VIIVA-NMP Audio System: *The design of a low latency and naturally interactive Ambisonic audio system for Immersive Network Music Performance*

Patrick James Cairns

MSc by Research

Electronic Engineering
University of York

February 2021

ABSTRACT

Network Music Performance (NMP) technology allows remote performers to play music together by sharing sound over the internet. Current developments in NMP practice and design have begun to explore the integration of Virtual Reality (VR) and other immersive technologies.

In contribution to this new context, this thesis presents an immersive audio NMP system for group vocal performance (audio only), named the Vocal Interaction in an Immersive Virtual Acoustic (VIIVA)-NMP system. This design was used to evaluate immersive audio NMP vocal performance, and explore the practical challenges and opportunities of immersive audio NMP systems.

A prototype implementation of the VIIVA-NMP audio system design was developed using open source resources including Jacktrip, Kronlachner VST, Open Air SIR, and the SADIE II Binaural Database.

In order to provide practical validation of the audio system design, and investigate the effect of acoustic environment and latency on performance synchrony and Perceived Immersion, the prototype implementation was deployed to 10 musicians across Europe. These musicians participated in testing by forming connections between their respective homes and providing remote duet vocal performance and questionnaire response.

Analysis of data collected in testing demonstrates that the prototype implementation can achieve full system latency of 30ms or less and allow for natural musical interactivity. Technical analysis demonstrates the requirements of achieving this, and provides estimation of the practical range within which this may be achieved. An effect of Room on perception of performance using the VIIVA-NMP audio system prototype was also identified, providing exciting direction for future study

As one of the first systems to provide immersive audio, latency of 30ms or less, and allow for natural musical interactivity, the VIIVA-NMP audio system provides an original contribution. The practical use-case testing and associated technical specification for Immersive NMP audio systems provides further contribution to the field.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my gratitude to XR Stories for funding this studentship. Thank you for making this project not only possible, but accessible.

I would also like to thank the staff and fellow researchers at the AudioLab. First and foremost Gavin Kearney and Helena Daffern who supervised this project. To both of you: your expansive knowledge, enthusiasm, support, encouragement and guidance has been greatly appreciated both academically and personally, as has your positive and holistic approach to leading research.

These recognitions are extended to Damian Murphy, who, as thesis advisor, has proven a consistent friendly face and excellent source of valuable feedback.

Indeed, to all AudioLab colleagues: Thank you, for your warm welcome to the AudioLab, and thank you, for making the workplace exciting, interesting and enjoyable (even when that workplace is Zoom calls). I learn something new every day, and am constantly inspired by you.

Last, but certainly not least, I would like to thank my family. Your support and encouragement means the world to me, and has made all of this possible.

DECLARATION

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

STATEMENT OF ETHICS

All testing detailed in this thesis, and the management of associated data, was approved by the University of York Physical Sciences Ethics Committee.

1. INTRODUCTION

A virtual space, or ‘virtual environment’, can be broadly defined as a synthetic space which is presented through interactive sensory display [1]. This can be something as simple as a telephone call, where two or more people are presented with a synthetic space to communicate in through interactive auditory display. This can also be something as complex as online Virtual Reality (VR) platforms, where thousands of people gather and interact in virtual worlds, which are presented through interactive visual, auditory, and haptic display.

Within the diversity of virtual worlds exists virtual performance spaces: virtual environments designed for musical performance experiences to take place in. These virtual environments provide virtual musical performance experiences, which may simulate live performance, or deliver new novel performance experiences. These virtual performance environments are also often Virtual Acoustic Environments [2], where the presented interactive auditory display includes simulation of an acoustic space.

Virtual performance experiences are useful in providing access to experiences which are otherwise difficult or impossible to access physically, or in providing access to novel new virtual performance experiences.

Virtual and Augmented Reality (VR/AR) technology is capable of delivering such an experience [3], [4], which can be provided to individuals, who perform with locally rendered virtual accompaniment, or to groups of performers by sharing audio over Local Area Networks (LANs). Such experiences may be extended to groups of remote performers by sharing audio over Wide Area Networks (WANs) [4], a discipline described as Network Music Performance (NMP) [5], [6].

NMP, at the base level, is an enabling tool which aims to allow geographically distant musicians to engage in synchronous musical interactions together, a functionality lends itself to a range of practical applications.

In education platforms NMP allows access to musical tuition where a performer may be unable to access a suitable tutor for the instrumental or musical training required locally [7]. Access is also facilitated to group performance where no suitable musical ensemble is accessible locally. In this manner NMP offers a utility to improve the level of educational service offered to musicians with potential for great impact particularly when services in remote locations are considered [8].

Potential applications in remote healthcare have been identified, where NMP extension of virtual performance experiences may allow for remote access to VR technology which can provide virtual exposure therapy to access the health and wellbeing benefits of group singing [9], [10], [11] or to help manage performance anxiety [12].

Further service provision is apparent in NMP where telepresence technology allows access to virtual acoustic performance spaces. Extension of VR performance experiences [3], [13], [14] to the NMP context offers an avenue of great potential. Such digitisation of the performance space provides a level of convenience in terms of practical access and travel considerations well familiar to professional musicians. Indeed the use of virtual acoustic performance spaces allows for access to acoustic simulations of performance environments which are inaccessible to most musicians or even performance environments which no longer exist [15], [16].

Beyond practical functionality NMP offers a revolutionary avenue for many conceptual aspects of musical performance. The internet as a medium for the arts has allowed the globalisation and sharing of musical cultures across many mediums. NMP provides an avenue for such global social communications through the medium of musical performance.

The NMP medium presents a new environment for concert performance with global online audiences [17] using live streaming platforms or online VR and gaming applications. Such technology offers the potential to take advantage of the faculties of human-computer interfaces in order to access innovative workflows [18] such as interactive audience participation [19], virtual reality instruments [20], gestural haptic control interfaces [21], [22] and novel VR/AR enhancements of musical composition and performance experiences [23].

NMP also has an impact on a personal level. During the Covid-19 pandemic, where the gathering of groups of musicians has been predominantly inaccessible to all, NMP provides the ability to interact and engage with others through music, bring communities together, deliver joy and entertainment, and support the flourishing of the Arts.

Though current NMP state-of-the-art offers many opportunities, a number of challenges can also be identified. Indeed many remote performers have expressed that the NMP solutions which are accessible to the typical home user offer an experience where the presentation of the performance and rhythmic interactivity between performers provides many discrepancies in comparison to live musical interactions. It is still common to hear from remote performers paraphrasing the sentence: "we have tried (various existing NMP solutions) but it takes a bit of work to keep a stable tempo, and it just does not feel like you are in the room with each other". The integration of VR performance technology to the NMP context offers not only an opportunity to provide a more immersive performance experience, but a design which allows access to the typical home user may indeed be the key to moving NMP from bespoke installation applications to mainstream use.

This thesis explores the potential for integration of virtual performance technology to the network performance context (defined here as Immersive NMP, INMP) through the investigation of the design, practical implementation and analysis of an audio system for Immersive NMP Vocal Performance.

1.1 Network Music Performance and Immersive Applications

The primary concern in Network Music Performance is the latency induced by the transport of audio data between network nodes [24] (*Figure 1.1.1*). This latency has been identified as having a significant effect on the ability of remote musicians to perform synchronously [25], [26], [27], [28]. As such NMP systems have traditionally considered enabling musicians to achieve synchronous performance the target of design [6].

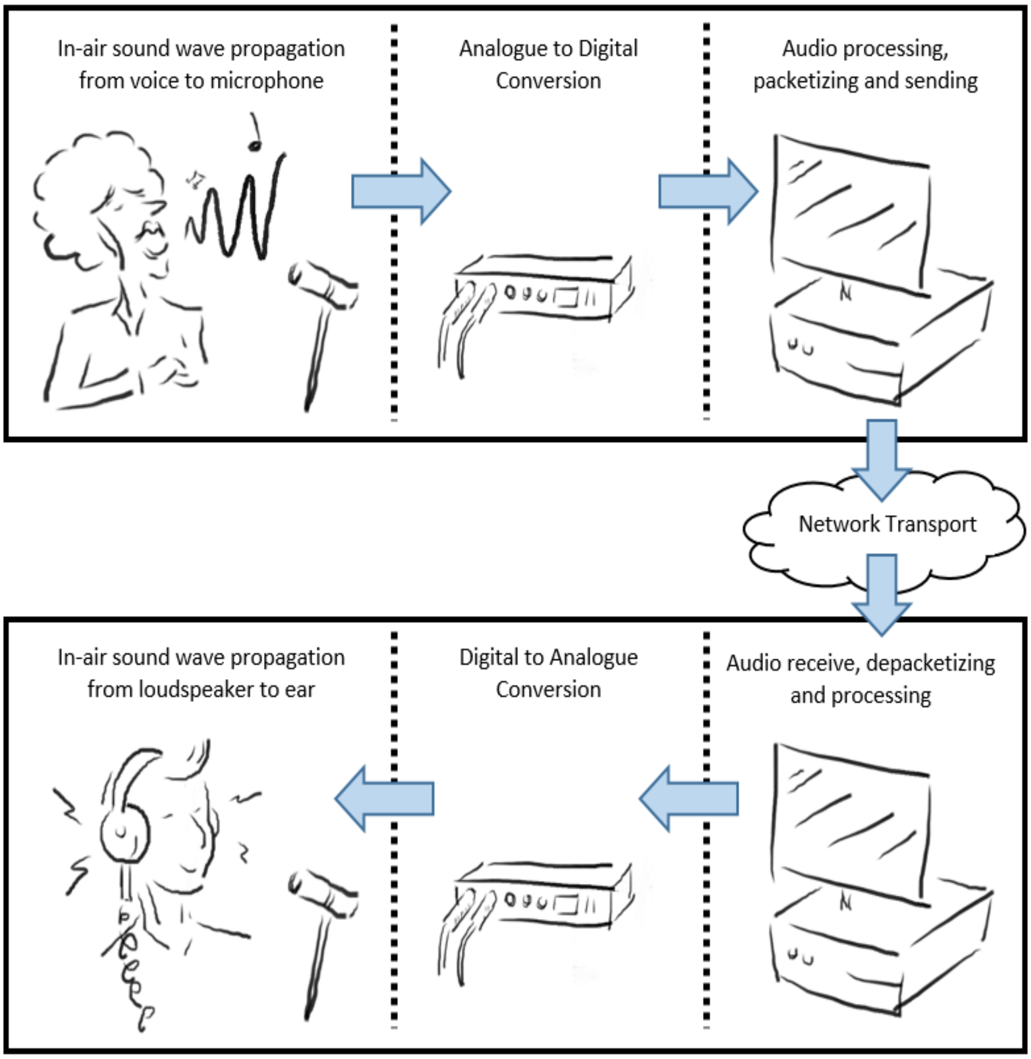

In-air sound wave propagation from voice to microphone
Analogue to Digital Conversion
Audio processing, packetizing and sending
Network Transport
In-air sound wave propagation from loudspeaker to ear
Digital to Analogue Conversion
Audio receive, depacketizing and processing

Figure 1.1.1 Breakdown of significant NMP latency contributors, adapted from [29].

In order to manage the effect of latency NMP designs follow one of two approaches, defined by Carot et al [6]:

- **Realistic Jam Approach**

This approach prioritises real-time interaction between performers, where the primary concern is minimising latency to a range where performance synchrony is not affected.

- **Latency Accepting Approach**

  This approach considers real-time interaction between performers impossible and facilitates synchronous performance through latency-coping methods.

With respect to immersive applications such as virtual performance systems it can be noted that latency is a key concern when delivering audio [30], where noticeable latencies and the additional cognitive load of latency-coping strategies detract from the immersive quality of the virtual performance experience. As such Realistic Jam presents an appropriate approach for immersive applications.

The target latency threshold for Realistic Jam NMP systems is defined by the Ensemble Performance Threshold (EPT) [27], [28], detailing "*the level of delay at which effective real-time musical collaboration shifts from possible to impossible*" [28]. Performance synchrony metrics provide a point of comparison, where equivalent synchrony to the live musical performance implies a level of musical interactivity between remote performers equivalent to live conditions. Empirical EPT research in the form of hand-clapping [27], [31], [32] and instrumental [33], [34], [35], [36] performance experiments largely agree on an EPT upper limit of 25-30ms, above which tempo deceleration and eventually collapse of natural performance synchrony occurs [29] (*Figure 1.1.2*).

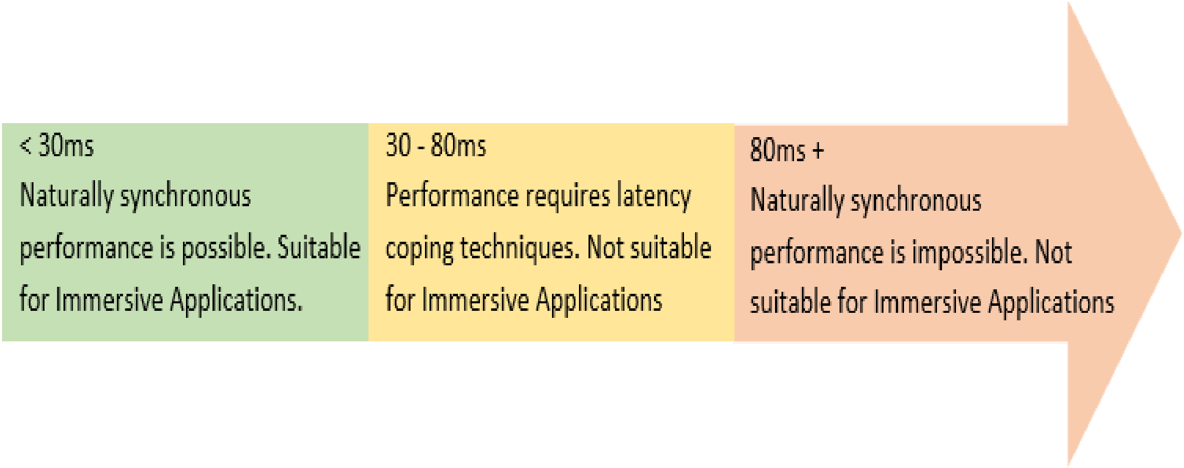

< 30ms
Naturally synchronous performance is possible. Suitable for Immersive Applications.
30 - 80ms
Performance requires latency coping techniques. Not suitable for Immersive Applications
80ms +
Naturally synchronous performance is impossible. Not suitable for Immersive Applications

Figure 1.1.2 NMP latency thresholds with respect to suitability for Immersive Applications.

This thesis explores the design and application of extending VR performance experiences to the NMP, where suitable levels of performer immersion, and thus implementation within these latency constraints defined by the EPT, is required. Considering this adaptation of the immersive virtual performance experience to an NMP deployment, several key areas of interest can be identified.

1.1.1 Immersive Application Driven Frameworks

The 'state of the art' approach to NMP from an immersive applications perspective currently follows a telepresence-motivated design which utilises streaming of audio and video content. Such video-streaming systems inevitably operate at latencies greater than the EPT [29], and have designed accordingly to present practical methods of managing this large latency, and suitably affect Presence [37] (the sensation of 'being in' a virtual environment) in the performance experience.

Virtual performance technology, however, aims to simulate the live performance experience with Immersion [38] rather than Presence as the design target, thus requiring 'Realistic Jam' approaches for extension to NMP deployment. It can be considered that telepresence-NMP frameworks, though extremely useful, are not compatible with the requirements of VR/AR performance experiences [30]. It is therefore prudent to specify Immersive NMP as discrete from telepresence-NMP, and present a new framework and blueprint for the design of Immersive NMP systems which are compatible with integration of immersive VR/AR virtual performance technology.

Qualitative assessment of NMP systems has ranged from simple subjective rating to comprehensive Presence-motivated analysis [4], [36], [39], [40], [37], [41]. In the same manner the development of a comprehensive immersion-motivated analysis is required to deliver validation of proposed Immersive NMP systems.

1.1.2 EPT and Virtual Reverberant Performance Spaces

Though the upper limit of the EPT is largely agreed upon there are some dependencies which require further investigation in the context of virtual vocal performance. Firstly observable tempo acceleration, dubbed the 'Chafe Effect', is apparent for network latencies below 11.5ms [27], [42].

It is postulated that this may be a result of negative mean asynchrony [31], [43], [44], a phenomenon where nearby musicians tend to count ahead of the guiding tempo of performance. It is also considered that this may be a result of anechoic test conditions [27], giving reason to believe that acoustic properties may influence performance synchrony in NMP systems.. Indeed, Farner et al [32] indicate that greater performance synchrony is achieved under virtual reverberant conditions than under anechoic conditions, and Schuett [28] also observes that stable tempo was achievable at higher levels of temporal separation between musicians in live reverberant conditions than with digital delay in anechoic test conditions.

The effect of acoustic simulation of reverberant environments on EPT is, however, wholly unquantified, and requires further investigation.

1.1.3 EPT and Instrumentation

Both Barlette [33] and Rottondi [34] observe a significant effect on the EPT limits for acceptably synchronous performance as instrumentation is varied. This can be identified with respect to instrument characteristics such as short attack slope and spectral flatness. This conforms to research on the perceptual onset of musical events. Within each musical note it can be recognised that there is a degree of temporal variation as to where the onset is located [25], [48]. This is well recognised as dependant on the amplitude and spectral enveloped of the musical event [48], [49], [50]. As the sequential perception of onsets is the indicator of tempo in musical performance it follows that this in turn will affect management of perceived onsets and thus the EPT in an NMP context. Voice is an instrument which is capable of exhibiting the characteristics of a range of instrumentation types. As such it is can be considered that a more robust understanding of instrument-specific EPT may be required for effective implementation of an Immersive NMP vocal performance system.

1.1.4 Practical Deployment Using Typical Home Internet Connections

NMP research which operates within the EPT upper limit has been historically restricted to high speed academic networks [6], [51] (JANET, GEANT, Internet 2) due to the poor bandwidths and lack of Quality of Service agreement offered by commercial Internet Service Providers. The general difficulty of achieving latencies below the EPT and unviability for the typical home user somewhat explains the lack of integration between NMP and immersive applications. Recent advances in typical home internet connection bandwidth and quality [52], however, have now reached the point where the deployment of NMP streaming solutions which achieve latencies below the EPT are a viability. As this is a relatively recent development it can be understood that little information on practical implementation of Realistic Jam NMP systems from typical home internet connections exists in academic research. Considering that typical home internet connections still differ significantly from academic networks a practical deployment and analysis of Immersive NMP systems is required to validate the capacity of the average home connection in supporting such technology.

1.1.5 Natural Musical Interactivity

It can be acknowledged that the EPT divides networked musical interactions between remote performers into two discrete conditions.

When latency does not exceed the EPT, the musical interactions conform to the expectations of live performance, and this results in synchronous performance which conforms to expectations again derived from live performance. When latency is above the EPT the inverse is true: musical interactions and achieved performance does not conform to the expectations derived from live performance.

It is recognised that what is being described by the EPT is the degree to which musical interactions and synchronous performance is ‘natural’.

Naturalness [53], [54], [55] is appropriately defined in an immersive audio context as the degree to which the stimulus under evaluation conforms to “an internal reference that relates to memories of characteristics of natural environments” [53],

Indeed this is coherent with evaluation of performance in NMP literature, which typically provides perceptual rating of performance with comparison to live conditions, or objective comparison of performance synchrony (as will be discussed in detail in Ch. 2 and 3).

For this reason the terminology is introduced ‘natural musical interactivity’, which describes musical interactivity (and associated synchronous performance) expectations of live performance.

It is noted that the naturalness of musical interactivity is discrete (though also a component of) the Naturalness of the complete immersive performance experience provided by an INMP audio system (which will again be discussed in more detail later in this thesis).

1.2 Project Objectives

The objective of this project is to develop and deploy an audio system for Immersive NMP vocal performance in order to study the challenges and opportunities presented in the design and practical implementation of such a system.

The hypothesis of this thesis is therefore:

It will be possible to design and implement an audio system which is suitable for Immersive NMP vocal performance.

Where the requirements of Immersive NMP vocal performance are considered:

- Facilitating natural musical interactivity between performers.
- Providing immersive audio which is suitable for VR musical performance applications.
- Being accessible to users operating from typical home internet access.

In order to achieve this objective several sub-objectives were developed. These sub-objectives identify key areas which support investigation of Immersive NMP systems, and address specific tasks relevant to the primary objective.

- To conduct pilot study of performance synchrony and EPT for vocal performance under the influence of varying virtual acoustic reverberant performance spaces. This will ensure that EPT provides an accurate estimation of latency thresholds for Immersive NMP vocal performance, and investigate the potential of reverberant environments to affect performance synchrony (and thereby EPT definition).

- To provide an immersion-driven framework for NMP audio system design, which will allow for validation of the Immersive NMP audio system presented in this thesis as suitably immersive.
- To conduct a pilot study in the practical deployment of audio systems for Immersive NMP and identify the challenges and opportunities presented using 'real-world' internet connections. This will allow for validation of the proposed system as a practical solution which is accessible to the typical home user.

1.3 Project Motivations

Previous research at the AudioLab, University of York, has investigated the design and application of VR vocal ensemble performance experiences for individuals and groups. This has culminated in the development of the Vocal Interaction in an Immersive Virtual Acoustic (VIIVA) system [3].

The VIIVA system provides a VR simulation of taking part in a pre-recorded group vocal performance. An individual may use this system, and be presented with 360 video visual display of the captured vocal performance from the perspective of one of the performers. The user is also provided real-time spatial audio auralisation of the user's voice in the virtual acoustic performance space presented by the visual display. Three degrees of freedom are provided in visual and auditory playback, to allow for an interactive VR experience.

Beyond offering VR ensemble singing experience, the VIIVA system acts as an effective research tool, having been used to identify the potential of VR exposure experiences to access the Health and Wellbeing benefits of group singing [9]. The VIIVA system also acts as an access enabling tool for the *Sing From Your Seat* project [56], [57] which provides a group singing experience to remote singers with a focus on the mobility impaired.

The project 'Lag free audio communication in a multi-user virtual reality environment' as detailed in this thesis is motivated primarily by the extension of the VIIVA system to remote users by the creation of a VIIVA-NMP adaptation.

Previous work at the AudioLab has proposed that any such extension will require consideration rendered environments and performer avatars [9], in order to overcome the challenges of video streaming (which will be discussed in Ch. 2).

Though this thesis is concerned with audio system design, the work described maintains consideration that nay successful audio system design should be suitable for implementation alongside VR visual display with performer avatars in future work.

It is also acknowledged that in an NMP context the use of performer avatars presents an unexplored solution, and the potential effect of such visual contact is an unknown which could prove a control issue in evaluation of the immersive audio NMP vocal performance using the VIIVA-NMP design. Indeed in live performance visual contact is well understood as having an impact on performance [58]. For this reason it was decided that evaluation of vocal performance using the VIIVA-NMP system should be conducted as an 'audio-only' experience, in order to control evaluation, and provide a baseline for future work which includes VR visual display, as well as an original audio-only evaluation.

This undertaking presents a relatively unexplored avenue of research which will provide a pilot study for further academic research and a blueprint for system design in the field of Immersive NMP.

1.4 Thesis Structure

Chapter 1: This chapter provides an Introduction to the topic of Immersive NMP and provides a summary of the design of the project discussed in this thesis.

Chapter 2: This chapter considers the design of an audio system for Immersive NMP vocal performance. A thorough presentation and discussion is provided regarding audio streaming and audio rendering design and technical specification for Immersive NMP. This is followed by the presentation and prototype implementation of the VIIVA-NMP audio system solution.

Chapter 3: This chapter considers the analysis of Immersive NMP systems. A review of subject and objective analysis methodology for Immersive NMP systems is presented. This is followed by the presentation of the testing protocol developed for user testing with the VIIVA-NMP prototype, and presentation of the developed analysis framework and tools. This includes presentation of the Onset Detection algorithm developed for synchrony analysis, which is based on the TIMEX [59] and detailing of the deployment of the VIIVA-NMP audio system prototype.

Chapter 4: This chapter presents the results from this testing protocol including synchrony analysis and Immersive Performance Experience questionnaire analysis for VIIVA-NMP system application from typical home networks under the influence of varying latency and virtual acoustic environment.

Chapter 5: This chapter provides a summary of the work undertaken in the project 'Lag free audio communication in a multi-user virtual reality environment' and draws conclusions from the results from testing (discussed in chapter 5) with relevance to the hypothesis presented in Chapter 1. A discussion of future work in the field of Immersive NMP is then presented.

1.5 Original Contributions

The amalgamation of audio system design for NMP and immersive performance technologies in an Immersive NMP vocal performance context presented in this project provides several novel contributions:

- The introduction of the concept of Immersive Network Music Performance, a presentation of the relevant design criteria, and a blueprint for the design of systems for Immersive Network Vocal Performance.
- The investigation of the influence of reverberant environment on latency perception and management in a vocal performance context.
- The presentation of an immersion-driven framework for Immersive NMP audio system analysis and an investigation of the effect of latency and reverberant environment on the immersive quality of the Immersive NMP aural experience.

1.6 Related Publications

- 'Immersive Network Music Performance: Design and Practical Deployment of a system for Immersive Vocal Performance' Patrick Cairns, Helena Daffern, Gavin Kearney, Audio Engineering Society 149th Convention, October 2020.

2. DESIGN OF AN AUDIO SYSTEM FOR IMMERSIVE NMP VOCAL PERFORMANCE

If one considers the context of NMP for immersive applications such as virtual performance it is first necessary to define what constitutes an immersive application. The exact definition of immersion has been the topic of some debate [60], [38]. A definition provided by Agrawal et al [38] is given as:

> ***"Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world."***

Rumsey surmises Agrawal et al's definition of Immersion as a ***"cognitive phenomenon"*** [60]**.** This definition provides useful clarification in that it considers various aspects of the experience of immersion.

Definitions of System Immersion describes immersion with respect to the ability of technology to provide "***illusion of reality to the senses of the human participant"*** [61]**.** Here the focus is on only the sensory input, without considering the effect on the immersive experience of the cognition of these stimuli and perceptual response of the individual.

Perceived Immersion, conversely, is defined as the "***"impression of being submerged into, enveloped or surrounded by the (virtual) environment"*** [55]**.** As a measurement of 'impression', Perceived Immersion focuses only on an individual's *awareness* of their own perception.

Psychological Immersion, characterised by an individual being '***involved, absorbed, engaged and engrossed"*** [62] in an experience provides a definition which focuses on the actual psychological response, rather than simply awareness of perception.

Agrawal et al's definition of Immersion as a cognitive phenomenon shows recognition of a more comprehensive range of immersive qualities, considering the stimuli provided, the psychological state of the individual (or Immersive Tendency [38]), the actual response to cognition of this stimuli, and the perceptual response to cognition of this stimuli. In the context of this project this definition is particularly useful as it conforms to VR definitions which express absolute immersion as *"**tricking one's brain into believing what they are sensing"*** [66], or Mixed Reality definitions which suggest an absolutely immersive stimuli as one which cannot be differentiated from real stimuli [67].

In order to provide evaluation of immersion, it is necessary to have a means by which immersion can be measured.

Definitions such as System Immersion argue that immersion can be measured in terms of the technical functionality of an immersive system. To a degree this can be accepted. For example, auditory immersion is associated with an explicit set of system functionalities [30], such as providing interactive 3-dimensional auditory display. This method of measuring immersion, however, can be acknowledged as providing an 'incomplete picture', as the definition does not consider aspects of Perceived Immersion or the cognition of stimuli presented by auditory display.

Robust evaluation of Immersion, rather than attempting to rate the experience in terms of technical system components, aims to model the complete immersive experience [38], [55], [61], [63], [68], including perceptual and personal aspects of the experience. Modelling the immersive experience is achieved by defining qualities of immersion. Qualities of immersion describe discrete, clearly defined aspects of the experience, and are used as 'building blocks' which can be used to construct a complete model of the immersive experience.

Individual immersive qualities may be sorted into groups, in order to define 'high level' structures within the immersive experience model. Robust models of the immersive experience typically divide the immersive experience into Plausibility, Interactivity, and Interestingness factors, each defining a range of associated immersive qualities [68].

Plausibility [64], [68] describes the degree to which an experience satisfies expectations [68]. This can be noted as discrete from concepts such as Naturalness [53] [55] [69] or Realism [63], which define reality as the reference for comparison. Plausibility instead also considers expectations that may not be derived from real-world reference. Plausibility is associated with physical, personal [68] and social presence [70] in virtual environments, where Presence may be defined as the sensation of being in a virtual environment [63].

Interactivity [68] details the ability of an experience to allow an individual to exert control [63], and the ability of the experience to respond to actions of an individual. This is again associated with Presence factors, but here more specifically with Social and Self Presence [68], and associated qualities such as Embodiment [21]. Interactivity also encompasses how the individual can physically or virtually involve [63] themselves with control and response in an experience.

Interestingness relates to narrative elements of the experience [68]. Discrete from Interactivity, Interestingness instead encompasses the narrative interaction with an immersive experience. With respect to narrative, Interestingness describes the ability of the experience and content to affect Involvement [63] and Attention [38] in narrative elements and the wider experience.

It can be recognised that the high level qualities Plausibility, Interactivity and Involvement, and the associated discrete immersive qualities, are all intertwined to a degree [68], and are also affected by subjective factors [68] native to the individual (as described by Immersive Tendency [38]). For example

Plausibility defines sensorimotor engagement factors which affect Presence in the experience, Interactivity defines task-oriented and motor engagement factors. If there is no narrative involvement, defined by Interestingness, however, an individual may not give any Attention to the aspects of the experience which provide sensory, motor and task-oriented engagement, thus impacting Plausibility and Interactivity. The degree of narrative involvement in the experience, in turn, is subject to the personal preferences, perception, awareness, willingness to engage, and internal reference of the individual.

In NMP literature rating of performance experiences is achieved in most cases through comparison to the live performance [29], either in objective measurement of synchrony or subjective rating (though Presence-motivated evaluations exist in high-latency contexts [37]). Indeed it can be acknowledged that it is reasonable to expect a musician to have a strong internal reference describing live performance, which they may use for comparison with virtual performance experiences. For this reason Naturalness can be identified as a common thread in the evaluation of INMP experiences.

As previously discussed the EPT essentially provides definition of a Naturalness boundary in terms of latency. Above the EPT (where performance is not natural) it can be identified that the high delay creates an Interface Awareness [37] issue which affects Engagement [65] in the performance experience: The effect on Attention may change narrative engagement (Interestingness). The effect on Naturalness and Realism may change the sensory engagement (Plausibility). The effect on the response of the system to control from the user may change motor and task-oriented engagement (Interactivity).

In immersive audio the affectation of auditory immersion is well understood (*Table 2.0.1*) and immersive audio design frameworks for the delivery of aural experiences which are suitable for immersive applications are well established [30]. System design components such as spatial audio [73], [74] delivery, acoustic simulation [75], [76], [77], [78] and interactive acoustics with 3/6 Degrees of Freedom (3DoF/6DoF) are well understood, commonplace, and deliver an immersive aural experience which is suitable for VR/AR applications [30], [66].

Immersive virtual musical performance technology has seen a high degree of study [79], [80], [13], [14], [20], [81], [3], [82], [10], [9], [83], [4] and the technical requirements of audio rendering for the immersive performance experience can be readily identified. A blueprint for immersive virtual vocal performance can be found in the VIIVA system [3] developed at the Audiolab, University of York.

Though the immersive musical performance experience is well documented, NMP frameworks have seen somewhat incomplete amalgamation of immersive audio design principles.

Existing NMP technology (*Table 2.0.2*), a good overview of which is presented by Rottondi et al [29], is historically concerned with audio streaming. This presumably avoids the pigeon-holing of audio streaming solutions, allowing for audio rendering to be implemented on a bespoke basis. This design focus has resulted in somewhat sparse research regarding the union of NMP and Immersive Audio in a low-latency context which is suitable for immersive applications such as VR/AR virtual performance technologies.

Factor of Auditory Immersion	**Example**
Audio Rendering	• Spatial audio [30] • Acoustic simulation [30] [37] [60] [55] [84] • Localisation [54] • Apparent source width [53] [54]
Interactivity	• Head-tracking [30] • 3/6 Degrees of Freedom (3DoF/6DoF) [30] • Latency [30]
Personalisation	• Individualised spatial audio rendering [30] • Headphone correction filters [30]
Perceptual Quality	• Presence [37] [63] • Embodiment [21] • Naturalness [53] [54] [69] • Clarity [69] • Envelopment [53] • Externalisation and source separation factors [53] [54] [69]

Table 2.0.1 Factors of auditory immersion in immersive audio applications.

Farner et al [32] provide a pilot investigation of spatial audio and acoustic simulation in an NMP context through investigation of the use of static Binaural Room Impulse Responses (BRIRs). Tempo was compared, and reportedly more stable in virtual acoustic conditions than with dry audio, though no conclusions could be confidently made due to lack of control on loudness. Novel interactive acoustic rendering investigation is presented by Chafe et al through the use of distributed Feedback Delay Network (FDN) reverberators [85], [86] to provide ‘internet rooms’ utilising an echo-based system which incorporates the network delay into the FDN design [87], [88], [89]. Though an interesting avenue of research in acoustic simulation for NMP, such systems again do not address criteria of immersive applications such as spatial audio delivery and realistic interactivity between performer and simulated acoustic environment.

Gurevich et al [39] approach a system appropriate for immersive applications in the presentation of a design of a networked system for interactive spatial audio delivery using a 3-D loudspeaker configuration in order to achieve an interactive aural environment which satisfies many requirements of immersive audio. Though a suitably immersive networked audio system is presented here, it can be identified that such a system is not practical for typical home users, and that no acoustic simulation, latency measurements, or musical performance case study is presented for this system, which functions more as an immersive audio installation than an NMP tool.

The experience of NMP performance has been investigated in a range of empirical studies. Chew [36], Farner [32], Carot [35], and Rottondi [34] all implement a form of self-reported subjective rating from performers, indicating the perceived performance conditions and tolerability of network latency. Barlette [33] extends the subjective rating to the perceived musicality of the performance, and Driessen [90] retrieves direct subjective rating of perceived performance synchrony. Olmos [91] provides simple subjective ratings of the emotional connection between participants in performance experiments.

Such research into the subjective quality of NMP experiences has led to the development of system frameworks which address the immersive qualities of the NMP using telepresence-motivated designs [71], [72]. Projects such as Distributed Immersive Performance [36], [92], Soundjack [93], Musinet [94], [95] Diamouses [96] LOLA [41], and Intermusic [37] have all presented solutions to telepresence NMP (*Figure 2.0.1)* where video streaming is utilised to enable gestural communication between remote performers and affect Embodiment and Presence in the performance experience [21]. Though some investigation into the use of immersive audio methods in the telepresence design has been conducted [97], [94] it is recognised that immersive audio rendering is generally considered difficult to implement in telepresence NMP designs due to the strict latency limits of networked interactive performance experiences relative to the existing throughput latency presented by many immersive audio rendering methods [37]. Indeed practical implementation of telepresence NMP designs will inevitably incur one-way latencies above 60ms [29] and further require the use of high-speed academic networks in order to effectively stream audio and video for groups of performers [41], [29]. As in practice latency is well above the 30ms EPT [29], telepresence NMP systems recognise that the performance experience is different from natural live musical interactions, and have presented many practical Latency Accepting approaches [6] to achieving synchronous performance using latency-managing techniques. Telepresence NMP systems also recognise that as the performance experience is significantly cognitively different from the live performance experience [98], it is difficult to design systems which aim to affect immersion by accurately simulating the live performance experience. Telepresence systems therefore target Presence and Embodiment in networked musical interactions rather than Immersion.

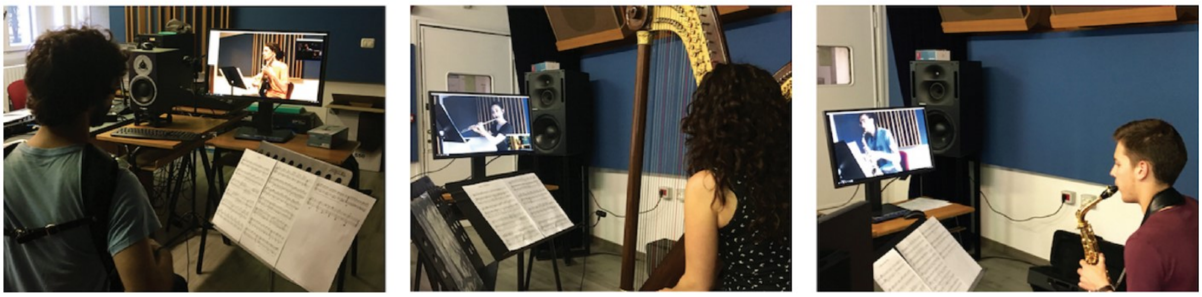


Figure 2.0.1 Remote musicians engaging in musical performance using a tele-presence NMP system. Auditory display is presented over loudspeakers, and visual contact is made available through tele-conferencing screen display [37].

Telepresence systems provide a practical solution as long-distance NMP will always incur large network latencies with current technology. If the extension of immersive virtual performance technologies to the NMP context is considered, however, it is clear that development of a new branch of design frameworks for systems which provide appropriate latency and audio delivery for networked immersive virtual performance applications is required. In this thesis this new design framework is defined as Immersive NMP.

Cutting-edge NMP research has begun investigation into proposed video rendering solutions for Immersive NMP which circumvent the latency issues of video streaming by presenting performer avatars which may be controlled using haptic technology [22] to provide gestural interaction and embodiment in the Immersive experience [21]. Research by Yoon et al [70] demonstrates that in collaborative virtual environments even simplified avatar representations of VR system users can affect suitable presence in group social interactions. The application of virtually rendered musicians [79] in the Immersive NMP system [4] has seen some pilot study using LAN systems [10] using dry mono audio to provide a baseline for future study. The missing component of such Immersive NMP systems can be identified as a practical framework for WAN audio streaming and rendering solutions which are compatible with existing VR/AR technology used for virtual performance applications.

A range of low latency audio streaming NMP solutions such as Soundjack [99] and Jacktrip [100] exist which are capable of achieving network latencies between performers which is below the 30ms EPT threshold associated with an immersive level of interactivity between musicians. Considering this, it is easy to assume that the solution to providing an audio system for Immersive NMP is simply the connection of iterations of existing virtual performance technology using these low latency audio streaming tools.

It can, however, be identified that immersive audio rendering solutions typically consider an audio rendering latency of 20ms as the target for ‘real-time’ application, and indeed some rendering methods rely on the use of large windows which utilise this full 20ms latency overhead to deliver high quality immersive audio [101], [102], [103]. With the 30ms full system latency target for Immersive NMP in mind it is clear that such audio rendering solutions only typically consider an allowance of 10ms for network transport, which is a far from realistic target for reliable network transport of audio data in any practical Immersive NMP deployment.

Furthermore, though the streaming of immersive audio [104], [105], [106] is an area of current interest in the field, it can be noted that some commonplace methods in immersive audio streaming [101] operate well beyond the 30ms full system latency target for Immersive NMP.

System Type	Low Latency Audio Streaming	Immersive Application Driven Framework	Spatial Audio Delivery	Acoustic Simulation	Interactive Acoustics (3DoF/6DoF)	Suitable for VR/AR	Practical for Typical Home Deployment
NMP	Yes	ST	ST	ST	ST	ST	ST
Telepresence NMP	No	Yes	ST	ST	ST	No	No
Avatar Based VR NMP	No	Yes	No	No	No	Yes	?
Virtual Performance	No	Yes	Yes	Yes	Yes	Yes	ST

Table 2.0.2 Overview of Immersive NMP technical requirements with respect to current state of the art and empirical research. 'ST' indicates 'sometimes'and '?' indicates 'unknown'.

This chapter aims to bridge the gap in design of immersive virtual performance VR/AR audio technology and NMP technology as required for Immersive NMP. This is done through the specification of technical requirements of, and the presentation of a design for, Immersive NMP audio systems. The solution devised is the VIIVA-NMP (discussed in Chapter 2.4) audio system, designed to be accessible to typical users operating from typical home internet connections. The prototype implementation of the VIIVA-NMP audio system design is then presented.

In this manner this chapter addresses the first part of the hypothesis:

It will be possible to design and implement an audio system which is suitable for Immersive NMP vocal performance.

2.1 Design Brief

In order to provide a technical specification of the system components in an audio system for Immersive NMP, a design brief was drafted to identify the technical requirements these system components must satisfy. Though the design presented is an audio-only system, the wider context of INMP, and associated technologies, is considered throughout system design.

2.1.1 Latency

The 30ms [29] EPT provides an indication as to the absolute system one-way throughput for Immersive NMP audio. Any system which operates beyond this latency limit is ultimately a cognitive experience unlike musical interaction between performers in the live environment [98], and is unsuitable for application with virtual performance technologies which aim to affect immersion through appropriate simulation of the live performance experience. This 30ms latency overhead requires division into audio streaming and audio rendering allotments. Average network latencies between sites, however, provides a hard limit on the amount of latency overhead which must be provisioned for network transport of audio data. In practice if an audio streaming solution wishes to achieve national (for example UK to UK) or continental (for example to Europe) range it will be required that a minimum of 20ms latency is provisioned solely for audio streaming to account for typical network delays. This leaves only 10ms for audio rendering processes, highlighting the issue with immersive audio delivery identified in telepresence NMP literature [37].

The technical requirements for the Immersive NMP audio system can therefore be identified as:

- ***20ms latency overhead should be reserved for audio streaming in Immersive NMP audio systems.***
- ***10ms latency overhead should be reserved for audio rendering in Immersive NMP audio systems.***

2.1.2 Head Mounted Displays

Immersive virtual performance technology, specifically where rendering of performer avatars [10] is considered, typically requires the use of VR/AR head-mounted displays [67]. As this represents the wider context in which the contribution of immersive audio may improve the quality of experience, the use of head-mounted displays requires consideration in system design, such that the immersive audio NMP system developed is suitable for implementation alongside such visual displays in further work.

As Eaton notes [107] ***"the most common audio playback method for consumer VR content is via headphones".*** Indeed the consideration of headphone audio playback as a VR standard is commonplace [67]. As Eaton discusses in depth, head-mounted displays typically imply headphone playback due to the access requirements associated with multichannel arrays used for immersive audio delivery. Specifically it is noted that the requirements of multichannel array immersive audio playback is that the user physically has such an array present, and that the user will have the expertise required to correctly calibrate such a system for correct operation.

It should be noted that it is possible to achieve spatial audio playback over stereo loudspeakers, using binaural or stereo methods [74]. Such methods, however, present some loudspeaker calibration requirements for correct operation. This is relatively straightforward for stereo reproduction, however is decidedly more complex for binaural reproduction [74], [108].

In the context of NMP the use of any loudspeaker array has an inherent complication: the distance between transducers and the remote performer. This represents an air propagation delay which is added to the signal path between remote performers [29]. As it is desirable to minimise latency in NMP system design, loudspeaker methods of any form can be considered impractical.

For INMP system design it is therefore prudent to make the specification of audio playback via headphones. VR/AR headset technology also typically allows for 3/6DoF [30] for movement in the virtual environment visual rendering and should provision for the same functionality in audio rendering. In Immersive NMP performance experiences it is generally suitable to assume stationary performers, and therefore simulation of head rotations only is suitable as a minimum requirement for audio rendering using VR/AR headsets, which can be achieved using head-tracking functionality to pass head rotation metadata to audio rendering systems.

The technical requirements can thereby be stated:

- ***Immersive NMP audio should be delivered via headphones.***
- ***Immersive NMP audio rendering should facilitate 3DoF using head-tracking functionality.***

2.1.3 Immersive Audio

Immersive applications aim to provide a cognitively convincing simulation of the real-world target experience [38]. In a virtual performance context this requires that audio rendering effectively emulates real world acoustic environments and the response of the performance space to performer actions in order to provision similar cognitive involvement. The design target for immersive audio delivery in virtual vocal performance systems is provided by the VIIVA system for VR vocal performance [3]. This system conforms to immersive audio requirements for immersive VR/AR applications in the provision of real-time acoustic simulation and spatial audio delivery. One key deviation from the VIIVA system audio rendering framework in the Immersive NMP audio system framework presented in this thesis is that the VIIVA systems provides simulation of the directionality of the singer’s voice as well as relative location of the other performers’ voices in the 3DoF method. The acoustic measurements required to provide this functionality are not commonplace, and introduce an additional processing overhead that is not practical for the design of an Immersive NMP system which is intended to be accessible to typical home users who are likely to have limited processing capacity. As such the Immersive NMP framework presented here considers it sufficient that in immersive audio rendering functionality head-locked audio is sufficient for a performers own audio, and 3DoF in audio rendering considers sound sources omnidirectional. Immersive audio design [66], design of virtual performance technologies [14], and specifically the VIIVA system (which is the design target) relevant to Immersive NMP for vocal performance, provide a base set of audio rendering functionalities which will be required in Immersive NMP frameworks:

- ***Immersive NMP audio should provide spatial audio delivery with 3DoF***
- ***Immersive NMP audio should provide real-time acoustic simulation with 3DoF***
- ***3DoF functionality can consider sound sources omnidirectional in minimum requirement for Immersive NMP (though efficient provision of sound source directionality would of course improve immersive audio quality)***

2.1.4 Immersive Audio and Haptic Metadata Streaming

Delivering immersive audio content requires consideration of the streaming of audio in a format where spatial audio and acoustic simulation information may be delivered between remote performers. This is generally broken down into scene-based approaches, where audio is rendered as part of a whole acoustic scene and transported in this format, or object-based approaches, where audio is transported as mono or stereo audio alongside the required information for encoding this sound source into an immersive audio rendering [105].

Specific to Immersive NMP frameworks it is recognised that avatar-rendering methods [10] require the inclusion of haptic metadata that must be transported between remote performers and incorporated to audio and visual delivery [21], [18], [22]. This haptic data should be rendered synchronously with immersive audio streams in order to present a synchronous immersive reproduction of remote musicians.

This allows for the Immersive NMP design specification:

- ***Audio streaming and rendering in Immersive NMP systems will require consideration of immersive audio streaming and compatibility with synchronous haptic metadata streams which may be enabled for avatar rendering haptic control.***

2.1.5 Practical Considerations

The design presented in this chapter is intended to provide an Immersive NMP audio system solution which is accessible by the typical home user. For this reason it is necessary to take into consideration certain aspects of the technology available to the typical home user. Firstly it is recognised that home users may have limited processing capabilities, and that any processing done by a practical Immersive NMP audio system should aim to operate within reasonable limits.

Though this project does not provide discrete measurement or evaluation of processing cost, it is still relevant to understand the processing requirements associated with aspects audio system design in an NMP context. For example, some system components will require a new instance for each remote performer. If this component includes processes which an audio engineer can recognise as being computationally taxing on a typical VR computer, then it can be recognised that multiple instances will be very taxing, and this will therefore limit the number of possible remote performers which can use the developed system. It can also be generally recognised that computationally expensive operations will require longer audio buffers to run smoothly. In an NMP context, this means additional latency, which can be recognised as limiting the potential geographic range of an NMP system. Though no discrete processing measurements or evaluation were made in this project, the system design process did include general consideration of processing requirement.

It is evident that typical home network bandwidths will provide a limiting factor in audio streaming, and that bandwidth consumption consideration will be required if the Immersive NMP technical framework presented is to prove a practical solution which is accessible to the typical home user.

It is therefore necessary that:

- ***Practical Immersive NMP audio system design should operate within the bandwidth and processing limitations of the typical home user.***

2.2 Audio Rendering Technology for Immersive NMP

Immersive NMP audio systems can be separated into audio streaming and audio rendering components. Audio rendering components must simply deliver acceptably immersive audio for virtual performance, but must accomplish this within the 10ms audio rendering latency allowance specified in the design brief, and in the context of a networked multi-user system.

Immersive audio for virtual performance experiences can be broken down into functionality as presented in the design brief, such that spatial audio delivery and real time acoustic simulation are achieved with 3DoF for each remote performer.

The VIIVA system [3] audio rendering design provides a suitably immersive target for the design of Immersive NMP audio rendering solutions which can be made suitable for practical Immersive NMP audio system implementation with some minor alterations to provide a VIIVA-NMP technical framework.

The technical design framework for audio rendering solutions in Immersive NMP systems is presented in this chapter, with discussion of the VIIVA-NMP audio rendering solution as compared to state of the art immersive audio alternatives and justification of the technical design choices made in the presented framework.

2.2.1 Spatial Audio

As the target of design in immersive applications is generally the real world experience, immersive audio aims to approach this target by rendering a simulation of the real world listening experience. In the live environment sound travels as a mechanical wave moving through a compressible medium (typically air) in 3 dimensional space. Sound waves can be described in terms of amplitude and frequency over time, but also in terms of the directional qualities of the sound wave [109].

In immersive applications as specified by Immersive NMP, particularly with virtual and augmented reality applications, accurate recreation of 3-dimensional sound is critical in achieving an immersive experience [66], [110], [111], [71]. The discipline of the capture, processing and delivery of sound with the inclusion of 3-dimensional directional characteristics is defined as spatial audio [73]. Though complex solutions to spatial audio exist, such as Wave Field Synthesis (WFS) [112], two primary methods of providing spatial audio have been used in practical immersive applications [66], [74]: Binaural stereo [113] and Ambisonics [114]. WFS is can be considered more complex than Ambisonic and Binaural methods simply due to the number of real or virtual transducers involved (or the calculation of relevant point source signals) [115]. It can be recognised that at extreme Ambisonic orders WFS and Ambisonics become similar, however such high Ambisonic order signals are not typically used in applications which aim to be accessible to typical home users due to the associated high processing requirement.

2.2.2 Spherical Coordinate Notation

Regardingsound in 3-dimensional space, common form dictates description using spherical coordinate representation (*Figure 2.2.2.1*). In this form a point in 3-dimensional space may be described relative to an origin point in terms of radial distance, *r*, azimuth (rotation on the horizontal plane), $\emptyset,$ and elevation (rotation on the vertical plane). θ [116].

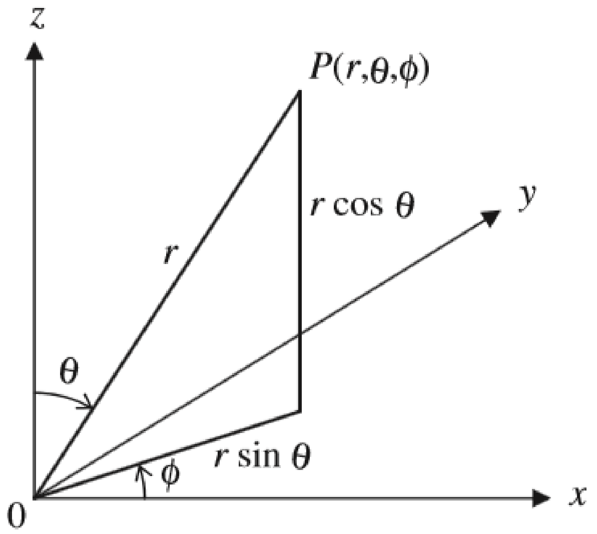


Figure 2.2.2.1 Spherical coordinate notation, sourced directly from [117].

A directional sound source will typically be described with respect to the directional Cartesian unit vector [118]:

$$u(\emptyset,\theta) = \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} = \begin{pmatrix} cos\emptyset cos\theta \\ sin\emptyset cos\theta \\ sin\theta \end{pmatrix} \qquad \textbf{\textit{(2.2.2.1)}}$$

With respect to the Cartesian unit vector ,$\boldsymbol{u}$, azimuth and elevation are defined in terms of unit vectors along spatial axis:

$$\emptyset = \arctan\frac{u_z}{\sqrt{u_x^2+u_y^2}} \qquad \theta = \frac{u_y}{u_x} \qquad \textbf{\textit{(2.2.2.2, 2.2.2.3)}}$$

2.2.3 Binaural Stereo

Binaural stereo considers the representation of 3-dimensional sound as modelled with respect to sound relative to the human auditory system [119] in order to provide the sound heard at either ear through audio playback via headphones. Binaural stereo modelling of the auditory system can be broken down into two primary components: Duplex Theory and Head Related Transfer Functions (HRTFs) [120].

Duplex theory [119] models sound source localisation using difference measures between the ears, namely Interaural Time Difference (ITD) and Interaural Level Difference (ILD) (*Figure 2.2.3.1*).

ITD models the propagation of sound waves, where the horizontal position of a sound source dictates the discrete amount of time taken for wave fronts to arrive at either ear. ITD provides an effective model for horizontal localisation below 1500Hz. Above 1500Hz the wavelength of sound propagating from a source begins to cause ITD values to no longer be specific to only one discrete horizontal direction, and ILD becomes the localisation cue in duplex theory. ILD describes the amplitude difference in a sound arriving at either ear, where sound source localisation is dependent on the attenuation of sound at one ear relative to the other. This attenuation is incurred by the difference in distance travelled between the sound source and either ear, and the shadowing effect of the physical extremities of the human head and shoulders on incoming sound waves.

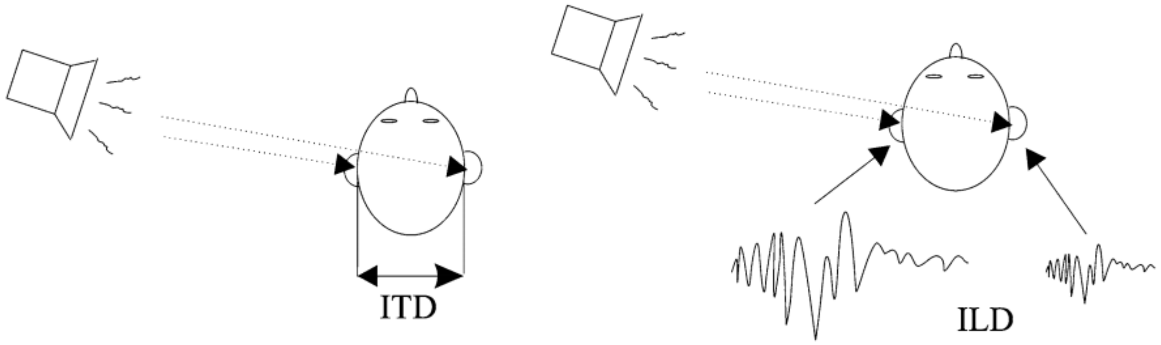


Figure 2.2.3.1 ITD and ILD illustration, sourced directly from [109].

Frequency dependencies in duplex theory, namely the shadowing effect of the human head, led to the development of HRTF representation of binaural localisation cues. HRTFs provide a comprehensive representation of binaural cues in the form of Impulse Response sets, where each discrete pair of Binaural Impulse Responses (BIRs) describes the difference between a sound source at a discrete 3 dimensional location relative to the listener and the sound apparent at each of the listener's ears. HRTF sets describe a full sphere of sound source locations relative to a listener in terms of angle to the median vertical plane (horizontal rotation, azimuth) and angle to the median horizontal plane (vertical rotation, elevation), thereby allowing for the reproduction of full 3-dimensional listening.

In its simplest form the encoding of the signal *s(t)* from direction $(\emptyset, \theta)$ to binaural stereo can be expressed in terms of the left and right headphone speaker feeds, $s_{left}(\emptyset, \theta, t)$ and $s_{right}(\emptyset, \theta, t)$ as HRTF convolutions:

$$s_{left}(\emptyset, \theta, t) = HRTF_{left}(\emptyset, \theta) * s(t) \qquad \textbf{\textit{(2.2.3.1)}}$$

$$s_{right}(\emptyset, \theta, t) = HRTF_{right}(\emptyset, \theta) * s(t) \qquad \textbf{\textit{(2.2.3.2)}}$$

As immersive applications such as Virtual and Augmented Reality typically require the use of head mounted displays [67] audio delivery via headphones is implied and binaural stereo will therefore be required at least for end-step delivery of spatial audio in Immersive NMP systems, where visual display using VR/AR avatar rendering is proposed.

Binaural processing methods which allow for 3DoF in audio rendering do however present a design concern in an Immersive NMP context. To achieve 3DoF with binaural processing methods it is required that sound sources rendered are made dynamic. As each HRTF left and right pair represents a static point in 3-dimensional space relative to the listener it becomes necessary to switch [121] , crossfade, or interpolate [122] between adjacent HRTF pairs in the HRTF set in order to produce an interactive audio scene with pure binaural methods. In an Immersive NMP context performers will typically provide a mono input which must be rendered as part of a discrete auditory scene presented to each performer. This requires that for each new performer an additional instance of dynamic HRTF filters will need to be created. As such, processing requirement for raw binaural methods will quickly increase [123] with the size of the performer group.

Though binaural stereo will be a requirement for any Immersive NMP system that aims to be compatible with AR/VR technology and modern immersive applications in terms of end-step audio playback, alternate spatial audio methods are more appropriate for audio processing prior to playback in the design of an Immersive NMP system which aims to be practically implementable from the typical home network, where processing capacity may be limited.

A more efficient and effective method for processing spatial audio for binaural playback can be found in Ambisonic methods [124] as utilised in the VIIVA system [3].

2.2.4 Ambisonics Theory

Ambisonic methods [114] achieve spatial audio processing and delivery by modelling 3 dimensional sound in terms of directional pressure vectors moving through a field of compressible medium, or ‘sound field’. The sound field is detailed with regard to directional and radial pressure functions applied on the surface of a unit sphere to provide a solution to the spherical expansion of the Helmholtz wave equation [125]. The solution to sound field pressure may be expressed by presenting the Fourier Bessel series as a product of Spherical Harmonics [126], through consideration of arbitrary directional pressure vector $u(\emptyset,\theta)$ at point $p(\emptyset,\theta)$ [116]:

$$p(\emptyset,\theta) = \sum_{m=0}^{\infty} j^m j_m(kr) \sum_{0\leq n\leq m,\sigma=\pm1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\emptyset,\theta) + \sum_{m=0}^{\infty} j^m h_m^{-}(kr) \sum_{0\leq n\leq m,\sigma=\pm1} A_{mn}^{\sigma} Y_{mn}^{\sigma}(\emptyset,\theta)$$

(2.2.4.1)

Where $Y_{mn}^{\sigma}(\emptyset,\theta)$ are the Spherical Harmonics (*Figure 2.2.4.1*), m and n are the degree and order respectively, $j_m(kr)$ represents the spherical Bessel series, $h_m^{-}(kr)$ represents the divergent spherical Hankel functions, *r* is the radial distance, *k* is the wavenumber defined as: *k = 2πf/c_{sound}* (where *c_{sound}* is the speed of sound), and B_{mn}^{σ} is a weighting function representing the influence of sound sources originating outside the sound field being considered.

A_{mn}^{σ} is a weighting function representing the influence of sound sources originating inside the sound field being considered. Ambisonic methods assume that no sound sources are apparent inside the considered sound field, and as such the weighting function A_{mn}^{σ} is zero and its products may be removed from the equation to yield:

$$p(\emptyset,\theta) = \sum_{m=0}^{\infty} j^m j_m(kr) \sum_{0\leq n\leq m,\sigma=\pm1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\emptyset,\theta) \qquad \textbf{\textit{(2.2.4.2)}}$$

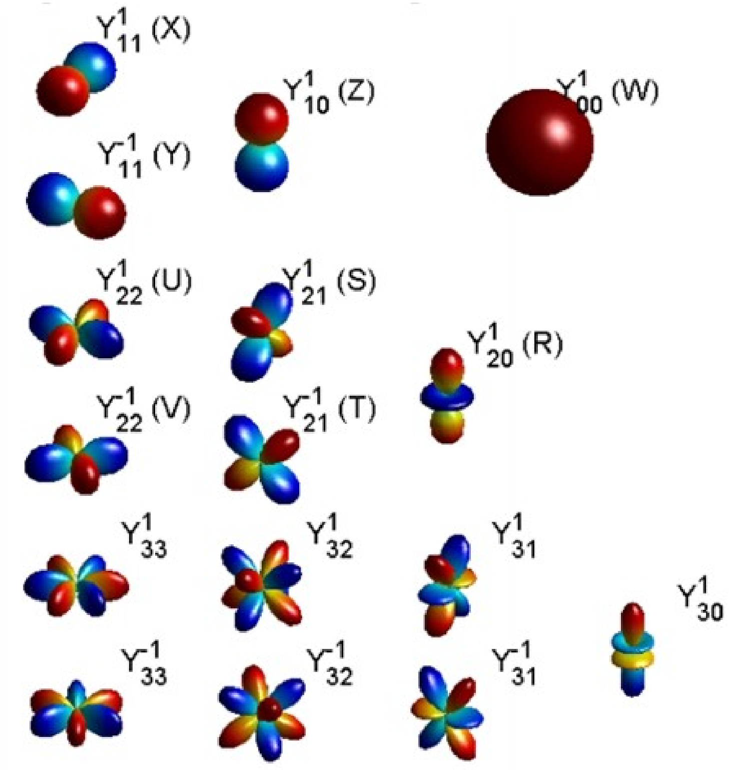


Figure 2.2.4.1 Ambisonic spherical harmonic components, sourced directly from [116].

2.2.5 Practical Ambisonics

Ambisonic methods manifest as a multichannel audio format where each audio channel represents the signal $p(\emptyset, \theta)$ for discrete values of order and degree of spherical harmonic. In its simplest form the encoding of the signal *s(t)* propagating in direction $(\emptyset, \theta)$ into the Ambisonic signals $s_{mn}(\emptyset, \theta, t)$ can be expressed as [127]:

$$s_{mn}(\emptyset, \theta, t) = Y_{mn}^{\sigma}(\emptyset, \theta) s(t) \qquad \textbf{\textit{(2.2.5.1)}}$$

Where the function $Y_{mn}^{\sigma}(\emptyset, \theta)$ represents the normalised spherical harmonics defined as [116]:

$$Y_{mn}^{\sigma}(\emptyset, \theta) = N_{mn} P_{mn}(sin\ \theta) \begin{cases} \cos(n\emptyset)\ if\ \sigma = 1 \\ \sin(n\emptyset)\ if\ \sigma = -1 \end{cases} \qquad \textbf{\textit{(2.2.5.2)}}$$

Where $P_{nm}(sin\ \theta)$ is the associate Legendre functions [125] and $\sigma = \pm 1$. N_{nm} is the Normalisation weighting function, which typically follows 3d Normalised (N3D) [116] and 3d Semi-Normalised (SN3D) [128], defined respectively as:

$$N_{mn}^{SN3D} = \sqrt{\frac{(2m+1)(2-\delta_{0,n})}{4\pi} \frac{(m-n)!}{(m+n)!}} \qquad \textbf{\textit{(2.2.5.3)}}$$

$$N_{mn}^{N3D} = \sqrt{(2m+1)(2-\delta_{0,n}) \frac{(m-n)!}{(m+n)!}} \qquad \textbf{\textit{(2.2.5.4)}}$$

Where $\delta_{0,n}$ is the Kronecker delta function, defined as:

$$\delta_{0,n} = \begin{cases} 1\ for\ n = 0 \\ 0\ for\ n \neq 0 \end{cases} \qquad \textbf{\textit{(2.2.5.5)}}$$

The multichannel format used to store this 3-dimensional sound field data is defined as the Ambisonic B-Format. In this format the Ambisonics Exchangeable (AmbiX) indexing method is used to define each channel in terms of discrete values of order and degree (*Table 2.2.5.1*). The Ambisonic Channel Number (ACN) identifying each discrete channel in AmbiX B-format is defined using the relationship [128], [129]:

$$ACN = m^2 + m + n\sigma \qquad \textbf{\textit{(2.2.5.6)}}$$

Beyond the encoding definition, Ambisonic audio may be recorded, typically using first order Ambisonic microphones, although designs do exist for higher order Ambisonic recordings [74] (*Figure 2.2.5.1*). First order Ambisonic microphones typically follow a design featuring 4 sub-cardioid microphone capsules mounted outwards-facing on the surfaces of a regular tetrahedron. These capsules are indexed Left Front (LF), Left Back (LB), Right Front (RF) and Right Back (RB) where the raw audio recording from this microphone is defined as Ambisonic A-Format, detailed in vector notation as [130]:

$$s_A = [s_{LF}(k), s_{RF}(k), s_{LB}(k), s_{RB}(k)]^T \qquad \textbf{\textit{(2.2.5.7)}}$$

In practice Ambisonic microphones will typically also enable recording into the B-Format to provide the 1st Order Ambisonic Channels in the AmbiX Format, although legacy Furse-Malham (FuMa) [129] nomenclature is still common in practice The differences between FuMa and AmbiX formats are channel ordering and nomenclature, where FuMa uses a $1/\sqrt{2}$ normalisation factor [129] rather than SN3D normalisation. In this 1st Order B-Format ACN 0 represents the omnidirectional component, and ACN 1-3 represent pressure vectors along the 3 spatial dimensions, which can be calculated from the A Format signals using sum and difference methods [73] *(Table 2.2.5.2).*

Ambisonic Order	**m, n, σ**	**ACN**	**SN3D Definition**
0	0, 0, 1	0	1
1	1, 1, -1	1	sin $\emptyset$ cos θ
	1, 0, 1	2	sin θ
	1, 1, 1	3	cos $\emptyset$ cos θ
2	2, 2, -1	4	$\left(\sqrt{3/2}\right) sin2\emptyset cos^2\,\theta$
	2, 1, -1	5	$\left(\sqrt{3/2}\right) sin\emptyset sin2\,\theta$
	2, 0, 1	6	$3sin^2\,\theta - 1/2$
	2, 1, 1	7	$\left(\sqrt{3/2}\right) cos\emptyset sin2\,\theta$
	2, 2, 1	8	$\left(\sqrt{3/2}\right) cos2\emptyset cos^2\,\theta$
3	3, 3, -1	9	$\left(\sqrt{5/8}\right) sin3\emptyset cos^3\,\theta$
	3, 2, -1	10	$\left(\sqrt{15/2}\right) sin2\emptyset sin\,\theta cos^2\,\theta$
	3, 1, -1	11	$\left(\sqrt{3/8}\right) sin\emptyset cos\,\theta(5sin^2\,\theta - 1)$
	3, 0, 1	12	$sin\,\theta(5sin^2\,\theta - 3)/2$
	3, 1, 1	13	$\left(\sqrt{3/8}\right) cos\emptyset cos\,\theta(5sin^2\,\theta - 1)$
	3, 2, 1	14	$\left(\sqrt{15/2}\right) cos2\emptyset sin\varphi cos^2\,\theta$
	3, 3, 1	15	$\left(\sqrt{5/8}\right) cos3\emptyset cos^3\,\theta$

Table 2.2.5.1 AmbiX channel definitions up to 3rd order, adapted from [129].

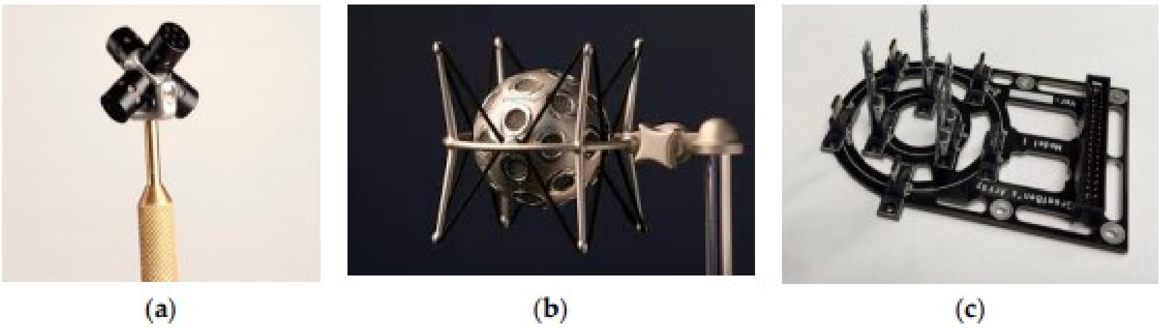


Figure 2.2.5.1 Three Ambisonic microphones: a: TetraMic, b: EigenMike, c: Planar Microphone Array, sourced directly from [74].

