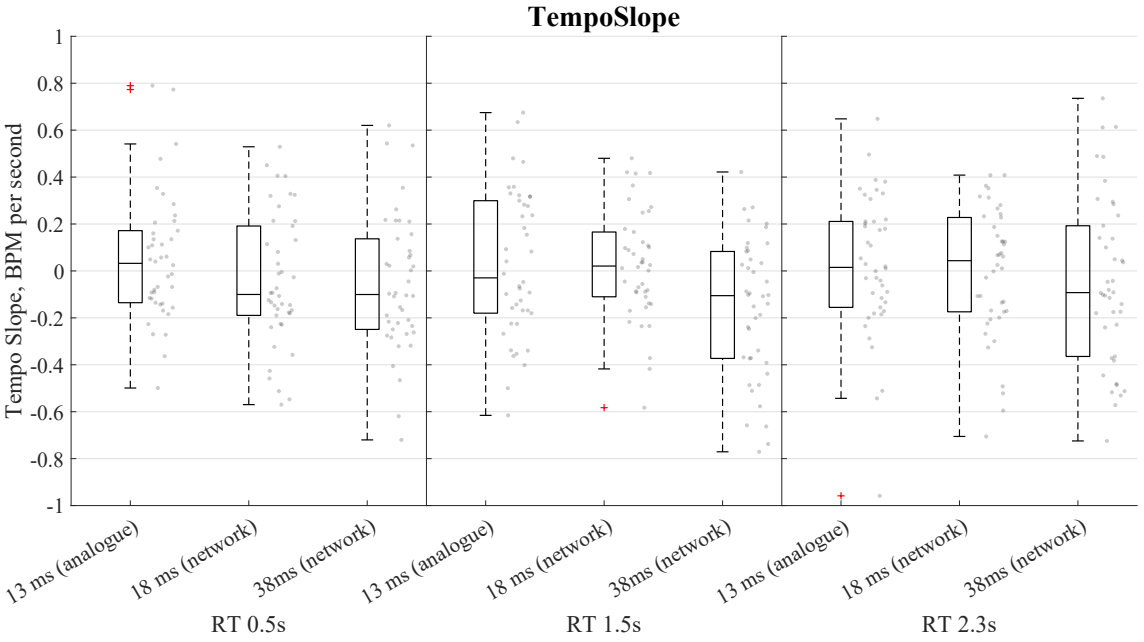


Figure 7.31: Box plot across RT and latency factors for Tempo Slope.

7.7.8 Within–Parts Synchrony: Tempo Slope

For measurements of Tempo Slope ANOVA conducted on results demonstrates a significant effect of latency ($F = 5.62, p = 0.0039, \eta^2 = 0.0305$) (Figure 7.31, Table 7.18). No effect of RT ($F = 0.24, p = 0.7836, \eta^2 = 0.0013$) or interaction effect of RT and latency ($F = 1.14, p = 0.3368$) was detected. Pairwise comparison with Bonferroni adjustment identifies a single significant difference between group *RT0.5, L13* and *RT1.5, L38*. Addressing the discrete factor of latency, pairwise comparison of latency conditions demonstrates significant difference between groups for *L13* and *L38*. Here it is demonstrated that *13 ms (analogue)* latency conditions allow for more stable tempo slope, whereas tempo slope for *38 ms (network)* is less stable and trends towards tempo deceleration (Figure 7.32, Table 7.19).

Given borderline significance and violation of ANOVA assumptions here, follow–up non–parametric pairwise comparison (Wilcoxon rank sum, Table 7.20) was also conducted. Here a significant difference is reported between *L38* and *L13* ($z = 3.1694, p = 0.0015$) and *RT18* ($z = 2.2582, p = 0.0239$). This would indicate that the lack of a significant difference detected between *RT1.5* and *RT2.3* in ANOVA pairwise comparisons is a Type II error. This would suggest that *13 ms (analogue)* and *18 ms (network)* latency allows for more stable tempo than *38 ms (network)* latency conditions.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	1	1	1	**0.0405**	1	1	1
RT0.5, L18	-	-	1	1	1	1	1	1	1
RT0.5, L38	-	-	-	1	1	1	1	1	1
RT1.5, L13	-	-	-	-	1	0.1309	1	1	1
RT1.5, L18	-	-	-	-	-	0.1602	1	1	1
RT1.5, L38	-	-	-	-	-	-	0.1699	0.6530	1
RT2.3, L13	-	-	-	-	-	-	-	1	1
RT2.3, L18	-	-	-	-	-	-	-	-	1
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.18: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for Tempo Slope.

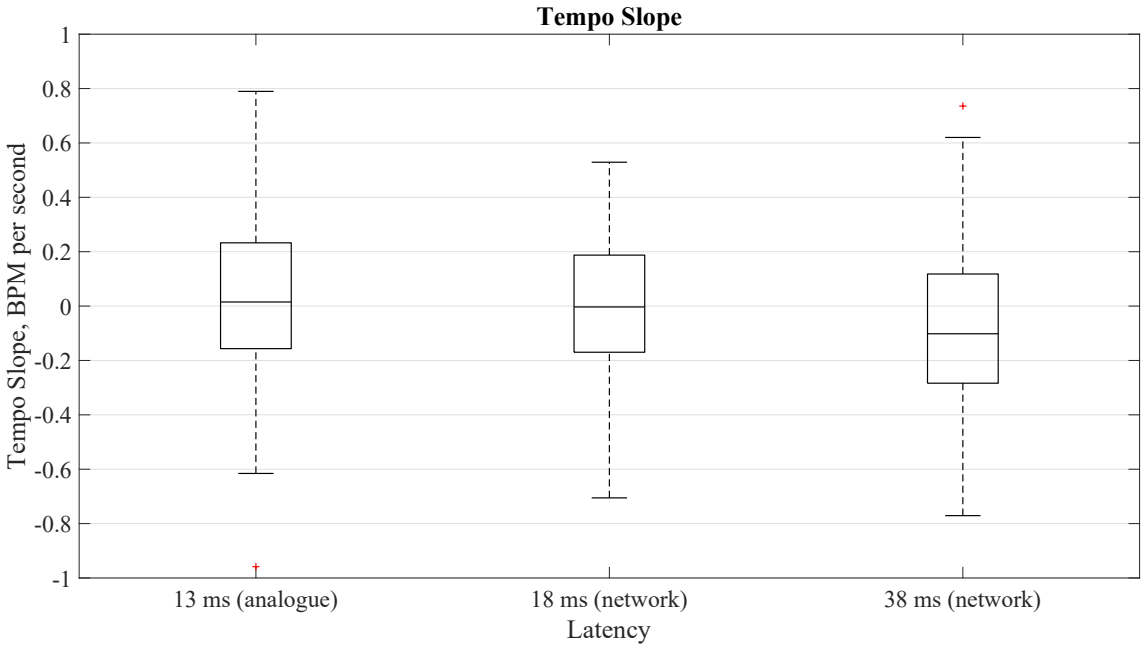


Figure 7.32: Box plot across the factor of latency for Tempo Slope.

	L13	L18	L38
L13	-	0.7328	**0.0031**
L18	-	-	0.0981
L38	-	-	-

Table 7.19: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of latency for Tempo Slope. Significant differences are highlighted in bold.

	L13	L18	L38
L13	-	0.3232	**0.0015**
L18	-	-	**0.0239**
L38	-	-	-

Table 7.20: Wilcoxon rank sum significance values (p) from pairwise comparison across the factor of latency for Tempo Slope. Significant differences are highlighted in bold.

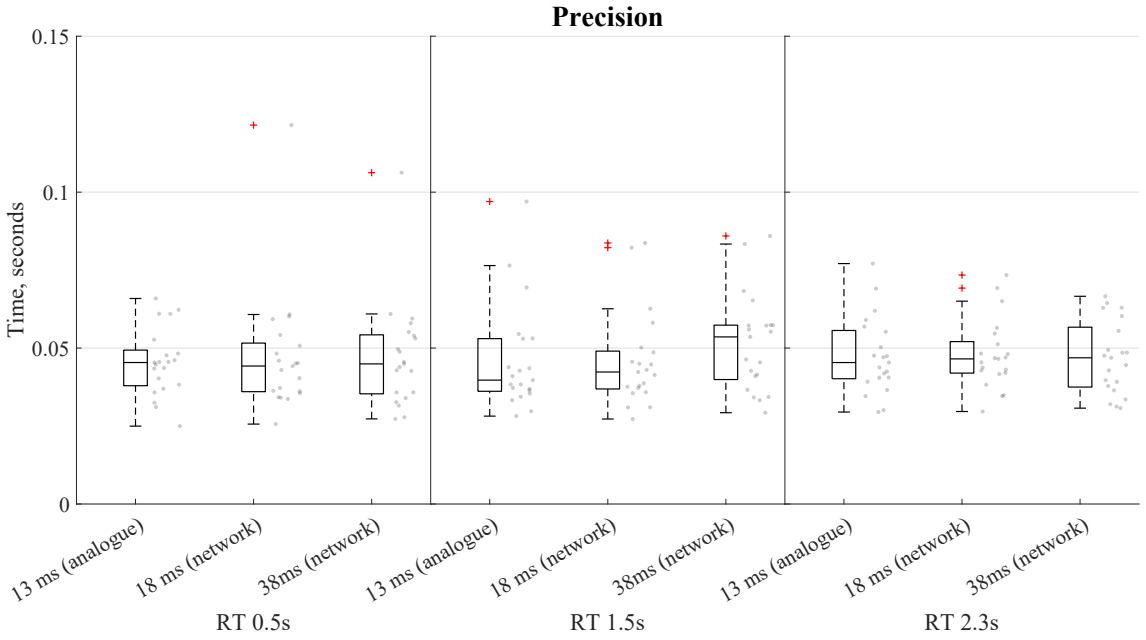


Figure 7.33: Box plot across RT and latency factors for Precision.

7.7.9 Between–Parts Synchrony: Precision

For measurement of precision (the mean of absolute OTDs) between parts, ANOVA results (Figure 7.33) demonstrate no significant effect of RT ($F = 0.09, p = 0.9129, \eta^2 = 0.0010$) or latency ($F = 0.39, p = 0.6793, \eta^2 = 0.0043$). No interaction effect for RT and latency was detected either ($F = 0.4, p = 0.8052$). As such no follow–up pairwise comparisons are presented in text. Notably, several cases of non–normality are identified in groups here (Table 7.2).

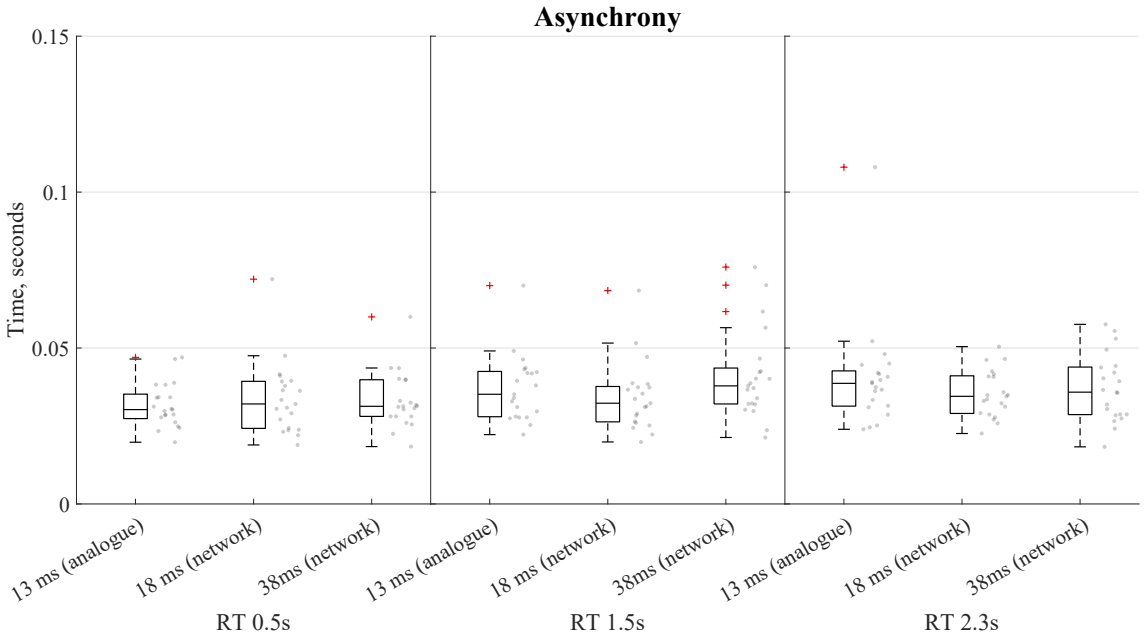


Figure 7.34: Box plot across RT and latency factors for Asynchrony.

7.7.10 Between–Parts Synchrony: Asynchrony

ANOVA results for the measurement of Asynchrony (the mean of standard deviations of absolute OTDs across an ensemble, and simply the standard deviation of absolute OTDs in a duet context as considered in this study) demonstrates a significant effect of RT ($F = 3.44, p = 0.0342, \eta^2 = 0.0382$). No significant effect of latency ($F = 0.87, p = 0.4218, \eta^2 = 0.0096$) or interaction effect of latency and RT ($F = 1.12, p = 0.3478$) was detected. Bonferroni–adjusted pairwise comparisons, however, indicate no significant differences between discrete groups (Table 7.21, Figure 7.34). Addressing the discrete factor of RT, pairwise comparison with Bonferroni adjustment indicates no significant difference across RT conditions (Figure 7.35, Table 7.22). This indicates that reports of significant effect of RT from ANOVA results are a Type I error and cannot be accepted.

Considering ANOVA assumption violations and the potential for Type II error non–parametric follow–up was conducted using Wilcoxon rank sum tests across RT groups (Table 7.23). Here results indicate that asynchrony for *RT0.5* varies significantly from *RT1.5* ($z = -2.2054, p = 0.0274$) and *RT2.3* ($z = -2.708, p = 0.0068$) and it is suggested that lower *asynchrony* is achieved for *0.5 s RT* conditions. It is therefore acknowledged that the lack of significance in pairwise comparison from parametric testing may be a Type II error.

	RT0.5, L13	RT0.5, L18	RT0.5, L38	RT1.5, L13	RT1.5, L18	RT1.5, L38	RT2.3, L13	RT2.3, L18	RT2.3, L38
RT0.5, L13	-	1	1	1	1	0.3168	0.5224	1	1
RT0.5, L18	-	-	1	1	1	1	1	0.9525	1
RT0.5, L38	-	-	-	1	1	1	1	1	1
RT1.5, L13	-	-	-	-	1	1	1	1	1
RT1.5, L18	-	-	-	-	-	1	1	1	1
RT1.5, L38	-	-	-	-	-	-	1	1	1
RT2.3, L13	-	-	-	-	-	-	-	1	1
RT2.3, L18	-	-	-	-	-	-	-	-	1
RT2.3, L38	-	-	-	-	-	-	-	-	-

Table 7.21: Bonferroni–adjusted significance values (p) from pairwise comparison across latency and RT factors for Asynchrony. Significant differences are highlighted in bold.

