Sponsoring Committee: Professor Agnieszka Roginska, Chairperson
Professor Morwaread Farbood
Doctor Jean-Marc Jot

ACOUSTICS AND COPRESENCE: TOWARDS EFFECTIVE AUDITORY VIRTUAL ENVIRONMENTS FOR DISTRIBUTED MUSIC PERFORMANCES

Andrea Felice Genovese

Program in Music Technology
Department of Music and Performing Arts Professions

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in the
Steinhardt School of Culture, Education, and Human Development
New York University
2023

ABSTRACT

The work addressed in this doctoral thesis concerns the use of *immersive audio* techniques within the application of *Networked Music Performances* (NMP). The core question postulated asks whether immersive audio technology improves the quality of a distributed network performance. In its larger scope, this research aims to guide the drafting of multidisciplinary study methodologies that adequately consider the multitude of aspects that come into play when determining whether an immersive system is beneficial in a given context. The illustrated investigation aims to shine a light on the relationship between *Immersive Quality* and other quality metrics within NMPs, making a case for the introduction of interaction-design strategies inspired by Virtual- and Mixed- reality applications.

The first part of this work is concerned with presenting the implementation of novel proof-of-concept applications that illustrate the merging of real-time distributed music performance networks with Virtual- and Augmented reality environments. A critical analysis of the work incurred in literature and by the author, during the doctoral program, presents the challenges faced by audio engineers working to implement immersive experiences for audiences and musicians.

The second, and larger part, of the dissertation, gives a central focus to a new empirical study designed to provide insight into the roles of latency and auralization, and their interactions, in eliciting the psychological construct of *Auditory Copresence*, intended as an extension of social telepresence, and explored in its relation towards subjective or objective measures of the quality of experience. In more detail, the study consists of a simulation of a music performance over the Internet using asymmetric node locations, in which remotely placed musicians are digitally connected between rooms that are different in acoustic character while being presented with different rendering strategies of virtual acoustic environments. Several layers of evaluation are investigated by collecting and analyzing data from in-session questionnaires, listening evaluation tests, and digital signal analysis.

Results derived from mixed-effects regression models show that latency was the biggest factor in degrading quality metrics across all observed layers, while the auralization strategies were partially identified as positive contributors to subjective dimensions of evaluation, including “copresence”, but not toward objective metrics relating to tempo stability and beat

synchronization. Secondary effects related to "learning effects", or "time within a system", and individual biases were also found to significantly contribute to explaining variability in the response. A correlation analysis among the response layers suggests that different dimensions of evaluation are not correlated, implying that improving the "immersive experience" in NMPs does not necessarily translate to improvements in the musical outcome.

When put into the context of future VR/AR NMP applications for traditional music performance, the results indicate that the engineering cost of creating an immersive experience for musicians may not always be a worthwhile contributor to the audience experience as the introduction of auralization and spatialization methods did not improve the objective musical outcome of performances. However, the subjective experience of performers can improve to a significant degree, provided that latency is kept within the established thresholds for time-critical music performance. The effects of auralization strategies on the musical outcome were found to be room-dependent, prompting further discussion on the interactions between the local acoustic character of a listening room and the acoustic character of an auralized virtual environment. Future work is directed toward a more parameterized investigation of artificial reverb and its interaction with local acoustics, dynamic 3DoF auditory displays, and multimodal immersive displays involving virtual or mixed-reality headsets.

The findings provided by this work have important validation and discovery value for connecting the factors affecting perceptual experience to the technical limitations of distributed systems in regard to signal latency and immersive displays. Furthermore, for the larger community, the data will be made available through a novel public dataset of distributed music performances. The gained insights can contribute to the larger conversation about the design of NMPs and help with the management of specific system-dependent "latency budget" according to the priorities set by applications.

“There is geometry in the humming of the strings.
There is music in the spacing of the spheres. ”
— Pythagoras

ACKNOWLEDGEMENTS

My doctoral journey would not have been possible without the incredible people whose efforts have been fundamental in enabling mine.

First of all, my deepest gratitude goes to my advisor, Prof. Agnieszka Roginska, who welcomed me and guided me throughout this doctoral adventure, helping me navigate the various aspects of academia with wisdom and fortitude. This gratitude also goes to the professors who provided key advice on my academic progress at various stages. Prof. Morwaread Farbood and Prof. Pablo Ripolles for their patient help with statistic matters, Prof. Brian McFee for help the librosa tools, Prof. Juan Bello for broadening my interests and helping me figure out the vision of our MARL laboratory, and Prof. Ken Perlin for showing me novel abstract uses of technology that are of great inspiration to my current work. To my dissertation committee and readers who provided me with exceptional feedback and supported me all the way, Dr. Jean Marc Jot, Prof. Robert Rowe, and Prof. Braxton Boren. I also thank all of my course teachers at NYU who broadened my knowledge from classical music to machine learning; I appreciate the incredible breadth of these experiences and the opportunity to understand how to train an AI system to simulate Beethoven.

This thesis would not have been feasible without the amazing support of the technical and administrative staff of the *Music and Performing Arts Professions* department and *NYU Steinhardt*, who patiently helped me through my equipment and facility booking requests, taught me how to operate the building connection hardware, and worked with me throughout various challenging stages of experimental setups. My warmest thanks go to Tom Beyer, Ernesto Valenzuela, Michael Oikonomidis, Parichat Songmuang, Drew Francis, Luis Mercado and all the Dolan engineers, Patio and Facilities staff who interacted with me. They all deserve medals for their incredible work and for enduring my endless hassling.

The next mention goes to my colleagues and collaborators who made this journey special for me. To Dr. Rachel Bittner, Dr. Michael Musick, Dr. Finn Upham, Dr. Charlie Mydlarz, and Dr. Claire Pelofi for mentoring me early in this journey and giving me the warmest welcome to doctoral life in New York. To my Ph.D. fellow warriors Marta Gospodarek, Andrew Telichan, and Jong-wook Kim, who have been through the trials of candidacy with me. Thanks to everyone else in the laboratory with whom I have interacted, Willie, Ana Elisa, Yu, Mark, Magdalena, and the whole list of Ph.D. students and Postdoc fellows. A warm recognition goes to my research collaborators and coauthors: Dr. Robert Pahle, Robert Hupke, Sripathi Sridar, Scott Murakami, Juan Simon Calle, Makan Taghavi, Gregory Reardon, and everyone else who participated in the 3D audio research group and the fantastic experience that was the Immersive Audio Group. Thanks to everyone who sat in the 6^{th} floor office with me and everyone who joined me for academic beers (it is a long list, but you are all remembered fondly).

There are many other people who deserve to be mentioned and I apologize if I forgot anyone. I thank all my students at NYU, the participants who bared with me throughout lengthy experiment

sessions, the whole of the Music Technology faculty and staff, all the MARL/CUSP/CLaME Ph.D. colleagues, postdoctoral researchers, and external collaborators. Also, thanks to my managers and colleagues at THX, Qualcomm, and Microsoft Research; Patrick Flanagan, Hannes Gamper, and Andre Schevciw, and everyone who patiently sat down with me to teach me about their work. I am very grateful for your daily contribution to my personal intellectual growth.

Like every researcher in the history of time, this work stands on the shoulders of giants. I want to express my respect and recognition to the research communities and organizations in audiology, audio engineering, music technology, and all the other fields that make it possible to study this fascinating multidisciplinary branch of human knowledge. You inspire me to work for the benefits of the arts and sciences.

Finally, to my family, my girlfriend Alessandra, my roommates family, and all my close friends - within and outside of the US. You gave me the motivation, energy, and support to keep crunching at the most difficult moments. Thank you for believing in me during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Glossary of Terms & Acronyms

Research Areas

- **Immersive Audio**: A multidisplinary branch of auditory sciences involving the study and production of audio content, technology and experiences capable of eliciting realistic "auditory immersion". The field mainly combines audio engineering and acoustics with psychoacoustics. Immersive audio is commonly used in applications such as virtual reality, gaming, and live performance to create a more immersive and engaging experience for the listener.
- **Psychoacoustics:** Interdisciplinary branch of cognitive psychology concerned with auditory perception and its physiological effects.
- **NMP:** Networked Music Performances. An internet-based collaborative system between two or more connected performers physically distant from each other. Also known in literature as *Distributed Performance Networks*.
- **Music Cognition:** The branch of psychology which investigates the understanding of the perception of musical qualities.
- **MIR:** Music Information Retrieval. Branch of signal processing that deals with the extraction of musical parameters and information from digital audio signals.

Acoustics

- **Sound-field:** a region in a material medium in which sound waves are propagating.
- **Acoustic free-field:** A situation in which acoustic reflections do not occur.
- **RIR:** Room-Impulse-Response. Response of an acoustic receiver to an impulse sound source, used to characterize the acoustic properties of a space as sound is reflected by hard boundaries. It can be measured or modeled.
- **HRIRs:** Head-Related-Impulse-Responses, a stereo acoustic filter which describes the path of a sound source to the human ears from a defined location.

- ◇ **BRIRs:** Binaural room impulse responses. HRIRs measured in the presence of reverberation.
- ◇ **ISM:** Image Source Method. A method used to model impulse responses using geometric, shoebox, and path calculations for sound reflections from a source to a receiver (Dance and Shield 1997).
- ◇ **Plausibility:** A perceptual measure of the extent to which an auditory simulation is in agreement with the listener's expectation of a corresponding real event. (Lindau and Weinzierl 2012).
- ◇ **SPL:** Sound Pressure Level.
- ◇ **Reverberation Time, T_{60}:** Time required for an impulse sound inside a room to decay by 60 dB.

Immersive Systems

- ◇ ***Presence*:** The feeling of "being there" (Heeter 1992). The perceptual illusion of non-mediation (Lombard and Ditton 1997).
- ◇ ***copresence*:** The feeling of being together in a shared space (Riva et al. 2003)
- ◇ ***Telepresence*:** The extent to which one feels present in the mediated environment, rather than in the immediate physical environment (Steuer 1992).
- ◇ **VR:** Virtual Reality. The computer-generated simulation of a three-dimensional image or environment that can be interacted with in a seemingly real or physical way by a person.
- ◇ **AR:** Augmented Reality. The general introduction of digital information about the real world around a user though a technological device.
- ◇ **MR:** Mixed Reality. The blending of digital elements with the real local environment using dedicated tracking technology. The rendering is locally adaptive and respectful of the physical boundaries of the space (Milgram and Kishino 1994).

- **XR:** eXtended Reality. Umbrella term to indicate the common fields and technologies of Virtual, Mixed and Augmented reality.
- **Ecological approach:** The act of testing and observing the effect of a technology in its intended use case scenario, as opposed to a controlled experiment.
- **HMD:** Head-Mounted Display. Wearable device provided with an occlusive (VR) or transparent (MR) stereoscopic screen provided with sensors.
- **3DOF:** 3-Degrees of Freedom. Attribute of technology capable of sensing and reacting to 3-dimensional rotation of a device using gyroscopes. For example, a 3DOF XR device can respond to a user moving or rotating the head (yaw, pitch and roll).
- **6DOF:** 6-Degrees of Freedom. Expansion on 3DOF by adding sensors capable of tracking the XYZ position of a device inside a space. A 6DOF XR system can respond to a user walking within a (usually delimited) space.

Distributed Networks

- **NMP:** Network Music Performance
- **LAN:** Local Area Network
- **WAN:** Wide Area Network
- **DAW:** Digital Audio Workstation, software environment to record and process digital audio signals.
- **UDP:** User Datagram Protocol. A transmission protocol used for real-time information, lenient to packet losses and unreliable connections.
- **VST:** Virtual Studio Technology, DAW processing plugin software.

Network Music Performance

- **BPM:** Beats Per Minute. Measure of tempo.
- **IBI:** Inter-beat interval. Range in seconds between the onset of a quarter beat and the next.

- **RIA:** Realistic Interaction Approach. The act of making music in a distributed network as if in the same room, that is, without applying particular musical compensation strategies (Carôt and Werner 2009).

Statistics

- **LMM:** Linear Mixed Model. Prediction models from the regression family that can account for "random factors" by introducing a random intercept or slope in the model fitting terms. Useful to account for response variability tied to subjective biases.
- **GLMM:** Generalized Linear Mixed Model. An expansion of linear mixed model capable to account for nonnormally distributed predictions (e.g. binomial variables or Poisson distributed variables).
- **PCA:** Principal Component Analysis. A statistical method used to transform correlated variables into principal orthogonal components, maximizing variance and facilitating the construction of a predictive model (Smith 2002).
- **ANOVA:** Analysis of variance. A statistical model used to detect significant differences among the means of group distributions and possible interactions between independent variables in the recorded responses. There are several variations of the ANOVA test.
- **JND:** Just Noticeable Differences. Quantifiable perceptual threshold of noticeability between two closely related events.

CHAPTER I

INTRODUCTION

The advent of immersive Virtual- and Augmented-reality technologies in the consumer market has opened new frontiers of digital applications to the public. *Immersive Technologies* are progressively being adopted in several dimensions of human-skilled practice, spanning several fields of science, medicine, engineering, and arts. In this landscape, academia plays a crucial role in advancing the field by researching new, future-oriented ideas and experimenting with novel proof-of-concept experiences, providing useful data that feeds back to the technical engineering of new tools. By taking on this role, academic researchers can delve deeply into studying the effects of this technology on target populations and unravel the technical and creative challenges that arise across diverse professional uses. Although the road is still long for the widespread adoption of extended-reality technology in daily life tasks, new potential horizons are opening as the technology develops, changing the nature of future digital interaction.

Within the realm of music, a lot of work is being done with the intent of merging immersive telecommunication with distributed music networks, in order to provide a socially immersive, plausible, and collaborative digital virtual environment for the creation of music. Immersive technology, such as mixed and virtual reality, can be used to create digital social interactive environments and new dimensions of musical collaborations. Their usage in distributed music networks has the potential of augmenting the experience towards a more "cohesive" or "realistic" interaction that brings the activity of making music over an internet-based network closer to that of a real-life traditional interaction.

Distributed music environments have been studied in academia for years, enabling musicians and researchers to connect, rehearse, and perform remotely in "real time", using internet-based audio-visual exchanges. Nevertheless, the combination of modern immersive technologies with distributed music performances over the internet is a relatively new and unexplored approach that is rapidly gathering interest. The renewed interest in the topic and

the technological progresses call for renewed research, capable of approaching the field through new artistic and scientific lenses. Although mixed-reality tools are still under development and not widely available, prototype systems can be used to study the requirements that need to be met by the industry for the development of high-quality collaborative experiences. The strong current demand for augmenting artistic expressivity, and the need to adapt artistic collaboration to today's use of remote-presence technology, has accelerated the development of new paradigms for interactive music enjoyment and production, whether offline or in real-time, whether in person or remotely.

One of the most promising aspects of immersive technology is the ability to create new forms of distributed musical collaborations and interactions that were previously only available in laboratory conditions. As this technology continues to mature, it is likely that we will see new and innovative applications emerge that will fundamentally transform the way that we think about music and the creative collaborative process. Already, we see a first generation of immersive music performance applications being published, including remote VR concerts (Charron 2017), augmented musical practice and education (Shahab et al. 2022; Baratè et al. 2019), virtual recording sessions and rehearsals (Cairns et al. 2022), interactive collaborative music experiences (Schlagowski et al. 2022), creative interfaces, and more (Loveridge 2020). Within this context, researchers are investigating how immersive technologies can be used to enhance the emotional and sensory experience of music, while also exploring how they can be integrated with traditional musical instruments and techniques for an effective performance outcome. Additionally, there is growing interest in studying the social and cultural implications of immersive musical experiences, such as how they may impact audience engagement, music education, and the distribution of musical content. As the field continues to evolve, it will be important for researchers, musicians, and industry professionals to work together to ensure that these technologies are used in ways that enhance, rather than diminish, the richness and diversity of musical expression.

The precise methods for eliciting inner psychological constructs associated with auditory copresence in immersive distributed music networks, as well as their influence on subjective experiences and technical results, remain largely unanswered inquiries. For example, one of the current unknowns is whether immersive NMP experiences translate to effective musical

quality in the traditional sense. While some early experiments have shown promising results, there is still much research to be done in order to fully understand the relationship between different realms of quality evaluations. Additionally, questions remain about how to design and implement distributed immersive musical experiences in ways that are engaging and meaningful for concurrent performers and audiences alike.

1 Problem Statement

This dissertation is concerned with the idea of *Immersive Network Music Performances* as applied to experimental applications involving different combinations of design constraints and organizations of musical performers and audiences. To summarize the main motivation behind the work of this manuscript, the following research question is posed: *"Does immersive audio technology improve the quality of a distributed network performance?"*. The nature of the problem is expanded through review of previous literature and projects and through a new empirical study produced for this dissertation. While extensive literature exists on the effect of latency on distributed performance, and on the effect of virtual acoustic environments on the "immersive experience", there is little material published on the combinations of these two elements and their combined effects, evaluated through different lenses of "quality".

Central to this dissertation is the conversation around how an immersive distributed music application can make the experience closer to that of a traditional music interaction. Subsequently, the subjective assessment of immersive qualities has to be put in the context of the application hereby examined. In this sense, it is sought to create an auditory illusion for two musicians remotely connected from different sorts of environments (eg. concert hall and studio booth) and measure the effects on performance. The goal is for the participants to come closer to the sensation of performing as if together in the same acoustical space. In the context given, musicians use technology to participate in a distributed performance that cannot happen in the natural world. Yet, one of the primary objectives of immersive technology is to simulate "reality" and remove the awareness of the medium. Since the main experiment illustrated by this document does not employ direct visual components linking the connected nodes, such as video or *avatar* representations, the reference to "immersion" is strictly limited to the auditory realm.

The psychological constructs that build the inner sensations of reality are usually identified in the "plausibility" realm of attributes. For example, a system can be rated based on how close is an audiovisual output to the expected sensorial experience of a person. However, another aspect that is central to a distributed interaction is that of *"Presence"* as transportation (Biocca et al. 2003; Nowak 2001) and virtual co-location (Mason 1994; Zhao 2003), or more specifically *"Copresence"*. Copresence is a central desirable outcome of interaction through a virtual or augmented medium and points to the inner belief of a user of "being present in the space with someone" whether virtually or in real life, and is increasingly being used as an evaluation metric for assessing the immersive success of a social augmented or virtual reality experience.

1.1 Significance of Study

The results attained by this work primarily expand the existing literature by providing insight into the effects of auralization strategies and latency levels (interacting with other inherent factors) on subjective and objective quality metrics, exploring relationships across evaluation dimensions. In addition, the work adds empirical data to the validation of immersive experience models (Lee 2020) as applied to the context of network music performance. The formulation of the specific hypotheses brought forward is driven by real challenges encountered while working on collaborative XR experiences dedicated to music interactions. The answers to the formulated hypotheses directly inform future iterations of experience design, showing the extent to which interventions targeting "immersive" attributes of a communication network translate to the objective of producing accurate music performances in distributed settings.

In practical terms, the conducted studies are intended to inform the future design of immersive distributed music networks and indicate how the investigated factors may affect the success of the collaborative interaction. These insights can help to guide the balance of the tradeoff between system complexity and "immersion quality", according to the target objective of an application. Moreover, the vast quantity of data collected through this work can be useful to the wider NMP community to expand the analysis to other hypotheses and secondary effects.

2 Dissertation Overview

This document is divided into two main parts. The first part illustrates a series of exploratory studies looking into the combination of immersive audio auralization methods within the context of distributed Music Performances (NMP), and the description of pilot "augmented NMP concerts" implementations involving several layers of media distribution. Each work is discussed with a dual-lens that combines the perspective of the scientific and engineering challenges with those pertaining to digital experience design, illustrating the driving theory and principles. The second part of this dissertation concerns an experiment designed to explore the specific role of auralization in relation to psychological constructs related to *auditory copresence* (a type of "social telepresence") and their impact on the musical outcome of a two-way distributed performance of a musical piece, simulating an internet-based collaboration. The common thread linking the works exhibited in this document is the pursuit of enhancing our comprehension of the technical and conceptual obstacles inherent in creating and executing immersive distributed music experiences, while also collecting valuable data for scientifically studying the impact of technology on individuals.

2.1 Dissertation Structure

The dissertation commences in Ch. II with an exposition of pertinent background literature related to research areas associated with the work presented in this dissertation, with a focus on some key studies that provide the motivation and theoretical background for the experiments portrayed in later chapters. This chapter includes a summary of auralization methods and their effects on interactive media and systems, and their "immersive qualities". While the literature primarily focuses on the auditory domain, the interplay between sound and the visual field of a listener is crucial to creating convincing auditory illusions. Research investigating the impact of visual context on sound perception is particularly relevant to multimodal systems. This discussion is anchored in the extensive body of literature on various aspects of *Presence* (such as *telepresence, social presence, copresence*, etc.), which represents the desired psychological state to induce in users of immersive virtual and augmented environments. Lastly, the dissertation delves into the subject

of *Network Music Performances,* discussing its current state-of-the-art, the standard challenges that are encountered, and novel experimental applications in the field.

The following two chapters, Ch. III illustrate previous relevant work conducted and published by the author and colleagues during the doctoral program at NYU's Music and Audio Research Lab. Previously accomplished work gravitates around audio engineering research aimed to investigate the interactions of real and virtual Acoustics in VR and AR experiences, the effects of head-mounted-displays on the local acoustic field, and examples of VR music experiences experimenting with integrating combinations of real and virtual sources with diverging embedded acoustic character. Furthermore, the chapter presents previous work in the realm of NMPs, primarily focused on the "Holodeck" - a complex multimodal experimental network architecture for real-time augmented interactions. Proof-of-concept reduced versions of the Holodeck have been employed to investigate augmented concerts that involve musicians and performers distributed across different locations within the network nodes and remote locations, as well as real-time avatar embodiment of stage performers, revealing a challenging intertwined combination of individual interaction paradigms to solve. The chapter also features summaries of more targeted studies that explore the characterization of latency and acoustic features of distributed audio networks, along with a study on the impact of spatial direction on musical interactions.

The core of the dissertation is contained in the chapters going from Ch. IV to Ch. VIII representing the second part of the document. This part presents a previously unpublished scientific experiment carried out for this dissertation. The principal goal of this study is that of exploring the effects of the interactions of audio transmission latency and auralization strategies over a distributed music network evaluated through different layers of "quality". This is explored through a set of auralization "modes" designed following principles drawn from the mixed-reality and virtual-reality fields, in the attempt to create a feeling of "immersion" and "presence" in users. In the process, hypotheses are made on how the subjective elicitation of "Auditory copresence" in participants of distributed music systems can be facilitated, and how it does correlate to other layers of evaluation that look at the technical outcome of a performance. The purpose is to understand better how "immersive quality" ratings can work as a proxy to predict task success within the NMP context. Ultimately, the study aims to make a case for introducing spatial audio

and virtual auralization processes within NMP systems and drive towards a digital experience that more closely emulates that of a "real" performance without sacrificing effectiveness.

In specific, Ch.IV lays the motivations behind the study goals and illustrates the design of different auralization environments applied at the network nodes, as well as the general study constraints applied. The following chapters, Ch. V and Ch. VI illustrate the methodology used for the technical setup of the experiment, including acoustic measurements used for the auralization effects, and the data collection process to acquire the primary data in the form of raw recordings. The primary data is used to extract three subsequent layers of evaluation data, consisting of direct subjective experience ratings from the actual study participants, objective tempo/beat performance metrics, and third-party annotations and ratings by musically-literate individuals. The secondary layers form the data being fed to a statistical analysis framework. This framework is described in Ch. VII and it involves the use of "Generalized Linear Mixed Effects Models" (GLMM for short) to estimate the relationship between the effects of auralization, latency, and other secondary factors, towards the results of the different evaluation layers. The analysis draws effect size estimations and identifies significant contributors to observed variances while accounting for random effect variations. A correlation study on the observed variables provides further insight on the connections between the inner subjective experience of the performer-participants and the objective musical outcome produced within the established network. The discussion portrayed in Ch. VIII analyzes the results in connection with the starting research questions and hypotheses, with a focus on the link between virtual and augmented auralization networks and "presence" constructs and whether these constructs can serve as a proxy to evaluate the success of the application task, that of an effective collaborative music performance over a distributed network.

The final chapter of the dissertation draws higher-level conclusions from the work presented, in the larger context of integrating XR and immersive audio in novel types of NMPs. From the viewpoint of engineers, performers, and digital experience designers, this chapter provides a summary of the array of scientific and artistic challenges encountered in the process of designing experiences.

3 Key Terms

The following definitions are provided to remind the context of the work portrayed in this chapter. This chapter defines "Virtual acoustics" as the field of study concerned with the simulation and reproduction of acoustic environments in a virtual or digital space. It involves using digital signal processing techniques to recreate the complex acoustic behavior of real-world environments, such as concert halls, recording studios, or outdoor spaces, in a simulated or virtual environment. With the term "Extended Reality" (XR) we describe immersive technologies that enable users to interact with digital content and the physical world in new and enhanced ways. It encompasses a range of technologies, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), which offer varying degrees of immersion and interaction, usually through a headset device.

Network Music Performances (NMPs) refer to musical performances that involve musicians located in different physical locations who collaborate and perform together in real-time using networked technology. This is usually achieved through the use of high-speed internet connections and specialized software that enables real-time audio and, optionally, video communication between musicians. Non-internet-based NMPs also exist and are achievable through dedicated infrastructure but present high costs and geographical limitations. NMPs can take many forms, ranging from small-scale improvised performances between a few musicians to large-scale concerts featuring multiple performers located in different parts of the world.

4 Related Academic Contributions

At the time of the distribution of this document, the principal study presented in this dissertation has not yet been submitted for peer-reviewed publication. However, the journey that led to this dissertation has helped produce the following related academic contributions:

XR Experience Design

- **Andrea Genovese**, Marta Gospodarek, and Agnieszka Roginska (2019b). "Mixed realities: a live collaborative musical performance". In: *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019; 5th International Conference on Spatial Audio; September 26th to 28th, 2019, Ilmenau, Germany*, pp. 159–164

- Marta Gospodarek, **Andrea Genovese**, Dennis Dembeck, Corinne Brenner, Agnieszka Roginska, and Ken Perlin (2019). "Sound design and reproduction techniques for co-located narrative VR experiences". In: *Audio Engineering Society Convention 147*. Audio Engineering Society
- Cindy Bui, **Andrea Genovese**, Trey Bradley, and Agnieszka Roginska (2020). "Multimodal Immersive Motion Capture (MIMiC): A workflow for musical performance". In: *Audio Engineering Society Convention 149*. Audio Engineering Society

Distributed Music Network Studies

- Robert Hupke, Sripathi Sridhar, **Andrea Genovese**, Marcel Nophut, Stephan Preihs, Tom Beyer, Agnieszka Roginska, and Jürgen Peissig (2019b). "A latency measurement method for networked music performances". In: *Audio Engineering Society Convention 147*. Audio Engineering Society
- Robert Hupke, **Andrea Genovese**, Sripathi Sridhar, Jürgen Peissig, and Agnieszka Roginska (2020). "Impact of Source Panning on a Global Metronome in Rhythmic Networked Music Performance". In: *2020 27th Conference of Open Innovations Association (FRUCT)*. IEEE, pp. 73–83

Acoustic Calibrations Methods

- **Andrea Genovese**, Gabriel Zalles, Gregory Reardon, and Agnieszka Roginska (2018). "Acoustic perturbations in HRTFs measured on Mixed Reality Headsets". In: *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society
- **Andrea Genovese** and Agnieszka Roginska (2019). "Hmdir: An hrtf dataset measured on a mannequin wearing xr devices". In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society
- **Andrea Genovese**, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J Tashev (2019a). "Blind room volume estimation from single-channel noisy speech". In: *ICASSP*

2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 231–235

- Braxton Boren and **Andrea Genovese** (2018). "Acoustics of virtually coupled performance spaces". In: *International Conference on Auditory Displays, ICAD*. Georgia Institute of Technology
- Julian Vanasse, **Andrea Genovese**, and Agnieszka Roginska (2019). "Multichannel impulse response measurements in MATLAB: An update on scanIR". in: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society

Events

In addition, the author contributed to the design and implementation of two academic events which showcased the "Holodeck" platform to the public and generated interest in the questions leading to the presented work:

- *"Concert on the Holodeck: Connecting Artists"* (Apr. 2018)
- *Ozark Henry on the Holodeck: Map to the Stars* (Oct. 2018)

CHAPTER II

BACKGROUND AND LITERATURE

This literature review provides the theoretical background and overview of existing research and scholarship relevant to the study. The purpose of the literature review is to establish an understanding of the research problem and the underlying theoretical framework that informs the study. While the main lens utilized for the studies is that concerning immersive audio engineering, a multidisciplinary perspective is necessary in order to perform an investigation where several fields interact for the creation of a future-oriented application.

The relevant theoretical background for the understanding of this work refers to the study of "presence" in digital mediums, the categorization of immersive technologies and their main differences, the connection between presence and room acoustics, an overview of distributed music performance networks and their challenges, the effect of room acoustics on music, and more. All of these topics intersect to form the background that informs the constituent theory behind this dissertation.

1 Virtual, Augmented, and Mixed Reality

An important area to define is that of "immersive technology", in particular *eXtended Reality (XR)*. The umbrella term "XR" is an inclusive term that incorporates the concepts of Virtual (VR), Augmented (AR), and Mixed Reality (MR) systems. This dissertation touches on all three of these technologies.

In recent years, research and engineering have broken new frontiers in XR at an accelerating pace. The current technological landscape has opened up new horizons and realms of applications for all types of performing arts. Although immersive technology has existed for decades (Ohta and Tamura 2014; Minsky 1980), it is only now that we can tangibly envision the use of such devices in our daily activities and profession-specific applications.

Head-mounted displays, tracking sensors, and machine learning have improved immersive multimedia technology to a level that might have seemed "futuristic" only a few years ago, giving developers the necessary tools for quick implementations. Within the new landscape of mixed and virtual reality, there is space for the enhancement of forms of artistic collaboration, such as distributed music performances. As for other forms of arts (Murray 2017), new interdisciplinary research fields are needed to support and guide the correct use and development of technology in relation to music and virtual presence.

Immersive audio technology is deeply involved in this branch of computer science, three-dimensional spatial audio allows us to localize and externalize sound with the intent of simulating the real world, while acoustic rendering methods aim to reproduce a high-fidelity virtual soundfield as close as possible to the expectations of a user given the present environment. When the sensorial expectations are met, the immersive system can be classified as "plausible" and considered capable of eliciting sensations of "presence" and "immersion".

1.1 eXtended Reality and Collaborative Environments

The definitions for "Virtual-", "Augmented" and "Mixed Reality" are often attributed to Milgram and Kishino. Milgram gave the definition of MR as the "merging of real and virtual worlds, somewhere along the *virtuality* continuum, which connects completely real environments to completely virtual ones. In (Wagner et al. 2009) it is added that "MR systems augment the real world with added virtual features (augmented reality, AR) or augment the virtual world with real features (augmented virtuality, AV)". This taxonomy space is illustrated in Fig. 1 showing the "*virtuality* continuum", declaring Mixed Reality to include the range of the continuum between, but excluding, Real Environments and Full Virtual Environments. This representation is a simplification of a design space that comprises three main factors: reproduction fidelity (of the mediated stimuli), the extent of presence (conditions under which the physical stimuli are received), and the extent of real-world knowledge.

Collaborative virtual environments are online spaces that allow multiple users to interact with each other and with digital content in real-time. Users can be represented by avatars, which are virtual representations of themselves that they control within the virtual environment. Users can communicate with each other using text, voice, or gesture-based commands and can

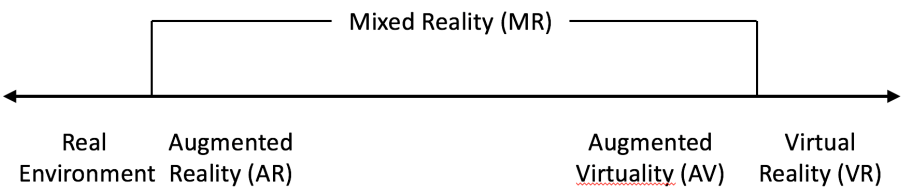


Figure 1: The virtuality continuum. Mixed-Reality comprises the range between real world (complete reality) and virtual reality (complete virtuality). Image from (Milgram et al. 1995).

work together to achieve shared goals, such as collaborating on a project, solving a problem, or participating in a virtual event. Laboratories such as the NYU Future Reality Lab have been developing early prototypes of what future interactions in XR would be like. In the Holojam project (*HOLOJAM* 2014), audiovisual sensors track multiple users and allow them to interact through new forms of augmented communication (Perlin 2016). The platform has been used to explore new forms and dimensions of immersive theater (Gochfeld et al. 2018), virtual collaborations (Xia et al. 2018), and narrative MR art installations (Lobser et al. 2017), paving the way for research in collaborative work and social interaction in virtual environments.

Another similar project is that of the "Holodeck" at NYU (Plass et al. 2022), a multi-room platform capable of transmitting multimodal data across different nodes through a central server (*NYU Corelink | Homepage* n.d.) that routes, records and processes the different types of data streams. The streams can include audiovisual captures, motion capture data, rendering configurations, and more. The range of applications that can be created on this network of nodes is very wide, going from music performance to education and other sorts of simulation environments. The particular approach of the platform is that of *asymmetric rendering,* where each node adapts the interpretation of data in a way that fits the local rendering requirements and conditions. Although the platform is still in development, some proof-of-concept work has been conducted on it and is presented in Ch. III.

1.2 Presence in Virtual Environments

One crucial aspect of evaluating collaborative virtual reality/mixed reality (VR/MR) environments is the concept of *presence*. *"Presence"* can be considered as a psycho-physiological attribute

that combines the concepts of *"plausibility"* and *"engagement"* into a single latent psychological construct, which can exist in various forms and topologies (Biocca et al. 2003).

The philosophical debate around the concept of "presence" led to different variations of the definition. Traditionally, presence has been described as the perceptual illusion of "being there" (Heeter 1992) or "illusion of non-mediation" (Lombard and Ditton 1997). "Presence" is often portrayed as a multidimensional encompassing construct (Lombard et al. 2009), impacted by the sensation of "transportation", "realism" of sensorial stimuli, "perceptual" and "cognitive immersion", "social richness", "social self-identification", and "illusion of medium as actor". Several alternative conceptual models have been proposed with different organizations of hierarchical structures of these elements (Lee 2020; Riva et al. 2003). Nevertheless, "presence" is considered an important desired outcome of any immersive virtual environment, and the debate focuses more on its possible measurement. Evaluation methods in general tasks related to presence have ranged from the evaluation of subjective psychological phenomenon (questionnaires) to the observation of objective biosignals (Slater and Steed 2000).

A renewed decomposition, relevant to the proposed work, is given by (Riva et al. 2003), where presence is divided into physical presence, "being in a place", and social presence, "being together with another person". Together, the two form *copresence*, the feeling, or illusion, of "being together in a shared space". This concept fits the purpose of acoustics in distributed music as the goal is to create a subjective shared acoustic space, asymmetrically at each node. Furthermore, *copresence* is also defined as a mutual exchange (Campos-Castillo 2012) where, in addition to feeling as "being together", the user also feels as "being perceived". In relation to this dissertation, *"auditory copresence"* is here defined as the illusion felt by a user of an immersive system (for example, a musician) when perceiving a connected user (co-performer) as "being here with me", or the illusion of being transported to a remote location where the connected user is present, essentially "being there with someone".

Although established and validated *presence questionnaires* are found in literature (Witmer and Singer 1998; Lombard et al. 2009; Lessiter et al. 2001), the field lacks a modern methodology with proven reliability, validity, and sensitivity, capable of capturing *auditory copresence* in immersive systems. In (Floridi 2005) a criticism is made that the methods of assessment of presence cannot be purely subjective, as the measurement must be "objective and observable",

nor purely objective since external observations must be related to internal mental states. It has been argued that it is possible to pair the success of presence with the success of actions in an environment for which "presence" is a support (Zahorik and Jenison 1998). In other words, the success of a distributed connection, where the meaning of the interaction is socially co-constructed through dimensions of presence, can relate, by proxy, to the successful psychological manifestation of presence (Mantovani and Riva 1999). Therefore, research methods must be multidimensional in nature and context-dependent. In their review paper (Wagner et al. 2009) the authors make a point that the best approach is a combination of ethnographic observations, interviews, analysis of artifacts (activities performed in MR), and presence questionnaires.

2 Background on Immersive Audio

Immersive Audio is a branch of audio engineering at the intersection of *acoustics* and *cognition* that aims to study the characteristics of sound environments and model their perceptual effects. In general, "the sense of immersion" can be achieved through a constructed soundscape of directional and non-directional sounds surrounding the listener" (Roginska and Geluso 2017). The study of human perception of sound, how we perceive the sense of space and distance, or how do we localize an auditory event around us, has led to great advancements in simulation technology where the goal is to digitally recreate a plausible auditory scene. By adding a three-dimensional layer of auditory reproduction, spatial audio technology has contributed to advancements in XR technology, entertainment media, navigation for the visually impaired, sonified information environments, audiology, and others. Several modern multimedia applications make use of immersive audio technology; for example, multichannel surround sound is a typical use case for immersive audio available to consumers. Binaural audio through headphones has brought spatial sound content to mobile devices by simulating interaural cues representing the location of a sound source with respect to a listener (Kendall 1995). Spatial audio can also improve the understanding of an auditory scene (Bregman 1994), a modern example is found in teleconferencing applications, where the sense of telepresence and intelligibility of the interactions have been shown to improve when the directional components were preserved (Pulkki 2007).

In recent years, the field of virtual reality and game audio has begun to look at spatial audio

technology as a key component in improving the virtual experience and enhancing the plausibility of virtual scenes (Friberg and Gärdenfors 2004). In fact, new immersive displays such as mixed and virtual reality experiences represent a very appropriate field of application for spatial immersive audio, as the goal is to enhance the local soundfield by rendering digital "object" virtual sounds with the illusion of belonging to the local listening space of a user. To create such illusions, "spatialization" and "auralization" techniques can be applied to simulate the directionality of a sound source, and its acoustic behavior within a reflective room, adding a "plausible" character of realism to the original sound material.