The key benefit to using the Ambisonic representation of 3-dimensional sound for Immersive NMP systems is found in the processing cost of enabling 3DoF for remote performers. If the pure binaural system is considered the inclusion of each new performer in a shared virtual scene requires the use of interpolation-based time-variant HRTF filters to provide rotation functionality, and a number of iterations equal to the group size multiplied by the group size minus one will be required to provide discrete rendering for every performer. It is evident that for an Immersive NMP design which accessible from typical home networks that it is necessary to consider the processing limitations of the typical home user, and the large processing costs of pure binaural methods make this an impractical choice.

ACN index	FuMa index	A Format to B Format Conversion
0	W	0.5 (LF + RF + LB + RB)
1	Y	0.5 (LF - RF + LB – RB)
2	Z	0.5 (LF – RF – LB – RB)
3	X	0.5 (LF + RF – LB – RB)

Table 2.2.5.2 First Order B-Format Microphone channel description.

If the Ambisonic system is considered in an Immersive NMP context then each performer may receive the remote performer signals as B-Format audio. The Ambisonic signal feed from each remote performer may be summed to provide a single B Format audio signal representing all remote performers relative to the local performer in a single aural scene. This allows for the provision of 3DoF by rotation of this sound field representing all remote performers relative to the local performer listener location at the centre of the sound field. In comparison to binaural methods this functionality provides vast reduction in processing requirement [123] and particularly when considering groups of performers audio being rendered in an Immersive NMP system this becomes a practical solution to enabling 3DoF.

The rotation matrices required for 1st Order Ambisonic B-Format allow for free rotation around the axis of the three spatial dimensions *x, y* and *z*, defined as roll, ψ, pitch, χ and yaw, Φ respectively [118]. In the simplest form for 1st order Ambisonic signals the rotation $R(\psi, \chi, \Phi)$ of the Cartesian unit vector, u to provide the rotated vector $\hat{\boldsymbol{u}}$ may be described [127]:

$$\hat{u} = uR(\psi, \chi, \Phi) \qquad \textbf{\textit{(2.2.5.8)}}$$

$$uR(\psi, \chi, \Phi) = \theta \left(\underset{x\ axis\ rotation\ (roll)}{\begin{pmatrix} 1 & 0 & 0 \\ 0 & cos\psi & -sin\psi \\ 0 & sin\psi & cos\psi \end{pmatrix}} \underset{y\ axis\ rotation\ (pitch)}{\begin{pmatrix} cos\chi & 0 & sin\chi \\ 0 & 1 & 0 \\ -sin\chi & 0 & cos\chi \end{pmatrix}} \underset{z\ axis\ rotation\ (yaw)}{\begin{pmatrix} cos\Phi & -sin\Phi & 0 \\ sin\Phi & cos\Phi & 0 \\ 0 & 0 & 1 \end{pmatrix}} \right)$$

$$\textbf{\textit{(2.2.5.9)}}$$

Ambisonic B-Format audio is traditionally designed for playback over loudspeaker configurations where the individual loudspeakers are placed at discrete positions on a spherical mesh around the listener position in the centre location. Individual loudspeaker feeds are calculated such that the summation of all loudspeaker outputs at the listener location provide a reproduction of the recorded sound field. The individual loudspeaker feeds are presented in terms of gains for each of the Ambisonic channels specific to each discrete loudspeaker location relative to the listener, $l(\emptyset, \theta)$. These discrete gain functions for each loudspeaker can be applied through multiplication of a gain matrix with the AmbiX B-Format channels to render the loudspeaker feed for each discrete loudspeaker. Obtaining the loudspeaker decoding matrix which contains the relevant Ambisonic channel gains for each loudspeaker feed is generally achieved in practice using the pseudoinverse decoding method.

Where loudspeaker locations are declared in the spherical harmonic domain as a matrix of *K* rows and *L* columns, ***C***, where *K* is the number of Ambisonic channels and *L* is the number of loudspeakers, and $Y_k = Y_{mn}^{\sigma}$ [131]:

$$C = \begin{bmatrix} Y_1(\emptyset_1, \theta_1) & Y_1(\emptyset_l, \theta_l) & \cdots & Y_1(\emptyset_L, \theta_L) \\ Y_k(\emptyset_1, \theta_1) & Y_k(\emptyset_l, \theta_l) & ::: & ::: \\ \cdots & \cdots & & \\ Y_K(\emptyset_1, \theta_1) & Y_K(\emptyset_l, \theta_l) & \cdots & Y_K(\emptyset_L, \theta_L) \end{bmatrix} \qquad \textbf{\textit{(2.2.5.10)}}$$

The mode-matching [132] loudspeaker decoding matrix, ***D***, can be defined using the pseudoinverse of ***C***, where ***T*** denotes transposition [123]:

$$D = pinv(C) = C^T(CC^T)^{-1} \qquad \textbf{\textit{(2.2.5.11)}}$$

This allows calculation of individual loudspeaker feeds, s_l, using loudspeaker decoding matrix, ***D***, applied to the discrete Ambisonic channel signal $s_k(\theta, t)$:

$$s_l = \sum_{k=1}^{K} s_k(\emptyset, \theta, t) * D_{kl} \qquad \textbf{\textit{(2.2.5.12)}}$$

Considering again the goal of designing an Immersive NMP system it can be easily identified that loudspeaker configurations for Ambisonic playback [133] are unlikely to be readily available to the typical home performer. As was considered earlier (Ch. 2.1.2) the context of INMP ultimately specifies audio playback over headphones in consideration of minimising latency and maximising accessibility.

In order to take advantage of the processing savings of using Ambisonic methods for providing 3DoF, particularly in a multi-user Immersive NMP context, and still provide headphone-suitable binaural stereo audio rendering the virtual Ambisonic approach [123] is used as in the VIIVA system design target. This method consider the Ambisonic loudspeaker feeds, s_l, and demonstrates that the audio heard at either ear as required by the Binaural Stereo format may be achieved by summing the convolution of the various loudspeaker feeds, s_l, for loudspeaker locations, $l(\emptyset,\theta)$, with the relevant HRTFs, $HRTF(\emptyset,\theta)$, for each ear:

$$s(t)_{binaural} = \sum_{l=1}^{L} HRTF(\emptyset_l,\ \theta_l) * s_l \qquad \textit{\textbf{(2.2.5.13)}}$$

This can be computed more efficiently in the spherical harmonic by attaining a decoder matrix for the Ambisonic channels which includes the binaural decoding information, $\boldsymbol{D_{SH}}$, by multiplication of decoding matrix, $\boldsymbol{D}$, with the relevant HRTFs and summing of the Spherical Harmonic representation for each loudspeaker:

$$D^{SH} = \sum_{l=1}^{L} HRTF(\emptyset_l,\ \theta_l)\, D_l \qquad \textit{\textbf{(2.2.5.14)}}$$

This allows calculation of the binaural stereo signal for each ear using the summing of the Ambisonic channels $s_k(\emptyset,\theta,t)$ convolved with the Virtual Loudspeaker Binaural Decoding matrix, $\boldsymbol{D_k^{SH}}$:

$$s(t)_{binaural} = \sum_{l=1}^{L} s_k(\emptyset,\theta,t)\ * D_k^{SH} \qquad \textit{\textbf{(2.2.5.15)}}$$

In practice this method allows for a computationally efficient Binaural Stereo spatial audio delivery method with 3DoF as specified in the design brief.

In the VIIVA system the implementation of the Virtual Ambisonic approach is coupled with an Auralisation method for simulation of performer voice directivity using higher order Ambisonic impulse response measurements. These methods combined incurred a throughput latency of ~20ms [3]. Though this is acceptable for local system implementation, extension to the Immersive NMP context requires that this throughput is significantly reduced. By removing the directionality function and using 1st order Ambisonic signals it is possible to drastically reduce the throughput of such a system. As a convolution operation, the Virtual Ambisonics method allows for use of overlap-add partitioning schemes using short window sizes [134], [135], therefore enabling the minimisation of local latency in achieving spatial audio delivery with 3DoF. In the VIIVA system 20ms throughput is associated with vector sizes of 512. It is possible, at the cost of extra processing overhead (thus the requirement for reduction in Ambisonic order), to reduce these buffer lengths to as low as 64 samples, which in practical application is associated with 1.3ms latency considering a standard audio sampling rate of 48 kHz.

2.2.6 Acoustic Simulation

In immersive audio applications, particularly in a virtual performance context, acoustic simulation of the virtual environment is required. Even where no visual accompaniment is present it can be considered that any performer in a performance simulation will expect to encounter the type of reverberant acoustic performance space typically encountered in live interactions. To provide a convincing performance simulation it is required that acoustic simulation of the reverberant performance environment is interactive and rendered in real-time.

Several commonplace methods of achieving real-time acoustic simulation of reverberant spaces [86] are available, namely Feedback Delay Network (FDN) [85] reverberators, Geometric methods [76] such as Ray Tracing [136], Beamforming [102] or Image Source [137], and Impulse Response convolution [75].

2.2.7 Feedback Delay Networks

Feedback Delay Network (FDN) reverberators [85] are networks of all-pass and comb filters used to provide parametric simulation of room acoustics. Though independent filter components typically provide artificial-sounding reverberation, large FDN recursive networks [138] have been found to provide acceptably natural sounding results.

With respect to Immersive NMP systems FDN reverberators offer a significant processing saving compared to other reverberation methods, which allows for practical implementation. As Immersive NMP requires spatial audio delivery, it is essential that FDN designs are appropriate for this application, specifically in terms of compatibility with the Ambisonic medium specified for Immersive NMP technical frameworks.

Though some investigation of Directional FDNs [139] and Scattering FDNs [140], [141] for Ambisonic reverberation has been conducted this has been a relatively recent development and FDN reverberation for Ambisonics in its current state is not the optimal solution for Immersive NMP and validation is required that such FDN designs can provide a perceptually equivalent degree of naturalness (an important quality in immersive applications) in comparison to other reverberation methods.

2.2.8 Geometric Acoustics

Geometric acoustic approaches consider that sound propagating radially through 3-dimensional space can be modelled effectively with a finite amount of discrete linear paths representing the radiation of acoustic waves. By computing the paths in terms of distance travelled and boundary reflections while calculating for diffusion and absorption of the sound the acoustics may be modelled by summing the divergent paths at any receiver location in the simulated environment.

The Image Source Method (ISM) [137] (*Figure 2.2.8.1*) presents the simplest form of geometric acoustics, where boundary reflections can be computed as simply as mirror images of coincident paths. Image source methods appropriate for Ambisonic applications have been proposed [142]. This approach, however is generally considered only appropriate for simulating simple geometric rooms, rather than the bespoke acoustic spaces found in musical performance.

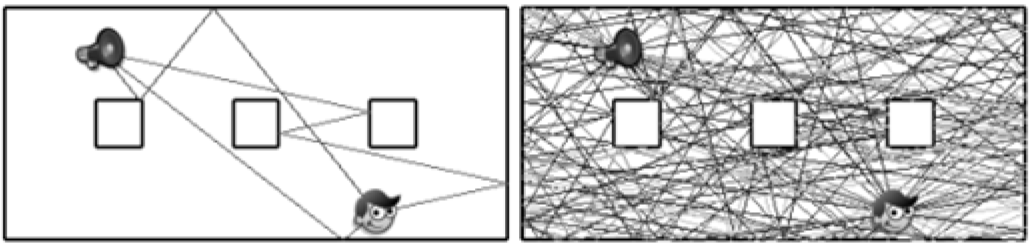
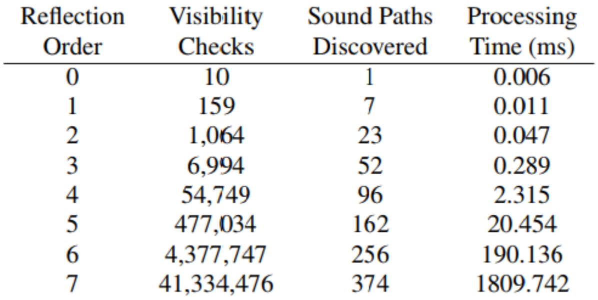


Figure 2.2.8.1 3rd order (left) and 8th order (right) reflections for a single sound path, sourced directly from [103].

More effective ray tracing and beamforming methods can provide natural sounding reverberation for spatial audio applications, however a base-level trade-off is apparent in all such approaches. The quality of the rendered result is entirely dependent on the number of paths computed *(Figure 2.2.8.1),* where denser computations provide superior results, but require much higher computational load and throughput latency to facilitate [76].

For Immersive NMP systems geometric approaches will generally incur a higher processing load than is practical for the typical home user, and incur higher latency [103] than can be reserved for audio rendering in Immersive NMP design specifications where high orders of reflection are considered *(Table 2.2.8.1),* and do not present an optimal solution to providing interactive acoustic functionality.



Reflection Order	Visibility Checks	Sound Paths Discovered	Processing Time (ms)
0	10	1	0.006
1	159	7	0.011
2	1,064	23	0.047
3	6,994	52	0.289
4	54,749	96	2.315
5	477,034	162	20.454
6	4,377,747	256	190.136
7	41,334,476	374	1809.742

Table 2.2.8.1 Geometric Acoustic processing times for a room with 10 surfaces, sourced directly from [103].

2.2.9 Spatial Impulse Response Convolution

Impulse Response (IR) Auralisation [75] methods model reverberant environments as filters describing source-receiver positions in a reverberant space. The sound source is considered the input, and the sound apparent at any receiver location is considered the output. The acoustic influence of the reverberant space is represented in the IR characterising the filter relationship [143]:

$$y(n) = x(n) \otimes h(n) = \sum_{k=-\infty}^{\infty} x(n)h(n-k) \tag{2.2.9.1}$$

Where *x(n)* is the sound source, *y(n)* the output at the receiver location, and *h(n)* the impulse response.

IRs may be rendered using simulation models and geometric approaches [86], where offline rendering may achieve high quality results which can later be used in convolution reverberators [15]. For the real performance spaces desired for simulation in Immersive NMP, IRs can be measured practically. The measurement process [144] involves the calibration of a loudspeaker at the sound source position, and microphone at the receiver (or 'listener') location. The loudspeaker is used to output an excitation signal, and the sound at the receiver location is recorded in order to measure the response of the acoustic environment to the excitation signal.

IR measurement [145] can be achieved using a range of excitation methods. Provided the excitation signal is spectrally and temporally even the differences apparent at the signal recorded at the receiver position will accurately describe the acoustic influence of the measured space. In practice Farina's sine sweep method [146] is typically used. Farina's excitation signal, *s(t),* is described by:

$$s(t) = \sin[Ke^{-\frac{t}{L}} - 1] \tag{2.2.9.2}$$

Where ***t*** is time, calculated with respect to duration, *T*, and sweep start and end frequencies $\boldsymbol{\omega_1}$and $\boldsymbol{\omega_2}$ define *K* and *L,* given as:

$$K = \frac{\omega_1 T}{\ln\left(\frac{\omega_1}{\omega_2}\right)} \tag{2.2.9.3}$$

$$L = \frac{T}{\ln\left(\frac{\omega_1}{\omega_2}\right)} \tag{2.2.9.4}$$

With recorded room signal *r(t),* impulse response, *h(t),* is calculated by convolution of the room response with inverse filter, *f(t).* The inverse filter used is the reversed excitation signal with the modulation signal applied:

$$m(t) = \frac{\omega_1}{\omega(t)} \tag{2.2.9.5}$$

Historically IR measurement standards specify the use of an omni-directional loudspeaker to output the excitation signal and omni-directional microphone to record the room response.

Spatial Impulse Response measurement [147], [148] follows the same measurement protocol, however an Ambisonic microphone, typically a first order B-Format microphone [149], [150] will be used (*Figure 2.2.9.1*) instead of an omnidirectional receiver *(Figure 2.2.9.1).*

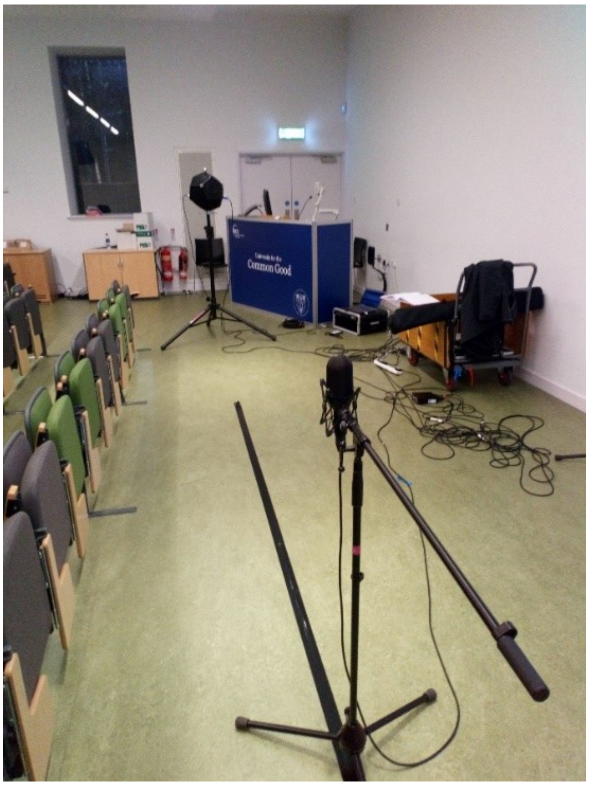


Figure 2.2.9.1 Example of SIR measurement setup.

SIR measurements typically are presented as 1st Order B-Format multichannel IRs, where each channel represents the room response with respect to the specified Ambisonic channel. A mono audio signal can be encoded through convolution with the measured SIR to render the mono input as a first order Ambisonic signal representing the sound source Auralised in the space described by the SIR. The position of the sound source relative to the receiver, or 'listener' position can be described in terms of the Ambisonic representation of directional sound sources such that the encoding of the signal *s(t)* may be expressed:

$$s_{mn}(\emptyset,\theta,t) = s(t) * h_{SIR}(\emptyset,\theta,t) = \begin{bmatrix} s(t) * h_{ACN0}(\emptyset,\theta,t) \\ s(t) * h_{ACN1}(\emptyset,\theta,t) \\ s(t) * h_{ACN2}(\emptyset,\theta,t) \\ s(t) * h_{ACN3}(\emptyset,\theta,t) \end{bmatrix} \qquad \textbf{\textit{(2.2.9.6)}}$$

In virtual performance technology this method is used to place performers in the Auralised performance space relative to one another, simulating the acoustic experience of the space and providing Spatial Audio representation of the sound experienced by performers. This convolution process is implemented in real-time audio applications using over-lap add partitioning algorithms [134] to allow for manageable latency and Cooley-Tukey Fast Fourier Transform (FFT) methods [151] to reduce processing overhead.

The overlap-add convolution method [135] *(Figure 2.2.9.2)* can be outlined as:

- Partition the input into frames of audio
- Zero-pad input audio frames and IR to an even length
- Zero-padded input and IR are transformed to the frequency domain using FFT. Multiplication is performed to calculate the convolution, then an inverse FFT is applied to resolve to the convolved audio in the time domain.
- Convolved audio output frames are summed into the output buffer at the relevant time indices

In immersive virtual performance experiences, SIR auralisation can be used to place performers in a range of measured performance spaces and provide high quality immersive audio playback with interactive acoustic simulation (using Ambisonic rotation methods) and spatial audio delivery, as in the VIIVA system [3]. SIR databases representing ranges of source-receiver locations within real measured performance spaces [152] can be used to auralise performers in these spaces with bespoke performer placement within the virtual acoustic space.

This method offers a practical opportunity for extending the virtual performance experience to Immersive NMP systems, where overlap-add algorithms allow for implementation using extremely short buffer lengths to minimise audio rendering latency while still providing a high quality immersive performance experience. The cost of such low-latency, high-quality immersive audio rendering is the increased processing cost of using short partition lengths in convolutions. In a practical Immersive NMP setting this will imply a limit on the size of a performance group associated with the processing overhead, with a similar limit to SIR Auralisation Ambisonic order.

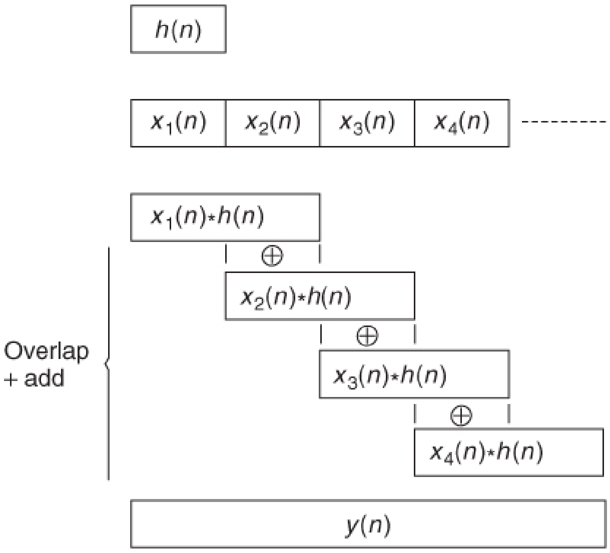


Figure 2.2.9.2 FFT Overlap-Add Finite IR filter design, sourced directly from [135].

2.2.10 The VIIVA System and Immersive NMP Audio Rendering

As discussed, the VIIVA [3] system presents a target for the technical design of audio rendering for Immersive NMP. The significant technical components that make up the VIIVA system outlined in this section.

The VIIVA system is a VR vocal performance implementation, where users engage in the experience of singing with a virtual performance group. This is implemented locally, where the user takes the place of the live virtual performer, and accompaniment is provided by interactive (3DoF) playback of 360 degree video and spatial audio recording of the virtual performance group.

The VIIVA system renders the users voice in the virtual performance using SIR convolution, Ambisonic rotation methods with head-tracking to enable 3DoF with the virtual performer playback, and Virtual Ambisonic binaural rendering for audio playback via headphones to provide high quality interactive acoustic simulation and spatial audio delivery. This system has proven suitably immersive, and received a positive response from experienced vocal performers who participated in system testing, making this a suitable design target for Immersive NMP. There are some key deviations from the VIIVA system design, however, which will be required to adapt this design to Immersive NMP.

Firstly, VIIVA system design only considers the real-time audio rendering of a single user, and has allocated processing overhead accordingly. SIR convolution provides 3rd order Ambisonic auralisation to achieve excellent immersive audio quality. This auralisation utilises an extension of the SIR convolution methods detailed in Section 2.2.9, where SIR sets describe directional sources. This involves the use of SIR sets which describe the instance of singers own sound, where the source and receiver location are the same, and rather than omnidirectional loudspeakers, directional loudspeakers are used to output IR measurement excitation signals in order to model the directionality of the sound source. These SIR sets describing ‘own voice’ in the virtual space at discrete directions describe the rotation of the singers own head, and are interpolated between using head-tracking data and convolved with the input audio.

The SIR measurements used in the VIIVA system are specialist measurements that must be made by audio professionals on a bespoke basis. Such directional SIR sets are not typically included in SIR performance space databases which might be utilised in Immersive NMP systems and which are accessible to the home user. . Though the VIIVA system presents a high quality auralisation, SIR sets which describe sound source rotations (which are required to auralise directional sound sources) are impractical in the context of INMP due to lack of accessibility..

Regarding real-time interpolation between higher order Ambisonic SIR, processing is also a concern for Immersive NMP. Rendering directionality for groups of performers will require high processing capacity as an instance of this function will be required for each performer.

Indeed in the VIIVA system development and testing buffer lengths of 512 samples, associated with 10.7ms latency at 48 kHz sample rate, were required for smooth functionality. In order to use the short buffer lengths in convolution-based systems which allow for the minimisation of audio rendering latency required for Immersive NMP, it is necessary that the processing incurred must be reduced significantly in order to provide a solution which is suitable for use with multiple users operating from typical home computers and network connections.

This processing reduction can be achieved by firstly considering remote performer voices as omni-directional such that time invariant FIR filters may be used for SIR convolution. The Ambisonic order may also be reduced to first order to significantly reduce the processing cost of SIR convolution (16 convolutions at 3rd order down to 4 at 1st order). The local performers own voice is head-locked, while remote performer inputs may be auralised and summed into a single scene for rotations in 3DoF functionality.

Though the quality of immersive audio delivery will be reduced in comparison to the VIIVA system, using omnidirectional performer voice and 1st order Ambisonic Auralisation it is possible to reduce the processing overhead to a manageable level where multiple instances of this functionality can be implemented in order to render reasonably sized groups of performers to all be rendered at each local instance of the audio renderer. 1st order Ambisonic reproduction can be considered appropriate for Immersive NMP as it is the typical order to which Ambisonic spatial audio will be typically implemented in consumer VR/AR applications.

The second main deviation from the VIIVA design in the technical specification of audio rendering for Immersive NMP is with respect to video rendering methods. Though the initial VIIVA design utilises 360 degree video playback, it is noted that the latency incurred by video streaming is unsuitable for Immersive NMP design. The VIIVA system development proposes a multi-user extension which should utilise avatar-based rendering [9], as specified in Immersive NMP design to circumvent the latency cost of video streaming. Though this is not an audio factor, it must be recognised that the use of avatar rendering systems will require the streaming of haptic metadata synchronous to audio streaming. Though this thesis is concerned with audio streaming and rendering design, this deviation from VIIVA system design and consideration for further work is worth stating. In this respect the consideration of performer voice as omnidirectional can reduce the bandwidth cost of audio streaming, as sound source directional information will not require inclusion in synchronous metadata streams, and streaming of only mono audio alongside head-tracking and other haptic data will provide a bare-bones but sufficient method of achieving the proposed full Immersive NMP implementation with inclusion of avatar rendering which audio rendering and streaming components must be compatible with.

2.3 Audio Streaming Technology for Immersive NMP

The implementation of the audio streaming component of an Immersive NMP audio system is described with respect to how remote performers are connected over networks, and how audio data describing the virtual performance is shared between these remote performers. The chief concern in audio streaming is latency, as achieving transport and delivery of audio within the 30ms EPT latency limit is integral to delivering an immersive level of musical interaction between remote performers.

Existing audio streaming solutions such as Jacktrip [100] are capable of providing this low-latency audio streaming functionality and describe a technical framework for the low-latency transport of audio data between endpoints on typical Internet Service Provider (ISP) backbones. In the context of Immersive NMP system design, however, it is important to consider how such audio streaming frameworks fit together with audio rendering components and other proposed components of Immersive NMP systems such as metadata streaming for avatar rendering, and the practical considerations which may make Immersive NMP technology accessible to the typical home user.

This chapter section discusses the technical design and configuration of audio streaming technology for Immersive NMP, with some discussion of potential solutions and justification of design choices made.

2.3.1 Transmission Medium

Considering audio transport across ISP backbones, it is important to recognise how nodes on the network are physically connected, and how audio data physically moves between these endpoints. Data traverses the WAN ‘roadmap’ where Routers are the intersections, Asymmetric Digital Subscriber Line (ADSL) connections are the B-roads, and fibre-optic the Autobahn. As in Immersive NMP high speeds and bandwidths are required, fibre-optic transport medium is specified.

The speed at which data is carried over fibre-optic cable is given as [6]:

$$speed_{fibreoptic} = 0.7 * c_l \qquad \textbf{(2.3.1.1)}$$

Where c_l is the speed of light (~$3*10^8$ ms^{-1}). The latency incurred by network transport of audio data can be broken down into distance travelled through fibre-optic medium, latency incurred by each ‘hop’ (throughput of each router encountered in network transport), and any additional latency incurred by transport processes. When combined with audio rendering latency, an absolute range on the effective distance of any Immersive NMP system can be described in relative to system latency and the 30ms full system latency allowance as:

$$distance_{range} = 0.7 * c_l * (0.03 - delay_{rendering\ process} - delay_{hops} - delay_{transport\ process})$$

(2.3.1.2)

Where c_l is the speed of light (~$3*10^8$ ms^{-1}).

This range is dependent on the network provisioning of the ISP backbone between remote performer access points, which is beyond the control of the typical home user. As such minimising latency of audio transport and rendering processes is integral to maximising system range. It is important to recognise that Immersive NMP ranges will never be global with current technology, and continental deployments will be the best possible within the 30ms latency limit presented by the EPT.

The size of the physical connection also provides limits on audio data transport. The amount of physical medium available to transport data is defined by the internet bandwidth. Pro audio applications specify a minimum of 16-bit at 44.1 kHz but can typically range up to 24-bit at 96kHz [51]. This implies a minimum of 0.7056Mb/s per audio channel, ranging up to 2.304Mb/s.

Bandwidth consumption in Immersive NMP system design impacts audio streaming format and performance ensemble size. If it is necessary to stream high numbers of audio channels as will be required with multichannel scene-based immersive audio streaming [105], [106], or include large performance groups, a high bandwidth consumption will be incurred. Recently measured (2017) global bandwidth averages at 7.2 Mb/s, yet top 10 country bandwidth all average above 18.7 Mb/s, and global trends in improvement are shown [52]. It can be seen that the average node in any internet-based NMP system can manage at least a handful of audio channels, yet will be restricted in terms of immersive audio Streaming method and group size. Some high-speed networks exist that are available through academia such as Internet2 (USA) [153], Geant (Europe) and Joint Academic Network (JANET, UK) [154] which offer connections of up to 200 Gb/s. Such connections are capable of handling large and diverse Immersive NMP audio streaming [155]. Considering that commercial network quality and bandwidth is ever-increasing and 1 Gb/s connections are available from current commercial Internet Service Providers (ISPs) the limitations of narrow-band networks are largely a temporary setback in the application of Immersive NMP systems [156].

2.3.2 Topology and Architecture

In order for audio data to be shared between remote performers in Immersive NMP audio is sent between from the device operated by each remote performer to the relevant LAN router, and along ISP backbones to other remote performers via each remote performers own discrete LAN router. This form of WAN configuration (*Figure 2.3.2.1*) is described as a 'Common-Carrier WAN', where the ISP provides the 'common carrier' connection between LANs [157].

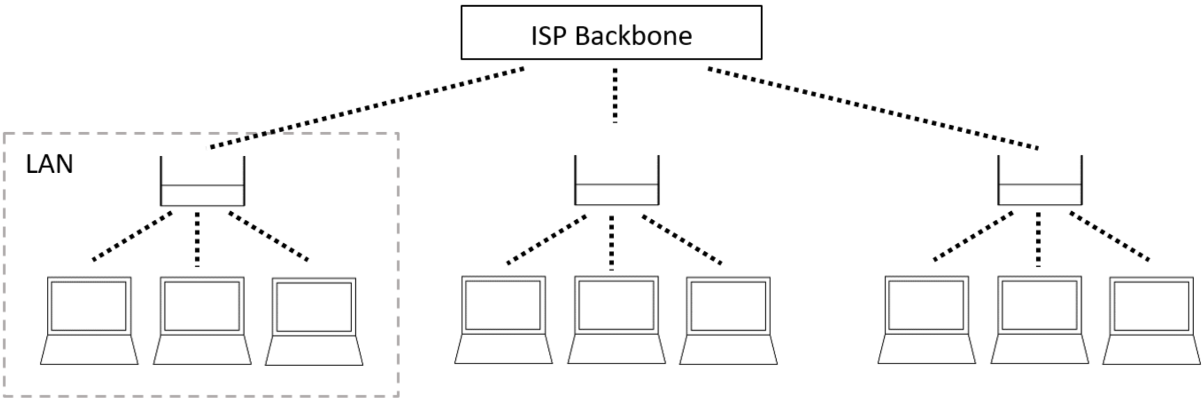


Figure 2.3.2.1 Common Carrier WAN configuration (adapted from [157]).

The way in which the connections between performer home routers are made in Immersive NMP, and the functionality required at each node on the network, is defined by the system network topology and architecture.

2.3.3 Peer-Peer Architecture

In Peer-Peer systems nodes on the network form direct connections with other nodes on the networks [158]. In Immersive NMP systems this means remote performers may stream audio directly between one another. In order to achieve the shortest path routing, fully meshed topologies may be implemented. In this instance each remote performer has a discrete send to and receive from each other remote performer (*Figure 2.3.3.1*). The shortest path routing allows for the minimisation of network latency by minimising the number of hops and distance of fibre-optic transport medium traversed. In this way the practical range of an Immersive NMP system may be maximised.

With respect to practical Immersive NMP audio streaming design Peer-Peer fully meshed architectures offer the best solution to making this technology accessible across practical ranges, and offer the most opportunity to maximise the range of potential remote performers with geographic range. As this method requires discrete duplex audio streams at each remote performer equal to the group size minus one, it is apparent that Peer-Peer systems require consideration of bandwidth consumption.

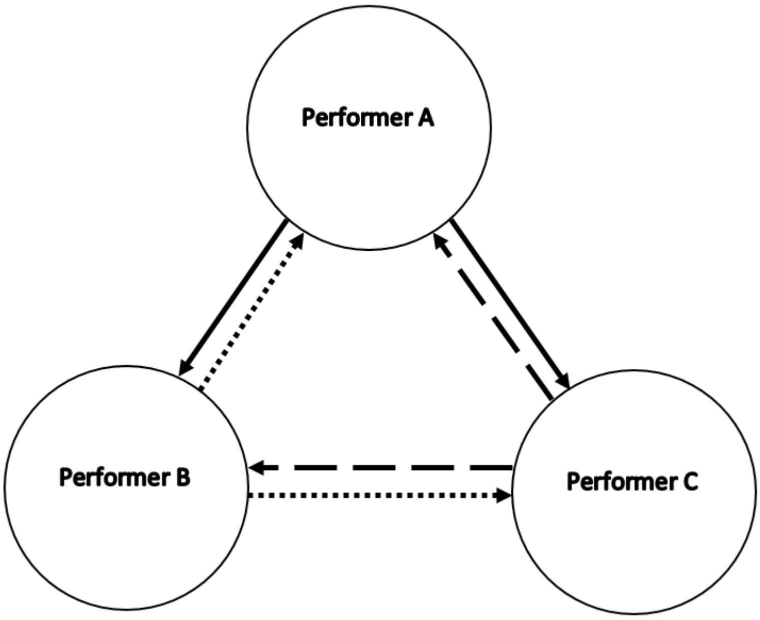


Figure 2.3.3.1 Peer-Peer fully meshed configuration for Immersive NMP.

It can be observed that scene-based immersive audio streaming methods [105], [106] such as the streaming of Ambisonic signals (as would be required for locally enabling 3DoF using Ambisonic rotation matrices as discussed in section 2.2) will require a large bandwidth consumption for even small group sizes. The use of object-based immersive audio streaming, where a mono signal is streamed and combined with metadata required for audio rendering, is needed for this Immersive NMP audio streaming design. In the virtual performance scenario, considering performer audio objects as omnidirectional, the only metadata required for audio rendering is performer location. The location of each discrete performer in an Immersive NMP ensemble may be pre-computed at each local performer node, such that no synchronous locational metadata streaming is required for systems which operate with 3DoF. This allows the minimisation of bandwidth consumption in audio streaming, and thereby maximisation of potential Immersive NMP group size, by simply streaming a mono audio signal between remote performers. This method does, however, come at the cost of additional processing requirement at each local performer, where audio streams from remote performers as well as local performer audio input are required to be simulated in the virtual acoustic performance space by an audio rendering instance at each performer node.

2.3.4 Client-Server Architecture

In Client-Server architectures [159] for Immersive NMP, each remote performer node on the network is connected in star topology to a central server node which may provide audio processing functionality, manage audio signal routing, and output audio streams to client nodes as required. In Immersive NMP system design Client-Server architectures present the opportunity to implement systems based around Server audio renderers. In this design a mono audio stream may be sent from each client to the central server, and audio rendering processes (as detailed in section 2.2) may be implemented to render a binaural mix-down of a group performance scene for each discrete remote performer. In order to achieve this it will be required that head-tracking metadata is streamed from each Client to the Server audio renderer in order to implement 3DoF in audio rendering. The binaural scenes, which represent the Immersive NMP performance group in the virtual acoustic performance space from the perspective of each discrete performer can be streamed from the server to the associated discrete remote performers in a scene-based immersive audio Binaural Stereo stream. In this manner it is possible to deliver audio for Immersive NMP to each remote performer using only one send channel and two receive channels. In terms of practical accessibility, the bandwidth consumption and processing requirement minimisation at remote performers present an optimised design, as processing and bandwidth requirement are minimised. By specifying high-processing power at audio renderer servers it is possible to further optimise the technical functionality of acoustic simulation using Higher Order Ambisonic signal rendering and processing, and further limit the latency incurred in audio rendering by providing a computer which can operate smoothly at extremely short buffer lengths.

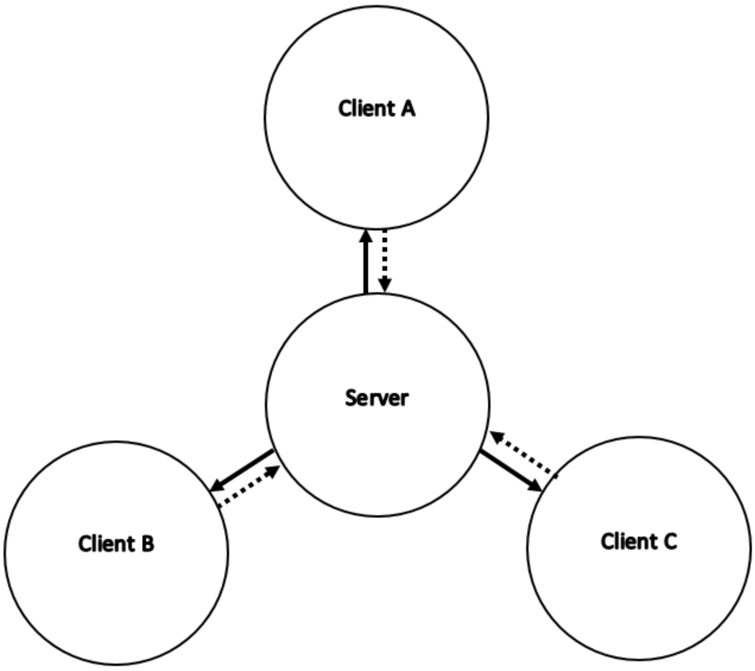


Figure 2.3.4.1 Client-Server architecture in star topology.

The difficulty presented by Client-Server architectures in Immersive NMP is twofold. Firstly, servers must be located at a high-bandwidth access point to act as host for large performance groups, managing 3 audio channels per remote performer. Such high bandwidth connections are not yet typical of home internet connections, though this is something that is expected to change over time. Considering current networks, Client-Server architectures are only possible with the use of high speed WAN connections such as that of academic internets. It can also be considered that this network configuration requires not just access to high-speed networks, but also to high power processors. In this way Client-Servers are impractical in Immersive NMP systems which aim to be accessible to the typical home user. The second difficulty presented by Client-Server architectures is the additional network latency incurred by routing sharing audio with other remote performers via the server. Audio transport across ISP backbones from remote performers to the server then from the server to each remote performer will present a mild to large (depending on the location of the server relative to remote performer internet access points) range reduction when compared to the Peer-Peer system outlined in section 2.3.3.

In these respects Client-Server architectures for Immersive NMP are not readily available to the typical home user, and Peer-Peer systems present the accessible and practical option. Client-Server architectures may indeed prove useful, however, where specific groups have access to facilities which meet the requirements of this network routing design, and all operate within a local region described by the effective server range. Such Immersive NMP designs should, however, be reserved for bespoke specialist implementation.

2.3.5 Network Addressing and Transport Protocol

Network addressing and transport protocol define how data moves from one discrete node to another through the various connections in Wide Area Networks. In Immersive NMP this describes how audio data is shared between performers across the network in order to render playback the shared virtual performance scene to each remote performer.

In order to identify the location of devices on the network, IP addressing [160] is the de-facto solution. This address identifies the device sending or receiving data, and provides a 'postal address' which directs data transport between devices. IPv4 [161] *(Table 2.3.5.1)* (or IPv6) datagrams provide the 'envelope' for IP addressing, and provides some additional management functionality in audio data transport across networks. Data sent using IP addressing is managed by a transport protocol [162], which exists as a packet structure within IP datagram payloads, and provide the rules governing the processing of data within these packets. As audio dropout presents a significant break in immersion, stable and high quality audio streaming is integral to the immersive quality of the virtual performance experience, making transport protocol and stream management major design considerations.

<table>
<tr><td><b><i>Version
(4bits)</i></b></td><td><b><i>IHL (4 bits)</i></b></td><td><b><i>DS Field
(6 bits)</i></b></td><td><b><i>ECN (2 bits)</i></b></td><td colspan="2"><b><i>Total Length (16 bits)</i></b></td></tr>
<tr><td colspan="4"><b><i>Identification (16 bits)</i></b></td><td><b><i>Flags (3 bits)</i></b></td><td><b><i>Fragment Offset (13 bits)</i></b></td></tr>
<tr><td colspan="2"><b><i>Time-to Live (8 bits)</i></b></td><td colspan="2"><b><i>Protocol (8 bits)</i></b></td><td colspan="2"><b><i>Header Checksum (16 bits)</i></b></td></tr>
<tr><td colspan="6"><b><i>Source IP Address (32 bits)</i></b></td></tr>
<tr><td colspan="6"><b><i>Destination IP Address (32 bits)</i></b></td></tr>
<tr><td colspan="6"><b><i>Options (up to 320 bits)</i></b></td></tr>
<tr><td colspan="6"><i>Payload (up to 524, 120 bits)</i></td></tr>
</table>

Table 2.3.5.1 IPv4 Datagram, adapted from [161].

2.3.6 Transmission Control Protocol

Transmission Control Protocol (TCP) is a ‘connection-oriented’ protocol which provides full management of connections and data flow between sender and receiver sockets [163]. The reliability of TCP is achieved through the inclusion of packet transmission acknowledgement functionality, implemented in the TCP packet *(Table 2.3.6.1)* headers, where sender and receiver sockets communicate to ensure that data is transported correctly. Errors in the data stream are managed by Automatic Repeat Request (ARQ) [164] error correction in this context, where receiver sockets send requests to the sender socket to retransmit missing or erroneous packets.

Although the reliability of TCP data streaming seems like an attractive solution to providing stable audio streaming between remote performers, the latency cost of retransmission schemes proves problematic where the Immersive NMP full system latency limit of 30ms is considered. The process of transmission, error detection, retransmission request, and retransmission of missing or erroneous packets comes at the cost of 1.5 Round Time Trip (RTT) of latency between remote performers. In practice, where it is required to consider the typical gateways available to the typical home user for Immersive NMP systems, such ARQ or other retransmission-based error correction [165] will result in latencies well above the 30ms full system limit [166].

TCP is however appropriate in non-real-time components of Immersive NMP, such as to transport control data for configuring the virtual performance setup (virtual room, performer location, session setup and stream setup) which can be computed prior to remote musical performance.

<table>
<tr><th colspan="10">Source Port (16 bits)</th><th>Destination Port (16 bits)</th></tr>
<tr><td colspan="11">Sequence Number (32 bits)</td></tr>
<tr><td colspan="11">Acknowledgement Number (32 bits)</td></tr>
<tr><td>HL (4 bits)</td><td>Resv (0) (4 bits)</td><td>CWR</td><td>ECE</td><td>URG</td><td>ACK</td><td>PSH</td><td>RST</td><td>SYN</td><td>FIN</td><td>Window Size</td></tr>
<tr><td colspan="10">Checksum (16 bits)</td><td>Urgent Pointer (16 bits)</td></tr>
<tr><td colspan="11">Options (variable)</td></tr>
<tr><td colspan="11">Payload (variable)</td></tr>
</table>

Table 2.3.6.1 TCP packet structure, adapted from [167].

2.3.7 User Datagram Protocol

User Datagram Protocol (UDP) [168] offers a 'bare-bones' approach to data transport and packet structure (*Table 2.3.7.1*). As a 'connectionless' protocol, UDP packets are simply sent from a sender to a receiver destination with no communication between send/receive sockets. Though UDP is typically considered an unreliable method of data transport many NMP streaming solutions [29] successfully achieve stable audio streaming using UDP transport through the inclusion of data flow management and error correction functionality implemented on top of UDP.

Source Port	***Destination Port***
Length	***Checksum***
Payload	

Table 2.3.7.1 UDP packet structure, adapted from [169].

Low-latency audio streaming frameworks such as Jacktrip [100] or Soundjack [93], which are capable of operating within the 30ms latency limit associated with Immersive NMP, utilise a custom packet structure which implements functions of Real-Time Protocol (RTP) [170] inside the UDP payload. The inclusion of sequence numbers allows for the detection of missing or out-of-order packets at the receiver socket. Error correction functionality can be achieved at the receiver end using Forward Error Correction (FEC) [171] methods (*Figure 2.3.7.1*), where sending overlapping audio buffers in UDP payloads allows for missing audio buffers to be retrieved from redundant packets stored at the receiver socket.

The sending of overlapping audio buffers in FEC implementation for NMP does induce an additional bandwidth consumption to provision for the streaming of redundant packets, however allows transport latency to be minimised by performing stream management and error correction at the receiver socket. This means that no retransmission is required, and only a single One Way Trip (OWT) between remote performers of latency (plus the length of the audio buffer overlap) is incurred in network transport between remote performers. In practice this method has shown the ability to achieve stable audio streaming functionality with only small FEC audio buffer overlap. It is also possible to achieve transport latencies between remote performers below the 20-25ms latency allowance specified in the design brief for Immersive NMP. This approach offers a practical solution to Immersive NMP audio transport.

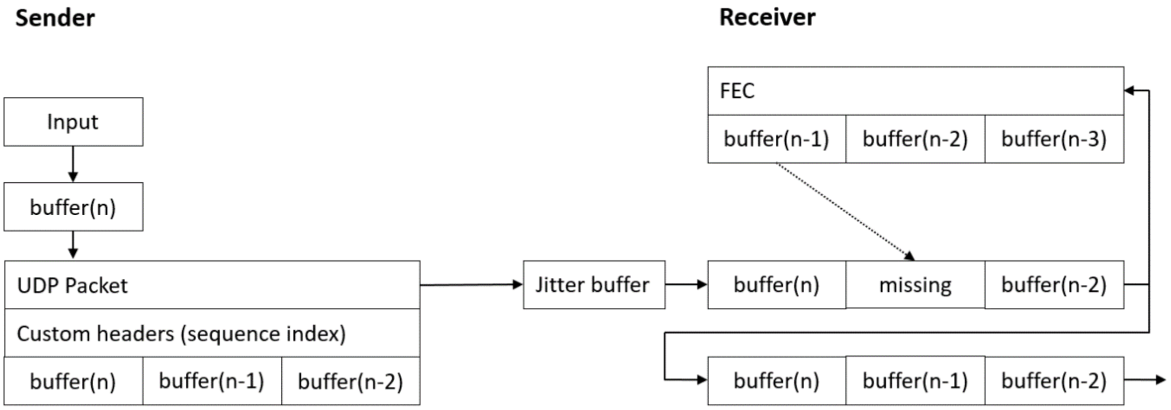


Figure 2.3.7.1 FEC example illustration with two redundant audio buffers used.

2.3.8 Jitter Buffering

The latency incurred by the transport of audio data across networks is not constant, and will typically vary with practical considerations of using typical home network connections, such as network congestion [172]. This variance in transport delay between nodes on a network is defined as network ‘jitter’. In Immersive NMP this means audio streams from remote performers will arrive at receive sockets at a variable rate. As audio data is required to be read from the buffer at a constant rate, management of network jitter is required to achieve the stable audio streams required by Immersive NMP.

The standard method of managing network jitter in this situation is Jitter Buffering (*Figure 2.3.8.1*), which is implemented as a large ring buffer at receiver sockets. Data is received into the ring buffer at a variable rate (non-blocking) until the buffer is full, at which point audio data is pushed to the output buffer where it can be read at a constant rate. The size of the ring buffer adds to the network transport latency, however will allow the establishing of a stable audio stream providing the jitter buffer is sufficiently sized.

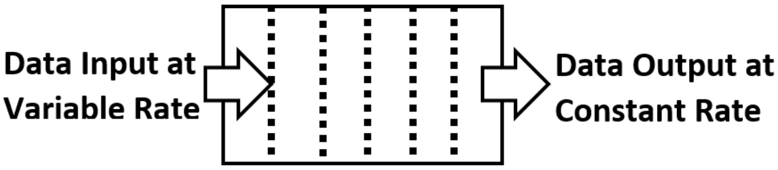


Figure 2.3.8.1 Jitter Buffer

Two error conditions may typically be encountered in jitter buffering for audio: over-run (where asynchronous clock drift causes the audio data to arrive at the receive socket while the jitter buffer is full) and under-run (where the jitter buffer must push audio data to be read before it is full, caused by jitter buffer sizing which is smaller than network jitter, or by packet loss). Low-latency audio streaming solutions such as Jacktrip [100] manage these conditions by making the jitter buffer non-blocking and resetting the read index at the jitter buffer for over-runs, as well as providing signal interpolation error correction or 'silent mode' options for under-run conditions.

2.3.9 Compression Codecs

As discussed in Section 2.3.1, bandwidth consumption is a significant concern in Immersive NMP which impacts immersive audio streaming and the potential number of performers in an Immersive NMP ensemble. Compression of the audio data being streamed allows for reduction in the bandwidth consumed by each audio channel, however does so at the cost of additional latency incurred by the encoding and decoding processes in compression codec application [173]. A range of low delay audio compression codecs (*Table 2.3.9.1*) are available such as AAC-LD/ELD [174], [175], Fraunhofer Ultra-Low Delay (ULD) [176], SILK [177], CELT [178] and OPUS [179], [180].

Codec	**Bit Rate**	**Sample Rate**	**Latency (min window size)**	**Quality**
Opus	6-510kb/s	Up to 48kHz	2.5ms	Poor - Excellent
ULD	64-80kb/s	32-48kHz	5.33-8ms	Good - Excellent
CELT	32-96kb/s	44.1kHz	8.7ms	Good - Excellent
AAC-LD	32-64kb/s	22-96kHz	16-129.3ms	Good – Excellent
SILK	6-40kbp/s	8-24kHz	20ms	Poor - Good

Table 2.3.9.1 Comparison of compression codecs.

Using low-bitrate compression it is possible for these codecs to minimise encoding/decoding latency, however a drop in audio quality is associated with the use of such low-bitrate perceptual codecs. In

Immersive NMP such audio quality reduction can detract from the immersive quality of the virtual performance experience, and may prove problematic when further acoustic simulation and spatial audio processing is required as discussed in section 2.2.

Though compression is a necessity in NMP systems which are accessible to users on narrow band networks [99], modern consumer internet has now reached speeds where the streaming of multiple channels of PCM audio is achievable at the typical home internet connection. Even low bitrate compression algorithms cut significantly into the 20-25ms latency allowance for audio data transport in Immersive NMP and will severely decrease the range of such a system. As the associated audio quality reduction is highly undesirable in any virtual performance experience. The use of compression is largely impractical for Immersive NMP, and streaming of PCM audio is preferable.

The use of PCM does restrict group size in Immersive NMP, and restricts audio streaming methods to object based approaches in order to maximise the potential group size. Trending increase in consumer internet speeds does however indicate that this will only be a temporary concern.

Compression codecs may prove useful for bespoke application in which a dedicated Immersive NMP system is deployed to support remote performance for large groups whereby all access points are within the limited system range. Applications of compression in Immersive NMP can also be found where online audiences are considered. Such immersive concert broadcast systems [106] can be achieved by streaming the virtual performance auditory scene as an Ambisonic signal to each discrete audience member, such that 3DoF and binaural stereo audio rendering can be achieved locally at each virtual performer. In such mass broadcast scenarios the large number of streams requires compression even on high-speed academic networks. As it is not required to provide the same level of musical interactivity as required in Immersive NMP between the performers and online immersive audience members, a greater latency can be accepted in audio playback to online audience members. Here codecs such as OPUS [181] will be appropriate in broadcast of the virtual concert immersive experience (*Figure 2.3.9.1*).

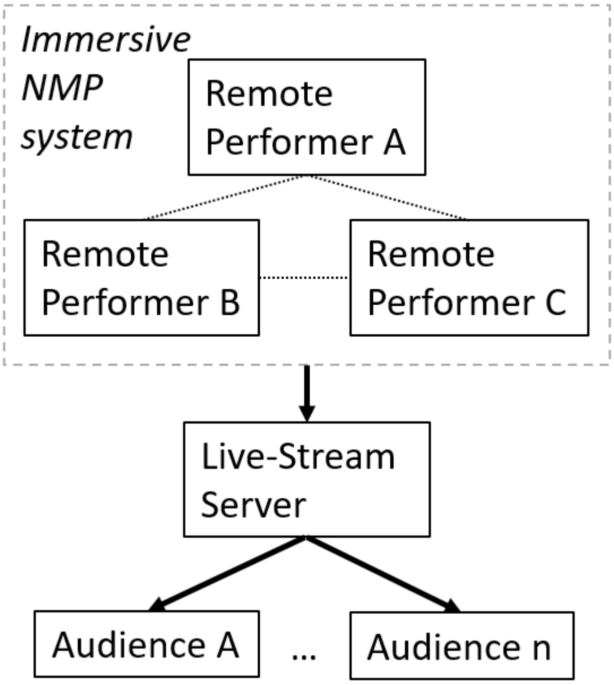


Figure 2.3.9.1 Immersive NMP concert broadcast example block diagram.

2.4 VIIVA-NMP Audio System Technical Design

The review of audio rendering and streaming technology for Immersive NMP presented in this chapter has facilitated the specification of a technical design for Immersive NMP which meets the requirements identified in the Design Brief.

This technical blueprint has been developed considering an Immersive NMP system for group vocal performance, following the objective of extending the VIIVA system [3] to the NMP context. The technical design presented in this thesis has accordingly been named the Vocal Interaction in an Immersive Virtual Acoustic Network Music Performance (VIIVA-NMP) audio system. Though specified for Immersive NMP for voice the VIIVA-NMP framework presented is adaptable to other instrumentation.

The VIIVA-NMP audio system provides a technical design for the audio component of proposed Immersive NMP systems which is appropriate for implementation alongside and integration with proposed avatar-based VR/AR visual rendering [9], [10], [4], and is practically accessible to the typical home user. In this way the VIIVA-NMP audio system presents a contribution to the development of next-generation Immersive NMP systems as a technical blueprint which can be referenced, followed and implemented. As such this system design allows for the accepting of the hypothesis presented in this thesis:

It will be possible to design and implement an audio system which is suitable for Immersive NMP vocal performance.

It can be noted that the system can only deliver an effective simulation of the aural experience of live performance within the geographic distance that full one-way system latency can be kept below 30ms. Beyond this range the musical interaction between remote performers is significantly different from live performance, and the virtual performance experience can low longer be considered ‘immersive’. Another limitation is imposed by bandwidth, which provides a hard limit on the number of performers who can take part in an Immersive NMP experience together. As such the hypothesis can only truly be accepted within these limits. In the typical use case geographic range is expected to extend to ~500km and ensemble sizes of ~10 performers should be achievable by the VIIVA-NMP audio system design. In this manner the range in which the VIIVA-NMP audio system is suitably immersive is more than enough to demonstrate Immersive NMP as a viable technology.

2.4.1 Equipment and System Requirements

The VIIVA-NMP audio system is developed with consideration of accessibility to the typical home user. For this reason it was specified that any practical design should function using only equipment which can be expected as typically accessible to the home user.

Capturing performer vocals requires that a microphone is available. The optimal solution specified is an omnidirectional headset microphone with an even frequency response to avoid coloration of the input signal, and to allow for optimised quality as specified in SIR Auralisation and spatial audio processing functions in audio rendering.

Signal filtering by hardware (such as microphone) response can act as an additional filter on top of HRTFs used in binaural decoding. This changes the response of the HRTFs, therefore distorting the spatial image provided by the HRTF response [119].

Such high quality headset microphones are generally available to professional vocal performers, however may not be available to members of non-professional vocal performance groups such as community choirs. Many commercially available and commonplace VR headsets such as the HTC [182] and Oculus [183] models do however have on-board microphone options, and many VR users may also have alternative headset microphones available. Such microphones can be acceptably used for the VIIVA-NMP audio system and can generally be expected as accessible to the typical home user, however it should be recognised that audio quality and accuracy of spatial sound reproduction will be diminished in comparison to high quality vocal mics if there is any coloration in the audio capture.

In application any stationary microphone such as traditional studio microphones which require microphone stands will be unsuitable, as free rotation of the performer head is required to provide 3DoF in rendering of the virtual performance experience in both the VIIVA-NMP audio system and proposed Immersive NMP frameworks complete with avatar-based visual rendering. In such complete Immersive NMP designs the use of haptics to enable embodiment through gestural communication is also proposed, and physical obstructions such as stationary microphones must therefore be avoided.

Microphone audio capture is input to the local computer either through USB connection or via an external audio interface. This audio input is then handled by either the dedicated audio interface drivers or the computer Operating System (OS) audio drivers. High quality audio interfaces and drivers will often allow for smooth operation at lower buffer lengths and higher sample rates than standard OS audio drivers. As each audio process in the VIIVA-NMP audio system is associated with a buffer of audio, reducing the buffer length and increasing the sample rate allows for minimisation of system latency (*Table 2.4.1.1),* and the use of high quality external audio interfaces with dedicated drivers will facilitate the best performance. Acceptable buffer lengths can however be achieved using standard Windows and Mac OS audio drivers, which will be accessible to typical home users.

Sample rate / Buffer Size	***64***	***128***	***256***
44.1kHz	1.45ms	2.90ms	5.80ms
48kHz	1.33ms	2.67ms	5.33ms
96kHz	0.67ms	1.33ms	2.67ms

Table 2.4.1.1 Associated latency for discrete buffer length and sample rates.

The VIIVA-NMP audio system performs audio rendering and streaming processes on user home computers. VR/AR headset technology which operates on home computers generally recommends system requirements upwards of [184], [185]:

- Quad core CPU with clock speed greater than 3GHz
- 8GB RAM
- 4GB VRAM (typically GTX 1060, equivalent or better)

Computers which meet these specifications can therefore be expected as typically available to an Immersive NMP home user, and the audio system processes should be easily computable on typical home computers meeting such requirements.

The VIIVA-NMP audio system streaming component is specified to transport audio data to and from home routers over ISP backbones with no required intermediate network provisions. A typical home router with a public IP address and a direct Ethernet connection between the home computer and router is the only networking hardware required of home users of the VIIVA-NMP audio system.

Audio playback in the VIIVA-NMP audio system is delivered via headphones as specified in compatibility with VR/XR headset technology which represents the wider context of INMP. In the context of vocal performance, open-back headphones will minimise the occlusion of direct sound from the singers own voice in comparison to closed-back models [186].

Headphone response acts as an additional filter on top of HRTFs used in binaural decoding [67], which presents the same issues as with microphone response. In order to minimise any potential HRTF distortion, headphones which provide minimal additional filtering are preferable [187] though headphone correction methods may also be used to avoid headphone response acting as an additional filter on audio playback [187].

2.4.2 System Overview

In order to maximise the geographic range in which the VIIVA-NMP audio system delivers an immersive virtual performance experience a Peer-Peer meshed design is specified. This requires an instance of the VIIVA-NMP audio system at each remote performer (*Figure 2.4.2.1).* Each performer inputs audio into the local VIIVA-NMP audio system instance via microphone input. This vocal performance audio capture is then passed to the audio streaming component, which shares audio between the Immersive NMP ensemble using a discrete duplex mono stream between the local performer and each discrete remote performer. These audio streams are formed between user home routers where the audio data is transported over commercial ISP backbones. Local performer audio capture is input to the audio rendering component alongside the discrete receive stream for each discrete remote performer. These audio inputs are rendered in the virtual acoustic performance space by the audio rendering component, which provides audio playback of the virtual acoustic performance scene via headphones for the local performer.

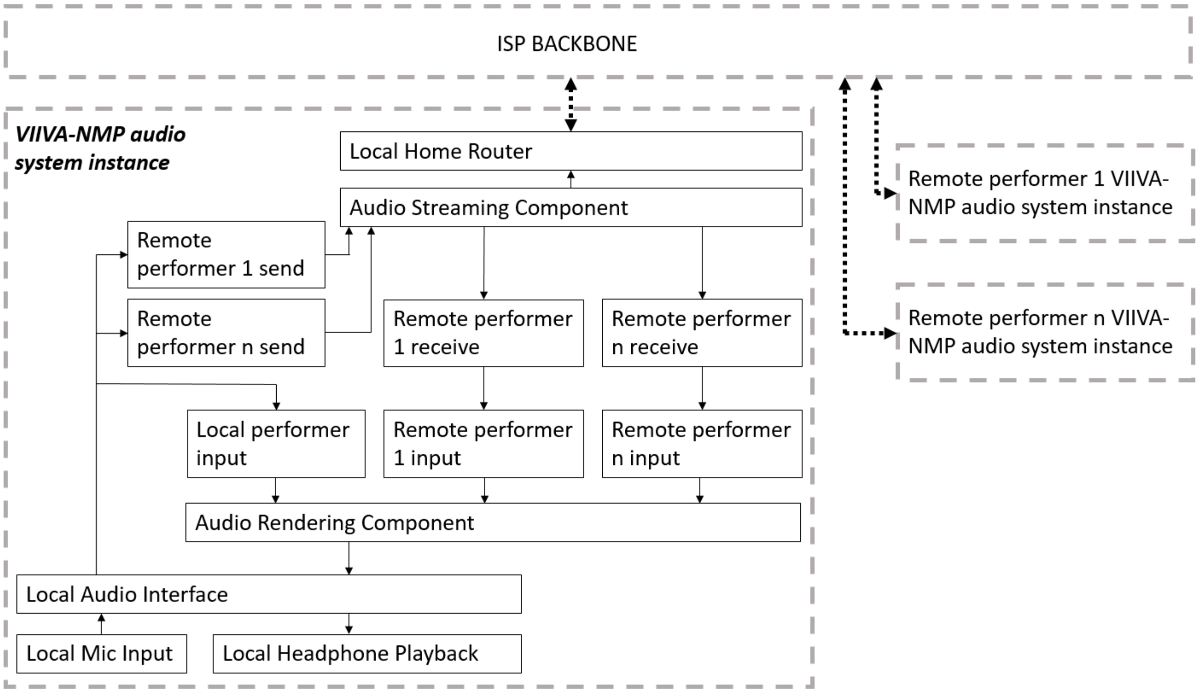

ISP BACKBONE
VIIVA-NMP audio system instance
Local Home Router
Audio Streaming Component
Remote performer 1 send
Remote performer n send
Remote performer 1 receive
Remote performer n receive
Local performer input
Remote performer 1 input
Remote performer n input
Audio Rendering Component
Local Audio Interface
Local Mic Input
Local Headphone Playback
Remote performer 1 VIIVA-NMP audio system instance
Remote performer n VIIVA-NMP audio system instance

Figure 2.4.2.1 VIIVA-NMP audio system overview.

2.4.3 Audio Streaming Component

The local performer input is transported to remote performers in the Immersive NMP ensemble by sending audio data in UDP packets directly (as possible over ISP backbones) from sender sockets on the local performer computer to specified ports at remote performers home routers using the relevant public IP address at each home router (*Figure 2.4.3.1*). This requires that each performer home router has port forwarding configured to pass data from the home router to the computer hosting the VIIVA-NMP audio system instance, and that the computer and home router firewalls are configured to allow this network traffic.

The direct UDP Peer-Peer streams allow for minimisation of audio transport latency, allowing again for maximisation of the geographic range within which the VIIVA-NMP audio system will deliver a suitably immersive experience. As erroneous and lost packets are an inevitability when considering streaming over commercial ISP backbones, as well as a degree of network jitter, jitter buffering and FEC are specified to allow for effective error correction and constant-rate audio reading at receive sockets. This allows the delivery of a stable audio stream between remote performers while minimising the additional bandwidth and latency cost of achieving this.

The UDP packets containing audio data payloads are specified as following a custom RTP-lite packet structure with the inclusion of additional packet headers:

- **Sample Rate**, **Buffer Size** and **Bit-Depth** describe the audio data contained in the UDP payloads, such that receive sockets can correctly depacketize this audio data and pass the receive streams to the audio rendering component.
- **Sequence Number** is required for the identification of erroneous or missing packets in FEC functionality.
- **FEC Overlap/Redundancy** describes the number of redundant buffers of audio data sent in the UDP payload, allowing for correct depacketization and FEC functionality at receive buffers.
- **Timestamp** allows for the implementation of clock synchronisation between remote performers and with the synchronous metadata streams that will be required in any implementation of Immersive NMP avatar rendering. In practice clock synchronisation can be implemented using low-cost GPS designs [188]. It is not expected that this will be readily accessible for the typical home user, nor will be strictly required in many use cases, however allowing for such implementation allows for compatibility with more complex installations.

The additional header information can also be considered to fulfil the bare-bones required fields for RTP audio transport and session description as specified by LAN interoperability standards such as AES67 [189], thereby making the VIIVA-NMP audio system compatible with bridging into pro-audio LAN installations and other complex networked AV deployments.

In order to ensure that receive sockets correctly hold the packet description parameters expected in the stream received from sender sockets a TCP handshake is required at the beginning of the audio stream setup. As multiple streams will need to be managed in any implementation other session description parameters such as group size and port allocation must be configured manually, or through TCP session setup communication.

In order to minimise latency, optimise audio quality, and conform with pro audio over IP specifications no compression is used and the VIIVA-NMP audio system is specified to stream uncompressed PCM audio. Streaming PCM audio can require anywhere from ~0.8Mb/s (44.1kHz, 16-bit, including some network overhead) to ~2.5Mb/s (96kHz, 24-bit, including some network overhead). In order to limit the bandwidth consumption of PCM streaming in Peer-Peer meshed configuration an object-based immersive audio approach is specified in the VIIVA-NMP audio system, requiring only duplex mono streaming between performers, Minimisation of bandwidth consumption allows maximisation of possible group size for INMP vocal performance ensembles.. As the VIIVA-NMP audio system considers sound sources omni-directional and performer locations are configured in session setup, this allows for further minimisation of bandwidth consumption, as all parameters required for object-based acoustic simulation and spatial audio rendering may be computed locally at each performer, and no additional audio metadata stream is required.