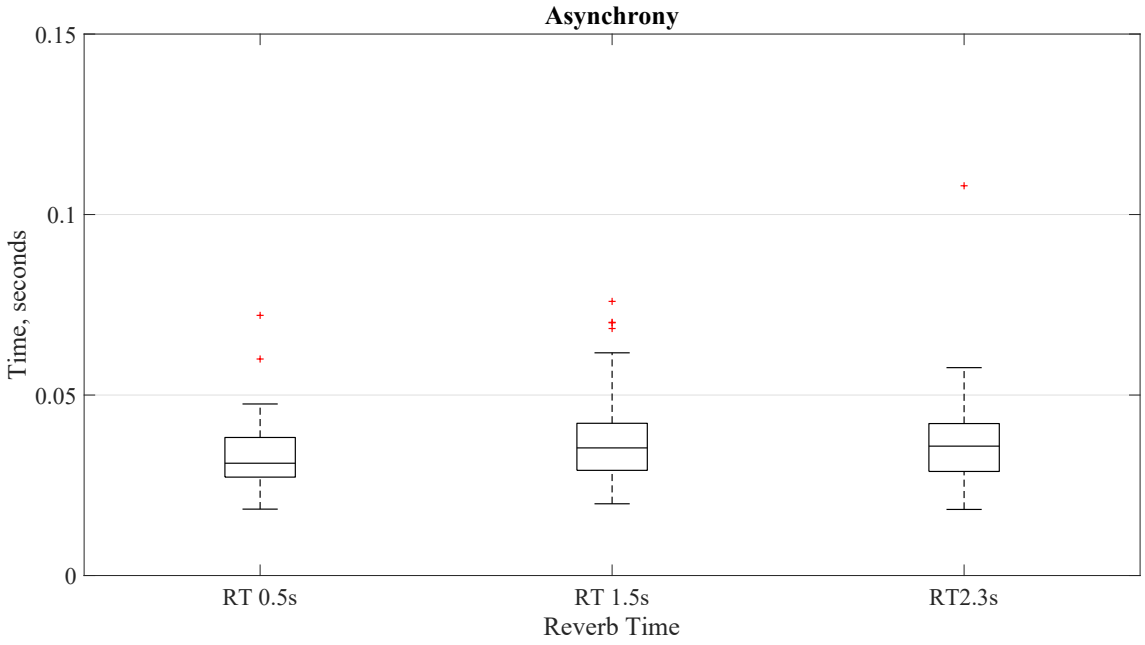


Figure 7.35: Box plot across the factor of RT for Asynchrony.

	RT0.5	RT1.5	RT2.3
RT0.5	-	0.0782	0.0683
RT1.5	-	-	1
RT2.3	-	-	-

Table 7.22: Bonferroni–adjusted significance values (p) from pairwise comparison across the factor of RT for Asynchrony.

	RT0.5	RT1.5	RT2.3
RT0.5	-	**0.0274**	**0.0068**
RT1.5	-	-	0.7585
RT2.3	-	-	-

Table 7.23: Wilcoxon rank sum significance values (p) from pairwise comparison across the factor of RT for Asynchrony.

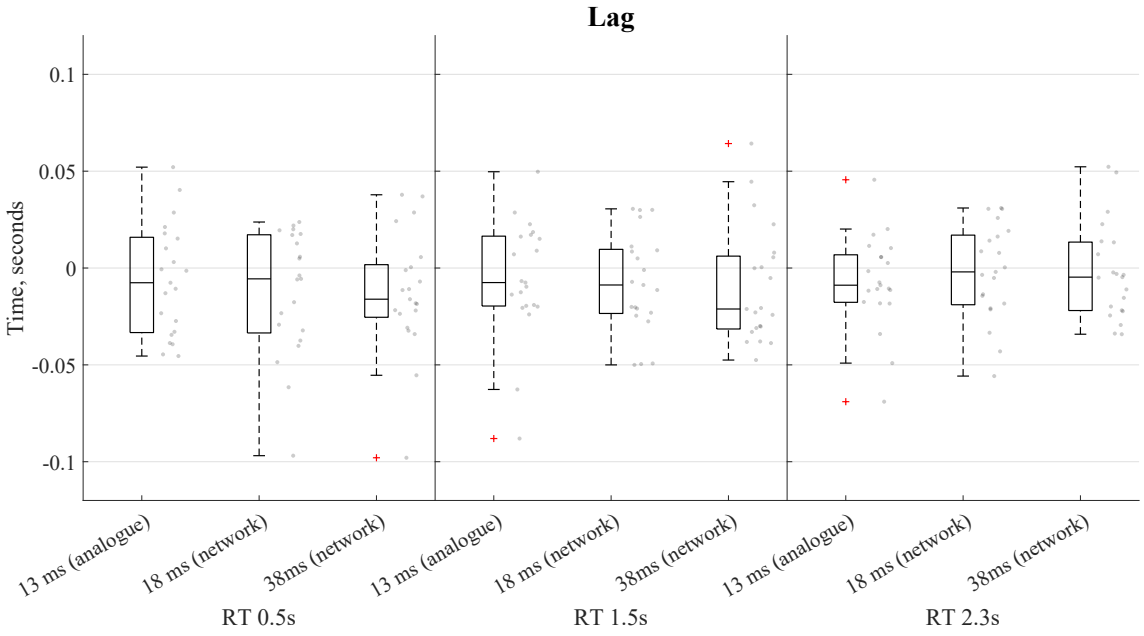


Figure 7.36: Box plot across RT and latency factors for Lag.

7.7.11 Between Parts Synchrony: Lag

ANOVA results for Lag (the median signed OTD, here expressing the lag of a *'follower'* part to a *'leader'* part) demonstrate no significant difference across groups associated with RT ($F = 0.88, p = 0.4186, \eta^2 = 0.0097$), latency ($F = 0.06, p = 0.9381, \eta^2 = 0.0007$), or interaction effect of RT and latency ($F = 0.3, p = 0.8753$) (Figure 7.36). Follow–up pairwise comparisons are therefore note presented in text.

7.8 Discussion

The following subsections present discussion of the results of the INMP system evaluation. This includes discussion of the effect of latency, and discussion of the effect of acoustics.

7.8.1 Effect of Latency

Results from rating items 1 and 2 (addressing perceived delay and synchrony respectively) demonstrate an impairment in rating of *38 ms (network)* conditions relative to lower latency levels, and no variation in rating between *13 ms (analogue)* and *18 ms (network)* latency levels. Here findings are consistent with expectations derived from prior work, where it is empirically understood that NMP experience will deteriorate in excess of 20–30 ms [6]. Focused prior studies have specifically demonstrated that NMP QoE rating of perceived delay and synchrony begins to deteriorate in excess of 40 ms [114], [118].

Less stable *tempo slope* is demonstrated for *38 ms (network)* latency relative to baseline latency conditions. It is also strongly suggested that tempo slope deteriorates at *38 ms (network)* latency relative to *18 ms (network)* latency. No tempo slope variation was detected between *18 ms (network)* and baseline latency conditions. These findings conform to expectations from prior work. Here it is empirically understood that latency below 20 ms will support similar synchrony to baseline conditions of approximately 10 ms, and that latency in excess of 30 ms may result in synchrony impairment, particularly detectable in tempo slope [6], [57], [58], [211].

Though impairment of *Tempo Slope* is demonstrated at *38 ms (network)* latency levels, it can be highlighted that no differences associated with latency were detected for *Precision*, *Asynchrony* or *Lag* measurements. This would indicate no relative impairment of synchrony across latency levels, and no indication of significant adoption of leader–follower coping strategies [129] in performance. Indeed it can also be noted that, while *Tempo Slope* is significantly less stable for *38 ms (network)* latency conditions, measured values do not exhibit tempo deceleration or instability associated with high–latency conditions in tapping–based studies (deceleration of mean value in excess of -0.2 BPM/s [57], [58], [211], or in excess of approximately $\pm 2\%$ [19]). This suggests that synchrony expectations of vocal performance using INMP systems are different from those described in tapping studies, and as such expectations from tapping studies may not be reasonable to assume in practical INMP vocal performance scenarios.

With respect to the objective of providing practice–oriented evaluation, *18 ms (network)* and *38 ms (network)* latency conditions are considered representative of 'best–case' and 'high–end' latency measurements respectively from practical WAN deployment of a prior iteration of the INMP system design [8]. Results here indicate that 'best–case' latency in practical WAN deployment over national ranges will result in no noticeable difference from analogue connection between endpoints (for example in studio monitoring systems), reflected both through positive experience rating of delay and synchrony, and objective performance synchrony. From synchrony, 'high–end' latency of national WAN deployment of the INMP system is still capable of supporting synchronous performance, albeit providing an impaired experience.

Though rating of perceived delay indicates some negative experience here, the ability to perform synchronously is still generally rated positively. Indeed prior work does include detail of high–latency[41] NMP deployments which support synchronous performance of rhythmic musical styles despite impairment of experience and interactivity between performers (for example [35]).

Considering testing of the hypothesis that a singer's experience and synchrony will vary for different latency levels it is evident that this is indeed the case. This is expected from empirical NMP study [6] and prior related practical WAN deployment of a previous iteration of the INMP system design [7], [8].

7.8.2 Effect of Acoustics

In the rating of perceived *clarity* in performance it is evident that experience of performance conditions for *2.3 s RT* conditions, though generally positive, are impaired relative to lower RT conditions. Inspecting the C50 clarity parameter of the three different acoustic conditions, clarity for *2.3 s RT* conditions is noticeable lower than other acoustic conditions[42] (Figure 7.37). Prior NMP evaluation has suggested that lack of clarity in monitoring with reverb is considered detrimental to the networked performance experience for drummers [119]. Indeed specific to prior development on which the INMP vocal performance system builds, evaluation of simulated performance experiences for single users report impaired rating of clarity in perceived performance conditions for RT of 2.2 s and 1.8 s relative to RT for 1.3 s [220]. As such impairment in the rating of clarity of perceived performance conditions for *2.3 s RT* in this study is coherent with the findings of prior work.

Subjective rating of the RT level and length clearly demonstrates that in both cases *1.5 s RT* conditions is preferable to both *0.5 s RT* and *2.3 s RT* conditions, and is likely to support a consistently positive performance experience. Though *1.5 s RT* conditions are significantly superior to alternatives, both negative and positive ratings were given for *0.5 s RT* and *2.3 s RT* conditions. As such these conditions should not be expected categorically to impair the INMP experience. Relevant prior work evaluating singer experiences of auralised acoustics report an impairment in rating of performance conditions for RT of 1.3 s relative to RT of 1.8 s and 2.2 s [220]. In other work, however, excessive RT values (here unspecified) are correlated with impairment of singer experience [219]. It can also be recognised that *1.5 s RT* conditions are most coherent with the RT of real concert hall spaces [213]. As such findings of this study are coherent with the findings of prior work, where excessive lack or surplus of reverberation is considered detrimental to perceived performance conditions.

[41] I.e., in excess of 30 ms [6]

[42] A JND of approximately 1.5 dB can be expected from prior work [234]

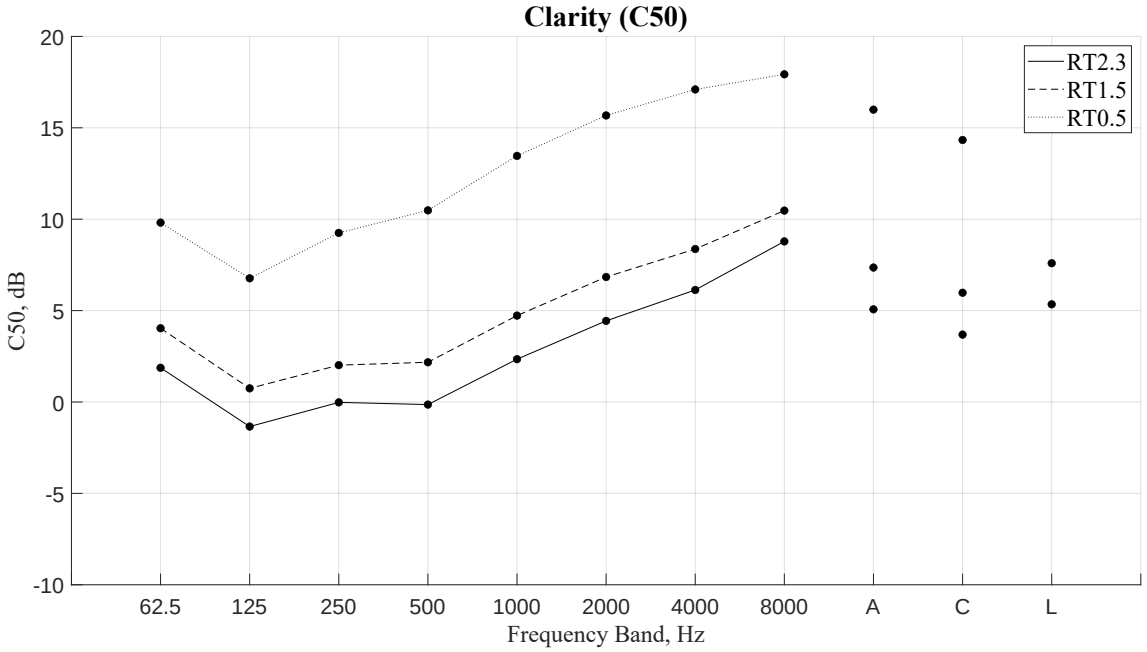


Figure 7.37: Clarity (C50) for the original SIR (RT2.3), and SIR with decay envelope adjustment applied to approximate RT of 1.5 s (RT1.5) and 0.5 s (RT0.5). A, C and L refer to A–weighted, C–weighted and Linear frequency weighting respectively.

Findings from the rating of perceived performance conditions are corroborated in the rating of perceived synchrony. Here *2.3 s RT* was demonstrated as supporting performance where the ability to perform synchronously, while still rated positively, was impaired relative to *0.5 s RT* and *1.5 s RT* conditions. Notably in synchrony evaluation it is strongly suggested in *Asynchrony* measurement that superior synchrony is achieved for *0.5 s RT* than other conditions. This supports ratings which describe deterioration in perceived performance conditions for *2.3 s RT* conditions. The generally superior *Asynchrony* measured for *0.5 s RT* here may reflect improved precision in performance relating to improved clarity. Indeed study of acoustic preferences of trumpet performers considers that shorter reverb times, while not preferred by musicians for concert performance and practice, were considered superior for technical performance practice [214]. No effect of RT is, however, reported in synchrony metrics other than *Asynchrony* in this study. Thus variation of synchrony with RT using the INMP system is only suggested, rather than robustly demonstrated by results.

Notably rating results regarding perception of delay strongly suggest that delay is considered more noticeable for *2.3 s RT* than other conditions. Indeed this would be coherent with the general expression of performance conditions deterioration for this RT level in other rating and synchrony analysis. It is robustly demonstrated here that no improvement for *1.5 s RT* and *2.3 s RT* relative to *0.5 s RT* is present. This provides contrary evidence to suggestions from prior work, which propose that auralisation of reverberant performance spaces in INMP contexts will improve the experience of latency [8], [57], [203], demonstrating that this is not the case (at least in the INMP vocal performance context presented).