2.1 Auralization

Immersive systems, such as mixed-reality or virtual-reality platforms, make extensive use of "auralization" as a method to simulate the acoustic behavior of a physical space or an acoustic system, with the intention of creating an auditory illusion where sound is perceived as originating from a target acoustic environment. In the field of immersive audio, "auralization" indicates the process of simulating the acoustic character of a target space (e.g. a particular room, an outdoor space, a cathedral, a cave, etc.) over an audio stream or media object. The simulation recreates aspects such as reflection patterns, diffuse reverberation field, frequency-dependent decays, and more (Kleiner et al. 1993). A common auralization technique relies on the application of spatial *Room Impulse Responses* (RIR), which in essence embed the reverberation patterns of a room in the form of stereo FIR filters, applicable through signal convolution processes that can superimpose the acoustic room character (made of diffused and directional components) onto a sound source in real-time.

Generally speaking, the acoustic path of a sound stimulus produced in a room, traveling toward a receiver, would acquire acoustic coloration depending on the position of the source and receiver within the room (itself affected by room dimensions, surface materials, and obstacles). The coloration is due to acoustic interferences between and among directional room reflections and the diffused reverberation field with the direct sound arriving at the receiver. For example, a church has a very distinct different sound than that of a recording studio, a RIR can describe the auditory cues that create that difference. RIRs can be measured in situ. By reproducing an impulse sound, for example, a balloon pop, we can capture the reflections and reverberation pattern of the

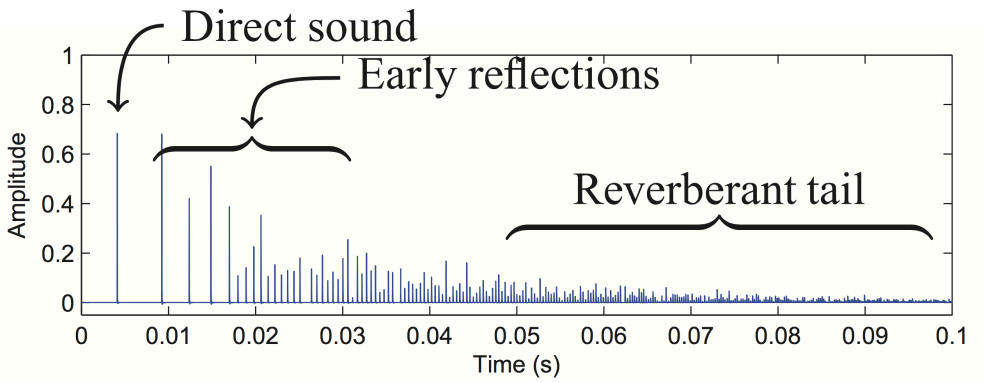


Figure 2: Characteristic elements of a room impulse response. Image from (Schimmel et al. 2009).

space, capturing its "fingerprint" or "character". Measurements are usually taken using a swept sinusoidal signal, which embeds equal energy at all frequencies between the desired start and end frequency points (Farina 2000). A deconvolution process between the recorded sinusoidal sweep signal and the original test signal can finally retrieve the RIR in the time domain.

A typical RIR (figure 2) is composed of a position-dependent part (subdivided into a direct path and a "early reflections" part) and a "diffused position-independent part " that describes the late reverberation curve. While the position-dependent part is specific to the location of the source and the microphone within the room, the diffused part response is theoretically identical at any unoccluded location within the space. RIRs are reference characterization curves that can also be used as convolution filters since they embed the transfer function describing the sound reflection behavior. Thus, a signal recorded under anechoic conditions can be simulated to sound as if it were in a different place. In practice, whenever real-time processing is required, the simulation takes place through buffered frequency domain multiplication (Cooley et al. 1967).

2.1.1 Spatialization

Spatial audio for headphones, known as "binaural audio", uses special signal processing filters to perceptually simulate the location and distance of a sound source around the head. When measuring an impulse signal using in-ear microphone capsules instead of a regular microphone, we can capture the acoustic transfer path of the source in relation to the ears. The resulting stereo measurement embeds the localization cues that tell the human brain where a sound is

located. The principal cues are *Inter-Aural Time Delay* (ITD), *Inter-aural Level Difference* (ILD), and spectral distortions caused by diffraction and shadowing effects of the head and torso, as well as resonances caused by the ears' pinnae (Blauert 1997). The response of a source placed at distance d, azimuth angle ϕ, and elevation angle θ (spherical coordinates) is described by transfer functions that encode the ITD, ILD, and spectral cues. These transfer functions are called Head-Related Transfer Functions (HRTFs) or Head-Related Impulse Responses (HRIRs) when in time-domain form. Although HRTFs can be recorded on dummy head microphones for generalized responses, or calculated using computational models (Algazi et al. 2002), the best degree of output quality is obtained through individually measured HRIRs by inserting the microphones into the ears of the intended listeners. These individual filters embed the localization cues caused by the subject's specific head and ear shape and size with higher-fidelity than a general HRIR measurement.

The combination of auralization and HRIR binaural filters can create the double illusion of a sound source perceived as "being in a certain place" and "being in a certain location". If the final acoustic character of the rendered audio matches the expectations of a listener, then the virtual recreation is likely to be subjectively deemed more *realistic* or *plausible*, as it merges with the local reflection patterns, while the time and level of arrivals to the two ears create a directional cue.

2.1.2 Binaural Capture

It is possible to directly capture the spatial acoustic field of a reverberant room by the measurement of *binaural room impulse responses* (BRIRs) which, like HRIRs, can be recorded for general fit (with a "dummy head" baffle microphone representing an average human body) or individual fit (with in-ear capsules placed in the listener's ears).

The acoustic path from a source in a room to a receiver, directionally originating from an azimuth angle θ, an elevation angle ϕ and a distance r represents a room transfer function. If the receiver is a human listener, a further stage of signal coloration is introduced by the interactions of the incoming direct and reflection wavefronts with the pinnae "receiver" ears. The geometrical offset of the ear receivers results in two different paths embedding the "binaural cues" relating to the original emitted sound and its reflections. The cues consist of time, phase, and level differences that the human brain can decode into a perceived sound location in three-dimensional space. To properly capture these spatial relationships, it is possible to utilize specialized binaural

microphones that comprise a rigid body "human head" in between a stereo-pair of receiver microphones ("dummy head microphones"). By directly measuring impulse sounds in a room with a binaural microphone, we obtain a *Binaural Room Impulse Response* (BRIR). BRIRs are thus FIR filters that describe a static acoustic relationship between a source in three-dimensional space within a room and a generalized human listener within the same room. In this document, the resulting time-domain acoustic path H_R of a source in room R is defined as follows:

$$H_R(t)\{\theta, \phi, r, \mathsf{ch}\} = \mathtt{BRIR}_R(t)\{\theta, \phi, r, \mathsf{ch}\} + \xi \tag{1}$$

Where θ and ϕ are the polar angles of incidence of the direct wavefronts, r is the distance between source and receiver, 'ch' denotes either the Left or Right ear signal channel, and ξ is an umbrella error term comprising coloration error introduced by the electronic equipment used for the reproduction of the emitter stimulus and the capture of the signals.

2.1.3 Application of Acoustic Filters

Any acoustic response filter, whether RIR, HRIR, or BRIR, can be transferred to an anechoic signal by the process of signal convolution between the FIR filter and a signal buffer (eq:2).

$$y(t) = h(t) * x(t) \tag{2}$$

Where $y(t)$ is the time-domain processed signal, $h(t)$ is the acoustic impulse response FIR filter, $x(t)$ is the original dry signal to be processed, and $*$ is the convolution operator. A simple fast convolution version exists in the form of multiplication of the Fast Fourier Transform (FFT) of the signal and impulse response, followed by an inverse FFT.

$$y(t) = \text{IFFT}(H(f) \times X(f)) \tag{3}$$

By convolution, we can make a signal sound as if recorded in a different space, or we can create the illusion of it coming from a particular direction.

2.2 Room Acoustics Modeling

A room-impulse-response can be decomposed into several parameters which describe its general shape and elements, those are mainly the "reverberation time" T_{60}, described as the required time for the source energy to decay by 60 dB, the "direct-to-reverberant ratio" (DRR), which acts as a contextual source-distance cue, calculated as the energy ratio between direct sound and diffuse reverberation, the "Initial Time Delay Gap" (ITDG), the time it takes for the first reflection to arrive after the direct sound is received, and many others like the early reflection's density, which describes the sparsity of the first reflections coming from walls, ceiling, and floor (Kuttruff 2014). The higher the number of parameters we know about locally measured RIRs, the better we can statistically reconstruct and model different RIRs that describe a different position within the same room. A geometric modeling strategy would be to use the precise dimensions and shape of the local room of the target room and the surface absorption coefficients. Computational models can then be employed to calculate the path of reflections within a room with different degrees of complexity. Geometrical dimensions are particularly useful for the direct and early part of the RIR, as they relate to reflection arrival time, which changes with receiver distance from the sound source and wall boundaries.

The simplest of the models is the Image Source Method (Allen and Berkley 1979; Dance and Shield 1997). This algorithm uses the three-dimensional shape of the room to calculate the sound reflection paths from a source to a receiver. Paths are calculated by linearly mirroring the direction path of the waveform from a phantom source reflection image to the receiver, up to a desired order of reflections. More advanced variations include surface absorption rates and a frequency-dependent decomposition. This model computes an overly-ideal pattern of reflection that ignores certain aspects of the physics of acoustics such as scattering effects, rough-edges diffraction, and real-world non-linearities.

Statistical models are very popular for the creation of artificial reverberation. A simple reverberator consists of a stochastic, exponentially decaying noise envelope model, which can be tuned to a few parameters representing the decay rate, the initial energy, and the noise spectrum (Schroeder 1962; Jot et al. 1999). Perceptually, an artificial stochastic reverb is theoretically indistinguishable from diffuse-field reverberation, but in practice, it is hard to determine the correct place at which the stochastic reverb should plug-in within a synthesized RIR. Synthesis

parameters can be extracted from either geometric measurements or statistical analysis of measurements in similar-sounding spaces. For example, the mixing time, the time at which the late reverberation part starts in a room can be approximated using the geometric cubic volume in m^3 as $M_t = \sqrt{V}$ (Howard and Angus 2017).

A more complex analysis framework is given by the "Reverberation Fingerprint" that characterizes the diffuse sound of a room, independent of a specific source-receiver configuration, directivity pattern or orientation (Jot et al. 1997). The elements that define the fingerprint are provided as frequency-dependent reverberation decays (T_{60}), called "Energy Decay Relief" ($EDR(t, f)$), and the initial power spectrum ($P(f)$, alternatively, the volume of the cubic room can be used). This framework makes it easy to adapt the fingerprint of a local, unmeasured, room from a reference EDR measurement; by using knowledge of the reference and local room volume, the following relationship can be applied:

$$P(f)_{\texttt{local}} = P(f)_{\texttt{ref}} \frac{V_{\texttt{ref}}}{V_{\texttt{local}}} \tag{4}$$

An entirely different family of RIR modeling is that of wave-solver computational models. These models rely on the physics of sound to derive spectral basis functions from analytical solutions, discretized into sampled partitions. "Boundary Element Methods" (BEM) can use arbitrary digital meshes of shapes to compute the way a wave propagates, bends, scatters, diffracts, and reflects around a shape or against a boundary. In addition to calculating RIRs (Habets 2006), BEM has been used to compute HRIRs by solving the acoustic field around the head (Katz 2001). In "Adaptive Rectangular Decomposition" (ARD) tools, the spatiotemporal reflections and scattering of virtual wavefronts against a rectangular voxel decomposition of a mesh scene can be computed (Raghuvanshi et al. 2009). Although a usually very expensive process, improvements in GPU technology have helped make this technique much faster and more palatable for sound designers (Mehra et al. 2012). These techniques are very popular in game audio applications and architectural acoustics, where a digital mesh of the structural environment is used to place virtual sources and virtual probes, to create virtual physical-modeled signals.

2.3 Object-Based Audio

Object-based audio is a widely adopted audio engineering rendering paradigm in which virtual audio sound sources are described by their content and by time-stamped metadata describing the intended three-dimensional location of a source within a scene (Tsingos 2017). Unlike the traditional channel-locked mixing process, sound objects can be flexibly rendered as emitting from any virtual location regardless of the configuration environment and reproduction equipment, although certain minimum specifications need to be met. The use of object-based paradigms can allow dynamic updates of a sonic environment, and navigation paradigms such as 3DOF and 6DOF can dramatically improve the quality of the experience in XR applications. New transmission codecs, such as MPEG-H (Herre et al. 2015), allow the encoding of spatial audio scenes into an object layer and an ambiance layer, allowing flexible decoding at the receiver side which adapts to the local reproduction configuration.

"Object sources" can be flexibly rendered in real-time using listener-tracking sensors. By adding an IMU (Inertial Measurement Unit, a combination of gyroscopes and accelerometers) head-tracker to a listener's headphones, we can lock rendered virtual sources into space. This means that as a user moves the head, a different HRTF filter will be used to process the new location of the sound in relation to the user's orientation. *Head-tracking* has been found to greatly improve the perceptual accuracy of a spatial audio display and quality of the experience (Begault et al. 2001). This is mostly due to the fact that binaural auditory cues are sometimes ambiguous. For example, the spatial ambiguities of sound sources placed on the "cone of confusion" around the head, where the ITD and ILD cues are identical at all locations, can be resolved by shifting the source location with head movement, resolving the confusion by creating a perceptual trajectory path. Realizing 6DOF systems is much more difficult. Besides rotational tracking, a positional tracking system is necessary to detect the user's proximity to walls and sources and allow a dynamic update of a room acoustics model. While the diffuse parts of a soundfield within a room is isotropic (location independent), the direct and early part of the sources' reflection patterns respond differently to the position of a receiver in the room. A dynamic virtual acoustics model thus usually operates on the separate dissected parts of a room model, adapting the virtual acoustic path of each audio element in response to the user's movement.

2.4 Measures of Quality

Regarding the evaluation of immersive audio systems, their success is often determined by subjective assessments of spatial audio qualities and immersive attributes, as well as technical accuracy rates. Several studies have been conducted in search of appropriate quality attributes to provide to listening test participants to rate sound in different categories of judgment. Some examples of agreed terms related to surround sound are "naturalness", "envelope", "timbral balance", and "presence" (Rumsey 2002). Binaural audio for headphones is usually evaluated for its ability to provide correct *localization* across several localization dimensions such as azimuth, elevation, distance, and hemisphere. Accuracy metrics are easy to analyze given the quantifiable rate of correctness between the perceived and intended source location. When the goal is to assess *externalization,* the sense of the sound being perceived as "outside of the head", the scale looks into a more abstract dimension of the sensorial experience (Reardon et al. 2018c). Despite the fact that externalization is always desired, it is hard to quantify levels of externalization and it is usually easier to formulate it as a "True/False" binary task. Work has also been done in terms of relationships between quality attributes and general preference; it was found that the choice of preferred rendering algorithm derives mainly from coloration-related attributes, although high content dependency (e.g. music vs. movie stimuli) was reported (Reardon et al. 2018a). However, the added dimensions of movement in 6DOF mixed reality have created the need to look for new types of multi-modal attributes which connect sound to perspective congruence and head movements (Olko et al. 2017). For these reasons, aspects such as "cohesion" and "stability" are becoming increasingly relevant in the field.

Another interesting way to look at the aspects of quality is the rating of *plausibility*. "Plausibility" is based on the general degree of belief felt in perceiving a virtual sound source as real (or more generally the credibility of a virtual scenario), which in technical terms means the accuracy with which sensorial expectations are met (Lee 2020). This aspect has been tested using guessing rate methods from listeners wearing headphones in the presence of a loudspeaker array (Lindau and Weinzierl 2012). By sending signals randomly to speakers or headphones, the guessing rate between "real" and "virtual" can be analyzed and used to rate the plausibility of a system. Reverberant environments have been reported to increase the rate of guessing against correct detections, indicating a higher plausibility of rendered content (Pike et al. 2014). A more

direct approach was taken in (Väljamäe et al. 2004) where subjects were asked to rate the quality of "presence", defined as a sensation of "being actually present in the virtual world", on a scale from 0 to 100 on rotating soundfields. Significantly higher presence ratings were found when individual HRTFs were used as opposed to generalized HRTFs from a dummy head. Presence has also been tested using 7-point Likert scales in a series of experiments that link soundfield movement and visual association to higher ratings (Ozawa et al. 2003b). Further multiple regression decomposition of psychological factors that affect presence found a correlation with attributes such as "naturalness" and "familiarity" (Ozawa et al. 2003a) of the displayed sound content.

3 Distributed Music

Making music on distributed music networks just as good as it can happen in real life is a great challenge of music technology. Issues related to network latency and the distant feeling of remoteness play a part in making this technology difficult to approach by musicians, from amateurs to experienced music professionals. The latencies that impact a system go through several stages; the delay introduced by the propagation of sound in physical space, AD/DA conversion, buffering and packaging on the sender/receiver side; the delay in data processing of the intermediate network nodes between the source and destination as well as the propagation delay over the physical transmission medium; and playout buffering which may be required to compensate the effects of jitter to achieve a sufficiently low packet loss rate.

Dedicated musical strategies and new contemporary genres are being developed in the world of academia to assimilate or compensate for the disadvantages of signal latency or asymmetrically mask the effects at one of the nodes (Carôt and Werner 2009). However, this type of performances are largely restricted to academic circles as popular and classical music genres are rarely attempted due to their stricter sensitivity to delay (Barbosa 2003). Research into this topic has been fairly sparse, with a few key projects, like CCRMA's SoundWire project (Chafe et al. 2000), leading the efforts in studying streaming protocols (Cáceres and Chafe 2010) and the musicians' behavior, while others have looked more into engineering-oriented analysis and solutions for latency, as well as musical coping strategies (Carôt et al. 2007). Other institutions like NYU and McGill have also been active in distributed music networks, with one of their

earliest experiments in the field involving the test of TCP and UDP transmission protocols for multichannel audio streaming (Xu et al. 2000).

The academic community of distributed music has released flexible ready-to-use software for the multichannel streaming of audio through the internet using UDP protocols. IP-based routing software like *Soundjack* (Carôt and Werner 2008) and *Jacktrip* (Cáceres and Chafe 2010) can be used to link signals from DAWs and interfaces to output ports from a transmitter computer to multiple listener nodes. Buffer size and sample rate are customizable in order to optimize latency stages, outside of the base network transmission latency, according to the available computational resources. The receiver node is able to route the incoming signal channel streams to a sound processing engine before final reproduction. Communication channels are also implementable as dedicated streams.

3.1 Effect Of Latency on Performance

The one-way latency threshold for cohesive integration of simultaneous sounds is usually reported to be between 20 to 30 ms (Hirsh 1959; Carôt and Werner 2009) depending on timbre, pitch, musical style, and other characteristics. This value corresponds to a physical distance of approximately $\sim 8.5mt$ for the propagation of sound in air at an average temperature. Previous research on latency impacts in Networked Music Performances (NMPs) (Chafe et al. 2004; Chafe et al. 2010; Farner et al. 2009; Chew et al. 2005) has primarily investigated rhythmic patterns in pairwise interactions based on hand clapping, examining factors such as consistency of tempo, synchronization, and time gaps. These studies revealed that delays under $10\,$ms to $15\,$ms result in accelerated performance tempos, as participants inherently tend to anticipate. Optimal synchronization with a steady tempo can be attained within a $10\,$ms to $25\,$ms range. Within the "usability range" of $25\,$ms to $65\,$ms, a deceleration in tempo becomes noticeable, and coping mechanisms can be employed; however, delays beyond this range significantly degrade performance quality. Although these results highlight general trends, the examined experimental task is somewhat unconventional for musical scenarios. More ecologically valid musical interactions have been studied by relating rhythmic intricacy and tonal instrument types to tempo fluctuations. In (Rottondi et al. 2015), it was discovered that intricate rhythms and greater spectral flatness (e.g., guitars, drums) led to more pronounced deceleration patterns. Additional

research linking tempo and latency demonstrated that factors such as genre characteristics, signal onset, musical interaction hierarchy, and musicians' familiarity with networked performance settings can influence both objective and perceived temporal synchronization (Bartlette et al. 2006a; Sawchuk et al. 2003; Delle Monache et al. 2019; Rottondi et al. 2016).

3.2 Musical Style Approaches

Extensive analysis of the signal stages affected by latency has led to the development of a taxonomy of musical strategies that can be employed according to the severity of the one-way delay between nodes and desired perspectives. In their review paper, Cârot & Wener (2007) describe the act of playing music as conventionally done when in the same room as the "Realistic Interaction Approach". In the presence of large latencies, an asymmetric "Leader-Follower" approach requires a "follower" node, supposedly where an audience is present, to play to the music as it is received, recreating perfect local synchrony, while the "leader" node produces the groove beat without being synchronized (but that would not matter for a concert since there would be no audience there). The "Laid-Back approach", which fits jazz-oriented musical styles, can be employed at latencies between 25 to 50 ms and consists of a slight behind-the-groove playing style, as done by choice in certain performances. One other strategy, interesting for large latencies scenarios, is the "Delayed Feedback Approach", which attempts to match the beat at each node by adding additional artificial latency, enough to match the sound at the follower node to be one beat, or measure, behind the leader node. This strategy may accommodate for the round-trip response at the leader node to also be on the beat, at twice the delay.

3.3 Evaluating Performance in NMPs

In distributed performances in the presence of latency, the most sensitive musical dimension is that of *rhythm*. The quality of performance over a network is usually analyzed through metrics related to tempo and beat stability (Rottondi et al. 2016). In (Chafe et al. 2004), the effect of latency on tempo has been tested through an analog bypass network in which latency was controlled as an independent variable. The evaluation metrics were based on tempo curve parameters obtained through linear regressions of inter-onset intervals (IOIs) of hand-clapping performers.

Specifically *tempo regression slope:* $b_{\hat{t}}$, as a measurement of acceleration against time, and *tempo jitter:* s^2, defined as the variance of the residual (eq: 5).

$$s^2 = \frac{\sum(\mathbf{t} - \hat{\mathbf{t}})^2}{n-1} \tag{5}$$

Where $\mathbf{t}$ is a vector of IOIs and $\hat{\mathbf{t}}$ is the linear regression prediction for a subject under given delay conditions. The tempo slope means, grouped by delay amount, revealed a negative linear relationship of deceleration with the delay amount. Remarkably, an acceleration effect on hand clapping interactions was found for latencies of < 11 ms, pointing out the possibility of a sweet spot where delay is beneficial. No significant interaction between initial metronome beat tempo and delay amount was found, indicating how the effect of latency on acceleration/deceleration is independent of tempo. In fact, the highly transient nature of the clapping task is likely to create recursive drags on tempo, where rather than performing as a self-correcting system, "players are often anticipating and pushing back on the drag" (Chafe et al. 2004).

Other evaluation systems require direct feedback from the participant. To test the tolerance limits of a network, a binary dichotomy paradigm "tolerable-intolerable" has been tested to find perceptual latency thresholds of different playing stratagems (Carôt et al. 2009). The study found that a maximum tolerable range falls within $35ms$ to $65ms$ with large individual tolerance variations dependent on combinations of beat pattern, tempo, and musical aptitude.

3.3.1 Subjective Evaluation in NMP

Subjective evaluations of qualities such as presence, enjoyment, and emotional connection have been previously explored using standard questionnaire forms related to the holistic experience. A 2009 study (Olmos et al. 2009) did not find a particular change in presence rating with different degrees of latency, but they found that rehearsal time had a small effect. However, the used paradigm made use of a video connection inclusive of a telematic conductor, so it is unclear how that rating would have changed in an auditory-only situation. A different approach to the problem was taken by (Bissonnette et al. 2016) in which the performers were asked to subjectively assess the level of anxiety and the quality of the performance after repeated rehearsal sessions in VR. A pre-experiment *Immersive Tendencies Questionnaire* (Robillard et al. 2002), aimed to test

the predisposition of individuals to feel immersed by asking about their concentration behavior during activities such as sports, gaming, etc. It was found that repeated exposures to VR can improve performance comfort and reduce anxiety, but no particular changes in subjective self-assessments of performance quality, concentration, or immersion were recorded.

3.4 Spatial Audio and Distributed Music

The application of spatial audio processing methods has not been extensively researched in the literature in relation to the field of distributed music performance. Generally, simple low-latency reverb processing units in receiver systems have been applied to a pipeline if so desired by a performer or to smear sharp signal transients (Chafe et al. 2000). More advanced pieces of technology, such as head tracking systems or individual HRTF filters, have the potential to enhance the immersive quality of an NMP system. However, the implementation of spatialization technology usually involves a trade-off between computational resources, the availability of calibration data, and fidelity. The biggest hurdle to the introduction of these advanced systems is the computational load that they would add to an already sensitive-latency communication paradigm.

Head-tracked rendering would be one of the most desirable features to implement to achieve 3DoF virtual environments. However, wireless Bluetooth transmission usually adds additional latency depending on the device clock speed, bit-rate, and codec employed (McPherson et al. 2016). New faster implementations such as *Bluetooth Low Energy* (BLE) are capable of reducing the latency down to $\sim 10ms$ in the best case scenario (no interference, small packet sizes) up to $140+$ ms in less optimal conditions (Tosi et al. 2017; Treurniet et al. 2015). Wired head-trackers may reduce latency further if a local machine is available to the node and the setup is not overly intrusive to the performative motions.

HRTF individualization is instead a difficult delicate tuning process to correctly apply and scale to many users, with previous attempts indicating that their introduction is not easily implementable (Zea 2012). New optical fitting systems partially respond to the complexity issue by approximating adapted filters from photographs or scans (Reichinger et al. 2013), this approach is yet to be tested for music performance. In terms of distributed networks, attempts have been made to use Ambisonics B-format streams for the purpose of reproducing ambience sound

of a connected node into another (Gurevich et al. 2011; Chafe et al. 2000), this approach is very attractive and interesting for loudspeaker-based reproduction at each receiving node and for optimizing the bandwidth required for transmission. However, loudspeaker setups can potentially lead to signal feedback issues if not properly tuned, while Ambisonics headphone reproduction still necessitates the use of HRTFs for binaural rendering.

4 The Immersive Experience

The work conducted during this dissertation concerns the combination of immersive audio technology and distributed networks, which are studied to inform the development of collaborative virtual or mixed-reality musical experiences of various kinds. The intersection of these fields is growing and several questions have not yet been answered by research. The main thread linking the relevant literature concerns the implementation of plausible spatial audio environments within distributed music applications and the research of methods that can be used to evaluate their contextual success. To this end, there are a few key studies that provided the inspiration for the studies presented later in the document.

4.1 Dimensions of "Immersion"

In (Lee 2020) the concept of *"Immersive Experience"* is proposed as a multidimensional model (Fig. 3), formed at a high level by subjective constructs of "presence" (physical, sensorial, and cognitive), and "involvement" in a narrative or in an application task (for example, the task of collaborating on a piece of music). Each dimension of these internalized constructs, alone or in combination, can help build the sensation of "immersion" into users of an immersive system. The more of these can be elicited during a display, the more immersive a system can be rated as such.

The elicitation of these constructs can be affected by technical factors like display accuracy and degree of interactivity of a system, and by confounding factors related to the user, such as the user's own reference experience, degree of skill, and preference. A distinction is made between "perceptual" factors, mainly a result of the quality of a system in creating plausible "realism" (itself subdivided into social and perceptual realism), and higher-level "cognitive" factors which respond to the integration and interpretation of perceptual stimuli within the contextual activity

undertaken within the system (Eaton and Lee 2019). Analogously, self-presence responds primarily to sensorial inputs, while social presence responds to contextual high-level interactions between senses and tasks.

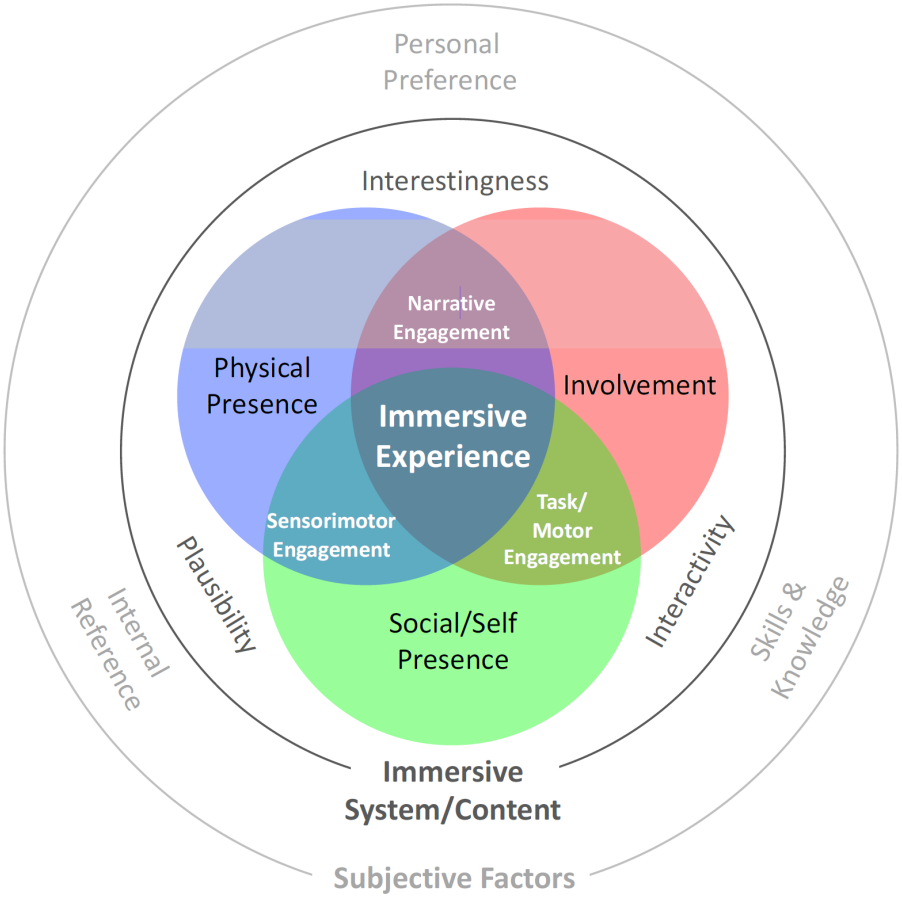


Figure 3: Lee's conceptual model of "Immersive Experience" from (Lee 2020). *Permission obtained from the original author*.

The implication of this multilevel model is that "immersion" is a function of several interacting factors, of which some can be predicted based on the technical performance and engagement quality of a system and its contents, and some are latent and dependent on the system user's bias. Looking at "social presence", it is hard to define a causal relationship with "immersion" since no direct measurement is possible. Different models mention different levels of interdependence hypothesized between various levels of "presence" and "immersion"

(Lombard et al. 2009). However, this specific framework model places "immersion" at a level above "presence", implying that the former is in theory caused by the latter.

This framework was crucial for the formulation of the experiment questionnaires presented in Ch. VI. The questionnaires were designed to capture different subjective dimensions of immersion quality and control for bias. Social presence, or "copresence" in particular is framed as a potentially measurable scale that may or may not correlate with task success within a system (Zahorik and Jenison 1998) and act as a link between the concepts of "immersive quality" and "musical outcome".

4.2 Multimodal Displays

The multimodal nature of human perception has often been found to influence quality ratings such as the sense of naturalness and plausibility of spatial audio (Begault and Trejo 2000). Dummy silent speakers, placed within a listener's field of view, have often been found to be crucial in activating a sense of *externalization* of auditory events when heard through headphones, as they are perceived as likely source emitters (Lindau and Weinzierl 2012). A related subjective dimension is that of *expectation*. "Auditory expectation" is a complex psychological construct, partly created by the current auditory experience of the present environment and partly by what our visual senses tell our brain about what sound should sound like (Valente and Braasch 2010; Blauert 1997), drawing from personal long- and short-term cognitive memory of similar spaces.

The impact of visual elements on distributed music-making has been an extensive object of research, albeit more under a musical engagement lens than an immersive experience lens. There is extensive evidence in music cognition research that players often rely on visual feedback for synchronization purposes (Bishop and Goebl 2015). Video transmission integration systems for the dual purpose of telepresence and synchronization aid have been part of numerous NMP performances (Olmos et al. 2009) but they suffer from large overheads in video stream latency and resources. The use of digital avatars can partially address that problem by instead requiring the transmission of low-bandwidth geometrical-point data which is used by a receiving system to render a virtual representation of the interacting body. Avatars are relatively new to the NMP field. An optical tracking system for music was first proposed in (Paradiso and Sparacino 1997) using laser-based hardware to track a music conductor and create abstract impersonations of a gestural

performer. In (Schroeder et al. 2007) the experimenters developed abstract non-humanoid visual avatars representing the haptic gesture of each connected musician, a relationship was found between the nature of the musical task and the perceived usefulness of the visual link. Scored pieces required the musician's attention to the instrument rather than the screen; however, improvisation-based pieces did indeed register high levels of glance behavior to the video screen. This suggests that visual feedback may be more useful in certain types of musical tasks, such as those that involve improvisation, compared to scored pieces where musicians need to focus more on their instruments. This dual type of response indicates that a musician's focus switches between self and co-performer, something that could be potentially addressed by non-obstructive visual links between connected nodes. Some work in this direction has been initiated with the use of projections in curated locations (Hupke et al. 2022). Transparent AR headsets would be an ideal future solution as they are linked to being the most appropriate display form for eliciting "presence" (Shu et al. 2019), but computational overhead and inherent latencies remain a challenge for traditional music performance. However, new promising research efforts and communities are being established (Turchet et al. 2018; Turchet et al. 2020) that can be expected to lead to innovations in the field.

4.3 Impact of Room Acoustics on Immersive Quality

In mixed-reality applications, information about the listening space can be used to adapt the acoustic character of a digital signal to better meet the expectations of a listener. This can be done by accurately simulating the local behavior of sound reflections and room reverberation through a room acoustics model (Kuttruff 2014) or by using virtualization techniques based on local acoustics measurements. The application of such processes has been shown to improve the auditory experience and spatial perception of virtual sound sources, both at a perceptual level in ratings such as "externalization" (Werner et al. 2016), and at a cognitive integration level for ratings such as "plausibility" (Thery et al. 2017).

A key study for the formation of this dissertation is that of Farner et al. (Farner et al. 2009), who investigate the effects of reverb in NMPs. A hand-clapping ensemble was subjected to different degrees of latencies and different types of BRIR-based auralizations. The clapping duo was first recorded when performing physically together in a real reverberant room, then

separated and recorded over a distributed network with artificial latency under anechoic and reverberant conditions (generalized BRIRs). Tempo-based metrics and a three-level judgment scale were used ("Good", "OK", "Bad"). Anechoic conditions were found to increase the rate of imprecision, indicating the positive effects of reverb over precision-based quality metrics. A side effect of using BRIR reverb was a lower *initial tempo*. No differences in subjective judgment were found. The study did not differentiate between congruent and non-congruent reverberation curves, raising the question of whether the "room divergence effect" assumes a latent role in performance. Moreover, there is room for exploration of different types of correlations between applicable metrics, for example, by evaluating social presence and immersion instead of general experience valence. A similar study (Carôt et al. 2009), investigating the effects of artificial reverb as a factor of mitigation of detrimental latency effects, had different observations, determining that the amount of reverb was found to be inconsequential to latency tolerance and not a preferred playing environment by musicians. However, the auralization intervention was only applied at delays already considered "intolerable" and it is unclear how it may have affected performance within the tolerable range.

4.3.1 Room Divergence Effect

An important phenomenon relevant to the field of mixed and augmented reality is the "room divergence effect" (Werner et al. 2016). The effect regards the judgment of virtual source "externalization" displayed through auralizations that are acoustically divergent from the local visual environment of a listener (non-congruent). Higher degrees of divergence are inversely correlated with the degree of externalization reported, which means that if the reference room used for auralization does not acoustically match an internalized "expectation" of acoustics and reverb, the spatial audio image degrades. On the contrary, auralizations through parameters designed to match the local room character worked positively towards the stability of the externalized image. The auditory expectation of a listener is affected by environmental factors and experience memory, but interestingly it can be overcome over time (Klein et al. 2017). The implication is that listeners training their externalization image on non-congruent auralizations can over time adapt to the contrasting visual factor. This effect was observed independently of whether the auralization was individualized (personal HRTFs) or not, and also independently

of the room of physical presence. The literature on this effect can be used to identify possible perceptual challenges that may occur when looking to create immersive audio experiences at nodes that diverge from the auralization settings.

The accurate tuning of a congruent auralization environment is a difficult challenge to solve. The high costs and engineering effort of collecting accurate acoustic measurements of a space make it difficult to flexibly apply signal adaptation over immersive systems, especially in mobile applications. Modern methods based on machine learning aim to synthesize the acoustic response of a space by dynamically extracting parametric acoustic information through electronic sensors (Eaton et al. 2015; Gamper and Tashev 2018; **Andrea Genovese** et al. 2018). These methods seek to "blindly" create a virtual environment that can acoustically match a local reproduction space in the absence of calibration measurements or prior geometric and spatial information, greatly reducing the engineering effort required. Auralization tuning through machine learning is a very promising technique that will be widely adopted in future mobile immersive systems. However, this is still a noisy process with limitations in accuracy and resolution. As of today, congruent auralizations are best achieved through measurements collected in situ.

4.3.2 Measuring Immersion

By definition (Milgram and Kishino 1994), mixed reality (MR) aims to blend the rendering and reproduction of digital, virtual media with the local present environment of a user. Unlike virtual reality, which looks to create the illusion of "being there" within a telematic medium (Steuer 1992) (i.e., *telepresence*) the goal of MR is to achieve the illusion of *copresence*. Copresence has been identified as an appropriate attribute of mixed reality, and defined as the feeling or illusion of "being together in a shared space" (Riva et al. 2003). Within the context of this work and the research areas involved, it is appropriate to instead refer to *auditory copresence*, since the visual elements are secondary to the problem to be addressed in this paper. In mixed reality systems, auditory copresence can be affected by factors such as audio reproduction methods, visual rendering, the number of users at each node, user orientation, and location within a room. In practice, all these factors have an impact on how data should be rendered and interpreted locally to maintain a cohesive perspective. By exploiting information on the geometry of the local room and tracking spatial relationships between users, boundaries, and virtual objects, it

is possible to create a *plausible* immersive experience in 6-degrees-of-freedom that can integrate real and virtual elements within the same audiovisual scene (Wagner et al. 2009).