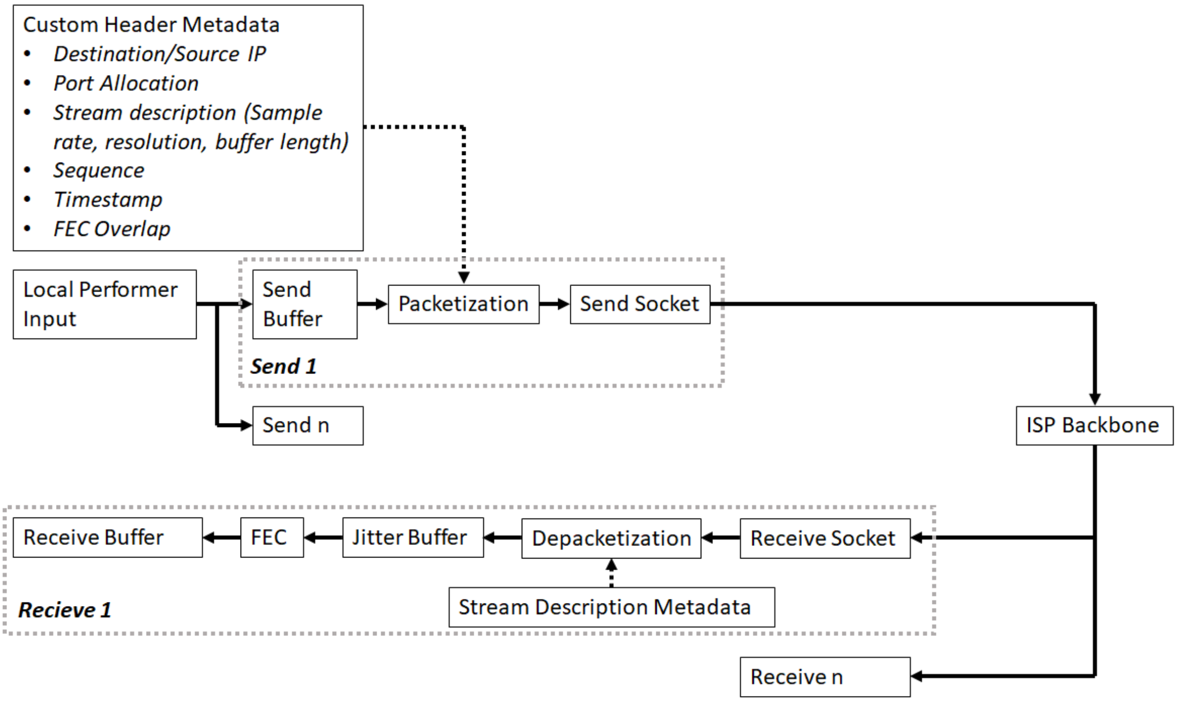


Figure 2.4.3.1 VIIVA-NMP audio streaming component design.

2.4.4 Audio Rendering Component

The audio rendering component takes audio input from the local performer microphone capture and the receive stream from each remote performer in the Immersive NMP ensemble.

These performer inputs are rendered as sound sources in the virtual acoustic performance space using SIR convolution. SIR measurement sets are used, where each discrete SIR describes a sound source relative to a receiver location within a real vocal performance space. Local performer input should be Auralised with the same source and receiver location where possible, however such SIR measurements may not be typically available. SIR measurements with minimal source-receiver relative distance should be used where such 'singer's own voice' SIR measurements are unavailable. This convolution provides the local performers own discrete audio feed output as a 1st Order Ambisonic signal of the vocal capture rendered in the virtual acoustic performance space. The input from each discrete remote performer can then be rendered as a sound source placed relative to the local performer in the virtual acoustic performance space using a discrete SIR convolution describing each discrete source-receiver relative location for each discrete remote performer, and outputting a discrete 1st Order Ambisonic signal representing each remote performer as Auralised in the virtual acoustic performance space.

These convolutions are implemented using standard over-lap add real-time convolution partitioning schemes. As the VIIVA-NMP audio system specifies sound sources should be considered omni-directional, the processing of source directivity is avoided. The significance of this is that no real-time interpolation is required in the SIR convolution stage of audio rendering, meaning smooth operation can be achieved at short buffer lengths, convolutions can be implemented as simple Linear Time Invariant (LTI) filters using short minimum partition lengths (typically as low as 64 samples) and the latency and computational load induced by the SIR convolution stage of the audio rendering component is minimised.

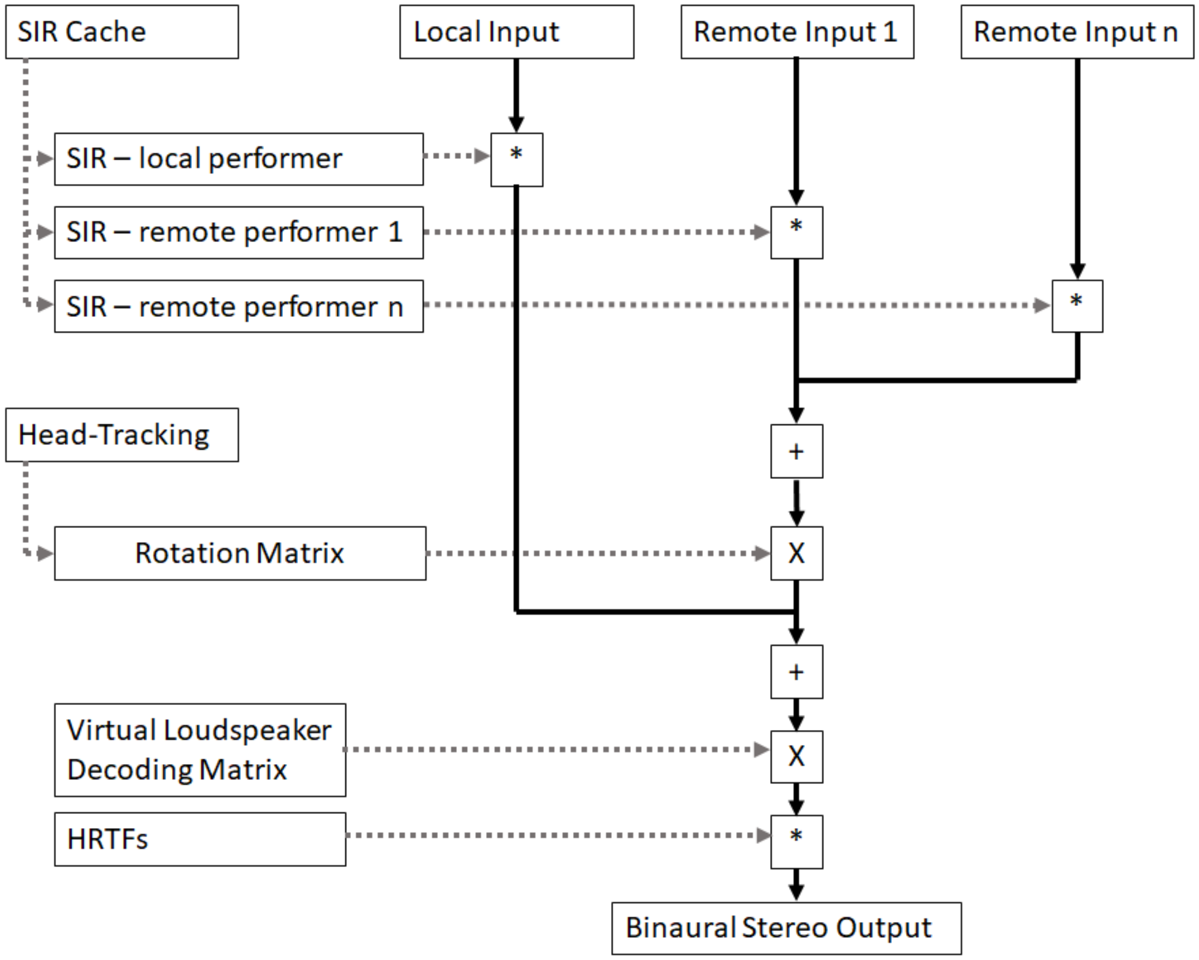


Figure 2.4.4.1 VIIVA-NMP audio system rendering component signal processing chain.

Considering sound sources omni-directional further allows for processing of head rotations independently of the SIR Auralisation stage. As it is necessary toavoid the externalisation of the local performers own voice, local performer input is headlocked. The Ambisonic signals representing each discrete remote performer may all be summed together to provide a single Ambisonic signal representing the virtual acoustic performance scene relative to the local performer (minus the local performers own audio). This Ambisonic scene representing all remote performers may be rotated relative to the local performer in response to local performer head rotations to provide 3DoF in audio rendering. This is accomplished by utilising head-tracking capabilities, typically built-in to VR/AR headsets, which can provide yaw-pitch-roll OSC control. The yaw-pitch-roll values are used to rotate the remote performer scene around the local performer using standard Ambisonic rotation matrix multiplication.

The head-locked local performer Ambisonic signal may then be summed with the post-rotation remote performer Ambisonic scene to output a single 1st Order Ambisonic signal representing the entire virtual acoustic performance scene. The virtual Ambisonics approach can then be used to provide binaural stereo playback to the local performer.

The audio rendering design (*Figure 2.4.4.1*) is specified such that the only processes involving LTI filter convolutions (SIR Auralisation and HRTF convolution) and matrix gains (rotation and loudspeaker decoding) are required to provide audio delivery which meets the requirements of virtual performance technology. This means that computational load is reduced throughout the system in order to achieve smooth operation with short buffer lengths. Each convolution or matrix gain need only incur 64 samples of latency. In practice binaural decoding via the virtual Ambisonic approach will save a further buffer through by including the loudspeaker decoding in the HRTF set required for Binaural Stereo rendering. In this way the audio rendering component of the VIIVA-NMP system minimises audio rendering latency to 3 buffers (SIR convolution, Ambisonic matrix rotation, binaural decoding). Considering the recommended sample rate and buffer size (*Table 2.4.1.1*) ranging from 96kHz, 64 sample to 44.1kHz 256 sample this implies the VIIVA-NMP audio rendering component exhibits a throughput latency of 2.1ms minimum to 17.4ms maximum. Ideal operation is near the minimum rated latency, which most typical users should be able to achieve.

2.4.5 VIIVA-NMP Design Prototype

In order to demonstrate the validity of the implementation of this technical blueprint and provide functioning example which may be utilised in user-testing a prototype VIIVA-NMP audio system was developed and deployed using existing open-source resource.

In the prototype configuration each performer's vocal performance is captured via local microphone input. DPA omni-directional vocal headset microphones were specified for the prototype design, with signal input via Focusrite Scarlett 18i20 gen 1 [190] or Focusrite Scarlett 2i2 gen3 [191] audio interface (and handled by the dedicated Focusrite audio drivers). As the prototype must be valid for practical use, a practical user instance was specified for prototype development: the primary researcher's home computer and internet access point. The home internet fibre-optic connection was provided by British

Telecom ISP featuring a download speed of ~50Mb/s and an upload speed of ~10Mb/s. The home computer hardware spec features 16GB DDR4 RAM and Intel i7-6700K CPU (quad-core, 4GHz).

Audio input is taken from the local performer microphone input from the audio interface using Jack Audio Connection Kit [192], [193] (Figure 2.4.5.1), which facilitates audio routing between the audio interface, audio streaming component and audio rendering component of the VIIVA-NMP prototype. The JACK router must be configured manually to operate at the desired buffer size, sample rate and resolution and to operate using the audio interface drivers (in this case the dedicated Focusrite drivers).

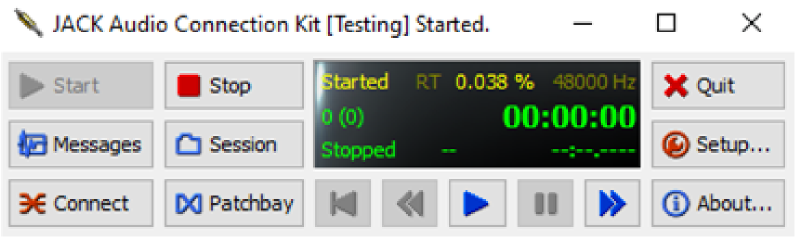
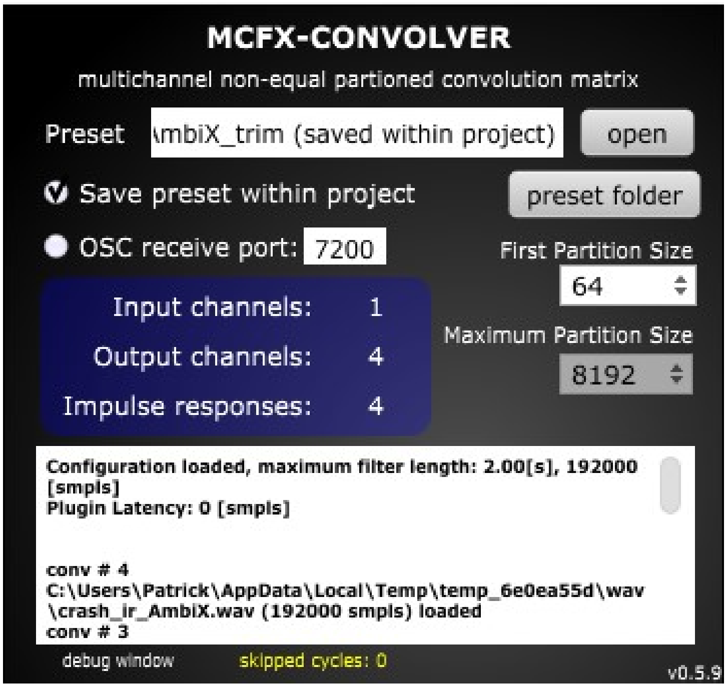


Figure 2.4.5.1 Jack Audio Connection Kit [193].

The audio streaming component uses Jacktrip [100], an open source streaming solution developed by Caceres and Chafe at CCRMA Stanford which follows (or can even be considered as having defined) the standard framework for low-latency audio streaming over WANs. This audio streaming solution allows for specification of FEC redundancy, number of channels, sample rate, buffer size, resolution, port allocation and jitter buffer length (amongst other stream parameters) manually as a console application which establishes discrete audio streaming between public IP addresses (home computers behind home routers with relevant port forwarding). A mono Jacktrip stream between each discrete performer pair in the group should is configured to create a meshed Peer-Peer WAN deployment as specified for the VIIVA-NMP audio system streaming component. Audio from local performer input can then be routed to Jacktrip send sockets for each remote performer, and Jacktrip receive streams can be routed from receive sockets to the audio rendering component using the Jack router.

The audio rendering component is hosted in Reaper [194], which can use the JACK router as an audio interface to allow for audio input and output. Inside Reaper the audio rendering signal chain is processed using VST. For each performer (local and remote) input a 4-channel track hosts an instance of the Kronlachner mcfx convolver [195] (*Figure 2.4.5.2*) which performs the SIR Auralisation using a multichannel partitioned over-lap add convolution algorithm.



MCFX-CONVOLVER

multichannel non-equal partioned convolution matrix

Preset \mbiX_trim (saved within project)

☑ Save preset within project

○ OSC receive port: 7200

First Partition Size: 64

Maximum Partition Size: 8192

Input channels: 1

Output channels: 4

Impulse responses: 4

```
Configuration loaded, maximum filter length: 2.00[s], 192000
[smpls]
Plugin Latency: 0 [smpls]


conv # 4
C:\Users\Patrick\AppData\Local\Temp\temp_6e0ea55d\wav
\crash_ir_AmbiX.wav (192000 smpls) loaded
conv # 3
```

skipped cycles: 0

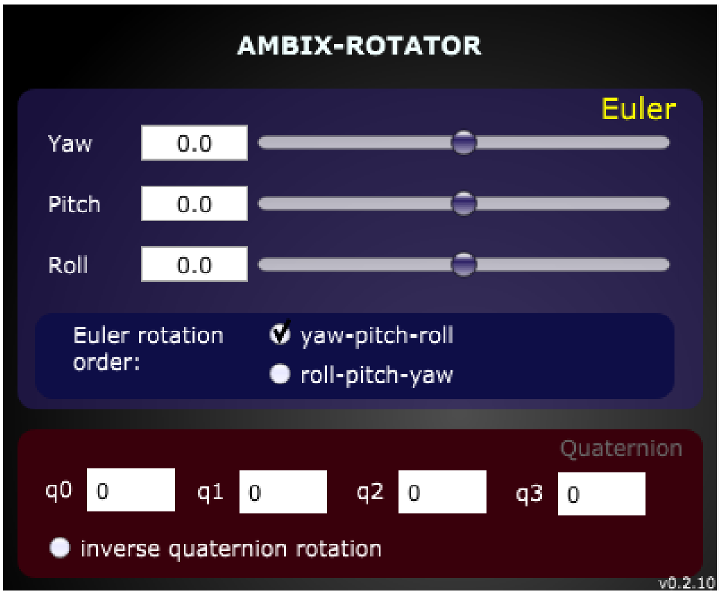
Figure 2.4.5.2 Kronlachner MCFX convolver VST [195].

A range of SIR measurement sets are available to choose from using the OpenAir impulse response library [152] which represent a range of source-receiver locations within real measured vocal performance spaces. SIR files may be manually specified to define the performer locations in the virtual acoustic scene, and are loaded into the mcfx convolver plugin using configuration files. Where SIR measurements were in legacy FuMa B-Format these SIR were converted by a simple Matlab script to AmbiX format using the standard 1st order conversion matrix [128]:

$$A_{(B\rightarrow ambiX)} = \begin{bmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \textbf{\textit{(2.4.5.1)}}$$

Configuration files were then generated specifying appropriate gain to avoid peaking with an input level of 0dB and specifying minimum (first) partition size of 64 samples (which will be automatically corrected for larger buffer lengths on SIR loading).

Local audio is kept head-locked, and all remote performer tracks are summed into a 4-channel track for rotation. The rotation is performed using the Kronlachner AmbiX Rotator VST [196] (*Figure 2.4.5.3*), which utilises the standard 1st Order Ambisonic rotation matrixes [118], [197].



Yaw 0.0

Pitch 0.0

Roll 0.0

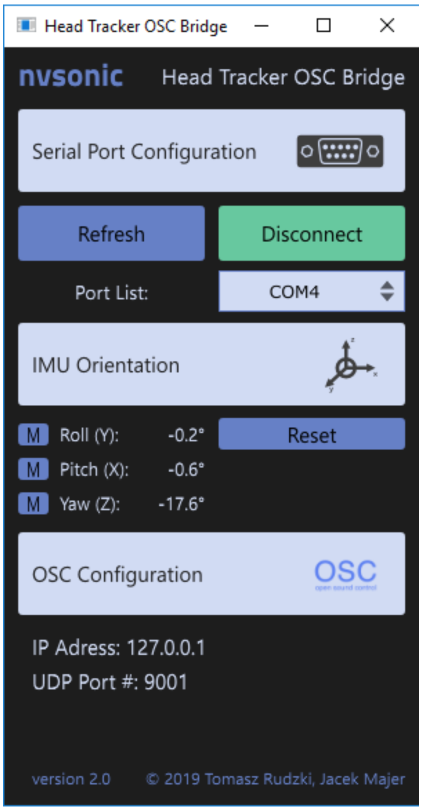
Euler rotation order: ☑ yaw-pitch-roll ● roll-pitch-yaw

q0 0 q1 0 q2 0 q3 0

● inverse quaternion rotation

Figure 2.4.5.3 Kronlachner AmbiX rotator VST [196].

Head-tracking is implemented using the nvsonic 3DoF head-tracker (*Figure 2.4.5.4*) design provided by Tomasz Rudzki [198] which uses an Arduino microcontroller design to track head rotations which are input to the local computer via serial port connection and passed to the AmbiX Rotator as yaw-pitch-roll OSC control data using UDP port 9001 [196], defining the matrix gain used in rotation of the remote performer scene around the local performer (3DoF).



Port List: COM4

Roll (Y): -0.2°

Pitch (X): -0.6°

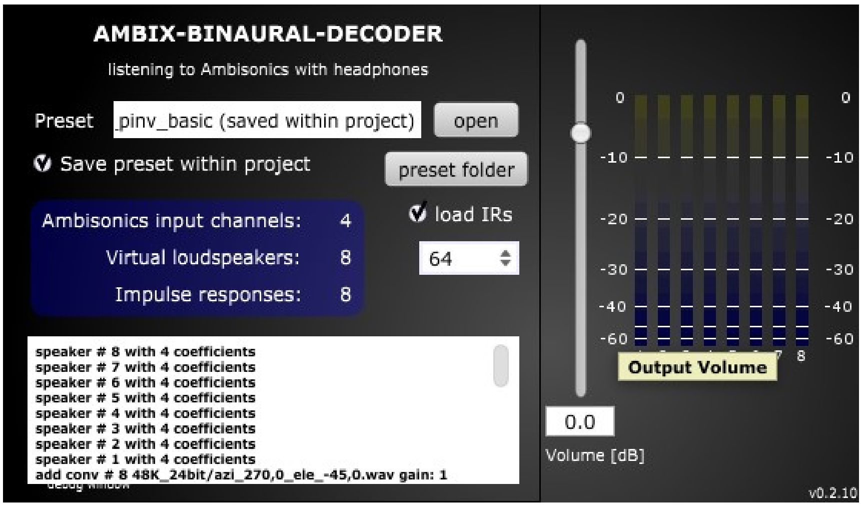
Yaw (Z): -17.6°

IP Adress: 127.0.0.1

UDP Port #: 9001

Figure 2.4.5.4 nvsonic 3DoF Headtracker OSC bridge [198].

The rotated remote performer scene and head-locked local performer Auralisation are summed into a single 1st order Ambisonic signal and passed to the master audio track which provides binaural decoding of the virtual acoustic performance scene. This is achieved using Kronloachner AmbiX Binaural Decode VST [196] (*Figure 2.4.5.5*) which uses the virtual Ambisonic approach. The virtual loudspeaker configuration and HRTF resources are taken from the state-of-the-art SADIE II Binaural Database [199]. This database contains a range of individual and generic HRTF sets for a range of Ambisonic speaker configurations. In the VIIVA-NMP audio system prototype the Neumann KU 100 [200] dummy head HRTF measurement set was utilised with a Birectangular (8 loudspeaker) virtual loudspeaker configuration using standard SN3D normalisation.



Preset pinv_basic (saved within project)

✓ Save preset within project

✓ load IRs

64

Ambisonics input channels: 4

Virtual loudspeakers: 8

Impulse responses: 8

```
speaker # 8 with 4 coefficients
speaker # 7 with 4 coefficients
speaker # 6 with 4 coefficients
speaker # 5 with 4 coefficients
speaker # 4 with 4 coefficients
speaker # 3 with 4 coefficients
speaker # 2 with 4 coefficients
speaker # 1 with 4 coefficients
add conv # 8 48K_24bit/azi_270,0_ele_-45,0.wav gain: 1
```

Volume [dB] 0.0

Figure 2.4.5.5 Kronlachner AmbiX Binaural Decoder [196].

The binaural stereo rendering of the virtual acoustic scene is then routed from Reaper to the soundcard output using the Jack router, for playback over headphones. In the prototype design Beyerdynamic DT990 headphones were used for binaural stereo playback. These open-back headphones allow for optimal spatial transparency of the singers own voice amongst the remote performer virtual sources.

This prototype implementation of the VIIVA-NMP audio system was deployed between two sites to ensure correct operation prior to any further testing. One site was located at the primary researcher's home residence in Glasgow, Scotland, and the second site that of another Audiolab postgraduate researcher's home residence in York, England. Jacktrip streams were configured between the sites following the specifications of the VIIVA-NMP audio system design, and the remote researcher pair were able to establish stable audio streaming with real-time acoustic simulation and spatial audio delivery as rendered locally at each performer with a one-way full system latency of ~28ms across the ~300km of displacement between sites.

Given that the prototype implementation of the VIIVA-NMP audio system design *(Figure 2.4.5.6)* was capable of covering over 300km when operating below optimal system specifications, and considering conservative choice of remote sites (such that typical representation of performance utilising the network provision found around small city sites is achieved) the VIIVA-NMP audio system was considered to have an appropriate range to validate the usefulness of Immersive NMP systems.

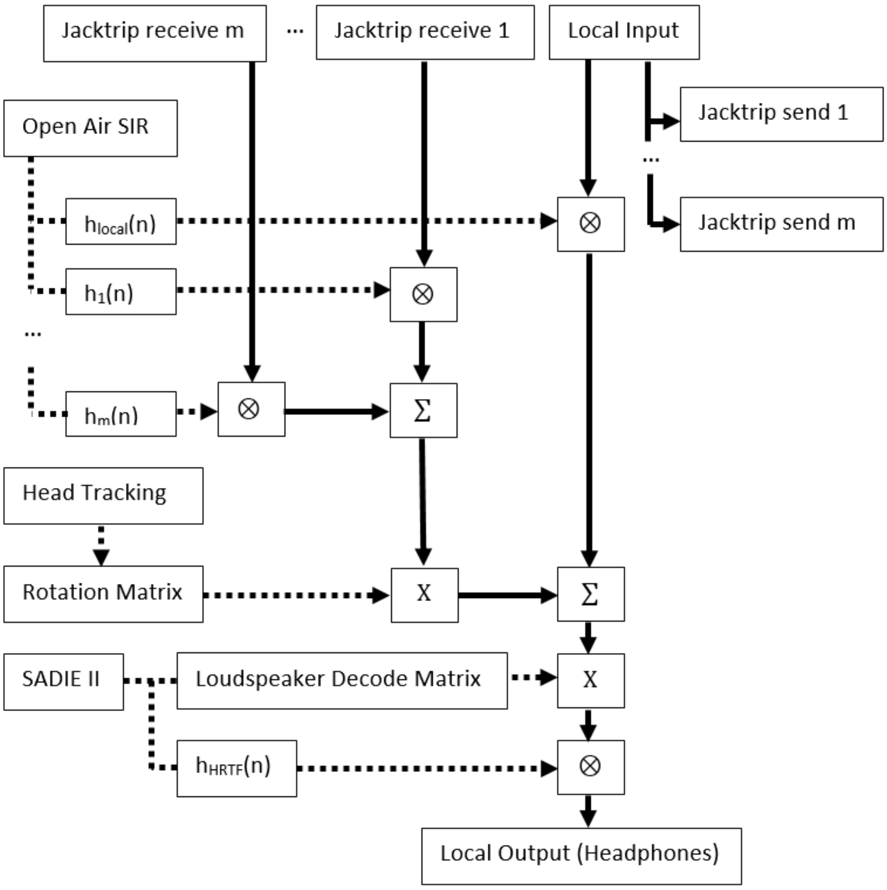


Figure 2.4.5.6 VIIVA-NMP audio system prototype implementation system diagram.

Real-time acoustic simulation with spatial audio delivery and 3DoF in the VIIVA-NMP audio system is notably of reduced technical functionality in comparison to the original VIIVA system. VIIVA-NMP utilises 1st order Ambisonics rather than higher orders as used in the VIIVA system, and does not render performer voices as directional sources as the VIIVA system does. Though the immersive audio quality in the virtual acoustic performance experience is reduced compared to the original VIIVA-system, the VIIVA-NMP audio system does still deliver a suitably immersive aural experience, which conforms to typical level of quality which is practically implemented in premium immersive audio rendering solutions for existing VR and AR technology such as Resonance [201] or Steam Audio [202]. Indeed limiting processing overhead and reducing buffer size is both integral in successful adaptation of the VIIVA design in both achieving useful Immersive NMP ranges and allowing this technology to be accessible to typical home VR/AR users.

With this prototype implementation of the VIIVA-NMP audio system design it is demonstrated that this design is practical for the home user, is compatible avatar-based visual rendering as proposed for Immersive NMP systems, can deliver a suitably immersive multi-user virtual acoustic performance experience across reasonable geographic ranges and with reasonable group size, and may be implemented practically and deployed with using existing technology.

Given the positive outlook for the VIIVA-NMP audio system design presented the prototype design was deployed further for user testing in order to help identify the challenges and opportunities, and limitations and potential of the VIIVA-NMP audio system design. The testing design is discussed in Chapter 3, with the results and discussion of system validation in Chapter 4, and conclusions and future prospects for the VIIVA-NMP audio system explored in Chapter 5.

3. VIIVA-NMP AUDIO SYSTEM TESTING: METHODOLOGY

Following preliminary testing by the lead researcher (with the assistance of AudioLab research staff), the VIIVA-NMP audio system prototype required the development of a test protocol devised to assess the limitations and potential of the system relevant to practical implementation. Several key aspects of the Immersive NMP aural experience were identified to undergo investigation. This chapter presents the testing design, associated developed resource, and relevant literature review.

It is required that the VIIVA-NMP audio system delivers a suitably immersive virtual acoustic performance experience. High immersive quality may be characterised by similarities in cognitive experience between the virtual performance and live performance [38]. Analysis may therefore be broken down into two key areas: subjective and objective comparison to live performance experience.

Subjective comparison to live musical performance provides a meter of Perceived Immersion [38] and is measured by perceptual rating of virtual performance. In modern NMP research the use of subjective rating (*Table 3.0.1*) is the primary source of measuring the subjective immersive qualities of an NMP performance experience. Robust subjective rating of immersive qualities is achieved through the use of extensive (often Likert-scale) questionnaires which assess specific aspects of the immersive group performance experience such as Presence [37], social interactivity between musicians [203], or subjective wellbeing [10] as utilised in the VIIVA system healthcare applications [9]. The rating of Immersive Audio subjective qualities [53] such as naturalness, envelopment or spaciousness may be assessed via questionnaire response, or in MUSHRA-type [204] listening test evaluation [205].

The accuracy of immersive quality assessment through self-report is, however, disputed [206]. As immersion is a cognitive phenomenon rather than purely perceptual, the ability of Perceived Immersion to reflect Cognitive Immersion is dependent on an individual's awareness of their own cognitive processes. This is apparent in NMP research, where it is not uncommon to observe examples of this such as musicians rating latency as unnoticeable up to 50ms [36], or not rating latency as having an 'above neutral' impact on the performance experience until delays well in excess of 60ms [37]. Empirical EPT research [31] can clearly demonstrate tempo deceleration and the use of latency coping strategies at such one-way delays, which presents a significant difference from musical interactions between performers in the live environment.

In order to provide a robust measure of the immersive experience, subjective metrics are often coupled with objective metrics. These objective metrics are defined in order to provide quantifiable measurements of the virtual performance experience which can be compared to measurements characterising live performance experiences. In this way such objective metrics provide discrete values which can be used to measure differences between the live and virtual experience, thereby providing a gauge of the immersive quality of a virtual experience.

Subjective Quality	• *Presence* • *Involvement, Embodiment, Attention* • *Naturalness, Realness* • *User Interface awareness* • *Coherence, Direct rating comparison to live performance*
Objective Quantity	• *Synchrony* • *Task Completion metrics (pass/fail, error %, time taken)* • *Gaze* • *Motion Capture (displacement, rotation)* • *Localisation*

Table 3.0.1 Subjective and Objective points of analysis for Immersive NMP.

In NMP the primary set of objective measures (*Table 3.0.1*) describe performance synchrony [29]. Synchrony measures detail how 'in time with each other' musicians are in ensemble performance [26]. Synchrony metrics which describe live performance may be used as comparison points in order to characterise the immersive quality of a virtual performance experience [59], and in order to characterise musical interactivity through the detection of latency coping strategies such as Leader-Follower interaction [37].

Though beyond the scope of this thesis, a range of objective measures are relevant for the analysis of Immersive NMP systems complete with visual avatar rendering. Telepresence NMP systems have introduced performer gaze [206] as a determinant of social presence in networked musical interactions. Such systems have also been used effectively to provide objective measures of Immersive Audio experiences such as localisation [207]. Quantitative task response metrics [22] for haptic and tactile technology associated with VR avatars also provide a quantifiable metric to assess the virtual experience, as does motion capture data describing the gestural interaction between musicians in live performance [208], and in telepresence-NMP contexts [206].

In order to provide validation of the VIIVA-NMP audio system prototype, the capture of subjective rating of the performance experience, as well as performance synchrony data, will be required for a practical deployment use-case. The development, completion and analysis of this test protocol addresses the sub-objectives detailed in the introduction to this thesis:

- *To introduce an immersion-driven framework for NMP audio system design, which will allow for validation of the Immersive NMP audio system presented in this thesis as suitably immersive.*
- *To conduct a pilot study in the practical deployment of audio systems for Immersive NMP and identify the challenges and opportunities presented using 'real-world' internet connections.*

The Ensemble Performance Threshold utilises performance synchrony metrics to define the latency limit at which performers will no longer experience a natural level of musical interactivity. By recording performance synchrony in a vocal performance context this project also addresses the sub-objective relating to EPT variance with instrumentation [33], [34] and the testing of EPT discretely for voice. It was also identified that reverberation may have some interplay with performance synchrony and thus EPT [28], [32], [27]. As the VIIVA-NMP system is able to provide real-time acoustic simulation of virtual acoustic environments, repeating the test protocol with varying virtual performance spaces (and reverberant qualities) allows for further investigation into reverberation in NMP in the context of spatial audio delivery and Immersive NMP. In this way the final sub-objective of this thesis is addressed:

- *To conduct pilot study of performance synchrony and EPT for vocal performance under the influence of varying virtual acoustic reverberant performance spaces*

3.1 Synchrony Measurement for Immersive NMP Vocal Performance

Establishing synchronous performance relies on all musicians in the ensemble agreeing to an acceptable extent on the temporal location of musical events [26]. This agreement regarding where musical events are to be temporally located describes a common understanding of the tempo of performance between musicians in the ensemble. In live performance this common understanding can be established through musical interactions between performers [209]. Here, the homogeneous concept of tempo amongst performers is entirely native to the live musical interaction, namely the expression and perception of musical events [26] by performers within the ensemble.

It is, of course, possible to establish a common understanding of the tempo of performance through the delivery of a direct tempo description as dictated by a score, conductor or metronome device. In such cases the primary interaction is between each discrete performer and the common rhythmic descriptor, and cannot be said to describe the musical interactivity between performers in an ensemble [44]. By sampling performance synchrony metrics using the VIIVA-NMP audio system and comparing to typical synchrony measures of live performance, a quantitative measure of immersion is provided in terms of how 'natural' the musical interactivity possible using the VIIVA-NMP audio system is [31], [28].

3.1.1 Performance Synchrony and the Perception of Musical Events

As performance synchrony metrics describe the degree to which performers agree on the temporal location of musical events, a metric which quantifies a feature of the musical event is needed from any temporal disagreement. The expressed (where the musician initialising the musical event has intended it to be placed in time) and perceived (where a musician experiencing the musical event has identified placement in time) is not without some capacity for 'human error'.

The perceived temporal location [49], or 'Perceptual Centre' (P-Centre) within a musical note is found between its sound onset and the energy peak of a sound. The P-Centre is not a discrete temporal location, but rather a frame of time within which the probability of the P-Centre existing is maximised and relatively even [48]. The perceived temporal location of where a sound begins is detailed as the Perceptual Onset [210], [211], which again is defined as a statistical window rather than discrete temporal location.

In synchrony metrics standard practice dictates the description of differences between the temporal locations of musical events in terms of onset times. As such, it is crucial that any synchrony data capture correctly identifies these onset times. The Perceptual Onset of a musical event is well known to vary significantly with expression, involving characteristics such as attack time [210], note duration [48],

[50], and spectral content [210], [48]. This presents a complication in the vocal performance context considered by the VIIVA-NMP audio system, where vocal performance includes variance in the expression and perception of onsets. In synchrony data capture it is therefore by necessary to correctly control the task in the test protocol and implement appropriate onset detection to ensure consistency in captured data and the analysis of this data.

Provided it is possible to accurately measure each onset time for each musical event performed by each musician in an ensemble, a range of synchrony metrics may be useful.

3.1.2 Inter-Onset Interval Synchrony Metrics

Synchrony within a single performer is calculated with respect to the Inter-Onset-Interval (IOI) which describes the time difference, *t(n),* between physical onsets of adjacent notes in a musical sequence [26]. This may be described mathematically as [29]:

$$IOI_n = t_{n+1} - t_n \qquad \textbf{\textit{(3.1.2.1)}}$$

Where n indicates the onset index with associated time, *t,* typically measured in seconds.

This allows the simple expression of measure of synchrony within a single part, Tempo, given as [29]:

$$\overline{\delta}(n) = 60/IOI \qquad \textbf{\textit{(3.1.2.2)}}$$

This provides the instantaneous tempo described by the two onsets detailed in the IOI. It is worth noting, however, that this specific description considers whole notes in the IOI. A more accurate tempo description can be given with inclusion of an additional factor, *k,* which describes the beat division of the note:

$$\overline{\delta}(n) = k * \frac{60}{IOI} \qquad \textbf{\textit{(3.1.2.3)}}$$

Calculation of tempo in this manner yields a set of tempo points between onsets. These tempo points may be used to identify acceleration and deceleration within parts as Tempo Slope [31], [27] or may be compared between parts in order to provide Tempo Ratio [212].

Within a part the mean IOI across the performance is defined as the Pacing [29], with Regularity defined as the associated Coefficient of Variability within the IOI vector [33]. Like tempo metrics these measures may be sampled over a moving window to allow for calculation of slope trends within a part and ratio between parts.

3.1.3 Onset Time Difference Synchrony Metrics

Given a vector of onsets for each performer in an ensemble, the synchrony of this ensemble may be expressed in terms of asynchronies: disagreement in the temporal location of onsets which are intended to occur synchronously [26]. Asynchrony metrics are therefore based upon Onset Time Difference (OTD). OTD may be expressed between any two discrete parts as an absolute OTD, or expressed across a full ensemble in terms of the relative OTD, describing the time difference between each onset and the mean onset time for the musical event in question [26] *(Figure 3.1.3.1).*

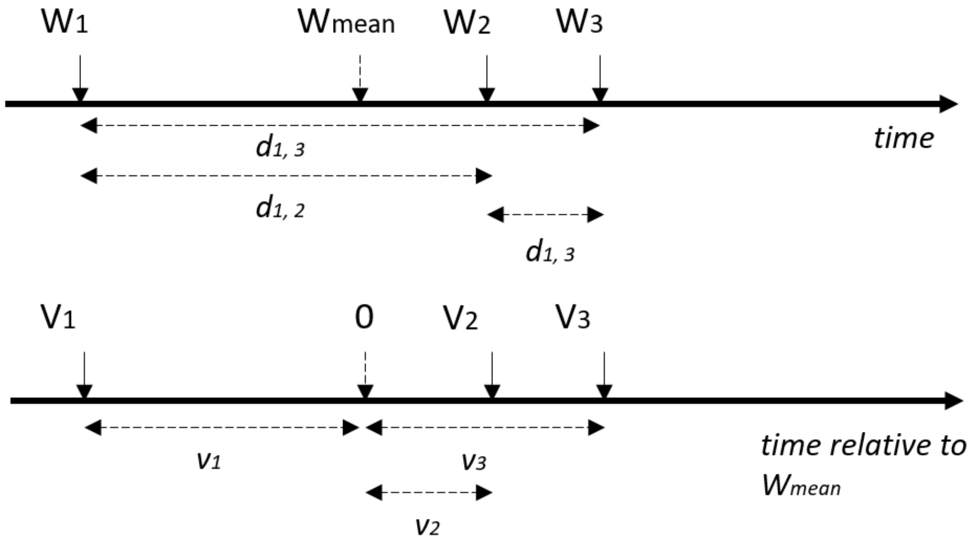


Figure 3.1.3.1 Absolute and relative OTD, adapted from Rasch [25]. W values indicate absolute onset times, where d values describe absolute OTD between parts. V values represent relative onset time as expressed relative to the mean onset time, allowing expression of asynchronies across large ensembles.

When assessing the synchrony between parts or across an ensemble throughout a musical performance, a range of metrics can be produced from absolute or relative OTD vectors. These metrics describe characteristics of the performance synchrony as single discrete values which characterise the performance in terms of OTD relationships between parts and variance across the performance.

The mean of the absolute OTDs between parts is defined in the context of vocal performance by D'Amario, Daffern and Bailes [59] as Precision. As a scalar metric this value does not provide information on any leader-follower relationships between performance parts, but simply a discrete value indicating how 'in time' the performance is.

Rasch, who provides much of the empirical research on performance synchrony, notes that the mean OTD will often present a small value which provides no indication of the state of synchrony over time [25]. Rasch states that variance of OTD is often more informative, and presents the empirical metric of Asynchrony. This is defined as a metric across a full performance ensemble, and is calculated by first finding the standard deviation of absolute OTDs between parts in the performance (or as relative OTDs for each part), then calculating the mean of these OTD standard deviations [26]. This provides a single scalar value detailing OTD variance throughout an ensemble performance. Between any two discrete parts in a performance the Asynchrony may simply be expressed as an absolute standard deviation. This metric is defined in the context of vocal performance again by D'Amario et al [59] as Consistency, which is complimented by similar calculation of Co-Variance. Farner et al. [32] define this metric in an NMP context as Imprecision.

Though asynchrony metrics are useful in providing absolute measures of performance synchrony, the expression of these metrics as scalar values removes data which can provide information on inter-performer relationships within the performance. For this reason metrics which measure the degree to which one part precedes another are relevant, which express this value as signed lead or lag. In NMP research this has predominantly taken the form of Asymmetry [29], detailing the mean of the time by which performer B lags behind performer A (where a negative value dictates leading rather than lagging). In the context of vocal performance D'Amario et al. [59] define Tendency to Lead, calculated as the median signed value by which performer A leads performer B.

3.1.4 Onset Detection

The calculation of performance synchrony metrics is reliant on the correct identification of the temporal location of onsets within parts. For this purpose, Onset Detection algorithms are used, which take audio recording of individual performers or group performance, and output a vector of onset events with associated temporal indices [213]. For single instrument recordings, as in NMP synchrony analysis, robust Onset Detection requires that a physical feature of the recorded audio signal may be used to characterise an onset, and that the algorithm can effectively recognise this feature of the audio recording. The Onset Detection process can be broken down into three discrete stages: pre-processing, reduction, and peak-picking (*Figure 3.1.4.1*).

Pre-processing functions are concerned with the preparation of the audio file for feature extraction in the reduction stage. Examples of Pre-Processing functions are: source separation [214], transformation to frequency domain [213], [215] and amplitude envelope following and modulation [213].

Reduction functions are responsible for the identification and extraction of the signal components which contain the information used to detect onsets in the audio signal [215]. These functions are described as 'reduction' as they will typically discard any part of the signal which is not useful to the identification of onsets, thereby reducing the input audio signal to onset-detection-relevant information. Reduction functions are predominantly concerned with analysis of Amplitude [213], Energy [215], Spectral Content and Phase [216], [215].

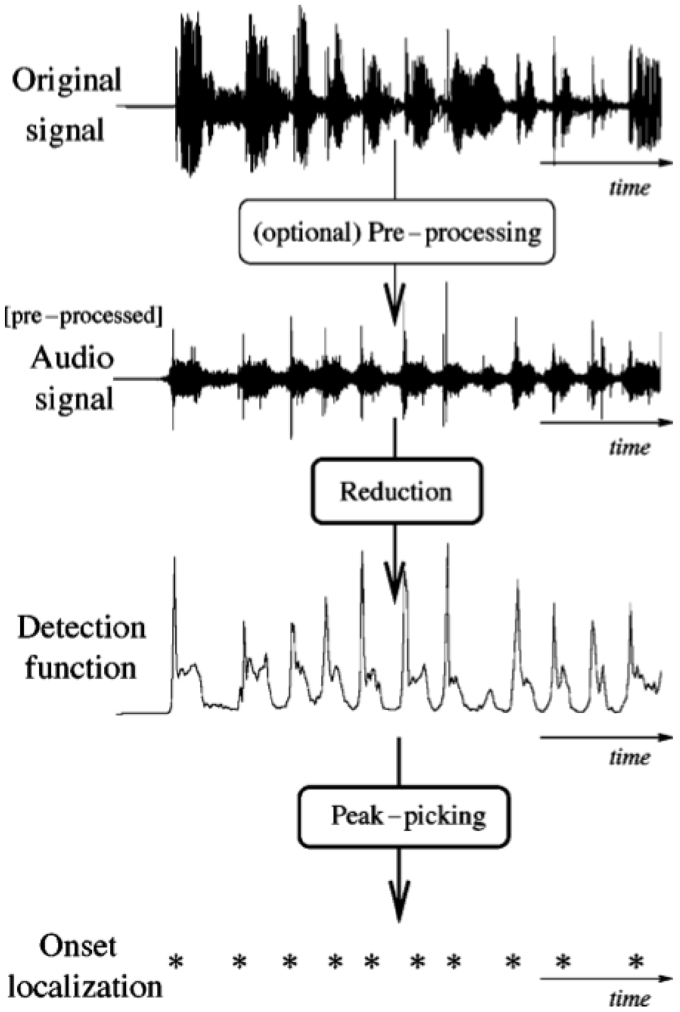


Figure 3.1.4.1 Onset Detection workflow, image sourced directly from [213].

The peak-picking component is responsible for the extraction of discrete onset temporal locations using the information passed by the reduction stage [213]. This is achieved by providing the peak-picking component a description of an onset, and a means of identifying the relevant features in the reduced signal. This will typically take the form of thresholding and the locating of maxima and minima in the reduced signal.

3.1.5 Pre-Processing Methods

The pre-processing stage of onset detection is an optional stage which may be included to improve the performance of processing later in the onset detection workflow [216]. Pre-processing functions primarily take the form of transformation to the frequency domain [213], [216], [215] in order to split the audio signal into frequency bands for onset detection based on spectral features, for the purposes of transient/steady-state separation [213], [217], or to extract pitch [218].

Transformation of audio from the time to frequency domain is accomplished through use of the Fourier Transform, typically implemented as the Short Time Fourier Transform (STFT) [219]. The STFT is computed by splitting the audio signal into frames in the time domain, then calculating the Discrete Fourier Transform (DFT), typically using the Radix 2 decimation [151], [220] (Fast Fourier Transform, FFT). This approach is typically used where the reduction component of the onset detection algorithm requires frequency-band energy analysis. Previous work has demonstrated use of this approach for onset detection based on physical spectral energy change [221], perceptually motivated systems [222], [223] which aim to emulate human hearing [210], or FFT estimation of fundamental frequency to output pitch [59].

Transient/ steady-state separation is associated with note segmentation functions in onset detection. This approach considers that onsets are characterised by transient components of the audio signal. Such transient components can be identified by fluctuation in pitch [59] or spectral energy [223]. Steady-state components are considered to characterise rests or continuing notes in the audio recording of musical performance. These steady-state parts of the signal may be identified through inclusion of offset detection to partition notes into frames and identify time frames between notes [59] or by thresholding changes in spectral energy [223]. Predictive models [224] using Linear Prediction [225] and Sinusoidal Modelling [226] can be used for transient/steady-state separation where transients can be identified by thresholding inconsistencies between the predicted signal and audio recording of the musical performance over time, considering that transients are highly unpredictable, whereas steady-state components should follow the predicted model.

3.1.6 Time Domain Reduction

Time Domain Reduction involves the processing of the amplitude envelope in order to identify onsets which are characterised by significant rises in amplitude or amplitude peaks [213]. This approach is widely used in empirical NMP latency research using 'tapping' studies [31]. Time domain reduction typically involves an amplitude envelope follower which rectifies and smooths (low-pass filter) the signal in the time domain in order to make amplitude peaks more apparent for peak picking. This process may be described mathematically, where *w(m)* is an *N*-point smoothing window of centre sample *m = 0* [213]:

$$E_0(n) = \frac{1}{N}\sum_{m=\frac{-N}{2}}^{\frac{N}{2}-1} |x(n+m)|w(m) \qquad (3.1.6.1)$$

Variations on this include local energy followers, where the sampled window of audio is squared rather than rectified. A psycho-acoustically motivated variation is to differentiate the logarithm of the envelope, in order to emulate perception of loudness in human hearing which has been shown as a relative measure of loudness [213]. Onset detection functions based on amplitude envelopes have been widely found as only acceptable for percussive onsets (such as tapping), however struggle with 'soft' onsets [227] or legato between notes [59], making this an unviable approach for vocal performance.

3.1.7 Spectral Domain Reduction

Spectral Domain reduction functions typically involve transformation of the audio signal to the spectral domain using the STFT [216] in the pre-processing stage. The spectral information of the audio signal may then be used to identify onsets through measures of sub-band energy, spectral difference measures, phase [216] and pitch [218] (or more specifically fundamental frequency). Spectral Domain Reduction is generally far more robust and effective than Time Domain Reduction for anything other than highly impulsive onsets [227].

High Frequency Content (HFC) methods of spectral domain reduction consider that as the majority of the energy of most audio signals is concentrated in lower frequency bands, then variation in high frequency bands occurs with greater relative magnitude than at low frequencies. Onsets can then be identified by analysis of change in spectral energy specific to high frequency bands, where onsets are recognised by the algorithm by peak-picking from an output HFC energy envelope. Energy envelope calculation may be implemented as such on discrete sub-bands [228], or may be weighted proportionally to frequency in order to diminish the masking of energy changes by overwhelming low frequency energy [216], [228]. Summing the weighted energy envelope provides the measure defined as the High Frequency Content (HFC) [227]:

$$\tilde{E}(n) = \frac{1}{N}\sum_{k=\frac{-N}{2}}^{\frac{N}{2}-1} W_k |X_k(n)|^2 \qquad (3.1.7.1)$$

The weighting function, W_k, is proposed as a linear weighting function [228], [229]:

$$W_k = |k| \qquad (3.1.7.2)$$

Though superior to time domain reduction, HFC reduction can still be noted as only appropriate for impulsive onsets and performs poorly for the detection of ‘soft’ [227] and legato [59] onsets [213].

Considering that transients may be identified by significant changes in energy it is possible to realise such transients by measuring energy changes. This may be accomplished by the provision of difference measures between adjacent STFT frames [216]. Differences between discrete frequency bins can be defined as the Spectral Difference [223] when calculated using the L2 norm, or Spectral Flux [216], [229], [230] when calculated using the L1 Norm. Spectral Differences describe the magnitude of difference in spectral energy between STFT frames. In implementation the spectrum is half-wave rectified such that only positive changes in energy, associated with Onset events, are considered in difference measures [213]. As with HFC reduction Spectral Difference metrics are generally only appropriate for percussive onsets [231] and perform poorly for detecting onsets with slow attacks or legato note onsets as will be found in vocal performance [59].

Spectral Difference (L2 Norm) can be given as [213]:

$$SD(n) = \sum_{k-\frac{N}{2}}^{\frac{N}{2}-1}\{H(|X_k(n)| - |X_k(n-1)|)\}^2 \qquad (3.1.7.3)$$

Spectral Flux (L1 Norm) is given as [230]:

$$SF(n) = \sum_{k-\frac{N}{2}}^{\frac{N}{2}-1}\{H(|X_k(n)| - |X_k(n-1)|)\} \qquad (3.1.7.4)$$

Where the half wave rectification is given by the function *H(x)* [213]:

$$H(x) = \frac{x+|x|}{2} \qquad (3.1.7.5)$$

3.1.8 Phase and Complex Domain Reduction

Phase-based reduction methods look for significant changes in frequency between adjacent STFT frames as estimated by the instantaneous frequency [213]. The phase, *ɸ(n)*, of a transformed signal can be considered as a coefficient of the STFT of a signal [213], [227]:

$$X_k(n) = |X_k(n)|\, e^{i\phi_k(n)} \qquad (3.1.8.1)$$

The instantaneous frequency, $f_k(n)$, is an estimate of the frequency of the kth STFT component, calculated as [213], [232]:

$$f_k(n) = \left(\frac{\phi_k(n) - \phi_k(n-1)}{2\pi h}\right) f_s \qquad (3.1.8.2)$$

Where *h* is the STFT window size and f_s is the sampling rate.

The basis of phase reduction methods is that if there is no deviation in phase then there is no deviation in instantaneous frequency estimation, and a 'steady-state' signal can be assumed [213]. Transients, associated with onsets, can then be identified by thresholding difference measures between adjacent STFT frames. A metric of phase deviation for onset detection is presented by Duxbery et al [233] is given as the Phase Deviation (PD) describing the mean absolute phase deviation across all frequency bands [227], [216], [230]:

$$PD(n) = \frac{1}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1}|\phi_k''(n)| \qquad (3.1.8.3)$$

An alteration on the PD metric is presented by the Weighted Phase Deviation (WPD) [230] where each frequency bin is weighted by its magnitude, $X_k(n)$:

$$WPD(n) = \frac{1}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1}|X_k(n)\phi_k''(n)| \qquad (3.1.8.4)$$

This metric may be normalised to give the Normalised Weighted Phase Deviation (NWPD) [230]:

$$NWPD(n) = \frac{\sum_{k=1}^{N} |X_k(n) \phi_k''(n)|}{\sum_{k=1}^{N} |X_k(n)|} \quad (3.1.8.5)$$

Phase reduction methods ultimately constitute pitch estimation, and are able to identify ‘soft’ and legato onsets with far greater success, though can be noted as less effective for Non-Pitched Percussive sounds [213].

Complex Domain (CD) reduction operates as a hybrid phase and energy difference measure between consecutive STFT frames [216], given as [216], [230]:

$$CD(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \left|X_k(n) - \widehat{X_k}(n)\right| \quad (3.1.8.4)$$

Where *Xk(n)* is the magnitude and phase of the transform of a signal as described by [227]:

$$X_k(n) = |X_k(n)|\, e^{i\phi_k(n)} \quad (3.1.8.5)$$

And $\widehat{X_k}(n)$ is the target value, where magnitude is equal [230]:

$$|X_k(n)| = \left|\widehat{X_k}(n)\right| \quad (3.1.8.6)$$

And phase conforms to the constant change relationship [230]:

$$\widehat{X_k}(n) = |X_k(n)| e^{\phi_k(n-1) + \phi_k{}'(n-1)} \quad (3.1.8.7)$$

Complex domain reduction is noted as performing well for impulsive, ‘soft’, or legato pitched onsets, but can be recognised as distinguishing poorly between onsets and offsets [213]. For this reason the CD output is often rectified an iteration described as Rectified Complex Domain (RCD) [216]. Though this method is generally effective, some complications still exist in the context of vocal performance. Expression of legato onsets with varying dynamics (for example legato from ***ff*** to ***pp)*** may be discarded or labelled offset in CD onset detection thresholding.

3.1.8 Pitch Reduction

Pitch Reduction [218] operates on a similar principle to Phase Reduction: looking for changes in fundamental frequency associated with transients. The key difference is where Phase Reduction looks for significant oscillation across multiple frequency bands as estimated by instantaneous frequency in order to observe an indication of change in fundamental frequency, Pitch Reduction provides more robust measures metering the fundamental frequency of a signal in order to directly observe variation in pitch associated with onsets [234]. The motivation of Pitch Reduction is the detection of ‘soft’ or ‘legato’ onsets where no indication of onset is present in measures of energy and phase as the note change is performed with constant power thereby leaving fundamental frequency as the only robust onset indicator [234].

Real-time pitch estimation can be accomplished using FFT methods [218]. Offline pitch detection provides much more robust measurement, and can be achieved using a range of effective algorithms such as Normalised Correlation [235], Cepstrum Analysis [236] and Harmonic Summation [237], [238].

3.1.9 Statistical and Artificial Intelligence Reduction

Statistical reduction methods aim to identify onsets through metering of deviation from a statistical model created using predictive methods [224] in the pre-processing stage. A small deviation from the statistical model represents a steady-state signal, whereas large deviation measures indicate a transient signal component indicative of an onset or offset which can be identified and sorted by the onset detection algorithm [213], [239], [240].

Recent onset detection advances have followed Artificial Intelligence and Machine Learning approaches such as Recurrent [231] and Convolutional Neural Networks [241]. These methods have been shown to provide exceptional onset detection performance in a range of contexts [231], [241]. Some limitations in the approach are, however, apparent. Machine Learning approaches rely on learning to detect onsets using training datasets. Though these datasets are widely available, producing this data requires the transcription of onset times for training the Neural Network, and machine learning approaches are thereby limited by the accuracy of onset annotation [242]. It is also recognised that musical expression can affect the performance of onset detection algorithms, where the performance will vary relative to the features extracted by reduction functions and the type of expression in performance [227]. Machine Learning methods have demonstrated better performance in onset detection for expressive performance, provided that training material includes performance or instrument-specific expressive techniques [243].

3.1.10 Onset Detection for Vocal Performance and TIMEX

As the VIIVA-NMP audio system is specified for group vocal performance, reliable onset detection for recordings of vocal performance undertaken using the VIIVA-NMP audio system is required. This will allow measurement of synchrony metrics, which may be compared to values which characterise ‘natural’ musical interactivity in system analysis.

Review of onset detection methods has demonstrated that the range of expression available to vocal performers makes even state-of-the-art onset detection methods difficult to implement for this purpose. Energy-based methods are most appropriate for percussive onsets, and will struggle with varying dynamics in vocal expression and legato onsets [213], [227]. Phase and fundamental frequency reduction presents the inverse problem, where ‘hard’ onset voicing and consonants follow an impulse-based onset model which such reduction methods struggle to handle [213]. Machine Learning presents possibilities for expressive vocal performance onset detection, however the creation of an appropriate dataset with annotated onsets, and the development, training and validation of a Neural Net onset detection algorithm is quite beyond the scope of this project.

Previous work at the AudioLab, however, has presented an effective method of onset detection for the purposes of vocal performance synchrony measurement in the form of the TIMEX [59] onset and offset detection algorithm. The TIMEX (*Figure 3.1.10.1*) system provides an extension of traditional pitch-based onset detection algorithms [218] which utilises voice science knowledge and biomedical equipment to provide robust onset detection for expressive vocal performance, specifically the capability to effectively manage legato onsets.

The individual performer's voice is recorded alongside non-invasive recording of vocal fold vibration using electrolaryngography (Lx) techniques. TIMEX then partitions musical phrases based on rests between phrases. Onsets within these partitioned phrases are then identified by fluctuation in fundamental frequency which characterise legato between note onsets in vocal performance. 'Note Endings' are identified as a fundamental frequency minimum preceding pitch fluctuation, and 'Note Beginnings' are characterised as a peak value in fundamental frequency after the pitch fluctuation, with the intermediate frame classified as legato. The pitch envelope used to identify these locations is smoothed with vibrato suppression techniques (where fundamental frequency deviation of less than 5Hz was discarded), and a threshold for fluctuation associated with legato onsets was defined (80 cents variation in the TIMEX system).

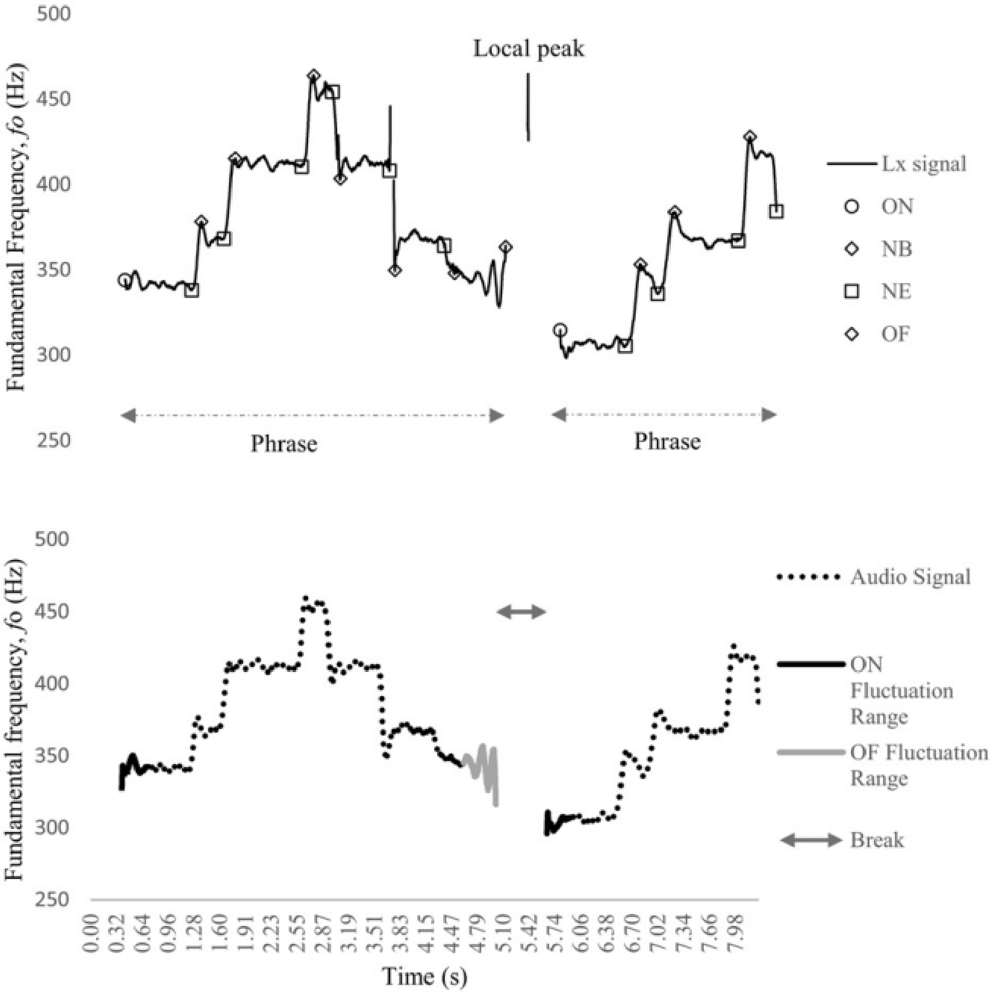


Figure 3.1.10.1 f_o profile and TIMEX onset detection event labelling, directly sourced from [59].

3.2 Testing Design: Synchrony Measurement

Review of synchrony measurement literature has facilitated the presentation of a synchrony measurement testing protocol and the development of the resource required to provide this. This allows a comparison of rhythmic interactivity between performance using the VIIVA-NMP audio system prototype and synchrony metrics characterising ‘natural’ interactivity. Where performance synchrony is acceptably similar it can be identified that the VIIVA-NMP audio system enables appropriately ‘natural’ interactivity between the performers, and where synchrony is significantly different, it is possible to establish a limitation of the VIIVA-NMP audio system and identify the causal factor for this limitation.

3.2.1 Stimulus

In order to measure synchrony a task which requires users attempt to engage in rhythmic vocal performance must be presented and undertaken. The task was a simple vocal duet of mostly coincident crotchets with additional passing quavers to ensure the singers were aiming to place musical notes together. The score uses the melody from the nursery rhyme ‘Freres Jacques’ in order to try and mediate any effect of the performer’s ability to learn melodies in testing by providing a typically familiar source. Performing as a round, with singer B entering after two bars allows for specification of certain notes in discrete parts as intended to occur at the same time (*Figure 3.2.1.1*).

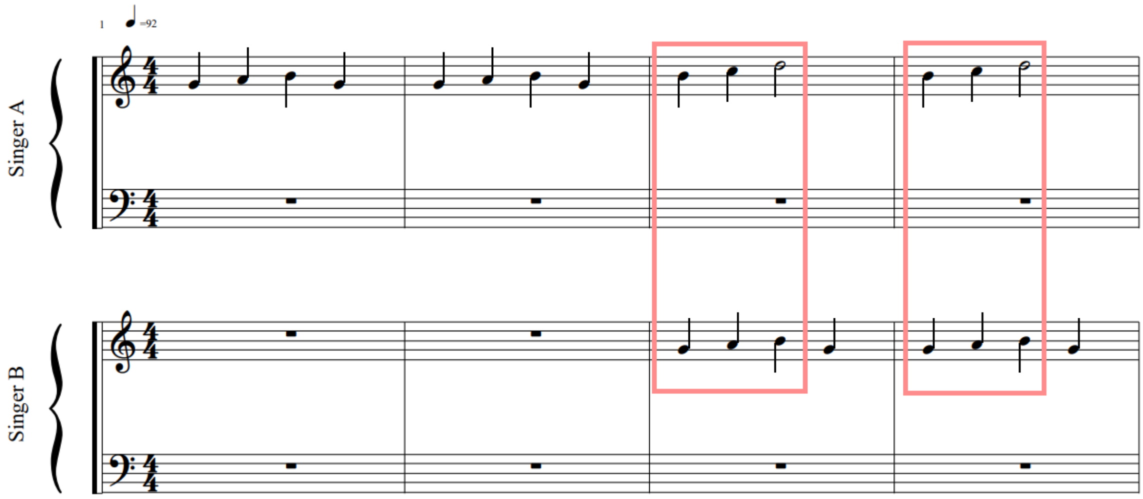


Figure 3.2.1.1 Prepared score excerpt with onsets intended to occur at the same time highlighted.

Sheet music for the score was provided, as well as an audio recording of indicative piano performance of the solo melody and duet performance. Though this material is indicatory and not used in testing, it serves to ensure familiarity. As tempo may affect the synchrony of performance in an EPT context [35] it is prudent to provide some level of control so as not to have the extreme tempo variance that could unduly influence results. The indicatory material also aims to place a loose control here by indicating an approximate desired performance tempo.

In order to improve onset detection it was decided that performers should only use the sound /ta/ for each musical note in performance. By using a voiced sound where it is possible to be confident of the approximate intended temporal location of the note onset, it is possible to provide measurement which can be confident is achieved with controlled regularity.

In this test design it is assumed that the /ta/ sound to be expressed and perceived similarly to the /tu/ sound measured in Morton's empirical P-centre study [49], as no explicit information relevant to /ta/ is readily available. Here the temporal frame in which the P-centre of the voiced sound may be acceptably placed is shown as extending from shortly after the beginning of the /t/ noise burst to shortly after the beginning of the completed progression to the pitched /u/ sound (*Figure 3.2.1.2*). Definition of the 'onset' as the beginning of the P-Centre frame affords two benefits. Firstly a common perceptual feature is identified, such that onset detection reflects perception of musical events experienced by performers. Secondly the physical characteristics of the transition from /t/ to /a/ presents a common physical feature which ensures consistency in measurement, namely the slope from /t/ noise to /a/ steady pitch. /Ta/ is used in testing rather than /tu/ as this was deemed more natural for vocal performance. It is also worth recognising that the definition of the onset of the voiced 'ta' sound in this way identifies a physical signal feature appropriately close to where state-of-the art onset detection for voice (TIMEX [59]) will place the note onset.

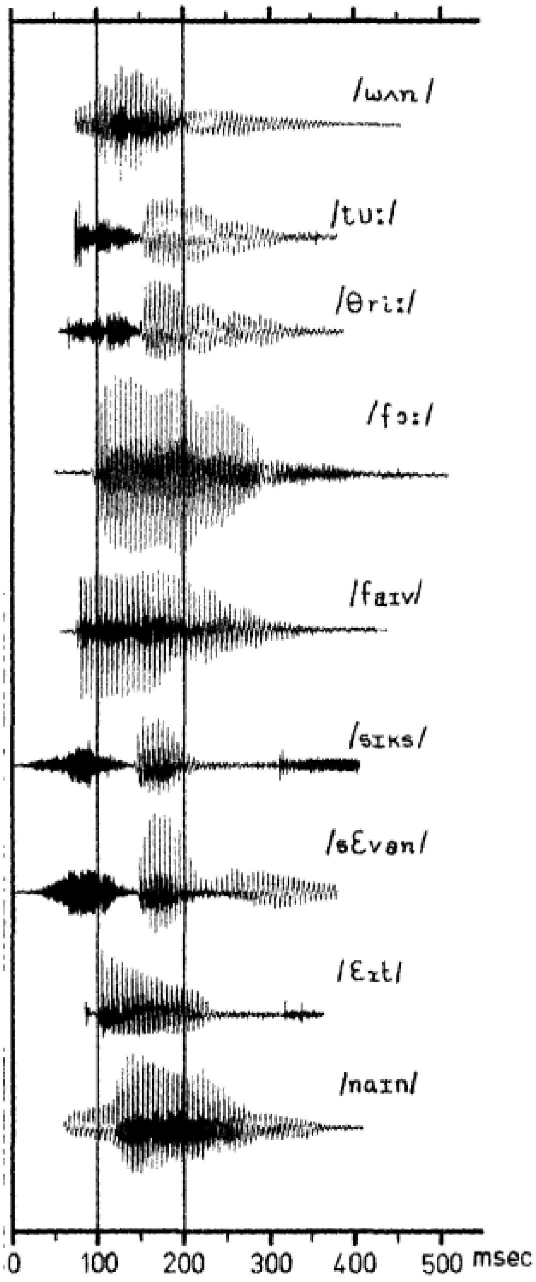


Figure 3.2.1.2 P-Centres for voiced numbers, sourced directly from [49].

3.2.2 Synchrony Metrics and 'Natural' Values

The synchrony measurement set specified for the testing protocol is specified as:

- Tempo (within parts)
- Tempo ratio (between parts) and Tempo difference
- Asynchrony
- Precision
- Tendency to Lead

For each of these synchrony metrics, values are identified which characterise 'natural' musical interactivity as measured in live or studio performance. Comparison to these discrete values then allows classification of the type of musical interaction present in performance during testing.