With respect to the primary hypothesis, testing for variations in INMP experience associated with RT, results robustly evidence that the perceptual experience of singers does indeed vary, including both perceived performance conditions, and perception of the NMP QoE factor of perceived synchrony. Considering INMP practice it can be surmised that the INMP experience for singers can be optimised through acoustic control, where acoustics may be characterised by RT. Where auralisation of a measured real or simulated virtual space is considered in INMP [16], [19], [20] this is particularly relevant. Where the native environment RT is excessive (here with 'excessive' being *2.3 s RT*), or *clarity* is poor, the decay envelope may be attenuated to improve perceived performance conditions. Where acoustics are particularly dry engineers may consider auralisation of higher RT acoustics which better support perceived performance conditions for singers. Parametric control of auralised acoustics is recommended to be included in INMP system deployment.

7.9 Hypothesis Revisited

At the beginning of this chapter the primary objective was presented:

> *To evaluate the influence of RT in duet vocal performance using a practical INMP system design.*

This objective was achieved through synchrony and QoE evaluation of vocal performance using an INMP system derived from a system which has been deployed over WAN connections in prior work.

The sub–objective was also presented:

> *To perform evaluation of the influence of RT in INMP duet vocal performance across different practical latency levels.*

This was achieved by conducting evaluation for latency conditions which were measured in WAN deployment of the INMP system in prior work. Additionally baseline latency conditions were evaluated.

To address the primary objective, the primary hypothesis was presented:

> *INMP QoE and/or synchrony for duet vocal performance will vary for different RT levels.*

As discussed in subsection 7.8.2, results of evaluation demonstrate differences across RT conditions in QoE evaluation, and a further effect of RT on synchrony is suggested. As such, the primary hypothesis is accepted in this study.

To address the sub–objective, the two secondary hypotheses were presented:

> *INMP QoE and/or synchrony for duet vocal performance will vary for different latency levels.*

> *Interaction between RT and latency levels will affect variation in INMP QoE and/or synchrony for duet vocal performance.*

As discussed in subsection 7.8.1, no difference in QoE or synchrony was found between baseline analogue latency levels, and latency levels which approximate the best–case levels found in WAN deployment of the INMP system. An expected impairment of both QoE and synchrony was found when latency increased in excess of 30 ms. No interaction effect of RT and latency was found. As such, the first of the secondary hypotheses (QoE and synchrony will vary across latency) is accepted. The second of the secondary hypotheses, however, is rejected.

7.10 Summary

The INMP system evaluation presented here provides a testing of the potential for INMP experience and synchrony to vary with RT. As a practice–oriented evaluation, this is achieved using a practical INMP group singing design, and evaluation is conducted across latency levels which are representative of practical WAN deployment. Results contribute novel evidence of INMP experience variation for different RT conditions, including perceived performance conditions and synchrony. Here acoustic parameter of both RT and clarity are identified as relevant to the INMP experience, where high RT and low clarity are associated with impairment of INMP experience.

It is additionally evidenced that RT does not improve experience of delay, as proposed in prior NMP literature, and in fact it is strongly suggested that the inverse is true, where high RT may influence impaired synchrony and perception and delay. This contribution informs INMP practice, where it is demonstrated that acoustic control utility in immersive audio monitoring for singers can allow for experience optimisation.

Chapter 8

Conclusion

And that's the end of that chapter

Detective Homer Simpson, The Simpsons S10E13 (1999)

8.1 Preface

This chapter provides the conclusion to this thesis. A summary is given of the work presented within, highlighting contributions to the field of INMP. The core hypothesis of this thesis is then revisited and a statement of acceptance is made with reference to the key findings of the presented work. Following this a summary of key areas for further work is provided. The implications of key findings for the modern media and events audio industry are then briefly surmised, and closing remarks are made.

8.2 Summary of Work and Contributions

The following subsections each provide a summary of the contents and contributions of each chapter of this thesis.

8.2.1 Chapter 1: Introduction

The opening chapter provides an introduction to the case of low–latency INMP as a practical remote performance solution. Immersive audio is presented as an enhancement of NMP systems, and as a design requirement when extending NMP with immersive media systems. The state–of–the–art in INMP is briefly surmised and the need for practice–oriented system documentation and evaluation is established. This leads to presentation of the objectives and core hypothesis of this thesis, which proposes that INMP systems are a practical remote performance workflow.

8.2.2 Chapter 2: Review of Network Music Performance

Here a review of NMP technology, practice and evaluation is contributed. The challenge of latency and need for low–latency solutions in practical remote performance workflows is outlined. Here detail is provided on a standard approach to low–latency audio streaming in NMP, popular and practical software solutions, evaluation approaches, and challenges identified in prior work.

8.2.3 Chapter 3: Review of Immersive Audio for Network Music Performance

In this chapter a review of immersive audio for NMP is contributed. The fundamentals of 3–D sound and simulation approaches for INMP are detailed, focusing on convolution–based Ambisonic and binaural stereo approaches. A summary of practical INMP system implementation and practice in prior work is given, and the largely theoretical state of low–latency INMP design and evaluation for practical applications is described.

8.2.4 Chapter 4: Evaluation of an INMP System for Remote Music Education

Chapter 4 presents a study which tests the viability of an INMP system designed for the practical application of remote music education. Three professional musicians participated in QoE evaluation of the performance experience, evaluation of performer engagement with visual display and interface, and evaluation of different auralisation approaches (namely auralisation of acoustics which are either congruent or divergent from the local physical environment at each endpoint). Results contribute a novel demonstration of the viability of the remote performance workflow under conditions which are representative of practical WAN deployment. Clear evidence of performer engagement with video display and interface is also contributed. Investigation of the experience of immersive audio monitoring contributes a novel demonstration of the potential for experience of immersive audio to vary with monitoring context in practical scenarios, and highlights the potential for high RT and low clarity to adversely affect the experience of INMP. Peripheral contributions here include novel documentation of the practical INMP system design.

8.2.5 Chapter 5: Evaluation of an XR INMP system for remote virtual production

Chapter 5 presents a study testing the viability of an XR INMP system designed for the practical application of remote virtual production. The four members of the professional rock band *Wolph* participated in evaluation, testing the ability of the system to support synchronous performance and high–quality QoE. Additionally, performer engagement with novel XR display and interface is tested, and participant experience of immersive audio rendering is investigated.

Results demonstrate that synchronous performance and high quality experience may be delivered using the novel XR INMP system design. It is demonstrated that immersive audio monitoring here can improve the experience with respect to spatial clarity and localisation factors. However, it is also demonstrated that experience of immersive audio varies with monitoring context in–practice, and may not always be experienced positively. Evidence of participant engagement with novel XR display and interface is contributed, however, it is also demonstrated as varying with monitoring context. Here results novelly identify factors of XR display and interface considered important to the INMP experience, including avatar representation and pass–through display perspective and quality. Peripheral contributions include novel documentation of the practical INMP system design.

8.2.6 Chapter 6: Evaluation of an INMP System for Remote Concert Performance

In chapter 6 a two–part study is presented. The first part tests the viability of an INMP system for remote concert performance through performer and audience QoE evaluation of a practical concert event. Here the eight members of *Ex Corde Vocal Ensemble* provide a comparative evaluation of live and networked performance, whereas audience members provide evaluation only for the networked concert event. Results contribute a demonstration of the ability of the INMP system to deliver a positive concert experience to both performers and audiences, and demonstrate that comparable performance quality to live conditions may be achieved. Further investigation novelly highlights the potential for experience improvement with immersive audio monitoring with respect to spatial clarity factors, and with respect to supporting tuning of performance delivery. This is, however, suggested as dependent on monitoring context, here with respect to virtual and physical positioning of singers. Investigation of the audience experience highlights several areas for improvement in loudspeaker reproduction of immersive audio, including position–independent listening, and naturalness of loudspeaker reproduction of voice.

In the second part of this study a listening test is conducted. Here participants provide an evaluation of performance quality for audio recordings of performances using the INMP concert system, and audio recordings of live reference performances. Results here contribute a novel demonstration that performance quality using the practical INMP concert system is comparable to live performance quality. Peripheral contributions of this chapter include a dataset of media recordings and rating data for the INMP concert production.

8.2.7 Chapter 7: Evaluation of the effect of Reverb Time using an INMP system for remote singing practice

The final chapter presents a practice–oriented evaluation of the impact of RT on vocal performance using an INMP system designed for the application of remote group singing. Here, pairs of singers use the INMP system to engage in duet vocal performance under different auralisation conditions (characterised by RT) and different latency conditions (including baseline conditions, and network conditions which approximate WAN deployment).

Evaluation is provided through synchrony analysis of recorded performance and participant QoE rating. Results contribute a novel demonstration of the impact of RT on perceived performance conditions, where excessive RT is identified as impairing rating of clarity, reverberation level and length, and NMP factors of perceived performance quality and perceived latency. Excessively high RT is also strongly suggested as impairing objective performance synchrony. Where RT is excessively low, impairment in rating of reverberation level and length is identified. Peripheral contributions here include original MATLAB implementations of SIR decay envelope adjustment, an original onset detection workflow and MATLAB wrapper function implementation, and a dataset of over 12,000 onsets.

8.3 Re–statement of Hypothesis

At the outset of this thesis the central hypothesis is given:

> *High quality musical performance can be achieved using immersive audio Network Music Performance systems designed for practical applications.*

This hypothesis was presented with the qualifying terms:

- *'High quality musical performance'*, characterised by synchrony which is comparable to reference values from live performance, and positive rating of performance quality, and the performance experience.
- *'Practical applications'*, where evaluation should be conducted using INMP designs which have been deployed in music industry applications, and where evaluation is conducted for conditions which are representative of practical WAN deployment.

The work presented in this thesis provides evidence which confirms this hypothesis. Practice–oriented evaluation of four INMP systems demonstrates that positive QoE, synchronous performance, and musical performance of comparable quality to live performance is supported. In three of the four cases (Chapter 4, Chapter 5 and Chapter 6), participants in evaluation were also professional musicians. It can also be acknowledged that in all cases performance of synchronous musical styles, including pop music, was possible, and was achievable by performers who did not have significant prior practice in remote performance. In this regard the qualifier of *'high quality musical performance'* is met.

The four INMP systems were all previously deployed over practical WAN connections. Evaluation of these systems was conducted either under latency conditions which approximate those experienced in WAN deployment, or in the case of Chapter 6, evaluated in practical application directly. In this regard the qualifier of *'practical applications'* is met. Cumulatively it is evidenced that low–latency INMP systems are indeed viable as a practical remote alternative to co–located performance workflows.

In this thesis, it is evidenced that immersive audio can support experience improvement with respect to factors such as spatial clarity, localisation, acoustic support of performance delivery, and plausibility of auralisation (where simulated acoustics allow performance to be monitored as heard in remote or virtual environments). It is, however, also evidenced that immersive audio monitoring can influence deterioration in experience quality with respect to factors such as clarity and RT in auralisation. Here it is identified that immersive audio monitoring is not a 'one–size–fits–all' solution, and that experience of immersive audio in practical INMP scenarios is highly dependent on monitoring context. This leads to practical recommendations to allow for engineer control over monitoring in INMP where both traditional stereo and immersive audio monitoring options are available, and parametric control of simulated acoustics is supported, such that it is possible to cater to different monitoring contexts encountered in practice.

8.4 Further Work

The following subsections provide detail on further work which is relevant to both the design and evaluation of INMP systems presented in this thesis.

8.4.1 Further INMP Evaluation

In the field of INMP, evaluation is often conducted through tapping studies [19], or through studies where performers perform to backing tracks [121]. It can be acknowledged that performance conditions in these studies are not necessarily representative of practical INMP scenarios. Indeed, across the studies presented in this thesis, it is highlighted that the INMP experience will vary across monitoring contexts considered for practical applications. For some monitoring contexts, the INMP experience did not necessarily conform to expectations which may be derived from tapping studies [19] and studies where performers play with backing tracks [121].

It is thus considered that it is not necessarily safe to assume that the findings of non–ecologically representative studies will extend to practical INMP scenarios. In the interest of developing a practical understanding of INMP which can inform best–practice in system design and engineer operation, practice–oriented evaluation is required in further work. Though the work detailed in this thesis contributes to this endeavour, it describes a limited set of monitoring contexts. Further practice–oriented evaluation is therefore expected to contribute to a fuller understanding, and establishment of best–practice recommendations in the emerging live sound production context of INMP.

Across this thesis a range of specific areas were identified that merit attention in further work. In Chapter 4 and Chapter 6 the potential for acoustics to impact the INMP experience is suggested. In Chapter 7 the impact of acoustics on the INMP experience is robustly demonstrated. Whereas the study presented in Chapter 7 addresses the impact of RT in INMP vocal performance, this is only one aspect of auralised acoustics. Further work should provide investigation of other acoustic parameters, and experience of acoustics in different monitoring contexts.

In Chapter 5 participants consider the experience of INMP immersive audio monitoring to provide improvements with respect to plausibility [125] and spatial clarity [180], however, these factors are not addressed in the rating questionnaire. In subsequent studies where factors of spatial clarity and plausibility are considered [19], [121] it is indeed demonstrated that experience improvement relative to a baseline mono audio system is detectable for these factors. It is therefore recommended that the INMP QoE evaluation is expanded to consider relevant rating items in future work.

In Chapter 6 participant comments and discussion highlights several significant aspects of the INMP concert experience which are not addressed in the rating questionnaire design. Items relating to spatial clarity and naturalness of loudspeaker reproduction [180] may be included in further evaluation. The experience of position–independent listening may also be the topic of further audience evaluation. Though NMP evaluation focuses on performance synchrony to describe performance quality [6], participants in Chapter 6 highlight other aspects of performance, particularly dynamic control, as relevant to the INMP experience. Indeed, dynamic, timbral and pitch qualities are evaluated in vocal performance studies [219], [220]. Further work may expand rating questionnaires to include evaluation of extended aspects of performance quality such as dynamics and pitch control.

In Chapter 4 the limitations of video display (AV asynchrony) are highlighted. In Chapter 5 the limitations of XR display (avatar representation and pass–through quality) are highlighted. Visual display is, however, not a focus of this thesis, and will require investigation in further work, particularly with respect to the impact of AV asynchrony and visual performance cues. Indeed, visual contact is well understood as important in communication of visual performance cues live group vocal performance [56], [224]. Successive work has already begun to further explore challenges associated with visual display in INMP [13], [19], [128]. Further evaluation is, however, required to establish a robust understanding of the impact of visual display and interface factors on INMP. Particularly the case of VR and XR performance applications [17], [25] are of interest to emerging industry practice in concert applications where INMP is considered as a solution [18], [20], [137].