The quest for the definition of appropriate metrics is still an object of debate in the immersive audio community (Rumsey 2002). Rating scales such as *naturalness* aims to quantify the degree of realism achieved by a spatial audio reproduction in comparison to a user's own internal reference or expectation. This is also referred to as *authenticity* (Lindau and Weinzierl 2012). The fundamental problem of this attribute is found when rated over non-natural sounds, for example, that of a synth instrument, which is never experienced in the natural world and therefore lacks a comparable reference. It is possible to decouple the "naturalness" of timbral qualities from that of spatial qualities, but also in this case the auditory expectation is heavily influenced by the visual field of the person rating the example (Kyriakakis 1998). While the task at hand, making music over the internet is inherently non-natural, it is possible that the process of matching the acoustical properties of a signal with the expectations created by the visual environment would meet the expectations of how a "synthetic sound would naturally be heard in that room" thus satisfying the perceptual requirements for plausibility, and by proxy, contribute to the feeling of immersion.

CHAPTER III

PREVIOUS WORK

This chapter summarizes a selection of the previously published research by the author that is directly relevant to the principal study described in the rest of the manuscript and conducted over time to deepen the understanding of the problem. The experience gained through the work here discussed led the way towards the formulation of the hypotheses and research questions later brought forward for the development of an empirical study aiming to uncover relationships between auralization methods, ratings of copresence, and quality of performance.

More in detail, this chapter describes prototype implementations centered on various VR/AR collaborative musical experiences, experiments looking at the use of sound directionality in network music performances, and the work that occurred towards the creation of an interactive multi-user augmented reality platform based on a network of specially dedicated rooms. This platform, called "Holodeck", was tested through two "proof-of-concept" experimental distributed concerts that helped to identify areas of challenges related to the implementation and usability of the system. The experience gathered through this work was fundamental in raising the questions and hypotheses that led to the dissertation work of Ch. IV.

1 The "Holodeck" Platform

The NYU-Holodeck project (*Holodeck - Experential Supercomputer* 2017; Plass et al. 2022) is a unique experiential supercomputing network platform that aims to virtually connect geographically remote locations using a variety of sensor arrays and XR reproduction devices. The participating laboratories are the NYU-X from Rory Meyers College of Nursing, NYU Steinhardt's Music and Audio Research Lab (MARL), NYU Courant's Future Reality Lab (FRL), NYU Tandon's MAGNET, and NYU Tisch's CREATE lab. The consortium was awarded a grant from the *National Science*

Foundation to build the platform and develop studies based on mixed-reality interactions of diverse type.

By exploiting a dedicated low-latency fiber optic infrastructure, present within the university's core infrastructure, real-time sensor data can be streamed between room nodes at ultra-low latency transmission (∼5 ms for round-trip delay). The data is parsed through a central relay server, which synchronizes various data stream types (audio, video, motion capture, haptics, etc.) and distributes them to the client nodes, which render the data according to desired configurations. This is achieved through a dedicated protocol built by associated laboratories called "CoreLink" (*NYU Corelink | Homepage* n.d.), a real-time data exchange framework capable of transferring, processing, and recording different types of data through a central network server. The framework provides an API for locally encoding and decoding various data streams at each node and for customizing the data exchange according to network speed capabilities and local rendering needs. This system effectively serves as a research platform for multi-user, multi-perspective, collaborative audiovisual interactions between remote locations. Figure 4, illustrates the star topology structure and a sketch of possible connections that may happen at a given time. Any number of nodes can connect to the relay server to send and receive data. The data is unwrapped at the receiver node and locally interpreted using space-specific information about the geometrical boundaries of the room and reproduction equipment.

The role of MARL in the project is to advise and contribute to the implementation of an audio capturing, streaming, and rendering protocol. The use of spatial audio is designed to be available for each node in its various forms, binaural stereo, ambisonics, and surround (Fig. 5. In the case of object audio sources, the data can be processed for headphones using local BRIRs and distance models, or upmixed into spherical harmonics domain and reproduced through loudspeakers using the appropriate configuration decoding parameters (Daniel et al. 2003). Similarly, soundfield audio can be transmitted between nodes to reproduce a directional environmental ambience through spatial audio rendering software.

Several engineering challenges apply at each client node, requiring the transmission and rendering of different audio-stream types (using codecs that allow to transmit object audio, multi-channel streams or higher order ambisonics (Herre et al. 2015)). Sound field data and individual mixes need to be curated for the individual needs of every system user, according to

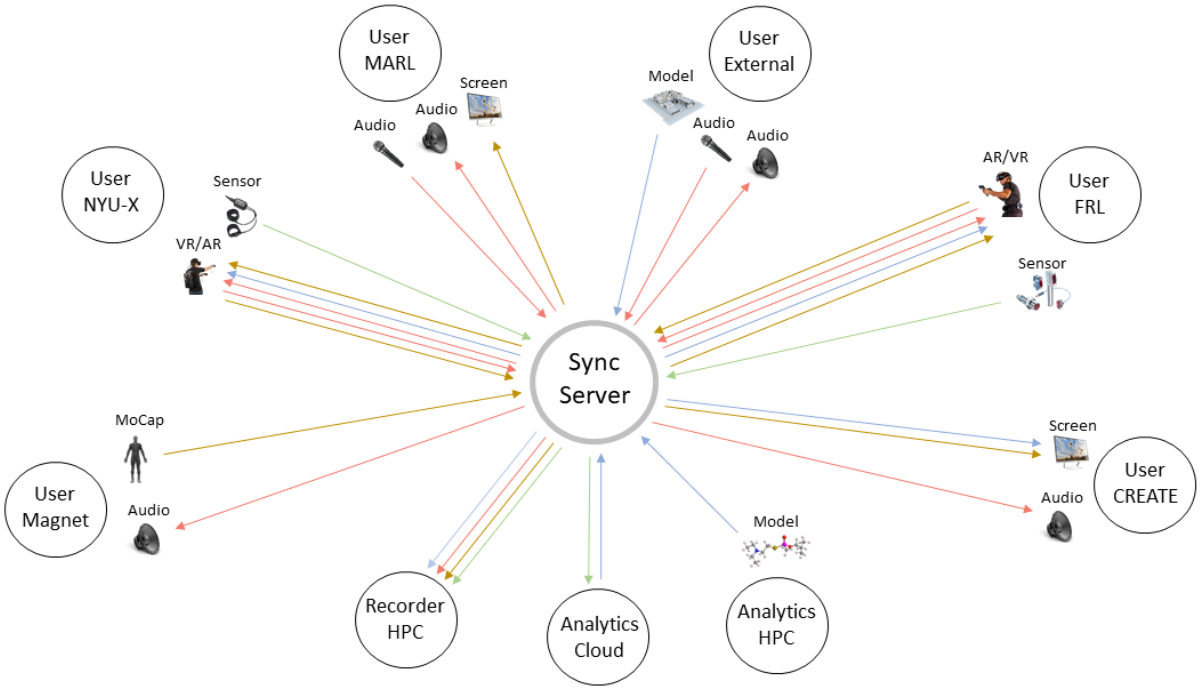


Figure 4: Concept diagram of the Holodeck network star-topology. A central server is in charge of managing low-latency synchronization, data distribution, record data and run analysis protocols. (Image from (*Holodeck - Experential Supercomputer* 2017))

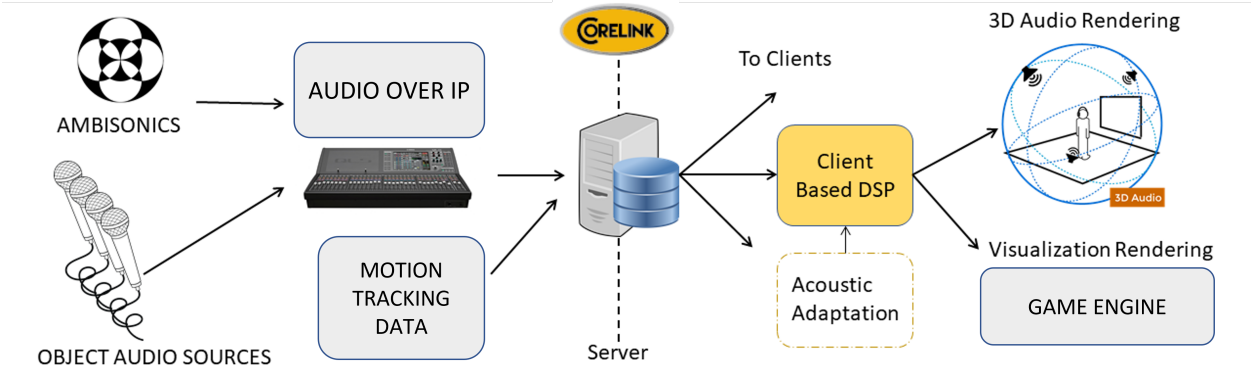


Figure 5: High-level audio connection diagram for the Holodeck audio transmission and rendering engine

their position and orientation within a room. The main challenge is to create a system which is flexible to the local needs of the spaces and the number of concurrent users-per-node, while maintaining good streaming rates and realistic rendering, avoiding sound feedback, unwanted coloration, and signal bleed.

Through the lens of the Music Technology field, the Holodeck provides an attractive infrastructure for the study of distributed augmented musical performance; musicians can be virtually brought together through audiovisual channels consisting of audio, video, and motion captured digital avatars. A long-term goal for this type of musical interaction is for distantly located musicians to be able to connect and perform "as if they were in the same room".

1.1 Concerts on the Holodeck: First Pilot

Beyond the challenge of the technical implementation of the platform, the role of each lab involved in the project was to prototype proof-of-concept applications capable of demonstrating the functioning of the system. In regards to audio-based applications, studies related to acoustics, audio, music, perception, and mixed reality were piloted by combining the experience gathered in musician's motion-capture studies and XR audio experience design.

The conjugation of the Holodeck project with distributed music materialized as a way to both test the early implementations of the platform and to explore the artistic space available for mixed-reality performances. The NYU Steinhardt Music Technology program has been involved for years in the topic of distributed music. Early collaborative performances were conducted by Prof. J. Gilbert (Ghezzo et al. n.d.) and continued under Prof. T. Beyer (Beyer 2016). These ongoing academic efforts created the right environment for large-scale distributed performance projects that raised the interests of several collaborators within the department. Thanks to these collaborations, the "Holodeck distributed concert series" came to life, with two pilot concert events that envisioned a series of artistic pieces showcasing the Holodeck functionality [1] [2].

A first pilot concert was held in April 2018, *"Concert on the Holodeck: Connecting Artists"* involving distributed music and dance [3]. This first iteration served to demonstrate the concept of

[1] First Holodeck concert https://wp.nyu.edu/immersiveaudiogroup/2018/04/19/Holodeck1

[2] Second Holodeck concert https://wp.nyu.edu/immersiveaudiogroup/2018/10/10/Holodeck2

[3] Video footage: https://www.youtube.com/watch?v=uTpXCKyWIqY

the various multimodal real-time elements involved. Testing each element as a separate entity. The setup involved two nodes (within the same building) organized hierarchically, once "concert" node where the audience and stage performers were present, and one remote node where parts of the musician ensemble and motion-captured dancers were located (Fig 6). Using the local Ethernet infrastructure, audio was streamed across nodes, where a mixing console created a dedicated mix for each musician's audio monitoring system. A visual connection between performers was created using a local video link, allowing musicians in the studio to see the stage via monitors, and stage musicians to see the studio via projectors in the concert room. A motion capture system was set up in the studio to capture the performance of two dancers (wearing a tracking suit) who reacted to the music. The system captured the digital skeleton points and linked them to a game engine software on a local machine, which rendered the performers as digital avatars. The rendered output was streamed to the concert room through a video link and transmitted on a projector. Care was taken to ensure that each distributed performer was able to see and hear the people at the opposite node. The setup permitted for very low transmission latencies and the application of classical, jazz and percussive music. The event was also broadcast over the Internet in both regular and 360 video formats.

No empirical data was collected for this event; however, post-concert informal interviews with a sample of the people involved revealed some of the technical and artistic challenges. It was suggested that the ability to rehearse was key fundamental to some performers, as it allowed one to gain more familiarity and comfort of performing in the physical absence of the musical partner(s). Others pointed out to the asymmetry of the experience as playing from the studio felt more clinical and less involved, but would also create less performance anxiety. Additionally, the fact that musicians used an earpiece monitor where the routed signals were dry, was reported to prevent the feeling of being "immersed into a cohesive sonic environment". This feedback helps to hypothesize that the quality of experience of a participant in a distributed concert could improve through training and through methods treating auditory cohesion.

1.2 Concerts on the Holodeck: Second Pilot

The second iteration of the concert experience took place in October 2018, titled: *"Ozark Henry on the Holodeck: Maps to the Stars"* (AES 2018) a seven-piece program involving musical instruments,

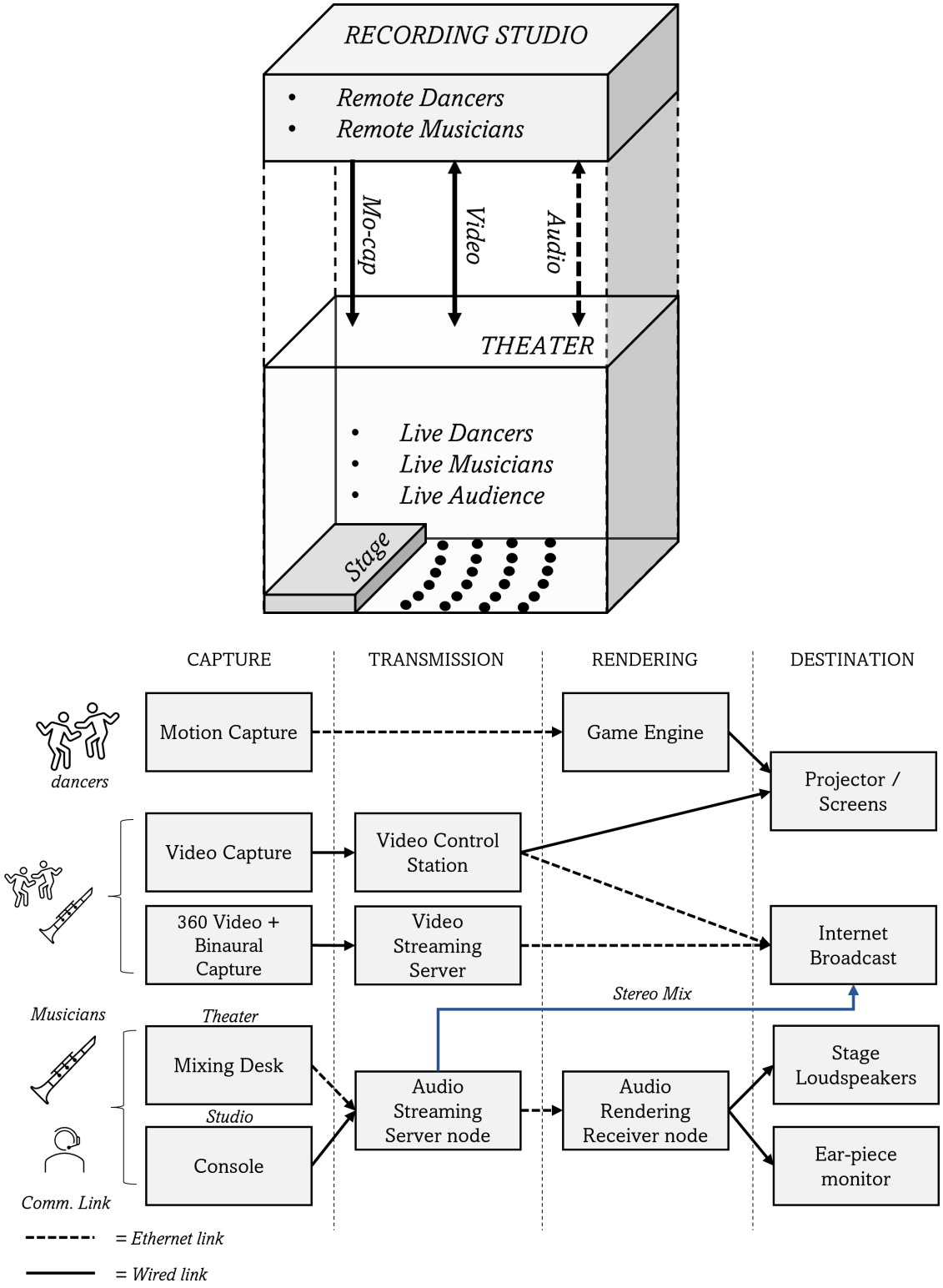


Figure 6: Organizational setup and high-level signal flow for the first “Holodeck” concert

choirs and dancers showcased during *AES NYC Convention*. The goals for this second pilot event were to introduce an alternative setup across nodes that were geographically distant. The experience was designed to involve the same local nodes of the first pilot, plus a node located on the university infrastructure network and one located overseas (Fig. 7). The CoreLink relay server was used to transmit dancers' motion capture data from the remote node on the inter-lab network infrastructure. Using a data wrapper script, the server code broadcasted data through the star-topology network and a listener node placed at NYU Steinhardt was able to unwrap data, parse it to a game engine, and render it as a 3D digital avatar scene to be streamed in the main theater. This test demonstrated the server ability to collect, wrap, and parse data to any listener node.

The music ensemble was divided between three locations; the theater node (musicians and choir), the studio node (choir), and the overseas node (musicians). A one-way connection was established between the choir in the studio facilities and the theater through an MPEG-H encoder[1] that could embed positional channel metadata for a spatialized reproduction at the destination node. The choir was captured through a soundfield microphone, which was passed to the encoder, streamed via Ethernet, and decoded at the destination node, where it was upmixed to the local theater PA using the metadata annexed to the stream. A two-way audio link with the overseas node (*Norwegian University of Science and Technology (NTNU)* was established through IP-based routing software (*Jacktrip* (Cáceres and Chafe 2010)).

Due to the geographical distance, high latencies were involved with the overseas node, and the delay was also asymmetric, leading to added difficulties and possible circular drags on the musical synchronization. To address the problem, a leader-follower approach was used between the musicians (with the overseas node designated as "leader") and between the musical mix and the remote dancers. Additional flexibility in the musical and dancing genre had to be adopted in the form of latency-coping mechanisms, because several stages of signal delay were present in the theater signal loop. The stage music signal had to be sent first towards the dancer's remote location, their reacting dance motion captured, and sent back to the studio location, where it was rendered by a local machine and streamed back to the theater projectors. To cope with this

[1] The MPEG-H system was set up thanks to the direct involvement of THX Ltd. and Qualcomm Technologies who provided the tools necessary

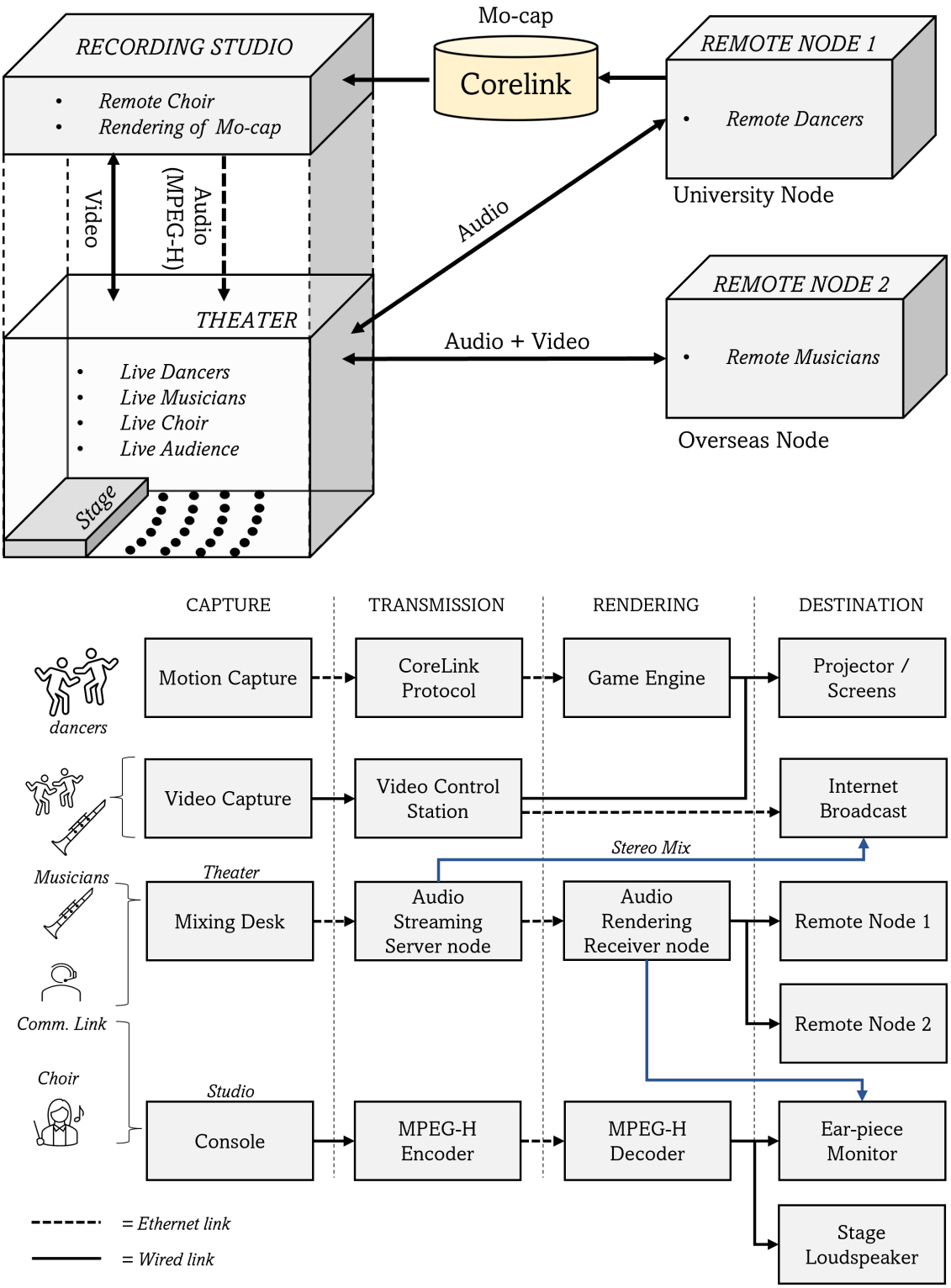


Figure 7: Organizational setup and high-level signal flow for the second “Holodeck” concert

situation and try to synchronize the beat, a "delayed-feedback" control (Carôt and Werner 2009) was implemented in the server to allow additional artificial latency to be added at the discretion of the receiving node, until a synchronization of the beat was found.

Informal subjective evaluation data were collected to capture the general audience and performer impressions and provide a baseline to compare against in future installments, two subjective evaluation questionnaires were conducted, one for the audience and one for the performers (all questionnaires are included in the appendix). The goal of these questionnaires was to conduct a qualitative investigation rather than to respond to specific hypotheses. Therefore, data collection was used to observe distributions, establish a reference baseline, and identify potential areas of problem to address in future stages.

Audience members (N = 100) responded to questions asking them to rate the quality of the audio and visual outcome of the experience, the cohesiveness of the musical and dance artistic components between the stage and the remote nodes, the level of "presence" they felt from the reproduction of the remote choir, and the overall rating of the event as a musical concert. Results, shown in Fig. 8 and 9 show that the visual component was not as cohesive or impactful as the auditory component (to an audience of mostly audio experts); however, it is not known if the lower ratings were due to the artistic quality of the choreography, the game scene artistic style, or due to signal latencies above noticeable thresholds (free-form feedback suggested a mix of all three). The question about the choir regarded the spatial rendering through the theater's PA, the distribution is skewed towards higher values, indicating that no major artifacts were created by the routing system. Overall, the audio and the experience were rated fairly high. Currently, these ratings do not have a reference to compare them with. However, they can serve as a baseline evaluation benchmark for future iterations of the concert.

Regarding the performer questionnaire (N = 18), Likert-type questions were used to poll general impressions among musicians and dancers. Figures 10 and 11 show that for this setup, latency was not perceived as a strong impact factor (although the musicians at the overseas node were not polled, so crucial data is here missing). Responses to a "presence" question showed high variance of opinion, as did responses concerning a question about the in-ear monitoring system (stereo or mono mix) investigating whether it affected the possibility of immersion. Most of the artists reported in general that the experience was pleasant. Other free-form feedback

revealed that performers in the studio felt more disjointed than performers on stage and that the opportunity to rehearse was crucial for the success and comfort of the distributed connection and musical approach.

As a result of these experiences, several areas of work were identified with the objective of understanding the subjective experience of a musician within a distributed system, the role of immersion, and its practical impact on performance. The observation that the experience was rated differently between people in the studio against people at the stage led to questions of room effect on presence and immersion. Asymmetric auralization at each node was discussed as a possible system to allow performers to feel more immersed in a concert experience. These considerations regarding the perspective of the involved performers formed the basis for the study designed and discussed from Ch. IV onward.

2 Mixed Reality and Distributed Performance

One of the data types handled by the Holodeck distribution system is body motion-capture data. Motion capture (also known as mocap) involves using specialized equipment such as sensors, cameras, and markers to capture the motion of a performer or an object and then translate that data into a 3D representation in a computer program. Potentially, this is a powerful tool for real-time mixed reality applications, allowing the embodiment of users into digital avatars able to interact within social collaborative virtual interactions. Furthermore, the data rates required for mocap can be lower than video streams (depending on frame rate and resolution), with the trade-off of computational resources needed at the destination node for the 3D rendering. In motion capture, a performer wears a special suit with markers that are tracked by cameras or sensors. The cameras or sensors capture the movement of the performer's body in real-time, and this data is used to create a 3D model of the performer's movements.

2.1 Motion Capture of Artists and Musicians

To learn about the practical challenges of motion capture as applied to musical performers, various tests were made internally involving different scenarios ranging from individual musicians and dancers to medium-sized ensembles, captured through soundfield arrays (e.g.

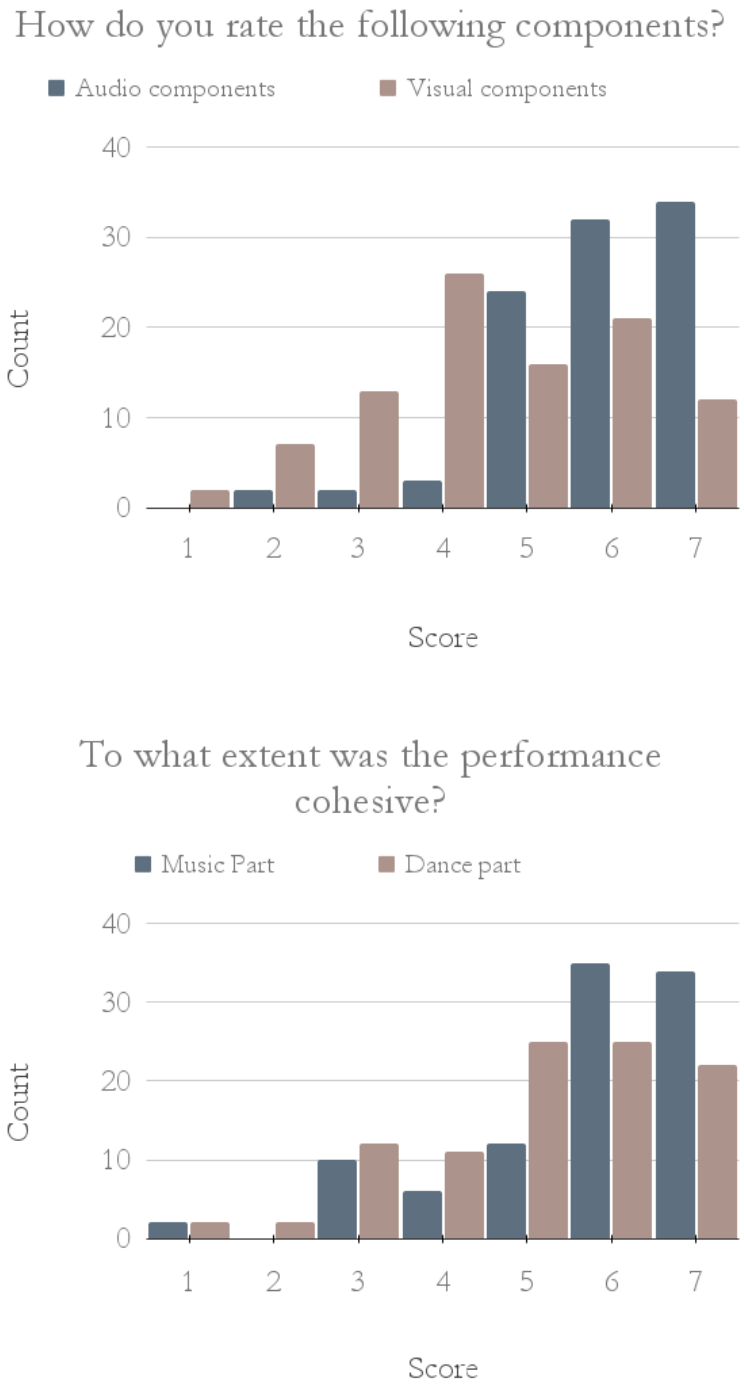


Figure 8: Distributions of audience scores for rating the quality and cohesiveness of the audio (music) and visual (dancers) components, collected during the second "Holodeck" concert

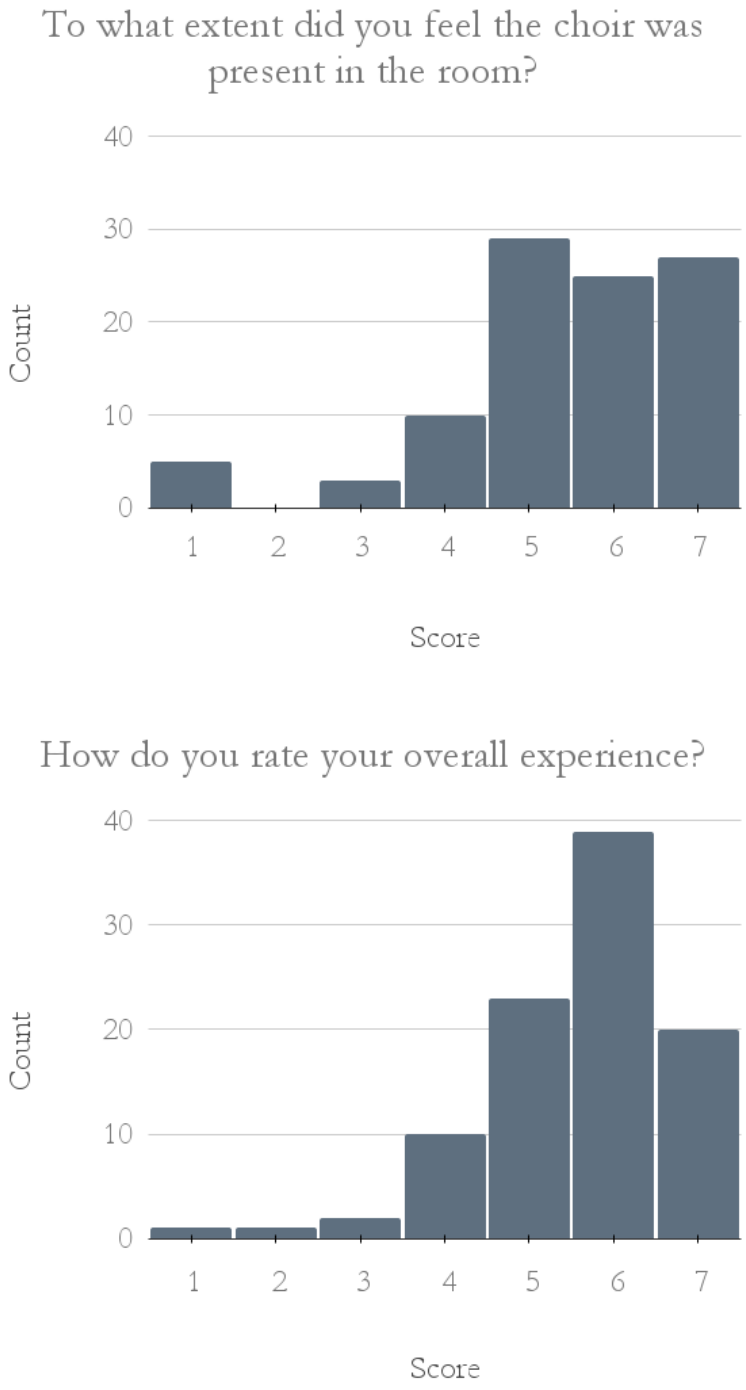
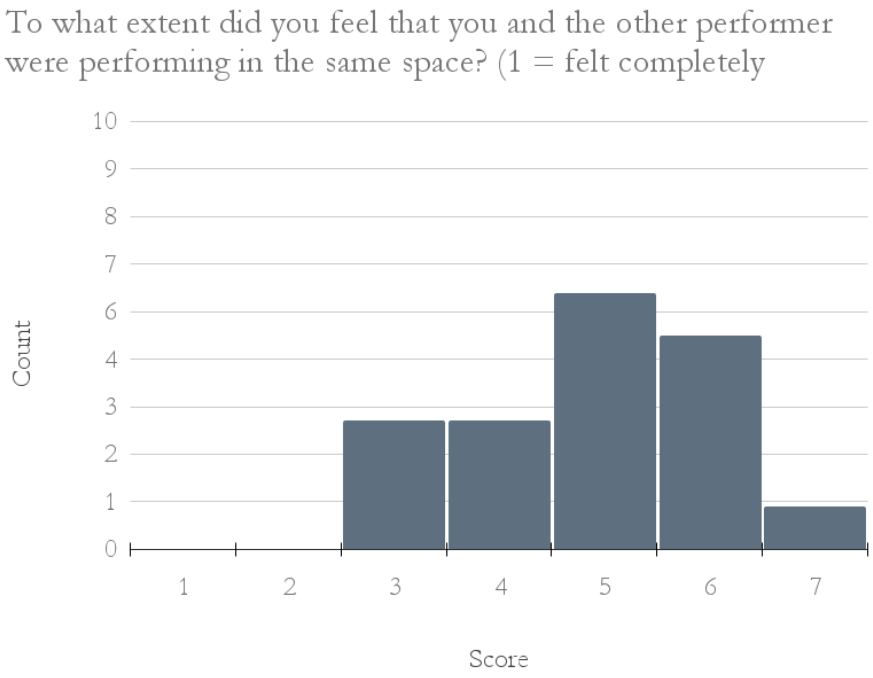


Figure 9: Distributions of audience scores rating the choir's "presence" and the overall rating of the experience



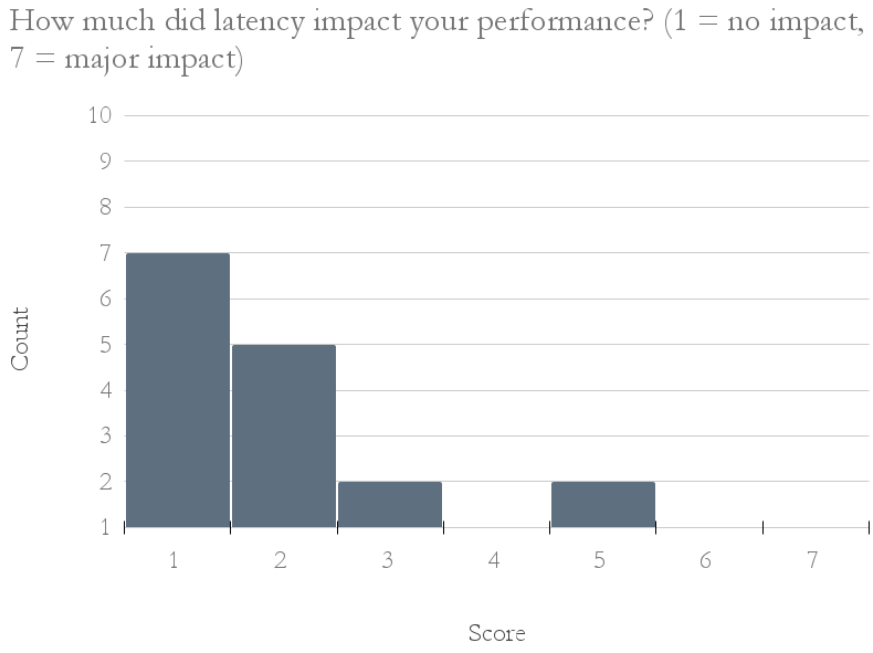


Figure 10: Distributions of performer scores for performance "presence" and "latency impact" collected during the second "Holodeck" concert

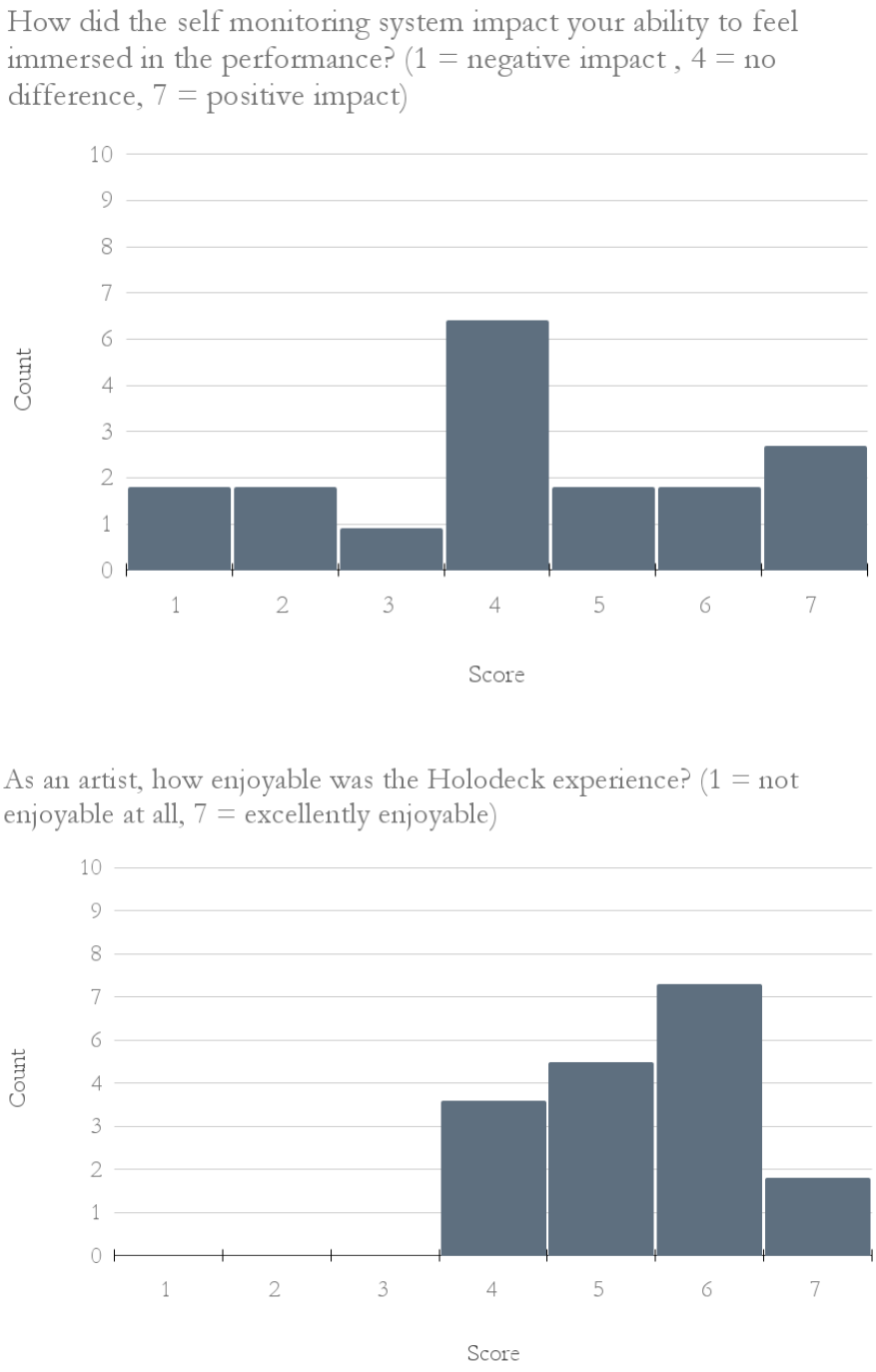


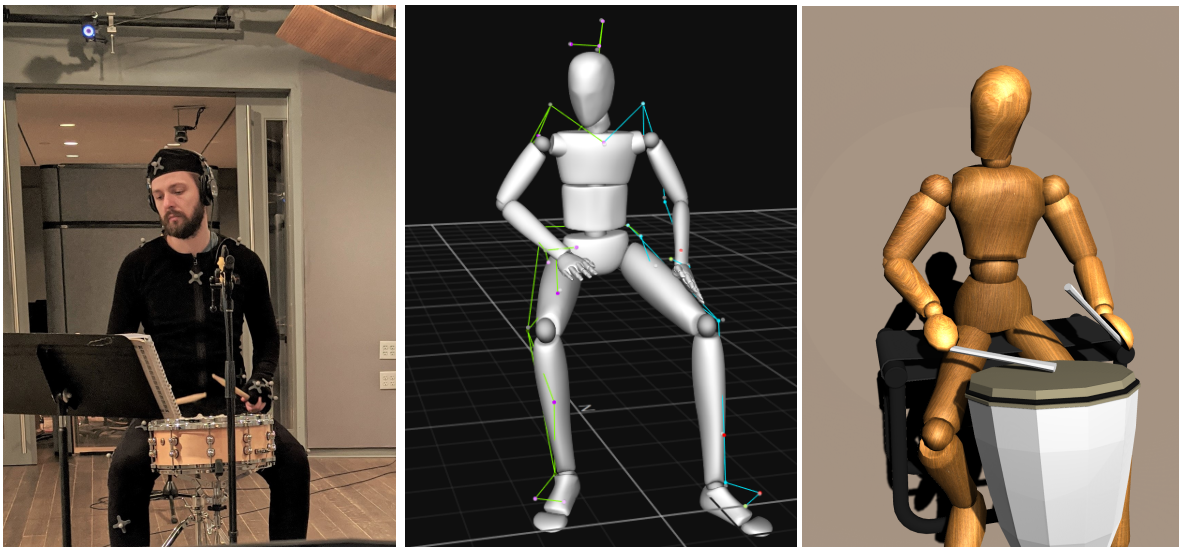
Figure 11: Distributions of performer scores for technology "immersion impact" and "enjoyment" collected during the second "Holodeck" concert

Hamasaki square, HOA spheres) and close-miking techniques. The biggest challenge encountered was that of simultaneous recording of multiple actors, necessary for the cohesiveness of the performance material. Having multiple bodies captured in a space can increase the chances of tracking errors (especially in the case of intersecting choreographies) requiring heavy data cleaning procedures not applicable to real-time data. Furthermore, shiny reflective surfaces such as microphones or instruments such as saxophones and flutes can interfere with the optical tracking methods available if the incidence of the room lighting is direct. A third limitation was that of large microphone arrays that obstruct the visual path of the cameras to the tracking suit markers. Nevertheless, having a properly calibrated light environment and by planning body motions such that points of contacts between actors were avoided, proved to be effective measures, with the skeleton capture turning out to be sufficiently stable for live applications. Data gathered during this phase of development was later used for several published projects, such as the creation of motion capture audiovisual musical drum loops (Bui et al. 2020), data wrapping and rendering tests through the Holodeck relay server (*NYU Corelink | Homepage* n.d.), AR multimedia displays, and mixed-reality rehearsal environments and concerts (**Andrea Genovese** et al. 2019b). This data can potentially be also applied for music pedagogy, 6DOF virtual experiences, or gaming. The experience gathered through this work helped to organize and plan the motion capture streams used in the pilot concert series.