Tempo within parts may be expressed using simply the Inter-Onset-Intervals (IOIs) from a vocal recording of a discrete performer. As a value describing 'natural' interactivity it is expected that tempo within each part to be relatively stable (without significant acceleration or deceleration). Tempo deceleration typically characterises the effect of large delays where the natural establishing of even tempo is impaired. The acceleration of tempo, however, is associated with latencies below ~11.5ms, described by the 'Chafe Effect' [42] in NMP contexts [31]. Latency will undoubtedly be well beyond this acceleration threshold in any practical instance of Immersive NMP deployment. If tempo acceleration is observed above this 'Chafe Effect' latency region it is possible that this is indicative of latency-compensation [31], where a remote performer has noticed a significant tempo deceleration due to latency and is adjusting their own performance to correct this [28]. In either instance (acceleration or deceleration) a significant bias is indicative that latency may be impairing the interactivity between performers, such that the performers are not achieving a 'natural' level of interactivity.

Tempi Ratio [212] may simply be expressed as ratio between tempo measurements within parts for each performer in the duet vocal performance task. Tempi Ratio is expected to be relatively even (around one) in live performance, providing a value to characterise 'natural' interactivity. In itself, Tempi Ratio does not fully characterise the effect of latency, as duet performers both experiencing latency induced tempo deceleration may still provide even Tempi Ratio provided both performers are decelerating at relatively equal rates. The observation of significant bias in Tempi Ratio (positive or negative) does, however, indicate that performers are struggling to achieve stable tempo through their own interactions, which may be due to latency inhibiting the rhythmic interactions between performers. A variation on this metric is Tempo Difference, detailing the distance between tempo BPM as measured for each performer.

Asynchrony [26] provides the empirical measure of performance synchrony, allowing a common point of comparison with other NMP research and studies on synchrony in live performance. As the context presented is vocal duet performance, the definition of Asynchrony is here simply the standard deviation of the absolute OTDs between parts (as taking the mean of a single value is meaningless). Asynchrony is expressed as a single scalar value in milliseconds. Rasch [25] observes typical asynchronies for live performance from vocal duet performance as falling into the 30-50ms range. D'Amario et el [59] provide the same measurement under the definition 'Consistency' in the context of vocal performance, demonstrating values in the range of approximately 30-40ms where vocal duos maintain visual contact, and approximately 40-50ms where vocal duos have no visual contact. In this way it is possible to define Asynchrony in this range as indicative of a natural level of interactivity between performers. Asynchronies near the upper range or in excess of the upper limit will in turn demonstrate that the level of interactivity between performers is impaired in some way.

Precision is defined as the mean of the absolute OTD between parts, and anchor values for live vocal pair performance are available in previous study by D'Amario, Daffern and Bailes [59]. Here the Precision of onsets in this context are observed in the 50-70ms range with no visual contact. This target value provides an indicator of natural interactivity in the same way as Asynchrony.

Acceptable performance synchrony can, however, be achieved using latency management techniques, the most common of which is the leader-follower method. Though leader-follower relationships in performance can be native to live musical interaction, it is common in NMP to see a shift in this relationship associated with one or more performer adapting their performance in response to becoming aware of the effect of large latencies. Though this is recognised as a significantly different cognitive experience from 'natural' musical interaction, performances which use this technique may still achieve suitable Precision and Asynchrony, and indeed even tempo slope and tempi ratio.

For the purpose of detecting this latency coping technique and thereby identifying impairments in immersive quality of performance, Tendency to Lead is included as a synchrony metric. This provides the median signed OTD for each performance. As the performance task is specified with performer A beginning before performer B a bias for performer A to precede performer B is established in onset times, though in such a way that performer B should be able to engage in the performance initiated by performer A with reasonable synchrony. Significant shift in this bias (where Tendency to lead outputs substantially negative values or shifts to extreme positive values) is indicatory of the use of latency-coping strategy in performance. Observance of Tendency to Lead in vocal duet performance by D'Amario et al [59] demonstrate that Tendency to Lead should likely vary between approximately -20ms to 40ms with no visual contact.

3.2.3 A TIMEX-Inspired Onset Detection Method

In order to calculate these synchrony metrics it is required that robust onset detection can be achieved using audio recordings of each performer attempting the vocal performance task. The performance task itself is specified such that the onset can be identified by a consistent physical feature of each performed note, namely the pitch contour from the /t/ noise burst transient to the /a/ pitched note steady state.

The literature review has demonstrated that energy based onset detection methods will likely place onsets at the start of the /t/ sound (the non-pitched percussive component), which can vary with performer expression, and may place the onset at too early a point in time with respect to the intended perceptual note placement. Phase, complex, and raw-pitch design will likely place the onset on the beginning of the /a/ steady state pitched component. Though this presents a consistent feature, this feature is more aligned with the centre of the perceived note (from P-Centre literature [49]), rather than the beginning of the sonic event defined by the onset.

For our specific purpose the TIMEX [59] algorithm presents a promising design form, where voice science knowledge allows the identification of a common physical characteristic of voiced onsets through pitch-based reduction methods. The TIMEX algorithm design does, however, present a significantly different context from that of this project, and implementation of a TIMEX-inspired onset detection algorithm requires some design modification.

Firstly TIMEX uses biomedical feedback in source separation. As in testing isolated individual voice recordings are made there is no need for this and only audio recording is used. Secondly, consideration of conducting testing from participant homes requires that the onset detection method is resilient to false negatives from environmental and accidental noise which would generally be controlled in lab conditions.

Firstly the original TIMEX implementation uses biomedical feedback, which requires resource which is not available to the typical home user expected for the VIIVA-NMP system. The biometric feedback is used to provide pitch tracking alongside the audio pitch tracking function. In order to provide a practical adaptation, it will be necessary to use only the audio recordings which are easy to achieve when distributing testing to participants performing from their own homes. Practical recordings of the performance task may need to be manually 'cleaned up' to remove such recording artefacts, and robust note partitioning is required to ensure that onsets labelling occurs within the correct regions of the performance.

With respect to 'real-world' use case testing it is also worth considering performer proficiency. The TIMEX algorithm is assessed in the context of practiced singers, who are able to deliver a high quality of performance. The VIIVA-NMP audio system testing considers the typical home user, where or participant roster contains amateur to professional level singers. As the adapted onset detection algorithm must manage this varying singer proficiency, labelling pitch peaks as onsets may be subject to error due to pitch correction in the delivered performance, due to the lack of control on singer proficiency in this study. In-practice, pitch slope provides a more robust feature which will be consistent in the contour from the /t/ noise burst to the /a/ pitched note as expressed by singers of varying proficiency.

The developed onset detection algorithm, with appropriate modifications from the original TIMEX design, is hereon referred to as 'TIMEX-Lite'.

3.2.4 TIMEX-Lite Onset Detection Algorithm

The TIMEX-Lite onset detection algorithm was developed as a MATLAB [244] function (see Appendix B). This function allows the selection of a mono audio file, where the audio file contains recording of a discrete performers' attempt at their part in the performance task. The algorithm then outputs a text file containing a list of onset times for the given performance as required. The process for achieving this can be broken down into several stages:

1. Manual pre-processing of the recorded audio files in DAW.
2. Noise removal
3. Note partitioning
4. Pitch reduction
5. Peak detection, peak sorting and Onset Labelling

3.2.5 Manual Pre-Processing

As the deployment of the testing protocol in a practical use case will involve lack of control over environmental noise, it will be necessary to remove any environmental noise or performance artefacts captured in recording to avoid erroneous onset labelling. During notes such recording artefacts can be disregarded by the onset region windowing function of the TIMEX-Lite algorithm. Between performed notes environmental noise or performance artefacts will cause false positive type errors. For this reason it was required to manually remove environmental noise and performance artefacts between notes manually in DAW (*Figure 3.2.5.1*).

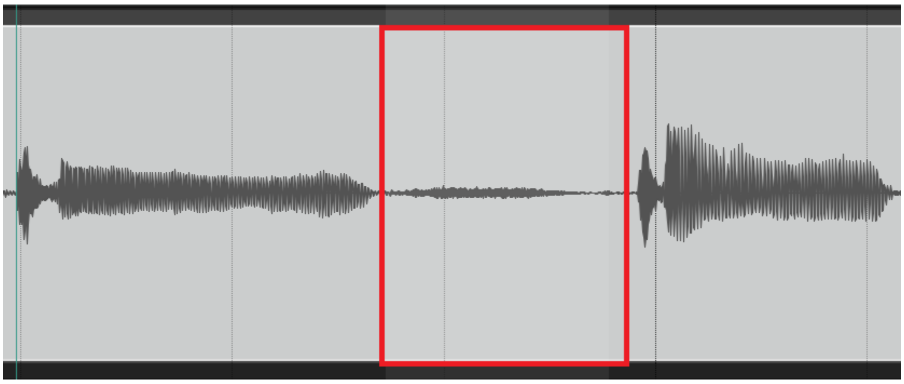


Fig 3.2.5.1 Example of performance artefact between notes (breath) which will cause amplitude-based note partitioning component of TIMEX-Lite to label a 'note on' region and search for an Onset to label. The manual removal of this part of the recording does not affect the recorded subsequent note, and therefore will not interfere with the Onset Labelling function.

3.2.6 Noise Removal

The noise removal component of TIMEX-Light simply consists of two hard gates (attack and release of 1 sample). The threshold for each of these two gates is specified as an independent input parameter when calling the TIMEX-Lite function. The output from the first gate is used by the note partitioning component. The second noise gate output is used by the Pitch Reduction component. Primarily these gates serve to remove recording and environmental noise from audio and facilitate note partitioning. It is recognised that this may affect onset labelling in the instance that /t/ is voiced with a slow attack. The nature of the /t/ sound itself, however, is designed to make a slow attack onset extremely unlikely.

The Note Partitioning component windows notes on the basis of 'on/off' regions, where it is specified to begin labelling 'on' regions a time frame prior to the sample triggering the 'on' toggle. Allowing for discrete gate values means the Pitch Reduction gate may have a slightly higher threshold than the Note Partitioning gate. This allows for the windowed region in which pitch is tracked to be slightly smaller than the note partition window, providing the practical functionality of removing a region at the beginning of the Note Partition window where any pitch reduction will simply detect noisy fluctuation which may cause errors in Peak-Picking and Onset Labelling.

As vocal performance typically exhibits a degree of variance in expression, even within the controls presented by the performance task, it can be expected that slightly different threshold values for these gates will be more or less appropriate depending on the delivery of the recorded performance. Allowing for some minor variation in the gate values does present a minor lapse of control in the onset detection algorithm. This component, however, simply affects the windowing of search regions for onset labelling, and not the onset labelling function itself. As such it should not affect the Onset Labelling function other than to filter out signal components which may cause an error. In practice gain thresholds ranging from -40dBFS to -20dBFS were effective in accomplishing this. As the nominal level of each performance was not monitored, this does present some lapse in control in this project, however the onset detection method did still work with appropriate accuracy, as detailed in Section 3.2.10.

3.2.7 Note Partitioning

The Note Partitioning component is responsible for dividing the audio recording into windows, where each discrete window is either labelled as 'On', where a performer is producing a musical note, or 'Off' where there is a gap between notes. This is achieved by cycling through the audio file samples. The default value for the Note Partitioning toggle is 'On'. At each zero crossing in the audio file a window of a specified size is taken as a sample from the audio file with the zero crossing as the start sample. This window is compared to a window of zeros of the same size. If the two are not equal then the event is discarded and the algorithm proceeds. If the windows are equal then the algorithm defines an 'off' region of the window size and proceeds. The size of the window of zeros which is used to define a 'note off' region is specified as a parameter when the TIMEX-Lite function is called.

This note partitioning scheme leverages the design of the performance task. Between each performed /ta/ sound the change from /a/ to the beginning of a new note /t/ involves a period of low amplitude in voice between the notes to form the new /t/ sound (*Figure 3.2.7.1*). By setting the Note Partitioning gate threshold to remove these low-amplitude signal components it is possible to identify these regions and thereby partition the notes, with each note window beginning a zero-mask minus one number of samples prior to the note, and ending on one sample before the subsequent 'off' region.

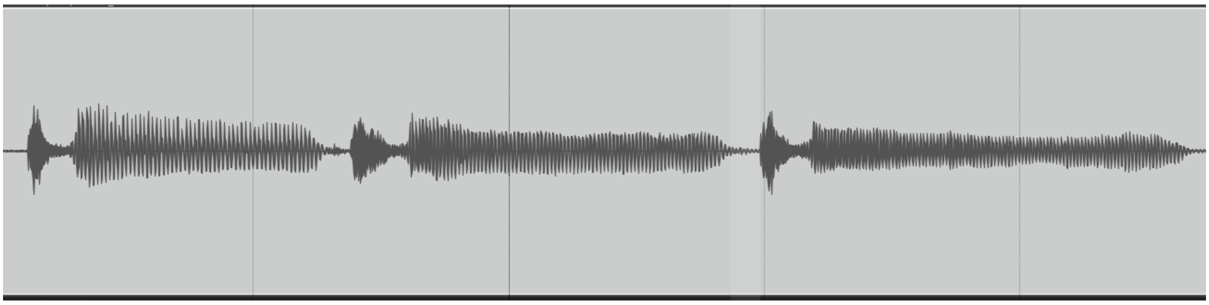


Figure 3.2.7.1 Performance task recording excerpt, where low-amplitude periods between each note can be observed.

In this manner it was possible to create a set of frames, each representing a note in the performance, which can be described by a vector of binary sample values across the audio file (*Figure 3.2.7.2*). Within these frames the pitch-reduction method is used to search for, sort, and label onset locations.

It was noted that the expression of the /ta/ sound by the performer could still cause some issues at this stage. If the gate threshold was too high (or conversely the performance soft), or in instances of vocal fry, it is possible for this partitioning scheme to place note off regions in the middle of notes.

Conversely, if the gate threshold is too low then noise in recording can cause the partitioning method not to label 'off' regions between notes as zero windows of the specified size are not present.

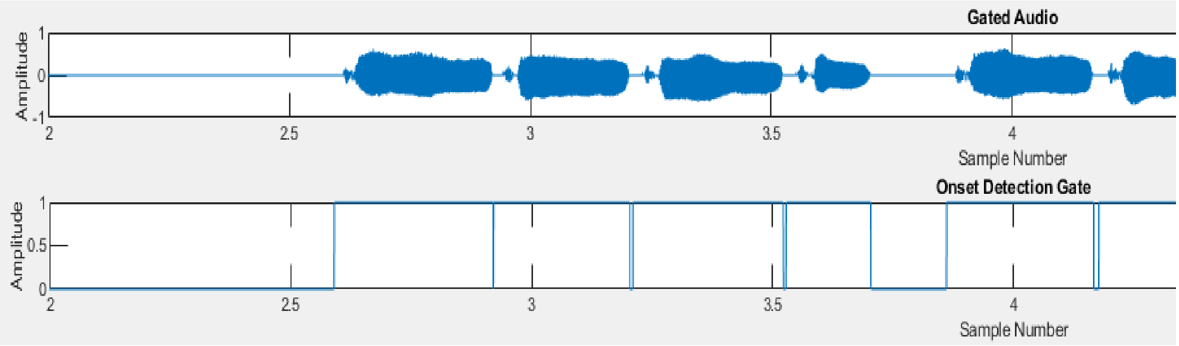


Figure 3.2.7.2 Gated Audio and associated Onset detection Gate with a 'off' region window size definition of 40ms (0.04)

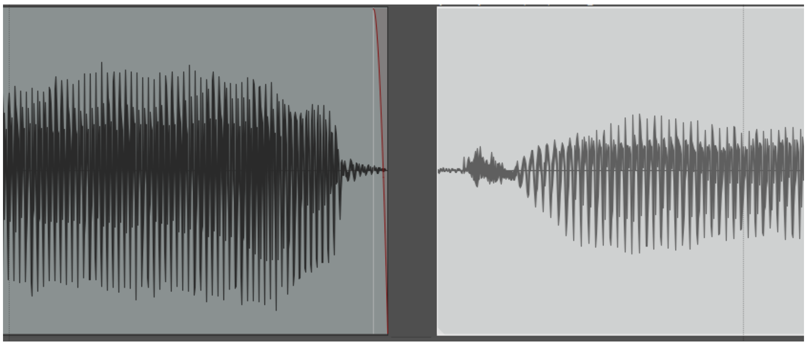


Figure 3.2.7.3 Example of manual audio file trimming between notes, such that a region of low amplitude of sufficient length exists between the trimmed region and next 'on' region. As a low amplitude region which will be removed by the noise gate in onset detection still exists prior to the next onset this trim only affects on/off state change, and off/on state change and onset detection processes are unaffected.

In practice it was found that consistent use of an 'off' region window size ranging 40-80ms was appropriate. Even with this allowance for variance in performance delivery the range of expression in vocals still presents some performances where this note partitioning scheme will present errors. In these cases rather than throw out data, it has been considered more appropriate to manually trim the audio recording at the offset of a note such that a gap of at least 40ms is present (*Figure 3.2.7.3*), where that the audio samples after this trim and before the next note will be zero after thresholding by the noise gate. This forces the note partitioning algorithm to operate correctly. Though another aspect of lack of control in the onset algorithm, only the offset of notes in the audio file are edited. This affects only the note partitioning with respect to the end of a note (the 'off' region label) and not the detection of any 'off' to 'on' region transitions in partitioning, or any Onset Labelling function. As such the toggle to 'on' to 'off' only needs to be an approximation such that the toggle value is registered for detection of the next ‘on’ state and the associated Onset Labelling. The Onset Labelling function itself remains unaffected by variance in the ‘on’ to ‘off’ state value which was edited in this case.

3.2.8 Pitch Reduction

The pitch reduction (*Figure 3.2.8.1*) component takes audio passed from the pitch tracking noise gate. Pitch tracking is achieved using the Normalised Correlation Function method [235] using a window size of 52ms, an overlap of 42ms and a median filter averaging of the estimated pitch with a filter length of 3 samples. The algorithm then zeros any locations with non-value samples (regions where the noise gate has removed all signal). Simple vibrato suppression is then applied in the form of a basic ‘pitch snap’ which rounds each pitch sample to the nearest musical note.

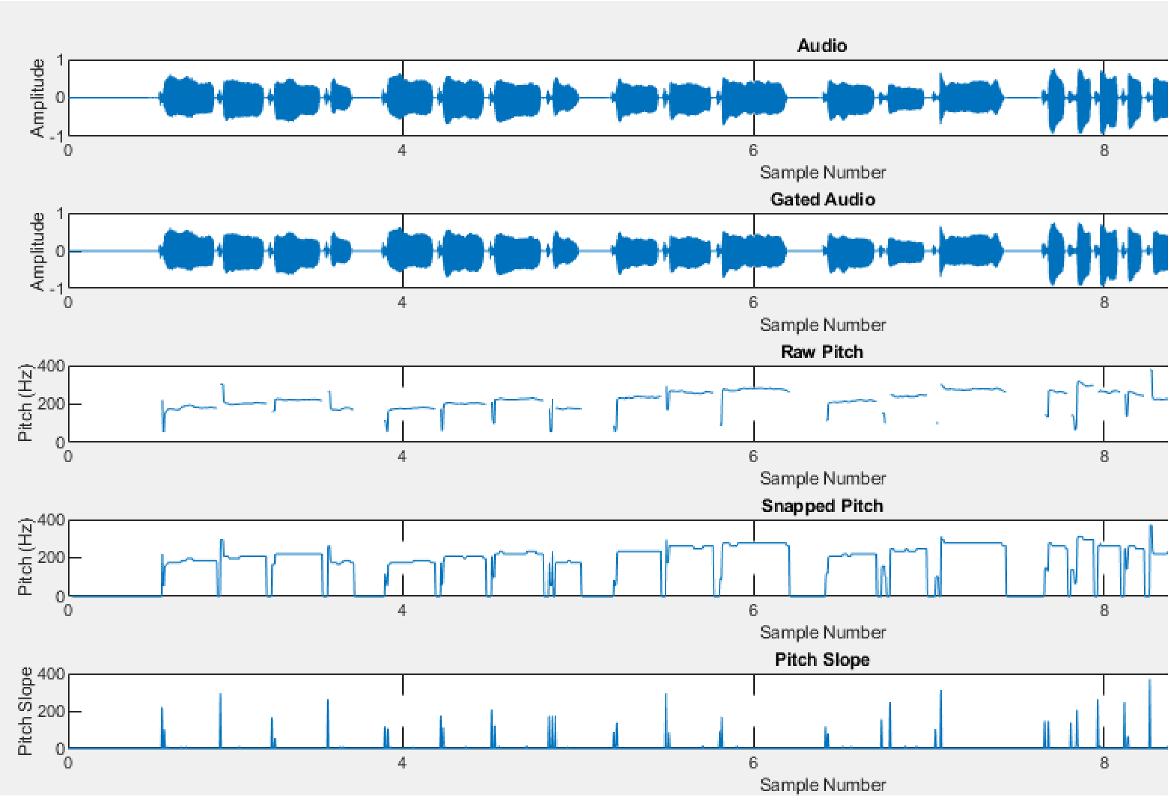


Figure 3.2.8.1 TIMEX-Lite Pitch Reduction Process plotted in MATLAB.

It was identified that the selection of pitch peaks may cause a degree of error due to pitch correction in the performance of non-professional singers. Instead the algorithm looks for the common feature of the pitch contour between ‘t-‘ and ‘-a’. This contour is best identified by a high value of pitch slope, and as such the next stage of the pitch reduction component is to calculate the pitch slope by taking the derivative of the estimated pitch. The calculated pitch slope is then further smoothed using a Savitzky-Golay filter [245]. The pitch slope trace for the audio file is then subjected to thresholding, where a default value of zero is used, such that the algorithm discards negative slopes (associated with offsets), and retain only positive slopes (where large values are associated with onsets). This output pitch slope trace for the audio file is then used to identify an onset time within each partitioned note.

3.2.9 Peak Picking

The pitch slope trace is input to MATLABs peak detection function to return a vector containing pitch slope maxima across the audio file. Within each note partition the TIMEX-Lite algorithm windows the pitch slope trace. The 3 peaks closest to the start sample of the note partition are then selected, and these peaks are then tested as being within a specified time window of the first sample of the note partition. This time window may be manually specified when calling the TIMEX-Lite function, and a default of 200ms is suggested. Any peaks which do not fall within this window are discarded, and the algorithm then picks the largest pitch slope peak from the 'first 3 peaks' selection from each note partition. This largest pitch slope is then identified by the algorithm as the contour between the 't-' onset and '-a' pitched note, and labelled as an onset (*Figure 3.2.9.1).* This process is repeated for each note partition in order to provide a vector of onset times for the audio file.

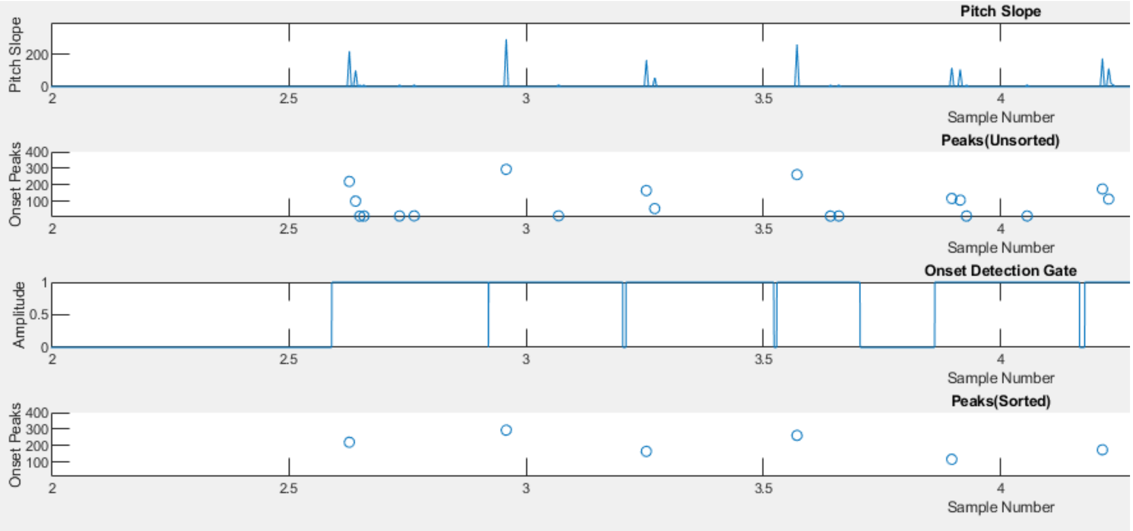


Figure 3.2.9.1 Peak-Sorting and Onset Labelling process within the TIMEX-Lite algorithm.

3.2.10 Timex-Lite Operation and Performance

The TIMEX-Lite algorithm may be called as a MATLAB function (see Appendix B). The input parameters when calling the function are:

- noiseThresh1: Detailing the amplitude threshold of the first noise gate (value between 0 and 1) which is used for note partitioning.
- noiseThresh2: Detailing the amplitude threshold of the second noise gate (value between 0 and 1) which is used for pitch tracking.

- peakThresh: Detailing the pitch-slope thresholding value, recommended as set to 0 to remove negative slopes.
- zeroMaskSize: Detailing the length of the window used to detect regions of zero amplitude in the audio signal which are used to partition notes. Specified in seconds with a recommended setting default of 0.04.
- peakSearchWindowSize: Detailing the length of the window from the beginning of each note partition in which the algorithm will consider peaks in the peak-sorting component. This value is specified in seconds and a recommended default value is 0.2s

The algorithm will then open a window where the user may select the mono audio file to be input. The algorithm will then take a short time to run and will output a graph detailing each stage of the onset detection process (*Figure 3.2.10.1*), and save a text file named 'onsets.txt' which contains a list of onset times for the audio file.

Onset detection performance is expressed in terms of Precision, *P*, (a ratio of correct, $N_{correct,}$ to erroneous detected onsets, or 'false positives', $N_{false\ positive}$) and Recall (a ratio of correct to undetected onsets, or 'false negatives', $N_{false\ negative}$). These metrics of onset detection accuracy are calculated with a +/- tolerance, and are surmised in the F-Measure, *F*, (the "harmonic mean of precision and recall" [59]).

Precision, Recall and F-measure may be calculated respectively as [59]:

$$P = \frac{N_{correct}}{N_{correct} N_{false\ positive}} \qquad [3.2.10.1]$$

$$R = \frac{N_{correct}}{N_{correct} N_{false\ negative}} \qquad [3.2.10.2]$$

$$F = \frac{2PR}{P+R} \qquad [3.2.10.3]$$

Three single vocal part recordings of the performance task presented in this thesis were taken as random samples for the purposes of providing sample validation for the TIMEX-Lite algorithm. As with the TIMEX F-measure an acceptable error of +/- 50ms is allowed. Onset times were manually annotated by the lead researcher. From these 3 random samples a total of 96 onsets were detected, with 3 false positives, and 0 false negatives, where a +/- 50ms tolerance was allowed. This indicates a Precision of 96.875%, Recall of 100%, and F-measure of 98.413%. Though far from a robust formal analysis, this is enough to demonstrate that the TIMEX-Lite algorithm appears to be suitably accurate for the purposes of synchrony analysis in this testing protocol and will identify onsets with consistency.

It is, of course, worth noting that this level of accuracy is entirely reliant on the exclusive performance task, which is designed to elicit the specific characteristics the TIMEX-Lite onset detection algorithm looks for. This level of accuracy is also dependant on the thorough manual preparation of audio files by the author exclusively for the purpose of input to the TIMEX-Lite onset detection algorithm. In the development of the algorithm it was recognised that if audio input was left unprepared that more false positives were measured (primarily due to environmental noise) and more false negatives were measured (primarily due to failure to recognise note ends in note partitioning). It is also worth considering that this performance relies on parametric control of the TIMEX-Lite algorithm. Though variation of onset detection parameters such as gate thresholds are limited to attempt not to unduly influence the onset detection process, this does present a point of lack of control.

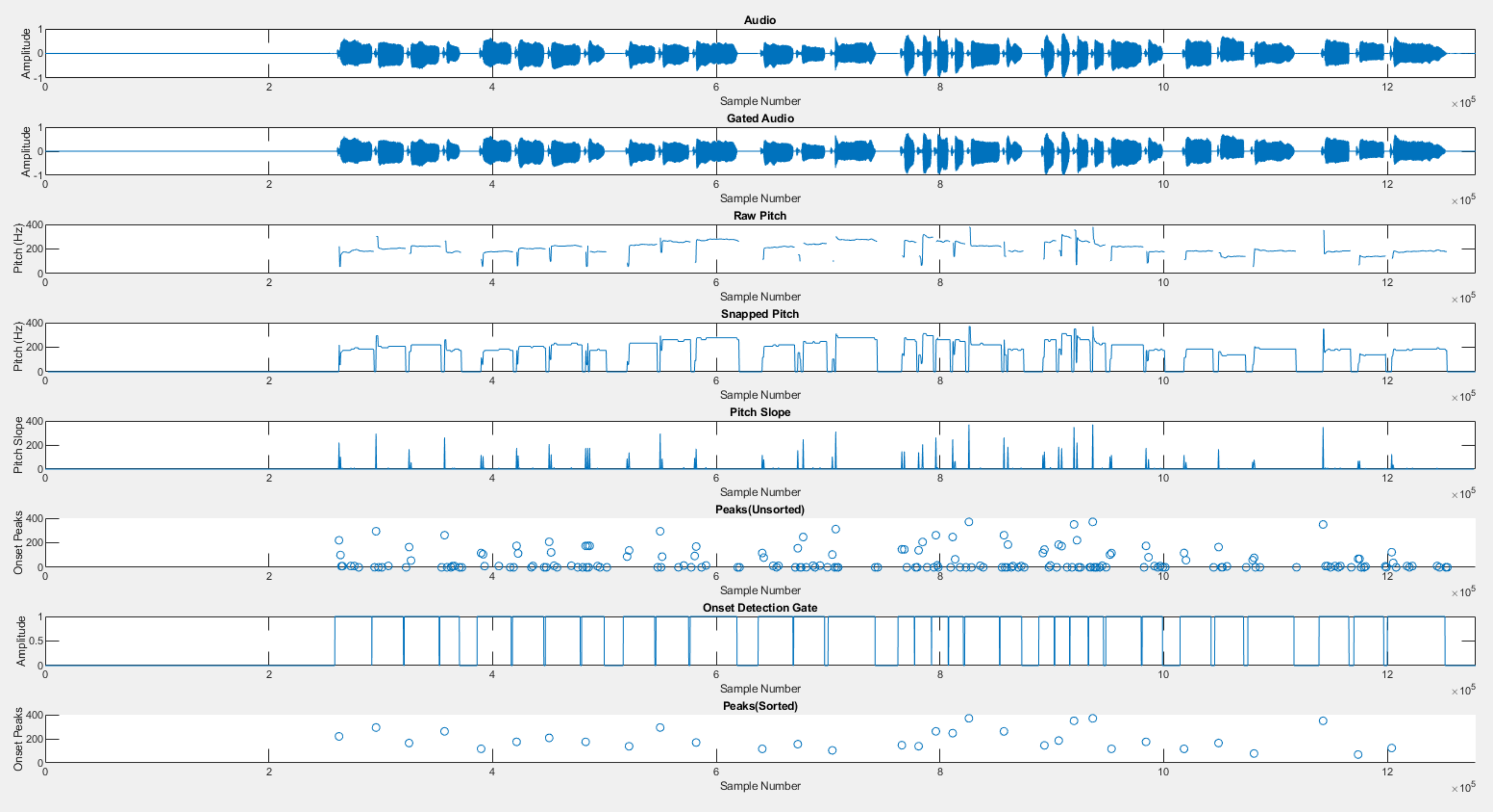


Figure 3.2.10.1 Graphic output from TIMEX-Lite

3.2.11 Synchrony Analysis Script

The list of onset times for each performance output from the TIMEX-Lite onset detection MATLAB function are used to provide the synchrony metric outputs required for analysis of performance using the VIIVA-NMP audio system prototype. This synchrony analysis is accomplished using two simple MATLAB functions. The first function provides tempo analysis of the performance task, and the second function provides synchrony analysis of the performance task.

The tempo analysis function (see Appendix D) takes as input two text files containing onset time lists for each discrete performer as output from TIMEX-Lite. These input text files are specified when calling the tempo analysis function. Mean tempo slope for each performer and mean tempi ratio between the two performers is saved in a text file named 'Tempo.txt'. Tempo (*Figure 3.2.11.1*), Tempo Slope (*Figure 3.2.11.2*) and Tempi Ratio (*Figure 3.2.11.3*) are then plotted. Tempi Ratio is calculated using only onsets from each performance which are intended to occur at the same time, whereas tempo and tempo slope are calculated using the Inter-Onset-Intervals within each part.

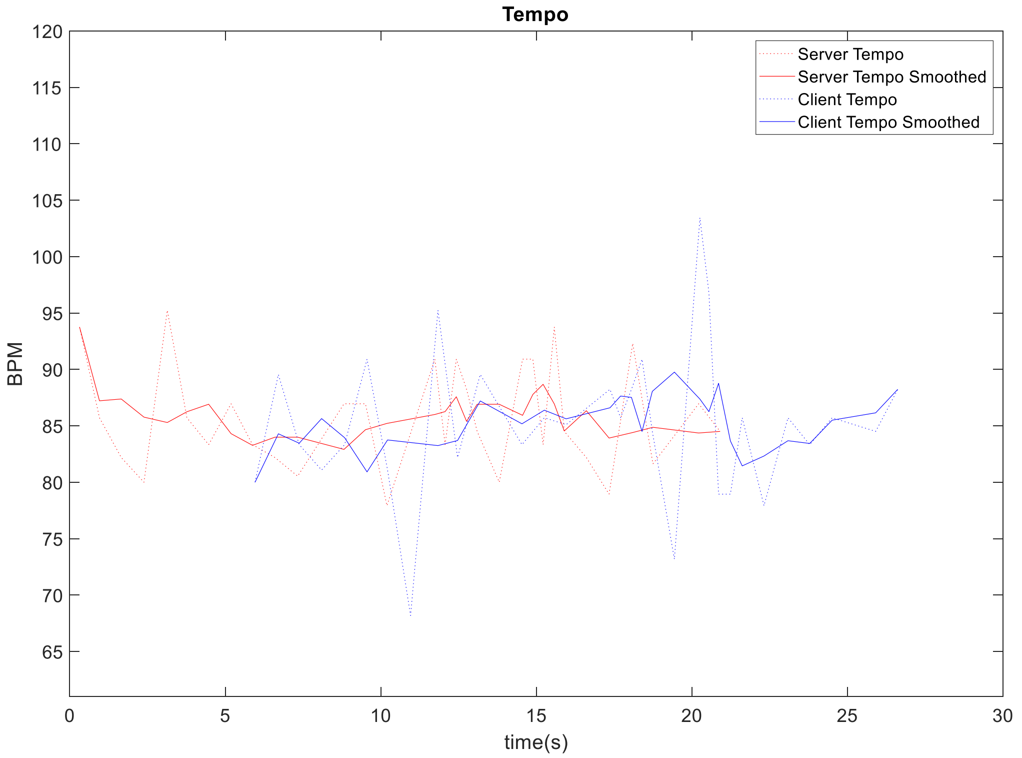


Figure 3.2.11.1 Tempo across performance. Showing tempo and smoothed tempo for performer A (server) and performer B (client) using the VIIVA-NMP audio system prototype.

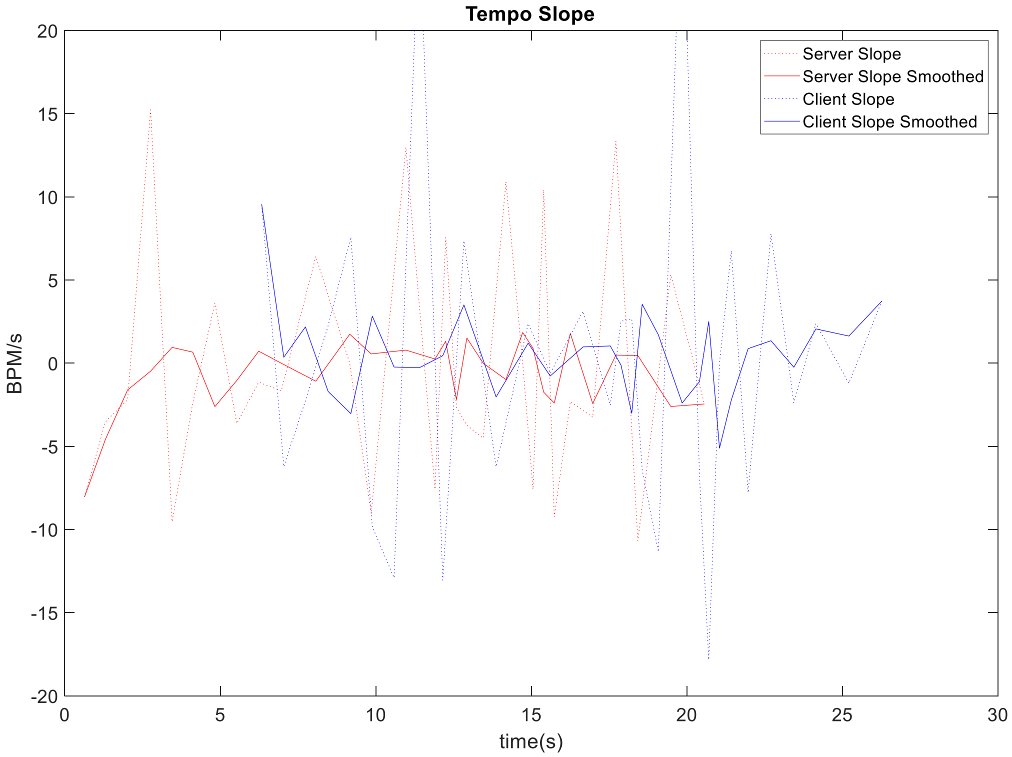


Figure 3.2.11.2 Tempo Slope across performance. Showing tempo slope and smoothed tempo slope for performer A (server) and performer B (client) using the VIIVA-NMP audio system prototype.

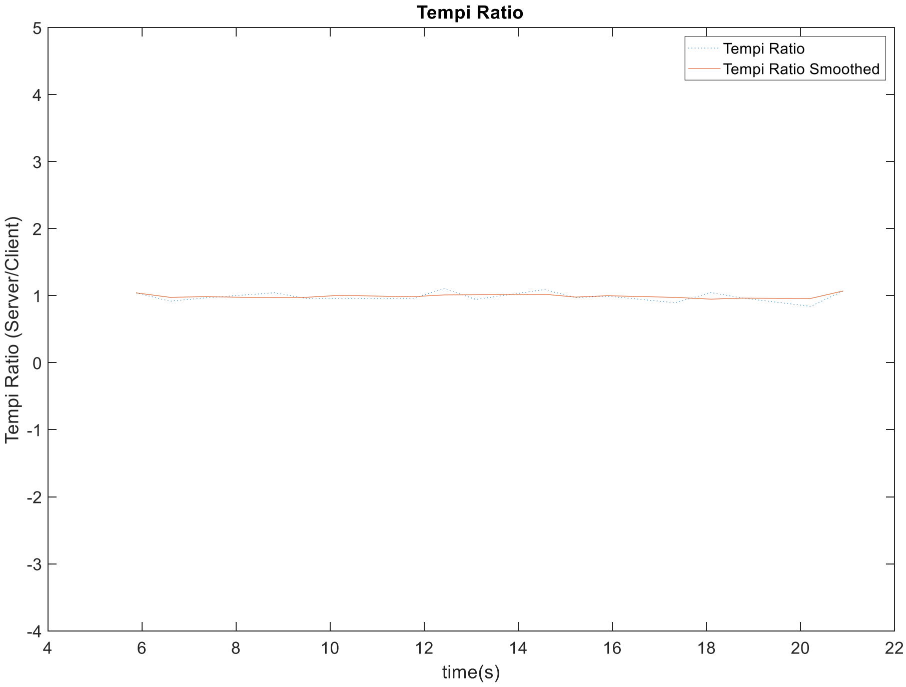


Figure 3.2.11.3 Tempi Ratio across performance. Showing tempi ratio and smoothed tempi ratio between performers using the VIIVA-NMP audio system prototype

The synchrony analysis function (see Appendix C) takes as input the two text files detailing onset times for each discrete performer undertaking the performance task and also the measured latency between performers (which will be discussed in Section 3.5: Testing Deployment). In NMP synchrony analysis, empirical research will typically separate two performers on a LAN, simulate a network delay between the two performers, and record the performance provided by each discrete performer into the same recording session with no delay [27]. In the context of the VIIVA-NMP audio system considerate is considered that a practical deployment with remote performers operating from their home networks. The performance at the Server (performer A) location is recorded, as the system features stable PCM audio streaming functionality. This means the Client (performer B) performance is recorded with a delay equal to the one-way network latency between the two remote performers. In order to provide the Client onsets at the time when they were performed it is necessary to simply shift the recorded Client performance by the measured One Way Trip (OWT) latency between sites for the purpose of synchrony analysis. It must be recognised that this method may be subject to slight inaccuracies due to clock drift between performers' audio interfaces [188].

The synchrony analysis function returns three values: the performance Asynchrony, Precision, and Tendency to Lead, provided in seconds.

Analysis of synchrony across varying latencies allows for validation of the suggested ~30ms latency upper limit associated with 'natural' interactivity between performers as suggested by empirical EPT research [29]. This will also allow identification of a hard limit on Immersive NMP technology: where audio latencies begin to interfere with 'natural' musical interactivity between performers the system is considered to have impaired immersive quality.

The characterisation of performance in terms of difference from 'natural' synchrony provides an estimation of Cognitive Immersive quality with respect to the qualities of Naturalness, Coherence [71], Realism [72] or Plausibility [73] and Interface Awareness and Quality [71].

3.3 Testing Design: Performance Experience Questionnaire

Whilst synchrony analysis provides a set of objective measures which characterises the naturalness of musical interactions between performers, questionnaires [38] are required for assessment of Perceived Immersion qualities.

3.3.1 Performance Experience Questionnaires in Empirical NMP Research

In empirical NMP research the application of such questionnaires has taken a range of forms. Some studies provide simple rating of the experience of performance tasks using NMP systems. These ratings will typically address discrete aspects of the performance experience such as:

- Discrete rating of the performance experience [34], [32], [36], [91].
- Discrete rating of the tolerability of network delay [34], [36].
- Direct comparison to live performance experiences [35].
- Discrete rating of the 'musicality' of performance [33].
- Discrete rating of 'emotional connection' in the performance experience [91].
- Discrete rating of perceived performance synchrony [90].

These simple subjective ratings measure the performer's awareness of characteristics of the performance. Such simple questionnaire design is often subject to noise in collected data as a result of the self-report errors, where individuals may not respond to questionnaire items with accuracy or consistency [38]. These empirical subjective measures in NMP do, however, indicate the immersive qualities which are most relevant to subjective NMP analysis.

Focus on discrete subjective rating of the performance experience or rating by comparison to the live performance experience demonstrates that Coherence [71] is an important immersive quality to address. In the context of Immersive NMP the characteristic of Coherence [71] (alternatively referred to as Realism [72] or Plausibility [73]) is one which can be defined in terms of the statement: a Coherent experience is one where the cognition of the Immersive NMP experience is similar to the cognition of live musical performance. Notably, in contrast to synchrony metrics, these Immersive Qualities are here measured in terms of Perceived Immersion rather than Cognitive Immersion.

Assessing the perception of network latency via questionnaire response addresses the immersive property of Interface Awareness and Quality [37]. This aspect of the immersive experience correlates strongly with Transportation and Disassociation from Reality [38]. Where a performer using an Immersive NMP system is aware of network latency, this performer is thereby actively aware the experience is a simulation. In this case a performer may feel Present, such that they feel 'in' the virtual performance experience, but be neither Transported nor Immersed, such that they are respectively aware the simulation is grounded in reality and do not feel submerged in the Immersive experience.

Emotional Connection between musicians in NMP most corresponds with Social Presence [246] and Communication [98] in multi-user (or co-located) virtual experiences such as NMP. In the context of group vocal performance, non-voice communication consists largely of embodied physical expression. Immersive NMP systems propose the use of avatar rendering to facilitate this, and Social Presence is likely to be only robustly measured where VR avatar rendering is implemented.

Subjective Musicality, as well as Emotional Connection, depends largely on the individual [38], and can be reasonably expected to correspond most with the personal preferences [247] of the individual. This corresponds most with Absorption of Narrative [38], including Emotional Narrative. In the context of Immersive NMP this may describe emotive or sensational response to the performance experience, or narrative plot contained in the performance piece.

Empirical questionnaire design addresses immersive qualities as features of Presence [63]. Such questionnaires have been adapted to the assessment in telepresence motivated NMP designs [37]. Though telepresence-motivated NMP analysis may report a high level of presence, it can be readily recognised that telepresence NMP systems will invariably operate at high latencies [29] which can be associated with poor immersive quality in VR/AR applications [30] specified by Immersive NMP. For this reason it is worth noting that there is no reason to expect similar Presence in Immersive NMP and Telepresence NMP systems when each system is used with equivalent parameters (for example, the same latency), and that it is sensible to expect performers to be more sensitive to Coherence issues such as latency in Immersive NMP VR/AR applications.

The same Presence questionnaire design which has been adapted to telepresence NMP has been adapted to use in VR choir applications [11]. Similar adaptation has been made for the analysis of immersive quality in virtual acoustic applications [55]. Virtual Reality performance technology which proposes Immersive NMP adaptation [9], [10] has been tested on Local Area Networks, providing questionnaires assessing Presence alongside questionnaire assessment of Health, Wellbeing and therapeutic value.

3.3.2 Performance Experience Questionnaire

The questionnaire devised for this testing protocol (see Appendix A) consists of 10 items.

Items 1-6 address immersive qualities of the virtual performance experience provided by the VIIVA-NMP audio system prototype. These questions are scored on 5 point Likert scales, where 1 indicates strongly disagree, and 5 strongly agree. 3 represents a neutral opinion.

These items are:

1. ***My performance experiences in the Virtual Acoustic Performance Space were consistent with my performance experiences in Real Performance Spaces.***

This item addresses the immersive quality of Coherence, and is amended from Item 12 of the Witmer and Singer Presence Questionnaire: "How much did your experiences in the virtual environment seem consistent with your real-world experiences?" [63]. This item looks to ensure that the VIIVA-NMP audio system may provide an experience users find realistic.

2. ***The Virtual Acoustic Performance Space was acoustically responsive to sounds I initiated and/or performed.***

Here Coherence is addressed again, but also Involvement and Attention in the virtual performance experience. Sensory immersive quality is also addressed, in that this assesses whether or not a user may provide input (vocal performance) and achieve expected output (the aural sensation of hearing one's self singing in a room). This also approaches Interface Awareness, in that a performer should not feel as though they are experiencing processed audio, but instead natural audio. This item is adapted from item 2 of the Witmer and Singer Presence Questionnaire: "How responsive was the environment to actions that you initiated (or performed)?" [63].

3. ***My Musical interactions with the acoustic environment in the Virtual Acoustic Performance Space seemed natural.***

Coherence is once again addressed by this item, as well as Interface Awareness, in that a natural interaction implies little awareness of simulation processes. This item is adapted from item 3 of the Witner and Singer Presence Questionnaire: "How natural did your interactions with the environment seem?"

4. ***My musical interactions with the other musicians in the Virtual Acoustic Performance Space seemed natural.***

This item is also adapted from item 3 of the Witmer and Singer Presence Questionnaire. Again Coherence is addressed, but in this case also Communication and Social Presence are incorporated into the questionnaire item. Additionally Interface Awareness (network connection and audio rendering) between performers is addressed.

5. ***I experience delay between my actions and expected outcomes.***

This item is amended from item 25 of the Witmer and Singer Presence Questionnaire: "How much delay did you experience between your actions and expected outcomes?" Coherence is considered again here (in that no delay is expected in live performance), and also address Interface Awareness.

6. ***I feel included in the acoustical scene presented in the Virtual Acoustic Performance Space.***

Presence is directly addressed in this item, and additionally co-Presence in the context of this study, as the acoustical scene itself contains more than one remote performer. This item further assesses Transportation to the Virtual Acoustic Performance Space, though not Disassociation from Reality, as inclusion in the virtual environment does not automatically imply one is no longer feeling included in reality also. This item is adapted from Colsman et al's [55] questionnaire generation form for assessment of immersion in virtual acoustic environments.

Items 7 and 8 in the devised questionnaire are rated on a 4 point scale detailing the acceptability of the feature addressed. These points are noted Absolutely Acceptable (equivalent to live conditions), Acceptable (like live conditions), Unacceptable (not like live conditions), and Absolutely Unacceptable (completely unlike live conditions). This rating scale adapts to 4 points as for these items there is no possible 'neutral' statement, as anything which does not qualify as at least 'acceptable' is therefore 'unacceptable'. For this reason a strong/mild binary (positive/negative) scoring system is used for these items. These items were generated to facilitate direct response to performance and latency conditions as in empirical NMP research.

These items are:

7. ***How would you rate the performance conditions your virtual acoustic performance experience?***
    a. ***Audio performance conditions were equivalent to live acoustic performance conditions.***
    b. ***Audio performance conditions were acceptably close to live acoustic performance conditions.***
    c. ***Audio performance conditions were not acceptably close to live acoustic performance conditions.***
    d. ***Audio performance conditions were completely unlike live acoustic performance conditions.***

8. ***How would you rate the latency experienced during your virtual acoustic performance experience?***
    a. ***No latency.***
    b. ***Acceptable Latency.***
    c. ***Manageable Latency.***
    d. ***Unmanageable Latency.***

Items 1-8 address several aspects of the immersive experience, chiefly Coherence, yet also Presence, Transportation, Interface Awareness, Attention and Involvement, and Communication. These questionnaire items are complimented with a ‘free comment’ section where test participants are free to provide open comment relevant to each question item if they wish to.

Items 9 and 10 address an application specific issue, namely Engagement:

9. ***Have you enjoyed using the system?***
10. ***Would you use the system again and, if yes, in what situation would you use the system?***

These items provide a simple measure of the usability of the VIIVA-NMP audio system design, and provide direction for further work with the VIIVA-NMP system development and application.

3.4 Virtual Acoustic Performance Spaces

It has been postulated that reverberation may impact the synchrony of performance in the context of NMP [27], [32]. Reverberation is a significant aspect of the VIIVA-NMP audio system design, where sound sources are auralised in virtual acoustic performance spaces. The effect of different virtual acoustic spaces are therefore assessed in testing of the VIIVA-NMP audio system prototype. Previous research provides no indication of the expected effect of reverberation on performance synchrony other than to propose that reverberation may 'smooth' the perception of onsets in performance [28]. This would imply that affective modelling of room simulation may allow for enhancement of the performance experience in Immersive NMP.

Repeating the performance task, synchrony analysis, and Performance Experience Questionnaire with different SIR measurements in the convolution reverberation used for acoustic simulation in the VIIVA-NMP audio system prototype provides preliminary investigation of potential effects of different reverbs.

Three room measurement sets were selected for use in testing, where these three rooms represent the range of reverberations commonly encountered in a vocal performance context. These three rooms were defined as:

1. Very dry (for example, a recording studio booth, representing the least reverberant environment typically encountered in vocal performance).
2. Medium Hall (for example a community hall or small venue used for choir practice, representing common halls found in vocal performance).
3. Very reverberant (for example a cathedral, representing the most reverberant performance space typically found in vocal performance).

It is proposed that effects of reverberation will be observable in repeated synchrony measurements for each of these performance spaces.

In this manner this work addresses the sub-objective of the project brief: ***To conduct performance synchrony pilot study of EPT for vocal performance under the influence of varying virtual acoustic performance spaces.***

The SIR measurement sets used describe various source locations relative to a receiver location for each of these three performance spaces. This allows for the placement of remote performers relative to the local performer in the relevant virtual performance space by each instance of the VIIVA-NMP audio rendering component. The SIR measurement sets were sourced from the Open Air Impulse Response Library [152]. The three measured performance spaces were:

1. Genesis 6 recording studios at the AudioLab, University of York (very dry).
2. St Margaret's Church – National Centre for Early Music, York (medium hall).
3. Lady Chapel, St Albans Cathedral (cathedral hall).

3.4.1 Genesis 6 Recording Studios

The Genesis 6 Recording Studios are located at Genesis 6, University of York. This studio booth measurement set (*Figure 3.4.1.1*) where there are two discrete locations for singers to be placed in auralisation of this space. These locations are derived from measurement of a drum kit setup, where the crash and ride positions are used for two singers who are close to each other in the booth. The studio space is acoustically treated with MDF/Rockwool panels to dampen sound in the booth, making this an appropriate choice for a room representing dry vocal performance spaces (*Table 3.4.1.1*).

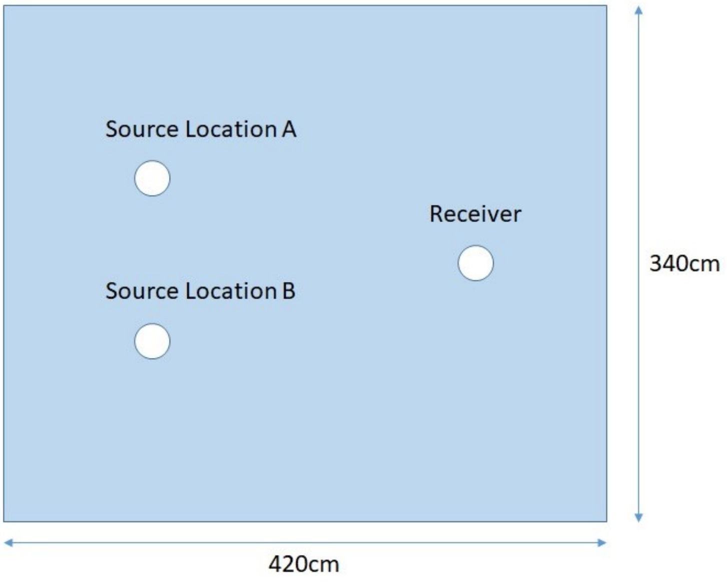


Figure 3.4.1.1 Genesis 6 Recording Studios SIR measurement setup, adapted from *[248]*.

This SIR measurement set was sourced from the Open Air Impulse Response Library [248]. The SIR measurements were made using a logarithmic sine sweep [146] excitation signal. This excitation signal was output through a Genelec 8040 loudspeaker [249]. For the purposes of vocal performance the use of such a directional loudspeaker is appropriate, as voice is a directional instrument. The excitation signal is captured using the Soundfield ST350 Portable Microphone System [250] to provide 1st Order B-Format SIR measurements.

Octave Band (Hz)	RT 60 (s)	EDT (s)	D50	C50 (dB)	C 80 (dB)
31.25	0.8	0.98	0.26	-4.65	-2.33
62.5	0.57	0.73	0.46	-0.74	5.82
125	0.28	0.34	0.92	10.63	16.5
250	0.27	0.22	0.94	11.92	20.03
500	0.21	0.22	0.98	16.19	23.87
1 k	0.19	0.22	0.97	15.36	26
2 k	0.21	0.22	0.97	14.75	23.59
4 k	0.22	0.22	0.97	14.48	22.54
8 k	0.21	0.22	0.96	13.44	22.92
16 k	0.18	0.22	0.98	17.54	27.99

Table 3.4.1.1 Averages IR parameters for Genesis 6 Studio, showing Reverb Time (RT60), Early Decay Time (EDT), Definition (D50), and Clarity (C50 and C80). This table is adapted from the acoustic parameters table for this SIR set sourced from the Open Air Impulse Response Library *[248]*.

3.4.2 St Margaret's Church – National Centre for Early Music

St Margaret's Church (*Figure 3.4.2.1)* is located in York and is commonly used for choir performance and practice. This room is a medium sized hall with damping drapes and panels which may be deployed. The instance of measurement used the VIIVA-NMP audio system testing describes the room with all drapes and 75% of absorption panels in use. This setup was identified as most appropriate for vocal performance by singers who practice in this space regularly, and was confirmed as an appropriate representation of a medium hall (*Table 3.4.2.1*).

The SIR measurement set describes a range of sound source locations relative to a static receiver position in the space, allowing for placement of singers relative to one another in Auralisation of this space using the VIIVA-NMP audio system prototype. This measurement set was made using a sine-sweep excitation source, output from a Genelec S30D loudspeaker [251]. This directional loudspeaker is once again appropriate for making measurements to Auralise voice, which is itself a directional instrument. The excitation signal sounded in the room was recorded using a Soundfield SPS422B Studio Microphone System [252].

Notably in the Acoustic Parameter data provided for the SIR set by the Open Air library a 21s RT60 is recorded for the 31.25Hz octave band. This was identified as noise in the SIR measurement set. .

Octave Band (Hz)	RT 60 (s)	EDT (s)	D50	C50 (dB)	C 80 (dB)
31.25	2.1	2.39	0.06	-12.22	-3.16
62.5	2.69	2.13	0.22	-5.48	-3.53
125	1.82	1.62	0.41	-1.58	-0.03
250	1.6	1.87	0.22	-5.38	-1.41
500	1.49	1.62	0.37	-2.29	-0.58
1 k	1.4	1.49	0.37	-2.22	0.53
2 k	1.3	1.36	0.43	-1.22	1.52
4 k	1.15	1.11	0.46	-0.78	3.2
8 k	0.81	0.59	0.71	3.8	7.91
16 k	0.52	0.47	0.85	7.44	12.48

Table 3.4.2.1 Averages IR parameters for St Margaret's Church, showing Reverb Time (RT60), Early Decay Time (EDT), Definition (D50), and Clarity (C50 and C80). This table is adapted from the acoustic parameters table for this SIR set sourced from the Open Air Impulse Response Library *[248]*.

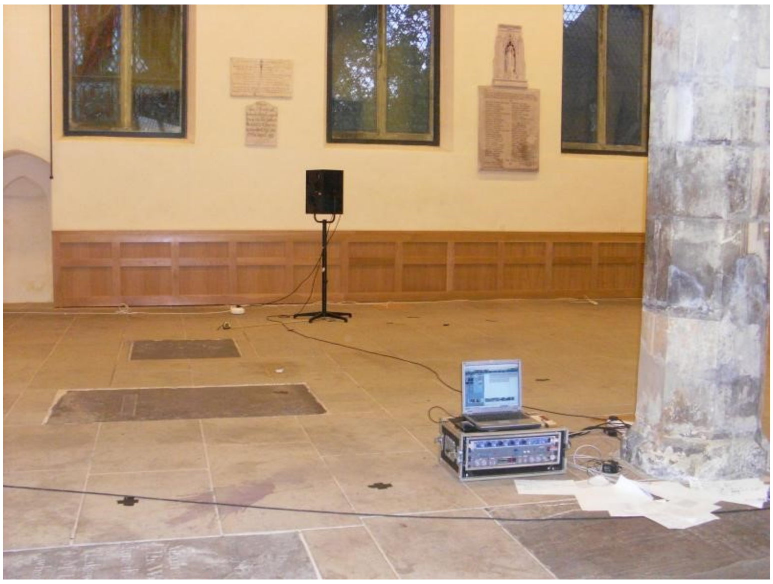


Figure 3.4.2.1 St. Margaret's Church SIR measurement setup, sourced directly from *[248]*.

3.4.3 Lady Chapel – St. Alban's Cathedral

Lady Chapel (*Figure 3.4.3.1*) represents the largest type of hall typically encountered in the context of vocal performance. The SIR measurement set for this space consists of two discrete source-receiver relative positions, again allowing placement of singers for testing in auralisation using the VIIVA-NMP audio system prototype.

Acoustic parameters of the averaged captured SIR set (*Table3.4.3.1*) demonstrate this space is an appropriate choice for 'extremely large hall'. The SIR measurements were taken using the same method and equipment as with SIR measurements at St. Margaret's Church (sine sweep excitation signal, Genelec S30D loudspeaker, and Soundfield SPS422B Studio Microphone System).

Octave Band (Hz)	**RT 60 (s)**	**EDT (s)**	**D50**	**C50 (dB)**	**C 80 (dB)**
31.25	3.44	2.14	0.13	-8.39	-5.66
62.5	3.28	2.52	0.34	-2.96	-2.13
125	3.31	2.4	0.19	-6.22	-0.65
250	2.78	1.76	0.56	0.99	2.74
500	2.49	1.5	0.57	1.3	2.74
1 k	2.34	1.5	0.63	2.35	3.79
2 k	2.1	1.24	0.67	2.98	4.44
4 k	1.71	0.6	0.85	7.66	9.22
8 k	1.05	0.35	0.91	10.08	12.2
16 k	0.62	0.22	0.96	14.4	17.47

Table 3.4.3.1 Averages IR parameters for Lady Chapel, showing Reverb Time (RT60), Early Decay Time (EDT), Definition (D50), and Clarity (C50 and C80). This table is adapted from the acoustic parameters table for this SIR set sourced from the Open Air Impulse Response Library *[248]*.

Figure 3.4.3.1 Lady Chapel, St. Alban's Cathedral, sourced directly from *[248]*.

3.4.4 Preparation of SIR for use in Testing

Having sourced three appropriate SIR measurement sets representing dry, medium hall and large hall it was required that these SIR were prepared for use with the VIIVA-NMP audio system prototype presented in Chapter 2 of this Thesis.

Firstly it was required that discrete SIR from each measurement set were selected for use in testing. In the defined vocal performance task (detailed in Chapter 3.2) two singers will be present. As the VIIVA-NMP audio system prototype renders audio locally at each performer it is required that two SIR are selected for use in each instance of the VIIVA-NMP audio rendering component. The first SIR will Auralise the singers own audio, and the second SIR will place the remote performer relative to the local performer. As each performer has their own discrete audio rendering process, it was decided that it was suitable to present each performer with the same acoustic scene (such that each performer experiences being in the same position with the same relative remote performer placement). This decision was made in order to impose a control on the experience of each performer.

A complication is present in that no SIR measurements are available which are measured with the source and receiver position being the same, such that the measurement is absolutely appropriate for the singers own voice. As SIR for singers own voice must be selected from sets where there is distance between source and receiver it is expected that some externalisation will be an issue, as the singers own voice will be Auralised as displaced from the listening position.

For Genesis 6 studios two appropriate source-receiver positions are available, described by the SIR measurements for the Crash and Ride positions relative to the static listener position. This presents two virtual singer locations in close proximity to one another. It was decided that at each instance of the VIIVA-NMP audio renderer the crash position would be used for the local performer's own audio, and the ride position for the remote performers audio.

For St. Margaret's Church a grid of 26 receiver positions are available relative to a static source (*Figure 3.4.4.1*). It was decided that positions 13 and 14, both nearest the source position, would be appropriate. Receiver position 13 was selected for use for local performer's own voice, and position 14 selected for Auralisation of the remote performer.

Lady Chapel presents only two SIR measurements, each describing a discrete source-receiver location (*Figure 3.4.4.2*). Source-receiver position A was selected for local performer's own voice and source-receiver position B selected for remote performers voice.

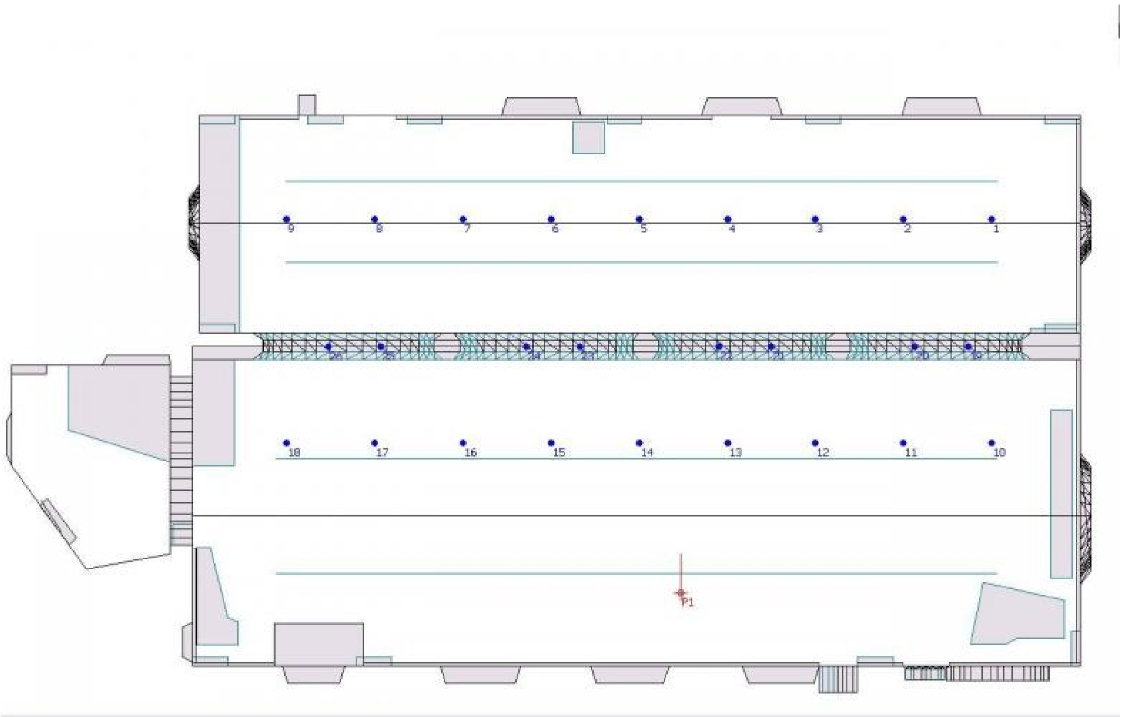


Figure 3.4.4.1 Receiver grid relative to source position for St Margaret's Church SIR measurement set, sourced directly from *[248]*.

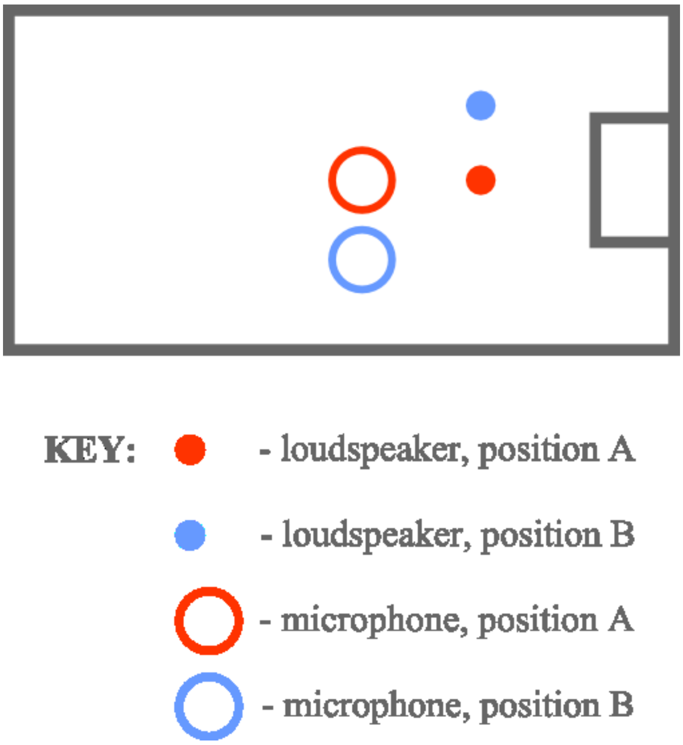

KEY:
● - loudspeaker, position A
● - loudspeaker, position B
○ - microphone, position A
○ - microphone, position B

Figure 3.4.4.2 Source-receiver locations A and B for Lady Chapel SIR measurements, sourced directly from *[248]*.