In the second study detailed in Chapter 6 a novel listening test approach is presented for performance quality evaluation of INMP. Notably, however, only a dry stereo mix is presented in performance material. Further evaluation may consider extending the listening test approach to include evaluation of immersive audio rendering systems in order to better represent the INMP context. With respect to difference between listening test conditions and performance conditions it is considered that virtual reproduction of the audience experience of INMP concert may be achieved, such that listening tests may better represent the INMP evaluation context. Indeed resources such as 360° video and Ambisonic performance recordings are available via the dataset associated with the networked concert event. Use of continuous rating scales as used in relevant audio quality rating standard [180] may be adopted in future work.

It can be acknowledged that the studies presented in this thesis describe a limited set of immersive audio rendering methods and monitoring contexts. Further evaluation may address alternative immersive audio methods both in performer monitoring and audience distribution. Given the flexibility of the Ambisonic format as a processing solution [197] for INMP concert and audience distribution this may be the focus of further development.

8.4.2 Standardising INMP Evaluation

Particularly highlighted in Chapter 2 is the lack of coherence in metric definition, both with respect to QoE rating items and synchrony metrics. It is recommended that researchers and engineers collaborate and work towards a standard here in order to improve cohesion and repeatability in research. With respect to the work described in this thesis 'agreement' is predominantly used as a common scale in QoE rating. It is, however, considered that a standard poor–excellent continuous numerical scale [19], [20], [35], [110], [118], rather than ordinal agreement scale, may be more suitable.

Both in the work detailed in this thesis, and successive research elsewhere in the field of INMP, some success has been achieved in detecting variations in experience using novel immersive audio rating factors, such as localisation and spatial clarity [19], [121]. Audio evaluation standards are identified which define these attributes of immersive audio monitoring [180] and are recommended as a common basis of definition in future evaluation.

It is highlighted in this thesis that synchrony is the focus of evaluation of objective performance parameters in study of NMP [6]. It can be acknowledged that other physical parameters of performance, such as pitch and dynamics, are considered in wider musical performance evaluation [219], [220]. Some NMP study has begun to address the dynamic [111] and timbral features of NMP [108], [114]. It is suggested that a wider set of physical performance parameters is considered in further evaluation.

8.4.3 Interactive Auralisation

The practical INMP systems presented in this thesis consider varying levels of interactivity. The systems presented in Chapter 4 and Chapter 6 provide head–locked binaural rendering in musician headphone monitoring. Whereas some QoE improvement may be expected with head–locked immersive audio systems [16], extension to include 3DoF interactivity may result in greater experience improvement. Indeed studies which include 3DoF rendering have reported significant QoE improvement for interactive immersive audio relative to mono or stereo baseline conditions [19], [121]. Particularly with respect to INMP systems which include video display (for example in Chapter 4), webcam–based systems may be suitable for enabling 3DoF tracking [134], [163].

The INMP systems described in Chapter 5 and Chapter 7 provide binaural rendering with 3DoF interactivity. For both these systems, and head–locked systems, further development may extend INMP designs to support 6DoF interactivity. Convolution–based designs may be adapted from existing HMD–based auralisation systems [188]. Parametric auralisation methods [17], [25], [159] may also be implemented, and usefully allow for parametric control of auralisation as recommended from the findings of this thesis.

Notably, 6DoF mobility does not allow for participants to be tethered by wired monitoring systems. Pro–audio wireless microphone and monitoring systems are the only solution here which does not add unacceptable amounts of latency in audio acquisition and wireless monitor mix distribution. Integration of such wireless systems will, therefore, be required in implementation of 6DoF for low–latency INMP systems.

8.4.4 Further System Development

Across Chapter 4, Chapter 5 and Chapter 6, performers indicated a desire for individual control over monitoring mixes. Indeed the findings of these chapters indicate that preferences in musician monitoring will vary with specific monitoring contexts. Additionally, in Chapter 7, it is demonstrated that control of acoustics can be applied to optimise the INMP experience. Further system development should therefore include the provision of individual control of musician monitor mixes. As this includes auralisation parameters, auralisation solutions which allow parametric control [159] should be considered.

In Chapter 7 musicians commented on the usefulness of plausible auralisation in monitoring (allowing performers to tune performance for reproduction in remote spaces by allowing monitoring with plausible auralisation of the relevant remote space). More complex auralisation methods than the simplified system used in this chapter should therefore be explored in further work.

Though beyond the scope of this thesis, the engineer experience of INMP should also be considered in further work. In the concert production detailed in Chapter 7, lack of integrated production inter- faces, and the multi–endpoint live production context, was impractical for engineers to manage with traditional production equipment. The development of production solutions for INMP applications may improve Quality of Life (QoL) for engineers in remote workflows.

8.4.5 Visual Display

The work in this thesis contributes a novel demonstration of performer engagement with both video (Chapter 4) and XR (Chapter 5) visual display and interface. In related evaluations, limitations in visual display were also highlighted. With respect to video display the primary challenge is the potential for AV asynchrony between low–latency audio and video, which incurs higher OWT delay between endpoints [15]. In some practical NMP concert cases AV asynchrony is not perceived as problematic [15]. In other practical cases performers may opt to ignore or remove visual display due to AV asynchrony [35]. Implementation of low–latency video systems is a topic of current work [15], [38], and should be considered in further INMP system development.

As demonstrated in Chapter 5, XR display can be used as an alternative to video display, and has significantly lower bandwidth requirements. Several limitations were also highlighted, which should be the topic of further INMP system development. This includes avatar representation and pass–through display quality. Indeed, subsequent research [13] provides coherent findings, and works towards such improvements in XR INMP systems [13], [19], [136].

8.4.6 Industry Uptake

This thesis demonstrates the viability of low–latency INMP as a remote workflow solution in a range of music performance applications. The logical next stage in development is industry uptake. Pro–WAN music production and recording solutions, for example AudioMovers[1] or Source Elements[2], are in widespread industry use. Interest in remote performance workflows is also demonstrated by pro–LAN industry such as Dante[3] [39], [40], [135]. Given the rising popularity of remote workflows, here and existing remote workflow utility, this may be a particularly suitable area for industry–application of INMP solutions.

Virtual concert experiences are also an area of rising popularity. This includes holographic live concert events such as the ABBA Voyage experience[4], 360° video concerts, such as the Meta Superbowl concert featuring the Foo Fighters[5], and various virtual concert experiences delivered in game environments such as Fortnight[6] or Roblox[7], or developed by bespoke industry practitioners such as Wave[8] or AmazeVR[9]. This may be another suitable area for industry uptake, particularly with respect to VR and XR INMP systems which allow remote artists and audiences to interact in real–time in physical or virtual environments. Indeed practical interest in this area has already been expressed by major industry such as Qualcomm [20], [137].

It can be acknowledged that INMP is still in a research and development state, with practical implementation being achieved on a largely ad–hoc basis. Practical industry uptake will require packaged solutions which are designed to be practically convenient to configure and operate by audio engineers, and which can be readily integrated with modern immersive media distribution systems. The Ambisonic format is common in INMP design and is flexible with respect to playback systems encountered in immersive live events and broadcast [197]. Ambisonics is also in widespread use in modern immersive media contexts, for example in Resonance Audio[10] for VR and XR development, in 360° video platforms such as HOAST[11] [191] and Youtube[12], and in open–source immersive media formats such as Eclipsa Audio[13]. A wide range of practical sound production solutions (as outlined in Chapter 3) are also readily (and freely) available. It is therefore proposed that Ambisonics is the basis of efforts towards practical industry uptake of INMP systems.

[1] https://audiomovers.com/
[2] https://www.source-elements.com/
[3] https://www.getdante.com/
[4] https://abbavoyage.com/
[5] https://www.oculus.com/experiences/event/3316682238614608/
[6] https://fortnite.fandom.com/wiki/Live_Events
[7] https://roblox.fandom.com/wiki/Category:Concerts
[8] https://wavexr.com/past-waves/
[9] https://www.amazevr.com/artists
[10] https://resonance-audio.github.io/resonance-audio/
[11] https://hoast.iem.at/
[12] https://www.youtube.com/
[13] https://github.com/AOMediaCodec/iamf-tools

8.5 Closing Remarks

This thesis has presented a practice–oriented evaluation of low–latency INMP systems. The outcome of this work is a robust validation of INMP as a remote performance solution, capable of supporting high–quality synchronous performance and high–quality performance experiences. Despite this, the field of practical low–latency INMP is still in relative infancy, and requires further development to bridge the gap to widespread industry usage. It is hoped that the work contained in this thesis provides a solid foundation to support this development.

Appendices

Appendix A

Digital Appendices Overview

Appendices for this thesis are presented digitally via Zenodo[1]. An overview of the contents is presented here, as is direction to relevant resource which is already publicly available.

A.1 Chapter 4 Supplementary Material

- QoE rating data.
- BRIR measurements used in auralisation.

A.2 Chapter 5 Supplementary Material

- QoE rating data.
- SIR measurements used in auralisation are publicly available online[2] [5].

A.3 Chapter 6 Supplementary Material

- For the first part of this study QoE data, audio and video recording of concert performance, and recording of interview discussion are publicly available via an online dataset[3].
- Audio excerpts used in the listening test for the 2nd part of this study.
- QoE rating data from the 2nd part of this study.

[1] https://zenodo.org/records/14933099
[2] https://zenodo.org/records/10020866
[3] https://zenodo.org/records/10402916

A.4 Chapter 7 Supplementary Material

- Relevant SIR processing Matlab scripts are publicly available online[4].
- The MM–BOP Matlab onset detection and synchrony analysis functions are publicly available online[5].
- Resource for the test protocol framework is publicly available online[6].
- QoE rating data.
- Onset detection, synchrony analysis, and statistical analysis Matlab scripts and related figures.

[4] https://github.com/SpaceCadetAlba/IRutilities
[5] https://github.com/SpaceCadetAlba/MM-BOP
[6] https://github.com/SpaceCadetAlba/NMP-Evaluation-Method

References

[1] P. Cairns, A. Hunt, J. Cooper, *et al.*, "Recording Music in the Metaverse: A case study of XR BBC Maida Vale Recording Studios," in *Proceedings of the AES 2022 International Conference on Audio for Virtual and Augmented Reality*, Redmond, Washington, USA: Audio Engineering Society, Aug. 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21842.

[2] P. Cairns, A. Hunt, D. Johnston, *et al.*, "Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 313–325, Jun. 2023, ISSN: 15494950. DOI: 10.17743/jaes.2022.0086. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22139 (visited on 02/25/2025).

[3] P. Cairns, H. Daffern, and G. Kearney, "A DAW-based approach to immersive audio system evaluation in Network Music Performance contexts," in *Proceedings of the 2023 AES International Conference on Spatial and Immersive Audio*, Huddersfield, United Kingdom: Audio Engineering Society, Aug. 2023. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22186.

[4] P. Cairns, T. Rudzki, J. Cooper, *et al.*, "Singer and Audience Evaluations of a Networked Immersive Audio Concert," *Journal of the Audio Engineering Society*, vol. 72, no. 7/8, pp. 467–478, Jul. 2024, ISSN: 15494950. DOI: 10.17743/jaes.2022.0145. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22644 (visited on 02/25/2025).

[5] G. Kearney, H. Daffern, P. Cairns, *et al.*, "Measuring the Acoustical Properties of the BBC Maida Vale Recording Studios for Virtual Reality," en, *Acoustics*, vol. 4, no. 3, pp. 783–799, Sep. 2022, ISSN: 2624-599X. DOI: 10.3390/acoustics4030047. [Online]. Available: https://www.mdpi.com/2624-599X/4/3/47 (visited on 07/03/2024).

[6] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2016.2628440. [Online]. Available: http://ieeexplore.ieee.org/document/7769205/ (visited on 04/27/2024).

[7] 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.

[8] P. Cairns, H. Daffern, and G. Kearney, "Parametric Evaluation of Ensemble Vocal Performance Using an Immersive Network Music Performance Audio System," *Journal of the Audio Engineering Society*, vol. 69, no. 12, pp. 924–933, Dec. 2021, ISSN: 15494950. DOI: 10.17743/jaes.2021.0040. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21541 (visited on 04/27/2024).

[9] R. Hupke, J. Dürre, N. Werner, and J. Peissig, "Latency and Quality-of-Experience Analysis of a Networked Music Performance Framework for Realistic Interaction," in *AES 152nd Convention*, New York, USA, May 2022. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21659.

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, en, 2nd ed. London, England: The MIT Press, Oct. 1996, ISBN: 978-0-262-26868-4. DOI: 10.7551/mitpress/6391.001.0001. [Online]. Available: https://direct.mit.edu/books/book/4885/Spatial-HearingThe-Psychophysics-of-Human-Sound (visited on 04/27/2024).

[11] M. Kob, S. V. Amengual Garí, and Z. Schärer Kalkandjiev, "Room Effect on Musicians' Performance," en, in *The Technology of Binaural Understanding*, J. Blauert and J. Braasch, Eds., Series Title: Modern Acoustics and Signal Processing, Cham, Switzerland: Springer International Publishing, 2020, pp. 223–249, ISBN: 9783030003869. DOI: 10.1007/978-3-030-00386-9_9. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00386-9_9 (visited on 04/29/2024).

[12] 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, ISBN: 9798350311044. DOI: 10.1109/I3DA57090.2023.10289169. [Online]. Available: https://ieeexplore.ieee.org/document/10289169/ (visited on 04/28/2024).

[13] 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, UK: Audio Engineering Society, Aug. 2023. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22183.

[14] L. Turchet, "Musical Metaverse: Vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, vol. 27, pp. 1811–1827, 2023. DOI: 10.1007/s00779-023-01708-1.

[15] A. Carôt, 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, ISBN: 978-1-72811-485-9. DOI: 10.1109/ICMEW46912.2020.9105995. [Online]. Available: https://ieeexplore.ieee.org/document/9105995/ (visited on 04/27/2024).

[16] A. F. Genovese, "Acoustics and Copresence: Towards Effective Auditory Virtual Environments for Distributed Music Performances," PhD Thesis, New York University, 2023. [Online]. Available: https://www.proquest.com/openview/fb49e48e7fdd20d010e2505cee3ae087/1?pq-origsite=gscholar&cbl=18750&diss=y.

[17] J.-M. Jot, T. Carpentier, and O. Warusfel, "Perceptually Motivated Spatial Audio Scene Description and Rendering for 6-DoF Immersive Music Experiences," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Bologna, Italy: IEEE, Sep. 2023, pp. 1–14, ISBN: 9798350311044. DOI: 10.1109/I3DA57090.2023.10289196. [Online]. Available: https://ieeexplore.ieee.org/document/10289196/ (visited on 04/27/2024).

[18] 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, ISSN: 15494950. DOI: 10.17743/jaes.2022.0096. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22345 (visited on 04/29/2024).

[19] R. Hupke, S. Preihs, and J. Peissig, "Immersive Networked Music Performance: Impact of Extended Reality on the Quality of Experience," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–9, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704092. [Online]. Available: https://ieeexplore.ieee.org/document/10704092/ (visited on 10/09/2024).

[20] A. F. Genovese, Z. Nguyen, M. Gospodarek, R. Pahle, C. Brenner, and A. Roginska, "Holodeck: A Research Framework for Distributed Multimedia Concert Performances," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–10, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704113. [Online]. Available: https://ieeexplore.ieee.org/document/10704113/ (visited on 10/09/2024).