2.2 Mixing Real and Virtual Sources

An early pilot exhibition test showcasing the possible usage of motion capture data in distributed performance applications was composed by mixing live and prerecorded audiovisual streams adapted to fit an exhibition space (**Andrea Genovese** et al. 2019b). In this scenario, a pseudo-live collaborative performance for a single-member audience was set up in a dedicated room and displayed through a VR headset. The ensemble consisted of a live motion-tracked percussion performer, rendered in real time as a 3D avatar, and three virtual co-performer "objects" that were pre-recorded in a studio through mo-cap and spot microphones.

The setup illustrated in Fig. 13 shows a live musician sharing the room with the audience. The virtual avatars (three musicians and dancers) were spatially located around the listener to form a virtual ensemble with the live performer. The audience was provided with transparent



(a) Mocapped performer (b) Raw mocap data (c) Rigged and rendered avatar

Figure 12: Capturing, cleaning, and rendering stages for a motion-captured snare drum performer

headphones to allow the local acoustic path and room reflections to be heard with as little obstruction as possible while the virtual sources were dynamically rendered in 3DOF binaural format. More specifically, the audio belonging to each virtual musician was composed of a double emitter object source (capture of the top and bottom parts of the percussion instrument, to preserve radiation width), captured in dry conditions and dynamically spatialized as such via HRTFs. Instead, the reverberant portion of the sound was created by auralization of the full set of channels through a diffused *reverberation fingerprint* (Jot and Lee 2016) measured in situ with an omnidirectional microphone pair at the exact location where the audience member was located during the exhibition. Through pre-processing with the diffuse room response, the pre-recorded sound material was rendered reverberant and delivered as a stereo stream that mixed in with the dynamic direct-path rendering. Early reflections were not simulated for this iteration of the experience.

Visually, the VR scene consisted of a visual recreation of the performance space (with the intention of creating an auditory "expectation" matching the sound of the live performer) and a spatial matching of the live avatar with the effective location of the live musician in regards to the

audience (Fig. 14)[1]. The combination of real elements with a calibrated virtual display falls into the "Augmented Virtuality" as defined in (Milgram and Kishino 1994).

A stereo, non-auralized, static mix of the prerecorded audio was separately provided for the live performer. The musical dynamic at play effectively mimicked a "leader-follower" scenario where the musicians at the destination node, and where the audience is also located, follow the signal coming from a remote node that is unaware of what happens at the destination. This approach ensures that the audience hears a time-aligned cohesive performance, provided that the musical material is hierarchically structured. Audience feedback was positive, with a high degree of auditory cohesion felt throughout a small set of listeners. The point of view of the musician instead revealed a different outcome. The musician's perspective was in this case treated as secondary, as no auralized spatial mix or visual environment was provided for that role. The absence of these elements was notably felt, post-event feedback revealed that the performance felt to the musician like a "one-way avenue of communication, where my job was to fit myself into this world that was created for the experience" and it "did not feel as organic as performing with other people in real-time", indicating that something was missing for creating a sense of "copresence" and "cohesion", pointing towards auralization as a desirable process for the musician as well as the audience. However, the component of having an interacting music network was bypassed in this project, so it is possible that having a pre-recorded audio base, rather than a live remote connection, latently affected the sensation of copresence.

This pilot framework served to understand more in-depth the acoustic challenges and computational resources needed for VR live-music experiences and inform future implementations of formalized empirical studies on mixed-reality and distributed performance that investigate the technical and cognitive aspects which regulate the subjective quality of experience from each role's perspective. The calibration cost associated with tuning the experience to a particular space or perspective is considerable, so it is important to understand how "success" is defined in a virtual experience according to a desired "target user" (performer or audience) for which the experience is tailored. A possible hypothesis may be that the improvement of the subjective experience of a musician could translate to the improvement of

[1] A video of the exhibition rehearsal is available at https://www.youtube.com/watch?v=-0VqIn1pTA0.

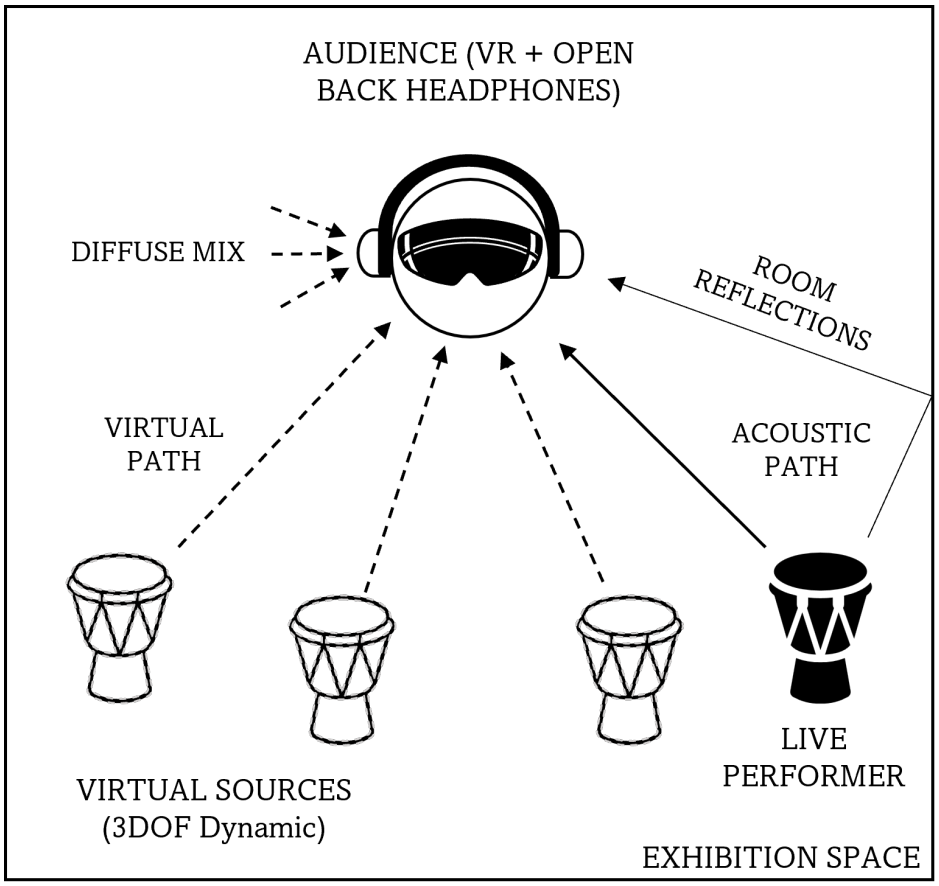


Figure 13: Design for the spatial arrangement of participants during the exhibition phase.

Figure 14: Exhibition trial. The point-of-view of the audience is shown in the background picture, while the overlayed smaller picture illustrates the external view of the live musician and the audience, seen from the experimenter.

the experience for the audience or to a general increase in musical performance. While both perspectives might tie quality to their sense of presence into the scene, the musician might seek something more keen to an intersection of "copresence" and "naturalness", as in the sense of "being performing together" to the fellow performer, in a setting comparable to real life. The evaluations from the two perspectives may or may not correlate. By establishing links between perspectives, it may be possible to determine whether the experience design requirements should be reduced or expanded, together with the implementation costs.

3 Collaborative Studies in Distributed Music

In parallel to Holodeck development, distributed music studies were conducted to learn more about the application of virtual environments and technological improvements to collaborative music networks. This series of studies was initiated as an academic collaboration between NYU and *Leibniz Universität Hannover* (LUH), located in Germany. The collaboration involved the planning of a geographically distant multimedia collaboration network. The network, depicted in Fig. 15 is characterized by the use of globally synchronous GPS timestamp data to generate a local click signal at both ends of a connection, representing a global time reference metronome simultaneously at each node (Hupke et al. 2019a). The GPS-based metronome is primarily intended to serve as a rhythmic synchronization tool, or performance conductor.

3.1 Latency Measurement Methodology

A first point of interest between the two universities was the establishment of a method able to measure the latency between geographically distant nodes, for the purposes of evaluating the feasible potential musical approaches that can be applied. The approach tested was based on the previously created GPS metronome system developed by LUH (Hupke et al. 2019a). Although originally meant for performance conduction, the method was found to be suitable for measuring one-way and round-trip delay times in NMPs.

The latency measurement methodology was set up by recording each node's own click, and the remote node click signal sent via a UDP-based streaming software (*Jackrouter*, Cáceres and Chafe 2010) using the transmission and recording setup illustrated in Fig. 16. The differences

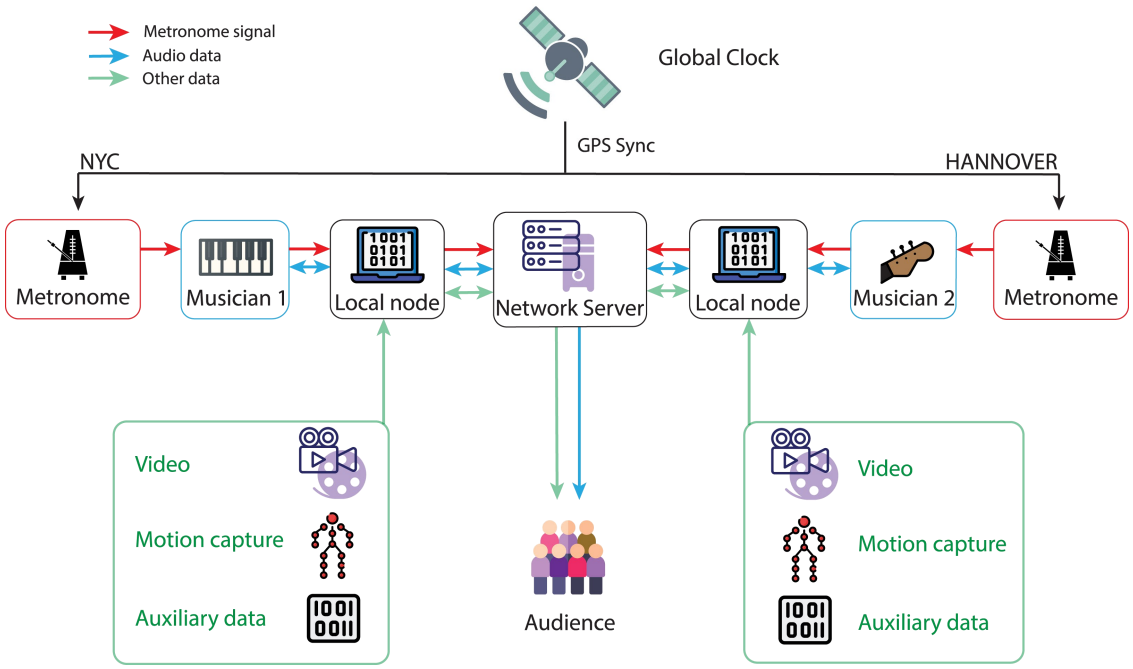


Figure 15: NYU-LUH Networked Music Performance Framework. Image from (Hupke et al. 2020)

between the two signals provide a comprehensive one-way delay between one node to the other with high precision. The dual-directionality of the setup allows to account for asymmetries in the one-way delay occurring according to differences in the quality of the equipment used at each node, CPU load, and the number of server "hops" that the data needs to go through when traveling through an IP-based link. Furthermore, the continuous stimulus used in the methodology allows to gather data for calculating the statistical distribution of jitter in order to evaluate the latency variance, indicating the stability of each connection path.

Several tests determined that buffer size settings and interface choice were the most influential factors in signal latency (Fig. 17, while the number of channels and the sample rate were found not to have an impact. In this particular case, the one-way latencies ranged from $\sim 45ms$ to $\sim 75ms$. The method was also able to capture equipment-dependent variability, caused by differences in internal sampling clocks; for example, one tested soundcard interface showed a click generation standard deviation of $\sim 56\mu s$. The equipment-related measure was achieved by local-network applications of the same setup that revealed significant differences in jitter across interface models. More details are published in (Hupke et al. 2019b).

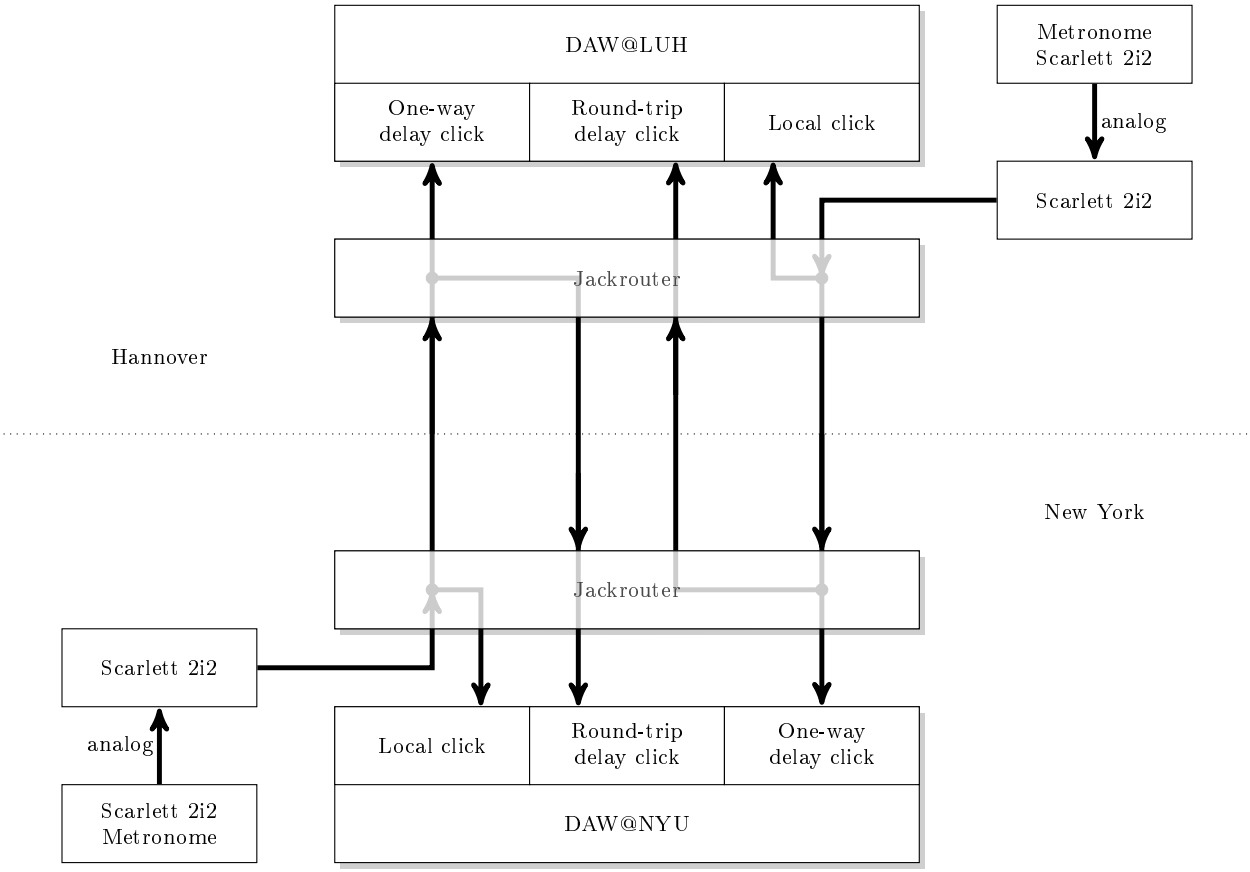


Figure 16: Latency measurement setup between NYU and LUH.

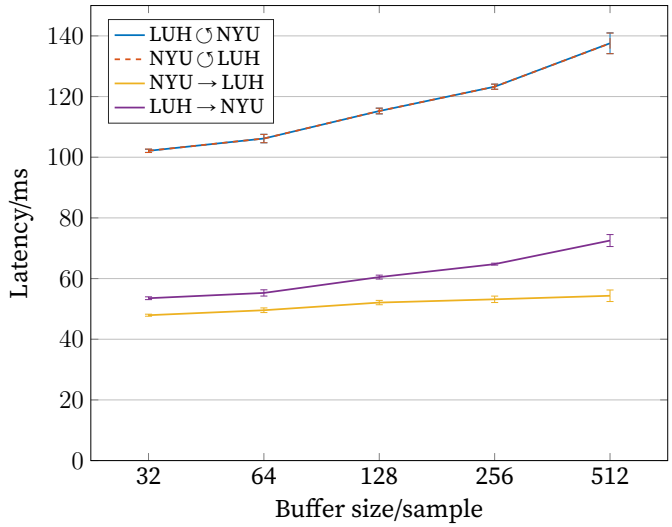


Figure 17: Measured round-trip latencies (LUH ↺ NYU, NYU ↺ LUH) and one-way latencies (NYU → LUH, LUH → NYU) for different buffer sizes.

3.2 Study on Metronome and Source Panning Interaction

Further interest in the effect of the global metronome in scenarios of distributed rhythmical performance led to the design of a study based on an ecologically viable "Realistic interaction approach" (Carôt and Werner 2009) focused on exploring the interactions of using the global metronome technology with stereo displays and their combined effects on performance. Spatial source separation of incoming streams has previously been indicated as a factor capable of reducing the cognitive load of a listener and improving auditory segregation of an auditory scene (Bregman 1994; Jung et al. 2000). The spatial separation of a performer's own monitor signal, from the coperformer stream, and potentially a metronome source, was thus hypothesized to aid the musicians' performance in directing the cognitive attention where necessary, possibly helping the accuracy of the musical outcome in the presence of latency and improve the immersive character of the experience.

A study was carried out to assess both the objective musical results and the subjective impressions of implementing these elements in network collaborations consisting of Djembe percussion duet pairs, under varying performance tempo and latency circumstances. A star-topology laboratory model network was created by setting up a local analog-based connection between two nodes and a central distribution node where an experimenter was able to control the precise degree of latency by injecting additional artificial delay in the stream. Through this system, the central node could activate source panning mixes for creating a stereo display reproduction at each node (over headphone playback) and/or enable global metronome signals. Participants were tasked with performing a 20 second beat sequence (4/4 beat) at tempos of either 90 or 120 BPM (Fig. 18), under representative one-way latency conditions of 10ms, 25ms, 50ms, and 100ms. Performances were repeated with different combinations of global metronome and source panning.

The objective results measured in metrics of tempo stability and synchronization (Rottondi et al. 2016) showed that the performances benefited significantly from the introduction of the global metronome when there was high latency (Fig: 19). There was an indication of a possible interaction of base tempo with the metronome as faster tempos were observed not to benefit from the introduction of the metronome as much as slower tempos, although a variance reduction was observed in both cases. In terms of performance pacing, mid-latency levels showed acceleration

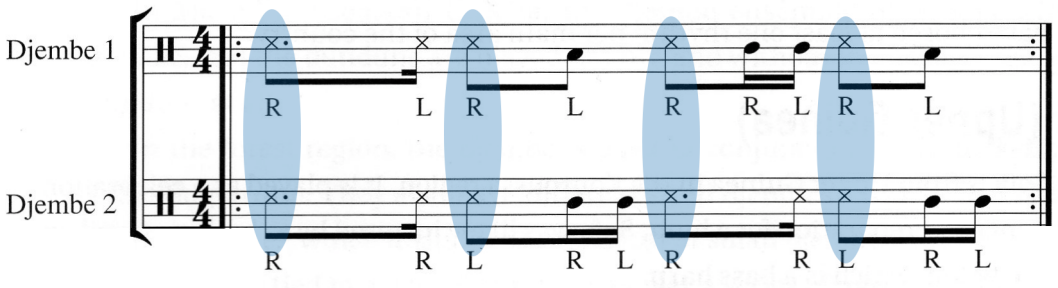
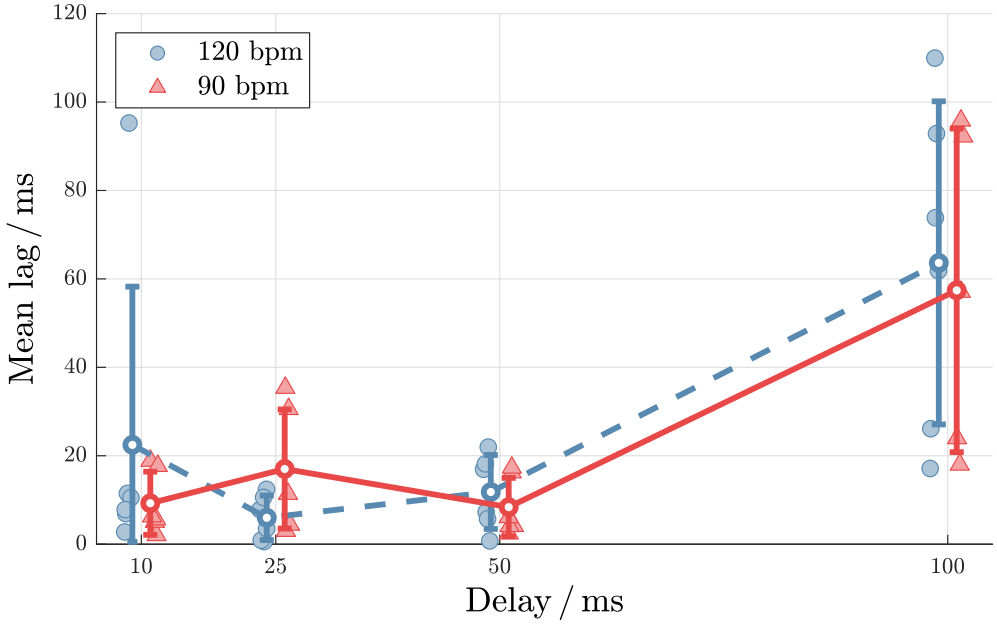


Figure 18: Rhythmic patterns used for the two Djembe performers in the "metronome and panning interaction" experiment. The synchronization onsets (blue highlights) are used to determine the objective beat tempo.

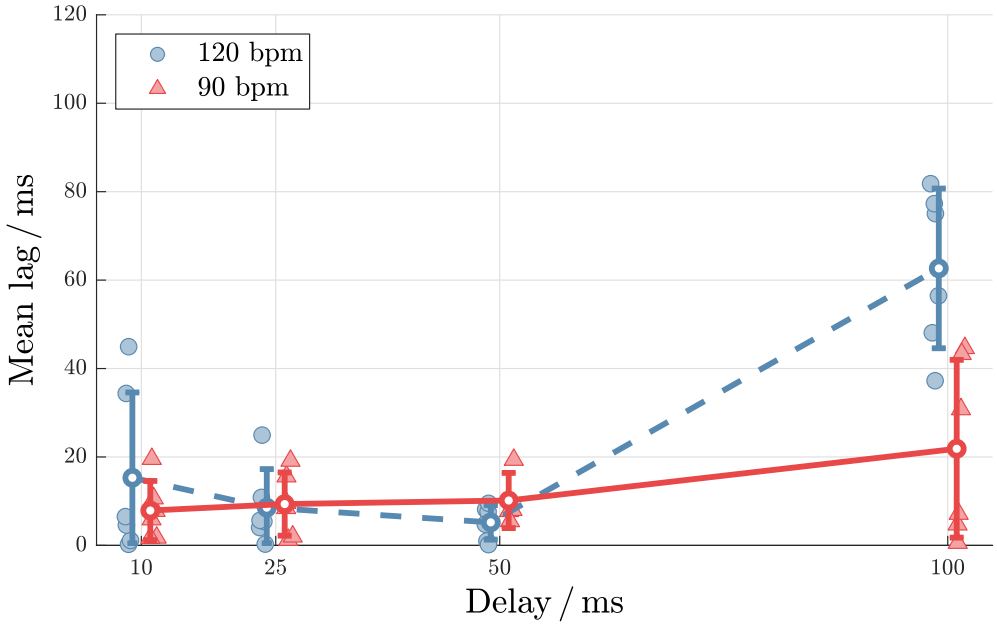
trends, while high-level latencies led to salient decelerations. The introduction of source panning was not found to have an effect in either direction on the objective metrics.

The subjective layer of evaluation (questionnaires, 5-point likert scales) showed clear trends of lower quality being associated with higher delays (with ratings of *interplay quality*, *auditory segregation*, and *difficulty*) but without an effect of performance tempo. Unlike objective observations, the metronome was not considered to have a perceivable effect on synchronization (Fig. 20). Source separation was rated as more impactful at lower latencies than at high latencies, suggesting an interaction effect in which the degradations brought on by latency overpowered the subjective improvements brought by the panning effect. In addition, the combination of panning with the metronome was rated higher than panning alone. Interestingly, some of the subjective results in regard to the effect of the metronome and source-separation were not mirroring the impact observed in the objective results. In these regards, the subjective and objective layers of evaluation were not always in agreement, indicating that the subjective experience in NMP does not always correspond to the observations of an objective analysis of the performance.

This experiment served as a preliminary step towards the thesis study and set some of the expectations in regards to the impact of latency and the correlations between objective and subjective metrics, providing a background over which to study more in depth the effects of different dimensions of evaluation concerning presence and the introduction of auralization effects. The study was published in (Hupke et al. 2020).

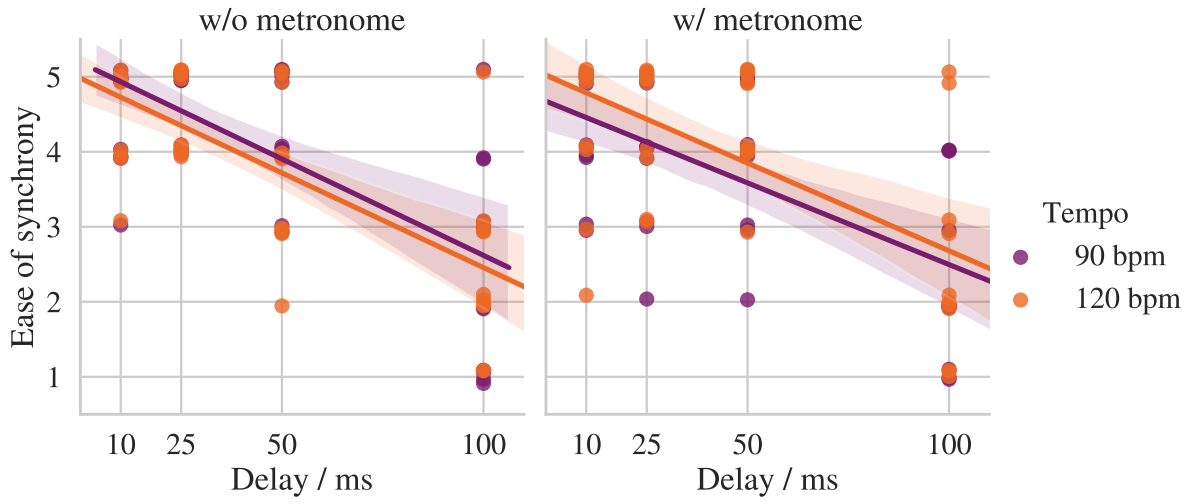


(a) Mean Lag measured across latencies w/o metronome

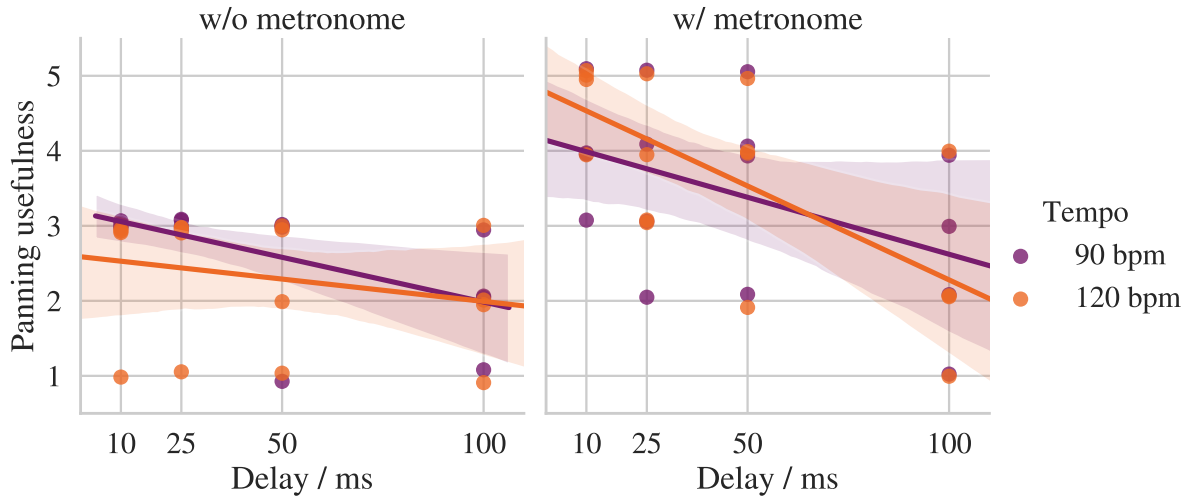


(b) Mean Lag measured across latencies w/ metronome

Figure 19: Mean lag for measured with and without metronome. The error bars show the mean and standard deviation of both predefined tempos (90 bpm and 120 bpm). Actual values are separated for both tempos (circle, triangle).



(a) Rated synchronization difficulty



(b) Rated source-separation usefulness

Figure 20: Questionnaire responses rating the “Ease of synchronization” (w/ and w/o metronome) and the “usefulness of the source panning” (w/ and w/o panning effect).

CHAPTER IV

INVESTIGATING LATENCY, AURALIZATION, AND COPRESENCE IN NMPS: OVERVIEW AND DESIGN

This chapter begins the part of the dissertation that concerns a previously unpublished empirical study designed and conducted during the course of the author's doctoral program. The experience gathered through the "Holodeck" project brought to light a series of interesting combinations of distributed collaborative networks intersecting different room types and purposes. While the concerts were designed to provide a cohesive experience for an audience, the perspective of the musicians had yet to be fully explored.

The chapters IV to VIII comprise the various phases of the experiment. The study concerns the impact of signal latency and auralization schemes; in relation to different types of quality metrics and the elicitation of latent psychological constructs related to auditory "copresence", in immersive *Network Music Performance* (NMP) settings. This chapter covers the conception and design of the study and formulates research hypotheses and case study design. In Ch. V, the measurement of the acoustic data required for the implementation of the auralization methods and the configuration of the distributed interaction network are illustrated. Ch. VI lays out the methodology applied for the collection of primary data regarding the study (audio recordings of distributed performances), and the subsequent steps applied to obtain secondary data consisting of three different quantitative evaluation layers. The evaluation layers consist of subjective responses from participants, ratings and annotations from third-party experts, and objective analysis metrics. The analysis portrayed in Ch. VII concerns this secondary data and shows the results of a Mixed-Effects Models framework over the dependent variables at hand. The outcomes of the experiment are discussed in Ch. VIII and put in relation to the initial research questions that indicate areas for future work and the value of the work in the larger context.

The study presented over the course of these chapters has been conducted under

IRB-FY2020-3945 (*"Impact of acoustic character manipulation on distributed music interactions"*) using data collected in Fall 2021.

1 Overview

An overview of the study is presented in Fig. 21, which shows a high-level introduction to the various components that formulated the flow of the study, from conceptualization to results discussions. In summary, the experiment looked at the effects of immersive auralization strategies and latency interactions on a distributed music network model that serves as a case study of interest. The effects are quantified through several types of quality evaluation layers that combine subjective and objective metrics with the goal of capturing both the quality of experience and the technical outcome of the distributed performance. A particular focus is given to the aspect of auditory "copresence" which is identified as a desirable attribute of an immersive system. The first chapter discusses the practical motivations and theoretical background driving this specific study and the general research questions derived from both the literature and previous work. Once the background is established, the case study paradigm is illustrated along the purposely designed "auralization schemes" (combination of virtual acoustic environments over two remote nodes) that form a central point of interest in the study. Other key elements of the case study, such as the modeling of latency interactions and the choice of musical material, are also presented here. Finally the specific hypotheses under test are formulated.