Having selected appropriate SIR for use it was required that these measurements are prepared for use with Kronlachner MCFX Convolver [195] as used in the VIIVA-NMP audio system prototype. All SIR measurement sets sourced were in legacy FuMa format, and required updating to AmbiX format [128]. This was achieved with a simple Matlab script (see Appendix F) using the conversion matrix detailed in Equation 2.4.5.1 [128]. Configuration files were then created, which allow for loading of SIR into the MCFX plugin [195]. Channel gains in the configuration file were set such that input signals of amplitude 1 will produce output of approximately -6dB in order to ensure no clipping in the convolution during the performance task.

3.5 VIIVA-NMP Prototype Deployment

The research presented in this thesis aims to assess the developed VIIVA-NMP audio system design in a practical use case. As such it is required that the VIIVA-NMP audio system prototype presented in Chapter 2 of this thesis is deployed to test participants at their homes in order to conduct the testing protocol. This requires that participants conduct the testing protocol using equipment which is readily available.

Home deployment and hardware variation presents two key issues in comparison to lab conditions.

Under the Covid 19 pandemic restrictions access to lab space is highly restricted and access to participant homes completely unfeasible. As the VIIVA-NMP audio system prototype requires the use of specialist software such as Reaper, plugins used in audio rendering, and Jacktrip, as well as some home networking knowledge, it was recognised that a system was needed to help participants to set up and operate the prototype design and successfully take part in testing. The optimal solution is of course a complete application operating behind an accessible user interface, however this itself is quite beyond the scope of this project. An extensive setup and operation manual was therefore created for participants, and participants were aided in setup and test completion by the remote guidance of the lead researcher.

Hardware was noted for each participant to assess performance of the VIIVA-NMP audio system design under equipment variance. Though this imposes a degree of lapse in testing control, hardware variance will be expected in any real-world use case. Collecting data with this allowance may enable better estimation of any hardware-relevant limitations of the VIIVA-NMP audio system design that would not be identifiable if hardware was completely controlled.

A key notable point on hardware variance is that the head-tracking component [198] used in the VIIVA-NMP audio system prototype design is a system component which will not be typically accessible to home users in the context of this study. As such the ***deployment of the prototype design for testing does not include head-tracking and 3DoF functionality***, providing instead static binaural renders during the performance task.

3.5.1 Test Protocol Distributable

A folder containing all soft resources required for the testing protocol was created (*Table 3.5.1.1*) in order to facilitate distribution of the testing protocol (see Appendix A). The distributable is specified for Windows 10 users, however provisional content is included such that participants using various Mac OS and older Windows systems may also be catered for. The contents of this folder include:

- ***Installer:*** *Where the content consists of installation files for the required software. This content is to be installed on participant's home computers during guided equipment setup with the help of the lead researcher and the provided Test Protocol Setup guide.*
- ***Preconfigured:*** *Where the content is pre-installed in the distributable folder and is ready for use in the testing protocol.*
- ***Backup for Troubleshooting:*** *Where content is included to allow easy troubleshooting if preconfigured content is not operating as expected. This is relevant to the setup process, where variation in home computer operating system is expected. Though the setup guide is specified for Windows 10 this is the content which allows guided variation of the setup process.*
- ***Peripheral:*** *This content is included for use by the participant outside the Setup and Test Protocol components of testing, generally providing supplementary information or fulfilling ethical requirements.*

Jack Audio Connection Kit [193] provides routing functionality between applications within the local computer (as discussed in Chapter 2). The installation of this application is straightforward, and can be easily accomplished following the installation wizard, the instructions found in the Participant Setup Guide, or with the remote guidance of the lead researcher. After installation, it is required that Jack is manually configured to operate with the relevant audio interface for each discrete test participant, including the manual registry of .dll files. As this must be achieved on a bespoke basis (audio interface is likely to vary) this process is covered in the Participant Setup Guide and is completed with the guidance of the lead researcher.

Jacktrip [100] provides the audio transport functionality between remote participants as detailed in Chapter 2. Installation for this application is again straightforward and easily accomplishable following either the installation wizard, Participant Setup Guide, or with the guidance of the lead researcher. After installation Jacktrip requires configuration to work with the relevant audio interface as with Jack. Guidance for this process is once again offered both in the Participant Setup Guide, and with the assistance of the lead researcher.

Reaper [194] is included as a portable installation in the Testing Protocol Distributable. This allows for the application to be run and operated as instructed in the Testing Protocol Guide with no installation. This allows Reaper sessions to be preconfigured with Kronlachner plugins [195], [196] included and preconfigured. Each of the three virtual acoustic performance spaces defined for use in the testing protocol is prepared as a discrete Reaper session, such that to 'load' each room during the testing protocol, performers may simply load the relevant Reaper session and create connections with the audio interface and Jacktrip using Jack as directed in the Test Protocol Guide.

Resource	**Type**
Jack Audio Connection Kit	*Installer*
Jacktrip	*Installer*
Reaper	*Preconfigured*
Reaper Sessions	*Preconfigured*
Kronlachner AmbiX plugins	*Preconfigured, Backup for Troubleshooting*
Kronlachner MCFX convolver plugin	*Preconfigured, Backup for Troubleshooting*
Open Air SIR data	*Preconfigured, Backup for Troubleshooting*
SADIE II Binaural Decoding Data	*Preconfigured, Backup for Troubleshooting*
Performance Task Indicative Material	*Peripheral*
Test Protocol Guide	*Peripheral*
Setup Guide	*Peripheral*
Performance Experience Questionnaire	*Peripheral, Backup for Troubleshooting*
Performer Hardware and Setup Questionnaire	*Peripheral, Backup for Troubleshooting*
Participant Information Sheet	*Peripheral*
Participant Consent Form	*Peripheral*

Table 3.5.1.1 Testing Protocol Distributable Contents listed by content type.

Reaper [194] is included as a portable installation in the Testing Protocol Distributable. This allows for the application to be run and operated as instructed in the Testing Protocol Guide with no installation. This allows Reaper sessions to be preconfigured with Kronlachner plugins [195], [196] included and preconfigured. Each of the three virtual acoustic performance spaces defined for use in the testing protocol is prepared as a discrete Reaper session, such that to 'load' each room during the testing protocol, performers may simply load the relevant Reaper session and create connections with the audio interface and Jacktrip using Jack as directed in the Test Protocol Guide.

Though no head-tracking is used in the testing protocol deployment, it is worth noting that the AmbiX rotator plugin used to facilitate 3DoF functionality is still present in these Reaper sessions. This allows for this functionality to be considered in latency throughput calculation, even though it is not utilised for 3DoF implementation in this context.

SIR and Binaural decoding data are preconfigured in the prepared Reaper sessions as appropriate. This resource is also present in the folder to relink in the eventuality that the Reaper sessions do not load as expected. The same is true of VST resource which is also present to be relinked to the prepared Reaper sessions if required. If this sort of troubleshooting is required a 'setup' Reaper session is included, such that resource can be correctly linked to Reaper without affecting the configured sessions representing the three different virtual acoustic performance spaces.

Guide material for the Setup of the VIIVA-NMP audio system prototype at participant home computers and the conduction of the Testing Protocol are included, as well as some other supplementary material. Performance Task Indicative Material is presented for the participant to familiarise themselves with prior to conduction of the testing protocol. Participant Information Sheet and Consent Form are included here in accordance with the ethical approval obtained for this study and University of York ethical guidelines. The Performer Experience and Hardware and Setup Questionnaires are included. In the testing protocol these questionnaires are distributed as Google Forms. These questionnaires are included in the testing distribution to cover for the eventuality that a participant may not be able to use the Google Forms questionnaires, and can instead complete and return the PDF or Word Doc included in the testing deployment folder.

3.5.2 Participant Setup Guide

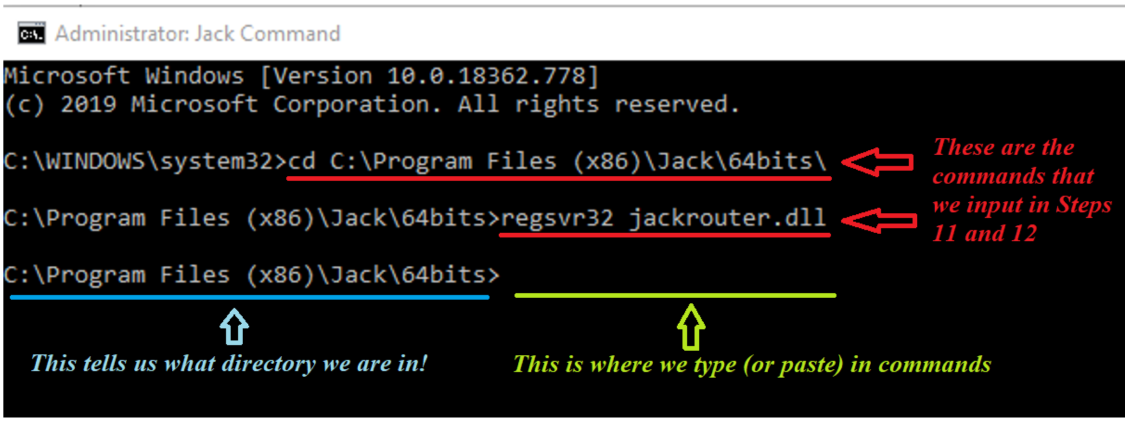
The Participant Setup Guide provides a 'step-by-step' instruction set, complete with graphic illustration of each step, in order to make the process as accessible as possible. These instructions were followed with the assistance of the lead researcher over videoconferencing (see Appendix A).

Installation and configuration of Jack Audio Connection Kit and Jacktrip are presented in detail (*Figure 3.5.2.1*) and involves some Command Line input. A 'paint-by-numbers' approach is used where participants are able to complete the installation process with no prior knowledge required provided the instructions are followed accurately. A graphic illustration is presented to accompany each discrete instruction in the Setup Guide such that participants need only copy what is shown in these illustrations.

One largely unavoidable facet of deploying NMP systems is the associated home network setup. With the required applications successfully installed it is required that the home computer is configured to allow audio data traffic via a home router connected to the computer via ethernet. This process requires that three components of the home network are correctly configured. Firstly it is required that the home computer is assigned a static IP address on the home network (*Figure 3.5.2.2*). Secondly port forwarding from the home router to this static IP must be configured on the home router for ports 4464 and 4465 (*Figure 3.5.2.3*). Finally local firewall rules must be made to allow traffic of audio data to the home computer through these ports using Jacktrip (*Figure 3.5.2.4*).

These processes are detailed with step-by step instructions and illustration in the Setup Guide provided. For this stage of setup it is generally required that the lead researcher is present via videoconference as variation in home router will cause some deviation from the setup Guide for the port forwarding process.

Upon completion of the Setup Guide participants were guided to test the Jacktrip connection and test that each of the Reaper Sessions could be loaded properly. This is accomplished working with the lead researcher remotely over videoconference. Having completed the installation and successfully tested the deployment a date is scheduled where pairs of participants may conduct the Testing Protocol, consisting of the Performance Task and Questionnaire Response.



Administrator: Jack Command

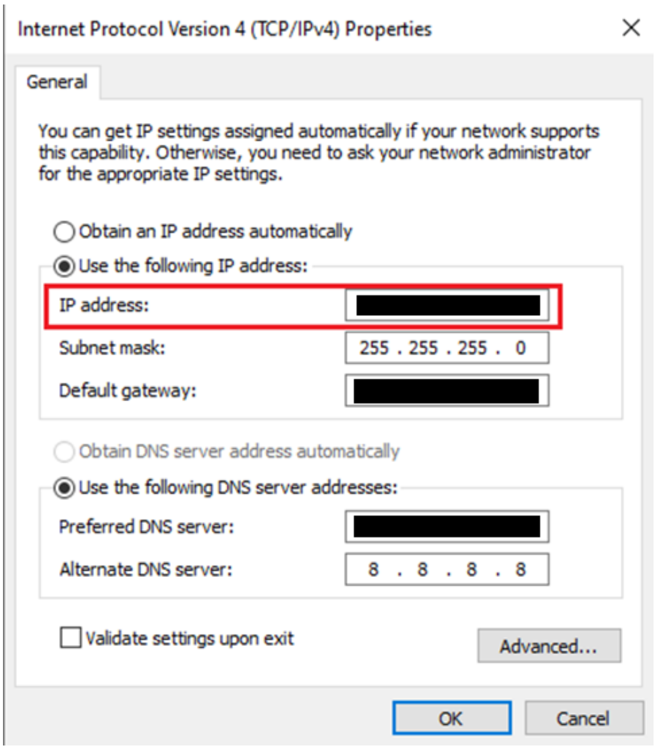
```
Microsoft Windows [Version 10.0.18362.778]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\Program Files (x86)\Jack\64bits\

C:\Program Files (x86)\Jack\64bits>regsvr32 jackrouter.dll

C:\Program Files (x86)\Jack\64bits>
```

These are the commands that we input in Steps 11 and 12

This tells us what directory we are in!

This is where we type (or paste) in commands

Figure 3.5.2.1 Example of Jack installation illustration accompanying clear instructions in the Participant Setup Guide.



Internet Protocol Version 4 (TCP/IPv4) Properties

General

You can get IP settings assigned automatically if your network supports this capability. Otherwise, you need to ask your network administrator for the appropriate IP settings.

○ Obtain an IP address automatically

● Use the following IP address:

IP address: [redacted]

Subnet mask: 255 . 255 . 255 . 0

Default gateway: [redacted]

○ Obtain DNS server address automatically

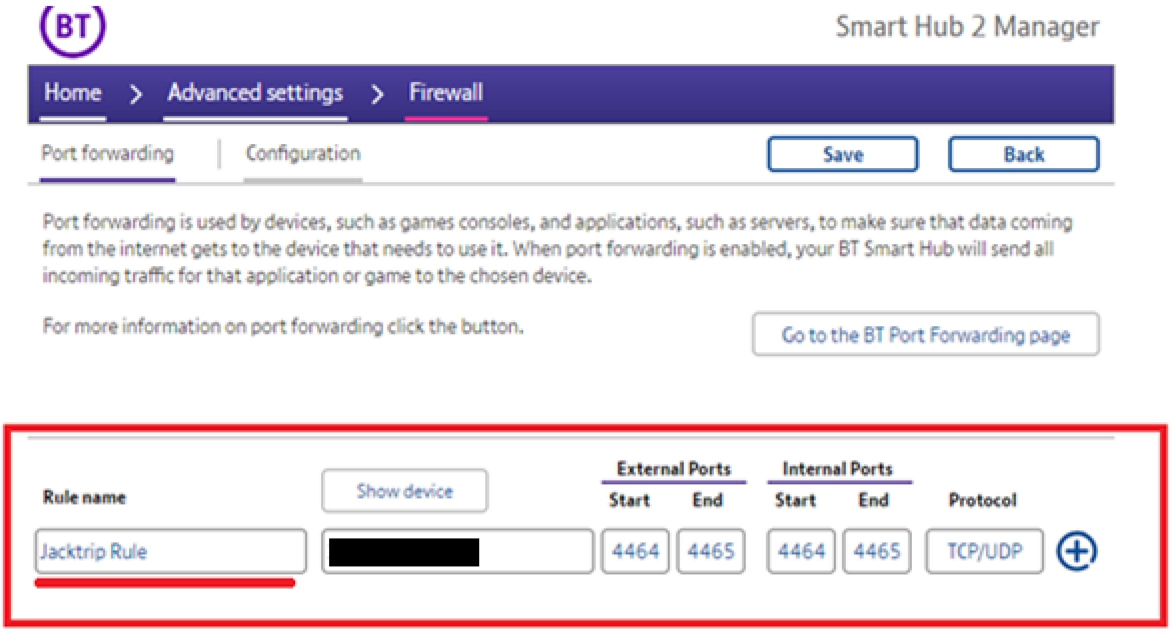
● Use the following DNS server addresses:

Preferred DNS server: [redacted]

Alternate DNS server: 8 . 8 . 8 . 8

□ Validate settings upon exit

Figure 3.5.2.2 Static IP assignment instruction illustration example from the Setup Guide.



BT
Smart Hub 2 Manager
Home > Advanced settings > Firewall
Port forwarding | Configuration
Save
Back
Go to the BT Port Forwarding page
Show device

Port forwarding is used by devices, such as games consoles, and applications, such as servers, to make sure that data coming from the internet gets to the device that needs to use it. When port forwarding is enabled, your BT Smart Hub will send all incoming traffic for that application or game to the chosen device.

For more information on port forwarding click the button.

Rule name		External Ports Start	External Ports End	Internal Ports Start	Internal Ports End	Protocol
Jacktrip Rule	[redacted]	4464	4465	4464	4465	TCP/UDP

Figure 3.5.2.3 Example illustration accompanying instructions for Port Forwarding in the Setup Guide.

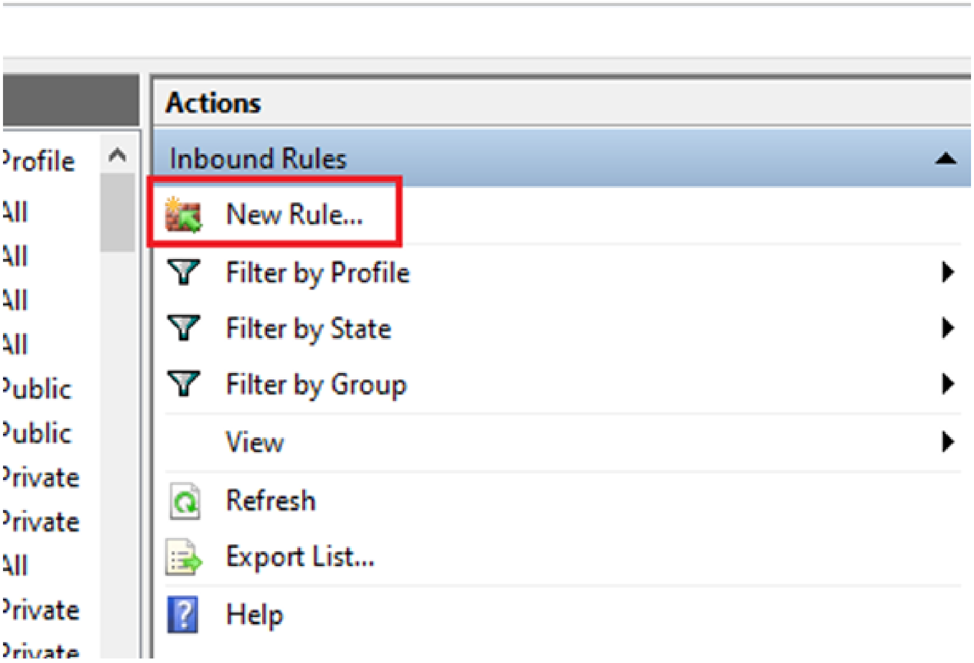
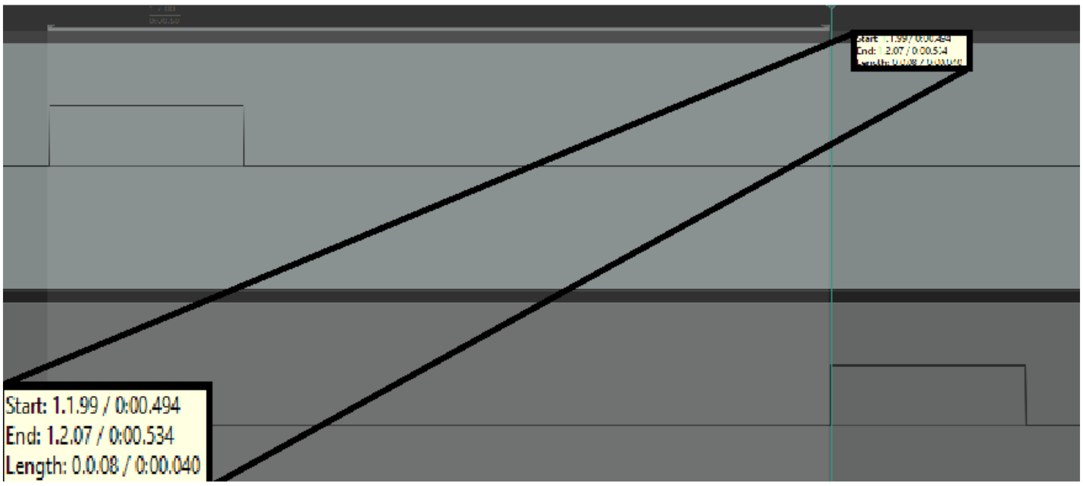


Figure 3.5.2.4 Example of illustration accompanying instructions for Firewall Configuration in the Setup Guide

3.5.3 Latency and Bandwidth Estimation

To provide latency measurement functionality a pulse signal is used (see Appendix F). This latency measurement is built into the pre-configured Reaper sessions (*Figure 3.5.3.1*) in order to be as unobtrusive to test participants as possible. This is achieved by designating one performer 'Server' (the one who initialises the Jacktrip session in the Test Protocol) and the other designated 'Client'. Different Reaper sessions and instructions are prepared for each of these designations. Between these two instruction sets a pulse signal is sent from the server Reaper session to the Client, then looped back to the Server where the loopback signal is recorded in the Reaper session. This allows measurement of the Round Time Trip (RTT) by comparing the leading edge of the pulse and recorded loopback pulse in the Reaper session. This latency is manually measured in Reaper in milliseconds. The One Way Trip (OWT) is then estimated as half the RTT. This latency measurement describes throughput for network transport between the two performers, but not local audio rendering throughput, which can be calculated for each performer based on buffer size and sample rate. For each OWT measurement the local processing latency (as detailed in Chapter 2) must also be added to provide the full OWT latency. As audio data is being sent and received at constant rate the latency measurement is assumed to remain constant throughout the testing protocol. It is of course worth noting that several control issues may be present where audio data underruns and overruns and clock drift using Jacktrip [188] may cause some variance in this latency measurement. It is expected that in the context of this study such variance will be minor and will not unexpectedly skew results.



Start: 1.1.99 / 0:00.494
End: 1.2.07 / 0:00.534
Length: 0.0.08 / 0:00.040

Figure 3.5.3.1 Loopback Pulse measurement example.

Bandwidth estimation is achieved using online bandwidth measurement tools. This provides a rough estimate of the available bandwidth, and will not provide as accurate a measurement as more robust bandwidth measurement methods such as iPerf tests. As the audio streaming method used in this study requires symmetrical upload/download speeds the Estimated Bandwidth Available is taken as the lowest measurement out of upload and download. The purpose of the bandwidth estimation conducted using web tools is simply to ensure that sufficient bandwidth is available to take part in testing, and to keep an approximate measurement of typical bandwidths encountered through the course of this study.

The metric 'Estimated Bandwidth Used' is calculable based upon the sample rate, resolution and FEC redundancy. Throughout testing audio resolution is 24-bit in order to conform to pro audio standards, though it recognised that bandwidth reductions may be made by reducing audio to CD quality (44.1kHz, 16-bit). FEC redundancy effectively adds to the packet size through the inclusion of additional buffers of audio data. This functionality ultimately amounts to a parallel stream of the same audio data delayed by an amount of buffers according to the overlap specified by the redundancy factor. As FEC will certainly be used in testing the Estimated Bandwidth Used is calculated simply as:

$$Estimated\ Bandwidth\ Used = Fs * 24 * (No.of\ redundant\ packets + 1) \quad \textbf{\textit{(4.1.4.1)}}$$

Where Fs is the Sample Rate. This value is then rounded up to the nearest 0.1Mb/s to allow overhead for packet headers and other associated network processes.

3.5.4 Participant Hardware and Setup Questionnaire

In addition to the soft resource provided in the testing deployment folder some hardware resource is required for usage of the VIIVA-NMP audio system prototype as detailed in Chapter 2. For the testing protocol it is required that participants have access to:

- Home computer and home router
- Minimum 10Mb/s bandwidth (upload and download)
- Audio Interface
- Headphones and Microphone

A Hardware and Setup Questionnaire was created (see Appendix A) to monitor variation in hardware and configuration for each discrete participant's home deployment. In Spatial Audio it is well known that the response of microphones and headphones can affect performance. Though this questionnaire is unlikely to reveal statistically significant dependencies, it proves useful in allowing estimation of guidelines for hardware and system configuration which may affect system performance.

The questionnaire items include:

- Sample Rate and Buffer Size: This will be used to calculate the latency of local audio rendering, and will provide a guide on the minimum system configuration for the VIIVA-NMP audio system prototype.
- Jitter Buffer Length: This will provide a descriptive measure for typical jitter buffer lengths required for stable audio streaming in a practical context.
- Audio Interface, Microphone and Headphone selection: Though differences will not be likely to be observed between models, categories of these hardware components describing general build quality may prove informative.

Also included in hardware measurements, though recorded directly by the lead researcher rather than through self-report, is the number of redundant packets used in FEC functionality for each performance.

Additionally this questionnaire records data on singer proficiency, asking each test participant to rate their proficiency as Amateur, Proficient, or Professional.

Amateur is given the definition: “I have no singing experience outside my own practice”.

Proficient was defined: “I have some experience in vocal performance with other, I have sung in bands”.

Professional was attributed the definition: “I am a trained singer. I have experience performing at a professional level”.

Finally the questionnaire asks whether or not the participant pair undertaking the duet vocal performance task have previously sung together. This question can be relevant or not relevant depending on whether both participants have consented to make their identities known to each other.

3.6 Testing Deployment Overview

Following the development of all the required resources for the deployment and setup of the VIIVA-NMP audio system prototype, the following protocol was undertaken.

3.6.1 Ethical Approval

Prior to deployment of this testing procedure Ethical Approval was obtained from the University of York ethics committee. Largely considerations for the experiments conducted in this study were relevant to security of personal information and the deployment of the system to performer homes. Prior to involvement in the project informed consent is provided by each test participant who confirms they have read the provided information sheet.

3.6.2 Equipment and Setup

The equipment required for the testing protocol consists of:

- *Home Computer*
- *Home Router and Internet Access*
- *Audio Interface*
- *Microphone*
- *Headphones*
- *Peripheral connectors and equipment (XLR, Ethernet cable, microphone stand if appropriate)*
- *Test Protocol Distributable*

Prior to the Test Protocol the participant and the lead researcher engage in a Setup session. Both parties meet via online videoconferencing and the participant proceeds through the Setup Guide with the assistance of the lead researcher. On completion of this setup process the participant will connect the VIIVA-NMP audio system prototype with the lead researcher and test the connection and the pre-configured Reaper sessions to ensure the system is operating correctly. When this is completed a time is scheduled to conduct the Testing Protocol.

Prior to testing participants are encouraged to familiarise themselves with the indicatory material for the piece of music used in the performance task which is contained in the testing distributable.

3.6.3 Test Protocol

The test protocol is guided by the lead researcher who joins remote performers online in video conferencing. One performer is designated ‘Server’ (who will initialise the Jacktrip session) and the other performer is designated ‘Client’.

Following the Test Protocol Guide (which provides discrete instructions for Server and Client) and with the assistance of the lead researcher the pair will open a preconfigured Reaper session and form a connection between the remote sites.

Three Reaper sessions are prepared, each representing a reverberant performance space. Each of these spaces are denoted by a letter: A is Genesis 6 studios, B is St. Margaret's church, and C is Lady Chapel. Participant pairs will attempt the performance task in each of these three virtual rooms by loading the appropriate Reaper sessions. The order in which performer pairs proceed through the virtual room sequence is based on a circular rotation: The first pair will proceed A, B, C. The second pair B, C, A, and the third C, A, B. Consecutive pairs then repeat this circular rotation of room order. This was decided so as not to unduly influence results with room order by consistently progressing linearly through the room sizes and reverberation times.

With the first Reaper session (representing the first virtual room in the pairs discrete sequence) loaded, participants are then instructed to increase their microphone input gain until a peak level of approximately -6dBFS is achieved (as observable visually in the Reaper session). Participants are then requested to set headphone output level to approximately -25dBFS.

The participant pair then begin recording in the Reaper sessions, prompting the loopback pulse latency measurement. The performance pair are then requested to provide three attempts of the performance task. On completion, recording is stopped and Reaper sessions saved such that audio recording of each performer is present for data extraction.

On completion of the three performance task attempts each participant is requested to complete the Performance Experience Questionnaire via Google Forms.

This process (three performance task attempts and one Performance Experience Questionnaire) is then repeated for each Reaper session representing a virtual acoustic 'room' according to the relevant room sequence for the group.

During each attempt of the performance task participants are requested not to look at the computer screen, and instead try to find a non-distracting visual focus (such as a blank area of wall). This is decided as the VIIVA-NMP audio system prototype does not provide visual playback. Any attention to the high-latency video conferencing with the lead researcher and other participants, or two graphic display in Reaper, may cause performers to lose focus on the task and thereby presents a lack of control. In order to control this a neutral unobtrusive and unrelated visual focus is chosen.

Upon completion of three attempts of the performance task in each of the three virtual rooms the test participants each complete the Hardware and Setup Questionnaire via Google Docs. Saved Reaper recordings of the performance task are then forwarded to the lead researcher and the test protocol is complete.

3.6.4 Data Retrieval

Data is retrieved from the Performance Task via the recordings saved with the Reaper sessions. For each performer pair latency is measured from the loopback pulse recorded in Reaper sessions. The audio files representing discrete performers for each performance task attempt are exported and input to the developed TIMEX-Lite onset detection algorithm (after any required manual pre-processing). With an onset list provided for each audio file, pairs of audio files representing each attempt of the performance task are input to synchrony measurement script to output synchrony metrics. Data from questionnaires is retrieved directly from participant response via Google Forms.

3.6.7 Participant Recruitment

Participants for this study were recruited based upon several perquisites:

- The participant has prior singing experience.
- The participant has normal hearing.
- The participant has access to typical home computer, and wired internet connection of minimum 5Mb/s (upload and download).

Further desirable features in participant recruitment were:

- The participant has an interest in audio technology and/or VR/AR technology.
- The participant is actively engaging with NMP systems.
- The participant's home computer uses Windows 10 operating system (though other systems were present in testing).

3.7 Summary

Conclusively, a thorough testing and analysis protocol is presented in this chapter. The development of the required resource and relevant literature are detailed, allowing the provision of a means by which it is possible to assess the VIIVA-NMP audio system design in a practical use-case. Objective synchrony comparison to values characterising 'natural interactivity' and questionnaire analysis of Perceived Immersion qualities will allow the identification of the conditions under which this Immersive NMP audio system design is optimal, and under which condition Immersive qualities are impaired. Analysis of this information alongside relevant technical specification data will allow statement of the minimum and recommended technical requirements of Immersive NMP audio, and provide a guide as to the expected performance of Immersive NMP systems.

As a practical use-case study of Immersive NMP with real-time acoustic simulation and spatial audio delivery which conforms to the standards of state-of-the art consumer VR technology this study presents an original contribution to the field. Supplementary investigation of the effect of reverberation further contributes by aiming to improve knowledge of role of reverberation in NMP.

4 VIIVA-NMP AUDIO SYSTEM TESTING: RESULTS AND DISCUSSION

The data retrieved from completion of the testing protocol using the VIIVA-NMP system can be broken down into three discrete categories:

- Setup, Hardware and Session data.
- Performance Synchrony data.
- Performance Experience Questionnaire data.

Setup, Hardware and Session data detail the discrete configuration of the deployed VIIVA-NMP prototype for each instance of the testing protocol. This collected data, combined with synchrony analysis, allows for characterisation of the VIIVA-NMP audio system technical requirements, thus addressing a sub-objective of this project: ***To conduct a pilot study in the practical deployment of audio systems for Immersive NMP and identify the challenges and opportunities presented by 'real-world' deployment.***

Synchrony data provides the primary focus of this study, where comparison to values which characterise 'natural' interactivity allows validation that naturally synchronous performance can be achieved using the VIIVA-NMP audio system, and provides an objective estimate of immersive quality. In that the technical design of the VIIVA-NMP audio system provides suitable Immersive Audio and is designed for operation with VR technology, validating that 'natural' musical interactivity is achievable is the final point in accepting the hypothesis: ***It will be possible to design and implement an audio system which is suitable for Immersive NMP vocal performance.***

Synchrony analysis is accompanied by response to the Performance Experience Questionnaire, which provides indicatory compliment to synchrony analysis by estimating quality of Perceived Immersion.

Repeating the performance task for each of the three different rooms detailed in Chapter 3 addresses the sub-objective: ***To conduct pilot study of performance synchrony and EPT for vocal performance under the influence of varying virtual acoustic reverberant performance spaces.*** This measure is predominantly intended to simply identify whether or not the potential effect indicated in previous research is present. If so, reverberation can be identified as a suitable topic for further study.

4.1 Participant, Hardware, Setup and Session Data

10 participants took part in the testing protocol, 9 participants from recruitment via email contact, and the lead researcher. Participants were organised into duet pairs for the testing protocol. Each participant was allowed to choose their own 'codename' for the purposes of anonymity. These 'codenames' were then translated to usable IDs for the purposes of data processing. Each participant ID is defined with a capital letter, denoting the group index, and a lower case letter indicating designation within the pair (s, server or c, client). The technical parameters for each group are retrieved from the Hardware and Setup Questionnaire responses, and notes taken by the lead researcher during the setup and test procedures (*Table 4.1.1*). The 5 pairs of participants took part in testing remotely from various locations across Europe (*Figure 4.1.1*). Glasgow (Scotland) was used as a common location to connect for location pairing. Other locations were York (England), Oslo (Norway) and Barcelona (Spain).

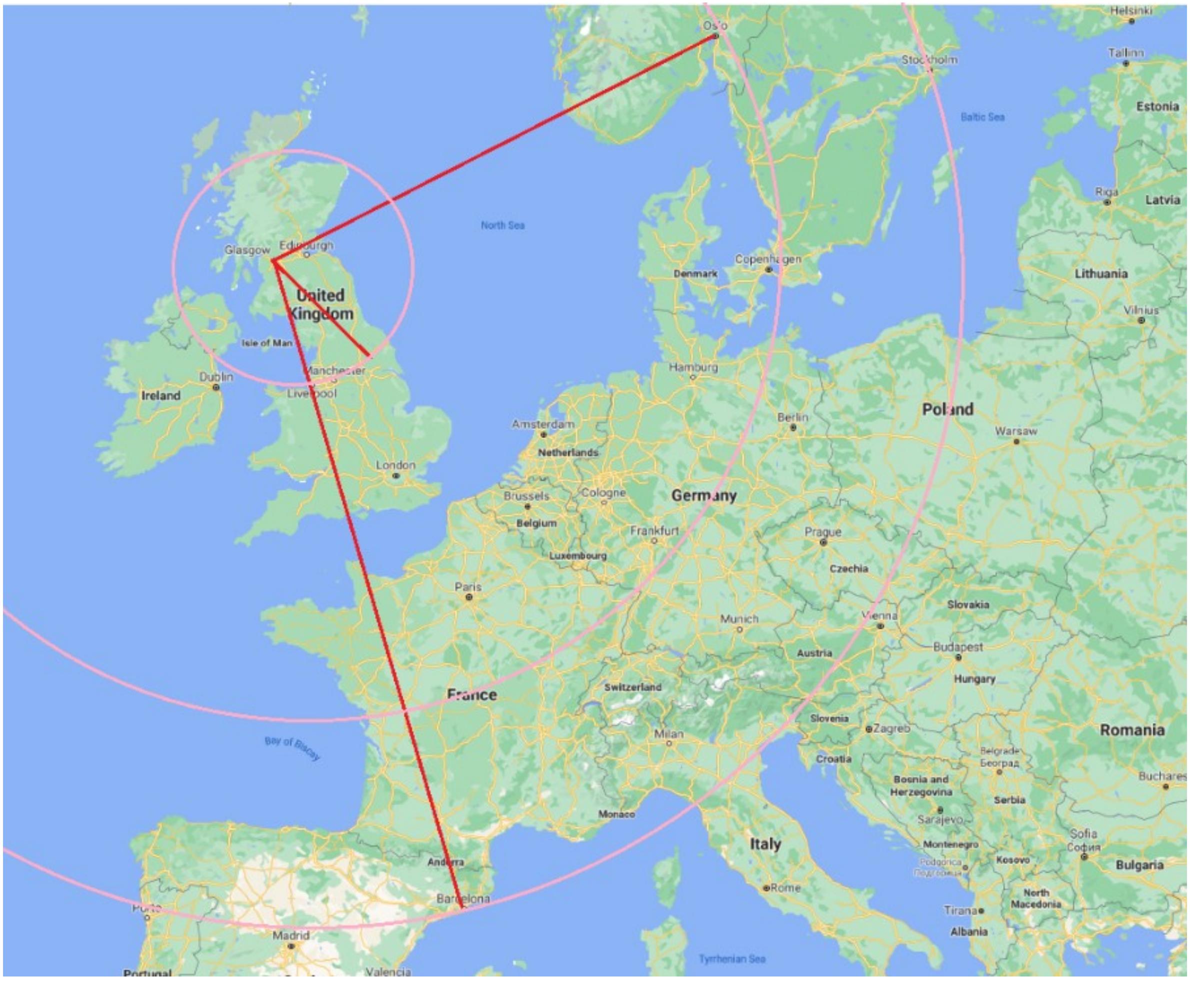


Figure 4.1.1 Remote testing location across Europe.

Group	***A***		***B***		***C***		***D***		***E***	
ID	***As***	***Ac***	***Bs***	***Bc***	***Cs***	***Cc***	***Ds***	***Dc***	***Es ****	***Ec***
Location	Oslo	Glasgow	Glasgow	York	Glasgow	Glasgow	Glasgow	Barcelona	Glasgow	York
Distance	~1000km		~300km		~1km		~1700km		~300km	
Network Latency (RTT)	83 ms		46 ms		39ms		68ms		40 ms	
Local Latency	13.3ms		6.7 ms		3.3 ms		3.3 ms		13.3 ms	
Estimated OWT	54.8 ms		29.7 ms		22.8 ms		37.3 ms		33.3 ms	
Bandwidth Estimate	~100Mb/s	~10Mb/s	~10Mb/s	~15Mb/s	~10Mb/s	~20Mb/s	~10Mb/s	~90Mb/s	~10Mb/s	~300Mb/s
Estimated Bandwidth Used	~7.0 Mb/s		~4.7Mb/s		~4.7Mb/s		~4.7Mb/s		~3.5 Mb/s	
Sample Rate	96kHz		96kHz		96kHz		96kHz		48kHz	
Buffer Size	256		128		64		64		128	
Jitter Buffer Size	3072		1024		640		1408		512	
Jitter Buffer (ms)	32		10.667		6.667		14.667		10.667	
FEC Redundancy	2		1		1		1		2	
Audio Interface	18i20 g2	18i20 g1	18i20 g1	2i2 g2	18i20 g1	828 mk2	18i20 g1	Red 4 Pre	18i20 g1	2i2 g3
Headphones	HD25	DT990	DT990	DT990	DT990	DT770	DT 990	DT 990	DT 990	DT 990
Microphone	AT2050	SM57	NT1a	MM1	NT1a	M2	NT1a	KM148	SM57	DPA6066
Operating System	Win10	Win10	Win10	Win10	Win10	Snow Leopard	Win10	Catalina	Win10	Win10
Singer Proficiency Level	Proficient	Amateur	Proficient	Amateur	Proficient	Proficient	Proficient	Professional	Amateur	Professional
Sung with Paired Participant before?	No		No		No		No		No	

Table 4.1.1 Session, Hardware and Setup details for all instances of the Test Protocol. * Indicates the ID associated with the Lead Researcher, who was one of the participants. A hardware list is provided below detailing specified models abbreviated in the table.

Table 4.1.1 Hardware List

Headphone List:

- *Beyerdynamic DT990*
- *Beyerdynamic DT770*
- *Sennheiser HD 25*

Microphone List:

- *Audio Technica AT2050*
- *Beyerdynamic MM1*
- *Sure SM57*
- *Rode NT1a*
- *Rode M2*
- *Neumann KM148*
- *DPA 6066 Omni*

Interface List:

- *Focusrite Scarlett 18i20 generation 1*
- *Focusrite Scarlett 18i20 generation 2*
- *Focusrite Scarlett 2i2 generation 2*
- *Focusrite Scarlett 2i2 generation 3*
- *Focusrite Red 4 Pre*
- *Motu 828 mk2*

4.1.1 Headphones, Microphone and Audio Interface

Audio interfaces used in testing (*Table 4.1.1*) were all of professional audio quality, and it was considered that variation here was highly unlikely to influence the results of testing. Headphones used in testing predominantly consisted of Beyerdynamic DT 990s, as this model was specified as preferred if possible. The two discrepancies are Beyerdynamic DT 770 and Sennheiser HD 25 models.

DT 990 [253], DT 770 [254] and HD 25 [255] models all have varying response. Indeed it can also be considered that the transfer function of headphones changes with each mounting instance [187]. As previously noted, headphone response can affect the quality of binaural reproduction. Though it was beyond the scope of this study to measure the transfer function associated with headphone mounting for each performance, this was not expected to significantly affect performance synchrony. As slight lack of control is, however, present in perceptual evaluation.

One significant difference, however, is apparent in the headphone design: DT 990 features an open-back, whereas DT 770 and HD 25 are closed-back models. The difference between open and closed back designs is specifically that closed back designs will occlude the direct sound of a singers own voice in the real world than open back headphones.

Microphone selection showed more hardware variance than with headphones and audio interface. Though all microphones are of professional audio quality as requested, it can be noted that there is variation in quality, pickup pattern, frequency response, and dynamic/condenser type. This variation is recognised as a potential control issue in the study. It is recognised that this will affect the perception of acoustic simulation and spatial audio delivery.

Though hardware variance allows identification of potential control issues in the perception of audio rendering in the testing protocol it is considered firstly that cursory investigation did not highlight any particular issues in the testing procedure. It is secondly considered that this control issue is likely only to be apparent in the Performance Experience Questionnaire and repeated measurements with different virtual acoustic spaces. As these components of the test protocol aim to provide a broad investigation it was considered unlikely that hardware variation would significantly skew these results. It was, however, duly noted that any future work involving parametric reverberation investigation or thorough investigation of Perceived Immersion will require that hardware variation is strictly controlled.

Hardware variance reduces control in this study to a degree, however it is expected, though there is room for improvement, that control is still satisfactory in the context of this study.

4.1.2 Singer Experience

No participant pair reported prior experience of singing together. There was, however, variance in the reported Singer Proficiency Level. It was considered that this presents a potential control issue in one of two ways: Professional level singers will be well practiced in adapting their own performance and providing this performance with stable tempo. This could mean that professional performers exhibit greater synchrony than less experienced counterparts. Alternatively it has been considered that professional singers may be more sensitive to differences from live singing conditions they are well naturalised to. In this instance it is possible that this potential sensitivity could cause performance by professional level singers to be more disrupted by performance conditions than less experienced counterparts.

4.1.3 Latency and Bandwidth Measurement

OWT latency measurement in this study (*Table 4.1.1*) consisted of 5 discrete values, each associated with a participant pair (identifiable by group): Group A (54.8ms), Group B (29.7ms), Group C (22.8ms), Group D (37.3ms) and Group E (33.3ms).

Results from this study (*Table 4.1.1*) demonstrate estimated Available Bandwidth ranging from 10Mb/s-300Mb/s, and estimated Bandwidth Used ranging from 3.5-7.0Mb/s.

4.1.4 Discussion: Participant, Hardware, Setup and Session Data

Several potential control issues have been identified in hardware and setup data, namely Singer Experience, as well as microphone and headphone choice. Measurement of Bandwidth Available can of course be more robust, and consideration of clock synchronization between remote sites should be included in latency measurements in future research.

Hardware measurements provide some useful indication of the technical considerations of practical deployment. Estimated Bandwidth Used can be considered a guideline for the cost of each remote performer connection for Immersive NMP audio systems. In this study practical deployment to test participant homes required bandwidths from 3.5-7.0 Mb/s. The available bandwidth measured in this study ranged from 10-300 Mb/s. This would imply that in a practical use case scenario using typical home internet connections Immersive NMP audio systems following the VIIVA-NMP audio system design will be capable of facilitating ensemble sizes of 2 to ~40 performers. Considering Immersive NMP proposes avatar rendering visual accompaniment in VR/AR systems, some network overhead must be allowed for associated metadata. For this reason the upper ensemble size limit estimate should be reduced to a more conservative ~30 performers. This bandwidth usage estimate allows for clarification of estimates from current Immersive NMP research which would suggest a bandwidth consumption of 1Mb/s is practical [10]. Estimates such as 1Mb/s are reserved for 16-bit audio at a sample rate of 44.1kHz or 48kHz with no error correction, which is only achievable on networks with zero packet loss.

Bandwidth consumption can of course be minimised through sample rate and resolution reduction, however sample rate reduction can be shown by Local Latency measurement to reduce the potential system range by increasing Local Latency, thereby reducing the latency overhead allowance allocated to network transport.

4.2 Synchrony Metrics

Synchrony metrics retrieved from the performance tasks were:

- Mean Tempo Slope (within parts)
- Mean Tempi Ratio (between parts)
- Tendency to Lead
- Asynchrony
- Precision

From Mean Tempo Slope (within parts) data a further metric is calculated:

- Mean Tempo Slope Difference (detailing the difference between parts in a performance)

This data is recorded for each undertaking of the performance task, and analysed with respect to two predictors:

- Latency (as identified by group ID)
- Room (as identified by room index)

It should be noted that in this study each discrete latency is associated with a discrete group. In this regard latency measurement may represent a nominal category (group) or a numerical value (latency). For this reason data is presented in this thesis in the form Group (latency) to identify these populations, where Group is the group identified in Table 4.1.1, and latency is the associated Estimated OWT from the same table.

Analysis of synchrony data allows characterisation of the level of musical interactivity between performers. First data is analysed with respect to Mean Tempo Slope Difference and Tendency to Lead across groups in order to identify cases where a Leader-Follower relationship is present.

Tempo Slope, Tempi Ratio, Asynchrony and Precision are then analysed in order to identify whether or not these metrics characterise musical interactivity using the VIIVA-NMP audio system prototype as ‘natural’. Where this is the case the hypothesis of this thesis can be effectively accepted.

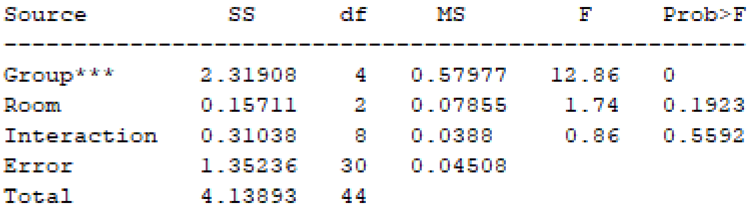
4.2.1 Mean Tempo Slope Difference

Mean Tempo Slope Difference (*Table 4.2.1.1*) close to zero is considered to characterise 'natural' interactivity, as in live performance parts should move at approximately equal tempo. Significant deviation from this expected value therefore represents a collapse in synchrony associated with high latency, where performers are simply performing with vastly different rate of tempo variation, or a leader-follower latency coping mechanic, where one performer is attempting to respond to noticeable tempo deceleration in the other part and influence a more stable tempo.

Group	**Room**	N	**Mean**	**Median**	**Std. Deviation**	**Max**	**Min**
A	**A***	*3*	*0.874*	*0.876*	*0.087*	*0.960*	*0.787*
	B	*3*	*0.934*	*0.903*	*0.272*	*1.220*	*0.678*
	C	*3*	*0.703*	*0.442*	*0.527*	*1.310*	*0.358*
B	**A**	*3*	*0.136*	*0.091*	*0.124*	*0.276*	*0.040*
	B	*3*	*0.296*	*0.245*	*0.181*	*0.497*	*0.146*
	C	*3*	*0.417*	*0.302*	*0.215*	*0.666*	*0.046*
C	**A**	*3*	*0.146*	*0.166*	*0.100*	*0.235*	*0.038*
	B	*3*	*0.409*	*0.510*	*0.240*	*0.583*	*0.136*
	C	*3*	*0.604*	*0.339*	*0.111*	*0.394*	*0.180*
D	**A**	*3*	*0.152*	*0.0731*	*0.151*	*0.326*	*0.057*
	B*	*3*	*0.247*	*0.279*	*0.060*	*0.285*	*0.178*
	C	*3*	*0.462*	*0.404*	*0.167*	*0.650*	*0.332*
E	**A**	*3*	*0.199*	*0.152*	*0.108*	*0.323*	*0.123*
	B	*3*	*0.259*	*0.152*	*0.230*	*0.523*	*0.103*
	C	*3*	*0.235*	*0.155*	*0.154*	*0.413*	*0.139*

Table 4.2.1.1 Mean Tempo Slope Difference results, measured in BPM/s. * Indicates a standard error or 0.05 or less.

Two-Way ANOVA (***Table 4.2.1.2***) of data sorted by Group (latency) and Room demonstrates significant difference across Group (latency), and that no effect of room is present in Mean Tempo Slope Difference.



```
Source            SS        df      MS          F      Prob>F
-------------------------------------------------------------
Group***        2.31908      4    0.57977    12.86    0
Room            0.15711      2    0.07855     1.74    0.1923
Interaction     0.31038      8    0.0388      0.86    0.5592
Error           1.35236     30    0.04508
Total           4.13893     44
```

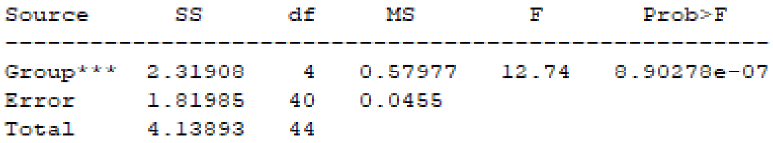
Table 4.2.1.2 Two-Way ANOVA of Mean Tempo Slope Difference sorted by Group (latency) and Room. *** indicates a significant difference between populations evaluated at p = 0.001.

Sorting data by Group (latency) including all Room data (*Table 4.2.1.3*), ANOVA (*Table 4.2.1.4, Figure 4.2.1.1*) demonstrates a clear significant difference between Group A (54.8ms) measurements and values for each other Group (latency) evaluated at a significance of $p = 0.001$.

Worth noting here is that for data sorted this way, Group E represents a not-normal distribution as indicated by Shapiro-Wilk test evaluated at significance $p = 0.05$, though from the notch plot (*Figure 4.2.1.1*) it is evident that this has no bearing in showing Group A (54.8ms) as significantly different from other Groups. The non-normality is to be expected using N=3 measurements within each combination of independent variables. Fortunately ANOVA analysis is inherently unlikely to yield false-positive results even where non-normal distributions are present in populations and this is unlikely to influence results [256] .

Group	**N**	**Mean**	**Median**	**Std. Deviation**	**Max**	**Min**
A	*9*	*0.837*	*0.876*	*0.317*	*1.310*	*0.358*
B	*9*	*0.283*	*0.276*	*0.197*	*0.666*	*0.40*
C	*9*	*0.287*	*0.235*	*0.182*	*0.583*	*0.038*
D	*9*	*0.287*	*0.285*	*0.180*	*0.650*	*0.057*
E	*9*	*0.0502*	*0.152*	*0.151*	*0.523*	*0.103*

Table 4.2.1.3 Results for Mean Tempo Slope Difference measured in BPM/s when sorted by Group (latency) including all room data.



```
Source        SS       df      MS          F        Prob>F
-------------------------------------------------------------
Group***   2.31908     4    0.57977    12.74    8.90278e-07
Error      1.81985    40    0.0455
Total      4.13893    44
```

Table 4.2.1.4 ANOVA of Mean Tempo Slope Difference by Group (latency) including data for all Rooms. *** indicates significance evaluated at p = 0.001

Cursory observation of the plotted Tempo traces for performance by this group (*Figure 4.2.1.2*) demonstrates that the difference in Mean Tempo Slope between performers for Group A (54.8ms) is likely due to the performer designated ‘Client’ in the group accelerating tempo in response to noticeable deceleration within the ‘Server’ part.

Mean Tempo Slope Difference therefore allows identification of a leader-follower latency-coping strategy in Group A (54.8ms). The magnitude of this difference would suggest partial synchrony collapse for this group also. Other Groups all exhibit Mean Tempo Slope Difference with means less than 0.3 BPM/s, which is assumed suitably close to zero as not to characterise any latency-coping strategies.

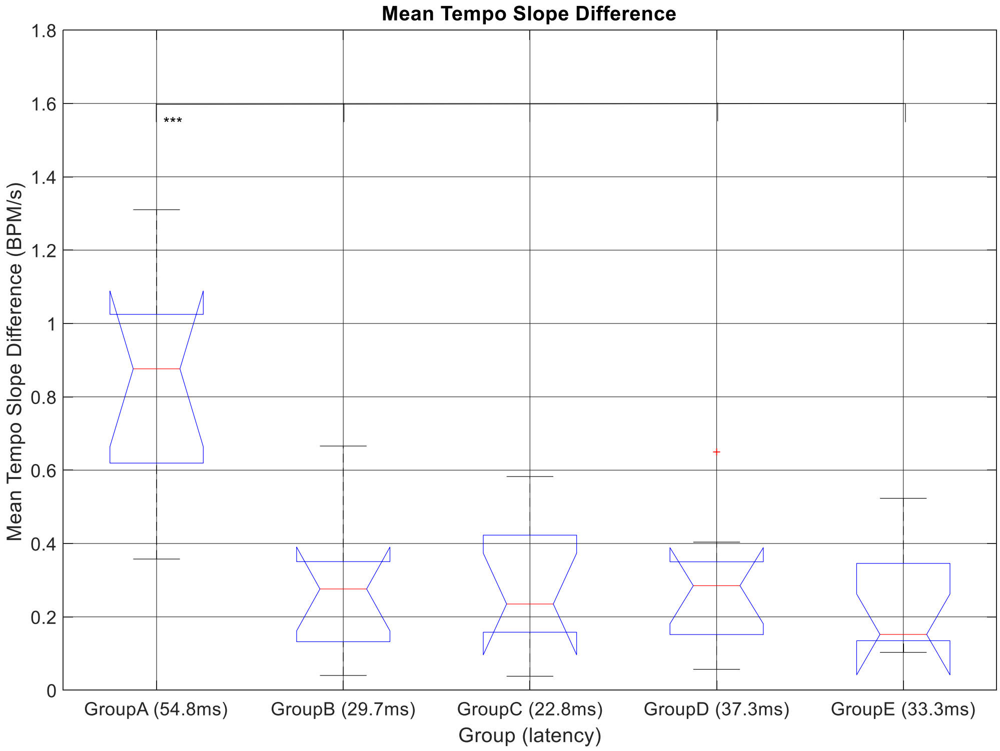


Figure 4.2.1.1 Notch plot of Mean Tempo Slope Difference sorted by Group (latency) including data for all rooms. *** indicates that Group A (54.8ms) can be observed as significantly different from other groups, evaluated at a significance of p=0.001.

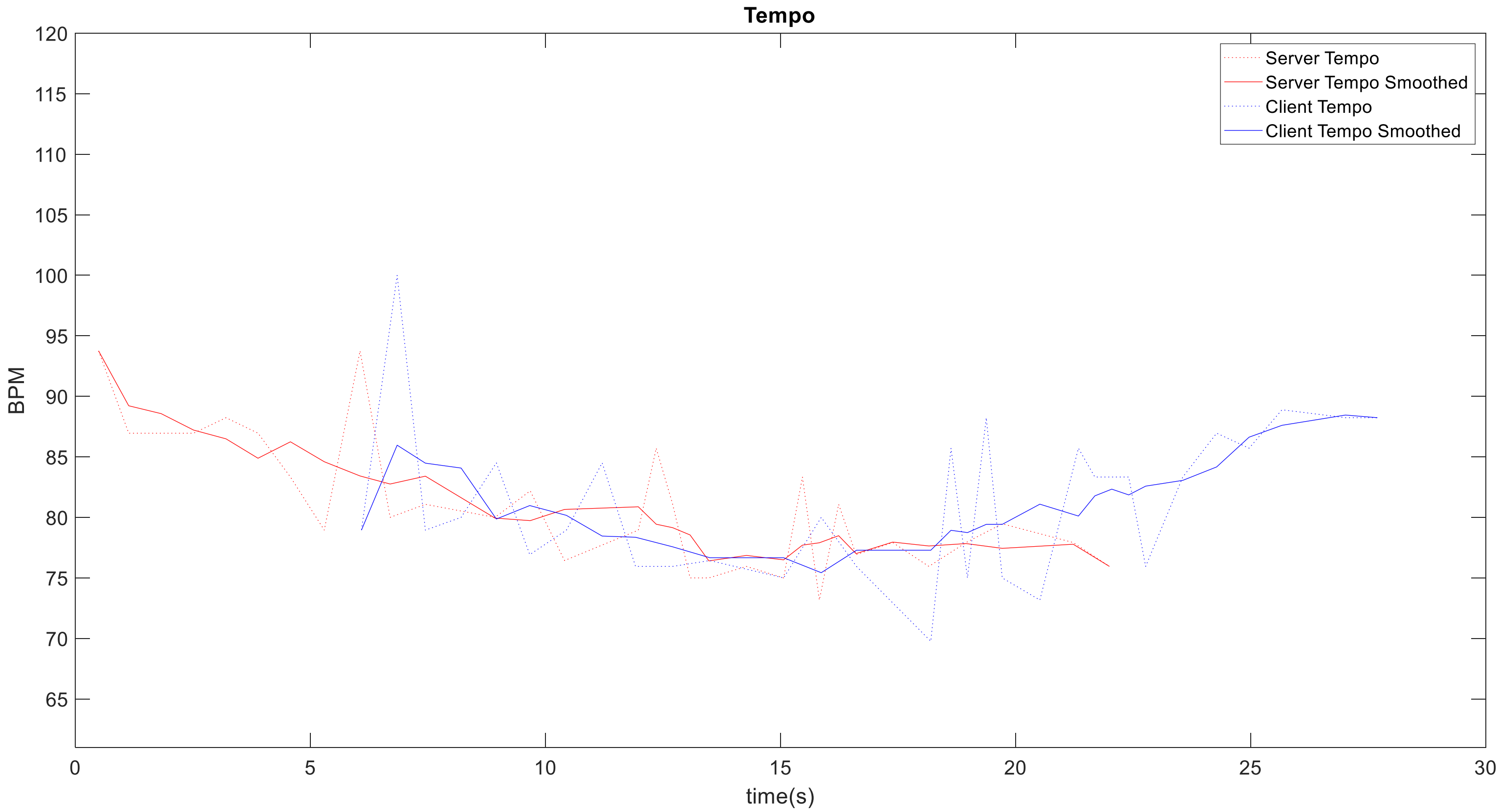


Figure 4.2.1.2 Tempo Trace sample (Group A, Room B, Take 1) demonstrating tempo difference associated with the Client adopting the 'Leader' designation and speeding up to counteract tempo deceleration, and Server stabilising tempo by adopting a 'Follower' designation.

4.2.2 Tendency to Lead

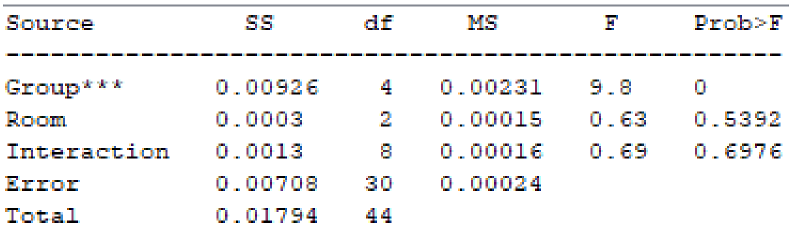
From previous study by D'Amario et al. [59] a Tendency to Lead between -20ms and 40ms was selected as value for Signed Tendency to Lead characterising natural interactivity. These thresholds can therefore be considered identification of a Leader-Follower level of interactivity, where the latency compensation relies on an exaggerated Tendency to Lead.

Notably of results for Tendency to Lead in this study, measured in milliseconds, ms, (*Table 4.2.2.1*) several populations represent non-normal distribution from Shapiro-Wilk testing, again likely due to analysing at N=3. Namely these populations of data are Group A (54.8ms), Room A, Group B (29.7ms), Room B and Group C (22.8ms), Room A. For Group D (37.3ms), Room C measurement yields repeated values, where with N=3 no testing of normality or significance is therefore possible.

Group	Room	N	Mean	Median	Std. Deviation	Max	Min
A	A	3	0.023	0.015	0.015	0.040	0.015
	B*	3	0.021	0.030	0.019	0.035	0.000
	C*	3	0.015	0.015	0.020	0.030	0.010
B	A**	3	0.001	0.005	0.015	0.015	-0.015
	B	3	-0.008	-0..005	0.014	-0.005	-0.030
	C**	3	0.006	0.005	0.008	0.015	0.000
C	A	3	0.006	0.008	0.003	0.008	0.003
	B**	3	-0.011	-0.017	0.017	0.008	-0.022
	C**	3	-0.007	0.002	0.009	-0.002	-0.017
D	A*	3	-0.023	-0.018	0.033	0.007	-0.058
	B**	3	-0.029	-0.033	0.010	-0.018	-0.038
	C	3	-0.028	-0.028	n/a	-0.028	-0.028
E	A**	3	-0.008	-0.012	0.010	0.003	-0.017
	B**	3	0.002	-0.002	0.010	0.013	-0.007
	C*	3	-0.007	-0.012	0.018	0.013	-0.022

Table 4.2.2.1 Results for Tendency to Lead, representing the time in milliseconds by which the performer designated 'server' precedes the performer designated 'Client'. * Indicates a standard error less the 0.05, and ** indicates a standard error of less than 0.01.

Two-way ANOVA of Tendency to Lead sorted by Group (latency) and Room (***Table 4.2.2.2***) demonstrates clear significant difference with varying Group (latency), though no associated variance with Room.



```
Source            SS       df      MS         F      Prob>F
-----------------------------------------------------------
Group***       0.00926     4    0.00231     9.8     0
Room           0.0003      2    0.00015     0.63    0.5392
Interaction    0.0013      8    0.00016     0.69    0.6976
Error          0.00708    30    0.00024
Total          0.01794    44
```

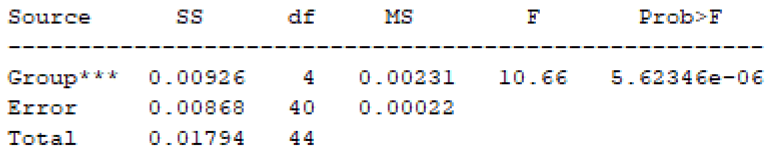
Table 4.2.2.2 Two-way ANOVA for Tendency to Lead sorted by Group (latency), columns, and Room. *** indicates significance evaluated at p = 0.001.

Sorting data by Group (latency) including all Room data for each group (*Table 4.2.2.3*) removes non-normalities in data by analysing at N=9. ANOVA of data sorted by Group (latency) (*Table 4.2.2.4, Figure 4.2.2.1*) demonstrates that both Group A (54.8ms) and Group D (37.3ms) exhibit significantly different results (evaluated at p = 0.001), with means of 19ms and -27ms respectively.

Group	**N**	**Mean**	**Median**	**Std. Deviation**	**Max**	**Min**
A**	*9*	*0.019*	*0.015*	*0.017*	*0.040*	*-0.010*
B**	*9*	*-0.002*	*0.000*	*0.014*	*0.015*	*-0.030*
C**	*9*	*-0.004*	*-0.002*	*0.012*	*0.008*	*-0.022*
D**	*9*	*-0.027*	*-0.028*	*0.017*	*0.007*	*-0.058*
E**	*9*	*-0.004*	*-0.007*	*0.013*	*0.013*	*-0.022*

Table 4.2.2.3 Results for Tendency to Lead (measured in seconds, s) sorted by Group (latency) including all Room data. ** indicates a standard error of 0.01 or less.

Group A (54.8ms) has already been identified as latency-coping and operating with partial collapse of synchrony by analysis of Mean Tempo Slope Difference. It is likely the low mean here is the result of uneven tempo ‘cancelling out’ across performances, given that results for this group has a min-max range of 50ms and comparatively high standard deviation of 17ms. Group D (37.3ms) demonstrates a mean Tendency to Lead which exceeds the thresholds associated with ‘natural’ interactivity, and can be classified as using a leader-follower latency coping strategy. Groups B (29.7m), Group C (22.8ms) and Group E (33.3ms) all demonstrate Tendency to Lead ranging from means of -2ms to -4ms. In these cases no leader-follower technique is detected and performance can be classed as ‘natural’ in this regard.



```
Source       SS       df     MS         F        Prob>F
-------------------------------------------------------
Group***   0.00926     4   0.00231   10.66   5.62346e-06
Error      0.00868    40   0.00022
Total      0.01794    44
```

Table 4.2.2.4 One-Way ANOVA of Tendency to Lead by Group (latency) for including all room data. ***indicates significance evaluated at p = 0.001

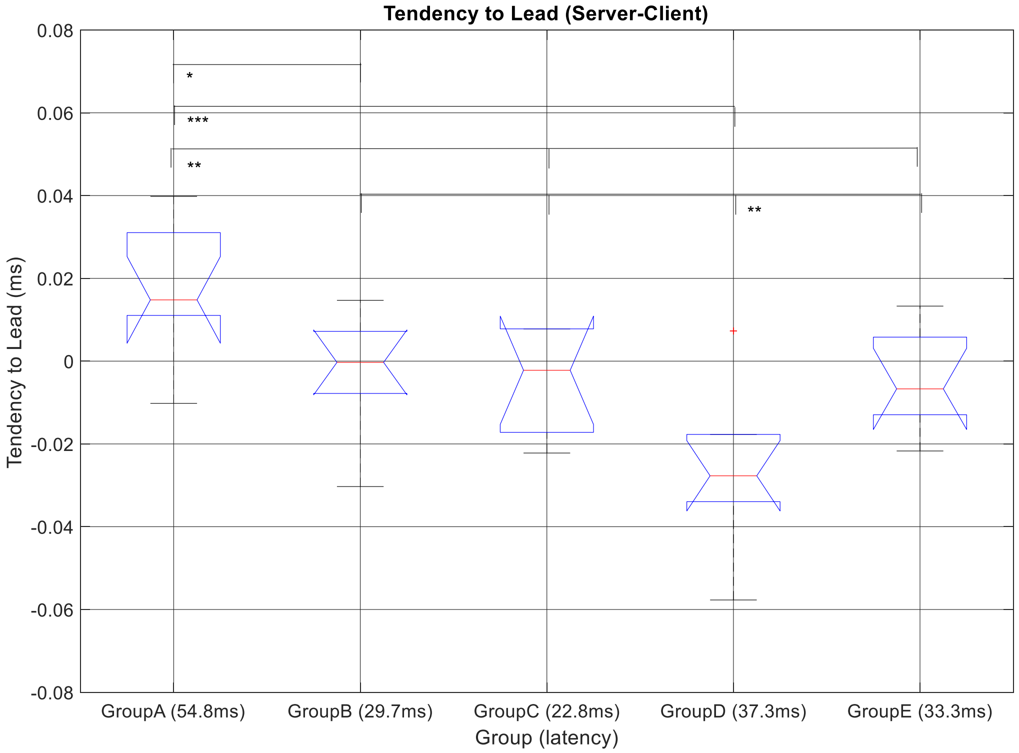


Figure 4.2.2.1 Notch Plot of Tendency to Lead by Group (latency) including all room Data. ***, **, and * indicate significance evaluated at p = 0.001, 0.01 and 0.05 respectively.