[21] M. Gurevich, D. Donohoe, and S. Bertet, "Ambisonic spatialization for networked music performance," in *17th International Conference on Auditory Display*, Budapest, Hungary, Jun. 2011.

[22] W. Ritsch, "ICE - Towards Distributed Networked Computer Music Ensemble," in *Joint Proceedings of the International Computer Music Conference 2014 and the 11th Sound and Music Computing Conference*, Athens, Greece, Sep. 2014.

[23] R. Hoy and D. Van Nort, "A Technological and Methodological Ecosystem for Dynamic Virtual Acoustics in Telematic Performance Contexts," en, in *Audio Mostly 2021*, virtual/Trento Italy: ACM, Sep. 2021, pp. 169–174, ISBN: 978-1-4503-8569-5. DOI: 10.1145/3478384.3478425. [Online]. Available: https://dl.acm.org/doi/10.1145/3478384.3478425 (visited on 04/29/2024).

[24] B. Mróz, P. Odya, P. Danowski, and M. Kabaciński, "A commonly-accessible toolchain for live streaming music events with higher-order ambisonic audio and 4k 360 vision," in *AES International Conference on Spatial and Immersive Audio*, Huddersfield, UK, Aug. 2023.

[25] 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, ISBN: 978-1-66540-998-8. DOI: 10.1109/I3DA48870.2021.9610971. [Online]. Available: https://ieeexplore.ieee.org/document/9610971/ (visited on 04/29/2024).

[26] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1/2, pp. 2–10, Feb. 1973.

[27] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer Topics in Signal Processing), en. Cham: Springer International Publishing, 2019, vol. 19, ISBN: 9783030172091. DOI: 10.1007/978-3-030-17207-7. [Online]. Available: http://link.springer.com/10.1007/978-3-030-17207-7 (visited on 04/29/2024).

[28] A. Barbosa, "Displaced Soundscapes: A Survey of Network Systems for Music and Sonic Art Creation," en, *Leonardo Music Journal*, vol. 13, pp. 53–59, Dec. 2003, ISSN: 0961-1215, 1531-4812. DOI: 10.1162/096112104322750791. [Online]. Available: https://direct.mit.edu/lmj/article/63347 (visited on 08/11/2024).

[29] D. Konstantas, "Overview of a telepresence environment for distributed musical rehearsals," en, in *Proceedings of the 1998 ACM symposium on Applied Computing - SAC '98*, Atlanta, Georgia, United States: ACM Press, 1998, pp. 456–457, ISBN: 978-0-89791-969-2. DOI: 10.1145/330560.330857. [Online]. Available: http://portal.acm.org/citation.cfm?doid=330560.330857 (visited on 08/13/2024).

[30] A. Xu and J. Cooperstock, "Real-Time Streaming of Multichannel Audio Data over Internet," in *AES 108th Convention*, Paris: Audio Engineering Society, Feb. 2000. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=9218.

[31] C. Chafe, S. Wilson, R. Leistikow, D. Chisholm, and G. Scavone, "A Simplified Approach to High Quality Music and Sound Over IP," in *DAFX-00*, Verona, Italy, Dec. 2000.

[32] J. Lazzaro and J. Wawrzynek, "A case for network musical performance," en, in *Proceedings of the 11th international workshop on Network and operating systems support for digital audio and video*, Port Jefferson New York USA: ACM, Jan. 2001, pp. 157–166, ISBN: 978-1-58113-370-7. DOI: 10.1145/378344.378367. [Online]. Available: https://dl.acm.org/doi/10.1145/378344.378367 (visited on 08/12/2024).

[33] A. A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to Distributed Immersive Performance," en, in *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence - ETP '03*, Berkeley, California, USA: ACM Press, 2003, p. 110, ISBN: 978-1-58113-775-0. DOI: 10.1145/982484.982506. [Online]. Available: http:

//portal.acm.org/citation.cfm?doid=982484.982506 (visited on 08/13/2024).

[34] R. Zimmermann, C. Kyriakakis, C. Shahabi, C. Papadopoulos, A. Sawchuk, and U. Neumann, "RMI system: Internet meets the future home theater," en, *IEEE Multimedia*, vol. 11, no. 2, pp. 48–57, Apr. 2004, ISSN: 1070-986X. DOI: 10.1109/MMUL.2004.1289041. [Online]. Available: http://ieeexplore.ieee.org/document/1289041/ (visited on 08/13/2024).

[35] 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, ISSN: 15494950. DOI: 10.17743/jaes.2021.0056. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21542 (visited on 04/27/2024).

[36] C. Hoene, I. Howell, and A. Carôt, "Networked Music Performance: Developing Soundjack and the Fastmusic Box During the Coronavirus Pandemic," in *2021 AES International Conference on Audio Education*, Online: Audio Engineering Society, Jul. 2022. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=21216.

[37] L. Turchet and C. Fischione, "Elk Audio OS: An Open Source Operating System for the Internet of Musical Things," en, *ACM Transactions on Internet of Things*, vol. 2, no. 2, pp. 1–18, May 2021, ISSN: 2691-1914, 2577-6207. DOI: 10.1145/3446393. [Online]. Available: https://dl.acm.org/doi/10.1145/3446393 (visited on 08/14/2024).

[38] S. Ubik, J. Halák, M. Kolbe, J. Melnikov, and M. Frič, "Lessons Learned from Distance Collaboration in Live Culture," en, in *Human Interaction, Emerging Technologies and Future Applications IV*, T. Ahram, R. Taiar, and F. Groff, Eds., vol. 1378, Series Title: Advances in Intelligent Systems and Computing, Cham: Springer International Publishing, 2021, pp. 608–615, ISBN: 9783030855390. DOI: 10.1007/978-3-030-74009-2_77. [Online]. Available: https://link.springer.com/10.1007/978-3-030-74009-2_77 (visited on 08/14/2024).

[39] P. Ferguson and C. Chafe, "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," in *AES 148th Convention*, Online, May 2020.

[40] A. Peleg, G. Dickins, and P. Killianey, *Peltrix Takes Dante Over Distance*, Webinar, Sep. 2021. [Online]. Available: https://go.audinate.com/resources/assets/blue-note-peltrix-webinar-2021-ty.

[41] J. C. Wu and S. L. Miller, "Recording Arts and Studio Engineering for High-fidelity Multitrack Network Music Production," in *AES 152nd Convention*, The Hague, Netherlands: Audio Engineering Society, May 2022.

[42] M. Iorwerth and D. Knox, "Challenges of using Networked Music Performance in education," in *UK 26th Conference: Audio Eduction*, Aug. 2015. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=17857.

[43] B. Redman, "Evaluating the use of LoLa in European conservatoires: The SWING project," en, *Journal of Music, Technology & Education*, vol. 13, no. 2, pp. 199–217, Dec. 2020, ISSN: 1752-7066, 1752-7074. DOI: 10.1386/jmte_00023_1. [Online]. Available: https://intellectdiscover.com/content/journals/10.1386/jmte_00023_1 (visited on 08/23/2024).

[44] M. Iorwerth and D. Knox, "The application of Networked Music Performance technology to access ensemble activity for socially isolated musicians," in *Web Audio Conference 2019*, Trondheim, Norway, Dec. 2019.

[45] J. Tamplin, B. Loveridge, K. Clarke, Y. Li, and D. 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," en, *Journal of Telemedicine and Telecare*, vol. 26, no. 6, pp. 365–375, Jul. 2020, ISSN: 1357-633X, 1758-1109. DOI: 10.1177/1357633X19828463. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1357633X19828463 (visited on 04/27/2024).

[46] T. Kostas, M. Borella, I. Sidhu, G. Schuster, J. Grabiec, and J. Mahler, "Real-time voice over packet-switched networks," *IEEE Network*, vol. 12, no. 1, pp. 18–27, Feb. 1998, ISSN: 08908044. DOI: 10.1109/65.660003. [Online]. Available: http://ieeexplore.ieee.org/document/660003/ (visited on 08/12/2024).

[47] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, pp. 586–593, May 1991, ISSN: 07338716. DOI: 10.1109/49.81952. [Online]. Available: http://ieeexplore.ieee.org/document/81952/ (visited on 08/12/2024).

[48] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, Jun. 2015, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00731. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00731/abstract (visited on 08/13/2024).

[49] M. Baldi and Y. Ofek, "End-to-end delay analysis of videoconferencing over packet-switched networks," *IEEE/ACM Transactions on Networking*, vol. 8, no. 4, pp. 479–492, Aug. 2000, ISSN: 10636692. DOI: 10.1109/90.865076. [Online]. Available: http://ieeexplore.ieee.org/document/865076/ (visited on 08/12/2024).

[50] J. E. Boland, P. Fonseca, I. Mermelstein, and M. Williamson, "Zoom disrupts the rhythm of conversation.," en, *Journal of Experimental Psychology: General*, vol. 151, no. 6, pp. 1272–1282, Jun. 2022, ISSN: 1939-2222, 0096-3445. DOI: 10.1037/xge0001150. [Online]. Available: https://doi.apa.org/doi/10.1037/xge0001150 (visited on 08/10/2024).

[51] H. Daffern, K. Balmer, and J. Brereton, "Singing Together, Yet Apart: The Experience of UK Choir Members and Facilitators During the Covid-19 Pandemic," *Frontiers in Psychology*, vol. 12, p. 624 474, Feb. 2021, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.624474. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.624474/full (visited on 08/25/2024).

[52] A. Martínez-Hernández, "Conservatory and music schoolteachers' experiences with videoconferencing software during and after the COVID-19 pandemic," en, *Distance Education*, vol. 43, no. 3, pp. 388–407, Jul. 2022, ISSN: 0158-7919, 1475-0198. DOI: 10.1080/01587919.2022.2088475. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01587919.2022.2088475 (visited on 08/25/2024).

[53] B. H. Repp, "Sensorimotor synchronization: A review of the tapping literature," en, *Psychonomic Bulletin & Review*, vol. 12, no. 6, pp. 969–992, Dec. 2005, ISSN: 1069-9384, 1531-5320. DOI: 10.3758/BF03206433. [Online]. Available: http://link.springer.com/10.3758/BF03206433 (visited on 08/25/2024).

[54] B. H. Repp and Y.-H. Su, "Sensorimotor synchronization: A review of recent research (2006–2012)," en, *Psychonomic Bulletin & Review*, vol. 20, no. 3, pp. 403–452, Jun. 2013, ISSN: 1069-9384, 1531-5320. DOI: 10.3758/s13423-012-0371-2. [Online]. Available: http://link.springer.com/10.3758/s13423-012-0371-2 (visited on 08/25/2024).

[55] R. A. Rasch, "Synchronization in Performed Ensemble Music," *Acta Acustica united with Acustica*, vol. 43, no. 2, pp. 121–131, Sep. 1979.

[56] S. D'Amario, H. Daffern, and F. Bailes, "Synchronization in Singing Duo Performances: The Roles of Visual Contact and Leadership Instruction," *Frontiers in Psychology*, vol. 9, p. 1208, Jul. 2018, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.01208. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2018.01208/full (visited on 08/25/2024).

[57] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan, "Effect of Time Delay on Ensemble Accuracy," in *International Symposium on Musical Acoustics*, Nara, Japan, Apr. 2004.

[58] C. Chafe and M. Gurevich, "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry," in *AES 117th Convention*, San Francisco, California, USA: Audio Engineering Society, Oct. 2004.

[59] D. M. Howard and J. A. S. Angus, *Acoustics and psychoacoustics*, eng, 4. ed. Amsterdam Heidelberg: Elsevier, 2009, ISBN: 978-0-240-52175-6.

[60] F. Rumsey and T. McCormick, *Sound and recording: applications and theory*, eng, 7th edition. New York: Focal Press, 2014, ISBN: 9780080953960.

[61] A. Carôt, P. Rebelo, and A. Renaud, "Networked Music Performance: State of the Art," in *AES 30th International Conference*, Saariselka, Finland: Audio Engineering Society, Mar. 2007.

[62] A. Carôt, U. Kramer, and G. Schuller, "Network Music Performance in Narrow Band Networks," in *Audio Engineering Society 120th Convention*, Paris, France, May 2006. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=13528.

[63] R. Hupke, S. Sridhar, A. Genovese, *et al.*, "A Latency Measurement Method for Network Music Performance," in *AES 147th Convention*, New York, USA, Oct. 2019.

[64] U. Zölzer and X. Amatriain, *DAFX: digital audio effects*, eng. Chichester, England: Wiley, 2002, OCLC: 51795031, ISBN: 978-0-470-84604-9.

[65] M. Tomasetti, A. Farina, and L. Turchet, "Latency of spatial audio plugins: A comparative study," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Bologna, Italy: IEEE, Sep. 2023, pp. 1–10, ISBN: 9798350311044. DOI: 10.1109/I3DA57090.2023.10289279. [Online]. Available: https://ieeexplore.ieee.org/document/10289279/ (visited on 08/27/2024).

[66] B. Grimes, *Networked AV systems*. New York: McGraw-Hill Eduction, 2014, OCLC: ocn857402127, ISBN: 978-0-07-182573-3.

[67] L. Turchet and P. Casari, "Assessing a Private 5G SA and a Public 5G NSA Architecture for Networked Music Performances," in *2023 4th International Symposium on the Internet of Sounds*, Pisa, Italy: IEEE, Oct. 2023, pp. 1–6, ISBN: 9798350382549. DOI: 10.1109/IEEECONF59510.2023.10335359. [Online]. Available: https://ieeexplore.ieee.org/document/10335359/ (visited on 08/14/2024).

[68] 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, ISSN: 2327-4662, 2372-2541. DOI: 10.1109/JIOT.2023.3288818. [Online]. Available: https://ieeexplore.ieee.org/document/10162178/ (visited on 07/04/2024).

[69] J. Dürre, N. Werner, S. Hämäläinen, *et al.*, "In-depth latency and reliability analysis of a networked music performance over public 5G infrastructure," in *Audio Engineering Society Convention 153*, New York, USA: Audio Engineering Society, Oct. 2022.

[70] 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, ISSN: 2644-125X. DOI: 10.1109/OJCOMS.2024.3407708. [Online]. Available: https://ieeexplore.ieee.org/document/10543134/ (visited on 07/03/2024).

[71] D. Comer and D. L. Stevens, *Internetworking with TCP/IP*, 3rd ed. Upper Saddle River, N.J: Prentice Hall, 1995, ISBN: 978-0-13-216987-5.

[72] J. Postel, "Internet Protocol," en, RFC Editor, Tech. Rep. RFC0791, Sep. 1981, RFC0791. DOI: 10.17487/rfc0791. [Online]. Available: https://www.rfc-editor.org/info/rfc0791 (visited on 09/30/2024).

[73] S. Deering and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification," en, RFC Editor, Tech. Rep. RFC8200, Jul. 2017, RFC8200. DOI: 10.17487/RFC8200. [Online]. Available: https://www.rfc-editor.org/info/rfc8200 (visited on 09/30/2024).