The document then proceeds with the layout of the methodology that was used for the implementation of the two-node case study model. The auralization "modes" designed for the experiment are driven by binaural room acoustic measurements collected in spaces chosen to represent actual distributed performance spaces or locations of intended "remote" telepresence. Ch. V displays the acoustic and calibration procedures and the parametric results that describe the character of each room. This is followed by the implementation of the actual distributed network connecting two nodes (a theater and a studio booth) on top of a pre-existing analog infrastructure connecting facilities across a building. The network was built with the capability of simulating an internet-based connection mediated by a central server, with the advantage of being able to control the transmission delay between nodes. The following chapter (Ch. VI) moves on to describe the various data collection methodologies applied to the primary and secondary

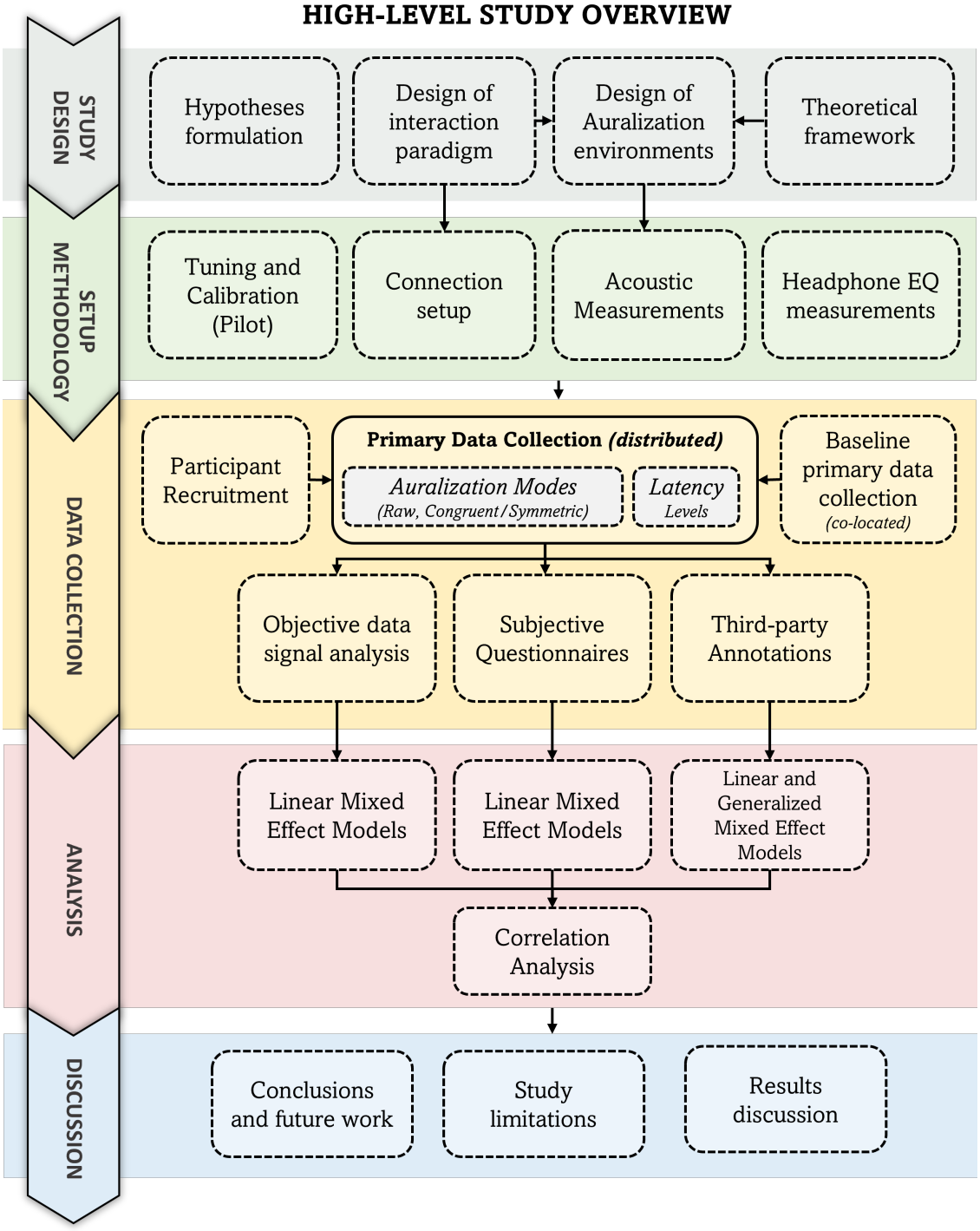


Figure 21: High-level overview of the empirical study on immersive NMP, illustrating the flow of the dissertation chapters.

layers of data. The "primary" data consist of distributed performances executed by pairs of musically-literate musicians through the network while being exposed to the various auralization environments and latency interactions. The audio signals of the performances are recorded at the central node for later analysis. The primary data also comprises "co-located" recordings of the musician pairs taken prior to the start of the experiment to capture "baseline" data of traditional interactions to help control for each pair's base musical abilities, and also to provide an internal reference point to each participant about the feeling of "presence" in a musical exchange. The primary data are used for the extraction of three different secondary evaluation layers consisting of different quantitative "realms" of evaluation. The first evaluation layer is collected from the participants themselves during the experiment and consists of a trial-based questionnaire polling individual impressions of auditory presence, auditory cohesion, and other perceived attributes of each performance as it happened. The second layer of data is composed of objective evaluation metrics, extracted from the raw signal recordings, obtained through beat-tracking and tempo estimation algorithms. The third and final layer of evaluation data is obtained through ratings and annotations from third-party expert listeners, tasked with evaluating the musical quality of each performance and listing the occurrence of perceivable performance inaccuracies.

The full set of evaluation data is used as the source for the analysis framework described in Ch. VII. The analysis framework relies on the use of selected "Mixed-Effects" models in both linear and generalized form (according to the nature of each observed variable) that are able to account for "fixed" effects of interest while controlling for "random" effects representing confounding factors that may have a potential impact on the statistical observations. The impacts of the main effects, namely "auralization" and "latency", as well as secondary effects, are therefore explored for their statistical link to each evaluation layer. The results of the analysis are depicted through model summaries and trend plots. Finally, a correlation analysis is applied between the secondary layers to explore potential links between copresence-related measures and performance quality metrics. The final assessment of the study hypotheses in relation to the results is tackled in Ch. VIII, which also discusses the significance of the findings to the larger research questions. The resulting insights are therefore discussed in light of the limitations encountered and the trajectories for future expansions of immersive NMP studies. In addition to making a case for introducing immersive technology strategies in the field of NMP, the exploration set forth by

this study traces directions toward the development of forms of assessment that bring together immersive technology and distributed music networks.

2 Study Motivations

The need for this study is motivated by the gap in literature found when looking to understand the relationships between immersive-audio environments and *auditory copresence* as a metric of social interaction quality in a distributed music system, and the relationship between subjective ratings of copresence and the technical quality of a musical performance following a "Realistic Interaction Approach" (Carôt and Werner 2009).

The improvement of auditory copresence is subject to trade-offs between complexity and fidelity in which different technological assets are able to provide different degrees of flexibility and accuracy to a measurable ground truth (Jot 1997). It is important to have a clear vision of the application-specific targets and requirements that need to be met in order to evaluate and compare the performance of different implementations. Perceptual studies serve the role of providing insights into the subjective tolerance and response to signal enhancements or artifacts by means of controlled user studies. While a lot of work has been done for speech-oriented applications such as teleconferencing (Sondhi et al. 1995), the goal here is usually that of improving signal *intelligibility* and *recognition* rather than *presence* or *immersion* (Rumsey 2002), making the validated literature not extensively applicable to musical applications. On the other side of the coin, as research in musical networks is mainly concerned with the effects of *latency* (Rottondi et al. 2016), or how to cope with it (Carôt and Werner 2009), mixed reality is a relatively new topic in the music communities and there has not been much reason, up to now, to study in depth the impact of auditory presence and immersion on distributed performance. Thus, it is still unclear, for collaborative musical applications in MR, if a high-fidelity acoustic adaptation of an incoming signal is effective for the improvement of the subjective experience, or even desirable. Furthermore, given the technical activity of the application at hand, it is important to assess how the quest for higher copresence and immersion affects the musical performer and the quality of the musical output. Such a study would inform the fields of immersive audio, computer science, and telematic music on what are the objective and subjective effects of acoustic adaptation in musical interactions, what are the perceptual tolerances, and what is the optimal balance between

complexity/fidelity that should be targeted by future systems. This document proposes a study in controlled conditions aimed to explore these questions.

2.1 Case Study

The particular case study brought forward concerns the handling of collaborative network music performances in which two connected nodes are established in acoustically divergent rooms, with asymmetric acoustic properties. It is a reasonable assumption that these scenarios are realistically common in NMPs, both within controlled concert or studio environments or personal Internet links between interested parties. As experienced during the previous work of setting up interactive collaborations for the Holodeck concert streams, NMP concerts usually involve a mix of theater stages, recording booths, and music halls of different kinds and sizes. The range of absorptive or reflective surface materials usually changes widely between the rooms employed. This can result in a "disjointed" experience, far from the realism of a regular rehearsal, where the auditory experience deals with the simultaneous cognitive processing of different acoustic characters (see "room divergence effects" (Werner et al. 2016; Klein et al. 2017)). In general terms, the common case within NMPs is that the experience of performance within the medium "feels" different than the experience in real life.

The fields of virtual and augmented reality can provide inspiration towards solutions designed to mitigate the acoustic mismatch and improve the subjective experience of musicians over a distributed network in different ways. By applying combinations of interventions based on auralization and spatialization with the purpose of eliciting *co-presence* in either the "local" or "remote" direction, it could be possible to obtain a more realistic interaction that gets closer to the auditory experience of a traditional musical exchange, and by proxy, a better music performance experience. However, there is no extensive literature looking into the effect of such interventions on the musicians' experience or on the success of the musical intention. *Quality* itself exists on many levels; objective quality of the musical result or subjective "quality of experience" are two examples. Within the world of NMP, we do not yet have enough hard data to make this case, and we lack understanding of what causes "copresence" during a distributed musical task, how copresence relates to subjective experience, and how the objective of producing an accurate music

performance is affected by it. It is therefore important to look deeper into what is the relationship between immersive audio techniques and social telepresence within musical applications.

2.2 Defining "Auditory Copresence"

This additional background section covers definitions of "auralization" and "copresence" as applied to the study.

A successful social immersive experience has the power of virtually "bringing" people to a shared feeling of *presence,* or in the social sense, *copresence* (Riva et al. 2003). The way a mixed-reality system would think of an interactive experience is to create an adapted rendering of the received audiovisual streams tailored to each receiving node. In other terms, the signals are processed so that a receiving user would believe that the audiovisual objects are "plausible" and belong to the physical current display location. A virtual reality system would instead approach this problem by creating a virtual shared environment, not necessarily grounded on the actual physical surroundings of a user, where both users are virtually "transported". Either way, the aspect of *copresence* is the key component of the experience design. In the MR/AR approach, the desired copresence space can therefore be defined as "local", ideally concurrently for each node involved. For the VR approach, the copresence space is instead "remote".

To achieve this, the auditory aspect is fundamental. The acoustic character of the streamed signals needs in some way to "match". Immersive audio techniques such as auralization and spatialization are widely employed in order to modify the sound signal to allow it to feel realistically cohesive to a target space, whether it is a real or virtual destination. *Auditory copresence* is here defined as either the illusion felt by an immersive system user (musician) when perceiving a connected user (coperformer) as "being here with me", or the illusion of being transported to a remote location where the connected user is present, essentially "being there with someone". This bidirectional exchange of presence can be explored in several ways, one such way is to acoustically adapt each stream, through auralization, to fit the local acoustics of each performer's own local physical space, thus achieving a "cohesive" rendering that conduces to each nod-user to potentially feeling copresence in their own space. This is the typical design principle of mixed/augmented reality applications, which factor in the local physical reality of the receiving node when rendering media content, addressing the so-called "room-divergence-effect" (Werner

et al. 2016). The virtual-reality approach would instead create a third "non-local" virtual acoustic environment common to both users, where performers at both nodes experience a remote version of copresence, which itself can be symmetric (both nodes experience the same remote virtual room) or asymmetric (different virtual rooms at each node). In practice, the choice of auralization strategy in an immersive distributed experience would depend on several application factors like the acoustical quality of the available performance environments, the presence of an audience, hierarchical relationships among the nodes, and technical constraints.

3 Research Questions

The starting driving hypothesis that motivates this study is that there are potential benefits to discover in the application of immersive auralization techniques (or virtual "treatments") in distributed music systems. Such methods, often applied in social mixed-reality and virtual-reality applications, have not been yet explored in depth in traditional distributed music networks. It is therefore sought to investigate auralization methods capable of enhancing the immersive qualities of the interactive music experience and evaluate the impact of the treatments over subjects and over the success of the musical outcome. A further layer regards the study of latent internal constructs of social telepresence as an indicator of general "immersive quality" (Lee 2020).

Fig. 22 summarizes the theoretical framework on top of which the hypotheses are formulated. The combination of "Latency" levels and specifically designed "Acoustic Environments" (expressed in the form of different "auralization modes", detailed in Sect. 4.2)) represent characteristics of an "immersive system" and are hypothesized to play a key part in eliciting or degrading latent inner psychological constructs of auditory *copresence* and *cohesion*, which are themselves expected to be correlated according to the literature on the subject. In the experiment, the constructs are measured through direct reporting from participating subjects as exposed to different conditions of distributed performance environments defined by the main effects, or independent variables, of "latency" and "auralization mode". In addition, the effects are hypothesized to have an impact on other observed dependent variables relating to the evaluation of the performances produced under the conditions under test. These layers of evaluation involve both expert-listener subjective assessments of the musical "quality" of the performances and

objective metrics extracted from the raw audio recordings (such as *tempo stability, synchronization metrics* etc.).

It is further hypothesized, that there is an existing correlation between the latent constructs and the evaluation layers. Meaning that the ratings of copresence and cohesion could predict the ratings of the evaluation layers. If such a correlation exists, then a case could be made that the successful elicitation of copresence and cohesion can serve as a proxy to enhance distributed performances when assessed through the proposed observable scales. In other words, the enhancement of copresence through technology might translate to the enhancement of "quality" as seen under a variety of lenses. This exploration can be further decomposed in the observation of how different modalities of copresence ("local" vs "remote") elicited by the acoustic environment treatments would impact the measurable metrics, and how the effects interact with the effects of latency. The hypothesis in this regard is that higher levels of reported *copresence* and *cohesion* can correlate to higher values of measurable assessment metrics pertaining to objective and subjective realms.

3.1 Hypotheses Formulation

The different layers of hypotheses under test are here formulated in terms of main- and sub-hypotheses to a related research question. The statistical analysis of the study is later formulated to test the reciprocal null hypotheses. Ultimately, the posed question is designed to conduce towards possible evidence that the introduction of immersive technology in distributed music networks is beneficial.

- ◇ **Are virtual acoustic environments, applied through auralization treatments, effective in eliciting auditory copresence? How can they impact distributed performance networks?**

H1. Auralization treatments, inspired by mixed and virtual reality systems, have a measurable positive impact on distributed music performance networks

H1.1. *Auralization treatments have a significant impact on participants' subjective ratings of quality of experience (i.e., copresence and cohesion) compared to the absence of auralization*

H1.2. *Auralization treatments have a significant impact on subjective evaluations of the musical outcome of a distributed performance compared to the absence of auralization*

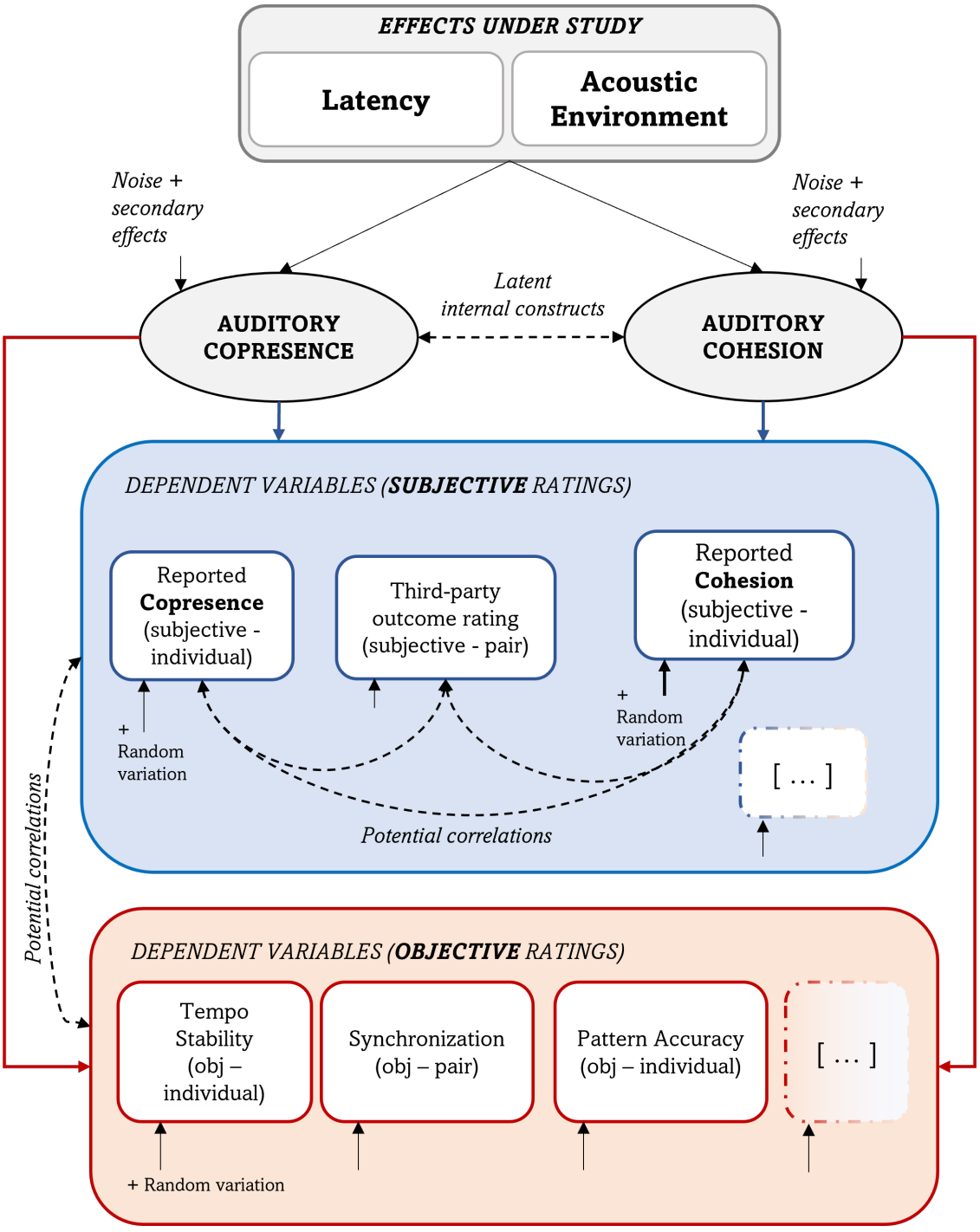


Figure 22: Visualization of the hypothesis space driving the study. The latent psychological constructs of auditory *copresence* and *cohesion* may affect measurable metrics in NMPs. For clarity, the figure only shows some examples from the set of possible causalities and correlations.

H1.3. *Auralization treatments have a significant impact on objective performance quality metrics of a distributed performance (i.e., tempo and beat metrics) compared to the absence of auralization*

◇ **How does the factor of signal latency interact with the auralization treatments, in regards to different "quality" aspects?**

H2. Latency effects can significantly degrade the quality of a distributed music experience.

H2.1. *In contrast to low-latency levels, high latency negatively impacts all layers of quality evaluations*

H2.2. *Compared to low-latency levels, high-latency levels can degrade the effects of auralizations on subjective ratings of quality of experience such as copresence and cohesion*

◇ **Is there a relationship between copresence and other observable measures of "quality"?**

H3. There exist positive correlations between *copresence* and other dependent variables

H3.1. *The rating of Copresence is positively correlated to the performers' ratings of other subjective indicators of quality*

H3.2. *The rating of Copresence is positively correlated to the subjective evaluation of distributed performances musical quality from external listeners*

H3.3. *The rating of Copresence is positively correlated to objective performance quality metrics of a distributed performance*

4 Study Platform Design

The case study of interest for this dissertation is the study of asymmetric real-time NMP connections established between rooms with divergent geometric and acoustic characteristics. The existence of this problem was first raised during the Holodeck concert series as observed between players connected between the theater stage and studio booth (see Ch. III, Sect. 1), a situation in which the level of asymmetry was evidently salient. This base scenario was taken as

an interesting canvas for the design of strategies aimed at eliciting *immersion* and *copresence* using VR- and AR-inspired principles.

The main requirement of the study was the design of combinations of auralization strategies that could be implemented over a distributed music network, with the intent of eliciting different variations of "auditory social telepresence" or *copresence* within performers. The system had to be able to record the performances for later evaluation through different assessment methods in order to observe different kinds of effects and provide answers to the research questions. The second requirement of the system was to introduce and control latency at different magnitudes, in order to recreate realistic NMP scenarios and study the interactions of latency with auralization.

A key consideration important for the design of this study is that the discussed auralizations are designed for the potential benefit of the musicians, rather than an external audience. This permits the study to allow for control of audio reproduction via headphones rather than loudspeakers and decreases the number of acoustic engineering challenges that interfere with the design process. However, it is not excluded that improvements to the musicians' quality of experience can translate into improvements towards the final musical outcome as experienced by an offline audience; see hypothesis H3.2..

In the proposed model, the two nodes are represented by two location types identified as "Theater" and "Booth". These locations types are representative of a typical connection topology that can occur in a distributed music network live applications, including the Holodeck concerts described in Ch. 4. A reverberant "Theater" location where an audience is potentially present is connected to an acoustically dry "Booth" studio location containing a remote musician "tuning in" the concert or the rehearsal. This is just one of the types of spatial relationships that can arise in a mixed-space network music performance but is a particularly interesting one. First of all the degree of difference among the nodes represents the most acoustically divergent scenario encountered during previous studies on NMP conducted by the author and laboratory colleagues. The acoustical divergence makes the situation challenging but at the same time very appropriate for the introduction of "immersive techniques" or auralization interventions dedicated to augmenting copresence between participants. Secondly, the particular theater-booth combination was observed to be a common occurrence in NMP systems experienced in professional or academic environments (although not particularly

common in mass-commerce applications), therefore a plausible candidate case study to bring forward as a model for future immersive NMP applications.

4.1 Interaction Paradigm

The interaction paradigm was designed with the Holodeck model in mind and with the goal of creating a controlled smaller-scale experiment platform that could be used to study remote interactions between users. For practical purposes, the early design stages started with the requirement that such a platform model had to exist within a single building location. Furthermore, a local study environment could be built without the need of introducing internet-based transmissions (see Ch. V for details on the actual implementation) by taking advantage of studio quality analog-transmission facilities. Using a controlled, reduced, local model network allowed to focus the objectives on evaluating the empirical effects of an immersive system on future-oriented applications without the technical overhead and limitations imposed by larger scale systems and their inherent transmission latencies.

Fig. 23 expresses a miniaturized version of an audio-only star-topology network, inspired by the Holodeck architecture. A central location is in charge of collecting data streams from the connected nodes and acts as the "distributor" of processed data. In this particular case, the data is simply formed by single-channel audio signals out from user nodes A/B on the way to the central node ($x(t)$ and $y(t)$), and two-channel processed audio signals on the way out of the central node, towards the user nodes where they are reproduced over headphones. The outward signals consist of dedicated mixes, individually rendered accordingly to the needs of each user (e.g. auralization or spatialization). The mixes contain processed communication signals from one node to the other, embedding a certain degree of network delay ($X(t-\tau)$ from A to B and $Y(t-\tau)$ from B to A), and also a self-monitor feedback signal, which is also potentially rendered according to some application requirements, sent back to the originating node. Optionally, secondary communication channels can be opened across nodes (e.g. voice channels broadcasts from the central node). Additionally, all passing signals are recorded at the central node location for later analysis.

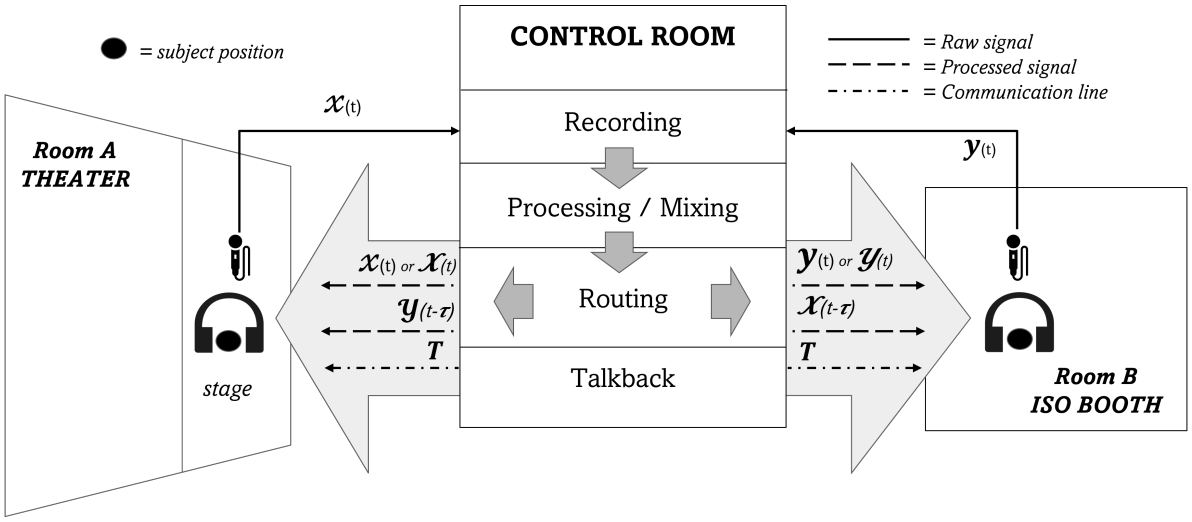


Figure 23: A three-nodes distributed music network, modeled on the star-topology paradigm. A central node collects data from two remote nodes, processes it, and distributes the processed versions back. The central node also acts as a signal processing server for self-monitor signals that are sent back to the originating nodes.

4.2 Auralization Modes

The driving design goal behind the proposed auralization approaches was identified as the elicitation of a feeling of auditory "copresence" between two connected users placed in acoustically divergent nodes. Four different auralization approaches were conceptualized to address the different directional modes of copresence and to cover a realistic set of structural topologies found in NMP systems while maintaining controlled laboratory conditions.

The approaches are hereby referred to as "auralization modes", categorized as combinations of either *congruent vs. divergent* strategies, with *symmetric vs. asymmetric* implementations. In principle, the congruent designs are grounded on the achievement of "acoustic fit" between the acoustic character of a virtual sound source and the local physical environment of a user, thus relying on an audio-visually cohesive rendering in the attempt to elicit a local copresence illusion, also responding to the "room divergence effect" (Werner et al. 2016). The divergent modes are instead relying on a non-local auditory virtual environment designed to transport the listener's illusion of presence towards a remote location, different from the one they are currently physically present in. Both congruent and divergent modes of auralization can be applied symmetrically at each node (by applying the same virtual room to the auralization process) or asymmetrically (in

different virtual rooms). The choice of including aspects of symmetry/asymmetry is motivated by the possibility of hierarchical organizations of distributed networks (e.g. virtually transporting musicians towards a target "concert" room) or potential application limitations (e.g. auralization filters are only available for one location).

The combinational organization of the different modes is summarized in Fig. 24, which collocates each strategy in the design domain characterized by two axes of reference. The resulting matrix shows the combination of *symmetric* and *asymmetric* treatments with the MR/AR inspired *congruent* strategy and the VR-inspired *divergent* strategy. Although no particular starting conjecture is here made in regards to the ability of each mode in eliciting copresence, it can be assumed that an optimal VR-oriented experience (i.e., divergent) would present a symmetric environment, while an optimal mixed-reality interaction (i.e., ideally congruent) would adopt an asymmetric strategy. Nevertheless, for statistical comparison and exploratory reasons, also the "non-optimal" strategies were included in the study. The figure also references the "raw" mode, where no auralization method is applied to the audio signals of the distributed network. The "raw" condition served as a baseline control for testing the formulated hypotheses. The factor of latency is considered at equal levels in each scenario.

The realization of each auralization mode relies on combinations of spatial *Binaural Room Impulse Response* filters (BRIRs) measured in target rooms at the source positions where each performer is virtually located. BRIRs can thus be used to transfer the virtual acoustic properties of the source-receiver relationship to a non-reverberant sound stream. For more background on BRIRs, please refer to Ch. II.

Through headphone reproduction (required for spatial audio playback when using BRIRs), each performer would be presented with a self-monitor feedback signal processed with a BRIR measured at a "near" position, where a performer's own clap occurs in relation to their ears. The co-performer streams, sent separately along each individual mix, are instead processed with a BRIR filter taken at a "far" position. Representing the intended virtual location of the collaborative partner within the virtual target room. The choice of BRIR room origin at each destination thus determines whether the applied condition is congruent (BRIR measured *in situ*) or divergent (BRIR measured elsewhere). As mentioned in earlier paragraphs, these can either be applied symmetrically (same BRIR room origin at each node) or asymmetrically (different room origin).

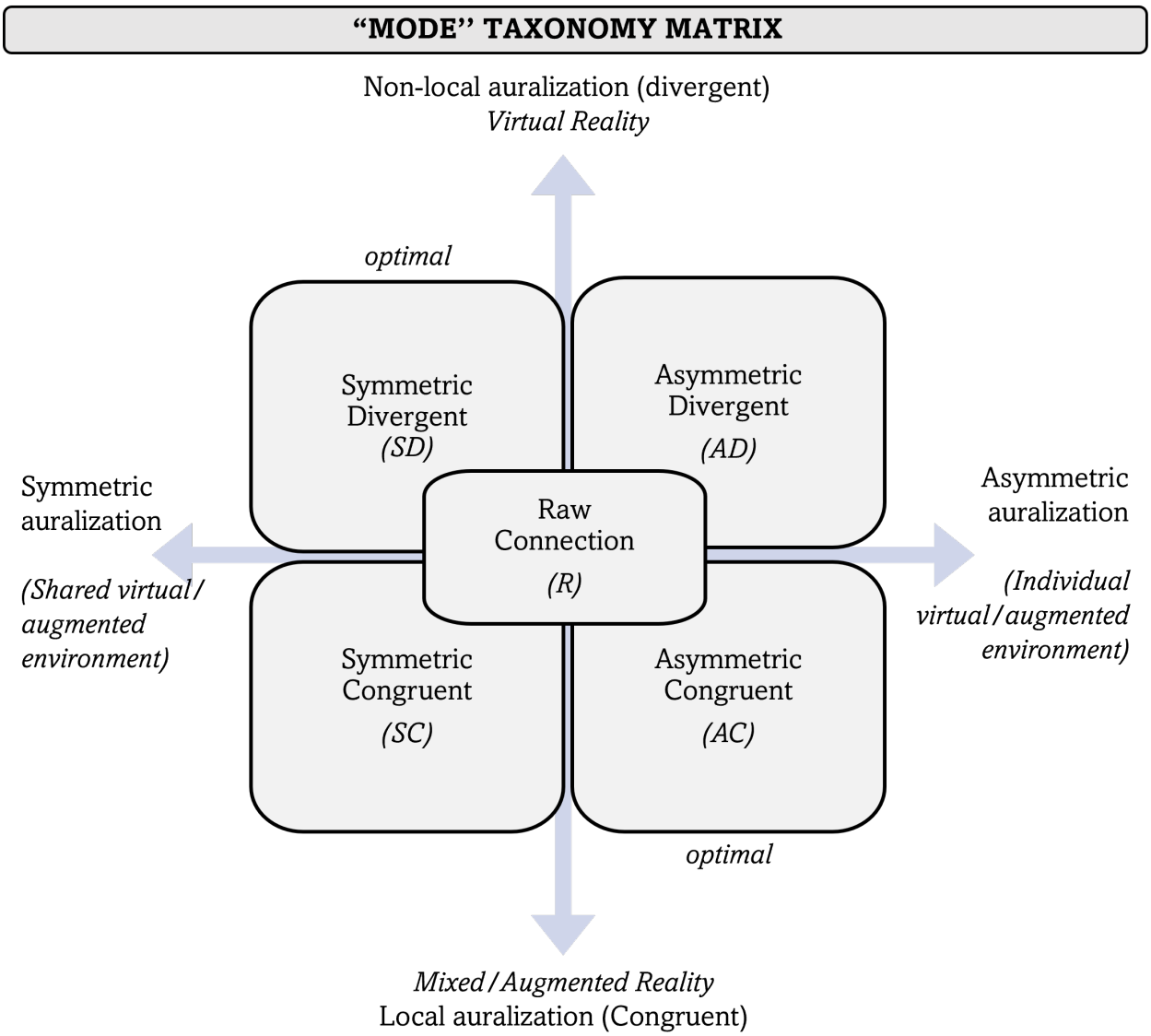


Figure 24: Classification taxonomy of virtual acoustic treatments that may be applied to an NMP environment. The treatments are not necessarily mutually exclusive if a hierarchy of nodes is established (for example a concert room may act as a reference room for acoustic adaptation). The conditions tested in the experiment in this chapter are designed accordingly to cover these potential strategies.

Following the definition of Eq. 1, a "near" BRIR taken in room R is here labeled as H'_R (Eq. 6), while a "far" BRIR taken in room R is labeled as H''_R (Eq. 7). These definitions provide a reference for the understanding of the figures representing each individual mode. The particular directional coordinates are applied identically throughout every room measured (see Sect.2.1 in Ch. V for the measurement methodology). The "near" position for the self-monitor filter is defined as being at approximately 12 inches distance from the listener, at an elevation offset of $\phi = -45^\circ$. The "far" position is instead set at a front position, at a distance of 8ft (deemed as a plausible distance between two performers in a room). Both positions are centered on the median plane ($\theta = 0^\circ$).

$$\texttt{(near position)}\ \ H'_R(t) = \texttt{BRIR}_R(t)\{\theta = 0^\circ, \phi = -45^\circ, r = 12''\} + \xi \tag{6}$$

$$\texttt{(far position)}\ \ H''_R(t) = \texttt{BRIR}_R(t)\{\theta = 0^\circ, \phi = 0^\circ, r = 8'\} + \xi \tag{7}$$

The definitions of equations 6 and 7 apply throughout the mode design illustrations shown below. In general terms, each incoming stream ($x(t)$ and $y(t)$) is processed at the central node of the network through convolution with target-room BRIRs according to the mode applied. Each signal is processed twice in parallel, once with the "near" position for the self-monitor signal going back to the originating room, and once with the "far" position for the transmitted signal going to the connected node. In mathematical terms, at each node, the self-monitor signals result in $H'_{R1}(t) * x(t)$ and $H'_{R2}(t) * y(t)$ and the transmitted signals result in $H''_{R2}(t) * x(t-\tau)$ and $H''_{R1}(t) * y(t-\tau)$, with rooms R1 and R2 defined by the mode and τ representing artificial delay injected in the stream to simulate internet-based connections (see Sect. 4.3). All processing happens in stereo to account for binaural spatial cues. Furthermore, to avoid interactions with unwanted captured local reflections getting through the microphone during performance, a signal gate stage is applied before auralization to preserve only the most prominent performer sound (a process deemed acceptable for the signal quality by the fact that the chosen musical material consisted in transient clap sounds, see Sect. 4.5). Ch. V provides more details on the measurement of the BRIR filters and the equipment chosen to minimize the coloration error, also including the measurements of equalization filters to correct the headphone response.

4.2.1 Raw Mode (R) (Control)

The study control condition consisted of a "raw" mode, representing a typical NMP where no auralization processing is applied to any of the signals being parsed through the network. The unprocessed signals are labeled $x_r(t)$ and $y_r(t)$ and are distributed to the network as captured by the microphone (including possible local room reflections that seep through in the signal capture). The only processing that occurs in this mode is the injection of artificial delay from one node to the other (see Sect. 4.3), in a similar fashion to every other mode. Fig. 25 illustrates the signal distribution between the theater and booth node (upper panel) and the hypothesized "auditory copresence image" (lower panel) showing the expected presence effect from the perspective of each node. In this particular case, the expected auditory copresence image in regards to room B as experienced by room A (Copresence $(B|A)$) and in regards to the signal of room A as experienced by room B (Copresence $(A|B)$) is disjointed and undefined at each node. Each user, therefore, experiences their own acoustic character as locally present, and the acoustical character of the co-performer as present in the connecting room.

4.2.2 Asymmetric Congruent Mode (AC)

Asymmetric auralization modes apply different sets of BRIR filters to each node's streams. Each self-monitoring signal is processed with a BRIR acquired at the "near" position while each coperformer signal - received at the opposite node - is convolved with the BRIR belonging to the "far" position. In the congruent case, the filters are acquired in situ at both rooms and applied to the streams being sent to the originating rooms. The *asymmetric congruent* mode represents the ideal mixed-reality solution. The signals at each node are subjected to a processing scheme that aims to achieve cohesiveness at each node. As a result, the copresence image at every node works in a "local" direction, meaning that both users can experience the "being here together" version of the copresence psychological construct (Fig. 26).