Analysing Tendency to Lead as a signed value allows identification of which performer in the group is the Leader. Notably in Group A (54.8ms) the Server tends to lead, in Group D (37.3ms) the Client adopts the Leader designation. Linear modelling of Tendency to Lead results over latency (*Tables 4.2.2.5 - 4.2.2.7*, *Figure 4.2.2.2*) is most significantly expressed as Absolute Tendency to Lead, detailing the magnitude of the Leader-Follower time gap (evaluated at p = 0.05). A plateu can be observed where groups operating above 35ms appear more likely to have a higher magnitude Tendency to Lead, associated with Leader-Follower strategies. It should be noted that this plot only represents data collected in this study, which is too sparse for robust modelling of the effect of latency on Tendency to Lead as a wider Effect.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.364[a]	.132	.112	.011750

a. Predictors: (Constant), Latency

Table 4.2.2.5 Model Summary for Linear Regression of Absolute Tendency to Lead over latency.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.001	1	.001	6.563	.014[b]
	Residual	.006	43	.000		
	Total	.007	44			

a. Dependent Variable: AbsTendencyToLead

b. Predictors: (Constant), Latency

Table 4.2.2.6 Associated ANOVA for Linear Regression of Absolute Tendency to Lead over latency.

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	.001	.006		.204	.839
	Latency	.000	.000	.364	2.562	.014

a. Dependent Variable: AbsTendencyToLead

Table 4.2.2.7 Coefficients for Linear Regression of Absolute Tendency to Lead over latency.

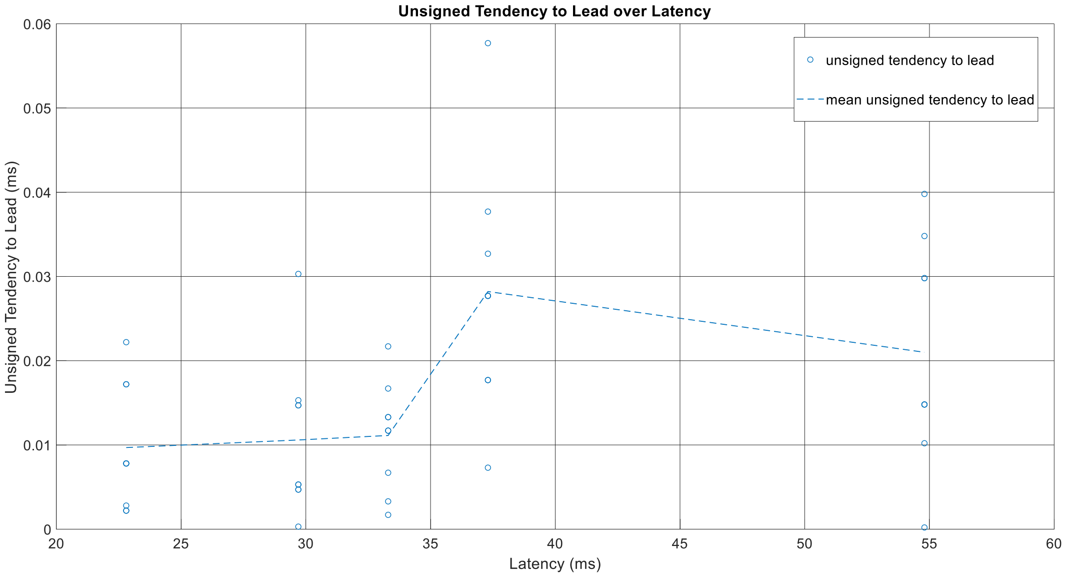


Figure 4.2.2.2 Linear plot of Absolute Tendency to Lead illustrating a plateau around 35ms in Tendency to Lead data recorded in this study.

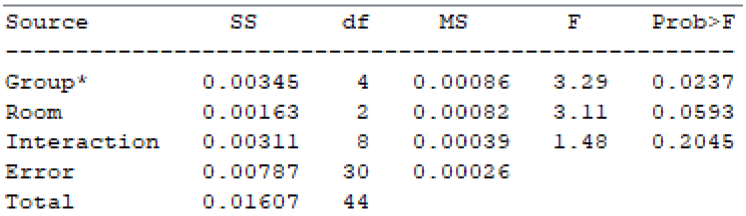
4.2.3 Mean Tempi Ratio

Data analysis of Mean Tempi Ratio (***Table 4.2.3.1***) describes the ability of performer pairs to perform 'in time', specifically at the same tempo. Performers who are absolutely in time with one another are expected to exhibit a Mean Tempi Ratio of 1. It is worth noting that Mean Tempi Ratio does not describe tempo deceleration or acceleration, simply how 'together' the tempo delivered by performer pairs is. The ratio is expressed with respect to the server/client performer.

Group	Room	N	Mean	Median	Std. Deviation	Max	Min
A	**A***	*3*	*1.003*	*1.001*	*0.022*	*1.023*	*0.979*
	B**	*3*	*1.012*	*1.013*	*0.011*	*1.023*	*1.001*
	C*	*3*	*1.015*	*1.024*	*0.021*	*1.030*	*0.991*
B	**A****	*3*	*0.980*	*0.982*	*0.010*	*0.989*	*0.970*
	B**	*3*	*1.004*	*1.003*	*0.008*	*1.012*	*0.667*
	C*	*3*	*1.017*	*1.022*	*0.019*	*1.033*	*1.000*
C	**A***	*3*	*0.990*	*0.986*	*0.029*	*1.020*	*0.963*
	B**	*3*	*0.990*	*0.990*	*0.003*	*0.993*	*0.990*
	C*	*3*	*1.008*	*1.014*	*0.022*	*1.025*	*0.983*
D	**A***	*3*	*0.986*	*0.988*	*0.017*	*1.003*	*0.968*
	B**	*3*	*0.969*	*0.976*	*0.015*	*0.979*	*0.951*
	C*	*3*	*1.000*	*1.003*	*0.018*	*1.016*	*0.981*
E	**A****	*3*	*1.007*	*1.005*	*0.006*	*1.014*	*1.003*
	B**	*3*	*1.013*	*1.015*	*0.004*	*1.017*	*1.008*
	C**	*3*	*0.998*	*1.002*	*0.013*	*1.009*	*0.985*

Table 4.2.3.1 Mean Tempi Ratio results. * and ** indicate a standard error of 0.05 and less or 0.01 and less respectively.

Two-way ANOVA (*Table 4.2.3.2*) of data sorted by Group (latency) and Room demonstrates a significant effect of Group (latency), evaluated at p = 0.05. It is worth noting that at 0.056 an effect associated with Room is close to significant evaluated at the same p value.



```
Source           SS       df      MS         F      Prob>F
----------------------------------------------------------
Group*         0.00345     4   0.00086    3.29    0.0237
Room           0.00163     2   0.00082    3.11    0.0593
Interaction    0.00311     8   0.00039    1.48    0.2045
Error          0.00787    30   0.00026
Total          0.01607    44
```

Table 4.2.3.2 Two-way ANOVA of Mean Tempi Ratio sorted by Group (latency) and Room. * indicates significance evaluated at p = 0.05.

Repeated Anova measuring Group (latency) as the main effect (*Table 4.2.3.3*) demonstrates that both Group (latency), and the product of Group (latency) and Room, are significant (evaluated at $p = 0.05$) in explaining Mean Tempi Ratio values within subjects. It is worth noting that the significant difference between Groups here may not indicate an effect associated strictly with latency, as this effect may have other influences, such as personal preferences [38] and perceptual experience within performer pairs.

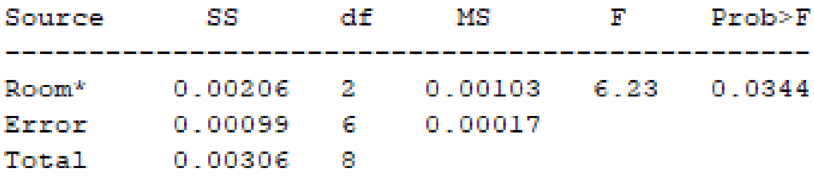
Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
GroupLatency	Sphericity Assumed	.003	4	.001	6.051	.002	.502
	Greenhouse-Geisser	.003	2.480	.001	6.051	.009	.502
	Huynh-Feldt	.003	4.000	.001	6.051	.002	.502
	Lower-bound	.003	1.000	.003	6.051	.049	.502
GroupLatency * Room	Sphericity Assumed	.003	8	.000	2.728	.027	.476
	Greenhouse-Geisser	.003	4.960	.001	2.728	.061	.476
	Huynh-Feldt	.003	8.000	.000	2.728	.027	.476
	Lower-bound	.003	2.000	.002	2.728	.144	.476
Error(GroupLatency)	Sphericity Assumed	.003	24	.000			
	Greenhouse-Geisser	.003	14.880	.000			
	Huynh-Feldt	.003	24.000	.000			
	Lower-bound	.003	6.000	.001			

Table 4.2.3.3 Test of Within-Subject Effects for Mean Tempi Ratio.

Evaluation of the effect of Room within Group (latency) by ANOVA (*Table 4.2.3.4, Figure 4.2.3.1*) demonstrates a significant difference (*evaluated at p = 0.05*) can be identified for Group B (29.7ms). In this case Room B can be identified as providing measurement of Mean Tempi Ratio closer to 1 than for Room A, and Room C is also identified as significantly different from Room A.



```
Source       SS        df      MS          F       Prob>F
---------------------------------------------------------
Room*      0.00206     2     0.00103     6.23     0.0344
Error      0.00099     6     0.00017
Total      0.00306     8
```

Table 4.2.3.4 ANOVA of Mean Tempi Ratio by Room for Group B (29.7ms). * indicates significance evaluated at p = 0.05.

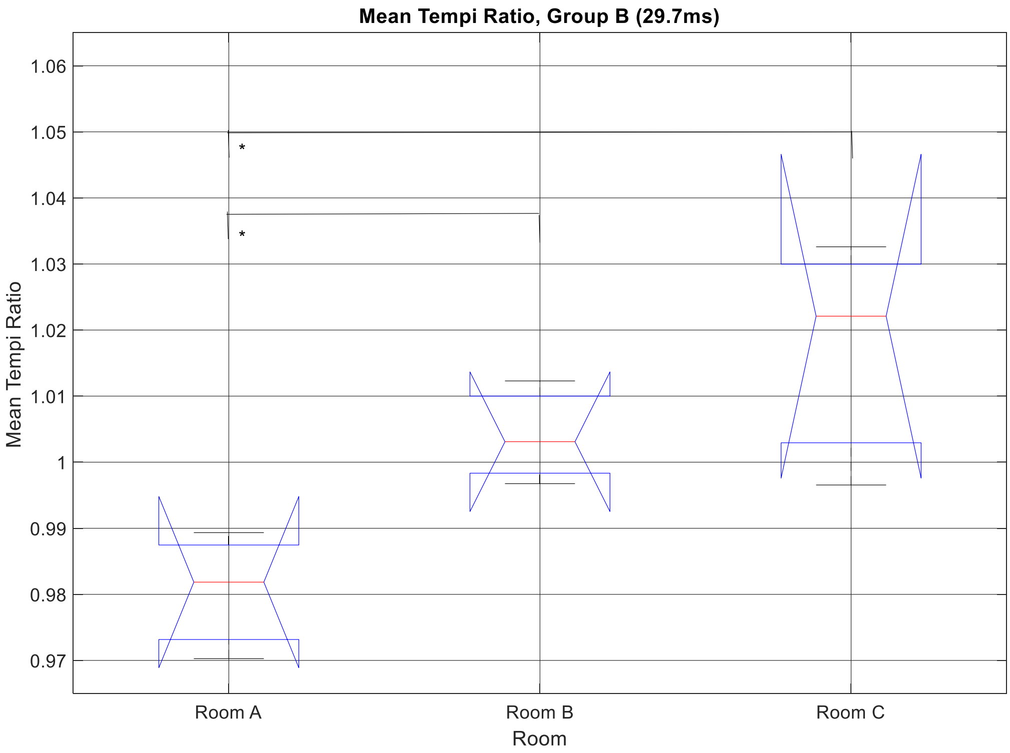


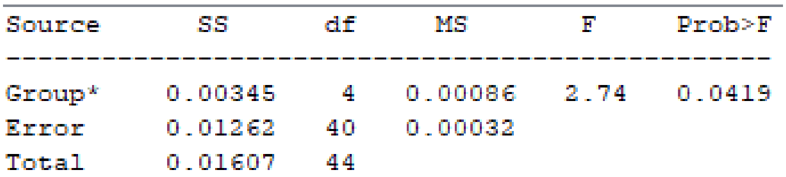
Figure 4.2.3.1 Notch plot of Mean Tempi Ratio by Room for Group B (29.7ms). * indicates significance evaluated at p = 0.05.

It should be noted that this difference, at N=3, may simply just represent the narrowness of measurement. It should also be recognised that the significant difference here indicates a tendency for the Client performer to move faster than the Server performer in Room A, and not a difference in absolute deviation from a ratio of 1 between the pair.

ANOVA (*Table 4.2.3.6, Figure 4.2.*3.2) of Mean Tempi Ratio sorted by Group (latency) including data for all Rooms (*Table 4.2.3.5*) demonstrates some significant difference between Group (latency), evaluated at *p = 0.05*. Again the caveat should be included that these differences are representative of which performer moves faster, rather than difference in magnitude of deviation of Mean Tempi Ratio from 1.

Group	**N**	**Mean**	**Median**	**Std. Deviation**	**Max**	**Min**
A**	*9*	*1.010*	*1.013*	*0.017*	*1.030*	*0.979*
B**	*9*	*1.001*	*0.997*	*0.020*	*1.033*	*0.970*
C**	*9*	*1.000*	*0.990*	*0.020*	*1.025*	*0.963*
D**	*9*	*0.985*	*0.981*	*0.020*	*1.016*	*0.951*
E**	*9*	*1.006*	*1.008*	*0.010*	*1.017*	*0.985*

Table 4.2.3.5 Mean Tempi Ratio sorted by Group (latency) including data for all Rooms. ** indicates standard error of 0.01 or less.



```
Source         SS       df      MS         F      Prob>F
--------------------------------------------------------
Group*      0.00345      4   0.00086     2.74    0.0419
Error       0.01262     40   0.00032
Total       0.01607     44
```

Table 4.2.3.6 ANOVA of Mean Tempi Ratio sorted by Group (latency) including data for all Rooms. * indicates significance, evaluated at p = 0.05.

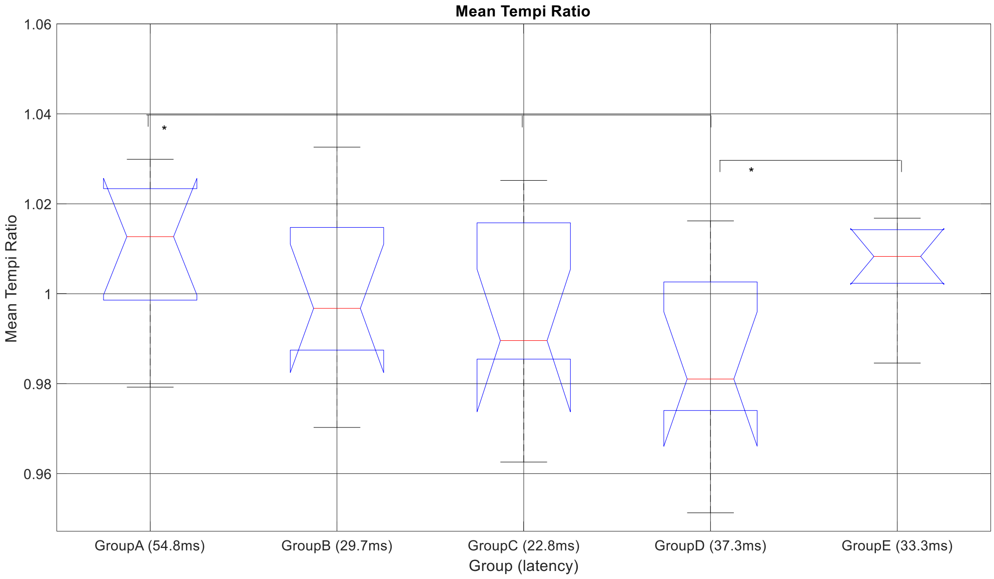


Figure 4.2.3.2 Notch plot of Mean Tempi Ratio sorted by Group (latency) including data for all Rooms. * indicates significance evaluate at p = 0.05.

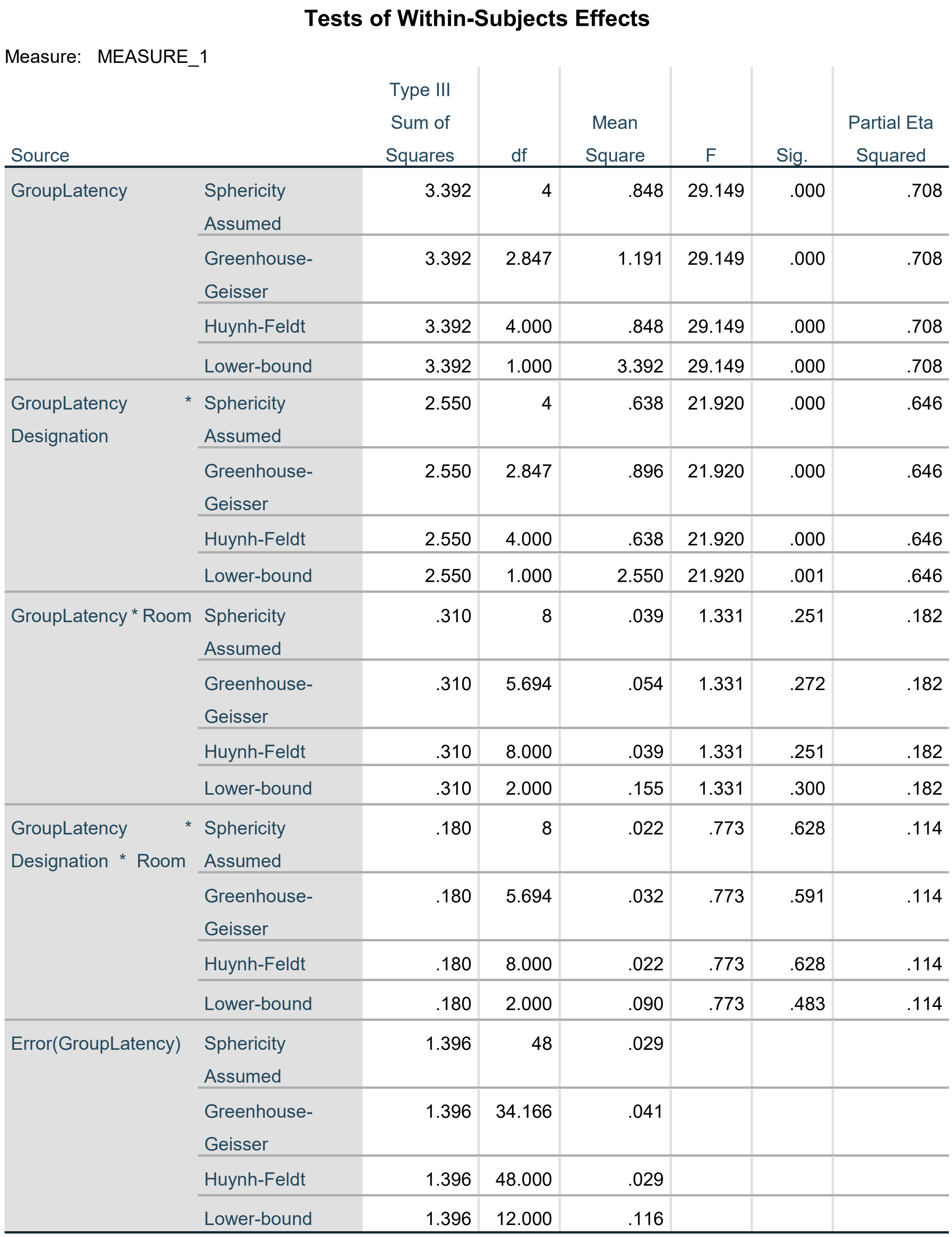
With respect to 'natural interactivity' no recorded values demonstrate deviation from a Mean Tempi Ratio of 1 to classify impairment of natural interactivity.

4.2.4 Mean Tempo Slope

Discrete Tempo Slope measurements (*Table 4.2.4.1*) are made for both Server and Client. As such data is sorted by Performer in this case, identifiable by Group (latency) and designation 'Server' or 'Client'. Notably at N=3 some non-normal data is present. For the Client designation, Group A (54.8ms), Room C, and Group B (29.7ms), Room C are not normal. For Server Group D (37.3ms) Room B and Group E (33.3ms) Room A are also not normal in this measurement set. For Mean Tempo Slope 'natural interactivity' is characterised by Mean Tempo Slope of approximately zero BPM/s (stable tempo). Repeated ANOVA demonstrates that Mean Tempo Slope within subjects is significantly affected by both the Group (latency) and Performer designation (*Table 4.2.4.2*) as evaluated at $p = 0.001$. Room, in this instance, demonstrates no significance.

Group	Performer	Room	N	Mean	Median	Std. Deviation	Max	Min
A	Server	A	3	-0.334	-0.238	0.178	-0.226	-0.539
		B	3	-0.416	-0.410	0.174	-0.245	-0.593
		C	3	-0.121	0.043	0.367	0.136	-0.541
A	Client	A	3	0.540	0.561	0.110	0.638	0.421
		B	3	0.517	0.433	0.260	0.810	0.310
		C	3	0.583	0.493	0.162	0.770	0.485
B	Server	A	3	0.101	0.136	0.089	0.168	0.000
		B	3	-0.065	-0.452	0.077	0.000	-0.150
		C	3	-0.060	0.045	0.216	0.084	-0.308
	Client	A	3	0.177	0.208	0.119	0.276	0.045
		B	3	0.231	0.245	0.124	0.347	0.101
		C	3	0.157	0.358	0.358	0.369	-0.256
C	Server	A*	3	0.134	0.136	0.083	0.215	0.050
		B	3	-0.302	-0.308	0.164	-0.136	-0.463
		C	3	0.152	0.226	0.242	0.347	-0.119
	Client	A	3	0.144	0.098	0.124	0.285	0.050
		B	3	0.107	0.047	0.147	0.275	0.000
		C	3	0.073	0.047	0.135	0.220	-0.047
D	Server	A	3	-0.179	-0.214	0.094	-0.073	-0.251
		B	3	-0.143	-0.214	0.124	0.000	-0.214
		C	3	-0.171	-0.137	0.218	0.027	-0.404
	Client	A	3	-0.331	-0.271	0.222	-0.146	-0.577
		B	3	-0.391	-0.395	0.110	-0.279	-0.499
		C	3	-0.364	-0.469	0.324	0.000	-0.622
E	Server	A*	3	-0.352	-0.396	0.079	-0.261	-0.400
		B	3	-0.460	-0.425	0.178	-0.306	-0.653
		C	3	-0.501	-0.410	0.281	-0.276	-0.816
	Client	A	3	-0.153	-0.109	0.109	-0.073	-0.277
		B	3	-0.201	-0.151	0.105	-0.130	-0.321
		C	3	-0.358	-0.404	0.089	-0.256	-0.404

Table 4.2.4.1 Mean Tempo Slope data, measured in BPM/s. * indicates standard error of 0.05 or less.



Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
GroupLatency	Sphericity Assumed	3.392	4	.848	29.149	.000	.708
	Greenhouse-Geisser	3.392	2.847	1.191	29.149	.000	.708
	Huynh-Feldt	3.392	4.000	.848	29.149	.000	.708
	Lower-bound	3.392	1.000	3.392	29.149	.000	.708
GroupLatency * Designation	Sphericity Assumed	2.550	4	.638	21.920	.000	.646
	Greenhouse-Geisser	2.550	2.847	.896	21.920	.000	.646
	Huynh-Feldt	2.550	4.000	.638	21.920	.000	.646
	Lower-bound	2.550	1.000	2.550	21.920	.001	.646
GroupLatency * Room	Sphericity Assumed	.310	8	.039	1.331	.251	.182
	Greenhouse-Geisser	.310	5.694	.054	1.331	.272	.182
	Huynh-Feldt	.310	8.000	.039	1.331	.251	.182
	Lower-bound	.310	2.000	.155	1.331	.300	.182
GroupLatency * Designation * Room	Sphericity Assumed	.180	8	.022	.773	.628	.114
	Greenhouse-Geisser	.180	5.694	.032	.773	.591	.114
	Huynh-Feldt	.180	8.000	.022	.773	.628	.114
	Lower-bound	.180	2.000	.090	.773	.483	.114
Error(GroupLatency)	Sphericity Assumed	1.396	48	.029			
	Greenhouse-Geisser	1.396	34.166	.041			
	Huynh-Feldt	1.396	48.000	.029			
	Lower-bound	1.396	12.000	.116			

Table 4.2.4.2 Test of Within-Subject Effects for Mean Tempo Slope.

As Mean Tempo Slope shows a significant effect of the measurement for one performer on the measurement for the paired performer, effect of Tempo was organised into discrete Server (*Table 4.2.4.3*) and Client (*Table 4.2.4.4*) categories. For each performer designation data was sorted by Group (latency) including data for all Rooms. ANOVA of Mean Tempo Slope by Group (latency) for the Server Performer (*Table 4.2.4.5, Figure 4.2.4.1*) and the Client Performer (*Table 4.2.4.6, Figure 4.2.4.2*) demonstrate significant difference across Group (latency) for both these independent groups.

Linear Regression of Mean Tempo Slope by Latency for both Server (*Tables 4.2.4.7 – 4.2.4.9*) and Client (*Tables 4.2.4.10 – 4.2.4.12*) performers independently demonstrates it is possible tocreate significant linear models of our data (evaluated at p = 0.05). Linear plotting of means across latency (*Figure 4.2.4.3*) demonstrates that for Group B (29.7ms) Group C (22.8ms) and Mean Tempo Slope is close to zero. Group E (33.3ms) suffers Tempo Deceleration, whereas Group D (37.3ms) suffers from less severe Tempo Deceleration. Group A (54.8ms) demonstrates complete collapse of synchrony in performance, with performer tempo slope varying massively across delivery from discrete parts. Notably in Group B (29.7ms), Group C (22.8ms) and Group E (33.3ms) the Client Performer exhibits a greater Mean Tempo Slope than the Server Performer, however for Group D (37.3ms) the inverse is true.

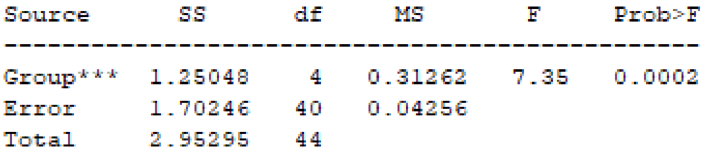
Regarding ‘natural interactivity’ Group B (29.7ms) and Group C (22.8ms) demonstrate tempo slope appropriately close to zero for characterisation as ‘natural’. Other Groups experience Mean Tempo Slope deviates from stable tempo, and can be considered to have impaired musical interactivity between performers.

Group	N	Mean	Median	Std. Deviation	Max	Min
A	*9*	*-0.291*	*-0.245*	*0.258*	*0.136*	*-0.593*
B*	*9*	*-0.008*	*0.000*	*0.148*	*0.168*	*-0.308*
C	*9*	*-0.006*	*0.050*	*0.270*	*0.347*	*-0.463*
D*	*9*	*-0.165*	*-0.214*	*0.135*	*0.027*	*-0.404*
E	*9*	*-0.438*	*-0.400*	*0.183*	*-0.261*	*-0.816*

Table 4.2.4.3 Mean Tempo Slope for the Server Performer. Data is sorted by Group (latency) including data for all Rooms, measured in BPM/s, and * indicates standard error of 0.05 or less.

Group	N	Mean	Median	Std. Deviation	Max	Min
A	*9*	*0.547*	*0.493*	*0.165*	*0.810*	*0.310*
B	*9*	*0.188*	*0.245*	*0.201*	*0.369*	*-0.256*
C*	*9*	*0.108*	*0.050*	*0.121*	*0.285*	*-0.047*
D	*9*	*-0.362*	*-0.395*	*0.206*	*0.000*	*-0.622*
E*	*9*	*-0.237*	*-0.256*	*0.128*	*-0.073*	*-0.415*

Table 4.2.4.4 Mean Tempo Slope for the Client Performer. Data is sorted by Group (latency) including data for all Rooms, measured in BPM/s, and * indicates standard error of 0.05 or less.



```
Source        SS       df      MS        F      Prob>F
-------------------------------------------------------
Group***   1.25048     4    0.31262    7.35    0.0002
Error      1.70246    40    0.04256
Total      2.95295    44
```

Table 4.2.4.5 ANOVA of Mean Tempo Slope by Group (latency) including all Room data for the Server performers. *** indicates significance evaluated at p = 0.001.

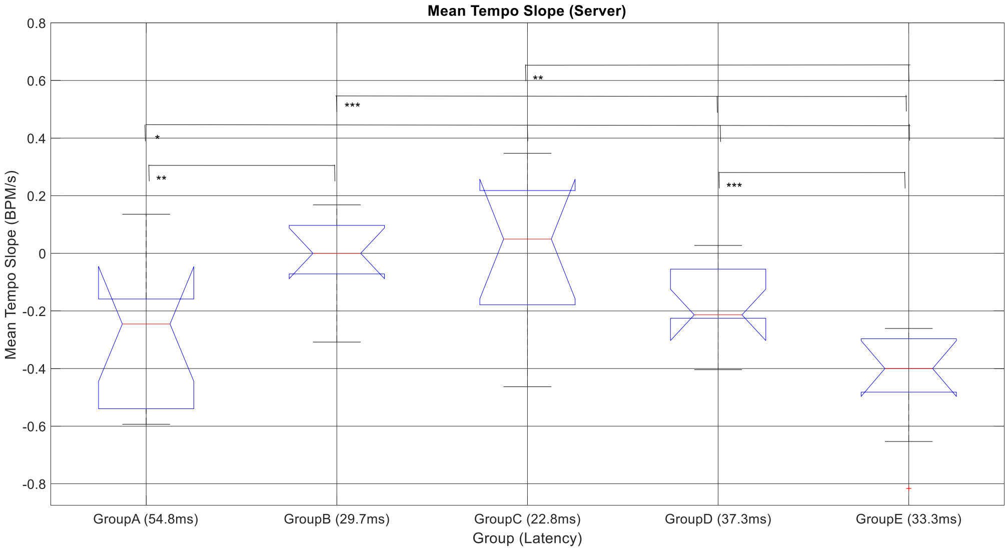
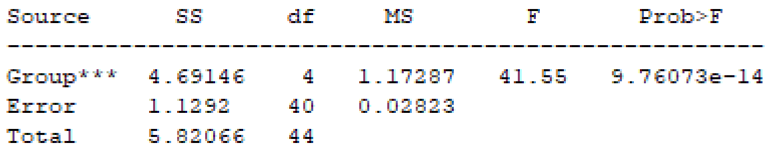


Figure 4.2.4.1 Notch plot of Mean Tempo Slope sorted by Group (latency) for Server Performers. *, ** and *** indicate significance evaluated at p = 0.05, 0.01 and 0.001 respectively.



```
Source        SS        df      MS          F        Prob>F
-------------------------------------------------------------
Group***   4.69146     4    1.17287    41.55    9.76073e-14
Error      1.1292     40    0.02823
Total      5.82066    44
```

Table 4.2.4. 6 ANOVA of Mean Tempo Slope by Group (latency) including all Room data for the Client performers. *** indicates significance evaluated at p = 0.001.

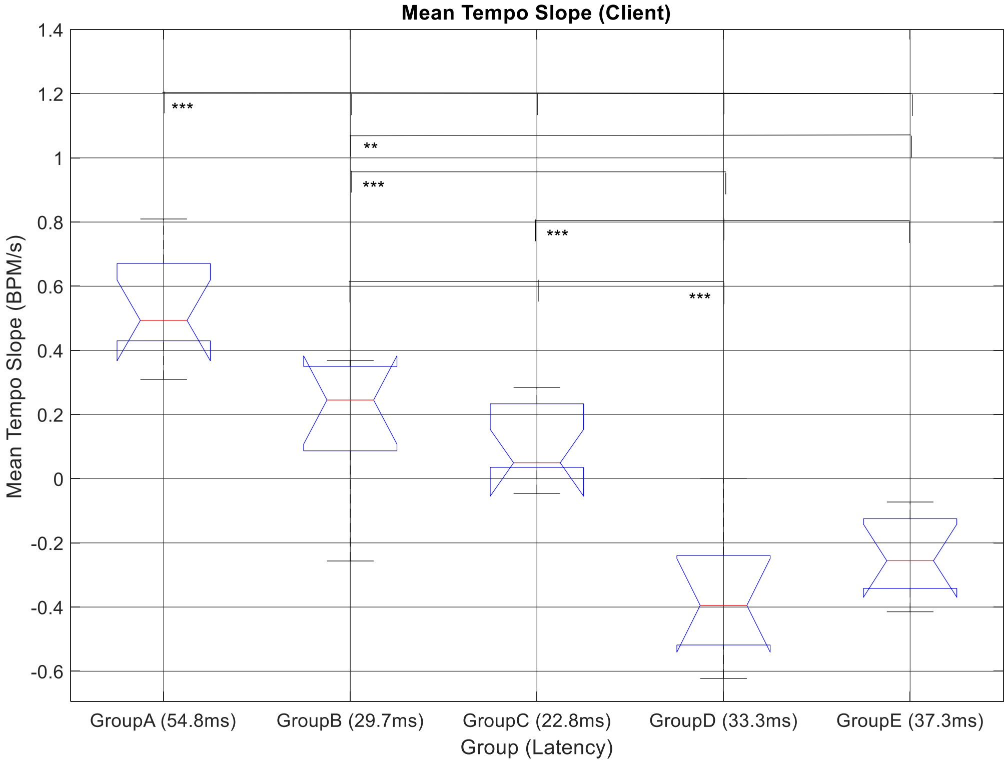


Figure 4.2.4.2 Notch plot of Mean Tempo Slope sorted by Group (latency) for Client Performers. ** and *** indicate significance evaluated at p = 0.01 and 0.001 respectively.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.346[a]	.119	.099	.24591	1.225

a. Predictors: (Constant), Latency

b. Dependent Variable: MeanTempoSlopeServer

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.353	1	.353	5.833	.020[b]
	Residual	2.600	43	.060		
	Total	2.953	44			

a. Dependent Variable: MeanTempoSlopeServer

b. Predictors: (Constant), Latency

Table 4.2.4.7 Model Summary (top left),

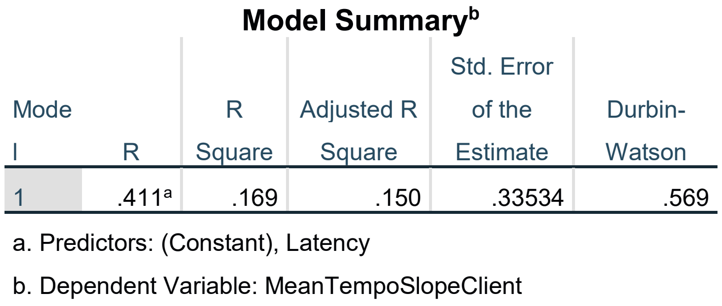
Table 4.2.4.8 ANOVA (top right) and

Table 4.2.4. 9 Coefficients (below) for Linear Regression of Server Mean Tempo Slope over varying Latency.

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.112	.127		.885	.381	-.144	.368					
	Latency	-.008	.003	-.346	-2.415	.020	-.015	-.001	-.346	-.346	-.346	1.000	1.000

a. Dependent Variable: MeanTempoSlopeServer

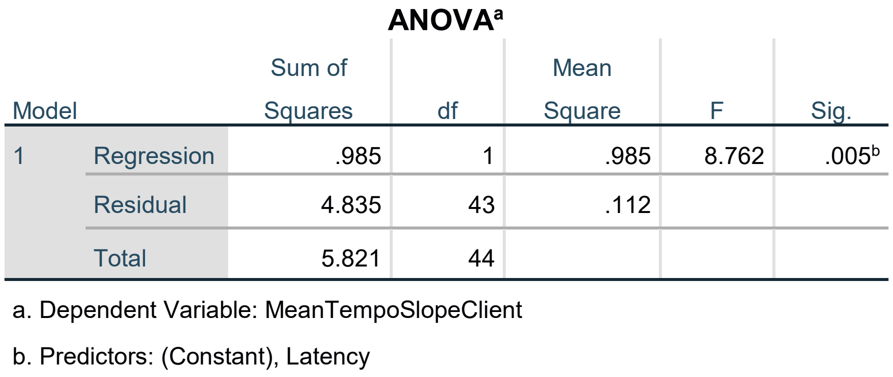


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.411[a]	.169	.150	.33534	.569

a. Predictors: (Constant), Latency

b. Dependent Variable: MeanTempoSlopeClient



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.985	1	.985	8.762	.005[b]
	Residual	4.835	43	.112		
	Total	5.821	44			

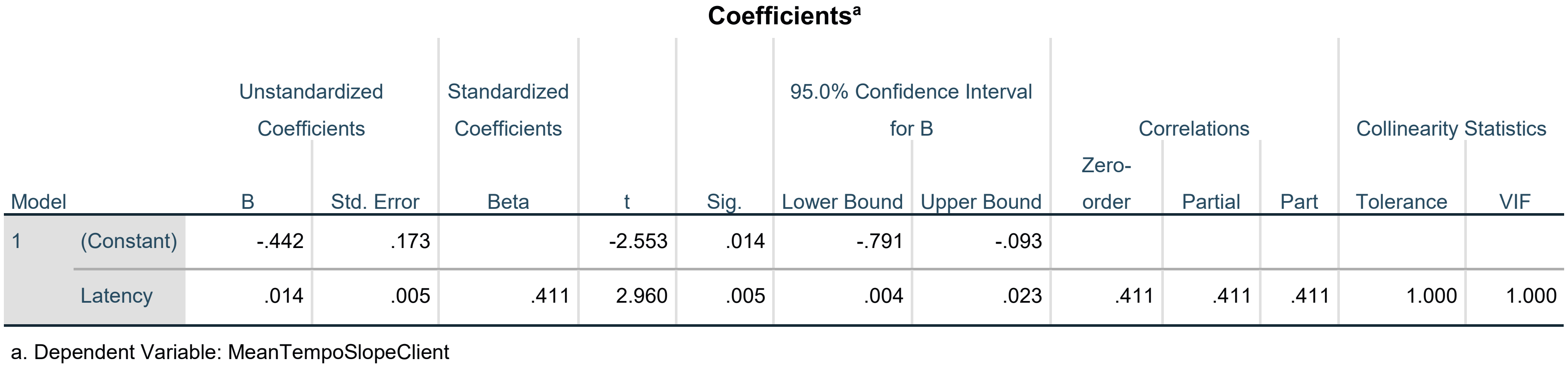
a. Dependent Variable: MeanTempoSlopeClient

b. Predictors: (Constant), Latency

Table 4.2.4.10 Model Summary (top left),

Table 4.2.4.11 ANOVA (top right) and

Table 4.2.4.12 Coefficients (below) for Linear Regression of Client Mean Tempo Slope over varying Latency.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-.442	.173		-2.553	.014	-.791	-.093					
	Latency	.014	.005	.411	2.960	.005	.004	.023	.411	.411	.411	1.000	1.000

a. Dependent Variable: MeanTempoSlopeClient

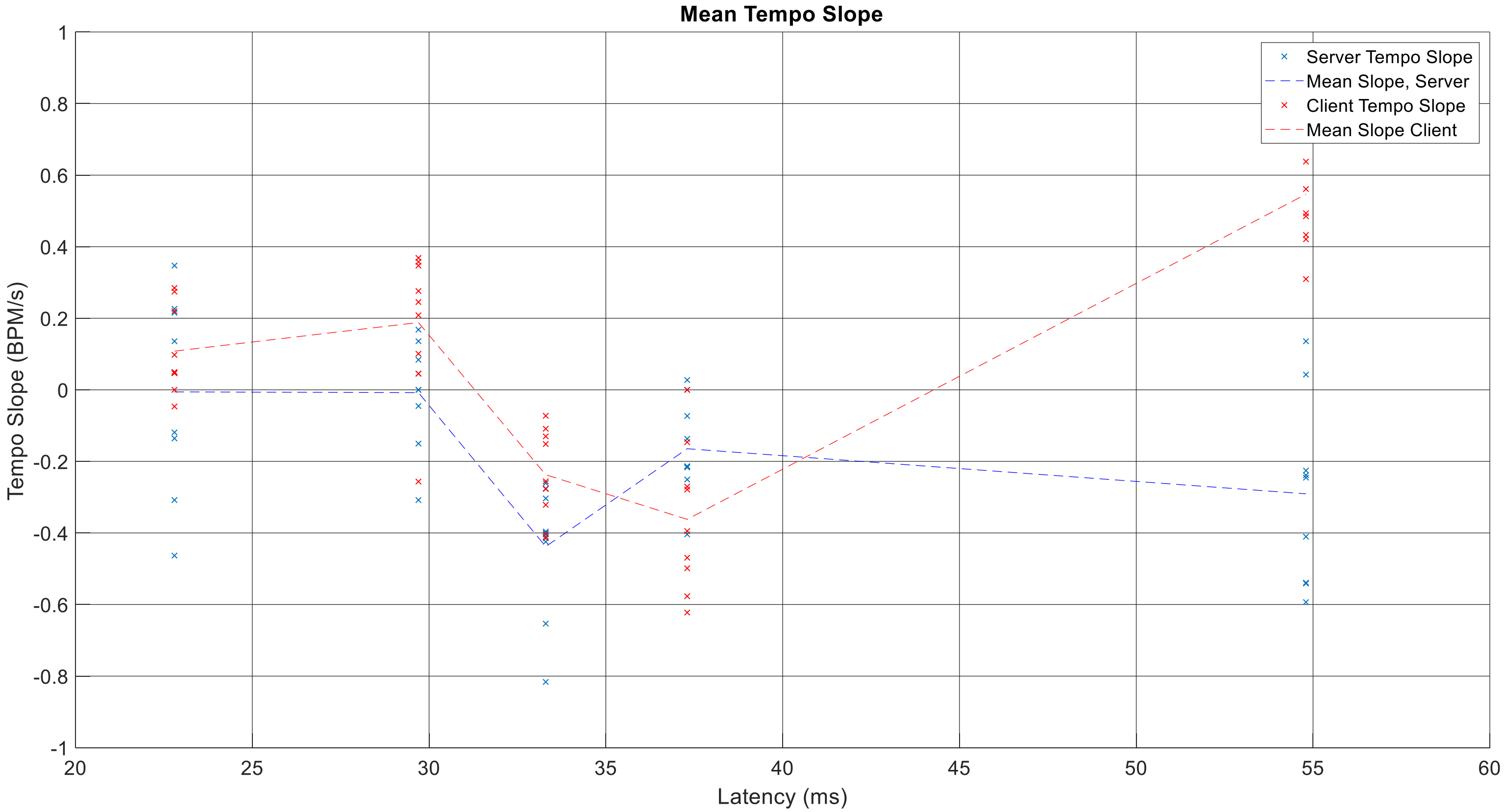


Figure 4.2.4.3 Scatter plot and means for Mean Tempo Ratio, with Client and Server data treated separately.

4.2.5 Asynchrony

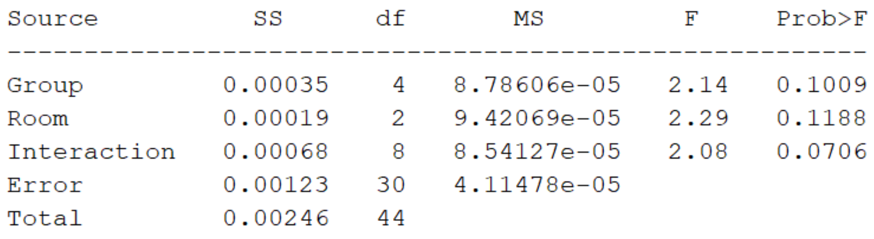
Measurement of Asynchrony (*Table 4.2.5.1*) describes disagreement in onset across performance. A value characterising natural interactivity was defined as 40-50ms [25].

In certain instances data deviates from normality or is close to statistically significant deviation from normal distribution. Notably Group C (22.8ms) Room C and Group E (33.3ms) Room B represent measurements which are not normally distributed.

Group	**Room**	**N**	**Mean**	**Median**	**Std. Deviation**	**Max**	**Min**
A	**A****	*3*	*0.033*	*0.033*	*0.004*	*0.038*	*0.029*
	B**	*3*	*0.027*	*0.029*	*0.004*	*0.030*	*0.023*
	C**	*3*	*0.040*	*0.037*	*0.012*	*0.053*	*0.030*
B	**A*****	*3*	*0.035*	*0.035*	*0.001*	*0.036*	*0.034*
	B***	*3*	*0.023*	*0.023*	*0.000*	*0.023*	*0.023*
	C**	*3*	*0.034*	*0.033*	*0.004*	*0.039*	*0.032*
C	**A****	*3*	*0.039*	*0.040*	*0.016*	*0.054*	*0.023*
	B**	*3*	*0.030*	*0.030*	*0.003*	*0.033*	*0.028*
	C	*3*	*0.026*	*0.025*	*0.002*	*0.029*	*0.025*
D	**A****	*3*	*0.041*	*0.038*	*0.006*	*0.047*	*0.037*
	B**	*3*	*0.038*	*0.039*	*0.003*	*0.039*	*0.035*
	C**	*3*	*0.037*	*0.041*	*0.008*	*0.042*	*0.027*
E	**A****	*3*	*0.030*	*0.030*	*0.007*	*0.036*	*0.023*
	B	*3*	*0.035*	*0.033*	*0.004*	*0.039*	*0.032*
	C***	*3*	*0.029*	*0.029*	*0.000*	*0.029*	*0.029*

Table 4.2.5.1 Asynchrony results, measured in milliseconds. ** and *** indicates standard error of 0.01 and 0.001 respectively.

Two-way ANOVA of Asynchrony sorted by Room and Group (latency) (*Table 4.2.5.2*) indicates no significant differences across Group (latency) or Room, however at p = 0.0706 (evaluated against p = 0.05) a possible interaction effect is just short of statistical significance. Notably all values conform to the 40-50ms upper limit associated with natural interactivity, such that this metric provides no indication of any level of interactivity variance with Latency.

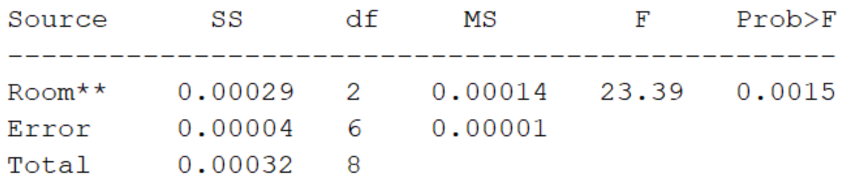


```
Source              SS       df        MS           F      Prob>F
------------------------------------------------------------------
Group            0.00035     4    8.78606e-05    2.14    0.1009
Room             0.00019     2    9.42069e-05    2.29    0.1188
Interaction      0.00068     8    8.54127e-05    2.08    0.0706
Error            0.00123    30    4.11478e-05
Total            0.00246    44
```

Table 4.2.5.2 Two-way ANOVA of Asynchrony sorted by Group (latency) and Room, showing no significance (evaluated at p = 0.05).

Though of no statistical significance in modelling data, interesting observations may be made from Asynchrony results *(Table 4.2.5.1)*. For Group B (29.7ms) and Group C (22.8ms) observation of means and medians suggests reverberant rooms B and C are more likely to provide lower values for Asynchrony (indicating more ‘in time’ performance) than for Room A. Observation of means and medians for Groups D (37.3ms) and Group E (33.3ms) show that this variation is no longer as pronounced. This may just be noise in data, or may suggest a close-to-significant interaction effect as suggested by two-way ANOVA *(Table 4.2.5.2)*.

Within Group (latency) measurements, ANOVA of data sorted by Room (*Table 4.2.5.3, Figure 4.2.5.1*) demonstrates statistically significant difference for Room B. Though Room B appears to yield significantly lower Asynchrony for this group, this may be due to the narrowness of measurement.



```
Source        SS        df      MS         F      Prob>F
-------------------------------------------------------
Room**      0.00029     2     0.00014    23.39    0.0015
Error       0.00004     6     0.00001
Total       0.00032     8
```

Table 4.2.5.3 ANOVA of Asynchrony by Room for Group B (29.7ms). ** indicates significance evaluated at p = 0.01.

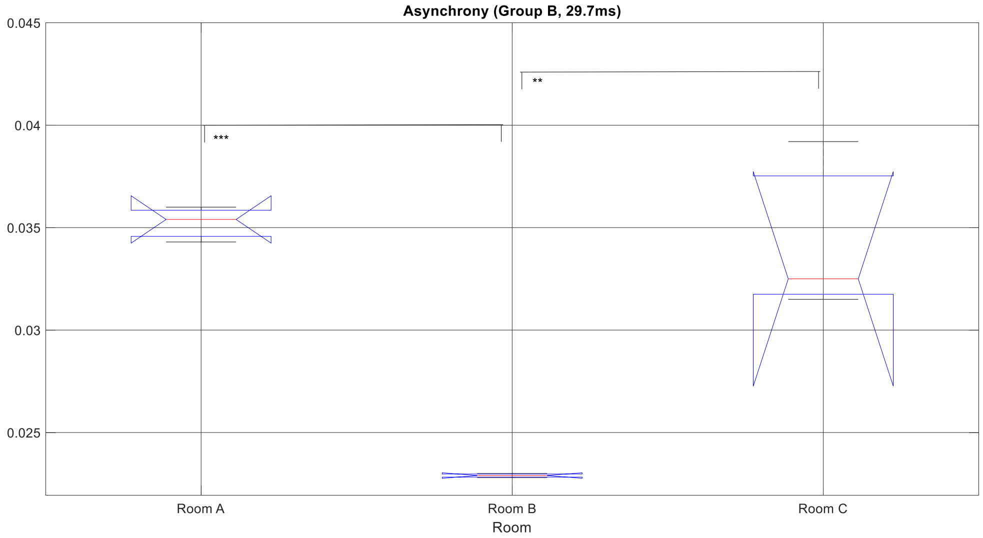


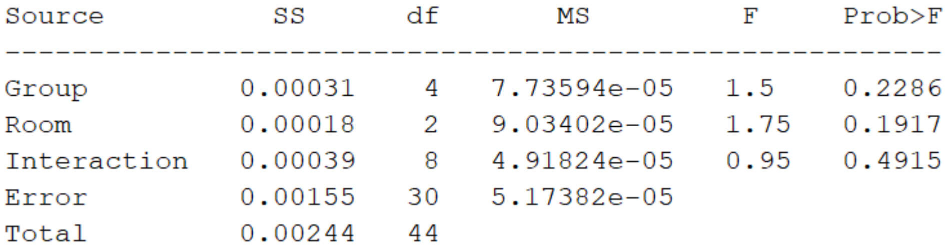
Figure 4.2.5.1 Notch plot of Asynchrony by Room for Group B (29.7ms). ** and *** indicate significance evaluated at p = 0.01 and 0.001 respectively.

4.2.6 Precision

Precision measurement (Table 4.2.6.1) yields little informative analysis, and represents no significant modelling of effects. Two-way ANOVA of Precision sorted by both Group (latency) and Room (*Table 4.2.6.2)* demonstrates no significant differences between populations and no potential interaction effect. All that can be stated from data collected measuring Precision is that means for all groups are well within the 50-70ms [59] upper limit presented by measurement of live musical interaction, and no indication of non-natural interactivity is presented by this metric.

Group	**Room**	**N**	**Mean**	**Median**	**Std. Deviation**	**Max**	**Min**
A	**A****	*3*	*0.043*	*0.045*	*0.008*	*0.050*	*0.034*
	B**	*3*	*0.040*	*0.042*	*0.008*	*0.046*	*0.031*
	C**	*3*	*0.046*	*0.046*	*0.009*	*0.055*	*0.037*
B	**A****	*3*	*0.048*	*0.047*	*0.005*	*0.053*	*0.044*
	B**	*3*	*0.036*	*0.036*	*0.004*	*0.039*	*0.031*
	C**	*3*	*0.041*	*0.041*	*0.005*	*0.046*	*0.036*
C	**A****	*3*	*0.044*	*0.045*	*0.008*	*0.051*	*0.035*
	B**	*3*	*0.042*	*0.039*	*0.006*	*0.049*	*0.038*
	C**	*3*	*0.037*	*0.033*	*0.010*	*0.048*	*0.029*
D	**A****	*3*	*0.052*	*0.050*	*0.007*	*0.061*	*0.046*
	B**	*3*	*0.050*	*0.046*	*0.010*	*0.060*	*0.042*
	C**	*3*	*0.042*	*0.047*	*0.010*	*0.049*	*0.032*
E	**A****	*3*	*0.043*	*0.044*	*0.005*	*0.047*	*0.037*
	B**	*3*	*0.046*	*0.046*	*0.003*	*0.048*	*0.043*
	C**	*3*	*0.040*	*0.040*	*0.007*	*0.047*	*0.033*

Table 4.2.6.1 Precision results (measured in milliseconds). ** indicates standard error of 0.01 or less.



```
Source            SS       df       MS          F       Prob>F
---------------------------------------------------------------
Group          0.00031     4    7.73594e-05    1.5     0.2286
Room           0.00018     2    9.03402e-05    1.75    0.1917
Interaction    0.00039     8    4.91824e-05    0.95    0.4915
Error          0.00155    30    5.17382e-05
Total          0.00244    44
```

Table 4.2.6.2 Two-way ANOVA of Precision by Group (latency) and Room. No significance is indicated (evaluated at p = 0.05).

4.2.7 Discussion: Synchrony Results

Analysis of Mean Tempo Slope Difference identified Group A as significantly different from other Groups. Casual observation of tempo over time for these performances illustrates that the large Mean Tempo Slope Difference in this group is likely the result of the application of latency-coping strategies. Namely the Client exhibits a tendency throughout performance by this group to significantly speed up performance to compensate for tempo deceleration in Server performance. This fundamentally amounts to quite an extreme Leader-Follower interaction type, which can be identified as prevalent in this performance group. Poor adoption of Leader-Follower strategy may be a result of the high latency [31] for this group (54.8ms OWT) and partial collapse of synchrony, but could also be influenced by performer proficiency. In this case, the naturalness of the musical interactivity of this performance pair has been impaired.

Tendency to Lead results illustrated that both Group A (54.8ms) and Group D (37.3ms) are significantly different from other groups. It is concluded that for Group D (37.3ms) a Leader-Follower technique is being applied across performances. A variant Leader-Follower technique has already been identified for Group A (54.8ms) by analysis of Mean Tempo Slope Difference.

For Group A (54.8ms) the Server tends to lead performance, conforming to measurement of Client tempo acceleration as a response to the Server tempo slope. For Group D (37.3ms) the Client adopts the Leader designation. This could be due to the 'professional' proficiency level of the Server performer of Group D (37.3ms). This may also explain the difference in relationship between Server and Client Performer Mean Tempo Slope for this group (*Table 4.2.4.3)*. As in the performance task the Server Performer starts first, it is expected that the Server Performer will lead performance (to an extent which is native to live performance) and thus be the more stable performer. As the Client performer adopts the Leader designation in Group D (37.3ms) this may explain the fact that for this group the Client has appeared the more stable performer.

The latency of Group D (37.3ms) conforms to latencies at which Leader-Follower latency-coping strategies are identified in empirical EPT research [28], [31].

Tendency to Lead in Group D falls outside the range of 'natural' musical interactivity (-20-40ms) [59], thus performance interactivity and naturalness can be identified as impaired. As the magnitude of the Tendency to Lead is within the limits of measures characterising natural interactivity (less than 40ms) it can be considered an effective Leader-Follower relationship has been adopted by this group. Other groups all show no indication of leader-follower strategy across performance, and no indication of impaired performance interactivity is presented for these groups

Linear plot of means for Absolute Tendency to Lead (*Figure 4.2.2.1*) demonstrates that in our data Groups operating at latencies above a 35ms threshold are using some form of latency coping technique, conforming with empirical study of NMP where modelling indicates leader-follower strategies being adopted at and above this latency [31], [28].

Mean Tempi Ratio provides a measure of the ability of performers to 'move together' with respect to tempo (though does not describe acceleration or deceleration other than to identify which performer has faster tempo). Here a significant combined effect of Group (latency), and also Room, impacts this aspect of performance in this study. This cannot be confidently considered representative of the interaction between latency and room, as each discrete latency also represents a discrete performer pair. For this reason it is considered that this interaction effect may alternatively reflect preferences within groups.

For Group B (29.7ms) it appears significant that the Client Performer tended to deliver at faster tempo than the Server performer when operating in the dry studio booth simulation, Room A, than when performing in reverberant environments Rooms B and C. This will be discussed further after presentation of participant questionnaire results.

Latency is demonstrated as having a significant effect on Mean Tempo Slope in performance, with a strong dependency between performers within Groups, as would be reasonably expected. Modelling of Mean Tempo Slope demonstrates that performance is possible with even tempo below 30ms latency. Above this point Tempo deceleration is expected, though this may be stabilised through the application of appropriate Leader-Follower latency coping strategies.

Group E (33.3ms), identified as experiencing slight tempo deceleration as is expected above the 30ms EPT [29], [28], [31], can be identified as having impaired naturalness of musical interactivity between performers. Mean Tempo Slope for Group B (29.7ms) and Group C (22.8ms) is appropriately even, and allows characterisation of natural musical interactivity between performers for these groups.

A potential combined effect of Group (latency) and Room is possible though short of significant for Asynchrony measurement. Within Group B (29.7ms) Room B, the medium hall, yields measurements of smaller Asynchrony than in Room A (studio) or Room C (cathedral) (*Figure 4.2.5.1*). This may be simply due to narrow measurement sets (N=3), or alternatively may be suggestion of an effect of room on performance synchrony for this group. This is best discussed after presentation of Performance Experience Questionnaire results.

Asynchrony measurements for all groups (*Table 4.2.5.1*) conformed to the 30-50ms upper limit identified as characterising natural interactivity [25], [59]. In this regard Asynchrony provides no indication of impairment in musical interactivity between performers.

Precision measures (*Table 4.2.5.2*) are largely uninformative other than to state that means for all groups again conform to values below the 50-70ms upper threshold identified as characterising natural musical interactivity [59]. As such this synchrony metric provides no indication of impairment in musical interactivity between performers.

The total of the synchrony measurement data allows the categorisation of 4 discrete categories of interactivity type arranged by latency which are adapted from previous literature [28], [31]:.

1. Below 30ms 'natural' interactivity can be achieved, where synchrony metrics conform to characteristic values. In this case the VIIVA-NMP system can be considered as providing optimised Naturalness, Coherence and Plausibility in musical interactions between performers, and Interface Awareness due to latency is minimised.
2. Above 30ms two categories are present:
   a. Tempo deceleration, where no latency adjustment is made. In this case a degree of Interface Awareness is present, and quality of Naturalness, Coherence and Plausibility is impaired. Though decelerating the performers may still move 'together' and achieve synchrony measurement which identifies no deviation from natural interactivity. This will be detectable in Mean Tempo Slope. This conforms to expectations from previous NMP research detailing the Ensemble Performance Threshold [31].
   b. Leader-Follower relationships may be adopted [28], [31] to stabilise tempo slope with varying success. In this case similar slight impairment of immersive qualities will be apparent. This will be apparent in Tendency to Lead.
3. Above 50ms performance synchrony begins to fall apart, and serious impairment of immersive qualities may be experienced. This is apparent throughout tempo measurements between parts.

A finite geographic range as well as associated technical recommendations can be associated with each category (*Table 4.2.7.1*). It is worth noting that network latency and jitter is largely dependent on network infrastructure rather than geographic distance, and operation from large technical hubs will provide superior performance.

Crucially, synchrony analysis does indicate that the VIIVA-NMP audio system can indeed facilitates natural musical interactivity and function from typical performer homes. It also appears that with the addition of acoustic simulation and spatial audio data still conforms to expectations based upon empirical modelling of the effect of latency on dry audio in EPT research.

.

Category	1	2a	2b	3
Interaction Type	Natural	Tempo Deceleration	Leader-Follower	Synchrony Collapse
Latency	<30ms	>30ms	>30ms	>50ms
Estimated Range	300km (min) – 500km + (recommended)	-	-	-
Sample Rate	96kHz	-	-	-
Buffer Size (minimum)	128	-	-	-
Buffer Size (recommended)	64	-	-	-
Jitter Buffer Size	~10.7ms	-	-	-
FEC Redundancy	1-2	-	-	-
Bandwidth per Channel	4.7 – 7.0 Mb/s	-	-	-

Table 4.2.7.1 Categories of Interaction Type with associated latency limits. Associated technical specification is provided for ‘natural’ musical interactions category.

4.3 Performance Experience Questionnaire

Participants provide one questionnaire response per Room. As such this provides one response per combination of nominal categories, and any sort of statistical analysis is impossible. Instead Performance Experience Questionnaire response provides indicatory information regarding Perceived Immersion [38] and invites comment from participants.

Data is sorted by Group (latency), Room and Performer (identifiable by designation server/client)

Response data has been coded into numerical representation.

The numerical code used for items 1-6 is:

1. Strongly Disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly Agree

For items 7 and 8 the numerical code used is:

1. Completely Unacceptable (completely unlike live conditions)
2. Unacceptable (not like live conditions)
3. Acceptable (like live conditions)
4. Absolutely Acceptable (equivalent to live conditions)

4.3.1 Item 1

Item 1 addresses the statement:

My performance experiences in the Virtual Acoustic Performance Space were consistent with my performance experiences in Real Performance Spaces.

Notably in response to this item (*Figure 4.3.1.1*) the higher latency groups (namely Groups A and D) tend to rate this item poorly in comparison to lower latency groups. This is likely due to Interface Awareness factors, namely noticeable latency. It can also be observed (for example in Group A server measurements) that a performer who scores one room higher or lower than the mean is also likely to do the same for other rooms respectively. This indicates an individual preferential factor described by Immersive Tendency [38].

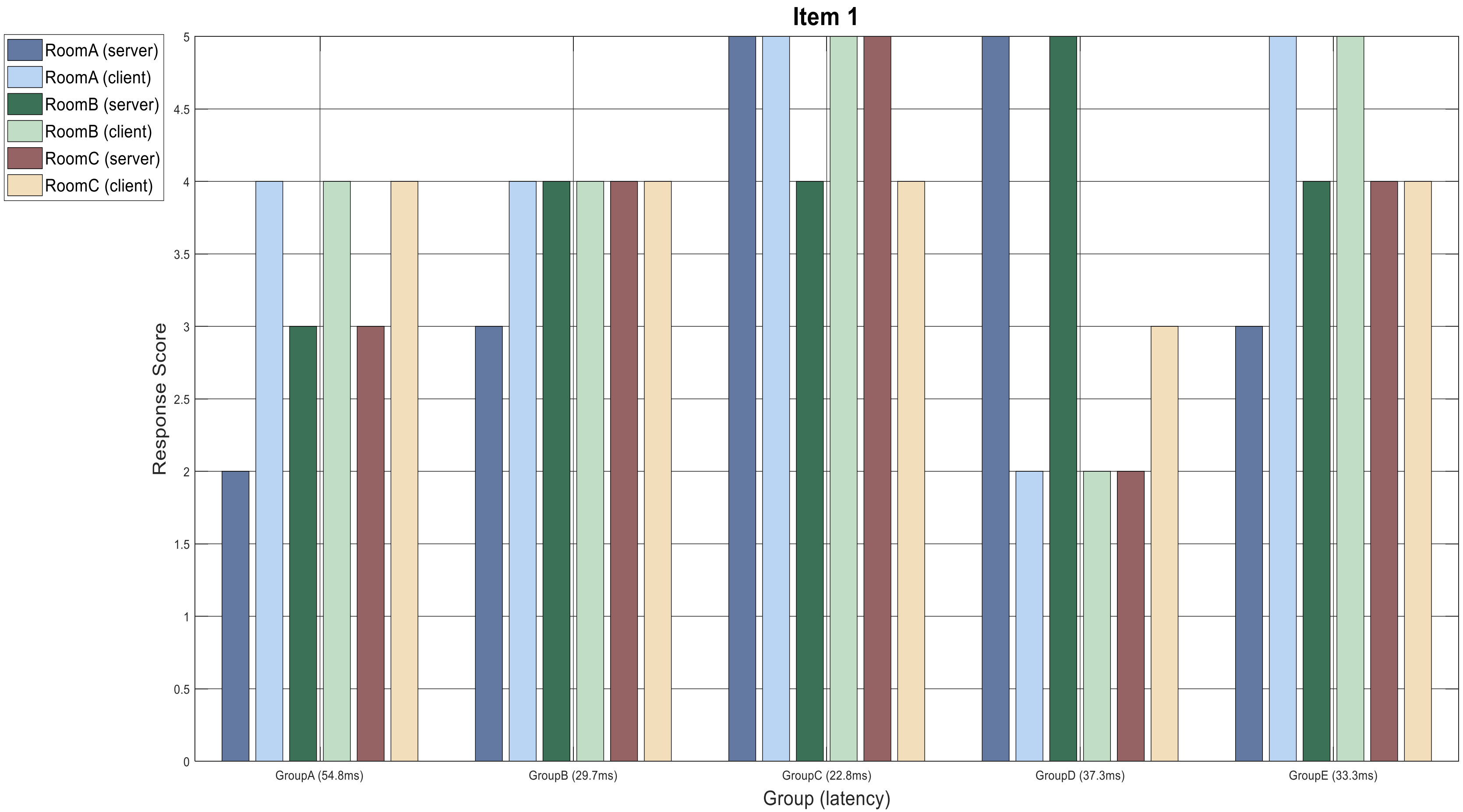


Figure 4.3.1.1 Response to questionnaire item 1.

Groups B, C and E, all with latency below 35ms seem to indicate Agreement or Strong Agreement with this question item, suggesting that there is an effective perceptual Coherence between the VIIVA-NMP audio system performance experience and live interactions.

Notable free comments on this question item indicated that in one case ‘reverberation did not seem consistent with the closeness of the voice’. This case is identified for Room B, notably from a participant using closed-back headphones in Group A.

Other comments indicate that Room B was ‘comparable’ with real-life experiences in similar rooms, and identified Room A as ‘similar to practice room experience’, as detailed in Group B response.

A participant from group C commented that the dryness of Room A was somewhat uncomfortable for singing and that more reverb was required, while commenting that Room B provided a ‘good balance of atmosphere and separation’.

Group D free comment indicated a noticeable delay in Room A response (potentially intended to identify this throughout performance). This perceived delay is to be expected as the OWT latency for this group is in excess of the EPT at 37.3ms. The professional level singer also indicated that better performance could be achieved if accompanied by another professional singer. In response to Room B a participant noted that the reverb noticeably was a digital effect when compared to experience in live acoustics.

4.3.2 Item 2

The second questionnaire item addresses the statement:

The Virtual Acoustic Performance Space was acoustically responsive to sounds I initiated and/or performed.

Again rating across the board typically scores this item highly (*Figure 4.3.2.1*), indicating good Coherence and Realism. The notable exception is some low rating of Room A in Groups B and E. These latencies are well within the limits where Synchrony metrics would indicate good Coherence and Realism so this is unlikely to be an Interface Awareness issue. It is considered that in some ways this question may be inherently loaded against Room A, in that the questionnaire item addresses ‘responsiveness’ and Room A, a dry studio booth, is inherently damped and thereby naturally unresponsive. This could present as a skew in results.