[74] J. Postel, "User Datagram Protocol," en, RFC Editor, Tech. Rep. RFC0768, Aug. 1980, RFC0768. DOI: 10.17487/rfc0768. [Online]. Available: https://www.rfc-editor.org/info/rfc0768 (visited on 09/29/2024).

[75] J.-P. Cáceres and C. Chafe, "JackTrip: Under the Hood of an Engine for Network Audio," en, *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, Sep. 2010, ISSN: 0929-8215, 1744-5027. DOI: 10.1080/09298215.2010.481361. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09298215.2010.481361 (visited on 08/12/2024).

[76] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," en, RFC Editor, Tech. Rep. RFC3550, Jul. 2003, RFC3550. DOI: 10.17487/rfc3550. [Online]. Available: https://www.rfc-editor.org/info/rfc3550 (visited on 09/30/2024).

[77] J.-C. Bolot and A. Vega-Garcia, "Control mechanisms for packet audio in the Internet," in *Proceedings of IEEE INFOCOM '96. Conference on Computer Communications*, vol. 1, San Francisco, CA, USA: IEEE Comput. Soc. Press, 1996, pp. 232–239, ISBN: 978-0-8186-7293-4. DOI: 10.1109/INFCOM.1996.497898. [Online]. Available: http://ieeexplore.ieee.org/document/497898/ (visited on 10/01/2024).

[78] IEEE, *IEEE 802.1BA-2021: IEEE Standard for Local and Metropolitan Area Networks - Audio Video Bridging (AVB) Systems*, New York, USA, Mar. 2023. [Online]. Available: https://standards.ieee.org/ieee/802.1BA/10547/.

[79] AES, *AES67-2023: AES standards for audio applications of networks - High-performance streaming audio-over-ip interoperability*, New York, 2023.

[80] V. Shankarkumar, L. Montini, T. Frost, and G. Dowd, "Precision Time Protocol Version 2 (PTPv2) Management Information Base," en, RFC Editor, Tech. Rep. RFC8173, Jun. 2017, RFC8173. DOI: 10.17487/RFC8173. [Online]. Available: https://www.rfc-editor.org/info/rfc8173 (visited on 10/10/2024).

[81] D. Mills, J. Martin, J. Burbank, and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification," en, RFC Editor, Tech. Rep. RFC5905, Jun. 2010, RFC5905. DOI: 10.17487/rfc5905. [Online]. Available: https://www.rfc-editor.org/info/rfc5905 (visited on 09/30/2024).

[82] C. Stais, Y. Thomas, G. Xylomenos, and C. Tsilopoulos, "Networked music performance over information-centric networks," in *2013 IEEE International Conference on Communications Workshops (ICC)*, Budapest, Hungary: IEEE, Jun. 2013, pp. 647–651, ISBN: 978-1-4673-5753-1. DOI: 10.1109/ICCW.2013.6649313. [Online]. Available: http://ieeexplore.ieee.org/document/6649313/ (visited on 10/01/2024).

[83] D. K. Goot, H. Chaubey, T. Y. Hsu, and W. S. Deal, "A Perceptual Evaluation of Music Real-Time Communication Applications," *IEEE Access*, vol. 11, pp. 46 860–46 870, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3271525. [Online]. Available: https://ieeexplore.ieee.org/document/10110963/ (visited on 10/03/2024).

[84] D. Akoumianakis, C. Alexandraki, V. Alexiou, *et al.*, "The MusiNet project: Addressing the challenges in Networked Music Performance systems," in *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Corfu, Greece: IEEE, Jul. 2015, pp. 1–6, ISBN: 978-1-4673-9311-9. DOI: 10.1109/IISA.2015.7388002. [Online]. Available: https://ieeexplore.ieee.org/document/7388002 (visited on 10/01/2024).

[85] W. Eddy, "Transmission Control Protocol (TCP)," en, RFC Editor, Tech. Rep. RFC9293, Aug. 2022, RFC9293. DOI: 10.17487/RFC9293. [Online]. Available: https://www.rfc-editor.org/info/rfc9293 (visited on 09/29/2024).

[86] J. Rosenberg, H. Schulzrinne, G. Camarillo, *et al.*, "SIP: Session Initiation Protocol," en, RFC Editor, Tech. Rep. RFC3261, Jun. 2002, RFC3261. DOI: 10.17487/rfc3261. [Online]. Available: https://www.rfc-editor.org/info/rfc3261 (visited on 09/30/2024).

[87] J. P. Cáceres and C. Chafe, "JackTrip/SoundWIRE Meets Server Farm," *Computer Music Journal*, vol. 34, pp. 29–34, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:13809219.

[88] 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, ISBN: 9798350382549. DOI: 10.1109/IEEECONF59510.2023.10335270. [Online]. Available: https://ieeexplore.ieee.org/document/10335270/ (visited on 04/27/2024).

[89] A. Carôt, C. Hoene, H. Busse, and C. Kuhr, "Results of the Fast-Music Project—Five Contributions to the Domain of Distributed Music," *IEEE Access*, vol. 8, pp. 47 925–47 951, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2979362. [Online]. Available: https://ieeexplore.ieee.org/document/9027818/ (visited on 10/01/2024).

[90] J.-M. Valin, G. Maxwell, T. Terriberry, and K. Vos, "High-Quality, Low-Delay Music Coding in the Opus Codec," in *135th AES Convention*, New York, USA: Audio Engineering Society, Oct. 2013.

[91] JM. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," en, RFC Editor, Tech. Rep. RFC6716, Sep. 2012, RFC6716. DOI: 10.17487/rfc6716. [Online]. Available: https://www.rfc-editor.org/info/rfc6716 (visited on 10/01/2024).

[92] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy, and G. Kearney, "Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes," en, *Applied Sciences*, vol. 9, no. 13, p. 2618, Jun. 2019, ISSN: 2076-3417. DOI: 10.3390/app9132618. [Online]. Available: https://www.mdpi.com/2076-3417/9/13/2618 (visited on 10/01/2024).

[93] 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, Apr. 2023, ISSN: 15494950. DOI: 10.17743/jaes.2022.0068. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22037 (visited on 10/01/2024).

[94] A. Carôt, "Musical Telepresence - A Comprehensive Analysis Towards New Cognitive and Technical Approaches," PhD Thesis, University of Lübeck, Lübeck, May 2009. [Online]. Available: https://www.researchgate.net/profile/Alexander-Carot/publication/267206490_Fundamentals_and_Principles_of_Musical_Telepresence/links/622f09223c53d31ba4c33dfc/Fundamentals-and-Principles-of-Musical-Telepresence.pdf?_tk=u7I0mPo6GF8EUW2KEzZ8u8K1hBnl7XOB1Vee0TDSQL8-1754753968-1.0.1.1-62RxdIvs2Pm4IZEPKHzkvd2MfvlszOBHybL_VtMyZGU.

[95] U. Kramer, G. Schuller, S. Wabnik, J. Klier, and J. Hirschfeld, "Ultra Low Delay audio coding with constant bitrate," in *Proceedings of the Audio Engineering Society 117th Convention*, San Francisco, California, USA: Audio Engineering Society, Oct. 2004.

[96] E. Chew, R. Zimmermann, A. A. Sawchuk, C. Kyriakakis, and C. Papadopoulos, "Musical Interaction at a Distance: Distributed Immersive Performance," in *4th Music Network Open Workshop*, 2004. [Online]. Available: https://webspace.science.uu.nl/~fleis102/MusicNetwork-IMSC.pdf.

[97] J.-P. Cáceres, R. Hamilton, D. Iyer, C. Chafe, and G. Wang, *To the Edge with China: Explorations in Network Performance*. Porto, Portugal, 2008, Publication Title: ARTECH 2008: Proceedings of the 4th International Conference on Digital Arts, ISBN: 978-989-95776-3-3.

[98] I. T. Union, *ITU-T H.264: Advanced video coding for generic audiovisual services*, Aug. 2024. [Online]. Available: https://www.itu.int/rec/t-rec-h.264.

[99] C. Drioli, C. Allocchio, and N. Buso, "Networked Performances and Natural Interaction via LOLA: Low Latency High Quality A/V Streaming System," in *Information Technologies for Performing Arts, Media Access, and Entertainment*, D. Hutchison, T. Kanade, J. Kittler, *et al.*, Eds., vol. 7990, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 240–250, ISBN: 9783642400490. DOI: 10.1007/978-3-642-40050-6_21. [Online]. Available: http://link.springer.com/10.1007/978-3-642-40050-6_21 (visited on 10/03/2024).

[100] B. Loveridge, "Networked Music Performance in Virtual Reality: Current Perspectives," *Journal of Network Music and Arts*, vol. 2, no. 1, 2020, ISSN: 2690-2044.

[101] L. Turchet, N. Garau, and N. Conci, "Networked Musical XR: Where's the limit? A preliminary investigation on the joint use of point clouds and low-latency audio communication," en, in *AudioMostly 2022*, St. Pölten Austria: ACM, Sep. 2022, pp. 226–230, ISBN: 978-1-4503-9701-8. DOI: 10.1145/3561212.3561237. [Online]. Available: https://dl.acm.org/doi/10.1145/3561212.3561237 (visited on 04/29/2024).

[102] 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., vol. 3, Series Title: Current Research in Systematic Musicology, Cham: Springer International Publishing, 2017, pp. 125–145, ISBN: 9783319472140. DOI: 10.1007/978-3-319-47214-0_9. [Online]. Available: http://link.springer.com/10.1007/978-3-319-47214-0_9 (visited on 07/04/2024).

[103] Cockos, *Up and Running: A REAPER User GUide v 7.20*, Aug. 2024. [Online]. Available: https://www.reaper.fm/userguide.php.

[104] H. von Coler, N. Tonnatt, V. Kather, and C. Chafe, "SPRAWL: A network system for enhanced interaction in musical ensembles," in *Proceedings of the 18th Linux Audio Conference*, Bordeaux, France, Nov. 2020.

[105] T. A. Rushton, R. Michon, and S. Letz, "A Microcontroller-Based Network Client Towards Distributed Spatial Audio," in *Proceedings of the 2023 Sound and Music Computing Conference*, Stockholm, Sweden, Jun. 2023.

[106] 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, ISBN: 9798350382549. DOI: 10.1109/IEEECONF59510.2023.10335480. [Online]. Available: https://ieeexplore.ieee.org/document/10335480/ (visited on 04/28/2024).

[107] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005, ISSN: 1063-6676. DOI: 10.1109/TSA.2005.851998. [Online]. Available: http://ieeexplore.ieee.org/document/1495485/ (visited on 10/06/2024).

[108] C. Rottondi, M. Buccoli, and M. Zanoni, "Feature-Based Analysis of the Effects of Packet Delay on Networked Musical Interactions," *Journal of the Audio Engineering Society*, vol. 63, no. 11, pp. 864–875, Dec. 2015, ISSN: 15494950. DOI: 10.17743/jaes.2015.0074. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18047 (visited on 10/03/2024).

[109] C. Bartlette, D. Headlam, M. Bocko, and G. Velikic, "Effect of Network Latency on Interactive Musical Performance," en, *Music Perception*, vol. 24, no. 1, pp. 49–62, Sep. 2006, ISSN: 0730-7829, 1533-8312. DOI: 10.1525/mp.2006.24.1.49. [Online]. Available: https://online.ucpress.edu/mp/article/24/1/49/62315/Effect-of-Network-Latency-on-Interactive-Musical (visited on 10/03/2024).

[110] S. Farner, A. Solvang, A. Saebo, and U. P. Svensson, "Ensemble Hand-Clapping Experiments under the Influence of Delay and Various Acoustic Environments," *Journal of the Audio Engineering Society*, vol. 57, no. 12, pp. 1028–1041, Dec. 2009. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=15235.

[111] A. Barbosa and J. Cardeiro, "The Influence of Perceptual Attack Times in Networked Music Performance," in *AES 44th International Conference*, San Diego, California, USA: Audio Engineering Society, Nov. 2011. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=16133.

[112] N. Schuett, "The Effects if Latency on Ensemble Performance," Ph.D. dissertation, Stanford University, Stanford, California, May 2002. [Online]. Available: https://ccrma.stanford.edu/groups/soundwire/publications/papers/schuett_honorThesis2002.pdf.

[113] S. Delle Monache, L. Comanducci, M. Buccoli, *et al.*, "A Presence- and Performance-Driven Framework to Investigate Interactive Networked Music Learning Scenarios," en, *Wireless Communications and Mobile Computing*, vol. 2019, pp. 1–20, Aug. 2019, ISSN: 1530-8669, 1530-8677. DOI: 10.1155/2019/4593853. [Online]. Available: https://www.hindawi.com/journals/wcmc/2019/4593853/ (visited on 10/03/2024).

[114] K. Tsioutas and G. Xylomenos, "On the Impact of Audio Characteristics to the Quality of Musicians' Experience in Network Music Performance," *Journal of the Audio Engineering Society*, vol. 69, no. 12, pp. 914–923, Dec. 2021, ISSN: 15494950. DOI: 10.17743/jaes.2021.0041. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21540 (visited on 10/03/2024).

[115] R. Hupke, M. Nophut, S. Preihs, and J. Peissig, "Toward Professional Distributed Performances: Effects of a Global Metronome on Networked Musical Ensemble Interactions," *Journal of the Audio Engineering Society*, vol. 69, no. 10, pp. 720–736, Oct. 2021, ISSN: 15494950. DOI: 10.17743/jaes.2021.0037. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21467 (visited on 10/09/2024).

[116] E. Chew, A. Sawchuk, C. Tanoue, and R. Zimmermann, "Segmental Tempo Analysis of Performances in User-centered Experiments in the Distributed Immersive Performance Project," in *2nd Sound and Music Computing Conference*, Salerno, Italy, Nov. 2005.

[117] ITU, *Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems*, Oct. 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1116-3-201502-I/en.

[118] K. Tsioutas, G. Xylomenos, I. Doumanis, and C. Angelou, "Quality of Musicians' Experience in Network Music Performance: A Subjective Evaluation," in *Proceedings of the AES 148th Convention*, Online: Audio Engineering Society, Jun. 2020. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=20774.

[119] A. Carôt, C. Werner, and T. Fischinger, "Towards a Comprehensive Cognitive Analysis of Delay-Influences Rhythmical Interaction," in *International Conference on Mathematics and Computing*, Montreal, Canada, 2009.

[120] A. Olmos, M. Brule, N. Bouillot, *et al.*, "Exploring the role of latency and orchestra placement on the networked performance of a distributed opera," in *12th Annual International Workshop on Presence*, Los Angeles, California, USA, 2009.

[121] M. Tomasetti and L. Turchet, "Playing With Others Using Headphones: Musicians Prefer Binaural Audio With Head Tracking Over Stereo," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 501–511, Jun. 2023, ISSN: 2168-2291, 2168-2305. DOI: 10.1109/THMS.2023.3270703. [Online]. Available: https://ieeexplore.ieee.org/document/10124058/ (visited on 04/28/2024).

[122] M. Iorwerth, "Playing together, apart: An exploration of the challenges of Networked Music Performance in informal contexts," PhD Thesis, Glasgow Caledonian University, Glasgow, Mar. 2019.