4.2.3 Asymmetric Divergent Mode (AD)

The *asymmetric divergent* mode instead applies BRIRs collected at third arbitrary locations (i.e., not in the performer's own spaces). The copresence image is "remote" at both nodes, meaning

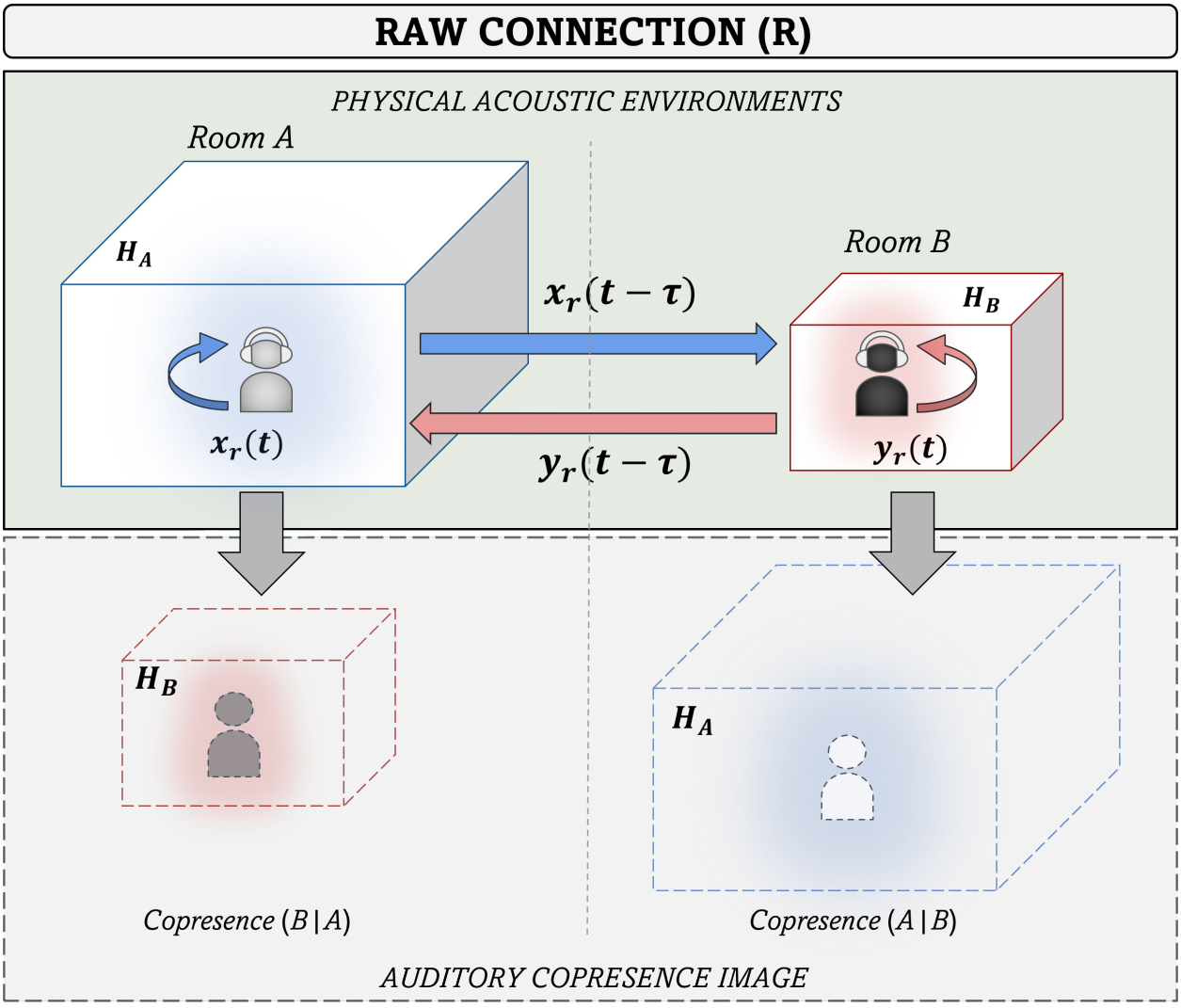


Figure 25: **Raw Connection** mode (R) - No auralization applied. Musicians in rooms A and B hear themselves and each other as captured. Local room reflections may pass through embedding the acoustic path captured by the microphone. The copresence image is disjointed at both nodes.

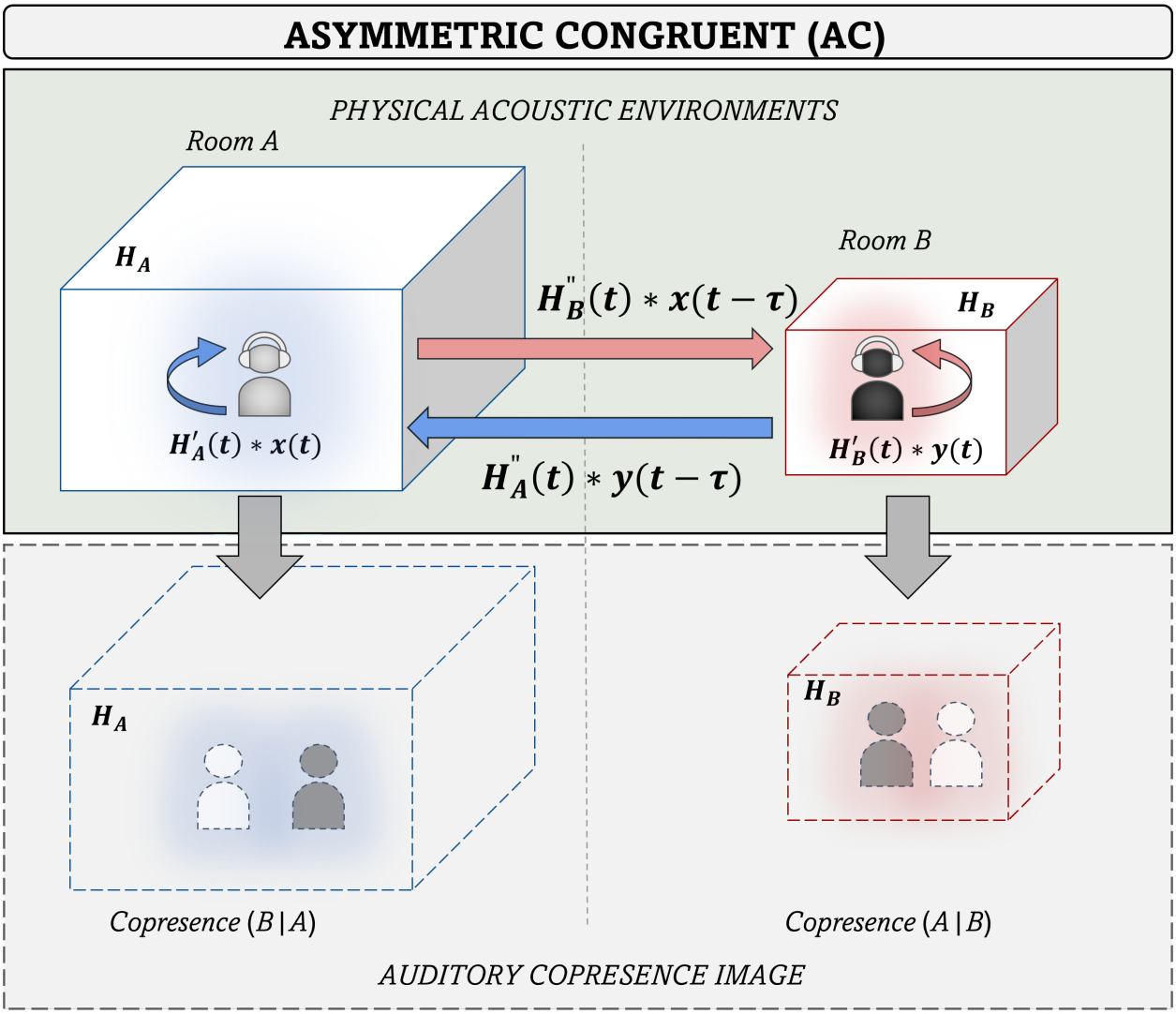


Figure 26: **Asymmetric Congruent** Mode (AC) - In this scenario the signals are "asymmetrically" adapted to their destination rooms. Within this condition, audiovisual cohesion is maximized as the acoustic character is intended to fit the local environment and acoustic expectation of each node.

that both users are prompted towards a virtual copresence experience of "being there together", although "there" in this case is not the same place at each end of the network (Fig. 27). The scenario represents a "mismatch" situation where each nodes handles rendering in their own way, either because there would be no common rendering data available or due to the user's preferences. In practice these third room locations are chosen in the experiment to be roughly similar in size, but not identical, to avoid biasing the response at each node based on strong acoustic decay features.

4.2.4 Symmetric Congruent Mode (SC)

The symmetric modes are processed through the applications of the same set of BRIR filters among the two directions of interactions. The self-monitoring signal is processed with the BRIR acquired at the "near" position while the coperformer signal received at each node is convolved with the BRIR acquired at the "far" position. The *Symmetric Congruent* mode, adopts the BRIRs measured in the space of one of the two nodes, in this case being the larger, more reverberant, theater space. While in this mode the choice of room is symmetric, the audiovisual cohesion of the acoustic character is not. Since the applied filters are collected in Room A, a congruent experience (where the acoustic character fits the visual space of a listener) only applies to the user of room A, who experiences a "local" presence. The user of room B is instead experiencing a divergent situation that only permits "remote" copresence to occur (Fig. 28). The (SC) auralization mode is a hierarchical scheme, as in more importance is given to the room from which the BRIRs were acquired (possibly a concert space), and connected streams are "brought" to this location. Another way to look at this mode is to consider it a combination of an AR experience in node A with a VR one in node B.

4.2.5 Symmetric Divergent Mode (SD)

The *symmetric divergent* mode aims to bring both users to a common-shared environment, not grounded in the real physical space of any of the two nodes, but pertaining to a third "virtual" location. Since "divergence" is equally applied at both locations, the copresence image is identical (Fig. 29) and expected to happen in its "remote" version, meaning that both participants may experience a feeling of "being there". This scenario, where the rooms are given equal importance,

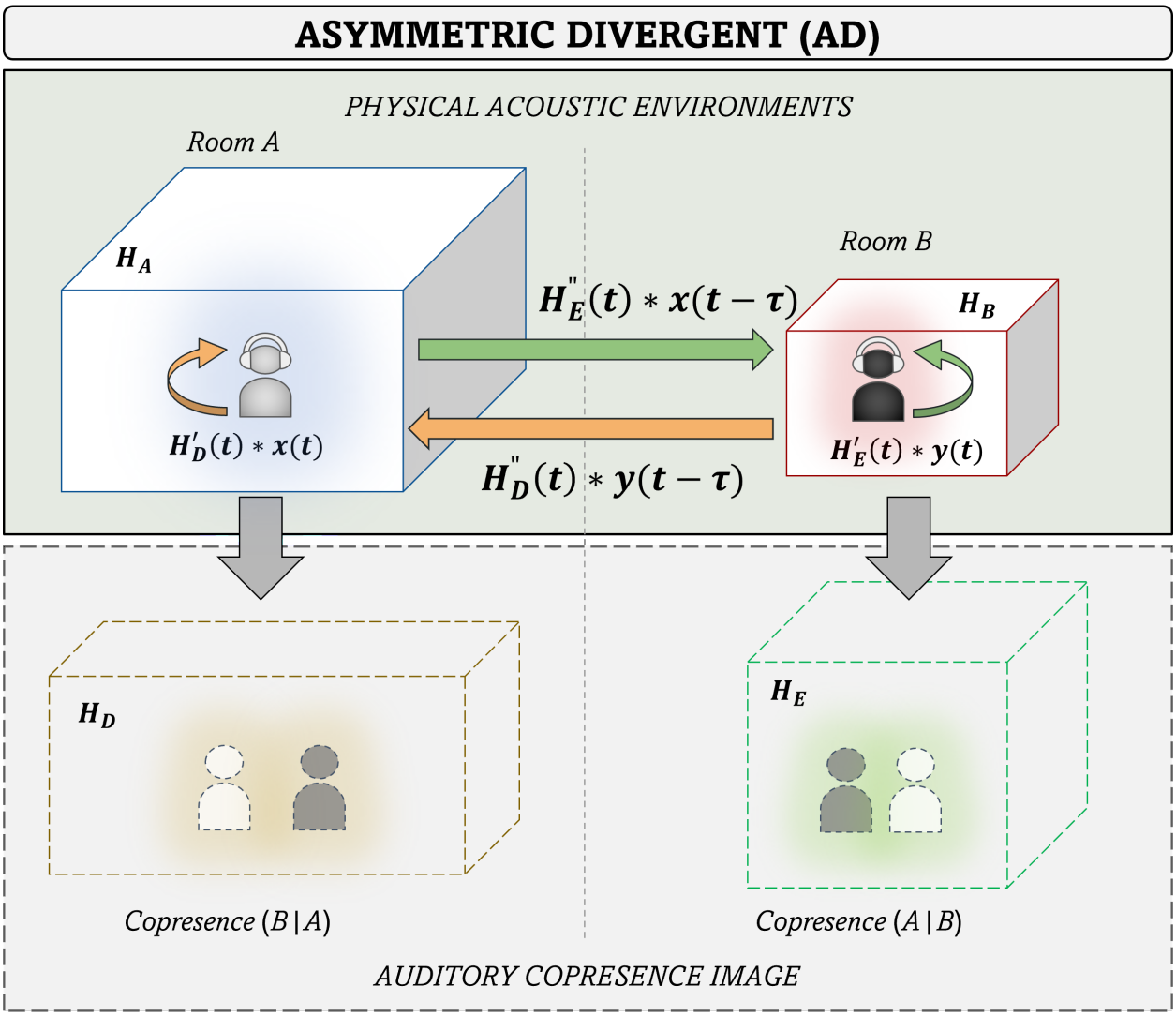


Figure 27: **Asymmetric Divergent** mode (AD) - In this scenario, the signals at each node are processed with non-congruent BRIRs at each end. From each node, the experience is that of "remote" copresence, towards a shared virtual location, albeit a different one at each end.

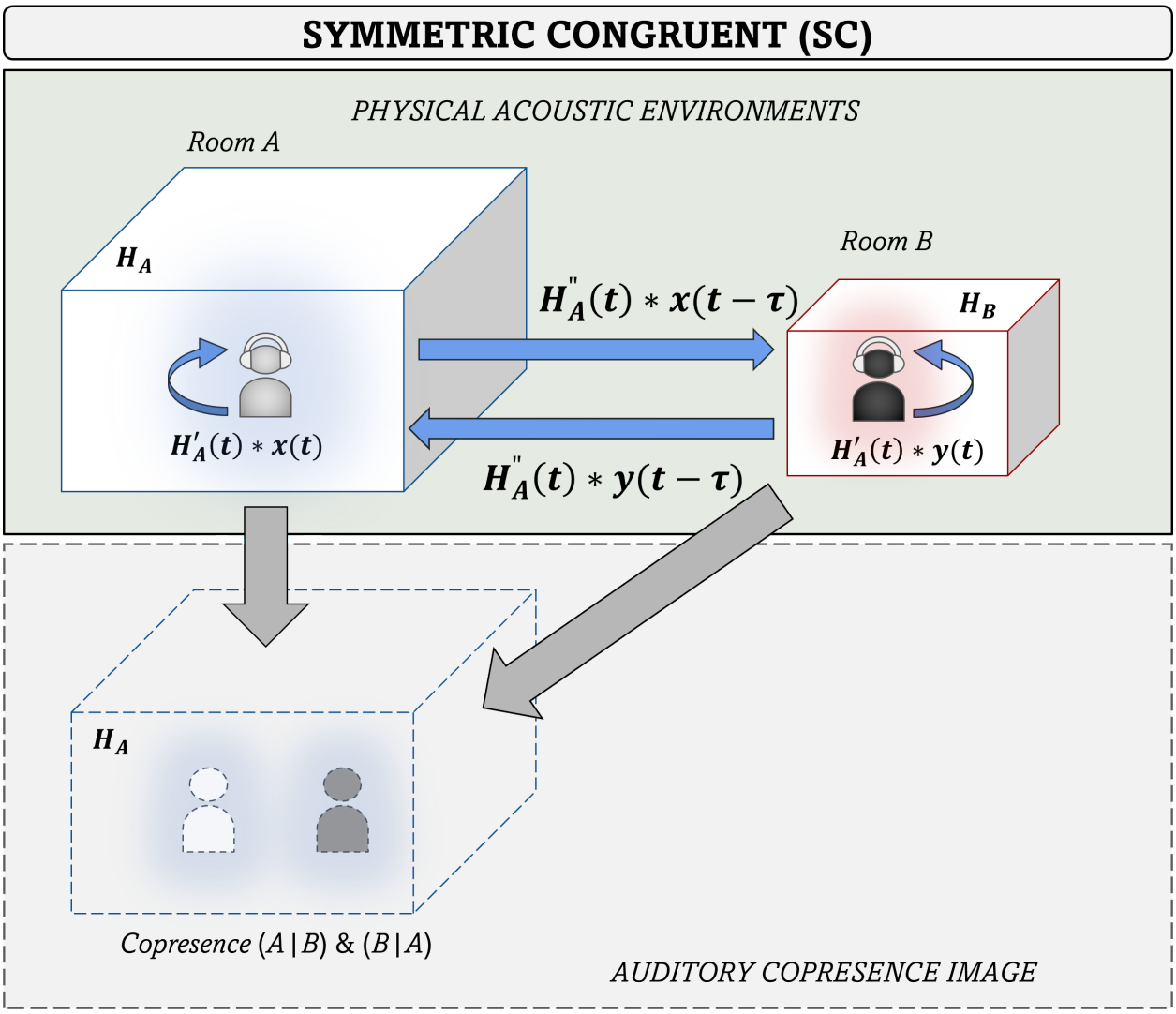


Figure 28: **Symmetric Congruent** mode (SC) - signals are treated symmetrically with the same set of BRIR filters. However, cohesive congruence is only experienced at a "concert" node from where the BRIRs were acquired.

is analogous to a virtual reality collaborative application that employs arbitrary acoustic spaces according to the user's desires or application design. When observed individually per room, the experience is not that different from that of the (AD) divergent mode, with the difference that a shared environment may conduct a similar musical response rather than two different ones.

4.3 Latency Effects

Researching immersive distributed music networks can hardly ignore the issue of latency and the way it affects the quality of musical interaction. Transmission latency is inherent to internet-based networks and can heavily affect the ability of musicians to rhythmically synchronize with each other. The use of auralization can potentially improve the quality of the musical experience and outcome, but the underlying latency conditions may hinder or degrade any benefits obtained. Therefore, it is important to consider the impact of latency when evaluating the effectiveness of immersive distributed music networks.

Within the analog infrastructure over which the model network is built, the actual transmission latency is reduced to a much lower system's baseline latency (measurements described in Ch. V, Sect. 4.2). So, artificial latency is added to the transmissions in order to achieve representative latency levels of interest to the study. The levels are identified as "acoustic-latency", "mild-latency", and "high-latency". The lowest level is designed to match the acoustic physical wavefront traveling time from a performer standing at a rough distance of 8ft from another. This results in a 7ms latency, which is within the typical range of ensemble interactions, and below the 10ms threshold of optimal delay response (Chafe et al. 2004). Drawing from literature (Rottondi et al. 2016), the *Ensemble Performance Threshold*, the one-way latency threshold at which regular performance is usually not impaired, is given to be on average around 30ms, with variations depending on circumstantial factors such as beat tempo, score complexity, and instrumentations. Following these figures, which are typical of clapping experiments, latency values right below and right above this threshold were deemed to be representative of "mild" and "high" conditions. Thus, the "mild" level is set to reach a one-way latency of 20 ms, serving as a noticeable yet tolerable level for a rhythmic performance. The "high" level is instead approximately set at the upper limit of the playable range where latency is expected to heavily disturb a performance, this is identified as a 40ms one-way latency.

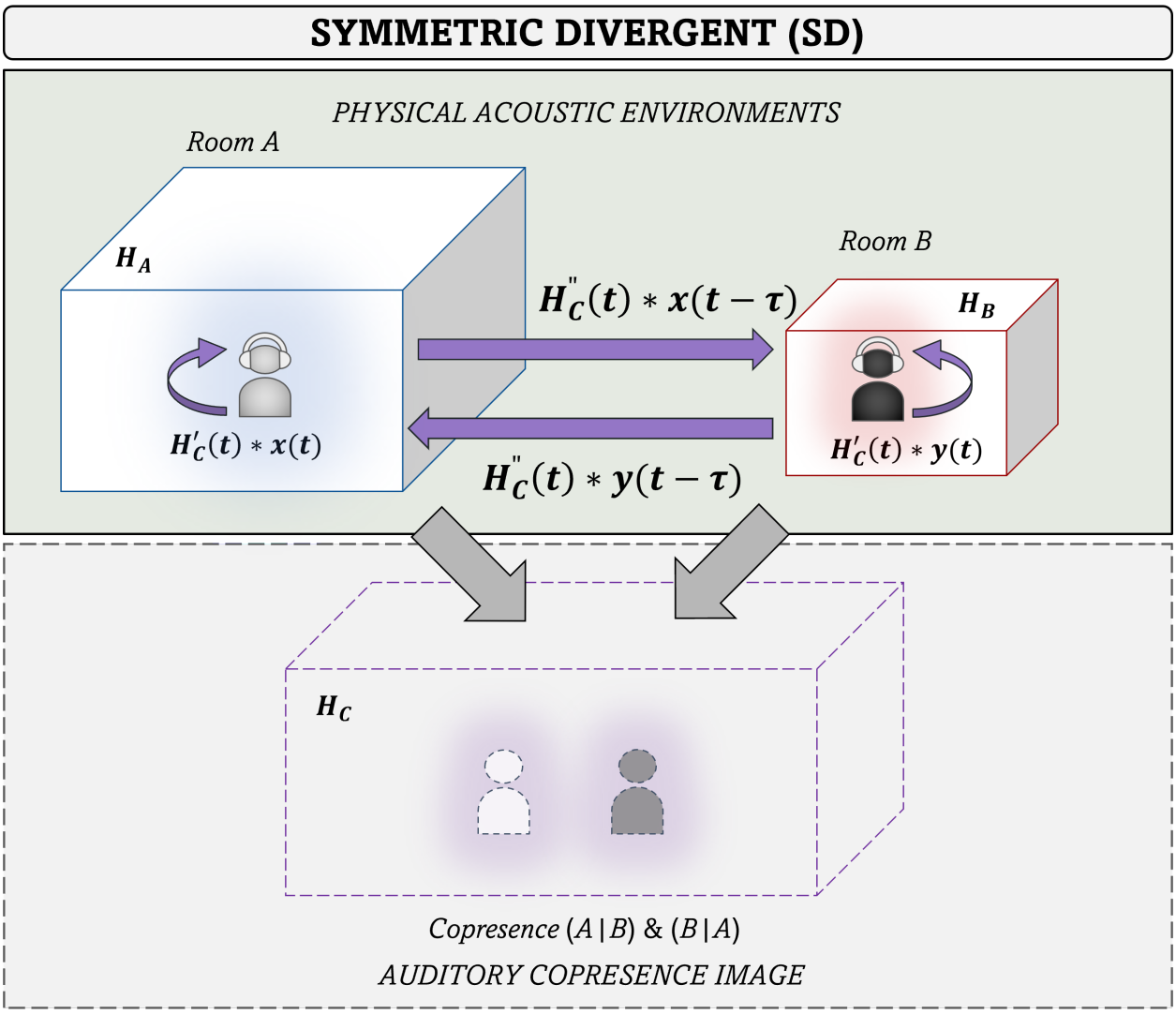


Figure 29: **Symmetric Divergent** mode (SD) - signals are treated symmetrically with the same set of BRIR filters belonging to an arbitrary room. From each node, the experience is that of "remote" copresence, towards an equivalent shared virtual location.

Table 1: Summary of designed experiment conditions

Auralization modes			
Acronym	**Name**	**Copresence** $(B\|A)$	**Copresence** $(A\|B)$
(R)	Raw Connection	*n/a*	*n/a*
(AC)	Asymmetric Congruent	Local (A)	Local (B)
(AD)	Asymmetric Divergent	Remote (α)	Remote (β)
(SC)	Symmetric Congruent	Local (A)	Remote (A)
(SD)	Symmetric Divergent	Remote (γ)	Remote (γ)
Latency Levels			
Denomination		**One-way latency amount**	
"Acoustic delay"		7ms	
"Mild latency"		20 ms	
"High latency"		40 ms	

Table 1: Summary of auralization treatment conditions and latency levels. Combinations of these two factors represent the set of conditions under study. The *raw* auralization mode and the 7ms *acoustic delay* represent control conditions. Letters *A* and *B* denote the two connected nodes (Theater and Booth) and $(A|B)$ indicates the copresence induced in room B in regards to signals originating from A. The letters α, β and γ represent virtual room locations.

4.4 Summary of Conditions

Table 1 summarizes the set of conditions that form the main effects of interest of the study. A total of five auralization modes and three latency levels compose the set of "treatments" under study within a distributed, immersive, musical collaboration.

4.5 Musical Material

For the purposes of this experiment, no sound-producing instrumentation was considered other than clapping hands, played through a "Realistic Interaction Approach" (Carôt and Werner 2009). The main reasons for this choice of approach were to remain in comparative terms with other experiment methodologies used in related literature (Chafe et al. 2004; Chafe et al. 2010; Hupke et al. 2019a; Hupke et al. 2020) and to control for a number of confounding variables that would have made this experiment harder to analyze (e.g., playing style, instrumentation class, timbre, interactions with room acoustics). Furthermore, percussive transient sounds present signal characteristics that are more easily and robustly tracked by modern beat-tracking algorithms. Their broadband spectrum property also makes them ideal stimulus signals for eliciting the

full-frequency character of the local room acoustics. The acoustic energy of impact sounds occupies almost the whole audible frequency range, giving a better chance for the room acoustics to be elicited in its full spectrum, maximizing its effect.

Regarding the musical aspect, in order to promote a fully mentally-engaging activity, it is important to steer away from overly simplistic repetitive beat patterns and avoid performers executing their task more out of mechanical memory rather than attentive musical collaboration. At the same time, an overly complicated musical piece can present issues related to fatigue and low repeatability, making data collection more noisy and the analysis less robust (Grosche et al. 2010). Given the prospected duration of the experiment, these considerations played a key role in selecting a piece. Another aspect, was that of the hierarchical relationship of the musical parts. It is sometimes the case in rhythmic interactions to rely on a "leader-follower" dynamic (Boerner et al. 2004), where the leader parts functions independently from the rest. This was an undesirable characteristic as it would play against the synchronization efforts that musicians would occur into when playing over the internet, essentially making only one node "care" about synchronization. It was ultimately decided to aim for a realistic and neutral musical interaction, steering away from NMP latency-coping strategies and hierarchical relationships, in furtherance of keeping a better study focus on the auralization and latency effects rather than the effects of musical strategies.

The musical piece that was deemed appropriate for this study was *"Clapping Music"* by S.Reich (Reich and Hartenberger 1980). The piece was first suggested by potential study participants from NYU's percussion program, who already engaged in the piece through their academic curriculum. The choice was thus made on the basis of the number of advantages this piece presented, mainly being a piece entirely based on hand clapping. This musical piece is divided into two playing parts, hereafter referred to as *"Static"* and *"Shifting"* parts. The Static and Shifting players both start on the same beat, played in a compound quadruple time of $\frac{12}{8}$, and repeat it a number of times after which the Shifting player circularly shifts its rhythm pattern by a one-eight note to the left. The performance sequence progresses in this circular shift pattern until the Shifting player rejoins the Static player's beat alignment[1].

Before running the principal study, the piece was tested in pilot trials with students of

[1] A video of a demonstrative performance can be found at this link: https://youtu.be/YPU5XrmORCQ . Consider that the actual experiment shortened the length and tempo of the piece.

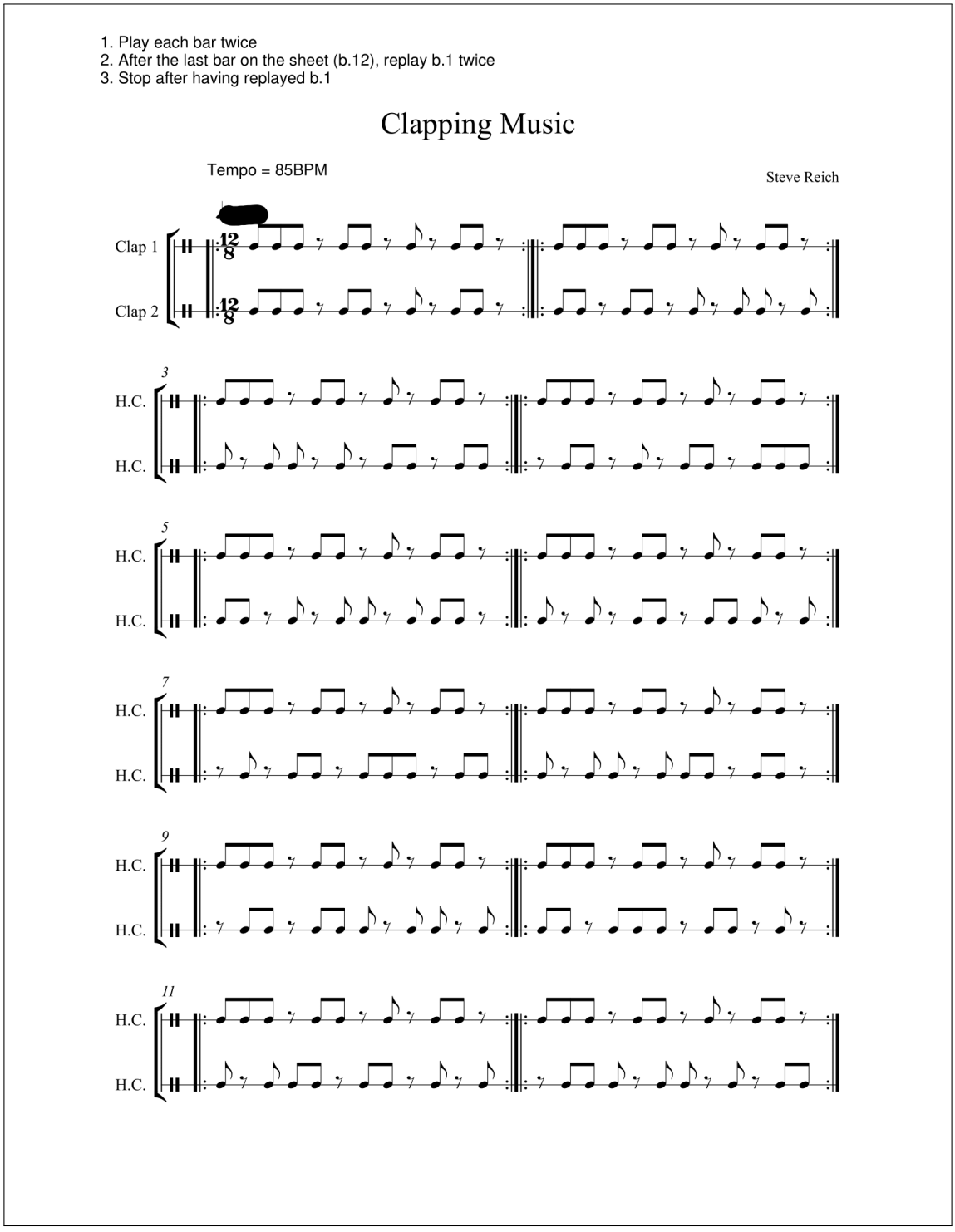


Figure 30: *"Clapping Music"* score used in the experiment, with annotated modifications. *Original from https://sites.ualberta.ca/~michaelf/SEM-O/SEM-O_2014/Steve's%20piece/Clapping%20Music.pdf.*

NYU's percussion program (see Sect. 6). This step was taken to attest to the feasibility of the score in this context. The pilot trials highlighted the challenging, but not overly so, nature of the piece, making it land in the "sweet spot" of difficulty being engaging (especially for the Shifting player) but easily repeatable multiple times. However, some adjustments were applied to keep the average performance time within the one-minute mark. To achieve this, each bar's repetition iterations was set to two repetitions instead of four. The working tempo was also adjusted to 85 BPM. It was also noted that, according to the musical proficiency of the participant, the beat accent was interpreted as either simple quarter-beat accents or compound triplet-beat accents. This aspect was deemed not influential for the purposes of the experiment and the calculation of relative metrics, so it was left to the participants to decide as more comfortable for them.

For reference, the complete annotated score is shown in Fig.30, where "Clap1" corresponds to the Static-part performer and "Clap2" to the Shifting-part performer.

5 Limitations

It needs to be noted that the presented hypotheses are constrained to the particular case study brought forward, that of a star-topology network with two nodes with divergent acoustic environments. Therefore no generalization claim is possible beyond said topology. Other fixed factors that could potentially change the resulting observations include the choice of alternative musical materials with different instrumentation or degree of complexity, alternative choice of node rooms that are acoustically significantly different from the chosen combination, and alternative choice of virtual "remote" environments. Factors such as the introduction of headsets for multimodal rendering, multiple ensemble members, or specific responses to reverberation parameters are also not part of this study. These are elements that are nevertheless interesting pieces for future explorations that can complete the puzzle of decoding "immersion" plausibility and quality in NMPs.

In regards to technical implementation, the auralization techniques implemented through static generalized BRIRs do not represent a state-of-the-art immersive system since the implementation does not include individualized HRTF user-fit nor head-tracking 3DOF rendering, which are important elements for improving the immersive experience (Roginska and Geluso 2017). However, the introduction of individualization elements did not prove practically feasible

(individually measuring the BRIRs of participants in different rooms is a big engineering cost). In the case of head-tracked spatial audio rendering, its implementation would require a different network topology to allow the rendering process to be executed at the end nodes rather than a central node, meaning a higher need for resources and the likely introduction of additional latency beyond the maximum viable levels specified by the study. Nevertheless, the proposed lower-complexity study platform still stands as a valid and valuable model since these missing elements can only improve the immersive quality of a system.

An interesting remark about divergent cases is that there is the potential for a "double-slope-decay" effect (Boren and **Andrea Genovese** 2018), where the "net" reverberation tail is a function of the interaction of connected reverberant rooms with different acoustic characters. The interaction of the two environments can be minimized through signal gating essentially making the effect negligible. However, when using open-back headphones there is the potential that the smaller room between the local and virtual environment would provide the earliest reflections, while the more reverberant room would provide more late-reverb energy. In practice, external-source coloration, and gain differences would make this a variable situation requiring its own measurement for determining the magnitude of this interaction. It is beyond the scope of this dissertation to explore the actual degree of this impact, but it is worth to consider this when thinking about the combinations of certain reverberant environments with others.

CHAPTER V

TECHNICAL SETUP METHODOLOGY

Disclaimer: All distance and size units in this chapter are reported using the imperial system, using terms such as "feet" and "inches", rather than the metric system. This is done to keep consistency with the measurement equipment used during the collection of the data presented in this chapter.

This chapter illustrates the methodology used for the technical setup of the experiment design described in Ch. IV. The chapter first covers the selection of facilities needed for implementing the connection setup, the measurement of the room acoustic filters (in the form of BRIRs) used to apply the auralization environments described in section 4.2, and headphone calibration for equalized audio reproduction. Next, the chapter covers the anatomy of the finalized analog connection system and the digital audio software environment used for recording, processing, and routing the performer's signals. In addition to providing static room acoustics simulation processing, routing software also applies an additional artificial delay to simulate the latency levels found in typical internet-based remote connections, applying different degrees of severity. Finally, the chapter covers a summary of the complete set of equipment employed in all stages of the experiment and a summary of the pilot trials used for establishing a tuning procedure for the signal levels.

The general goal of this stage was to implement a low-latency distributed network environment capable of reproducing the auralization environments designed for studying copresence in NMPs. Figure 31 represents the implementation target adapted to the facilities available at the University department's building, adapted from the conceptual model shown in Ch.IV, Sect.4.1. While all the processing and recording happens in digital format, the system is designed to leverage real-time transmission over an analog network, over which internet-based connections could be simulated with the addition of artificial latency. In essence, a central node location is in charge of collecting audio signals from connected rooms, recording them,

processing them with effects, and distributing the appropriate output mixes at each node. Each mix would consist of the self-monitor signal of the performer, with or without room acoustic processing using a "near-field" BRIR filter ($x(t)$ or $X(t)$), mixed with the co-performer signal processed with a "far-field" BRIR and occasional additional delay ($Y(t-\tau)$). The specific choices of BRIR filters for each routing depend on the auralization "mode" examined at each trial. The combination of modes and latency level represent the main effects under study. To allow amplitude calibration for the participants (necessary to account for different performance styles), headphone amplifiers are introduced at each location for controlling the master mix audition level relative to each node. Recording capture levels and mix levels are instead controlled at the experimenter node. A communication channel is also included to allow the experimenter to provide procedural instructions to the performers present at each node and send metronome cues at the onset of trials.

1 Selected Locations

From its conception stage, the experiment was designed with the facilities of NYU's Music and Performing Arts Professions (MPAP) in mind. The *Steinhardt Education Building*, located at 35 West 4$^{\text{th}}$ Street in Manhattan, New York, is equipped with a network of acoustically treated rooms dedicated to recording and producing music and hosting live concerts. The facilities comprise a recording studio with annexed "Live Room" (Fig: 35), three ISO booths (Fig: 33), sound production class spaces, lecture halls, and a large reverberant theater located on the ground floor (Figs: 32,33).

The key feature of these acoustically diverse rooms is that they are interconnected through an analog "copper" network that allows sound signals to travel through the building at an analog transmission speed. The network can be accessed and routed from a central location without disturbing parallel work occurring in other connected spaces. As a result, these facilities were optimal for assembling a controlled-latency NMP environment that could be isolated from the actual transmission delays, jitter, and protocol overheads inherent to internet-based communication. Moreover, the facilities in the MPAP department also benefit from the availability of professional and high-quality sound engineering technical equipment.

A total of six rooms played a role in the experiment. The two rooms selected for the distributed phase of the experiment comprise a theater (Figs: 32 and 33) and an ISO booth (Fig:

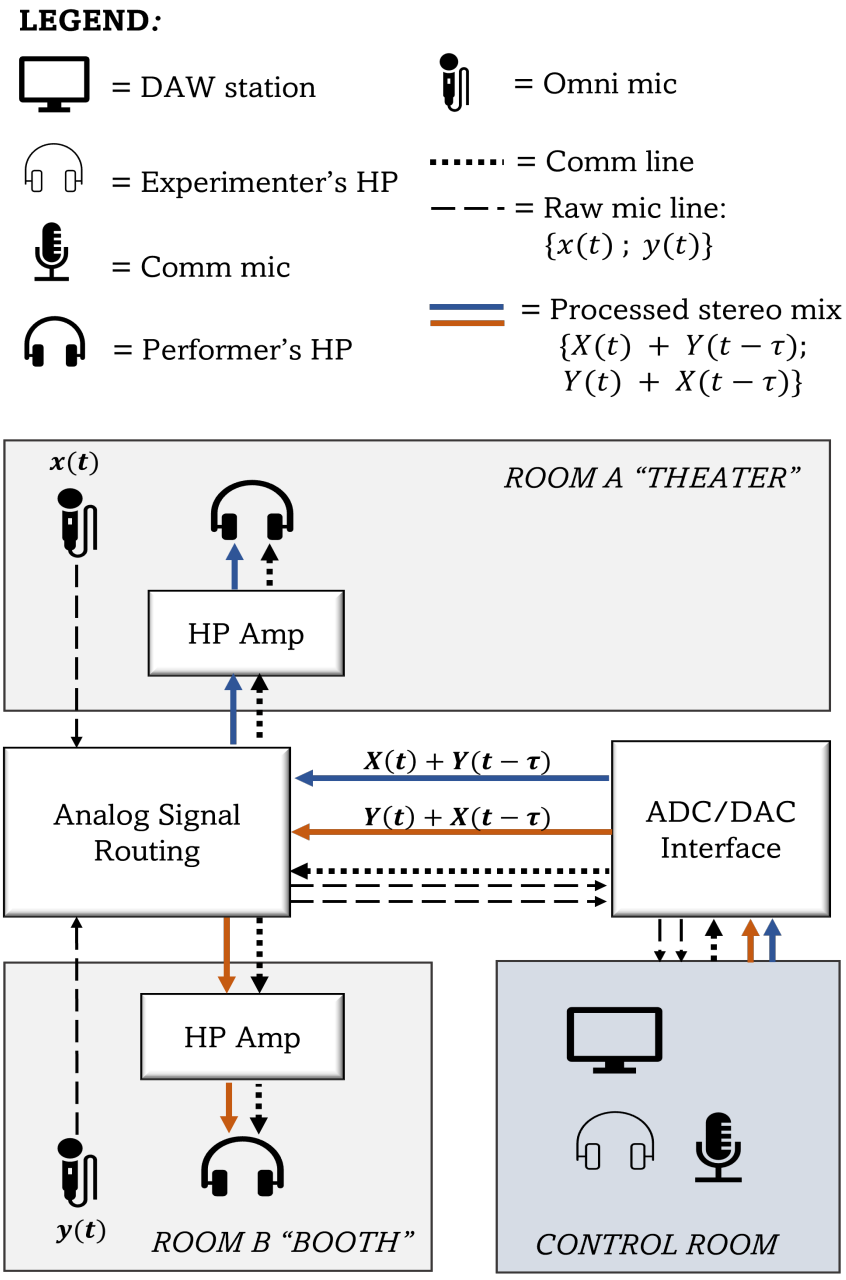


Figure 31: Conceptual implementation target of a three-node star topology network involving two performing locations (*Theater* and *Booth*) and a central control node in charge of recording the raw audio signals, processing the signal with latency and room acoustics effects, and route them towards the opposite node. Each node also receives their own feedback signal (without added latency) with or without room acoustics processing, according to the acoustic environment mode under examination. Reproduction levels are controlled both at the experimenter station and at each node individually.