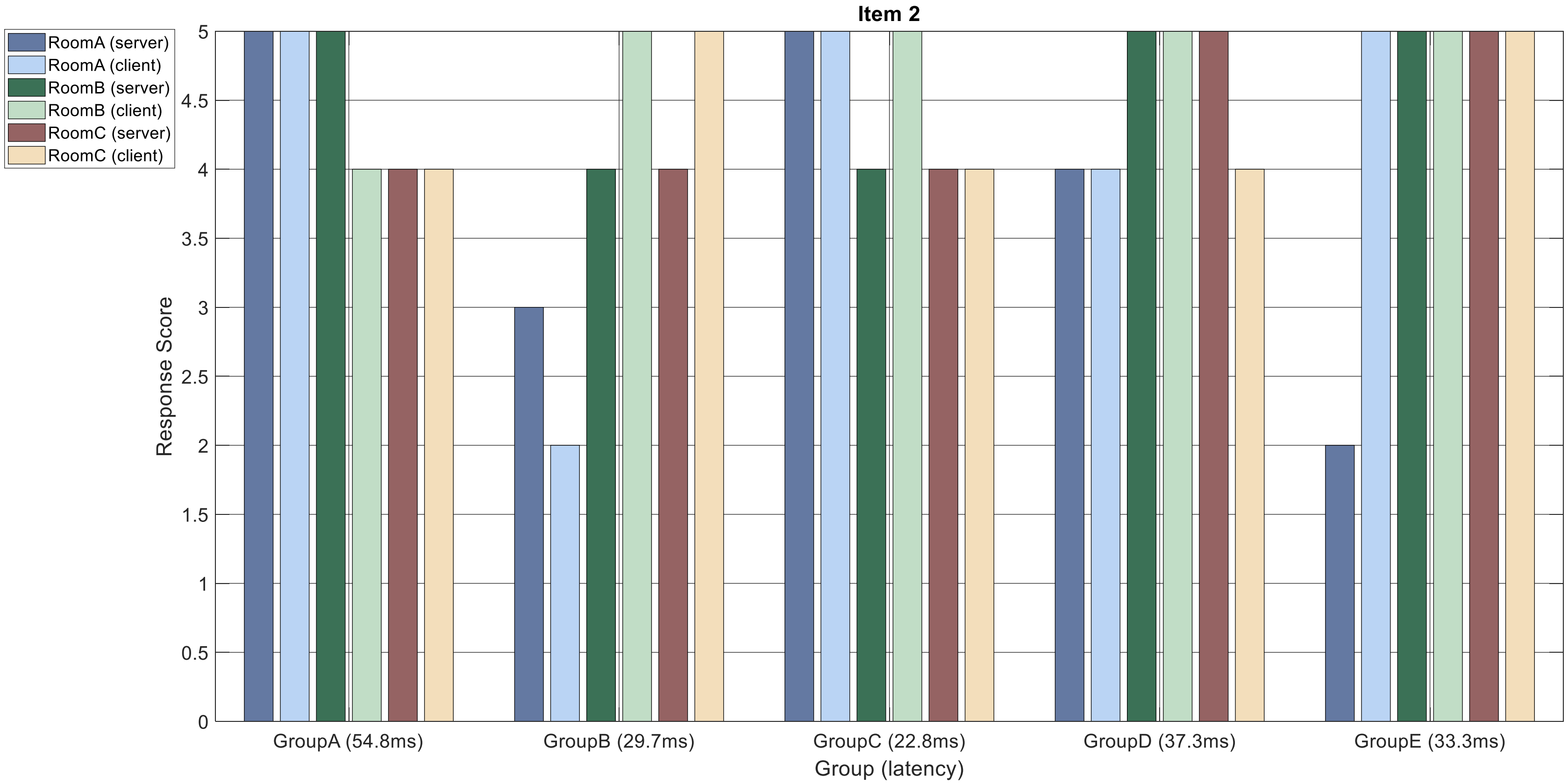


Figure 4.3.2.1 Response to questionnaire Item 2.

Free comment from Group B on item 2 of the questionnaire identified that Room A was 'as responsive as small bathroom acoustics could be' indicating that this item may be rated lower simply due to the wording of the questionnaire item, which specifies 'responsiveness'.

Response from Group C indicated that the responsiveness of Room C 'felt really good' in one instance and that in one instance Room A was noted as feeling like 'natural voices in a room'. This same participant noted Room B acoustic responsiveness 'felt clear, tight, and with natural room atmosphere'.

A Group D participant made another statement regarding Room A, stating that it felt 'dead', again indicating that response for this room may be loaded for this questionnaire item.

4.3.3 Item 3

Item 3 of the questionnaire gathers response on the statement:

My Musical interactions with the acoustic environment in the Virtual Acoustic Performance Space seemed natural.

General high scoring indicates that the aural experience provided to performers here is perceived as natural, however in several cases performers have identified either neutrality or disagreement with this statement (*Figure 4.3.3.1*). It can be observed that there is a slight tendency for the higher latency groups to rate this item lower. Neutral and Disagree responses, however, are isolated to rating of discrete rooms by performers within groups. Though no clear pattern is apparent it is hypothesised that this may reflect personal preferences regarding Room.

Free comment from a performer in Group A indicated that Room B was 'more natural' than Room C, providing a clear example of personal preferences between different virtual acoustic environments with respect to Naturalness. Another preferential example is apparent in free comment response from Group B, where a performer indicates that Room B was the 'more natural' room in comparison to Room C, while noting that the damped studio booth, Room A, was 'a bit muffled'.

A performer from Group C indicated that Rooms A and C 'felt very natural' and Room B 'felt very responsive and natural', once again illustrating a unique preferential ordering of response to different rooms within a group.

A group D participant noted that Room A 'felt natural but also a little odd as the room sound had changed'. As Room A was the first room used by this performer this presumably refers to a difference between the virtual acoustic space and the real space the performer is located in. In this case this raises an interface awareness issue that is to be expected in the audio-only case presented in this study. The solution to this issue is incorporation of the proposed use of VR avatar technology and virtual rooms to address this non-coherence between what is seen and what is heard.

The same performer from Group D notes that Room B felt like a 'more processed sound' than Room A. This may merely be the identification of a digital reverb being used, though may also suggest that the audio quality of the acoustic simulation and spatial audio rendering can be improved upon.

The non-researcher participant from Group E commented regarding Room C 'that much reverb felt unnatural compared to my surroundings, but musically it felt natural'. This once again directly highlights the interface awareness issue associated with the audio-only nature of testing.

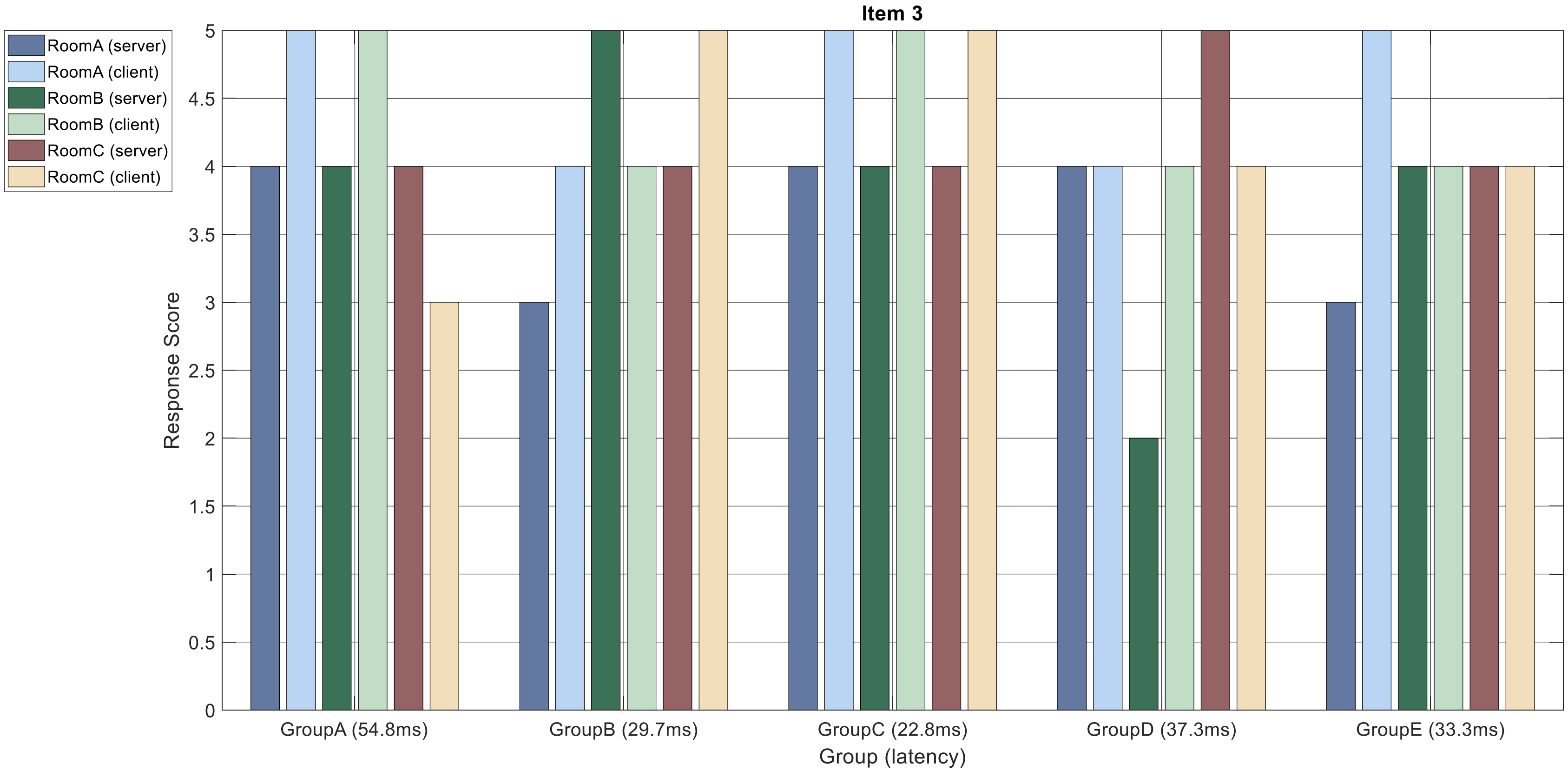


Figure 4.3.3.1 Response to questionnaire item 3

4.3.4 Item 4

Item 4 of the questionnaire addresses the statement:

My musical interactions with the other musicians in the Virtual Acoustic Performance Space seemed natural.

A slight tendency can be observed (*Figure 4.3.4.1*) for worse scoring of this item by the higher latency groups. Below 35ms response is largely positive, whereas high latency groups border on tendency to disagree. This would indicate that there may be a tendency for latency to impair Naturalness with respect to Social Presence and Communication.

Varied response can again be observed within discrete performers as Room varies, though no clear pattern is apparent here. This may once again be representative of the effects of personal preferences from individual performers.

In free comment response a participant from Group A notes 'I felt no real problems but real time visual contact could have helped' indicating that a lack of naturalness in interactions between performers is present, and specifically this is an Interface Awareness issue similar to the one suggested in response to questionnaire item 3.

For Room C the same participant states that the reverb was too much and made performance 'a little difficult to follow'. This does not exclusively raise a Naturalness issue, as spaces with long reverbs certainly exist. Instead this comment may pose a limit on sensible acoustic environment choice for Immersive NMP, as it is indicated that 'cathedral' type reverb times make synchronous performance more difficult to a degree where performers perceive this issue.

Other response from this participant for Room B stated the performer 'found it difficult to sync up at times'. Most likely this comment refers to an impairment of Naturalness in Communication and Social Presence caused by the high latency for this group (54.8ms).

Response from Group B makes comment specific to Room A relative to Room B and Room C: 'I think the ensemble singing experience was more precise. Pitch and time misalignment was more clear'. This response raises an important issue: variation in virtual acoustic environment is affecting the participant's perception of the ensemble performance. It would logically follow that if a perceptual difference is present then the way in which a remote performer responds may also vary. This would conform with the potential variation with Room identified in synchrony measurement, where Room category appears to affect the ability of performer pairs to 'move together' in performance.

Group C responses provide indication of at least perceptual variation between rooms, stating the Room A was 'almost too clean', Room C 'felt natural', whereas when using Room B it was 'easy to feel the other singer, to keep in time, and tune in to each other'. This would indicate that the performer felt that the naturalness of their musical interactions were somewhat impaired in Room A, whereas the reverberant environments presented in Room B and C made performance interactions more natural. The naturalness of performance interactions may indeed be affected and causing a resultant effect on the physical ability of remote performers to ‘move together’ in performance.

A performer in Group D identifies that they ‘felt like there was some lag between us at points’ but were unsure whether or not this was ‘just due to timing’. This potentially identifies an impairment of Naturalness in Communication between performers as a result of the high latency for this group (37.3ms), suggesting that there is an Interface Awareness issue present here.

The non-researcher participant from Group E noted that ‘breathing together’ assisted the Naturalness of musical interactions between remote performers. This participant also noted that interactions ‘felt more natural’ with the ‘dry acoustics’ in Room A. This may speak towards a variation on performance synchrony through the perception of musical events as Room varies, or alternatively may simply identify a preference from the individual performer.

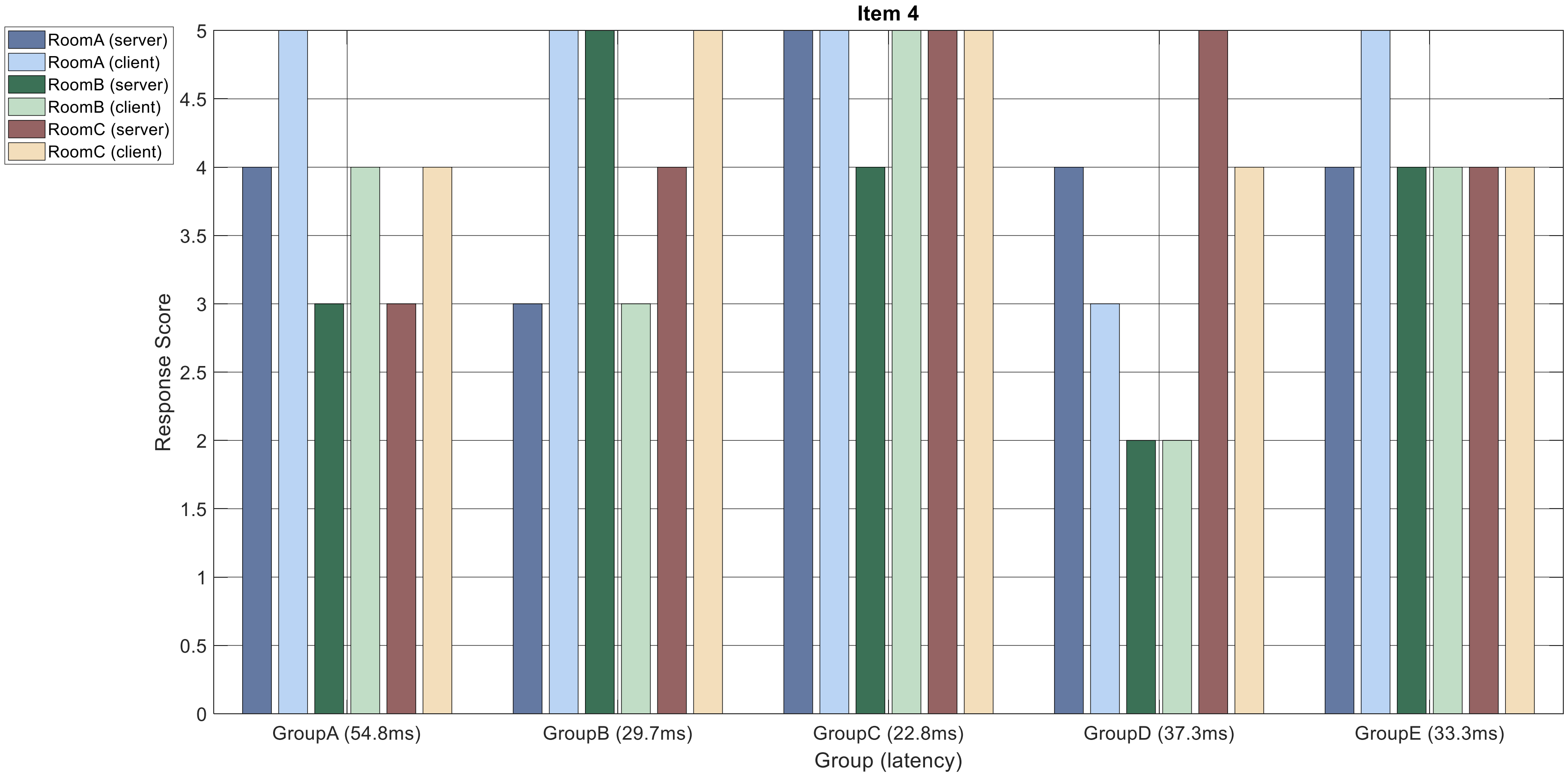


Figure 4.3.4.1 Response to questionnaire item 4.

4.3.5 Item 5

Item 5 investigates agreeability of the statement:

I experience delay between my actions and expected outcomes.

The score of this item aims to rate Interface Awareness, namely through latency. Converse to other question items where a high score denotes high Perceived Immersion qualities, in this case the inverse is true, as low Interface Awareness is desirable.

Specific to this question the latency aspect directly addresses local audio rendering, as the action (performing a musical event) and the expected outcome (hearing the musical event in the virtual acoustic space) only considers local throughput and not network transport.

Response to this item ranges from neutrality to strong disagreement, indicating that with respect to latency there is little Interface Awareness of local audio rendering processes.

Free comment on this question item collected similar statement from across the groups that performance 'felt tight', there was 'no noticeable' or 'not clearly noticeable' delay, and that the experience was 'just the reverb, no unnatural delay'.

The notable exception is the Client performer from Group D, who expresses neutrality towards the statement in all cases. As the local audio rendering latency for this group is 3.3ms it is likely that this performer has perhaps understood the statement as descriptive of the remote performer response as the 'expected response', and is referring to full system latency rather than local latency.

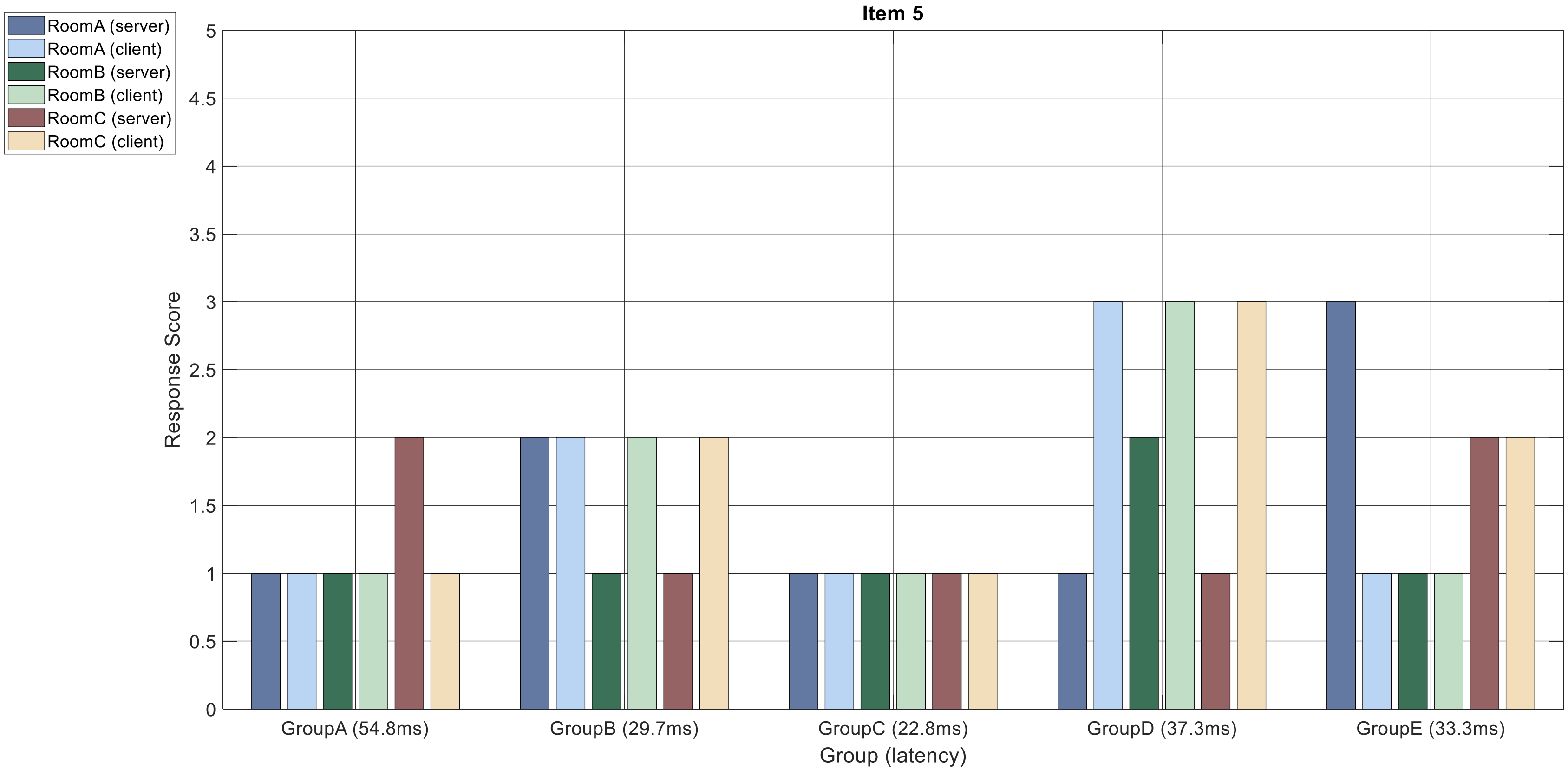


Figure 4.3.5.1 Response Score for Questionnaire Item 5

4.3.6 Item 6

Item 6 of the questionnaire tests agreement with the statement:

I feel included in the acoustical scene presented in the Virtual Acoustic Performance Space.

This item largely achieved positive rating with the exception of Group D where the Client performer provides neutral response for all rooms. Notably this participant belongs to a high latency group (37.3ms) and this rating may be the effect of latency as it is consistent across rooms. This may describe variation of Presence with latency, where high latency may affect worse Presence rating.

Some variation can again be observed between rooms within groups, potentially indicating that a preference is affecting scoring across varying Room. Room A appears to generally be more likely to be rated as less agreeable for this statement, potentially indicating that some level of reverberant environment is required to achieve superior Presence in the immersive performance experience.

A performer from Group A indicated that for Room A the acoustic environment ‘feels a little close for a performance space’. This demonstrates a Coherence and Plausibility issue affecting Presence, where the participant indicates that hosting a group singing in a damped booth would be unusual, and that it may be important to deliver simulation of typically encountered environments such as halls.

Other response in Group A indicated that ‘the experience with headphones does make it a little uncomfortable’. This raises an interface awareness issue with respect to required hardware. In the context presented the use of headphones is standard and largely unavoidable in a practical use case. One thing worth noting is that this participant used closed-back headphones, and that the effect of wearing headphones on Interface Awareness may vary between open-back and closed-back headphones.

A performer from Group B commented that the aural experience was ‘a wee bit faint’. This indicates that Presence is impaired by the level of audio delivered. As participants were directed to control listening level within safe limits it may simply be that in this case the co-performer was singing quietly.

Free comment from Group C stated that in Room C performance ‘felt like we were in the room together’ though the performer ‘did not feel much separation’. It is considered that this statement may speak to clarity variation between rooms, where clarity and source separation may provide a better scoring of Presence. Indeed for Room A the same performer notes they ‘felt a bit more separation but (the room) was very dry’. This would conform to previous statements regarding source clarity in the dry studio booth simulation where individual sources are indicated as being perceived with greater precision. This performer notes that Room B provided the ‘best separation’ and that the remote performer felt ‘apart but close’. This would imply there is a ‘goldilocks effect’ present for at least this performer, where rooms that are too ‘dry’ or ‘wet’ inhibit perception of source separation and presence. This would imply that affective design can be applied here if preferences are correctly modelled in order to design for optimisation of Presence.

The non-researcher participant from Group E commented ‘It was remarkable how familiar the musical setting felt, once I had got over the initial strangeness of the situation’ in response to performance in Room B. Firstly this clearly indicates that there is an Interface Awareness, Coherence and Naturalness issue in that the performance experience seemed peculiar. Presumably this speaks to NMP as a whole, in that the remote nature and hardware configuration present a substantially different context than is presented in live conditions. This comment would indicate that an adjustment phase is necessary for performers who are new to the context of NMP. Fortunately the comment also suggests that adjustment may be made quickly, and that after this performance can be considered ‘familiar’, indicating good Coherence, Naturalness and Plausibility.

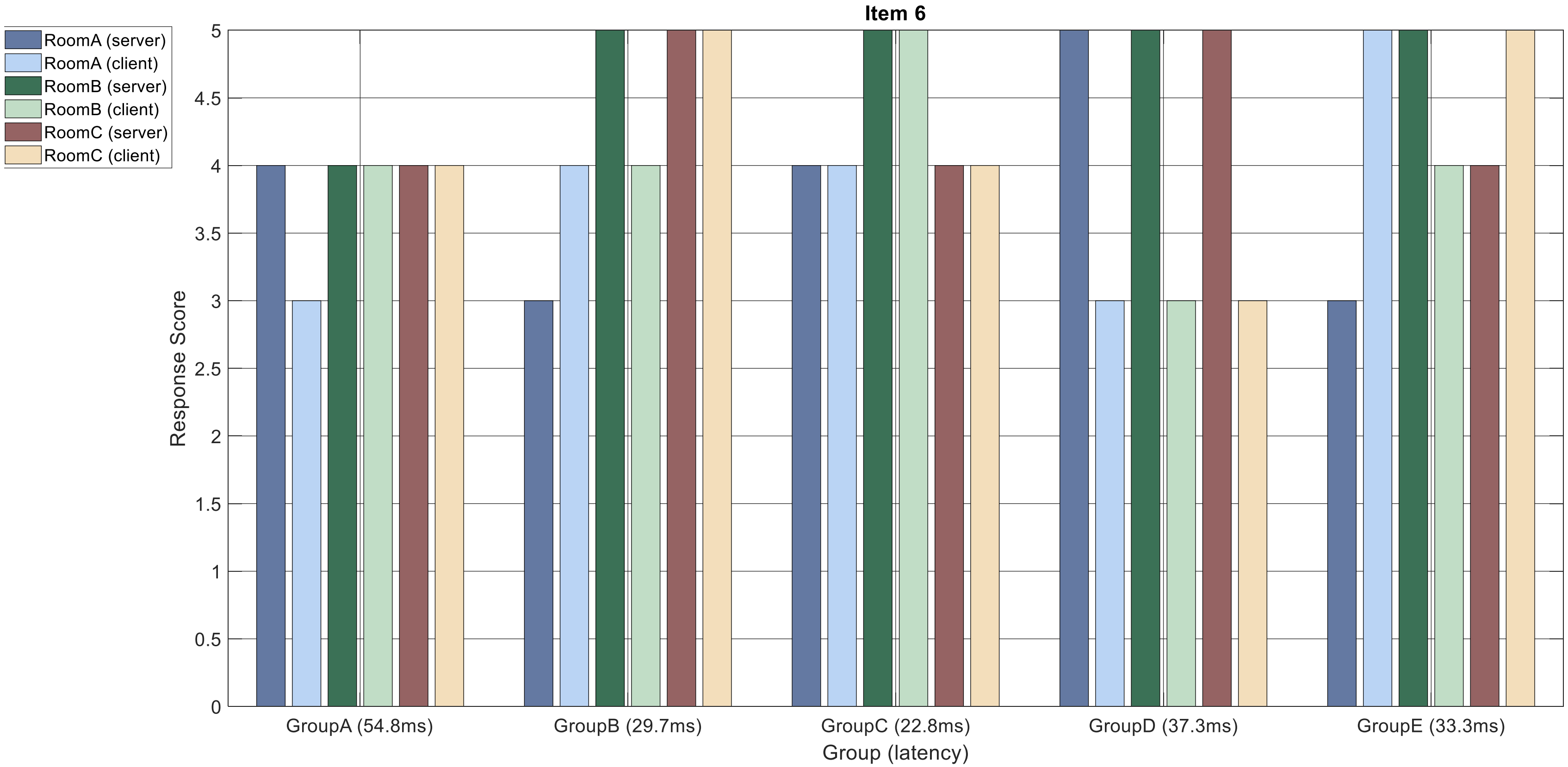


Figure 4.3.6.1 Response Score for questionnaire item 6

4.3.7 Item 7

The 7th questionnaire item asks for response to the statement:

How would you rate the performance conditions of your virtual acoustic performance experience?

Response demonstrates an overwhelming tendency to rate performance conditions 'acceptably close' or 'equivalent' to live performance conditions (*Figure 4.3.7.1*). Notably there is a slight tendency for rating as 'not acceptably close to live conditions' from Group D, potentially due to high latency (37.3ms). A performer from Group B and the researcher performer from Group E both rate exclusively Room A as 'not acceptably close' while rating other rooms 'equivalent' or 'acceptably close'. This would again indicate some form of preference with room choice may be present within groups.

Free comment from Group B indicated that rating here was impaired as audio delivered did not 'feel clear and amplified'. This is presumed further noting of an earlier identified level issue for this performer where loudness was considered too low.

Comment from Group C states that performance conditions were 'good enough for headphones'. Once again an inherent Interface Awareness, Coherence and Naturalness issue is raised in that the performance experience presented by any NMP system is different from live performance. Fundamentally what is described here is a hard limit in design presented by hardware.

Specifically with relevance to Room A this same performer states that performance conditions 'felt a little vibe-less' but was 'good for precision'. Regarding Room B the performer notes performance conditions 'felt great and sounded great'. These statements identify what may be a personal preference for reverberant environments over damped studio booths, and interestingly again raises the 'precision' issue identified in previous items. As this is exclusive to Room A it may well suggest that greater precision in perception and delivery can be achieved in some acoustic environments than others. Interestingly the performer also indicates that this preferential relationship may also affect Emotional aspects of Immersion.

Group D again provides comment on Interface Awareness issues, where 'on the head' listening via headphones presented a major difference that was not appropriately similar to live conditions.

The non-researcher performer from Group E indicates that the remote performer pair 'were as together as I would've expected a real performance to be'. This statement directly addresses perception of synchrony, however it contradicts synchrony measurement data which shows tempo deceleration in this group, and illustrates deviation from live conditions, where no tempo acceleration and deceleration should be expected.

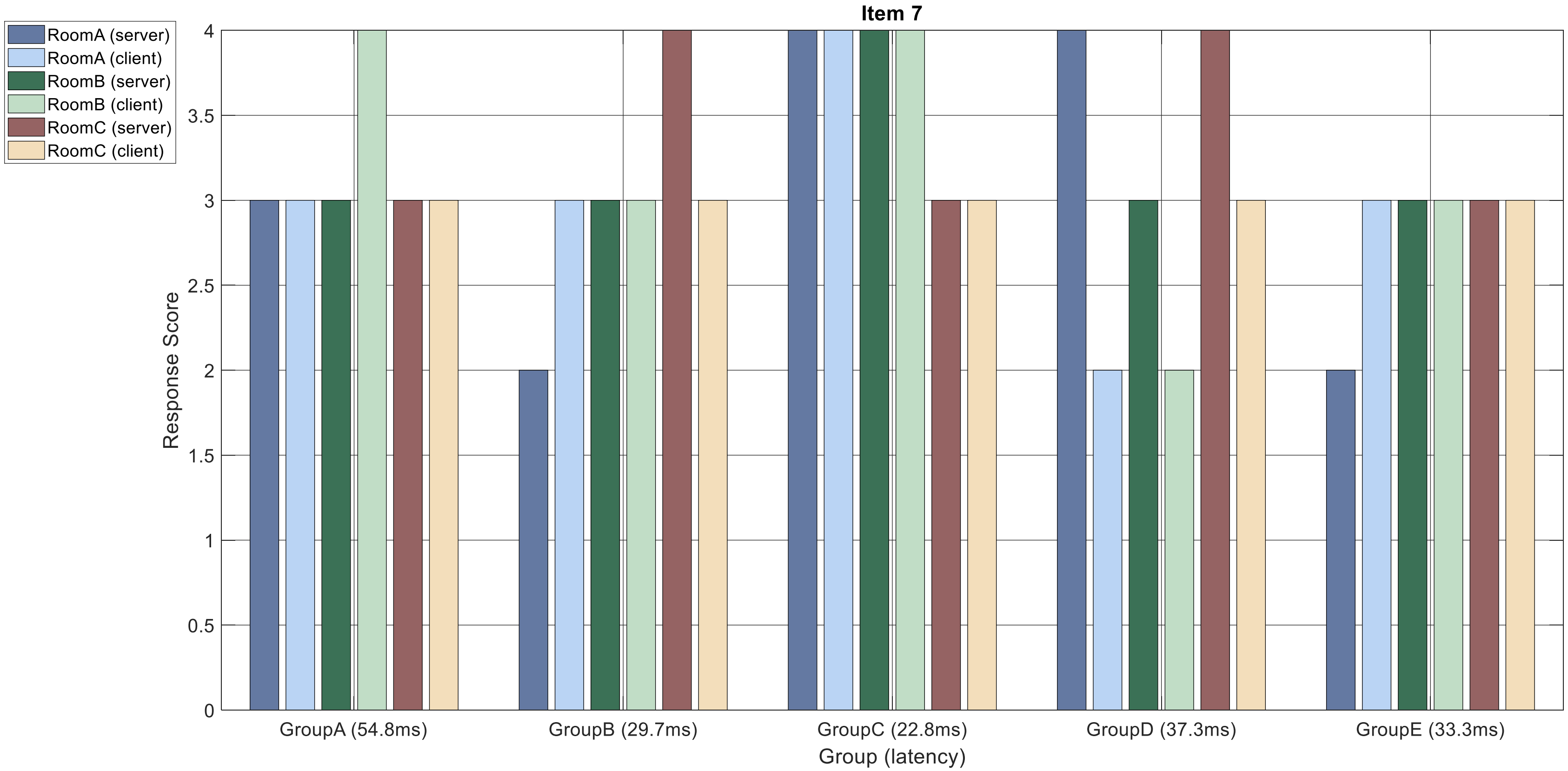


Figure 4.3.7.1 Questionnaire Item 7 response scores.

4.3.8 Item 8

Item 8 of the questionnaire poses the question:

How would you rate the latency experienced during your virtual acoustic performance experience?

Response to this questionnaire item (*Figure 4.3.8.1*) can again demonstrate substantial discrepancies between perceived response and task-performance measurement.

Group A rates latency as 'manageable' or 'no latency' where synchrony measurements can show clear deviation from real-world anchor values and can demonstrate with significance that high latency will a detrimental effect on this performance.

Notably Group D tends to rate latency unmanageable, which would imply that performers in this group are aware of synchrony and tempo issues caused by latency and illustrated in synchrony measurement.

It can be observed that Group C provides unanimous rating of 'no latency'. The latency of this group (22.8ms) may therefore indicate an approximate threshold below which latency is imperceptible in the context of NMP.

Group E sees some 'no latency' rating from the non-researcher performer in cases where synchrony and tempo analysis have demonstrated that performance is in fact impaired as a result of latency. This would indicate that there is a window in which latency does affect performance but this effect is not perceivable to the performer.

Only Group C left comment on this questionnaire item, expressing a positive response for all rooms and stating the latency conditions felt 'great' and 'tight'.

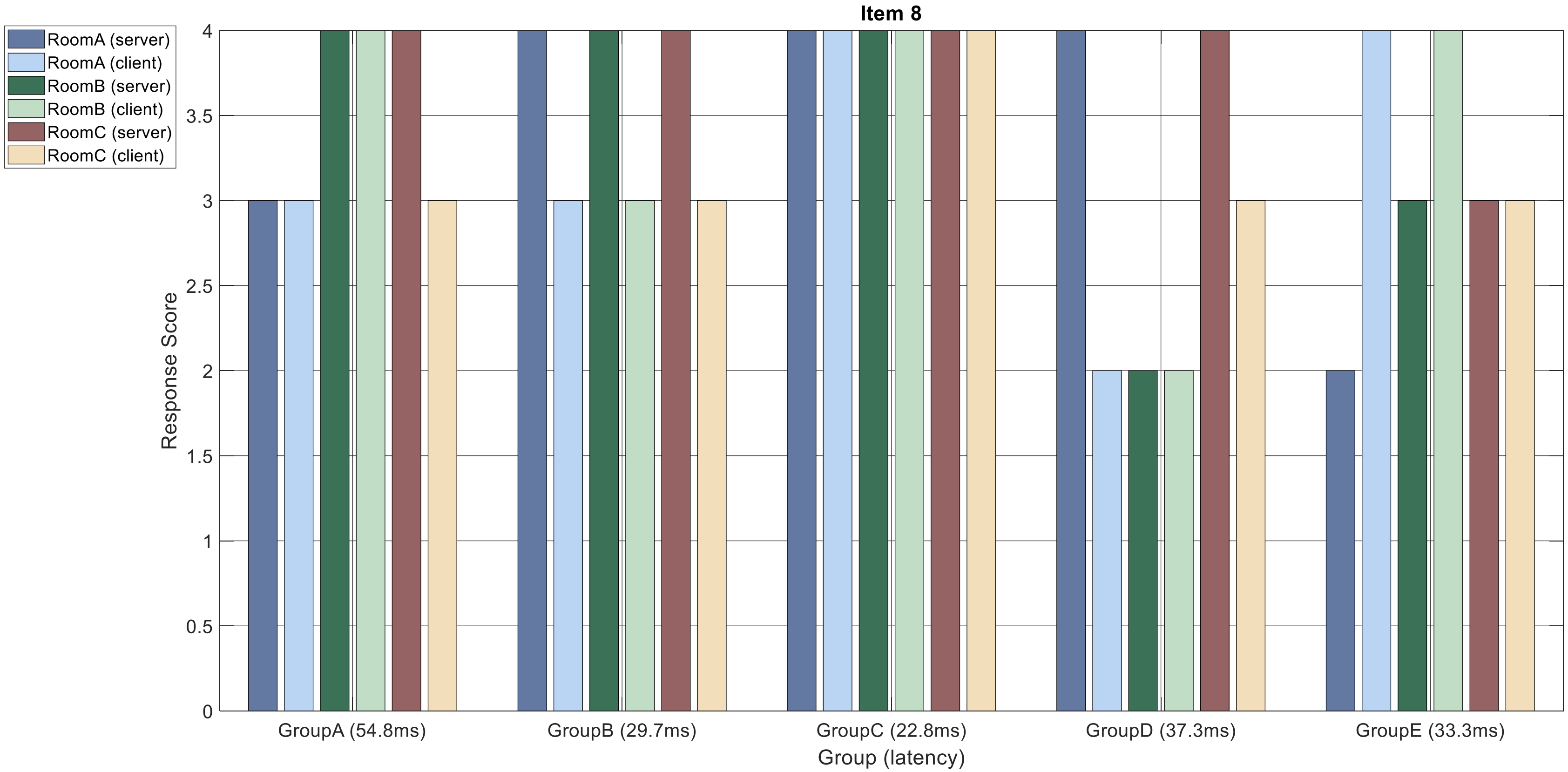


Figure 4.3.8.1 Response Scores for questionnaire item 8.

4.3.9 Items 9 and 10

Items 9 and 10 of the questionnaire simply ask for response to the questions:

9. Have you enjoyed using the system?

10. Would you use the system again, and, if yes, in what situation would you use the system?

Response to these questions (*Figure 4.3.9.1*) demonstrated only one out of 10 participants did not enjoy using the system and this was only the case for Room B (the performer indicated they enjoyed rooms A and C). This presents a total of one performance out of 30 which was not enjoyed.

Comments regarding item 9 demonstrate some perceptual variation between rooms. A performer from group B notes that Room A was less enjoyable than the reverberant rooms B and C. The same performer notes that Room C was more enjoyable than Room B, specifically noting this was due to the perception that 'the acoustics had more clarity and less (Low Frequency) reverberation. This discrete ordering of enjoyment by room directly specifies an entertainment preference in this instance.

Item 10 response was sorted into four categories:

- 'Blank' where the participant did not complete this questionnaire item.
- 'Maybe' where a condition is given under which the performer would use the system again.
- 'No'
- 'Yes'

Response demonstrates an overwhelming 19 responses indicating performers would use the system again. In one exclusive instance (Room B and Group A) a performer indicated they would not use the system again. This was the highest latency pair, Group A (54.8ms). The reason given was that the performer expects' dry voice over IP', indicating that this is a personal preference. 5 performance instances reported participants would maybe use the system again, and in 5 instances no response was given.

Amongst the 'maybe' conditions given a participant from Group B noted that the system would be useful 'if (the system) was easier to configure'. This speaks to the prototype design that was deployed to participants. It is noted that this design was developed for a researcher to operate under lab conditions and is far from a 'plug-and-play' user friendly solution. Despite the extensive guide material the setup process is still well beyond what one would expect a typical performer to have to manage. Though participants performed well in setting up and operating this equipment themselves it is noted that an accessible User Interface would vastly improve engagement.

Participants also took opportunity to make statements about room preference in 'maybe' responses. One performer from Group B indicated that they would maybe use Room A again but would be 'more likely to choose (reverberant) rooms'. Conversely a participant from Group D indicated a preference for Room A as Room B was considered 'too reverby'.

From 'yes' responses a range of potential applications for Immersive NMP were identified. These included:

- Remote Music Education
- Remote Practice and Performance
- Remote Recording
- Podcasts
- Social Teleconferencing

A professional level singer from group E (the non-researcher participant) expressed 'I would love to try this out with a four part choral ensemble' in response to reverberant rooms, however expressed a preference of reverberant over dry for singing, stating that Room A would be more appropriate for 'playing drums and bass with friends' or social teleconferencing.

A participant from Group C, the lowest latency group (22.8ms), expressed that it was 'amazing that you could do harmonies over the internet'.

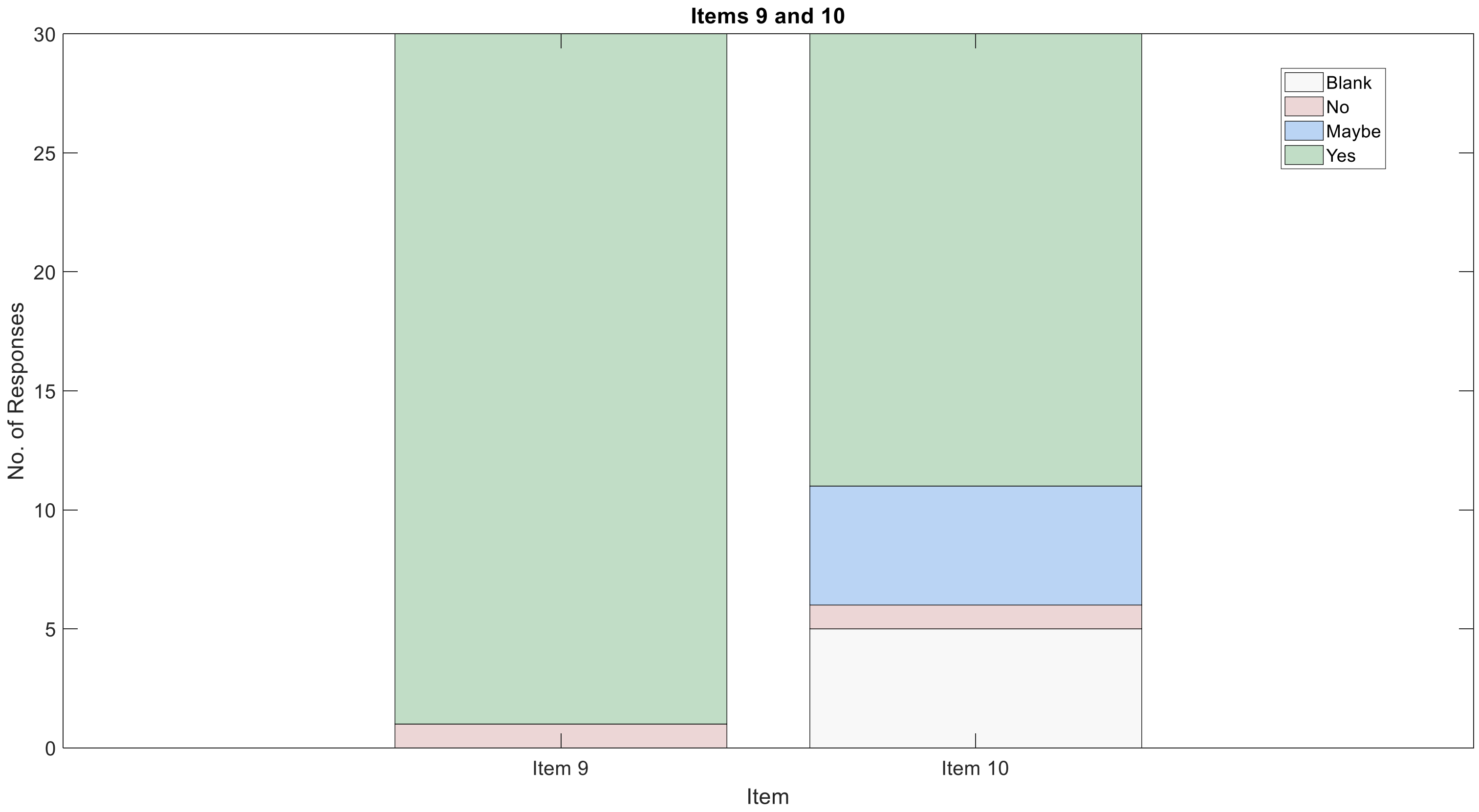


Figure 4.3.9.1 Response to questionnaire items 9 and 10, respectively ‘Did you enjoy using the system?’ and ‘Would you use the system again?’

4.3.10 Discussion: Performance Experience Questionnaire

Whilst there is not enough data for statistical analysis the questionnaire response provides a useful description of participant's individual perceptions of the system.

Firstly it can be identified that there is likely an effect of latency on Perceived Immersion. This appears to act via specific immersive properties. Interface Awareness is affected where high latency causes performers to notice a delay, and thus place some focus on the fact that their interactions are mediated by human-computer interface. Noticeable delays appear to be perceived as less Natural, making the performance experience using the VIIVA-NMP audio system prototype less Coherent with recall of live performance and less respectively Plausible. This would conform to expectation from previous research which indicates that latency is a major consideration in designing for Immersion in VR systems and that high latency will impair qualities of Immersion [30].

Secondly, it can be demonstrated through free comment response that the performance experience is a different kind of experience from live performance, and that this may require some adjustment from performers before the system begins to achieve high rating of Perceived Immersion qualities. This conforms to expectation from previous research [98], however is likely to be improved upon with the implementation of proposed Immersive NMP avatar rendering. Avatar rendering alongside virtual performance space visual rendering for VR will likely also yield improvements in interface awareness identified where performers felt what they heard did not match the room they were in.

Notably in response there are some instances where measurement of Perceived Immersion qualities in questionnaires fails to identify Naturalness, Coherence and Interface Awareness issues associated with latency that were successfully identified in synchrony measurement.

Some issues raised indicate that greater enjoyment and engagement can be achieved if the system is made more accessible to non-technical users. Largely this speaks to the challenges presented by the Covid 19 pandemic where the VIIVA-NMP audio system prototype, designed for use by researchers, is deployed remotely to a population of largely non-expert participants. Despite guide material provided and the assistance of the lead researcher the technical set-up and operation was challenging and unfamiliar. In either instance minimising the cognitive load on the performer in all stages of the technology can be readily identified as essential to Accessibility and Engagement with Immersive NMP technology.

4.4 On the Effect of Room

Response from the Performer experience questionnaire demonstrates multiple occasions where performers express a preferential relationship between Rooms, indicating that this affected their rating of Perceived Immersive qualities.

Of keen interest is the fact that on multiple occasions performers directly attributed differences between Rooms to differences in perception of the performance task and performance synchrony. In Group C (22.8ms) a performer considered Room B (medium hall) to help performers ‘keep in time’ and ‘tune in’. The same participant from Group C (22.8ms) notes that Room A (studio) was ‘good for precision’.

A participant from Group B (29.7ms) provides concurrent feedback, stating Room A was ‘more precise’ specifying that ‘pitch and time misalignment was more clear’. The participant from Group B (29.7ms) here not only attributes Room characteristics to perception of synchrony, but review of C50 values for each room (*Tables 3.4.1.1, 3.4.2.1 and 3.4.3.1: Rooms A, B and C respectively*) demonstrates this participant has correctly identified that Room A (studio) features the highest C50 amongst the Rooms. This same participant continues to state that Room C was ‘more enjoyable’ than Room B as it had ‘less (Low Frequency Energy)’ and ‘more clarity’. In this case this participant has again correctly identified that Room C (cathedral) featured shorter RT50 in low frequency bands and higher C50 across the spectrum when compared to Room B (medium hall).

The successful identification of room acoustic parameters and the direct statement that these parameters are affecting perception of performance synchrony would strongly suggest that the perception of the individual may be effectively modelled based on room acoustic parameters.

As some indication of variation of physical synchrony metrics between room it is hypothesised that the reported perceptual effect is potentially influencing an observable physical effect on performance synchrony (though in depth analysis of this will need to be the topic for future study).

Indeed observation of significant difference between Room was measured within Group B on two occasions. Asynchrony analysis demonstrates significantly lower Asynchrony for Group B in Room B (medium hall) than other Rooms. Analysis of Mean Tempi Ratio demonstrated that in the dry environment (Room A) the client performer (who made the above comments) tended to perform faster than the server performer, unlike with Rooms B and C.

Indeed in empirical EPT study it is noted that live (reverberant) environments may allow establishing of superior synchrony [28], and that reverberation may have a smoothing effect which can mediate tempo acceleration described by the Chafe Effect [31]. Indeed reverberation is known to ‘smooth’ sounds, fundamentally elongating the attack slope of onsets. Barbosa et al [212] identify that slower attack slopes can allow for improved synchrony in NMP when compared to performance with fast attack slopes. For Group B analysis of Mean Tempi Ratio appears to demonstrate that the rate of acceleration by the client is mediated by reverberant Rooms B (medium hall) and C (cathedral). It could well prove that appropriate reverberation may provide mediation of the Chafe effect and other tempo acceleration.

It is also considered that the smoothing effect of reverberation may be widening the P-centre frame [49], [50], [211] where performers perceive the onset of a musical sound. In this case such widening could, in low latency conditions, conceivably act as error concealment for performer perception by increasing the variability of where onsets can be acceptably placed.

Distance perception also provides an avenue of hypothesising. Room A (studio) provides SIR measurement with short source-receiver distance than in measurements used for Room B (medium hall) and Room C (cathedral). Further, absolute monaural distance cues include direct to reverberant energy ratio [46], [45]. As such a reverberant sound can be identified as having travelled further distance than a dry equivalent. Associated with this travelled distance is an air propagation delay.

This hypothesis considers two discrete listening conditions, one where at time t=0 a dry sound arrives at the outer ear, and one where at time t=0 a reverberant sound arrives at the outer ear. In the first condition, the dry sound represents minimal air propagation, and the sound can be assumed as occurring at approximately t=0. In the reverberant case, however, the reverberant energy implies a degree of air propagation prior to the sound arriving at the outer ear. In this case the sound can be recognised as initialising at t = -T where T represents the time taken in air propagation. In this case it may conceivably be possible that reverberation may actually smooth latency perception by simulating part of the network latency as air propagation. This topic provides a valuable pathway for further work.

5 CONCLUSIONS AND FURTHER WORK

The work presented in this thesis provides investigation into audio system design, implementation, and operation in the field of Immersive NMP. The central research question this project sought to address can be given:

- ***Is it possible to design an audio system suitable for Immersive NMP?***

Providing a robust answer to the project question required the investigation of several sub-objectives, which are surmised by the questions:

- ***Can the VIIVA-NMP audio system design facilitate natural musical interactivity between remote performers?***
- ***Does the effect of latency on performance synchrony using the VIIVA-NMP audio system design conform to empirical modelling of the effect of latency in NMP?***
- ***What is the potential and what are the limitations associated with the VIIVA-NMP audio system design in practical use cases?***
- ***Does the reverberant qualities of the virtual acoustic performance space simulated using the VIIVA-NMP audio system design affect the immersive quality of the system?***

In order to address these questions the VIIVA-NMP audio system design is presented, inspired by the design of the VIIVA system [3]. This system design was specified to provide audio data transport, acoustic simulation and spatial audio delivery which is suitable for implementation alongside consumer VR technology in future work (though the VIIVA-NMP audio system must be recognised as an audio-only design, and no VR visual display was used in this project).. Further specifications include the capability to provide full system one-way latency less than 30ms, and viability for operation by typical home users performing over consumer internet.

The VIIVA-NMP audio system design, being certainly amongst the first designs published which provides this functionality, provides the first original contribution of this thesis.

A prototype implementation of the VIIVA-NMP audio system design was then developed, using a range of open source resource including: Jacktrip [100], Kronlachner VST [196], [195], Open Air SIR Library [152], and the SADIE II Binaural Database [199].

The instance of the VIIVA-NMP audio system prototype deployed to test participants should be noted as a static binaural system. Though the VIIVA-NMP design includes 3DoF functionality, this simply could not be deployed to remote participants in the scope of this study and with practical limitations presented during the covid-19 pandemic. Evaluation of performance with 3DoF enabled will need to be the topic of future research.

This prototype was deployed to musicians across Europe who took part in testing, creating remote connections between their respective homes and delivering vocal duet performance. Analysis of this testing using the developed Timex-Lite onset detection tool demonstrates that natural musical interactivity can indeed be achieved using the VIIVA-NMP audio system, and that the effect of latency on the level of interactivity between performers conforms to the thresholds described in empirical EPT research [28], [31].

In the context of Immersive NMP (specifically with SIR Auralisation and virtual Ambisonic spatial audio delivery across latencies extending below the 30ms EPT) this study provides the second original contribution of this project.

The practical nature of the testing deployment presents a further point of originality in this study, as sparse data exists describing practical use cases in NMP, and little is identifiable in the field on Immersive NMP. The real-world testing scenario allows not only validation of system operation by typical home users, working on consumer internet, but also facilitates the presentation of technical specifications associated with 'typical home user' use cases. This technical specification describes the limits within which the VIIVA-NMP audio system may provide natural interactivity and optimised immersive quality, and out-with which performers can be expected to experience impaired interactivity and immersive quality.

Though no concrete conclusions can be drawn from investigation of the effect of different reverberations in this study, analysis of data collected here does still provide some interesting discussion and suggestion of direction for future work.

5.1 VIIVA-NMP Audio System Design: Conclusions

The audio system designed, prototyped, and tested (with disabled 3DoF functionality) in this project was developed to integrate immersive audio technology into an NMP system. The motivation of achieving this is, in the majority part, to provide an immersive audio system which is suitable for implementation alongside proposed VR visual displays in the emerging field of INMP. It should be noted that though VR NMP presents the wider context which motivates the research detailed in this thesis, no visual display was used in this project in order to control system evaluation by not introducing unknown visual contact factors.

At the outset of this project no systems which provided this functionality were identifiable, and at the point of concluding this thesis only one similar system has been identified as in development, though no publishing has been identified (Digital Stage [257], a telepresence-NMP system based off Soundjack [93], [258]. Digital Stage appears to provide FDN reverberation and Binaural or ORTF playback [259] using the Toolbox for Acoustic Scene Creation and Rendering, TASCAR [260]). Indeed current NMP literature would suggest that the design of an NMP solution which provides acoustic simulation and spatial audio delivery while remaining under the EPT of 30ms may well not be possible [261].

VIIVA-NMP therefore represents a design which is amongst the first Immersive NMP audio systems. This design fulfils the requirements of Immersive NMP audio systems in that it provides Immersive Audio (namely real-time acoustic simulation, spatial audio delivery and 3DoF) and is capable of providing audio transport and rendering at latencies lower than the 30ms EPT. This can be achieved by the typical home user using consumer internet.

Although no visual displays were used in this study, the design of the system included relevant functionality, such as provision for synchronisation of haptic metadata. Consideration of the wider INMP context allowed the design of a system which is suitable for implementation alongside immersive visual display in further work. Immersive NMP has proposed the use of Avatars in VR to circumvent the latency and bandwidth constraints of video streaming, and current research in the use of such visual display on LAN systems with no Immersive Audio suggests that this method shows promise. As such the VIIVA-NMP design has specified a transport solution such that timestamp and sequence packet fields may be used to synchronise the haptic metadata associated with avatar rendering to the audio stream. This packet structure also provides the minimal fields required to fill in RTP [170] packets, such that bridging to LAN standards is possible.

The design is based on the Jacktrip audio streaming framework, which was identified as providing the required functionality for prototype implementation, combined with an adaptation of the VIIVA audio rendering design. The key point of deviation from the VIIVA system design is the removal of any interpolation from the audio rendering, such that SIR convolution, rotation gain matrix, virtual loudspeaker decoding gain matrix, and HRTF convolution in the virtual Ambisonic signal chain may all be implemented as LTI filters with short buffers and with reasonable processing requirement. This allows the minimisation of audio rendering latency and allows for minimisation of bandwidth cost by streaming only mono audio between performers.

Application is also considered in system design, where use of the virtual Ambisonic approach allows access to the Ambisonic scene presented to each performer, or as rendered separately. For future work, which may involve online audiences, this means that the performance scene can be compressed using the OPUS codec [179], [181] and broadcast from an intermediate server to provision for Binaural decoding and playback over headphones with 3DoF locally for online audience members.

Conclusively, the VIIVA-NMP audio system provides an original, effective and practical solution for Immersive NMP. The system prototype demonstrates that NMP with immersive audio and latency below the 30ms EPT is viable for typical home users, which fulfils missing functionality in current Immersive NMP research. Whilst tested as an audio-only system, it has been designed to be suitable for implementation alongside VR visual display and avatars in future INMP development.

5.2 VIIVA-NMP Audio System Design: Related Further Work

Though the VIIVA-NMP audio system prototype proved effective in testing, some potential improvements on the design have been identified. The contribution presented in this thesis also highlights some promising areas for future research.

5.2.1 Auralisation of Singer's Own Voice and Remote Performer Voice Directivity

As SIR measurements representing the same source and receiver location (as required to accurately auralise a singer's own voice) are not used in the prototype implementation of the VIIVA-NMP audio system. SIR which are measured with receiver locations which are distant from the excitation source are used instead, and pose a potential externalisation issue. Singer's-own-voice SIR measurements are generally rare, however creating or simulating such measurements does provide a point for improvement in system design.

The removal of interpolation functions from the VIIVA design includes the interpolation between directional SIR measurements which are used to simulate the directionality of the singer's voice. In the VIIVA-NMP audio system 3DoF is provided by scene rotation, however head rotation does not provide this directionality simulation function. In this regard the current system presents remote performers as remaining directionally stable throughout performance. Providing simulation of source directivity therefore presents another area for improvement in the existing VIIVA-NMP audio system design.

5.2.2 User Interfaces, Accessibility and Engagement

One topic which is not discussed at great length in this thesis is the experience of deploying the prototype and testing protocol to participants remotely. Though a necessity due to Covid-19 related restrictions, it is clear that this process was time consuming and impractical, and can be improved upon in many ways.

In the case of home network setup for Peer-Peer systems there is little that can be done to make the process easier. The operation of Jacktrip and use of Reaper in testing however presented processes to participants which can be hidden. This statement can, to a degree, be extended to wider NMP technology. Processes such as jitter buffer sizing and setting FEC redundancy are generally open to users, however may be possible to hide through automated 'start up' processes.

Specific to the VIIVA-NMP prototype development the creation of an accessible user interface is something which can be identified as capable of increasing engagement and usability in the project. Ideally all any user should be presented with is 'phonebook', 'start/end' call, and 'room select' functionality.

5.2.3 Processing Reduction

Hosting audio rendering locally at each user presents a processing overhead concern when the number of convolutions required to manage large groups may begin to cause issues even on computers which meet the requirements of state-of-the-art consumer VR technology. In this regard, and with consideration of potential experiments with wireless 5G networks and mobile devices, reducing processing overhead may be hugely beneficial to system design.

Notably DFDN reverberation [139], [262] shows promise in recreating natural Ambisonic reverberation. Effective implementation of Ambisonic Schroeder reverbs has potential to drastically decrease the processing requirements of Immersive NMP audio rendering, and also allows parametric control of reverberation which may prove useful in future research.

5.2.4 Clock Synchronisation

It has been noted that, though largely inaudible, the Jacktrip streaming method does suffer from periodic underruns and overruns due to clock drift associated with operating between asynchronous systems. Fortunately a low-cost GPS clock design has been designed [188], and is something to be included in future VIIVA-NMP system development.

5.2.5 Avatar Rendering

As discussed, visual display with room and avatar rendering has been proposed [9] for Immersive NMP and us currently being tested on LAN based systems without immersive audio [10]. Having provided initial validation of the VIIVA-NMP audio system design the next step is to integrate this visual display and provide investigation into social and communicative immersive dynamics. In its current state the VIIVA-NMP system may be implemented alongside existing web-based social VR applications, allowing for more practical use-case study by deploying to typical home users using consumer internet again. The long term goal for this project, however, should be the integration of this technology into an accessible VR/AR ‘game’ format.

5.2.6 Client-Server Architectures

Though the latency measurements obtained in this study demonstrate that Immersive NMP systems operating on consumer internet are unlikely to achieve any cases where the full Round-Time-Trip will be less than the 30ms EPT associated with natural interactivity, recent research which uses academic networks suggests that this may be possible provided servers are located on academic internets such as JANET [258]. Such system configuration would allow for the use of high quality audio rendering servers, and would allow access to much larger group sizes, as bandwidth requirement in this case is only high at the server. Spatial audio scene-rendering methods can be provided to render audio with 3DoF and deliver to performers using only one mono stream from client to server and two channels back from Server to Client (Binaural Stereo) for any size of performance group.

5.2.7 Immersive Broadcast

The system is specified such that the Ambisonic scene representation of the performance can be used in Immersive broadcast to online audiences. Future research should include implementation and case study in this area. Associated with this, further research with the OPUS codec will be required to with respect to the compression of Ambisonic scenes [181], in order to identify appropriate compression bitrates associated with the provision of high quality immersive audio to remote audience members in this context.

5.2.8 LAN and RTP Bridging

As discussed, the VIIVA-NMP audio system is designed such that conversion to RTP packet structure is possible, and that implementation of this would allow for communication with LAN based systems operating on RTP-based LAN standards such as AES 67 [189]. Indeed such bridging efforts could also allow for communication between NMP frameworks using RTP packet structure as a common reference point. It is considered that the provision of such interoperability may circumvent a major obstacle in the adoption of NMP technology by the professional audio broadcast community and industry.

5.3 Natural Musical Interactivity, Synchrony and Latency: Conclusions

Though analysis of Precision and Asynchrony provided little information with respect to the known effects of latency on performance synchrony in NMP, the results of Tendency to Lead and Tempo-based metrics allows the observation of some significant results. Though data measured only represents the performance pairs who took part in testing, the fact that data collected for each pair conforms to the relevant expected effect of latency allows some confidence in the accuracy of these results, and suggests that the application of acoustic simulation and spatial audio is extremely unlikely to require redefinition of empirical modelling of the effect of latency in NMP.

Notably Group B (29.7 ms) and Group C (22.7 ms) performance exhibited synchrony metrics which conformed to values reported to indicate 'natural' musical interactivity for all measures. This high quality performance experience was also reflected in response to Performer Experience Questionnaire items for these groups, indicating an associated optimisation in Perceived Immersive qualities. Group E (33.3ms) was recognised as suffering from tempo deceleration, while Group D (3.7.3ms) was identified as utilising a Leader-Follower relationship to mediate the effect of latency with some success. Group A (54.8ms) demonstrated an unsuccessful Leader-Follower attempt and indicated some collapse of synchrony in performance. An associated impairment of Perceived Immersive qualities is suggested in questionnaire response across these higher latency groups, though notably with less sensitivity than synchrony analysis. Combined with data collected on technical configuration and hardware the study conducted allows not only some conclusive validation that 'natural' musical interactivity can be achieved using the VIIVA-NMP audio system design in a practical use case, but also the presentation of the associated technical requirements (*Table 4.2.7.1*) of achieving this.

Testing of the VIIVA-NMP system provides new evidence for the potential and limitations of Immersive NMP within the context of current research. Estimations of 1000km range with 1Mb/s bandwidth consumption per channel and under 30ms latency [10] are likely to be reserved for academic and other high speed networks where jitter and packet loss have been eliminated, and while using 16-bit, 44.1kHz or 48kHz audio. A practical estimate is provided by this study in that, if meeting minimum specified requirements, a range of 300km and up, and a bandwidth consumption of 4.7-7Mb/s per channel will be the requirements for the typical home user of Immersive NMP systems wishing to operate below the 30ms EPT and achieve natural musical interactivity. Above this range it is likely that performers will experience impaired musical interactivity and are likely to achieve an associated impairment in Perceived Immersive qualities as can be expected based on previous research on Immersive Audio for VR. This provides informative data for future implementation of potential Immersive NMP applications which are undergoing recent and current research [9], [10], [11], [8].

5.4 Natural Musical Interactivity, Synchrony and Latency: Related Further Work

This study presents a successful proof of concept for the design and implementation of the VIIVA-NMP audio system, and highlights several aspects of the wider NMP experience that require further investigation.

5.4.1 Further Data Collection

Although the analysis of synchrony data proves significant in describing the groups who participated in testing, no concrete confirmation is given that empirical EPT thresholds hold or differ in the context of immersive audio. Instead the fact that our data conforms to empirical EPT models allows us to present a hypothesis: ***Empirical EPT latency thresholds will remain accurate in the context of immersive audio***. As this study demonstrates that the implementation and configuration of acoustic simulation and spatial audio does have an effect on at least the perception of Immersive NMP musical interactions, this hypothesis must be given with the caveat: ***however, immersive audio is expected to have an impact on at least the perception of delayed musical interactions***. Further data collection with a larger population of participants will be required to test this hypothesis, though, as previously stated, it appears extremely likely that this hypothesis can be accepted.

The same can be said of the Performer Experience Questionnaire, which provides no statistical significance, but instead provides useful indication of Perceived Immersive quality. This data again conforms to expectations, as current immersive audio research is aware high latency will impair immersive quality [30]. Effective modelling of the effect of latency on auditory Perceived Immersion in the low latency range required by Immersive NMP will require further data collection in future study to see if this is possible.

5.4.2 Expanded Performer Experience Questionnaire

In addition to data collection from more participants, the Performer Experience Questionnaire could be expanded upon to provide a more robust measurement of Perceived Immersion. Notably, once the system is further developed to include visual display with avatars and room, Communication [98] and Embodiment [21] qualities will be extremely relevant to the questionnaire.

5.4.3 Eye-Tracking and Motion Capture

Particular to implementation alongside avatar rendering and visual display monitoring of gaze [206] and gesture [21], [19] will be relevant to providing physical metrics for measures of communicative immersive qualities. As such it will be relevant to implement appropriate eye tracking design [207] and motion capture methods.

5.5 Performance Experience, Synchrony, and Different Rooms: Conclusions

Though no conclusive statements can be made regarding the effect of different reverberations in testing, several indications are made which can be used to direct further research.

Although in two isolated cases (Asynchrony for Group B and Mean Tempi Ratio for Group E) significant difference between measurements for different simulated rooms is observed in one case this may simply due to narrow measurement sets, and the fact that this is not prevalent across different groups would suggest no robust statement regarding the effects of different simulated rooms on relevant synchrony metrics can be made. In one case (Mean Tempi Ratio) analysis of Within-Subject Effects (*Table 4.2.3.3*) does, however, indicate that an interaction effect of Room and Latency is present across the entirety of data collected.