[123] P. F. Driessen, T. E. Darcie, and B. Pillay, "The Effects of Network Delay on Tempo in Musical Performance," en, *Computer Music Journal*, vol. 35, no. 1, pp. 76–89, Mar. 2011, ISSN: 0148-9267, 1531-5169. DOI: 10.1162/COMJ_a_00041. [Online]. Available: https://direct.mit.edu/comj/article/35/1/76-89/94321 (visited on 10/03/2024).

[124] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments: A Presence Questionnaire," en, *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 3, pp. 225–240, Jun. 1998, ISSN: 1054-7460. DOI: 10.1162/105474698565686. [Online]. Available: https://direct.mit.edu/pvar/article/7/3/225-240/92643 (visited on 10/08/2024).

[125] H. Lee, *A Conceptual Model of Immersive Experience in Extended Reality*, Sep. 2020. DOI: 10.31234/osf.io/sefkh. [Online]. Available: https://osf.io/sefkh (visited on 10/08/2024).

[126] S. Zhao, "Toward a Taxonomy of Copresence," en, *Presence: Teleoperators and Virtual Environments*, vol. 12, no. 5, pp. 445–455, Oct. 2003, ISSN: 1054-7460. DOI: 10.1162/105474603322761261. [Online]. Available: https://direct.mit.edu/pvar/article/12/5/445-455/58924 (visited on 10/08/2024).

[127] R. Skarbez, F. P. Brooks Jr., and M. C. Whitton, "A Survey of Presence and Related Concepts," en, *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–39, Nov. 2018, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3134301. [Online]. Available: https://dl.acm.org/doi/10.1145/3134301 (visited on 10/09/2024).

[128] B. Loveridge, "Key Considerations for Duo Singing in Virtual Reality and Videoconferencing: An Exploratory Study with Bigscreen and Zoom," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–7, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704206. [Online]. Available: https://ieeexplore.ieee.org/document/10704206/ (visited on 10/09/2024).

[129] A. Carôt and C. Werner, "Network Music Performance - Problems, Approaches and Perspectives," in *Proceedings of the Music in the Global Village Conference*, Budapest, Hungary, Sep. 2007. [Online]. Available: http://www.carot.de/Docs/MITGV_AC_CW.pdf.

[130] J. Hacker, S. Schobel, and A. Carôt, "Let's Choir-laborate: Designing a Multisensory Technology that Supports Real-Time Interaction in Digital Environments," in *Proceedings of the 58th Hawaii International Conference on System Sciences*, Jan. 2025.

[131] C. Rinaldi and C. Centofanti, "The Musical Metaverse: Advancements and Applications in Networked Immersive Audio," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–7, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704154. [Online]. Available: https://ieeexplore.ieee.org/document/10704154/ (visited on 10/09/2024).

[132] P. Müller, R. Haefeli, J. Schütt, and M. Ziegler, "Telemersive Audio Systems for Online Jamming," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–15, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704096. [Online]. Available: https://ieeexplore.ieee.org/document/10704096/ (visited on 10/09/2024).

[133] D. Rossetti and C. C. Bomfim, "Live Electronics, Audiovisual Compositions, and Telematic Performance: Collaborations During the Pandemic," *Journal of Network Music and Arts*, vol. 3, no. 1, 2021.

[134] 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, ISBN: 9798350382549. DOI: 10.1109/IEEECONF59510.2023.10335395. [Online]. Available: https://ieeexplore.ieee.org/document/10335395/ (visited on 07/02/2024).

[135] S. Giacomelli, C. Centofanti, J. Santos, *et al.*, "Remote Immersive Audio Production: State of the Art Implementation, Challenges, and Improvements," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–10, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704192. [Online]. Available: https://ieeexplore.ieee.org/document/10704192/ (visited on 10/09/2024).

[136] R. Hupke, S. Preihs, and J. Peissig, "Immersive Room Extension Environment for Networked Music Performance," in *AES 153rd Convention*, New York, USA, Oct. 2022.

[137] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally Adapted Immersive Environments for Distributed Music Performances in Mixed Reality," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–10, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704217. [Online]. Available: https://ieeexplore.ieee.org/document/10704217/ (visited on 10/09/2024).

[138] ISO, *ISO 3740: Acoustics - Determination of sound power levels of noise sources - Guidelines for the use of basic standards*, Genva, Switzerland, 2019. [Online]. Available: https://www.iso.org/standard/45107.html.

[139] T. Rudzki, "Improvements in the Perceived Quality of Streaming and Binaural Rendering of Ambisonics," PhD Thesis, University of York, York, UK, May 2023. [Online]. Available: https://etheses.whiterose.ac.uk/33904/.

[140] J. Martin, D. Van Maercke, and J.-P. Vian, "Binaural simulation of concert halls: A new approach for the binaural reverberation process," en, *The Journal of the Acoustical Society of America*, vol. 94, no. 6, pp. 3255–3264, Dec. 1993, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.407232. [Online]. Available: https://pubs.aip.org/jasa/article/94/6/3255/854555/Binaural-simulation-of-concert-halls-A-new (visited on 10/24/2024).

[141] ISO, *ISO 3382: Measurement of room acoustic parameters*, 2022. [Online]. Available: https://www.iso.org/standard/40979.html.

[142] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," en, *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.1570439. [Online]. Available: https://pubs.aip.org/jasa/article/113/6/3233/548721/On-the-importance-of-early-reflections-for-speech (visited on 10/24/2024).

[143] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *Journal of the Audio Engineering Society*, vol. 60, no. 11, pp. 887–898, Nov. 2012.

[144] T. Rudzki, D. Murphy, and G. Kearney, "User Preference Evaluation of Direct-to-Reverberant Ratio of Virtual Ambisonic Listening Spaces," in *Proceedings of the AES 2023 International Conference on Spatial and Immersive Audio*, Huddersfield, UK: Audio Engineering Society, Aug. 2023.

[145] S. H. Nielsen, "Auditory Distance Perception in Different Rooms," *Journal of the Audio Engineering Society*, vol. 41, no. 10, pp. 755–770, Oct. 1993.

[146] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," en, *Nature*, vol. 397, no. 6719, pp. 517–520, Feb. 1999, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/17374. [Online]. Available: https://www.nature.com/articles/17374 (visited on 10/25/2024).

[147] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of Different Impulse Response Measurement Techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, Apr. 2002.

[148] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proceedings of the AES 122nd Convention*, Vienna, Austria: Audio Engineering Society, May 2007.

[149] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Proceedings of the AES 108th Convention*, Paris, France: Audio Engineering Society, Feb. 2000.

[150] M. Kleiner, B.-I. Dalenb¨ack, and P. Svensson, "Auralization - An Overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, Nov. 1993. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=6976.

[151] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," en, *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2746–2760, Apr. 2019, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.5096178. [Online]. Available: https://pubs.aip.org/jasa/article/145/4/2746/848382/A-round-robin-on-room-acoustical-simulation-and (visited on 10/25/2024).

[152] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, Melbourne, Australia, Aug. 2010.

[153] IEC, *IEC 61672-1: 2013 Sound level meters - Part 1: Specifications*, 2013.

[154] V. Välimäki, J. Parker, L. Savioja, J. O. Smith, and J. Abel, "More Than 50 Years of Artificial Reverberation," in *Proceedings of the AES 60th International Conference*, Leuven, Belgium: Audio Engineering Society, Feb. 2016.

[155] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty Years of Artificial Reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, Jul. 2012, ISSN: 1558-7916, 1558-7924. DOI: 10.1109/TASL.2012.2189567. [Online]. Available: http://ieeexplore.ieee.org/document/6161610/ (visited on 10/25/2024).

[156] L. Battisti, A. Farina, A. Bevilacqua, and A. Bevilacqua, "Implementation of non-equal-partition multi-channel convolver," en, *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 265, no. 6, pp. 1570–1581, Feb. 2023, ISSN: 0736-2935. DOI: 10.3397/IN_2022_0220. [Online]. Available: https://www.ingentaconnect.com/content/10.3397/IN_2022_0220 (visited on 08/08/2025).

[157] F. Wefers, *Partitioned convolution algorithms for real-time auralization* (Aachener Beiträge zur technischen Akustik Band 20), eng. Berlin: Logos Verlag Berlin GmbH, 2015, ISBN: 978-3-8325-3943-6.

[158] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *Proceedings of the Audio Engineering Society 90th Convention*, Paris, France: Audio Engineering Society, Feb. 1991. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=5663.

[159] B. Alary, A. Politis, and S. Bilbao, "Designing Directional Reverberators for Spatial Sound Reproduction," in *Proceedings of the AES 5th International Conference on Audio for Virtual and Augmented Reality*, Redmond, Washington, USA, Aug. 2024.

[160] L. McCormack and A. Politis, "SPARTA & COMPASS: Real-Time Implementations of Linear and Parametric Spatial Audio Reproduction and Processing Methods," in *Proceedings of the AES Conference on Immersive and Interactive Audio*, York, UK: Audio Engineering Society, Mar. 2019. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=20417.

[161] T. Carpentier, "Spat˜ : A comprehensive toolbox for sound spatialization in Max," *Ideas Sonicas, 2021, Electroacoustic Space - Reflections - Tools for its design*, vol. 13, no. 24, pp. 12–23, 2021.

[162] A. Oliveira, G. Campos, P. Dias, *et al.*, "Real–time dynamic image–source implementation for auralisation," in *Proceedings of DAFx-13*, Maynooth, Ireland, Sep. 2013.

[163] M. Dring and B. Wiggins, "WHAM: WEBCAM HEAD-TRACKED AMBISONICS," en, in *REPRODUCED SOUND 2020*, Virtual: Institute of Acoustics, Dec. 2020. DOI: 10.25144/13379. [Online]. Available: https://www.ioa.org.uk/catalogue/paper/wham-webcam-head-tracked-ambisonics (visited on 10/17/2024).

[164] L. Rayleigh, "XII. *On our perception of sound direction*," en, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, Feb. 1907, ISSN: 1941-5982, 1941-5990. DOI: 10.1080/14786440709463595. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14786440709463595 (visited on 02/08/2025).

[165] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency and Space," *Journal of the Audio Engineering Society*, vol. 49, no. 4, pp. 231–249, Apr. 2001. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=10196.

[166] A. W. Mills, "On the Minimum Audible Angle," en, *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, Apr. 1958, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.1909553. [Online]. Available: https://pubs.aip.org/jasa/article/30/4/237/738861/On-the-Minimum-Audible-Angle (visited on 10/28/2024).

[167] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," en, *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, Jul. 1977, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.381498. [Online]. Available: https://pubs.aip.org/jasa/article/62/1/157/751577/Model-for-the-interaural-time-differences-in-the (visited on 10/28/2024).

[168] H. Møller, M. F. Sorensen, D. Hammershoi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, May 1995. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=7949.

[169] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," en, *Applied Sciences*, vol. 8, no. 11, p. 2029, Oct. 2018, ISSN: 2076-3417. DOI: 10.3390/app8112029. [Online]. Available: https://www.mdpi.com/2076-3417/8/11/2029 (visited on 10/29/2024).

[170] M. Warnecke, S. Clapp, Z. Ben-Hur, D. L. Alon, S. V. A. Gari, and P. Clamia, "Sound Sphere 2: A High–resolution HRTF Database," in *Proceedings of the AES 5th International Conference on Audio for Virtual and Augmented Reality*, Redmond, Washington, USA: Audio Engineering Society, Aug. 2024. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22680.

[171] C. Armstrong, A. Chadwick, L. Thresh, D. Murphy, and G. Kearney, "Simultaneous HRTF Measurement of Multiple Source Configurations Utilizing Semi-Permanent Structural Mounts," in *Proceedings of the AES 143rd Convention*, New York, USA: Audio Engineering Society, Oct. 2017. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=19311.

[172] T. Rudzki, D. Murphy, and G. Kearney, "XR-based HRTF Measurement," in *Proceedings of the 2022 AES International Conference on Audio for Virtual and Augmented Reality*, Redmond, Washington, USA: Audio Engineering Society, Aug. 2022.

[173] K. Young and G. Kearney, "A High-Resolution Boundary Element Method Suitable Full Torso Mesh of KEMAR," *Journal of the Audio Engineering Society*, vol. 71, no. 7/8, pp. 492–501, Jul. 2023, ISSN: 15494950. DOI: 10.17743/jaes.2022.0085. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=22154 (visited on 12/05/2024).

[174] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions," in *Proceedings of the AES 16th International Conference*, Arktikum, Rovaniemi, Finland: Audio Engineering Society, Apr. 1999.

[175] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions," *Journal of the Audio Engineering Society*, vol. 69, no. 1/2, pp. 104–117, Feb. 2021, ISSN: 15494950. DOI: 10.17743/jaes.2020.0070. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=21019 (visited on 08/10/2025).

[176] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions," en, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, p. 5, Dec. 2019, ISSN: 1687-4722. DOI: 10.1186/s13636-019-0148-x. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-019-0148-x (visited on 10/28/2024).

[177] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3D Ambisonic Based Binaural Sound Reproduction System," in *Proceedings of the AES 24th International Conference*, Banff, Alberta, Canada: Audio Engineering Society, Jun. 2003. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=12314.

[178] R. Shukla, I. T. Radu, M. Sandler, and R. Stewart, "Real-time binaural rendering with virtual vector base amplitude panning," in *AES Conference on Immersive and Interactive Audio*, York, UK: Audio Engineering Society, Mar. 2019. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=20414.

[179] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," en, *Journal of the Audio Engineering Society*, vol. 65, no. 3, pp. 212–225, Mar. 2017, ISSN: 15494950. DOI: 10.17743/jaes.2016.0071. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18556 (visited on 12/06/2024).

[180] ITU, *ITU-R BS.2399-0: Methods for selecting and describing attributes and terms, in the preperation of subjective tests*, 2017. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-BS.2399-2017-PDF-E.pdf.

[181] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head-Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[182] J. Paterson and H. Lee, *3D Audio*, en, 1st ed. New York: Routledge, Jun. 2021, ISBN: 978-0-429-49121-4. DOI: 10.4324/9780429491214. [Online]. Available: https://www.taylorfrancis.com/books/9780429491214 (visited on 10/15/2024).

[183] F. Zotter, "Analysis and Synthesis of Sound-Radiation with Spherical Arrays," PhD Thesis, University of Music and Performing Arts, Austria, Sep. 2009. [Online]. Available: https://www.researchgate.net/publication/236024162_Analysis_and_Synthesis_of_Sound-Radiation_with_Spherical_Arrays.

[184] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "AmbiX - A suggested Ambisonic format," in *Proceedings of the 2011 Ambisonics Symposium*, Lexington, KY, USA, Jun. 2011.

[185] M. A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound," in *Proceedings of the AES 50th Convention*, London, UK, Mar. 1975. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=2466.

[186] E. Benjamin and T. Chen, "The Native B-Format Microphone," in *Proceedings of the AES 119th Convention*, New York, USA: Audio Engineering Society, Oct. 2005.