33). Moreover, BRIR measurements were taken in this room for usage in the *congruent* auralization environment designs, where the acoustic character of the routed signals is adapted to the local performance room. Both rooms can be accessed through a centralized access point to the analog network located in the room labeled "Control Room" (Fig: 34). A live studio room was selected for usage during the pre-experiment baseline stage (see Fig. 35), due to its neutral acoustic character and for being an optimal space for the concurrent recording of clean signals. In this phase participants would perform in a traditional co-located environment for establishing a base feeling of real "presence" while also collecting data signals for collocating objective performance metrics collected across the experiment in relation to the performer's pair musical ability. The two selected lecture halls, relevant for the "non-locally-sourced" *divergent* auralization environments, were selected because they had similar volumetric sizes but different surface materials, leading to different room tones (Figs: 36 and 37).

Table 2 shows details on the dimensions of the room and the average reverberation time for the rooms that were targeted for the study. More information on the auralization environment modes can be found on page 74 of Ch. IV.

Table 2: Rooms employed			
Room name	**Usage**	**Approximate Size**	**Avg. RT60**
Live Room: "Dolan"	Baseline phase	15' × 30' × 9'6"	0.44 s
Theatre: "Frederick Loewe theatre"	Distributed phase and measurements (*congruent modes*)	42' × 66' × 24' (stage area) 83' × 66' × 24' (total area)	1.13 s
ISO booth: "Research Lab"	Distributed phase and measurements (*congruent modes*)	15' × 12' × 8'3"	0.12 s
Large Lecture Hall: "Room 303"	Measurements (*divergent modes*)	36'8" × 32'7" × 13'	0.76 s
Medium Lecture Hall: "Conference Room"	Measurements (*divergent modes*)	33'6" × 30'7" × 11'2"	0.57 s
Mechanical Room: "CMR"	Experimenter control station	N/A	N/A

Table 2: Summary of rooms selected for the experiment, located at NYU's *"Education Building"* at 35W. 4th Street in New York City. All locations are within the same building, occupying different floors, and are connected via a copper wire infrastructure network. Dimensions are in **ft**, RT60 is calculated using the mean RT30 fit of the 500 Hz and 1000 Hz octave bands, taken from omnidirectional room impulse responses.

Figure 32: Theater: *Frederick Loewe Theater*. View from stage. Located at the ground floor of NYU's Steinhardt Education Building in Manhattan.

Figure 33: Theater: *Frederick Loewe Theater*. View towards stage from back corner. The theater is connected via analog wiring to the ISO Booth and Control Room. This space is used as the "Theater" room for the distributed phase of the study experiment.

Figure 33: ISO BOOTH: *Research Lab*. This room is placed a few floors above the Theatre and connected to it via analog wiring through the *Control Room*. Space used as "ISO Booth" room for the distributed phase of the study experiment.

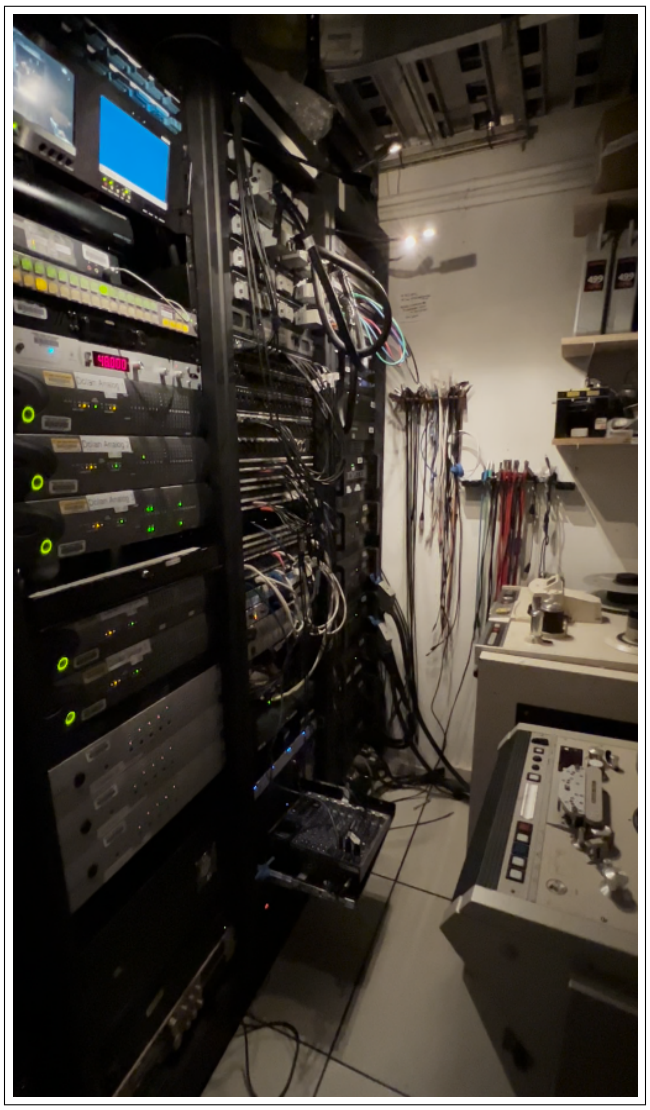


Figure 34: Routing and signal recording room: *Control Room*. The experimenter station was set up in the network wiring control room situated in the same building. This room provides easy access to the copper audio network across the building. All data routing, processing and recording was performed in this location.

Figure 35: Live Room: *Dolan's recording studio*. Used for the data collection process of the co-located baseline phase of the experiment. The room's reflectivity attributes can be controlled and manipulated through the removal or addition of absorption panels and acoustic diffusers.

Figure 36: Large lecture Hall: *Room 303*. This lecture/recital room was measured to collect BRIR acoustic filters employed for the "divergent" modes of auralization.

Figure 37: Medium lecture hall: *Conference Room*. This lecture room was measured to collect BRIR acoustic filters employed for the "divergent" modes of auralization.

2 Acoustic Measurements

In order to drive the virtual acoustic environments (auralization modes) designed for this experiment, *Binaural Room Impulse Responses* (BRIRs) were measured in four different rooms. The theater and ISO booth were measured for their usage in the *convergent* scenarios, while the lecture rooms were measured for the *divergent* scenarios. The goal was to capture the source-receiver acoustic paths representing both the self-produced sound of a performer (sound of their own clap as heard in each room), and the sound of a co-performer as heard by the reference performer within the room. The different examined modes would then be virtually recreated using combinations of room filters. This was achieved by measuring BRIRs at two different locations in each room, where the position of the emitter changed in relation to the microphone, itself collocated at an approximate human seating height (4 ft). A "near-field" measurement was first performed by placing the emitter at a height offset and length-distance offset of 12 inches from the front of the microphone position (elevation angle $\phi = -45^{\circ}$, azimuth angle $\theta = 0^{\circ}$), representing an average hand position used for a "seated clapping stance". The second measurement was placed in the front direction ($\phi = -45^{\circ}$, $\theta = 0^{\circ}$) at a distance of 8ft (meaning a wavefront arrival latency of $7ms$ in standard temperature and humidity conditions), this time representing the virtual location of a co-performer as heard by the reference performer.

The measurement equipment comprised a binaural stereo "dummy head" microphone (Neumann KU100) and a "flat-response" source emitter (Genelec Studio Monitor speaker, model 8030A). The motivation behind the use of a binaural microphone was to more accurately capture the spatial auditory cues embedded in the directional room reflections and uncorrelated diffuse field elicited by the measurement signal, as well as the capture of distance and elevation cues pertaining to the source-receiver positional relationship. In each room, the emitter reproduction level was calibrated to reach 80dB SPL at a distance of 2ft when playing a test pink-noise signal. The level was not changed for the far position to preserve the amplitude relation between the locations in the room. Figure 38 shows a sketch of the measurement setup for this phase. The setup was identical for each target room, the only difference being the reference location of the "center position" of the room or the theater stage where the microphone was set up.

The binaural room impulse responses were collected using the latest version of the

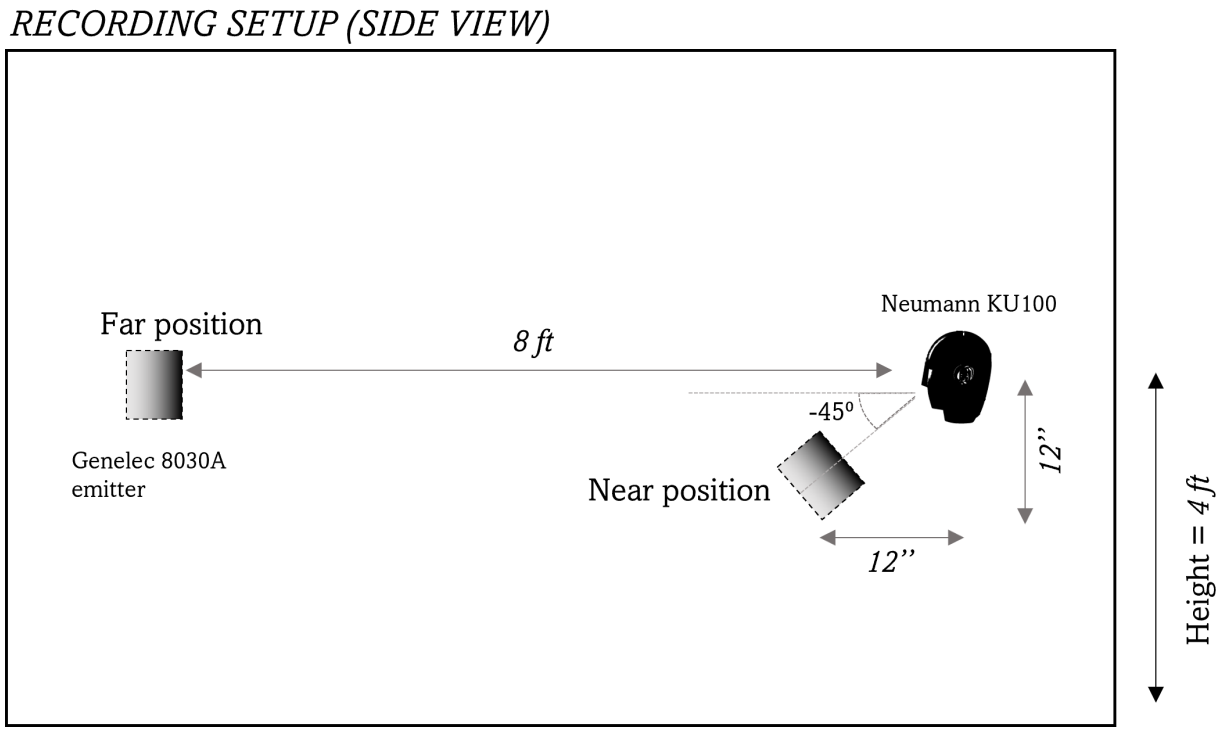


Figure 38: Sketch of the binaural impulse response measurement layout. In each room of interest, a binaural microphone was placed at an approximate seating height in the center of the room (or stage). A near-field measurement was taken at a horizontal and vertical offset of 12 inches and elevation $\phi = -45^\circ$ representing a "clapping" position. A second far-field measurement was taken at a distance of roughly 8ft representing the spatial location of a co-performer within the room. Both measurements were performed at the front direction, azimuth $\theta = 0^\circ$. Exponential sine-sweeps were used as stimulus signals.

"ScanIR" software (Vanasse et al. 2019). In this process, a logarithmic sine-sweep was used as a measurement signal due to its distortion-separation properties that best capture the frequency response of a linear time-invariant system (Farina 2000). The generated sine sweeps were computed to last 3 seconds, and range from 20 Hz to 20 kHz, produced at a sample rate of 48 kHz. For each measurement location, five takes were collected and the results were averaged for the purpose of reducing the influence of spurious noise events. The raw recordings were deconvolved in the frequency domain with the analogous loop-back measurement of the soundcard equipment performed with the original log sine sweep (Chan 2010). This allowed the signal to be equalized for the frequency distortions inherent to the emitter's driving amplifier's electrical components. An inverse FFT step was taken on the deconvolution results to retrieve the time-domain BRIR filters used in the following stages of the experiment. The BRIR measurements were later processed to remove the wavefront arrival latency in order to avoid delay interactions with the artificial delay module used in the routing software. Finally, the measurements were truncated at their reverberation time (chosen as the RT60 taken from the frequency band with the highest T30 value, see Table 3) to remove trailing silences and optimize the real-time processing needed for the distributed experiment.

In addition to the BRIR measurements, the rooms were also measured at the 8ft positions with a stereo pair of omnidirectional measurement microphones that provide an improved dynamic range (see Fig. 44 for technical reference). These measurements were used to extract the acoustic parameters that summarize the sound character of each room of interest. For the omnidirectional measurements, a "Maximum Length Sequence" signal was used (Schröder 1975) due to its appropriateness for parametric extraction tasks. The omnidirectional measurements were extended to the room used for the "Baseline" stage (*Dolan's Live room*). The extracted information is shown in Table 3. The table reports the RT60 reverberation time (average across channels in the 500 Hz and 1000 Hz frequency octave bands), the reverberation time for each octave band from 250 Hz to 4 kHz (average across channels), the *Early Decay Time* (EDT) and direct-to-reverberant ratio (DRR). Due to the limited dynamic range of the measurement microphones, the reverberation times reported are calculated using the linear fit estimation of the RT30 measure (decay time to -35 dB from -5dB), in some cases where the decay rate was very fast (i.e. ISO BOOTH) the RT20 measure was used instead.

Table 3: Measured acoustic parameters

	RT60	RT(250)	RT(500)	RT(1K)	RT(2K)	RT(4K)	EDT	DRR
Large Hall	0.76 s	0.89 s	0.82 s	0.71 s	0.70 s	0.69 s	0.69 s	6.14 dB
Medium Hall	0.57 s	0.59 s	0.58 s	0.56 s	0.53 s	0.56 s	0.46 s	4.04 dB
Live Room	0.42 s	0.38 s	0.41 s	0.42 s	0.47 s	0.39 s	0.38 s	0.57 dB
Theater	1.14 s	1.29 s	1.13 s	1.10 s	1.02 s	0.78 s	0.31 s	5.13 dB
ISO Booth	0.12 s	0.14 s	0.13 s	0.10 s	0.09 s	0.09 s	0.08 s	5.37 dB

Table 3: Acoustic parameters for each employed room, extracted for different octave frequency bands. Results were calculated from stereo omnidirectional measurements and averaged across channels. The RT60 metric is the average of the 500 Hz and 1 kHz band (T30 fit). Metrics extracted through the IOSR library from the University of Surrey (Hummersone 2017)

The final usage of the auralization filters landed on using the "Theatre" room for the *Symmetric Congruent* mode. Out of the divergent rooms, the "Large Hall" room was selected for the *Symmetric Divergent* mode (room 303) due to a less noisy background and the space being a dedicated concert room, while the other available space is defined more as a concert hall.

2.1 Measurement Plots

The following plots illustrate the omnidirectional room impulse response behavior in the time and frequency domain as measured in the three most salient rooms of the experiment. The room that was used for the baseline process, where the participants would perform together in a regular manner, and the two rooms that were used as locations for the distributed phase of the experiment. The frequency domain response is smoothed over $1/4^{\text{th}}$ octave bands and the time domain detail plot shows the first 120 ms from the onset of the impulse. The final plot represents the normalized "Energy Decay Relief" (EDR), also known as *Acoustic Fingerprint* of a room. The EDR represents the decay behavior of sound energy calculated from the Schroeder integration curve (Schroeder 1979) over logarithmically-spaced narrowbands providing a visualization of how fast sound decays in a room at different frequency points[1]. The plot of Fig: 42 shows the details of the RT30 fit used for estimating the reverberation time in the 500 Hz and 1 kHz frequency bands,

[1] Please note that for visualization purposes, the x-axis pertaining to "Time" is not equal across these plots, but capped at the calculated RT60 for each room

the final RT60 measure was derived by averaging the RT30 over these two octave bands and over the two signal channels.

For the full set of plots concerning the lecture hall rooms used for the "divergent" auralization modes, and the full visualization of the BRIR measurements at the "near" and "far" positions, please see section 2 of the Appendix.

3 Headphone Correction Filters

To achieve a neutral reproduction, that avoids unwanted stages of coloration, it is customary to correct the headphone output towards a target (in this case "flat") frequency response. Headphone equalization has been consistently shown to improve perceptual qualities of spatial audio reproduction (Schärer and Lindau 2009). Ideally, to obtain accurate equalization filters tailored to each user, the headphone correction filters should be measured individually over each subject (Pralong and Carlile 1996) for each headphone employed. However, the introduction of this step was not deemed appropriate for the flow of the experiment, making it rather preferable to obtain generalized equalization filters with a smooth response.

Two units of open-back *Sennheiser HD650* were measured and employed for the distributed experiment. The choice of open-back headphones allowed some degree of local diffuse room acoustic field to seep through to the performer while in the "Raw" mode of auralization environment, thus avoiding a directionally collapsed sound field at the reproduction output. This choice also helped to perceptually reinforce the directionality of the performer's own clap in support of the self-monitoring signal. Furthermore, the chosen brand of headphones is renowned for producing a reasonably neutral sound before any equalization (Rämö and Välimäki 2012). This neutral baseline translates to lower magnitude correction needs, making the impact of potential distortions and non-ideal individual fits of digital correction filters less severe. On the other hand, the drawback of utilizing open-back headphones is that the local field can seep through the acoustic path to the listener even when not intended to (i.e. when a BRIR pertaining to a different room needed to be applied). In spite of that, informal tests determined that this undesired effect was negligible, as the spatial auralized signal perceptually dominated and masked the local reflections.

On the other hand, if closed-back headphones were to be used, a potential challenge would regard the direct self-sound potentially not being instantaneously accurate (and the real instantaneous sound coming through occluded by the headphone body). A potential solution would be to separate the direct from the reverberant portion of the sounds, by having an optimized local direct monitoring that bypasses the central node (perhaps with some signal gating to remove reflections), mixed with the reverberant rendered sound (using an impulse response devoided of the direct transient) processed through the network. In an ideal implementation, participants would have been required to switch from open-back to closed-back headphones according to the auralization mode under examination. However, this was not considered practical for the feasibility of the experiment in its current form.

Frequency correction filters were measured by placing the target headphone units on top of a binaural microphone (i.e., Neumann KU100) in an acoustically dry space. Each headphone cup was measured individually using a one-second logarithmic sine sweep, spanning a frequency range from 20 Hz to 20 kHz (Farina 2000) sampled at 48 kHz, and repeated a total of ten times per unit before averaging results. To improve robustness to displacement variations, the headphones were removed and reseated on the dummy head microphone between each repetition of the measurements (Masiero and Fels 2011). The measurements were processed to extract inverse filters from the measured frequency-domain transfer functions, individually for each unit. High-shelf regularization (4 kHz cutoff frequency) and a flat-response reference target curve were applied using methods described in (Schärer and Lindau 2009; Lavoie et al. 2004). Results were converted to minimum-phase time-domain FIR filters. Finally, the filters were translated to equivalent IIR coefficients that could be applied to parametric EQ plugins using the tools published on the *AutoEQ* GitHub repository (Pasanen 2020). The equalization plugin consisted of a 10-band IIR using equivalent coefficients to the measured correction filters, applied within the DAW environment used for processing and routing.

4 Distributed Network Setup

The signal distribution network was built on top of the pre-existing low-latency analog routing channels communicating to the selected facilities. The routing was controlled from an access point location (“Control Room”) in which a machine was set up to collect the incoming audio

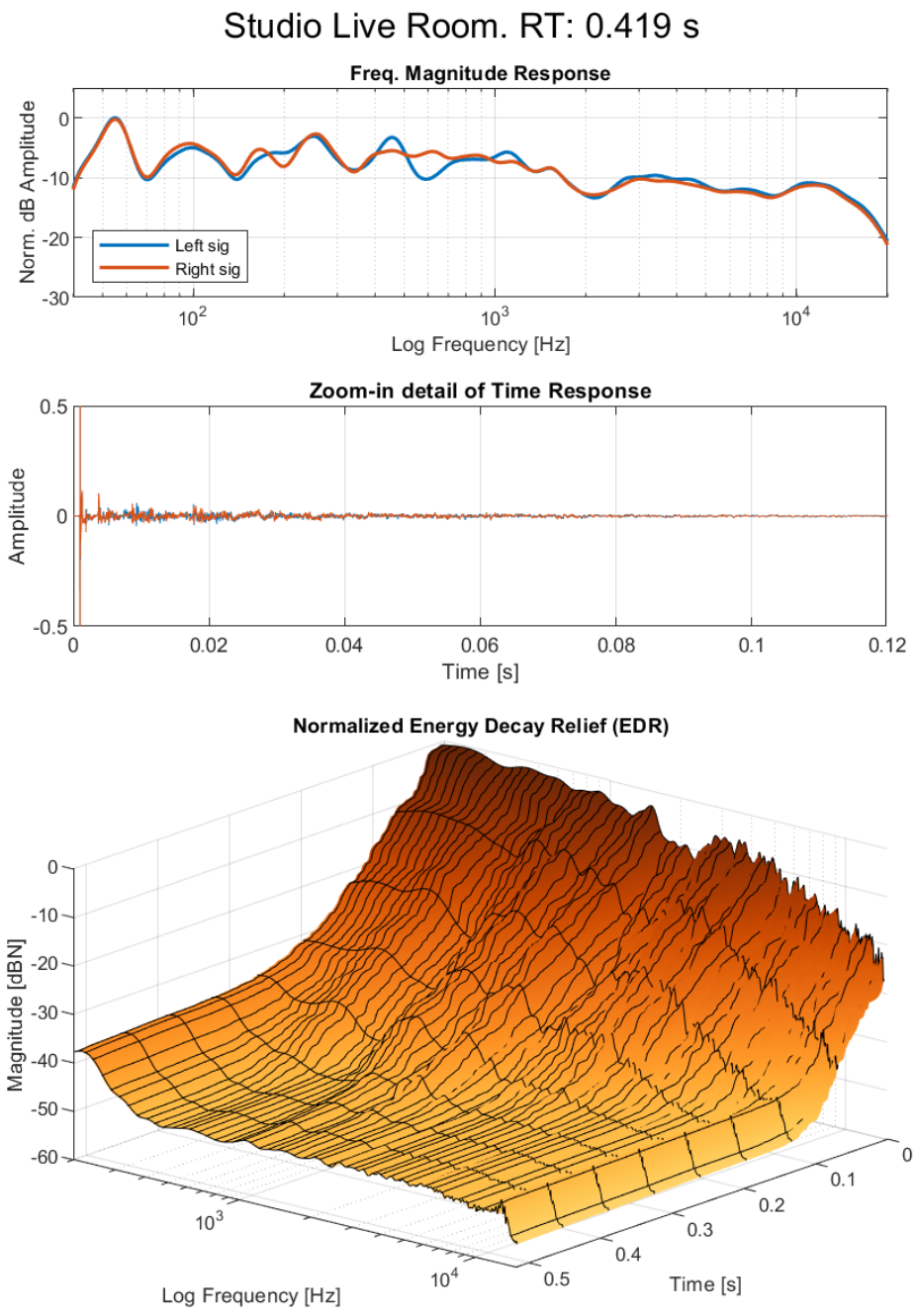


Figure 39: Frequency and Time behaviour of the Live Room used for the baseline study of the copresence experiment. Measurement taken with an omnidirectional pair at 8ft distance from the emitting impulse source. Frequency response is shown smoothed over $1/4$ octave bands.

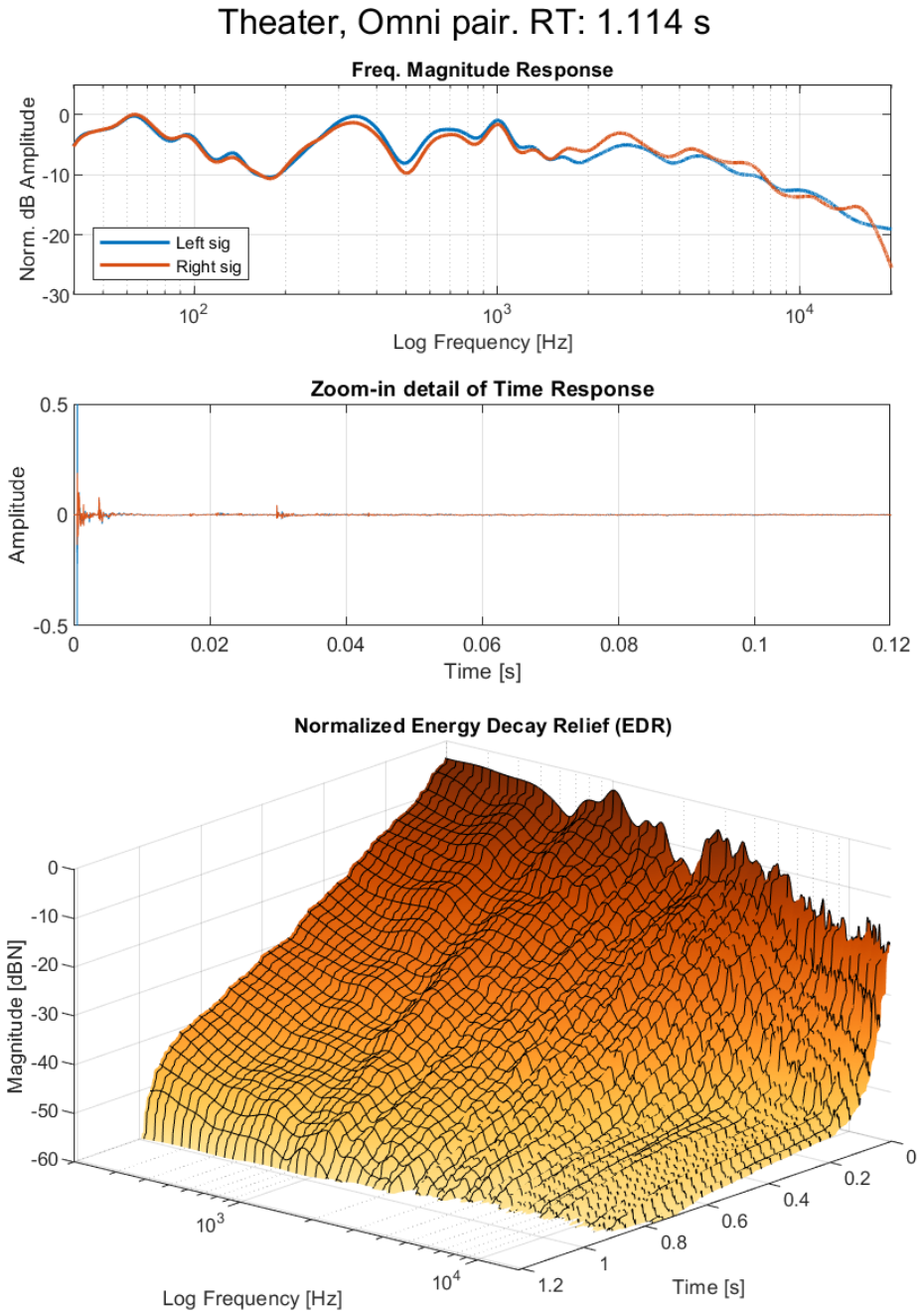


Figure 40: Frequency and Time behaviour of the Theater location ("F. Loewe Theater"). Used as one of the performer locations for the distributed performance phase. Measurement taken with an omnidirectional pair at 8ft distance from the emitting impulse source. Frequency response is shown smoothed over $1/4$ octave bands.

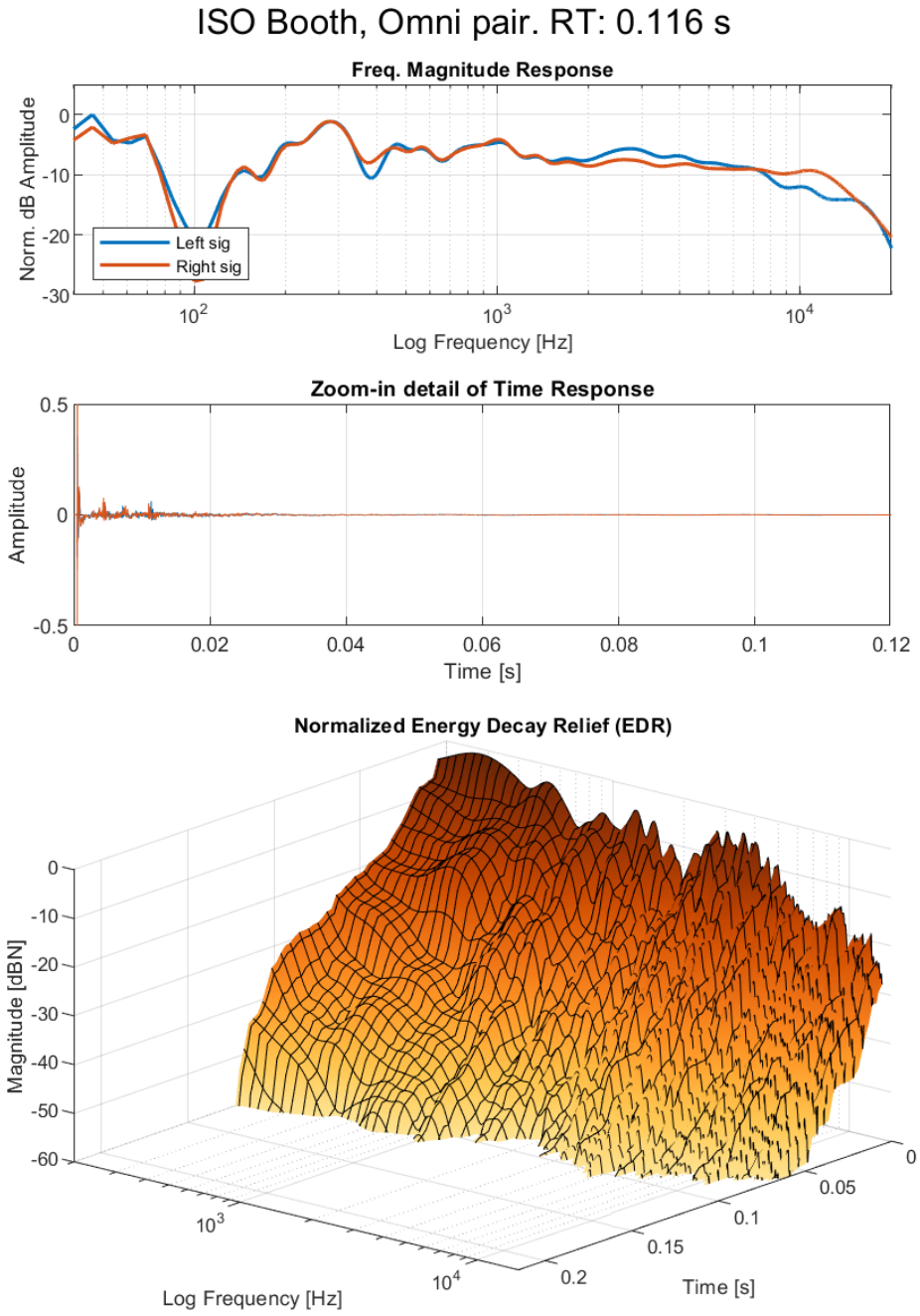


Figure 41: Frequency and Time behavior of the ISO Booth location ("Research Lab') Used as one of the performer locations for the distributed performance phase. Measurement taken with an omnidirectional pair at an 8ft distance from the emitting impulse source. Frequency response is shown smoothed over $1/4$ octave bands.

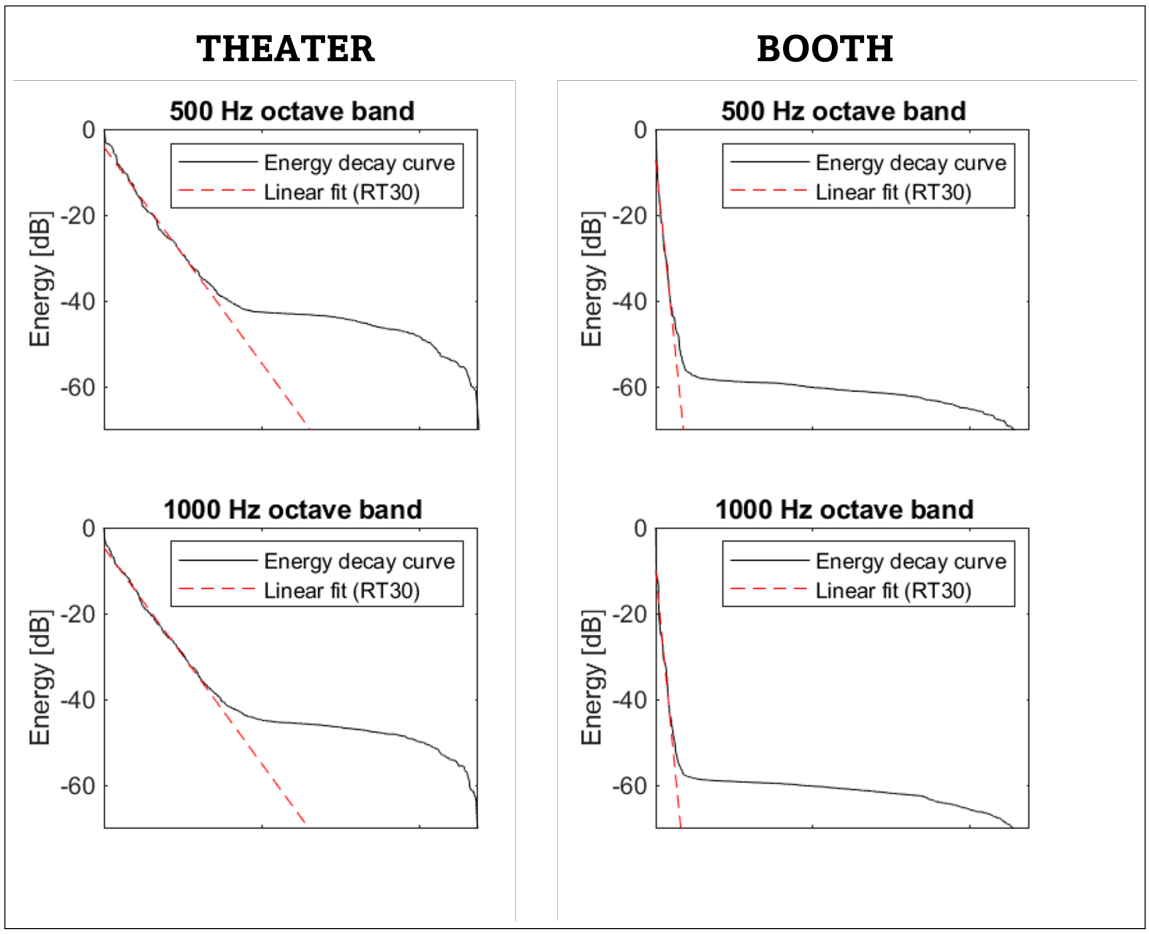


Figure 42: RT30 fit of the Theater and ISO Booth location at the 500 Hz and 1 kHz. These rooms corresponded to the performer locations for the distributed performance phase. Taken from the omnidirectional measurement of an impulse response from a source at 8ft.

signals, record them as raw inputs, process them with artificial latency and room acoustics filters (according to the auralization environment), and finally route them to their respective locations. Figure 43 illustrates the full hardware and software signal path from capture to reproduction. The connection setup described in this figure is illustrated in its finalized form as used for the data collection process described in VI.

The signal of each capture mic in each room is first injected into the analog routing network (EDAC link) and fed to a pre-amp to provide phantom power and control the gain levels of each microphone. The amplified analog signal is then sampled into digital form (sample rate: 48 kHz) through a low-latency high-quality converter unit which connects via an optical link to a USB audio interface. The interface is also a professional broadcasting-quality system able to handle buffers as low as 32 samples for real-time processing of audio channels. At this stage the signal enters the "software path" where it is internally routed into a DAW. The DAW environment is in charge of recording, processing, mixing and routing of the signal mixes dedicated to each receiving room (see sect. 4.1 for details on the processing DAW plugins). The performance recordings are taken raw before any processing is applied.