Questionnaire response is much more informative on this topic of investigation. Performer comment explicitly states a preferential ordering of the simulated rooms in multiple cases, and in some cases directly attributes this to correctly identified room acoustic properties such as Clarity [144].

Different preferences between Groups are also apparent from participant comments. As each Group represents a discrete latency, but also a discrete pair of performers, it is unclear whether this represents an interaction effect between latency and room, or performance pair preferences and room.

Considering this response, and considering that some statistically significant difference between rooms were observed in synchrony analysis (with the aforementioned caveat regarding the narrowness of measurement), it seems some effect is present relevant to room, though this requires further study.

A relevant hypothesis is presented for future research: ***The acoustic environment will have a perceptual or psychoacoustic effect on performance synchrony which is dependent on acoustic parameters and/or personal preferences.***

5.6 Performance Experience, Synchrony, and Different Rooms: Related Further Work

5.6.1 Preference Learning

Considering that variation in response due to room appeared most observable in questionnaire rating of Perceived Immersion this seems the logical point of focus for future research. Artificial Intelligence Preference Learning [247] methods have shown some success in explaining ranking based on perceptual features, and will be the method of choice in future research, specifically the application of interactive Mushra [204] variant testing will allow gathering of Room ranking data, where AI methods may then be applied to search for a predictive model based on reverb parameters.

5.6.2 Parametric Reverb Generation

In order to conduct a controlled study of the effect of virtual rooms the parametric generation of reverberations will allow for best control over this process. DFDN [139], [262] reverberators may prove the best option for this function, however the generation of SIR using image source [137], [142] or ray tracing [136] acoustic simulation methods may prove a more appropriate alternative.

5.7 Other Related Further Work

5.7.1 Telemedicine Applications

Several medical applications of virtual reality performance technology have been identified such as to access the beneficial health and wellbeing effect of group singing [9], as a complimentary therapy for medical procedures [10], or as exposure therapy for social anxiety or performance anxiety in group singing [11]. The development of an accessible user interface for the VIIVA-NMP design should allow for integration into existing projects such as Sing From Your Seat [56] and provide a valuable tool for telemedicine research.

5.7.2 Case Study with Small-Medium Groups

The VIIVA-NMP audio system, having undergone preliminary validation, is ready to be used in case study performances. Investigation with small-medium groups will allow for the collection of data which may be different from expectations extended from the duet context presented in this thesis. Inclusion of larger numbers of performers will also allow for some investigation into the processing limitations of the VIIVA-NMP design, where the convolution-heavy structure is expected to yield an associated limit on group size.

5.8 Summary

This project has demonstrated that Immersive NMP audio systems are not only viable for typical home users, but that latency below the 30ms EPT and natural musical interactivity is possible with high quality immersive audio. Comprehensive analysis provided understanding of the limits and potential of this technology, and indicates that the effect of latency in the context of Immersive NMP is likely to follow the modelling presented by empirical EPT study. The design which achieves this has been clearly presented for reference and repeatability in future Immersive NMP design and research. Though results regarding the effect of room are inconclusive some interesting directions for future research has been identified. The VIIVA-NMP audio system design and prototype and associated testing presents a novel and original contribution to the field of Immersive NMP and demonstrates: ***that it is indeed possible to design and implement an Immersive NMP audio system which provides immersive audio functionality, and can deliver natural musical interactivity.***

Moreover, the design of this system has considered the wider context of this thesis, namely INMP VR and avatar displays, and has presented an audio system which is suitable for implementation alongside such visual display in further work.

The evaluation provided validates immersive audio NMP as a discrete experience, and provides a baseline study against which future work, which shall include VR visual display, can be compared to.

APPENDICES

The appendices for this project consist of a digital appendices accompanying this thesis. The content of these digital appendices are as detailed in this section.

Appendix A: Testing Distribution

This supplementary material contains the testing distributable as provided to participants in testing. .

Within this directory is:

- ***/Binaural Decode:** This folder contains resource sourced from the SADIE II Binaural Database. This content consists of .wav and configuration files for the KU100 dummy head. The .wav files are sorted by sample rate.*
- ***/Guides:** This folder contains the guide material provided to participants. This material consists of the Setup Guide and the Test Procedure Guide as pdf and word documents. Please note that the public IP address used in this guide is dynamic no longer conforms to any known location.*
- ***/Impulse Response:** This folder contains resource sourced from the Open Air Impulse Response Library. This content consists of configuration files and associated .wav files for SIR measurements used in testing.*
- ***/Jack and Jacktrip:** This folder contains windows and mac installation files for Jack and Jacktrip and ASIO4ALL driver.*
- ***/Participant Consent and Information:** This folder contains the Participant Consent Form and Participant Information Sheet provided as pdf and word documents.*
- ***/Questionnaires:** This folder contains the Participant Hardware Questionnaire and Performer Experience Questionnaire (here listed as 'Participant Hardware Setup Questionnaire' and 'Participant Questionnaire' respectively. These are provided as pdf documents.*
- ***/Reaper (mac):** This folder contains Reaper installation files for mac users.*
- ***/Reaper Portable Install:** This folder contains the portable installation of Reaper for windows users.*
- ***/Reaper Sessions:** This folder contains the preconfigured Reaper files used by participants in the testing protocol. These Reaper sessions are sorted by room and by sample rate.*
- ***/Sheet Music:** This folder contains the pdf documents containing indicative sheet music for the performance task for the duet performance and for each part. Also contained in this folder is a .wav file providing indicative piano example of the performance task as a duet, and a .wav file indicating the melody for an individual performer.*
- ***/VST:** This folder contains the unpacked Kronlachner VST used to perform audio rendering processes. These VST are provided for Mac and Windows.*
- ***/VST Installers:** This folder contains Mac and Windows installation files for the Kronlachner VST used to perform audio rendering processes.*

- ***README:*** *This is an introductory message to testing participants upon receiving the distributable.*

Usage

A reader may choose to use this material to recreate the testing setup and protocol as detailed in the Setup Guide and Test Procedure Guide located in the /***Guides*** folder.

A reader may also wish to load the Reaper Sessions to experience the room simulations used in this study. For Windows users this is accessible via the /***Reaper Sessions*** folder which contains the pre-configured Reaper Sessions.

As these Reaper Sessions are organised by sample rate the reader should select either /48kHz or /96kHz in the subdirectories to match their audio interface. These sessions are also sorted by Client and Server as per the testing protocol pre-configuration.

Please be aware that as these sessions are configured to work with Jack and Jacktrip as per the testing protocol readers who are simply testing the room simulations may need to change the audio interface selected in the Reaper preferences to their own interface. This can be done in Options/Preferences/Devices in Reaper.

It is also possible that the VST may not load due to differences in home systems. In this case there is VST are contained in the ***/VST folder***. To relink these to the Reaper sessions first open the Reaper session named 'Setup', then go to Options/Preferences/VST in Reaper and add the VST folder for the relevant operating system (located in the ***/VST folder***) to paths. Save and try loading the Reaper Sessions again and the preconfigured session will operate correctly.

Appendix B: Onset Detection

This supplementary material contains data and resource as used in the onset detection component of this project.

Within this directory is:

- ***/Call Functions:*** *This folder contains a .txt file which details the Matlab call functions for the timexLite function as used in analysis of audio recordings of the performance task. These call functions additionally specify the parameters of the timexLite function used in analysis.*
- ***/Onset Detection Graphics:*** *This folder contains graphics detailing the onset detection process for each audio recording. These graphics are available as both JPEG and Matlab Figure documents.*
- ***/Onset Times:*** *This folder contains .txt files detailing onset time lists for each individual performer who took part in testing.*
- ***/Stems:*** *This folder contains audio recording of each take by each individual performer who participated in testing, stored as .wav files.*
- ***/TimexLite Matlab Script:*** *This folder contains the timexLite Matlab function used for onset detection in this project.*
- ***/TimexLite Sample F measure:*** *This folder contains the sampled onsets and relevant stems used for informal F-measure testing of the timexLite function.*

Usage

The reader may wish to repeat the onset detection process detailed in this thesis. To make this easy a list of call functions for the timexLite function is included in the **/Call Functions** folder. These functions may be input to Matlab. A UI box will open where you may select the appropriate .wav file. The timexLite function will then run and output a graphic detailing the onset detection process, and a .txt file containing a list of onset times named 'onsets.txt.'. Use of the timexLite function requires the Signal Processing and Audio Toolboxes.

Appendix C: Synchrony Analysis

This supplementary material contains data and resource as used in the synchrony analysis component of this project.

Contained in this directory is:

- ***/Call Functions:*** *This folder contains a .txt file which lists call functions for the synchrony analysis script, autoMetric3.*
- ***/Synchrony Analysis Matlab Script:*** *This folder contains the Matlab function, autoMetric3, which was used to calculate synchrony metrics from onset time lists.*
- ***/Synchrony Data:*** *This folder contains a .sav (SPSS) data document containing all synchrony data.*

Usage

The reader may also wish to repeat synchrony analysis using the autoMetric3 Matlab function. The call functions required to do this are listed in the .txt file in the **/Call Functions** folder. To do this it will be required that the text files detailing the relevant onset lists are in the same folder as the autoMetric3 function. These onset lists can be found as .txt files in **/Onset Times** folder detailed in **Appendix B**.

Appendix D: Tempo Analysis

This supplementary material contains resource and data from the tempo analysis component of this project.

This directory contains:

- ***/Call Functions:*** *This folder contains a .txt file which lists call functions for the tempo analysis script, autoTempi2*
- ***/Tempo Analysis Data:*** *This folder contains .txt files associated with each attempt of the performance task. These files each contain a list detailing the server mean tempo, client mean tempo, and mean tempi ratio in that order.*
- ***/Tempo Analysis Graphics:*** *This folder contains graphics for tempo and tempo slope for each performer, and tempi ratio between performers.*
- ***/Tempo Analysis Matlab Script:*** *This folder contains the Matlab function autoTempi2 which was used for tempo analysis in this project.*

Usage

The reader may wish to repeat the tempo analysis conducted in this project. The call functions used for the autoTempi2 function are included in the **/Call Functions** folder. In order to successfully repeat the tempo analysis the onset lists from **/Onset Times** detailed in **Appendix B** will need to be placed in the same path as the autoTempi2 function. When called the autoTempi2 function will output a .txt file containing server mean tempo, client mean tempo, and mean tempi ratio, as well as graphics detailing tempo and tempo slope for both server and client performers, and a final graphic detailing tempi ratio across the performance.

Appendix E: Statistical Analysis

This supplementary material contains statistical analysis data including descriptive statistics and repeated measure analysis for synchrony data. For each metric a .sav (SPSS) data file and .spv (SPSS) output file are presented.

This directory contains:

- ***/Asynchrony:*** *Data and statistical analysis for Asynchrony measurements.*
- ***/Mean Tempi Ratio:*** *Data and statistical analysis for Mean Tempi Ratio measurements*
- ***/Mean Tempo Slope:*** *Data and statistical andalysis for Mean Tempo Slope measurements.*
- ***/Mean Tempo Slope Difference:*** *Data and statistical analysis for Mean Tempo Slope Difference measurements.*
- ***/Precision:*** *Data and statistical analysis for Precision data.*
- ***/Tendency to Lead:*** *Data and statistical analysis for Tendency to Lead data.*

Appendix F: Other Matlab Scripts

This folder contains:

- ***/FuMa to AmbiX:*** *The Matlab script used to convert legacy FuMa SIR to current AmbiX standard.*
- ***/Pulse Generation:*** *The Matlab script used to generate the pulse signal used for latency measurement.*

Usage

The reader may wish to use the FuMa to AmbiX conversion script for udating their own FuMa SIR measurements. To do this simply run the script, and a UI will open to select the FuMa audio file. The script will then convert to AmbiX format and save this audio file with '_ambiX' added to the end of the original filename.

ABBREVIATIONS AND GLOSSARY OF USEFUL TERMS

Abbreviations

- ***3DoF: Three degrees of freedom.*** *Describing rotations around a static origin point.*
- ***AmbiX: Ambisonics Exchangeable.*** *Current standard Ambisonic channel definitions.*
- ***AR: Augmented Reality.*** *Systems where sensory stimuli is projected onto real-world experiences.*
- ***ARQ: Automatic Repeat Request***. *TCP error correction method based on the retransmission of missing or erroneous packets.*
- ***DbFS: Decibel Full Scale.*** *Decibel measurement for digital audio.*
- ***DFDN: Directional Feedback Delay Network***. Ambisonic multichannel variant of the Schroeder reverberator design, undergoing current research.
- ***FDN: Feedback Delay Network.*** Classic Schroeder reverberator design using delay lines.
- ***FEC: Forward Error Correction.*** *Data transport error correction through the sending of redundant overlapping packets in data streams.*
- ***FuMa: Furse-Malham.*** *Legacy Ambisonic B-Format channel definition.*
- HRTF: Head Related Transfer Function.
- ***ILD: Interaural Level Difference.*** *Describing level difference between the ears in binaural listening.*
- ***IMS: Image Source Method.*** Simple geometric acoustic simulation method.
- ***IOI: Inter-Onset-Interval.*** *Describing the difference between onset times within a part.*
- ***IR: Impulse Response***. *Acoustic measurement detailing the response of a real or simulated space to an excitation signal.*
- ***ISP***: *Internet Service Provider. Commercial supplier of internet connections found in the typical home.*
- ***ITD: Interaural Time Difference.*** *The difference between the time of arrival of a sound at one ear with respect to the other.*
- ***LAN: Local Area Network.*** *Local installation of networked devices connected by one or more switch. Typically hardwired via Ethernet or WLAN (wireless local area network) via one or more router.*
- ***MUSHRA: Multiple Stimulus test with Hidden Reference and Anchors.*** *Testing protocol for the evaluation of audio qualities.*
- ***N3D: 3-D Fully Normalised.*** *Ambisonic normalisation definition.*

- ***NMP: Network Music Performance.*** *Ensemble performance from remote musicians who share audio over the internet.*
- ***OTD: Onset Time Difference.*** *Detailing the disagreement in temporal location of onsets between parts in musical performance.*
- ***OWT: One-Way-Trip.*** *Detailing the latency between one location and another. Typically used to describe system throughput latency in Peer-Peer NMP systems.*
- ***RTP: Real-Time Protocol.*** *Protocol for the transport of real-time data typically used in LAN installations.*
- ***RTCP: Real-Time Control Protocol.*** *Accompanying control protocol for RTP.*
- ***RTT: Round-Time-Trip.*** *Detailing the latency from one location to another and back again. Typically used to describe the system throughput latency in Client-Server NMP systems.*
- ***SIR: Spatial Impulse Response.*** *Ambisonic variant of IR acoustic measurements.*
- SN3D: Schmidt semi-normalised.
- ***TCP: Transmission Control Protocol.*** *Connection oriented transport protocol, typically using ARQ error correction.*
- ***UDP: User Datagram Protocol.*** *Connectionless transport protocol which offers no inherent error correction or data stream management.*
- ***VIIVA system: Vocal Interaction in an Immersive Virtual Acoustic system***. *System for vocal performance in virtual reality with high quality acoustic simulation, spatial audio delivery, and 3-D video visual display.*
- ***VIIVA-NMP: Vocal Interaction in an Immersive Virtual Acoustic Network Music Performance audio system***. *NMP adaption of the VIIVA design for Immersive NMP as presented in this thesis.*
- ***VR: Virtual Reality.*** *Systems which present virtual sensory stimuli via computer display.*
- ***WAN: Wide Area Network.*** *Widely speaking any network of interconnected LANs, though in this thesis the WAN of interest is the Internet of Things.*

Other Useful Terms

Asynchrony (metric): *A synchrony metric, describing the mean of the standard deviations of absolute OTDs in ensemble performance.*

asynchrony (disambiguation): *A generalised term for synchrony measurement, speaking to generalised OTD disagreement within an ensemble.*

Immersion: *"a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world."* [38]**.**

iPerf: *Tool for measurement of network throughput.*

Immersive NMP*: Proposed NMP systems which combine NMP and VR applications.*

Perceived Immersion: *Awareness and perception of one's state of immersion.*

Precision (metric): *A synchrony metric, describing the mean of absolute OTDs in performance.*

Presence*: The sense of being in a place.*

precision (disambiguation*): descriptive as typically employed, detailing fineness of accuracy.*

System Immersion: *A rating of immersion in terms of the ability of a system to provide sensory stimuli.*

Tempi Ratio: *The ratio of the tempo of one performer to the tempo of another.*

Tempi Slope: *The rate of tempo acceleration or deceleration within a performance.*

Tendency to Lead (signed): *The median time by which one performer precedes another, where positive or negative values indicate which performer is leading.*

Tendency to Lead (absolute): *The median absolute time by which one performer precedes another.*

Telepresence-NMP: *NMP systems with video communication display included, typically on screen.*

Synchrony: *A generalised term, describing generalised OTD agreement within an ensemble.*

REFERENCES

[1] S. R. Ellis, "What are virtual environments?," *IEEE Computer Graphics and Applications,* vol. 14, no. 1, pp. 17-22, 1994.

[2] C. Armstrong and J. Brereton, "The Application of Flexilink in Muli-User Virtual Acoustic Environments," in *Interactive Audio Systems Symposium*, York, 2016.

[3] G. Kearney, H. Daffern, L. Thresh, H. Omodudu, C. Armstrong and J. Brereton, "Design of an Interactive Virtual Reality System for Ensemble Singing," in *Interactive Audio Systems Symposium*, York, 2016.

[4] B. Loveridge, "Networked Music Performance in Virtual Reality: Current Perspectives," *Journal of Network Music and Arts,* vol. 2, no. 1, 2020.

[5] J. Lazzaro and J. Wawrzynek, "A Case for Network Music Performance," in *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, New York, 2001.

[6] A. Carot, A. B. Renaud and P. Rebelo, "Networked Music Performance: State of the Art," in *AES 30th International Conference*, Saariselka, 2007.

[7] S. Duffy, D. Williams, J. Jansen, P. S. Ceasar, P. G. T. Healy, T. S. Stevens and I. C. Kegel, "Remote Music Tuition," in *9th Sound and Music Computing Conference*, Copenhagen, 2012.

[8] M. Iorwerth, D. Moore and D. Knox, "Challenges of Using Networked Music Performance in Education," in *26th UK AES Conference on Audio Education*, Glasgow, 2015.

[9] H. Daffern, D. Camlin, H. Egermann, A. J. Gully, G. Kearney, C. Neale and J. Rees-Jones, "Exploring the potential of virtual reality technology to investigate the health and well being benefits of group singing," *International Journal of Performance Arts and Digital Media,* vol. 15, no. 1, pp. 1-22, 2018.

[10] J. Tamplin, B. Loveridge, K. Clarke, Y. Li and D. Berlowitz, "Development and Feasability Testing of an Online Virtual Reality Platform for Delivering Therapeutic Group SInging Interventions for People Living with Spinal Cord Injury," *Journal of Telemedicine and Telecare,* vol. 26, no. 6, 2019.

[11] L. Bryce, M. Sandler, L. K. Andersen, A. Adjorlu and S. Serafin, "The Sense of Auditory Presence in a Choir for Virtual Reality," in *AES 149th Convention*, New York (online), 2020.

[12] J. Bissonnette, F. Dube, M. D. Provencher and M. T. M. Sala, "Virtual Reality Exposure Training for Musicians: Its Effect on Performance Anxiety and Quality," *Medical Problems for Performing Artists,* vol. 30, no. 3, pp. 169-177, 2015.

[13] T. Lokki, L. Savioka, J. Huipaniemi, R. Hanninen, T. Ilmonen, J. Hiipakka, V. Pulkki, R. Vaananen and T. Takala, "Virtual Concerts in Virtual Spaces - in Real Time," *Journal of the Acoustical Society of America,* vol. 105, no. 2, 1999.

[14] W. Woszczyk, D. Ko and B. Leonard, "Virtual Acoustic Environments for Music Performance, Rehearsal and Recording," *Journal of the Acoustical Society of America,* vol. 123, no. 5, 2008.

[15] S. Oxnard and D. Murphy, "Achieving Convolution-Based Reverberation Through Use of Geometric Acoustic Modeling Techniques," in *15th International Conference on Digital Audio Effects*, York, 2012.

[16] B. N. Postma, D. Poirier-Quinot, J. Meyer and B. F. Katz, "Virtual Reality Performance Auralization in a Calibrated Model of Notre-Dame Cathedral," in *Euroregio 2016*, Porto, 2016.

[17] J. L. Breese, M. A. Fox and G. Vaidyanathan, "Live Music Performances and the Internet of Things," *Issues in Information Systems,* vol. 21, no. 3, pp. 179-188, 2020.

[18] L. Turchet, C. Fischione, G. Essl, D. Keller and M. Barthet, "Internet of Musical Things: Visions and Challanges," *IEEE Access,* vol. 6, pp. 61994-62017, 2018.

[19] O. Hodl, G. Fitzpatrick and F. Kayali, "Design Implications for Technology-Mediated Audience Participation in Live Music," in *14th Sound and Music Computing Conference*, Espoo, 2018.

[20] S. Serafin, C. Erkut, J. Kojs, N. C. Nilsson and R. Nordahl, "Virtual Reality Musical Instruments: State of the Art, Design Principles and Future Directions," *Computer Music Journal,* vol. 40, no. 3, pp. 22-40, 2016.

[21] G. Hajdu, "Embodiment and disembodiment in network music performance," in *Body, Sound and Space in Music and Beyond: Multimodal Explorations*, London, Taylor and Francis, 2017, pp. 157-178.

[22] L. Turchet and M. Barthet, “Co-design of Musical Haptic Wearables for Electronic Music Performer's Communication,” *IEEE Transacrions on Human Machine Systems,* vol. 49, no. 2, pp. 183-193, 2018.

[23] R. Graham and S. Cluett, “The Soundfield as Sound Object: Virtual Reality Environments as a Three-Dimensional Canvas for Music Composition,” in *AES Conference on Audio for Virtual and Augmented Reality*, Los Angeles, 2016.

[24] B. Grimes, “Latency,” in *Networked AV Systems*, London, McGraw-Hill Education, 2014, pp. 328-329.

[25] A. R. Rasch, “Aspects of the Perception and Performance of Polyphonic Music PhD Thesis,” University of Groningen, Groningen, 1981.

[26] R. A. Rasch, “Synchronization in Performed Ensemble Music,” *Acoustica,* vol. 43, pp. 121-131, 1979.

[27] C. Chafe and M. Gurevich, “Network Time Delay and Ensemble Accuracy: Effect of Latency, Asymmetry,” in *AES 117th Convention*, San Francisco, 2004.

[28] N. Schuett, *The Effects of Latency on Ensemble Performance (Honours Thesis),* Stanford: CCRMA Stanford, 2002.

[29] C. Rottondi, C. Chafe, C. Allocchio and A. Sarti, “An Overview on Networked Music Performance Technologies,” *IEEE Access,* vol. 4, pp. 8823-8843, 2016.

[30] C. Eaton and H. Lee, “Quantifying Factors of Auditory Immersion in Virtual Reality,” in *AES International Conference on Immersive and Interactive Audio*, York, 2019.

[31] C. Chafe, J.-P. Caceres and M. Gurevich, “Effect of temporal separation on synchronization in rhythmic performance,” *Perception,* vol. 39, no. 8, pp. 982 - 992, 2010.

[32] S. Farner, A. Solvang, A. Saebo, Svensson and Peter, “Ensemble Hand-Clapping Experiments Under the Influence of Delay and Various Acoustic Experiments,” *Journal of the Audio Engineering Society,* vol. 57, no. 12, pp. 1028-1041, 2009.

[33] C. Barlette, D. Headlam, M. Bocko and G. Velikic, “Effect of Network Latency on Interactive Musical Performance,” *Music Perception: An Interdisciplinary Journal,* vol. 24, no. 1, pp. 49-62, 2006.

[34] C. Rottondi, M. Buccoli, M. Zanoni, D. Garao, G. Verticale and A. Sarti, “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, 2015.

[35] A. Carot, C. Werner and T. Fischinger, “Towards a Comprehensive Cognitive Analysis of Delay-Influenced Rhytmical Interaction,” in *Proceedings of the International Computer Music Conference*, Montreal, 2009.

[36] E. Chew, A. Sawchuk, R. Zimmerman, T. P. D. (. S. a. I. Tosheff), C. Kyriakakis, C. Papadopoulos, A. Francois and A. Volk, “Distributed Immersive Performance,” in *Annual National Association of the Schools of Music Meeting*, San Diego, 2004.

[37] S. D. Monache, L. Comanducci, M. Boccoli, M. Zanoni, A. Sarti, E. Pietrocola, F. Berbenni and G. Cospito, “A Presence and Performance Driven Framework to Investigate Interactive Networked Music Learning Scenarios,” *Wireless Communications and Mobile Computing,* vol. 2019, pp. 1-20, 2019.

[38] S. Agrawal, A. Simon, S. Bech, K. Baerentsen and S. Forchhammer, “Defining Immersion: Literature Review and Implications for Research on Immersive Audiovisual Experiences,” in *AES 147th Convention*, New York, 2019.

[39] M. Gurevich, D. Donohoe and S. Bertet, “Ambisonic Spatialization for Networked Music Performance,” in *17th International Conference on Auditory Display*, Budapest, 2011.

[40] W. Ritsh, “ICE - towards distributed networked computer music,” in *Proceedings of the International Computer Music Conference*, Athens, 2014.

[41] C. Drioli, C. Allocchio and N. Buso, “Networked Performance and Natural Interaction via LOLA: Low Latency High Quality AV Streaming,” in *Proc. of the International Conference on Information Technologies for Performing Arts, Media Access and Entertainment*, Porto, 2013.

[42] N. Darabi, P. Svensson and F. Snorre, “Quantifying the strategy taken by a pair of ensemble hand-clappers under the influence of delay,” in *AES 125th Convention*, San Francisco, 2008.

[43] B. H. Repp, “Sensorimotor synchronization: A review of the tapping literature,” *Psychonomic Bulletin and Review,* vol. 12, no. 6, pp. 969-992, 2005.

[44] B. H. Repp and Y.-H. Su, “Sensorimotor synchronization: A review of recent research,” *Psychonomic Bulletin Review,* vol. 20, no. 3, pp. 403-452, 2013.

[45] A. W. Bronkhorst and T. Houtgast, “Auditory Distance Perception in Rooms,” *Nature,* vol. 397, no. 6719, pp. 517-520, 1999.

[46] S. H. Nielsen, “Auditory Distance Perception in Different Rooms,” in *AES 92nd Convention*, Vienna, 1992.

[47] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea and S. Pardhan, “Auditory distance perception in humans: a review of cues, development, neuronal bases, and effect of sensory loss,” *Attention Perception and Psychophysics,* vol. 78, no. 2, pp. 373 - 395, 2016.

[48] A. Danielsen, K. Nymoen, E. Anderson, G. S. Camara, M. T. Langerod, M. R. Thompson and J. London, “Where is the beat in that note? Effects of attack, duration and frequency on the perceived timing of musical and quasi-musical sounds,” *Journal of Experimental Psychology Human Perception and Performance,* vol. 45, no. 3, pp. 402-418, 2019.

[49] J. Morton, S. Marcus and C. Frankish, “Perceptual Centres (P-centres),” *Psychological Review,* vol. 83, no. 5, pp. 405-408, 1976.

[50] S. M. Marcus, “Acoustic determinants of perceptual center (P-center) location,” *Perception and Psychophysics,* vol. 30, no. 3, pp. 247-256, 1981.

[51] C. Chafe, S. Wilson, R. Leistikow, D. Chisholm and G. Scavone, “A Simplified Approach to High Quality Music and Sound Over IP,” in *COST G-6 Conference on Digital Audio Effects (DAFX -00)*, Verona, 2000.

[52] akamai, “Akamai's state of the internet q1 2017 report,” 31 March 2017. [Online]. Available: https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/q1-2017-state-of-the-internet-connectivity-report.pdf. [Accessed 06 Febuary 2020].

[53] F. Rumsey, “Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning and a Scene-Based Paradigm,” *Journal of the Audio Engineering Society,* vol. 50, no. 9, pp. 651-666, 2002.

[54] F. Rumsey, “Subjective Assessment of the Spatial Attributes of Reproduced Sound,” in *AES 15th International Conference*, Copenhagen, 1998.

[55] A. Colsman, L. Aspock, M. Kohnen and M. Vorlander, “Development of a questionnaire to investigate immersion of virtual acoustic environments,” in *DAGA 2016*, Aachen, 2016.

[56] AudioLab, University of York, “Sing From Your Seat,” AudioLab, University of York, [Online]. Available: https://audiolab.york.ac.uk/sing-from-your-seat/. [Accessed 01 August 2020].

[57] J. Rees-Jones and H. Daffern, “The Hills are Alive: Capturing and Presenting an Outdoor Choral Perfomance for Virtual Reality,” in *AES Conference on Immersive and Interactive Audio*, York, 2019.

[58] 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, 2018.

[59] S. D'Amario, H. Daffern and F. Bailes, “A new method of onset and offset detection in ensemble singing,” *Logopedics Phoniatrics Vocology,* vol. 44, no. 4, pp. 143-158, 2018.

[60] F. Rumsey, “Immersive Audio: Defining and evaluating the experience,” *Journal of the Audio Engineering Society,* vol. 68, no. 5, pp. 388-392, 2020.

[61] M. Slater and S. Wilbur, “A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments,” *Presence: Teleoperators and Virtual Environments,* vol. 6, no. 6, pp. 603-616, 1997.

[62] M. Lombard and T. Bolmarcich, “Measuring Presence: The Temple Presence Inventory,” in *ISPR 12th Anual International Workshop on Presence*, Los Angeles, 2009.

[63] B. G. Witmer and M. J. Singer, “Measuring Presence in Virtual Environments: A Presence Questionnaire,” *Presence: Teleoperators and Virtual Environments,* vol. 7, no. 3, pp. 225-240, 1998.

[64] M. Gospodarek, A. Fenovese, D. Dembeck, C. Brenner, A. Roginska and K. Perlin, “Sound design and reproduction techniques for co-located narrative VR experiences,” in *AES 147th Convention*, New York, 2019.

[65] K. Tcha-Tokey, E. Loup-Escande, O. Christmann and S. Richir, “A Questionnaire to Measure the User Experience in Immersive Virtual Environments,” in *Virtual Reality International Conference 2016*, Laval, 2016.

[66] M. Altman, K. Krauss, J. Susal and N. Tsingos, “Immersive Audio for VR,” in *AES Conference on Audio for Virtual and Augmented Reality*, Los Angeles, 2016.

[67] A. Harma, J. Jakka, M. Tikander and M. Karjalainen, “Augmented Reality Audio for Mobile and Wearable Appliances,” *Journal of the Audio Engineering Society,* vol. 52, no. 6, pp. 618 - 639, 2004.

[68] H. Lee, *A Conceptual Model of Immersive Experience in Extended Reality (Pre-Print),* PsyArXiv Preprints, 2021.

[69] 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, 2017.

[70] B. Yoon, H. Kim, G. A. Lee, M. Billinqhurst and W. Woo, “The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration,” in *IEEE Conference on Virtual Reality and 3d user interfaces*, Osaka, 2019.

[71] M. J. Schuemie, P. Van Der Straaten, M. Krijn and C. A. P. G. Van Der Mast, “Research in Presence in Virtual Reality: A Survey,” *CyberPsychology and Behaviour,* vol. 4, no. 2, pp. 183-201, 2001.

[72] J. Steuer, “Defining Virtual Reality: Dimensions Determining Telepresence,” *Journal of Communication,* vol. 42, no. 4, pp. 73-93, 1992.

[73] F. Rumsey, Spatial Audio, Oxford: Focal Press, 2001.

[74] W. Zhang, P. N. Samarasinghe, H. C. and T. D. Abhayapala, “Surround by Sound: A Review of Spatial Audio Recording and Reproduction,” *Applied Sciences,* vol. 7, no. 6, p. 532, 2017.

[75] M. Kleiner, B.-I. D. And and P. Svensson, “Auralization - An Overview,” *Journal of the Audio Engineering Society,* vol. 41, no. 11, pp. 861-875, 1993.

[76] L. Savioja and P. U. Svensson, “Overview of Geometrical Room Acoustic Modelling Techniques,” *The Journal of the Acoustical Society of America,* vol. 138, no. 2, pp. 708-730, 2015.

[77] J. J. Embrechts, “Review on the Application of Directional Impulse Responses in Room Acoustics,” in *Congrès français d'acoustique*, Le Mans, 2016.

[78] M. Vorlander, Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality, Springer, 2008.

[79] R. Hanninen, L. Savioja and T. Takala, “Virtual Concert Performance - Synthetic Animated Musicians Playing in an Acoustically Simulated Room,” in *International Computer Music Conference*, Hong Kong, 1996.

[80] L. Savioja, J. Huopaniemi, L. Tapio and R. Vannanen, “Creating Interactive Virtual Acoustic Environments,” *Journal of the Audio Engineering Society,* vol. 47, no. 9, pp. 675-705, 1999.

[81] J. Janer, E. Gomez, A. Martorell and M. Miron, “Immersive Orchestras: audio processing for orchestral music VR content,” in *8th International Conference on Games and Virtual Worlds for Serious Applications*, Barcelona, 2016.

[82] D. Fancourt and A. Steptoe, “Present in Body or Just in Mind: Differences in Social Presence and Emotion Regulation in Live vs Virtual Singing Experiences,” *Frontiers in Psychology,* vol. 10, 2019.

[83] S. Serafin, A. Adjorlu, L. Andersen and N. Andersen, “Singing in Virtual Reality with the Danish National children's choir,” in *International Symposium on Computer Muic Multisiciplinary Research*, Marseille, 2019.

[84] F. L. Martin, “Sound localization training and auditory adaption: a review,” in *AES 150th Convention*, Online, 2021.

[85] M. R. Schroeder, “Natural Sounding Artificial Reverb,” *Journal of the Audio Engineering Society,* vol. 10, no. 3, pp. 219-223, 1962.

[86] V. Valimaki, J. D. Parker, L. Sacioja, J. O. Smith and J. S. Abel, “Fifty Years of Artificial Reverb,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 5, pp. 1421-1447, 2012.

[87] C. Chafe, “Distributed Internet Reverberation for Audio Collaboration,” in *AES 24th International Conference: Multichannel Audio, the New Reality*, Banff, 2003.

[88] C. Chafe and J. Granzow, “Internet rooms from internet audio,” *Journal of the Acoustical Society of America,* vol. 133, no. 5, 2013.

[89] C. Chafe, “I am Streaming in a Room,” *Frontiers in Digital Humanities,* vol. 5, 2018.

[90] P. F. Driessen, T. E. Darcie and B. Pillay, “The Effect of Network Delay on Tempo in Musical Performance,” *Computer Music Journal,* vol. 35, no. 1, pp. 76-89, 2011.

[91] A. Olmos, M. Brule, N. Bouillot, M. Benovoy, J. Blum, H. Sun, N. W. Lund and J. R. Cooperstock, “Exploring the role of latency and orchestra placement on the networked performance of a distributed opera,” in *Twelth Annual International Workshop on Prescence*, Los Angeles, 2009.

[92] E. Chew, R. Zimmermann, A. A. Sawchuk, C. Kyriakakis, C. Papadopoulos, A. R. J. Francois, G. Kim, A. Rizzo and A. Volk, “Musical Interaction at a Distance: Distributed Immersive Performance,” in *Music Network 4th Open Workshop on Integration of Music in Multimedia Applications*, 2004.

[93] A. Carot and C. Werner, “Distributed Network Music Workshop with Soundjack,” in *Proceedings of the 25th Tonmeistertagung*, Leipzig, 2008.

[94] D. Akaoumianakis, C. Alexandraki, V. Alexiou, C. Anagnostopoulou, A. Eleftheriadis, V. Lailot, A. Mouchtaris, D. Pavlidi, G. C. Polyzos, P. Tsakalides, G. xylomenos and P. Zervas, “The MusiNet project: Towards unraveling the full portential of Network Music Performance Systems,” in *IEEE 5th International Conference on Information, Intelligence, Systems and Applications*, Chania, 2015.

[95] D. Akoumianakis, C. Alexandraki, V. Alexiou, C. Anagnostopoulou, A. Eleftheriadis, V. Lailioti, Y. Mastorakis, A. Modas, A. Mouchtaris, D. Pavlidi, G. C. Polyzos, P. Tsakalides, G. Xylomenos and P. Zervas, “The MusiNet project: Addressing the challenges in Network Music Performance Systems,” in *5th International Conference on Information, Intelligence, Systems and Applications*, Corfu, 2015.

[96] C. Alexandraki and D. Akoumanakis, “Exploring New Perspectives in Network Music Performance: The DIAMOUSES Framework,” *Somputer Music Journal,* vol. 34, no. 2, pp. 66-83, 2010.

[97] A. A. Sawchuk, E. Chew, R. Zimmerman, C. Papadopoulos and C. Kyriakakis, “From Remote Media Immersion to Distributed Immersive Performance,” in *ETTP 03: Proceedings of the 2004 ACM SIGMM workshop on experimental telepresence*, 2003.

[98] M. Iorwerth and D. Knox, “The application of Networked Music Performance to access ensemble activity for socially isolated musicians,” in *Web Audio Conference - Diversity in Web Audio*, Trondheim, 2019.

[99] A. Carot, K. Ulrich and G. Schuller, “Network Music Performance (NMP) in narrow band networks,” in *AES 120th Convention*, Paris, 2006.

[100] J.-P. Caceres and C. Chafe, “Jacktrip: Under the hood of an engine for network audio,” *Journal of New Music Research,* vol. 39, no. 3, pp. 183-187, 2010.

[101] V. Pulkki, “Spatial Sound Reproduction with Directional Audio Coding,” *Journal of the Audio Engineering Society,* vol. 55, no. 6, pp. 503-516, 2007.

[102] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min and A. Ngan, “A beam tracing method for interactive architectural acoustics,” *Journal of the Acoustical Society of America,* vol. 115, no. 2, pp. 739-756, 2004.

[103] A. Oliveira, G. Campos, P. Dia, D. Murphy, J. Viera, C. Mendonca and J. Santos, “Real-Time Dynamic Image Source Implementation for Auralisation,” in *16th International Conference on Digital Audio Effects*, Maynooth, 2013.

[104] M. Domanski, O. Stankiewiez, K. Wegner and T. Grajek, “Immersive Visual Media - MPEG-I: 360 video, virtual navigation and beyond,” in *International Conference on Systems, Signals and Image Processing (IWSSIP)*, Poznan, 2017.

[105] J. Kares and V. Larcher, “Streaming Immersive Audio Content,” in *AES Conference on AUdio for Virtual and Augmented Reality*, Los Angeles, 2016.

[106] G. Jacuzzi, S. Brazzola and J. Kares, “Approaching Immersive 3D Audio Broadcast Streams for Live Performances,” in *AES 142nd Convention*, Berlin, 2017.

[107] C. Eaton, *Quantifying Factors of Auditory Immersion for Virtual Reality. Masters thesis.,* Huddersfeild: University of Huddersfield, 2020.

[108] K. Young, G. Kearney and A. I. Tew, “Loudspeaker Postions with Sufficient Natural Channel Sperarution for Binaural Reproduction,” in *AES Conference on Spatial Reproduction*, Tokyo, 2018.

[109] J. Sodnik and S. Tomazic, Spatial Auditory Human Computer Interfaces, 1st ed., New York: Springer, 2015.

[110] M. Kraus, “The Effect of Spatial Audio on Immersion, Presence and Physiological Response in Games A Master's Thesis,” Aalborg University, Aalborg, 2015.

[111] P. Larsson, A. Valjamae, D. Vastfjall, A. Tajadura-Jimenez and M. Kleiner, “Auditory-induced presence in mixed reality environments and related technology,” in *The Engineering of Mixed Reality Systems*, London, Springer, 2009, pp. 143-163.

[112] A. J. Berkhout, D. de Vries and P. Vogel, “Acoustic control by wave field synthesis,” *Journal of the Acoustical Society of America,* vol. 93, no. 5, pp. 2764-2778, 1993.

[113] H. Moller, “Fundamentals of Binaural Technology,” *Applied Acoustics,* vol. 36, no. 3-4, pp. 171-218, 1992.

[114] M. A. Gerzon, “Periphony: With Height Sound Reproduction,” *Journal of the Audio Engineering Society,* vol. 21, no. 1, pp. 2-10, 1973.

[115] S. Spors, R. Rabenstein and J. Ahrens, “The Theory of Wave Field Synthesis Revisted,” in *AES 124th Convention*, Amsterdam, 2008.

[116] J. Daniel, R. Nicol and S. Moreau, “Further Investigations of High Order Ambisonic and Wavefield Synthesis for Holophonic Sound Imaging,” in *AES 114th Convention*, Amsterdam, 2003.

[117] L. L. Baranek and T. J. Mellow, Acoustic Sound Fields and Transducers, London: Academic Press, 2012.

[118] M. Kronlachner, “Spatial Transformations for the Alteration of Ambisonic Recording (Master's Thesis),” Insitute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Graz, 2014.

[119] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition), London: MIT Press, 1999.

[120] C. I. Cheng and G. H. Wakefield, “Introduction to Head Related Transfre Functions: Representation of HRTFs in Time, Frequency and Space,” *Journal of the Audio Engineering Society,* vol. 49, no. 4, pp. 231-249, 2001.

[121] P. F. Hoffmann and H. Moller, “Audiobility of Time Switching in Dynamic inaural Synthesis,” in *AES 118th Convention*, Barcelona, 2005.

[122] K. Hartung, J. Braasch and S. J. Sterbing, “Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions,” in *AES 16th International Conference: Spatial Sound Reproduction*, Rovaniemi, 1999.

[123] M. Noisternig, A. Sontacchi, T. Musil and R. Holdrich, “A 3D Ambisonic Based Binaural Sound Reproduction System,” in *AES 24th International Conference on Multichannel Audio*, Banff, 2003.

[124] J.-M. Jot, V. Larcher and J.-M. Pernaux, “A Comparitive Study of 3-D Audio Encoding and Rendering Techniques,” in *AES 16th International Conference*, Rovaniemi, 1999.

[125] L. L. Beranek and T. J. Mellow, Acoustic Sound Fields and Transducers, London: Academic Press, 2012.

[126] J. Ahrens, Analytic Methods of Sound Field Synthesis, Berlin: Springer, 2012.

[127] F. Zotter and M. Frank, Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality, Cham: Springer Open, 2019.

[128] C. Nachbar, F. Zotter, E. Deleflie and A. Sontacchi, “Ambix - A Suggested Ambisonic Format,” in *Ambisonics Symposium*, Lexington, 2011.

[129] D. Malham, “Space in Music - Music in Space: Masters Thesis,” Uniersity of York, York, 2003.

[130] J.-M. Batke, “The B-Format Microphone Revised,” in *Ambisonics Symposium*, Graz, 2009.

[131] T. T. McKenzie, “High Frequency Reproduction in Binaural Ambisonic Rendering PhD Thesis,” University of York, York, 2019.

[132] M. A. Poletti, “Three-Dimensional Surround Sound Systems Based on Spherical Harmonics,” *Journal of the Audio Engineering Society,* vol. 53, no. 11, pp. 1004-1025, 2005.

[133] D. B. Ward and A. T. D, “Reproduction of a Plane-Wave Sound Field Using an Array of Loudspeakers,” *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 6, pp. 697-707, 2001.

[134] F. Wefers, “Partitioned Convolution Algorithms for Real-Time Auralization: PHD Thesis,” Logos Verlag, Berlin, 2014.

[135] U. Zolzer, Digital Audio Effects, Chichester: John WIley and Sons, 2011.

[136] A. Krokstad, S. Dtrom and S. Sorsdal, “Calculating the Acoustical Room Response by Using the use of a Ray Tracing Technique,” *Journal of Sound Vibrations,* vol. 8, no. 1, pp. 118-125, 1968.

[137] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small room acoustics,” *Journal of the Acoustical Society of America,* vol. 65, no. 4, pp. 943-950, 1979.

[138] J.-M. Jot and A. Chainge, “Digital Delay Networks For Designing Artificial Revererators,” in *AES 90th Convention*, Paris, 1991.

[139] B. Alary, A. Politis, S. J. Schlecht and V. Valimaki, “Directional Feedback Delay Network,” *Journal of the Audio Engineering Society,* vol. 67, no. 10, pp. 752-762, 2019.

[140] E. D. Sena, H. Hacihabiboglu, Z. Cvetkovic and J. O. Smith, “Efficient Synthesis of Room Acoustics via Scattering Delay Networks,” *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 23, no. 9, 2015.

[141] B. Wiggins and M. Dring, “AmbiFreeverb 2 - Development of a 3D Ambisonic Reverb with Spatial Warping and Variable Scattering,” in *Conference on Sound Field Control*, Guilford, 2016.

[142] S. McGovern, “The Image-Source Reverberation Model in an N-Dimensional Space,” in *14th International Conference on Digital Audio Effects*, Paris, 2011.

[143] A. V. Oppenheim and R. W. Schfer, Discrete Time Signal Processing, London: Prentice-Hall, 1999.

[144] British Standard, “Acoustics - Measurement of Room Acoustic Parameters,” British Standard, 2009.

[145] G.-B. Stan, J.-J. Ebrechts and D. Archambeau, “Comparison of Different Impulse Response Measurement Techniques,” *Journal of the Audio Engineering Society,* vol. 50, no. 4, pp. 249-262, 2002.

[146] A. Farina, “Simultaneous Measurement of Impulse Response and Distortion with a SIne Swept technique,” in *AES 108th Convention*, Paris, 2000.

[147] J. Merimaa and V. Pulkki, “Spatial Impulse Response Rendering I: Analysis and Synthesis,” *Journal of the Audio Engineering Society,* vol. 53, no. 12, pp. 1115-1127, 2005.

[148] A. Perez-Lopez and J. De Munke, “Ambisonic Directional Room Impulse Response as a new Convention of the Spatially Oriented Format for Acoustics,” in *AES 144th Convewntion*, Milan, 2018.

[149] S. Shelly, D. Murphy and A. Chadwick, “B-Format Acoustic Impulse Response Measurement and Analysis in the Forest at Koli National Park, Finland,” in *16th International Conference on Digital Audio Effects*, Maynooth, 2013.

[150] F. Stevens and D. Murphy, “Spatial impulse response measurement in an urban environment,” in *AES 55th International Conference*, Helsinki, 2014.

[151] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Mathematics of Computation,* Vols. EC - 15, no. 4, pp. 297-301, 1965.

[152] S. Shelly and D. T. Murphy, “OpenAIR: An Interactive Auralisation Web Resource and Database,” in *AES 129th Convention*, San Francisco, 2010.

[153] R. Barger, S. Church, A. Fukunfa, J. Grunke, D. Keislar, B. Moses, B. Novak, B. Pennycook, Z. Settel, J. Strawn, P. Wiser and W. Woszczyk, “AES White Paper 1001: Networking Audio and Music Using Internet2 and Next-Generation Internet Capabilities,” *Journal of the Audio Engineering Society,* vol. 47, no. 4, pp. 300-302, 304-310, 1999.

[154] JANET, “30 Years of the Janet Network,” 01 April 2014. [Online]. Available: https://www.jisc.ac.uk/sites/default/files/janet-news-24-pull-out-april-2014.pdf. [Accessed 06 Febuary 2020].

[155] D. Prior, F. Reuben, I. Biscoe and M. Rofe, “Designing a system for Online Orchestra: Computer hardware and software,” *The Journal of Music, Technology and Education,* vol. 10, no. 2-3, pp. 185-196, 2017.

[156] cisco, “Cisco Annual Internet Report (2018-2023) White Paper,” Cisco, 2020.

[157] B. Grimes, “Network Topologies,” in *Networked AV Systems*, London, McGraw Hill Education, 2014, pp. 8-13.

[158] B. Grimes, “Network Architectures,” in *Networked AV Systems*, London, McGraw Hill Education, 2014, pp. 13-16.

[159] K. R. Fall and W. R. Stevens, “Client Server,” in *TCP/IP Illustrated Volume 1: The Protocols*, London, Addison-Wesley, 2011, pp. 20-21.

[160] J. Postel, “DOD Standard: Internet Protocol RFC 760,” Information Sciences Institute, University of Southern California, Marina del Rey, 1980.

[161] K. R. Fall and R. W. Stevens, “The Internet Protocol (IP),” in *TCP/IP Illustrated Volume 1: The Protocols*, London, Addison-Wesley, 2011, pp. 181-231.

[162] B. Grimes, “The OSI Model,” in *Networked AV Systems*, London, McGraw Hill Education, 2014, pp. 21-25.

[163] J. Postel, “Transmission Control Protocol: DARPA Internet Program Protocol Specification,” Information Sciences Institute, University of Southern California, Marina del Rey, 1981.

[164] A. Xu, W. Woszczyk, Z. Settel, B. Pennycook, R. Rowe, P. Glanter, J. Bary, G. Martin, J. Corey and J. R. Cooperstock, “Real-Time Streaming of Multichannel Audio Data over Internet,” *Journal of the Audio Engineering Society,* vol. 48, no. 7-8, pp. 627-641, 2000.

[165] S. Pejhan, M. Schwartz and D. Anastassiou, “Error Control Using Retransmission Schemes in Multicast Transport Prtotocols for Real-Time Media,” *IEEE/ACM Transactions on Networking,* vol. 4, no. 3, 1996.

[166] N. Pekez and A. K. J. Popovic, “Performance analysis on TCP/IP Audio Streaming in Point-to-Point communication,” in *2019 Zooming Innovation in Consumer Technologies Conference (ZINC)*, Novi Sad, 2019.

[167] K. R. Fall and W. R. Stevens, “TCP: The Transmission Control Protocol (Preliminaries),” in *TCP/IP Illustrated Volume 1: The Protocols*, London, Addison-Wesley, 2011, pp. 579- 593.

[168] J. Postel, “User Datagram Protocol RFC 768,” Information Sciences Institute, Marina del Rey, 1980.

[169] K. R. Fall and W. R. Stevens, “User Datagram Protocol (USP) and IP Fragmentation,” in *TCP/IP Illustrated Volume 1: The Protocols*, London, Addison-Wesley, 2011, pp. 473-510.

[170] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson, “RTP: A Transport Protocol for Real-Time Applications,” The Internet Society, 2003.

[171] J.-C. Bolot, S. Fosse-Parisis and D. Tonsley, “Adaptive FEC-Based Error Control for Internet Telephony,” in *IEEE INFOCOM 99 Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now*, New York, 1999.

[172] D. E. Comer, “Jitter and Playback Delay,” in *Internetworking with TCP/IP Volume 1: Principles, Protocols and Architecture*, London, Prentice-Hall International (UK) Limited, 2000, pp. 541-542.

[173] M. Lutzky, G. Schuller, M. Gayer, U. Kramer and S. Wabnik, “A guidline to audio codec delay,” in *AES 116th convention*, Berlin, 2004.

[174] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Gieger, V. Ruoppila, P. Ekstrand, M. Lutzky and B. Grill, “MPEG- 4 Enhanced Low Delay AAC - a new standard for high quality communication,” in *125th AES Convention*, San Francisco, 2008.

[175] U. Kramer, J. Hirschfeld, G. Schuller, S. Wabnik, A. Carot and C. Werner, “Network Music Performance with Ultra-Low Delay Audio Coding under Unreliable Network Conditions,” in *AES 123rd Convention*, New York, 2007.

[176] U. Kramer, G. Schuller, S. Wabnik, J. Klier and J. Hirschfeld, “Ultra Low Delay audio coding with constant bitrate,” in *AES 117th Convention*, San Francisco, 2004.

[177] K. Vos, S. Jensen and K. Soerensen, “SILK Speech Codec,” Skype Technologies S. A., 2010.

[178] J.-M. Valin, T. B. Terriberry, C. Montgomery and G. Maxwell, “A High-Quality Speech and Audio Codec With Less Than 10-ms Delay,” *IEEE Transactions on AUdio Speech and Language Processing,* vol. 18, no. 1, pp. 58-67, 2010.

[179] J.-M. Valin, G. Maxwell, T. B. Terriberry and K. Vos, “High-Quality, Low-Delay Music Coding in the Opus Codec,” in *AES 135th COnvention*, New York, 2013.

[180] J. Valin, K. Vos and T. Terriberry, “Definition of the Opus Codec rfc6716,” Internet Engineering Task Force, 2012.

[181] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, J. Skoglund, D. T. Murphy and G. Kearney, “Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes,” *Journal of Applied Sciences,* vol. 9, no. 13, p. 2618, 2019.

[182] HTC, “VIVE Specs and User Guide,” 2020. [Online]. Available: https://developer.vive.com/resources/vive-sense/hardware-guide/vive-specs-user-guide/. [Accessed 2 November 2020].

[183] Oculus, “Oculus Devices and Audio Capabilities,” Oculus, [Online]. Available: https://developer.oculus.com/learn/audio-hardware/?locale=en_GB. [Accessed 2 November 2020].

[184] Oculus, “Occulus Rift and Rift S Minimum Requirements and System Specifications,” Oculus, [Online]. Available: https://support.oculus.com/248749509016567/. [Accessed 2 November 2020].

[185] HTC, “What are the system requirements?,” HTC, [Online]. Available: https://www.vive.com/us/support/vive/category_howto/what-are-the-system-requirements.html. [Accessed 2 November 2020].

[186] C. Schneiderwind, A. Neidhardt and D. Meyer, “Comparing the effect of different open headphone models on the perception of real sound sources,” in *Audio Engineering Society 150th Convention*, Online, 2021.

[187] B. Boren and M. Geronazzo, “Comparison of Distortion Products in Headphone Equalization Algorithms for Binaural Synthesis,” in *Audio Engineering Society 150th Convention*, Online, 2021.

[188] P. Ferguson, C. Chafe and S. Gapp, “Trans-Europe Express Audio: testing 1000 mile low-latency uncompressed audio between Edinburgh and Berlin using GPS-derived word clock, first with jacktrip and then with Dante,” in *AES 148th Convention*, Vienna, 2020.

[189] Audio Engineering Society, Inc, “AES standard for Audio applications of networks - High-performance streaming audio-over-IP interoperability,” Audio Engineering Society, New York, 2018.

[190] Focusrite, “Scarlett 18i20 User Guide,” 2013. [Online]. Available: https://customer.focusrite.com/en/support/downloads. [Accessed 02 November 2020].

[191] Focusrite, “Focusrite Control Scarlett 3rd Gen - User Guide,” 2019. [Online]. Available: https://customer.focusrite.com/en/support/downloads. [Accessed 02 November 2020].

[192] A. Knoth, F. Coelho, N. Arnaudov and S. Letz, “Jack Audio Connection Kit,” Jack Audio, 2020. [Online]. Available: https://github.com/jackaudio. [Accessed 11 November 2020].

[193] “Jack Audio Connection Kit,” JACK, [Online]. Available: https://jackaudio.org/. [Accessed 11 November 2020].

[194] Cockos, “Up and Running: A REAPER User Guide v 6.15,” October 2020. [Online]. Available: https://www.reaper.fm/userguide.php. [Accessed 02 November 2020].

[195] M. Kronlachner, “Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics,” University of Music and Performing Arts, Graz, Graz, 2014.

[196] M. Kronlachner, “Ambisonic plug-insuite for production and performance usage,” in *Linux Audio Conference*, Graz, 2013.

[197] M. Rumori, “Girafe - A versatile Ambisonics and Binaural System,” in *Ambisonics Symposium*, Graz, 2009.

[198] T. Rudzki, “nvsonic 3DOF Head Tracker,” trsonic, 2019. [Online]. Available: https://github.com/trsonic/nvsonic-head-tracker. [Accessed 26 July 2020].

[199] 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,” *Applied Sciences,* vol. 8, no. 11, 2018.

[200] Neumann, “Operating Instructions KU 100,” June 2020. [Online]. Available: https://en-de.neumann.com/file-finder. [Accessed 04 November 2020].

[201] Resonance Audio, “Resonance Audio,” Resonance Audio, [Online]. Available: https://resonance-audio.github.io/resonance-audio/. [Accessed 12 November 2020].

[202] Valve, “Steam Audio,” Valve, [Online]. Available: https://valvesoftware.github.io/steam-audio/. [Accessed 12 November 2020].

[203] M. Iorwerth and D. Knox, “Playing Together Apart: Musician's Experiences of Physical Separation in a Classical Recording Session,” *Music Perception,* vol. 36, no. 3, pp. 289-299, 2019.

[204] International Telecommunications Union, “Method for the subjective assesment of intermediate quality level of audio systems,” October 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534/en. [Accessed 12 November 2020].

[205] S. Djordjevic, H. Hacihabiboglu, Z. Cvetkovic and E. D. Sena, “Evaluation of the Percieved Naruralness of Artificial Reverberation Algorithms,” in *AES 148th Convention*, Vienna, 2020.

[206] M. Iorwerth, “Playing together, apart: An exploration of the challenges of Network Music Performance in informal contexts PhD Thesis,” Glasgow Caledonian University, Glasgow, 2019.

[207] T. Rudzki, D. Murphy and G. Kearney, “On the Measurement of Perceived Lateral Angle Using Eye Tracking,” in *AES International Conference on Audio for Virtual and Augmented Reality*, Online, 2020.

[208] W. Goebl and C. Palmer, “Synchronization of Timing and Motion Among Performing Musicians,” *Music Perception,* vol. 23, no. 5, pp. 427-438, 2009.

[209] P. E. Keller, “Musical Ensemble Synchronisation,” in *International Conference on Music Communication Science*, Sydney, 2007.

[210] J. W. Gordon, “The Perceptual Attack Times of Musical Tones,” *Journal of the Acoustical Society of America,* vol. 82, no. 1, pp. 88 - 105, 1987.

[211] J. Vos and R. Rasch, “The Perceptual onset of Musical Tones,” *Perception and Psychophysics,* vol. 29, no. 4, pp. 323-335, 1981.

[212] A. Barbosa and J. Cordeiro, “The Influence of Perceptual Attack Times in Networked Music Performance,” in *AES 44th International Conference*, San Diego, 2011.

[213] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler, “A Tutorial on Onset Detection in Music Signals,” *IEEE Transactions on Speech and Audio Processing,* vol. 13, no. 5, pp. 1035-1047, 2005.

[214] N. Collins, “A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions,” in *AES 118th Convention*, Barcelona, 2005.

[215] R. Zhou and J. D. Reiss, “Music Onset Detection,” in *Machine Audition: Principles, Alforithms and Systems*, Hershey, Information Science Reference, 2010, pp. 297-316.

[216] C. Rosao, R. Ribeiro and D. M. de Matos, “Comparing Onset Detection Methods Based on Spectral Features,” in *OSDOC 2012: Workshop Open Source and Design of Communication*, Lisbon, 2014.

[217] C. Duxbury, M. Davies and M. Sandler, “Seperation of Transient Information in Musical Audio Using Multiresolution Analysis Techniques,” in *COST G-6 Conference on Digital Audio Effects*, Limerick, 2001.

[218] N. Coollins, “Using a Pitch Detector for Onset Detection,” in *6th International Conference on Music Information Retreival*, London, 2005.

[219] J. B. Allen, “Short Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform,” *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vols. ASSP-25, no. 3, pp. 235 - 238, 1977.

[220] L. D, “The Discrete Fourier Transform, Part 2: Radix 2 FFT,” *Object Technology,* vol. 8, no. 5, pp. 21-33, 2009.

[221] M. Goto and Y. Muraoka, “Beat Tracking based on Multiple-agent Architecture - A Real-Time Beat Tracking System for Audio Signals,” in *Proceedings of the Second International Conference on Multiagent Systems*, Kyoto, 1996.

[222] A. Klapuri, “Sound Onset Detection by Applying Psychoacoustic Knowledge,” in *IEEE International Conference on Acoustics Speech and Signal Processing*, Phoenix, 1999.

[223] C. Duxbury, M. Sandler and M. Davies, “A Hybrid Approach to Musical Note Onset Detection,” in *5th International Conference on Digital Audio FX*, Hamburg, 2002.

[224] J. Glover, V. Lazzarini and J. Timoney, “Real-time detection of musical onsets with linear prediction and sinusoidal modeling,” *EURASIP Journal on Advances in Signal Processing,* vol. 1, no. 68, pp. 1-13, 2011.

[225] J. Makhoul, “Linear Prediction: A Tutorial Review,” *Proceedings of the IEEE,* vol. 63, no. 4, pp. 561 - 580, 1975.

[226] X. Amatriain, J. Bonada, A. Loscos and X. Serra, “Spectral Processing,” in *DAFX: Digital Audio Effects*, Chichester, John Wiley and Sons, 2002, pp. 373-438.

[227] C. M. T. Rosao, “Onset Detection in Music Signals Masters Thesis,” ISCTE Instituto Universitario de Lisoa, Lisbon, 2012.

[228] X. Rodet and F. Jaillet, “Detection and Modelling of Fast Attack Transients,” in *Proceedings of the International Computer Music Conference*, Havana, 2001.

[229] P. Masri and A. Bateman, “IMproved Modelling of Attack Transients in Music Analysis - Resynthesis,” in *Proceedings of the International Computer Music Conference*, Hong Kong, 1996.

[230] S. Dixon, “Onset Detection Revisited,” in *Proceedings of the th Interernational Conference on Digital Audio Effects*, Montreal, 2006.

[231] F. Eyben, S. Bock, B. Schuller and A. Graves, “Universal Onset Detection with Bideirectional Long Short-Term Memory Neural Networks,” in *OSDOC 2012: Workshop Open Source and Design of Communication*, Lisbon, 2010.

[232] M. Dolson, “The Phase Vocoder: A Tutorial,” *Computer Music Journal,* vol. 10, no. 4, pp. 14-27, 1986.

[233] C. Duxbery, J. P. Bello, M. Davies and M. Sandler, “A Combined Phase and Amplitude Based Approach to Onset Detection for Audio Segmentation,” in *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services*, Singapore, 2003.

[234] A. Holzapfel, Y. Stylianou, A. C. Gedik and B. Bozkurt, “Three Dimensions of Pitched Instrument Onset Detection,” *IEEE Transactions on Audio Speech and Language Processing,* vol. 18, no. 6, pp. 1517-1527, 2010.

[235] B. S. Atal, “Automatic Speaker Recognition Based on Pitch Contours,” *The Journal of the Acoustical Society of America,* vol. 52, no. 6b, pp. 1687-1697, 1972.

[236] M. A. Noll, “Cepstrum Pitch Determination,” *The Journal of the Acoustical Society of America,* vol. 31, no. 2, pp. 293-309, 1967.

[237] D. J. Hermes, “Measurement of Pitch by Subharmonic Summation,” *The Journal of the Acoustical Society of America,* vol. 83, no. 1, pp. 257-264, 1988.

[238] T. Drugman and A. Abeer, “Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Florence, 2011.

[239] A. v. Brandt, “Detecting and estimating parameter jumpes using ladder algorithms and likelihood ratio test,” in *International Conference on Acoustics Speech and Signal Processing*, Boston, 1983.

[240] S. A. Abdallah and M. D. Plumbley, “Probability as metadata:event detection in music using ICA as a conditional density model,” in *4th International Symposium on Independent Component Analysis and Blind Signal Seperation*, Nara, 2003.

[241] J. Schluter and S. Boeck, “Improved Musical Onset Detection with Convolutional Neural Networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, 2014.

[242] B. Stasiak and J. Monko, “Analysis of time-frequency representations for musical onset detection with convolutional neural networks,” in *Federated Conference on Computer Science and Information Systems*, Gdansk, 2016.

[243] I. Ali-MacLachlan, C. Southall, M. Tomczak and J. Hockman, “Improved Onset Detection for Traditional Irish Flute Recordings Using Convolutional Neural Networks,” in *7th International Workshop on Folk Music Analysis*, Malaga, 2017.

[244] Mathworks, "MATLAB," Mathworks, [Online]. Available: https://uk.mathworks.com/products/matlab.html?s_tid=hp_ff_p_matlab. [Accessed 20 December 2018].

[245] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry,* vol. 36, no. 8, pp. 1627-1639, 1964.

[246] C. S. Oh, J. N. Bailenson and G. F. Welch, "A Systematic Review of Social Presence: Definition, Antecedents, and Implications," *Frontiers in Robotics and AI,* vol. 5, 2018.

[247] G. N. Yannakakis and M. Maragoudakis, "Preference Learning for Cognitive Modeling: A Case Study on Entertainment Preferences," *IEEE Transactions on systems, man and cybernetics - Part A: Systems and Humans,* vol. 39, no. 6, pp. 1165-1175, 2009.

[248] University of York, Audiolab, "Open Air Library," University of York, Audiolab, [Online]. Available: https://openairlib.net/. [Accessed 26 July 2020].

[249] Genelec, "Genelec 8040B User Manual," 2015. [Online]. Available: https://assets.ctfassets.net/4zjnzn055a4v/67DMY6rUTSKsIwIM2iEWSw/ab188eec8e0e536eeae580f2968fd576/8040b_8050b_en_fi_opman_e.pdf. [Accessed 15 12 2020].

[250] Soundfield, "Soundfield ST350 Portable Microphone System User Guide v1.02," [Online]. Available: http://www.thesoundmanifesto.co.uk/Soundfield_ST350_man.pdf. [Accessed 15 December 2020].

[251] Genelec, "Genelec S30D Digital Monitoring System," 2001. [Online]. Available: https://assets.ctfassets.net/4zjnzn055a4v/7pY7N5HOCIcWwOaMGq4QI/4951df0c0b231b801d911661a899fafb/S30D_opman.pdf. [Accessed 20 December 2020].

[252] Soundfield, "Soundfield SPS422B User Guide," [Online]. Available: http://cdn.soundfield.com/assets/downloads/manual/SPS422B-manual.pdf. [Accessed 20 December 2020].

[253] Beyerdynamic, "DT 990 Pro Spec Sheet and Manual," [Online]. Available: https://europe.beyerdynamic.com/dt-990-pro.html. [Accessed 20 6 2021].

[254] Beyerdynamic, "DT 770 Pro," [Online]. Available: https://europe.beyerdynamic.com/dt-770-pro.html. [Accessed 20 06 2021].

[255] Sennheiser, "HD25 On Ear DJ Headphone," Sennheiser, [Online]. Available: https://en-uk.sennheiser.com/on-ear-dj-headphone-hd25. [Accessed 20 06 2021].

[256] M. R. Harwell, E. N. Rubinstein, W. S. Hayes and C. C. Olds, "Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effect ANOVA Cases," *Journal of Educational Statistics,* vol. 17, no. 4, pp. 315-339, 1992.

[257] Digital Stage, "Digital Stage," Digital Stage, [Online]. Available: https://digital-stage.org/?lang=en. [Accessed 20 January 2021].

[258] A. Carot, F. Sardis, M. Dohler, S. Saunders, N. Uniyal and R. Cornock, "Creation of a hyper-realistic remote music session with professional musicians and public audiences using 5G commodity hardware," in *IEEE International Conference on Multimedia and Expo Workshops*, London (virtual), 2020.

[259] G. Grimm, "digital-stage/ov client/src/ov_renderer_tascar.cc," 12 January 2021. [Online]. Available: https://github.com/digital-stage/ov-client/blob/master/src/ov_render_tascar.cc. [Accessed 22 January 2021].

[260] G. Grimm, J. Luberadzka, T. Herzke and V. Hohmann, "Toolbox for acoustic scene creation and rendering (TASCAR): Render methods and research applications," in *Linux Audio Conference*, Mainz, 2015.

[261] A. Carot, 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. 47925-47951, 2020.

[262] B. Alary and A. Politis, "Frequency-Dependant Directional Feedback Delay Network," in *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, 2020.