[187] A. Farina, P. Martignon, A. Capra, and S. Fontana, "Measuring Impulse Responses Containing Complete Spatial Information," in *Proceedings of the AES 22nd UK Conference*, Cambridge, UK: Audio Engineering Society, Apr. 2007.

[188] 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.

[189] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, L. McCormack, S. J. Schlecht, and P. Ville, "Perceptually informed interpolation and rendering of spatial room impulse responses for room transitions," in *Proceedings of the 24th International Congress on Acoustics*, Gyeongju, Korea: Acoustical Society of Korea, Oct. 2022.

[190] T. McKenzie and S. J. Schlecht, "Source position interpolation of spatial room impulse response," in *Proceedings of the AES 154th Convention*, Espoo, Helsinki, Finland: Audio Engineering Society, May 2023. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22044.

[191] B. Mróz, M. Kabacinski, T. Ciotucha, A. Ruminski, and T. Zernicki, "Production of six-degrees-of-freedom (6DoF) navigable audio using 30 Ambisonic microphones," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Bologna, Italy: IEEE, Sep. 2021, pp. 1–5, ISBN: 978-1-66540-998-8. DOI: 10.1109/I3DA48870.2021.9610970. [Online]. Available: https://ieeexplore.ieee.org/document/9610970/ (visited on 12/03/2024).

[192] M. A. Poletti, "A Unified Theory of Horizontal Holographic Soud Systems," *Journal of the Audio Engineering Society*, vol. 48, no. 12, pp. 1155–1182, Dec. 2000. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=12033.

[193] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, Oct. 2012.

[194] G. Kearney, E. Bates, F. Boland, and D. Furlong, "A Comparative Study of the Performance of Spatialization Techniques for a Distributed Audience in a Concert Hall Environment," in *AES 31st International Conference*, London, UK: Audio Engineering Society, Jun. 2007.

[195] P. Stitt, S. Bertet, and M. Van Walstijn, "Off-Centre Localisation Performance of Ambisonics and HOA For Large and Small Loudspeaker Array Radii," en, *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 937–944, Sep. 2014, ISSN: 16101928. DOI: 10.3813/AAA.918773. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=1610-1928&volume=100&issue=5&spage=937 (visited on 11/06/2024).

[196] M. Frank and S. Lutz, "Evaluation of Compressed Higher-Order Ambisonics at Off-Center Listening Positions," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany: IEEE, Sep. 2024, pp. 1–7, ISBN: 9798350366525. DOI: 10.1109/IS262782.2024.10704209. [Online]. Available: https://ieeexplore.ieee.org/document/10704209/ (visited on 10/09/2024).

[197] M. Frank and A. Sontacchi, "Case Study on Ambisonics for Multi-Venue and Multi-Target Concerts and Broadcasts," *Journal of the Audio Engineering Society*, vol. 65, no. 9, pp. 749–756, Sep. 2017, ISSN: 15494950. DOI: 10.17743/jaes.2017.0026. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19204 (visited on 04/27/2024).

[198] C. Schorkhuber, M. Zaunschirm, and R. Holdrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of DAGA 2018*, Munich, Germany, Mar. 2018.

[199] M. Kronlachner, "Ambisonic plug-in suite for production and performance usage," in *Proceedings Linux Audio Conference*, Graz, Austria, May 2013. [Online]. Available: https://www.matthiaskronlachner.com/wp-content/uploads/2013/05/ambisonics-lac2013_inet.pdf.

[200] M. Kronlachner, "Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics," in *Proceedings of the AES 136th Convention*, Berlin, Germany: Audio Engineering Society, Apr. 2014. [Online]. Available: https://www.matthiaskronlachner.com/wp-content/uploads/2013/01/kronlachner_aes_studentdesigncompetition_2014.pdf.

[201] J. Francombe, T. Brookes, and R. Mason, "Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences," *Journal of the Audio Engineering Society*, vol. 65, no. 3, pp. 198–211, Mar. 2017, ISSN: 15494950. DOI: 10.17743/jaes.2016.0070. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18555 (visited on 11/07/2024).

[202] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *Journal of the Audio Engineering Society*, vol. 65, no. 3, pp. 212–225, Mar. 2017, ISSN: 15494950. DOI: 10.17743/jaes.2016.0071. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18556 (visited on 11/07/2024).

[203] R. Gang, S. H. Shivaswamy, S. Roessner, A. Roa, D. Headlam, and M. Bocko, "Audio Latency Masking in Music Telepresence Using Artificial Reverberation," in *Proceedings of the AES 133rd Convention*, San Francisco, California, USA, Oct. 2012. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=16431.

[204] B. Redman and J. Impett, "The potential of videoconferencing and low-latency (LoLa) technology for instrumental music teaching," eng, *Music & Practice*, vol. 6, Jul. 2020, ISSN: 1893-9562. DOI: 10.32063/0610. [Online]. Available: https://doi.org/10.32063/0610 (visited on 10/03/2024).

[205] L. Comanducci, M. Buccoli, M. Zanoni, *et al.*, "Investigating Networked Music Performances in Pedagogical Scenarios for the InterMUSIC Project," en, in *2018 23rd Conference of Open Innovations Association (FRUCT)*, Bologna: IEEE, Nov. 2018, pp. 119–127, ISBN: 978-952-68653-6-2. DOI: 10.23919/FRUCT.2018.8587915. [Online]. Available: https://ieeexplore.ieee.org/document/8587915/ (visited on 10/03/2024).

[206] E. Hendrickx, M. Paquier, V. Koehl, and J. Palacino, "Ventriloquism effect with sound stimuli varying in both azimuth and elevation," en, *The Journal of the Acoustical Society of America*, vol. 138, no. 6, pp. 3686–3697, Dec. 2015, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.4937758. [Online]. Available: https://pubs.aip.org/jasa/article/138/6/3686/630731/Ventriloquism-effect-with-sound-stimuli-varying-in (visited on 12/24/2024).

[207] S. Stickland, R. Athauda, and N. Scott, "Design and Evaluation of a Scalable Real-Time Online Digital Audio Workstation Collaboration Framework," *Journal of the Audio Engineering Society*, vol. 69, no. 6, pp. 410–431, Jun. 2021, ISSN: 15494950. DOI: 10.17743/jaes.2021.0016. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21110 (visited on 01/19/2025).

[208] G. Kearney, H. Daffern, L. Thresh, H. Omodudu, C. Armstrong, and J. Brereton, "Design of an Interactive Virtual Reality System for Ensemble Singing," in *Proceedings of the Interactive Audio Systems Symposium*, York, UK, Sep. 2016.

[209] A. Southern, *Acoustic Parameters Toolbox*, Language: en, Apr. 2022. DOI: 10.5281/ZENODO.6505344. [Online]. Available: https://zenodo.org/record/6505344 (visited on 11/29/2024).

[210] O. Lartillot and P. Toiviainen, "A MATLAB Toolbox for Musical Feature Extraction from Audio," in *Proceedings of the 10th International Conference on Audio Effects*, Bordeaux, France: DAFx, Sep. 2007.

[211] C. Chafe, J.-P. Cáceres, and M. Gurevich, "Effect of Temporal Separation on Synchronization in Rhythmic Performance," en, *Perception*, vol. 39, no. 7, pp. 982–992, Jul. 2010, ISSN: 0301-0066, 1468-4233. DOI: 10.1068/p6465. [Online]. Available: https://journals.sagepub.com/doi/10.1068/p6465 (visited on 01/15/2025).

[212] A. Schmid, M. Ambros, J. Bogon, and R. Wimmer, "Measuring the just noticeable difference for audio latency," in *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures*, ser. AM '24, Milan, Italy: Association for Computing Machinery, 2024, pp. 325–331, ISBN: 9798400709685. DOI: 10.1145/3678299.3678331. [Online]. Available: https://doi.org/10.1145/3678299.3678331.

[213] L. Beranek, "Music and Acoustics," in *Concert Halls and Opera Houses:Music, Acoustics and Architechture*, New York: Springer-Verlag, 2004, pp. 1–18.

[214] S. V. A. Gari, "Investigations on the Influence of Acoustics on Live Music Performance using Virtual Acoustic Methods," PhD Thesis, Detmold University of Music, Germany, Sep. 2017. [Online]. Available: https://opus.hfm-detmold.de/frontdoor/deliver/index/docId/68/file/Dissertation_Amengual.pdf.

[215] B. Bacila and H. Lee, "Perceptual dimensions of listener envelopment (LEV) in a positional and directional-varying context," in *AES 2023 International Conference on Spatial and Immersive Audio*, Huddersfield, UK: Audio Engineering Society, Aug. 2023. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=22161.

[216] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound Externalization: A Review of Recent Research," en, *Trends in Hearing*, vol. 24, p. 2 331 216 520 948 390, Jan. 2020, ISSN: 2331-2165, 2331-2165. DOI: 10.1177/2331216520948390. [Online]. Available: https://journals.sagepub.com/doi/10.1177/2331216520948390 (visited on 02/03/2025).

[217] S. Uhrig, S. Möller, D. M. Behne, U. P. Svensson, and A. Perkis, "Testing a Quality of Experience (QoE) Model of Loudspeaker-Based Spatial Speech Reproduction," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, Athlone, Ireland: IEEE, May 2020, pp. 1–6, ISBN: 978-1-72815-965-2. DOI: 10.1109/QoMEX48832.2020.9123119. [Online]. Available: https://ieeexplore.ieee.org/document/9123119/ (visited on 02/04/2025).

[218] E. B. Union, *EBU: R128 Loudness normalisation and permitted maximum level of audio signals*, EBU Recommendations, Geneva, Aug. 2020. [Online]. Available: https://tech.ebu.ch/docs/r/r128.pdf.

[219] Y. G. Redman, J. D. Glasner, D. D'Orazio, and P. Bottalico, "Singing in different performance spaces: The effect of room acoustics on singers' perception," en, *The Journal of the Acoustical Society of America*, vol. 154, no. 4, pp. 2256–2264, Oct. 2023, ISSN: 0001-4966. DOI: 10.1121/10.0021331. [Online]. Available: https://pubs.aip.org/jasa/article/154/4/2256/2916171/Singing-in-different-performance-spaces-The-effect (visited on 04/29/2024).

[220] J. Brereton, "Singing in Space(s): Singing performance in real and virtual acoustic environments - Singer's evaluation, performance analysis and listener's perception.," PhD Thesis, University of York, Aug. 2014. [Online]. Available: https://etheses.whiterose.ac.uk/id/eprint/7877/1/JSBRERETON_PHD.pdf.

[221] H. Daffern, D. A. Camlin, H. Egermann, *et al.*, "Exploring the potential of virtual reality technology to investigate the health and well being benefits of group singing," en, *International Journal of Performance Arts and Digital Media*, vol. 15, no. 1, pp. 1–22, Jan. 2019, ISSN: 1479-4713, 2040-0934. DOI: 10.1080/14794713.2018.1558807. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14794713.2018.1558807 (visited on 02/13/2025).

[222] K. Ueno, K. Kato, and K. Kawai, "Effect of Room Acoustics on Musicians' Performance. Part I: Experimental Investigation with a Conceptual Model," en, *Acta Acustica united with Acustica*, vol. 96, no. 3, pp. 505–515, May 2010, ISSN: 1610-1928. DOI: 10.3813/AAA.918303. [Online]. Available: http://www.ingentaconnect.com/content/10.3813/AAA.918303 (visited on 04/29/2024).

[223] D. Cabrera, D. Lee, M. Yadav, and W. Martens, "Decay envelope manipulation of room impulse responses: Techniques for auralization and sonification," in *Australian Acoustical SocietyConference 2011, Acoustics 2011: Breaking New Ground*, Nov. 2011.

[224] S. D'Amario, H. Daffern, and F. Bailes, "A new method of onset and offset detection in ensemble singing," en, *Logopedics Phoniatrics Vocology*, vol. 44, no. 4, pp. 143–158, Oct. 2019, ISSN: 1401-5439, 1651-2022. DOI: 10.1080/14015439.2018.1452977. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14015439.2018.1452977 (visited on 10/06/2024).

[225] J. Schluter and S. Bock, "Improved musical onset detection with Convolutional Neural Networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy: IEEE, May 2014, pp. 6979–6983, ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854953. [Online]. Available: http://ieeexplore.ieee.org/document/6854953/ (visited on 02/12/2025).

[226] K. Milligan and N. Bailey, "A Review of Software for Note Onset Detection," in *ANIMUSIC 2015*, Tavira, Portugal, Dec. 2015. [Online]. Available: https://eprints.gla.ac.uk/124075/.

[227] D. Stefani, G. A. Giudici, and L. Turchet, "On the Importance of Temporally Precise Onset Annotations for Real-Time Music Information Retrieval: Findings from the AG-PT-set Dataset," en, in *Audio Mostly 2024 - Explorations in Sonic Cultures*, Milan Italy: ACM, Sep. 2024, pp. 270–284, ISBN: 9798400709685. DOI: 10.1145/3678299.3678325. [Online]. Available: https://dl.acm.org/doi/10.1145/3678299.3678325 (visited on 11/05/2024).

[228] S. Mullins, "Des voix du passé : The historical acoustics of Notre-Dame de Paris and choral polyphony," PhD Thesis, Sorbonne Université, Sep. 2024. [Online]. Available: https://hal.sorbonne-universite.fr/THESES-SU/tel-04879114v1.

[229] P. Cairns, *Marker Metadata-Based Onset Picking (MM-BOP)*, Jun. 2024. DOI: 10.5281/ZENODO.12581697. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.12581697 (visited on 02/13/2025).

[230] M. Blanca, R. Alarcón, J. Arnau, R. Bono, and R. Bendayan, "Non-normal data: Is ANOVA still a valid option?" *Psicothema*, vol. 4, no. 29, pp. 552–557, Nov. 2017, ISSN: 0214-9915, 1886-144X. DOI: 10.7334/psicothema2016.383. [Online]. Available: https://www.psicothema.com/pii?pii=4434 (visited on 02/18/2025).

[231] O. J. Dunn and V. Clark, *Applied statistics: analysis of variance and regression* (Wiley series in probability and mathematical statistics), eng. New York, NY: Wiley, 1974, ISBN: 978-0-471-22700-7.

[232] R. V. Hogg and J. Ledolter, *Engineering statistics* (Mathematics & statistics), Internat. ed. New York u.a: Macmillan u.a, 1989, ISBN: 9780029461327.

[233] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, en, 1st. New York, USA: Chapman and Hall/CRC, Jul. 2010, ISBN: 978-0-429-11188-4. DOI: 10.1201/9781439896129. [Online]. Available: https://www.taylorfrancis.com/books/9781439896129 (visited on 02/19/2025).

[234] F. Martellotta, "The just noticeable difference of center time and clarity index in large reverberant spaces," en, *The Journal of the Acoustical Society of America*, vol. 128, no. 2, pp. 654–663, Aug. 2010, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.3455837. [Online]. Available: https://pubs.aip.org/jasa/article/128/2/654/900764/The-just-noticeable-difference-of-center-time-and (visited on 02/21/2025).