The processing blocks comprise a *noise gate* to remove local reflections when detrimental to the realization of certain auralization modes, a Short-Time-Fourier-Transform (STFT) FIR stereo *convolution reverb*, an *artificial delay* plugin, and finally an IIR *EQ filter* derived from the stereo calibration measurements. Each node mix comprises a "self-monitoring" signal (not passed through the delay block) and a "co-performer" signal (plus the communication channel whenever used). According to the auralization mode activated at each trial, the self and co-performer signals are processed with different stereo BRIRs respectively belonging to the "near" and "far" measurements. An unprocessed hard-panned stereo mix of the captured signals is also created for the experimenter in order to monitor the ongoing process of experiment trials. The final mixes are routed out through dedicated hardware interface output ports, converted back into the analog domain, and sent through their respective destination channels (organized in L/R stereo for each output). The L/R channels are finally collected in each destination room and sent to locally-placed headphone amplifiers capable of providing mix-level controls at each node.

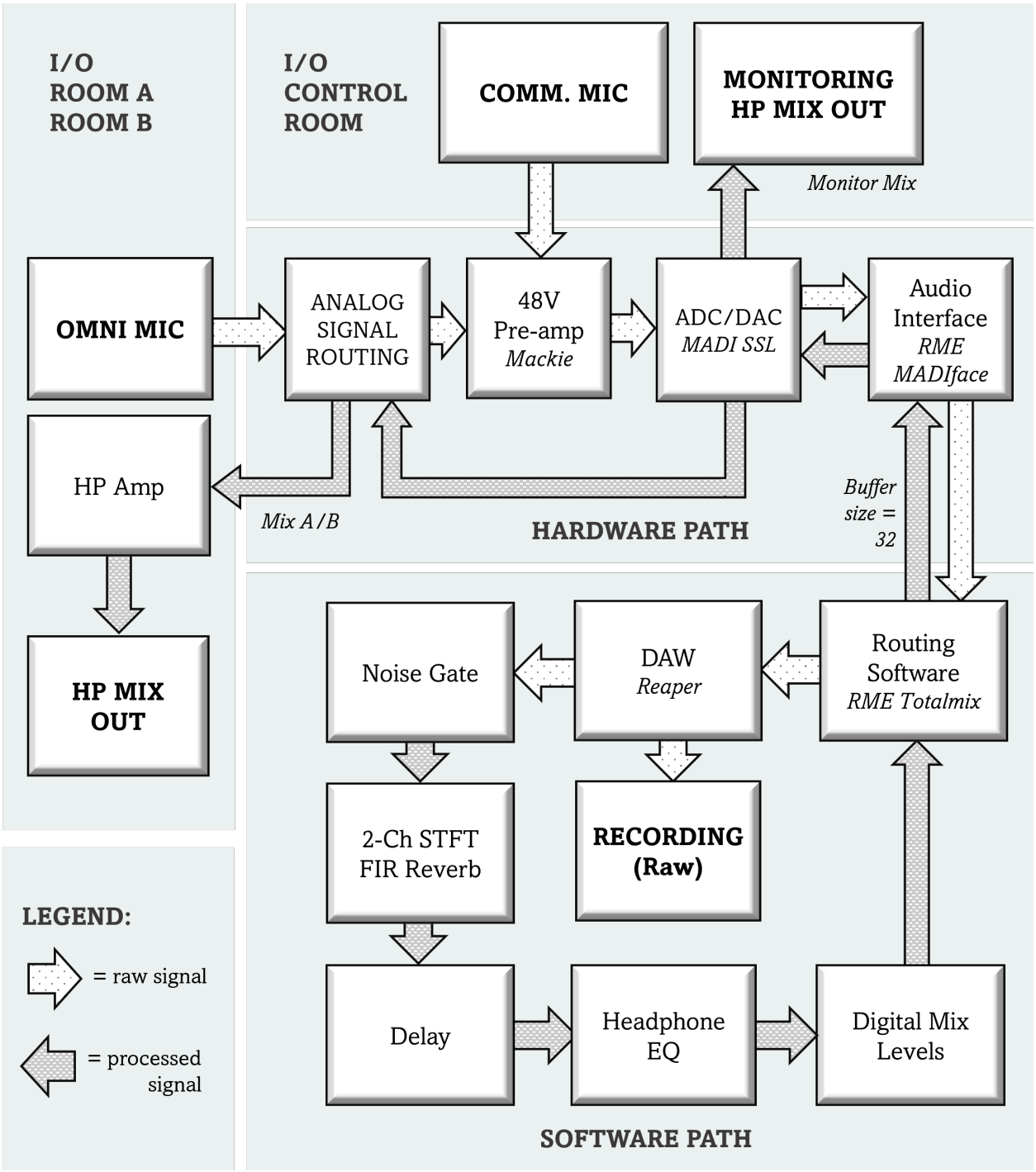


Figure 43: Full signal path showing the raw signal flow from the experiment rooms (Rooms “A” and “B”) to the recording station, while a processed version gets sent back to the connected rooms. Signals from the experimenter are also injected into the output hardware path to allow procedural instructions to be heard over headphones. The exact software processing path on each signal varied according to acoustic mode, originating room, and whether it was mixed as a “self-monitoring” signal or a “co-performer” signal.

4.1 Software Environment

The digital signal processing path is handled in the *REAPER* DAW environment (*REAPER, Digital Audio Workstation* n.d.). The DAW session is set up to handle three stereo output mixes (mix for "theater", mix for "booth", and "control room" mix) from three different inputs (mono capture in each room). From the perspective of each node, each mix contains the self-produced signal, processed according to the desired auralization mode using a "near" BRIR filter, and the co-performer signal, processed with a "far" BRIR filter plus one of three possible artificial delay levels. All processing is handled by using proprietary VST plugins optimized for usage with the REAPER DAW. A recording session template was created for the fast handling of different processing settings across trial conditions comprising different modes and latency levels. This is achieved by setting up easily navigable "regions" that can be automatized to activate different combinations of plugin settings according to the playback marker position in the session timeline. The region system is numbered accordingly to randomized lists of trial-execution orders specially created for each pair of participants. The experimenter's task is thus reduced to the manual navigation of the correct region at each trial in the list order.

In regards to the signal pipeline, the first block of digital processing consists of a noise gate set at -38dB which serves to remove low-amplitude reflections from the captured signal before applying BRIR processing. This step ensures that no collapsed directional reflection energy is fed through the BRIR filters creating perceptual artifacts due to interactions with the embedded spatial reflection binaural cues. This step is not necessary for the signal captured in the ISO booth room since the reflection levels pertaining to that space are negligible. Four instances of an STFT-based stereo convolution reverb plugin are set up to handle the different "near" and "far" filter combinations for each send channel. The plugin algorithmic latency is minimized via the *Low-latency (LL)* and *Zero-latency (ZL)* options which activate extra CPU threads for faster processing (Francis n.d.). Thanks to interface compatibility, 32-sample buffers could be used for minimized buffer delay. The noise gate and reverb processors are not active for the "Raw (R)" mode, as that specific modality is intended to send signals as they are captured.

Before mixing, the "co-performer" send of each mix is passed through an artificial delay plugin that inserted additional signal delay in accordance with one of three desired latency levels (namely "7 ms", "20 ms", and "40 ms") depending on the current trial activation. The base system

latency was taken into account when tuning the specific settings, to obtain the desired one-way total arrival latency at each node. For details on the system latency measurement please refer to section 4.2. The mixes are then passed through a bus-level parametric EQ effect, which applied the headphone correction filter in IIR form (tuned individually per headphone unit, so each bus has a different EQ setting) in order to compensate for the coloration introduced by the reproduction drivers.

To cue the start of a performance at each trial, a click metronome cue signal is placed on the mix bus pertaining to the performer assigned to the "Static Part", in practice this meant that for the first half of the trial set the metronome cue is sent to the mix destined to the ISO Booth, while for the second half of the trial-set, the cue is sent to the Theater mix. The metronome is set to play a 4-beat cue for four bars at 85 BPM. The third mix created for the experimenter is void of any processing besides hard-panning the performer signals to the left and right output channels. The purpose of this mix is to monitor the correct procedural progress of the experiment and quickly troubleshoot occasional faults.

4.2 System Latency

The round-trip network system base latency was measured in order to evaluate the pipeline implementation and to calibrate the actual artificial delay parameters needed to obtain the desired one-way interaction latency levels. The latency measurement was conducted by creating a loop-back connection to the target rooms (one at a time) through the network environment described in the previous sections. The REAPER DAW was configured to not compensate for the interface delay, while keeping the processing plugins active and configured in the same way intended for the experiment session (48 kHz sample rate), minus the artificial delay plugin (worst-case scenario). To include the electric process path of the playback and capture components, the signal was reproduced through a sound emitter in close contact with a microphone. An impulse signal was transmitted through this path and simultaneously recorded through a different track. In summary, the full loop-back latency measurement path started in the measurement software, including plugins, DAC converter, analog path, headphone amp, emitter, minimal acoustic path, back into the analog circuit path, pre-amp, ADC conversion, interface, and software recording.

To calculate the latency amount, the recording and stimulus signal was fed into a cross-correlation algorithm in MATLAB which provided the amount of samples displacement between the two. Since this number represented the round-trip latency of the system, the one-way latency measure was taken as half of that value. Values were found to be close enough across the Theater and Booth rooms, thus a unified value was taken as average. Over a number of takes, the average round-trip latency value was found to be 4.6 milliseconds, meaning a one-way system latency measure of roughly 2.3 milliseconds, of which 1.3 were assigned to the USB interface (reported through the driver software).

Table 4: System latency and actual delay level parameters

System latency	**Value**	
Round trip latency	4.6 ms	*(applied offset value)*
One-way latency	2.3 ms	
USB Interface	1.3 ms	
Trial condition	**One-way interaction latency**	**Plugin adjustment parameter**
"Acoustic latency" level	7 ms	2.4 ms
"Mild latency" level	20 ms	15.4 ms
"High latency" level	40 ms	35.4 ms

Table 4: Measured system latency and delay plugin calibration parameters for simulation of three network-latency levels

The round-trip path measured with the loop-back method begins at the central node, passes through an end node (A or B), and returns to the central node. Although in a different order, the same LTI components are utilized when a signal travels through this network from one room end node to the other. This makes the loop-back round-trip path equivalent to the sum of the one-way latency from node-A to the central processing node, plus the one-way latency from the central processing node to node-B (this total is labeled as the "one-way *interaction* latency").

Due to practical complexities, it was not possible to isolate the latency contribution of each system component and calculate the latency breakdown of each sub-path. The assumption that the one-way system latency is equal to half of the total system latency is an approximation because certain components are not symmetrical between the room-to-machine and machine-to-room directions. However, this is anticipated to be a negligible factor, as the routing process pipeline

is roughly symmetric from A-to-B and from B-to-A, with differences measured in the order of fractions of milliseconds. The working assumption is thus that the total latency time for the musician in room B to hear the signal produced by the musician in room A is approximately equal to the round-trip system latency measured with the loop-back method. Finally, the average round-trip system latency value was used to offset the parameters of the artificial delay module to achieve the desired one-way performer-to-performer interaction latency levels determined in the design stage.

Table 4 shows the measured system latency levels and the actual delay parameters used to recreate the desired latency levels under study. The system was deemed successful as the round-trip system latency was below 7 ms, which was the minimum one-way interaction latency level needed for the experiment design implementation (corresponding to the acoustic path of two performers placed at an 8ft distance). It was noted that the base system latency undesirably affected the self-monitor signals, which ideally should embed only the purely acoustic delay. Nevertheless, during pilot trials, this base latency was deemed low enough to allow the physical acoustic path of the performer's sound to temporally integrate with the monitor feedback without perceptually noticeable disturbances.

5 Equipment Summary

Table 5 illustrates the full list of equipment employed during the measurement stage described in this chapter, and the primary data collection stages described in Ch. VI.

Table 5: Full Equipment list

Item	Brand/model	Usage
MEASUREMENT STAGE		
Impulse Emitter	Genelec 8030A	Loudspeaker for emission of measurement signal
Receiver	Neumann KU-100	Binaural capture of BRIR responses.
Measurement Software	ScanIR (MATLAB) and REW	Measurement and processing of impulse responses (Vanasse et al. 2019; Mulcahy 2022)
SPL calibrator	-	Used to calibrate the measurement signal energy emitted from the source
Soundcard	Scarlett Focusrite 4i4	Analog-digital conversion and interface to computer
Stereo Headphones (x2)	Sennheiser HD650	Headphone filter correction measurement

continues on next page

BASELINE CAPTURE STAGE		
Cardioid Dynamic Microphones	Sennheiser MD421	Capture of live signals
Loudspeaker	Genelec 8030A	Metronome playback
DAW software	Reaper	Recording of signal and metronome playback
Audio interface	Behringer UMC404HD	Soundcard interface
EXPERIMENT STAGE: PERFORMER ROOMS		
Notebooks (x2)	-	Questionnaire filing
Omnidirectional microphone (x2)	Earthworks M30	Capture of live signals
Open-back Headphones (x2)	Sennheiser HD650	Routed signal reproduction and self-monitoring
Headphone amplifier (x2)	Behringer AMP800	Self-adjustment for signal level
EXPERIMENT STAGE: CONTROL & ROUTING ROOM		
Cardioid Microphone	Shure SM58	Talkback communication
Stereo Headphones	Sony MDR750	Signal Monitoring and communication
Soundcard interface	RME Madiface	Digital interface between the computer and analog system
Analog/Digital conversion	Madi SSL	ADC and DAC signal conversion
Mixer	Mackie 1202VLZ4	Phantom power and analog acquisition gains
DAW software	Reaper	Recording of raw signals, processing and routing of transmission signals
Soundcard interface software	RME Totalmix	Sample rate control and digital routing gains
Computer	iMac 2021 (M1 processor)	Central computing machine

Table 5: Complete list of equipment used for the measurement stage and data-collection stage of the methodology process.

5.1 On Capture and Reproduction Equipment

To complete the setup of the experiment, the appropriate capture equipment was selected based on specific design requirements. In the case of distributed performance rooms, omnidirectional flat-response condenser microphones were selected microphone to best capture the nuances of the diffused local reflections that occur when audio is bounced around a large space (Earthworks M30, reference schematic in Fig: 44). Additionally, flat-response microphones are designed to capture sound as accurately as possible, without adding any additional coloration to the signal.

The presence of local reflections in the signal transmitted from the theater towards the booth was necessary to implement the “Raw (R)” and “Symmetric Convergent (SC)” acoustic environment modes. For all other modes, reflections were cut off using a digital gate as shown in Fig: 43 to avoid the double processing of reflections with a BRIR filter. No particular requirements were applied to the choice of communication material for the experimenter’s talk-back channel.

For the case of the baseline, the “co-located” part of the experiment, it was not important to capture the local room sound as there was no transmission involved. However, it was necessary to capture clean signals for a smoother and more robust objective metric extraction performed in later stages (Ch. VI, Sect. 4), making it important to minimize signal bleed and background noise levels. Hence, dynamic cardioid microphones (Sennheiser MD421) with high SPL capacity were used to primarily capture percussive sound occurring directly in front, while rejecting sound that came from other directions.

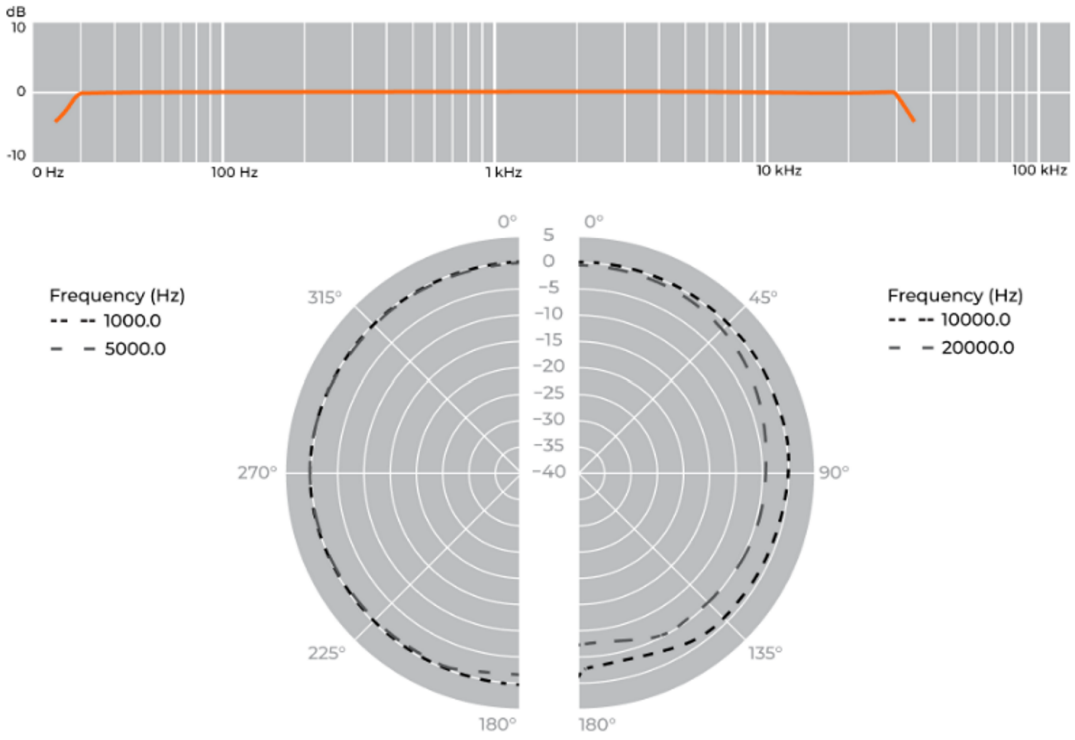


Figure 44: Frequency and directivity response of Earthworks M30 microphones for capturing and transmitting signals in the distributed phase of the experiment. Image from: (Earthworks 2022)

6 Pilot trials: Tuning and Validation

To verify the correct functioning of the system and perform tuning operations, a pilot test trial was carried out[1] with the assistance of students of NYU's Percussion program. The pilot mainly served to validate the correct setup of the experiment, perform adjustments of the signal flow (already presented in its finalized form in Sect. 4), calibrate recording levels, and troubleshoot minor equipment faults. Furthermore, participants in the pilot test helped assess the feasibility of the methodology, define the level calibration procedure concerning each node, and establish the procedure around the use of a metronome cue. During these trials, informal tuning of the maximum viable latency level took place, and it was decided to cap it at 40 ms for total one-way interaction latency.

A signal level calibration procedure was established with the participants to account for variations of clapping motions observed across subjects (i.e. preferred posture and strength of hand impact), minimize noise, and obtain consistent levels across the nodes. This was addressed by the implementation of a calibration process based on an exchange between the experimenter and the performers, where the self-feedback levels were set by the participant thanks to the headphone amplifier, and the routing levels were calibrated from the control room (more details are provided in Ch. VI, Sect. 2.3.1). It was also established that a count-in metronome could not be fed to both players in the presence of high signal latency as it created confusion at the onset of the performance, so the metronome signal cue was exclusively routed to the static-part performer with the instruction of providing a vocal "count-in" over the audio network. These considerations directly fed into the establishment of the data collection procedure; refer to figure 51 (p. 136) in Ch. VI for a more detailed explanation of the finalized experimental procedure.

After informal assessments, modifications to the score as previously described in Ch. IV, sect. 4.5 were applied to keep each repetition of the piece around a minute in length and to avoid over-fatiguing the performers and keep the total length of a session between the two and three hours mark. For the same reason, repetition trials of each combination of latency level (three levels) with the auralization modes (five categories) had to be limited to a maximum of two, thus

[1] An illustrative video taken during pilot tests can be found at this link: https://youtu.be/EtCHOFCylTc.

leading to the decision of having performers switch rooms once, between the completion of a full set of conditions and their repetition.

The final adjustment that resulted from the pilot trials regarded the room used for the Baseline-phase, which was initially different from the one later used in the experiment. Initially, the lecture hall ("room 303") was selected due to its smooth acoustics, which was deemed optimal to establish a baseline reference sense of co-located room acoustics in participants. However, this choice proved difficult to deal with in the signal analysis layer, where it was found that significant signal bleed was caught by the microphones because of the high acoustic reflectivity level. Although this could have been partially corrected by smart noise-gating, the variability of clap onset strength (due to the varying clapping energy and distance to the microphone) observed in the recorded signals led to the selection of the less reflective *Dolan's Live Room* for the baseline phase of the experiment. In this room, the issue of signal bleeding could be significantly reduced.

CHAPTER VI

DATA COLLECTION METHODOLOGY

This chapter provides a comprehensive overview of the multiple stages involved in data collection, along with the methodologies and procedures employed during each phase. Broadly speaking, the data is classified into two main categories: "Primary data", which comprises raw audio data gathered during the experimental sessions regarding the distributed network, and secondary "Evaluation data", which encompasses three distinct layers of assessments conducted on the primary data as well as the subsequent statistical analyses.

The three evaluation layers include subjective evaluations of individual trials conducted by the performers themselves, third-party annotations and ratings that offer an external perspective from expert listeners, and the extraction of objective performance metrics that allow for quantitative analysis. By combining these diverse sources of information, it is possible to obtain a more holistic understanding of the subject matter and derive valuable insights from the collected data.

1 Data Layers

1.1 Primary Data

The primary data consists of the raw unprocessed audio signals gathered during two experiment phases, the "baseline phase" where performances were recorded in a regular setting, and the "distributed phase" where performances occurred over the network described in Ch. V. The initial baseline phase consisted of recordings of co-located performances by the participant pairs executed in the traditional sense of musical interaction. The main purpose of this phase was to control for the technical musical ability of each pair of musicians. The metrics of the objective evaluation layer extracted from these signals would later be used as a reference criterion for the calculation of the distributed performance metrics observed under the examined conditions

under study. This allowed to portray the effect of the examined conditions on distributed performance in relative-deviation terms rather than absolute terms. The secondary purpose of the baseline phase was to elicit a recent inner sense of "copresence" into participants for later reference when evaluating the trial experience through the subjective evaluation layer.

The distributed primary data represent the bulk of the total gathered data. Each pair of participants performed the selected piece a total of 30 times: under three network latency levels and five auralization environment modes (in randomized order), repeated twice. Their captured audio signals were recorded as they arrived at the experimenter control station. In conjunction with the distributed-performance data collection, participants evaluated their experience through a questionnaire labeled *"Trial Questionnaire (Q2)"* that formed one of the evaluation layers. The distributed primary data was subsequently post-processed (through different pipelines) for use in the "third-party evaluation" layer and "objective metric extraction" layer.

1.2 Evaluation Data Layers

The full set of evaluation data incorporates the data layers over which the statistical analysis is ultimately performed in the form of Mixed Effects Models and correlation analysis (see Ch. VII). The principal focus of the analysis was to understand the impact of latency and acoustic scenarios on NMP performance in terms of each of the evaluation systems presented. Furthermore, part of the research questions posed in this dissertation is the relationship between different evaluation perspectives, in order to understand how the concept of "quality" in the context of NMP can be correlated among subjective and objective dimensions of evaluations, each able to highlight different aspects of the performance and establish if some metrics can be used as proxies to determine others (in particular "copresence"). Sections 3, 5, and 4 provide detailed explanations on each respective layer.

The first layer of evaluation data consists of subjective questionnaires collected in conjunction with the primary distributed data, representing the experience of the participant after each performance trial. After the completion of each trial, participants were instructed to complete a new evaluation instance of the questionnaire related to the trial they just completed. The related questions consisted mainly of scale ratings of their subjective feelings of "copresence" and "cohesion" (in different nuanced variations) and self-judgment of the quality of their

performance under the presented distributed environment conditions (latency and auralization mode).

The second layer of evaluations consists of objective metrics related to performance quality that were extracted from the primary data. Given the nature of the performance piece, the metrics were largely based on the rhythmic quality of each performance. The process revolved around the tracking of a *dynamic tempo curve* showing the individual tempo progress over time. A series of metrics summarizing tempo stability, accuracy, and deviation were then extracted from the dynamic tempo curve. Synchronization metrics were also extracted by computing the beat alignment from the perspective of each node in the network.

Finally, third-party annotations of the primary data signals were sourced from musically literate expert listeners. The material provided to the annotators consisted of pair-performance takes reprocessed in stereo, but without including any of the latency and acoustic-mode rendering heard by the performers at the time of collection. This ensured that the evaluation of the take would be "blind" to the condition under which the trials were recorded. This layer was subdivided into "mistakes/inaccuracies annotations" and "ratings". The mistake annotations provided binary data on whether the annotators heard certain types of technical mistakes, not easily detectable by objective signal analysis, in each evaluated recording. The ratings consisted of numerical scale evaluations concerning different qualitative aspects of the performance as judged by the expert listeners.

1.3 Support Data

Additional support data was collected in the form of a pre- and post-experiment questionnaire focusing on the demographics of the participants and qualitative feedback. These questionnaires are here labeled, respectively, as: *"Demographic Questionnaire (Q1)"*, and *"Debrief Questionnaire (Q3)"*. Both questionnaires were designed to capture possible confounding variables, as identified in (Lee 2020). Questions related to Q1 questionnaire were specifically designed to capture and identify possible emerging classifiable demographic attributes for categorizing participants into groups. The presence of certain participant attributes could indicate the presence of a potential response bias that would need to be considered as a random effect in the mixed effects models described in the Analysis chapter (Ch. VII). The underlying motivation is that the complexity

of the experiment would make it susceptible to possible interference of confounding factors. For instance, one such dependency that may arise could be related to participants' familiarity with the chosen performance piece (*Clapping Music*). It is plausible that individuals with high familiarity levels may exhibit a different response pattern compared to those with a lower degree of familiarity. This divergence could be particularly noticeable in aspects such as response to latency levels, where a person's prior knowledge of the piece or experience might directly influence their ability to perform under adverse conditions. Identifying these factors allows one to partially control for confounding variables (if enough data is available over the representative classes) and obtain a more accurate and reliable understanding of the true nature of the relationships being examined, allowing the development of more effective analysis models.

The debrief questionnaire (Q3) was instead directly related to the research questions explored in Q2 and served to provide high-level impression trends to support the data analysis with an additional validation layer and collect additional comments in regard to the distributed experience. This questionnaire was completed at the end of the experiment session. This additional layer of data, collected at the end of the experiment session, significantly enhances the depth and robustness of the analysis and ultimately contributes to a more thorough and accurate understanding of the subject matter. Within this questionnaire, an additional classification attribute (as with Q1) was collected in the form of *fatigue level*, in order to control for the potential effect of fatigue on performance metrics.

2 Primary Data Collection

This section concerns the methodology and procedural flow related to the "primary data" (i.e., the unprocessed recordings of the distributed performances under different conditions of latency and auralization environment modes) which is itself subcategorized as "baseline phase data" and "distributed phase data". Details of the experimental procedure of these two subphases of the data collection, an overview of the experimental session flow, and insights into participants' demographics are provided below. The data connected to the *"Demographic Questionnaire (Q1)"* and the *"Debrief Questionnaire (Q3)"* are addressed in this section. The primary data collection is also intertwined with the "Trial Questionnaire Data" as that evaluation layer was collected in conjunction with the experiment procedure. For clarity, details concerning the *"Trial*

Questionnaire (Q2)" are portrayed in its own section (Sect. 3). The questionnaires can also be found in the Appendix section of this document.

2.1 Participants

A total of 32 unique participants participated in this study, composed of 16 pairs of Static-part players matched with Shifting-part players. Participants were recruited through emails distributed to the NYU Music and Audio Professions programs (no external participants could be recruited due to the measures against the spread of COVID-19 taken by the university institution at the time). The two requirements for participation were first to have at least four years of music performance experience and secondly to have a musical literacy level high enough to read the score of the chosen piece. Once individual contact was established with the interested parties, each potential participant was provided with audiovisual material and a copy of the piece score to learn and choose a part to play (static vs. shifting role). Pairs were formed according to the part-playing preference and availability of each respondent. Once confirmed, each participant received a consent form and a first questionnaire (*"Demographic questionnaire (Q1)"*) to complete online. The purpose of Q1 was to collect descriptive attributes for each participant, for later factorization in the analysis by creating user-classification categories. Figures 45-46 illustrate some of the demographic information collected in Q1 on participants' fields of expertise and years of musical experience. The questionnaire also included Likert-style questions on the participants' self-ratings of potential response bias factors, categorized as "familiarity", "experience in NMP" and "expectation bias". Details are shown in Sect. 2.1.1.

On the day of the session, the participants had the opportunity to rehearse the piece together in the room where the baseline recording was planned, repeating the rehearsal as much as necessary to reach a degree of performance level of their satisfaction. Subsequently, the experiment proceeded with the baseline phase followed by the distributed phase of the experiment as indicated by the flow diagrams in Figs. 50 and 51 (a rest break was introduced in the middle of the distributed phase when the performers switched rooms for the second set of trials). The data from the *"Trial Questionnaire (Q2)"* was collected during the distributed phase as indicated by the flow diagram figure. At the end of their session, each participant was debriefed and guided through the completion of the third questionnaire *"Debrief Questionnaire (Q3)"*. The

participants were compensated $\$50$ for their participation. The average duration of an experiment session, including the distributed part and the baseline part (excluding the technical setup), was 2 hours and 30 minutes.

2.1.1 Demographic Questionnaire Responses

This section covers the responses collected from the *"Demographic Questionnaire (Q1)"*, taken by each participant before the primary data collection. In addition to collecting the "field of expertise" and "years of musical experience" from the subjects (Q0.1 and Q0.2, shown in Fig: 45 and 46), the questionnaire polled the participants on categories of questions related to "familiarity" (Q1.1 to Q1.3), "NMP experience" (Q1.4 to Q1.6) and "bias" (Q1.7 to Q1.8), using a 5-point Likert-style self-rating scale. Results were later aggregated per participant to get a single score value for each category and create different sets of classification groups to support the analysis. These categories were chosen to highlight potential sources of confounding interference (or random effect) toward objective performance quality or towards the responses provided in (Q2). The full set of questions covered in this questionnaire is shown in Table 6. Responses are shown in this chapter as they illustrate the classification attributes of the participants involved. Figures 47, 48 and 49 show the distributions of the ratings for each of the three categories. High response variance is found in the "NMP Experience" category and in the familiarity question concerning the familiarity with the musical piece.

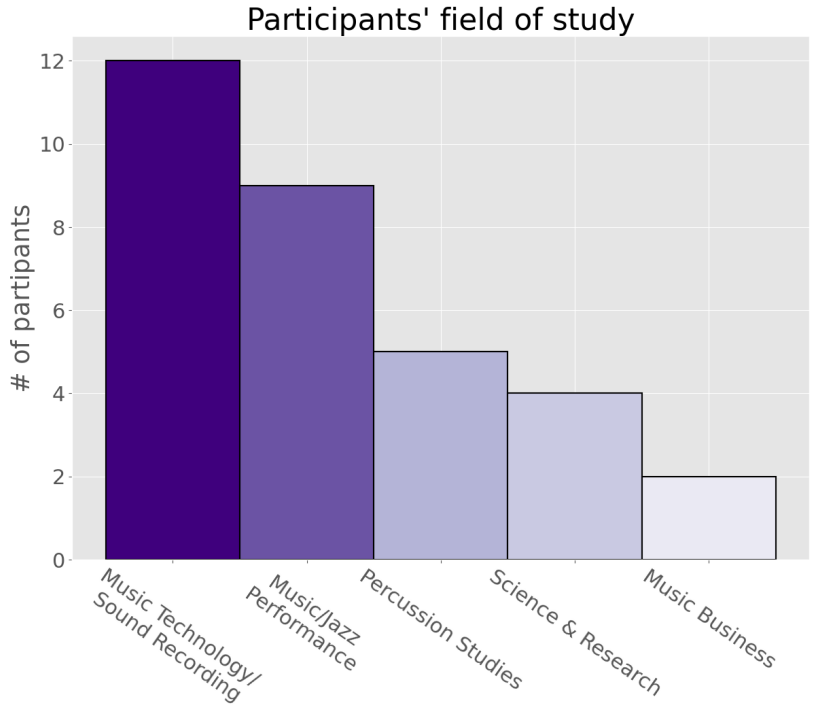


Figure 45: Barchart illustrating the field of study/profession for all participants

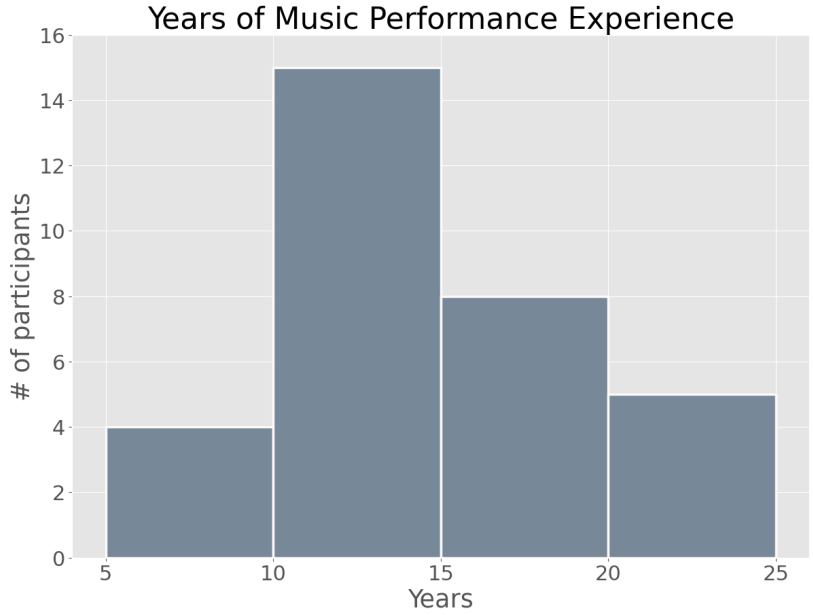


Figure 46: Histogram representing the years of musical experience of the participants

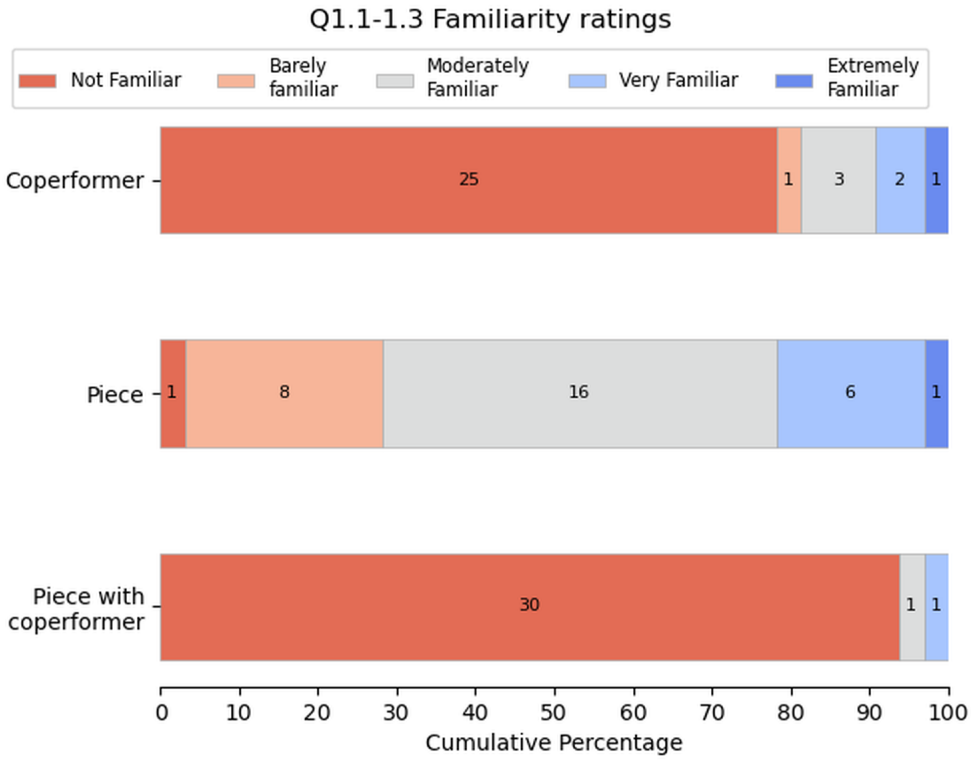


Figure 47: Responses for **Q1.1, Q1.2, Q1.3.** Proportionality-plot showing ratings distributions for participants' familiarity with the co-performer as musical partner, familiarity with the piece, and combination of the two

.

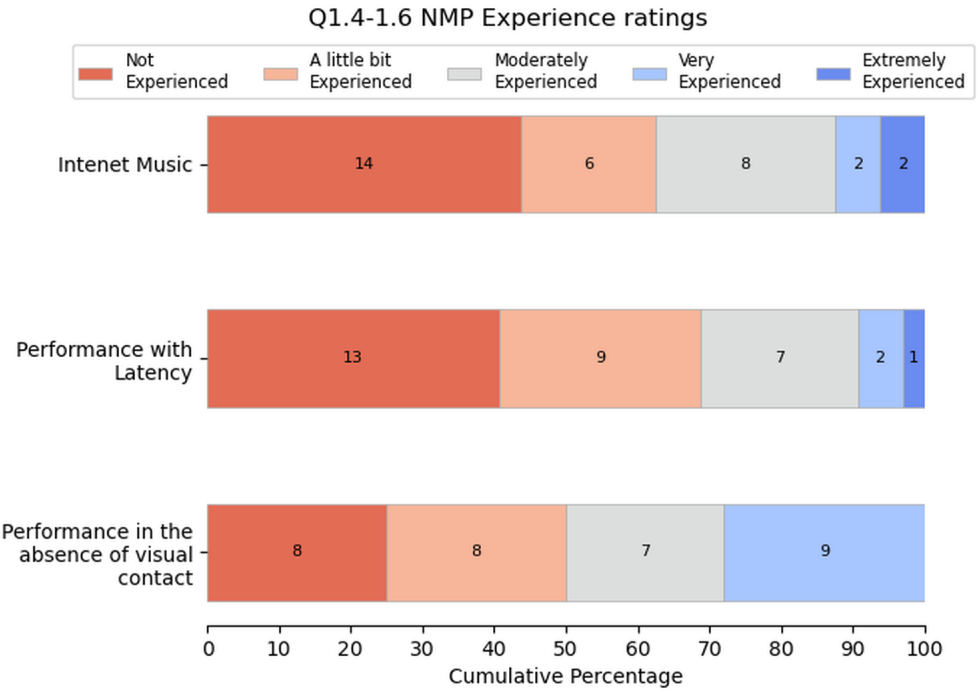


Figure 48: Responses for **Q1.4, Q1.5, Q1.6.** Proportionality-plot showing ratings distributions for participants' experience with NMP performances, performance over internet, in the presence of latency, and in the absence of visual contact

.

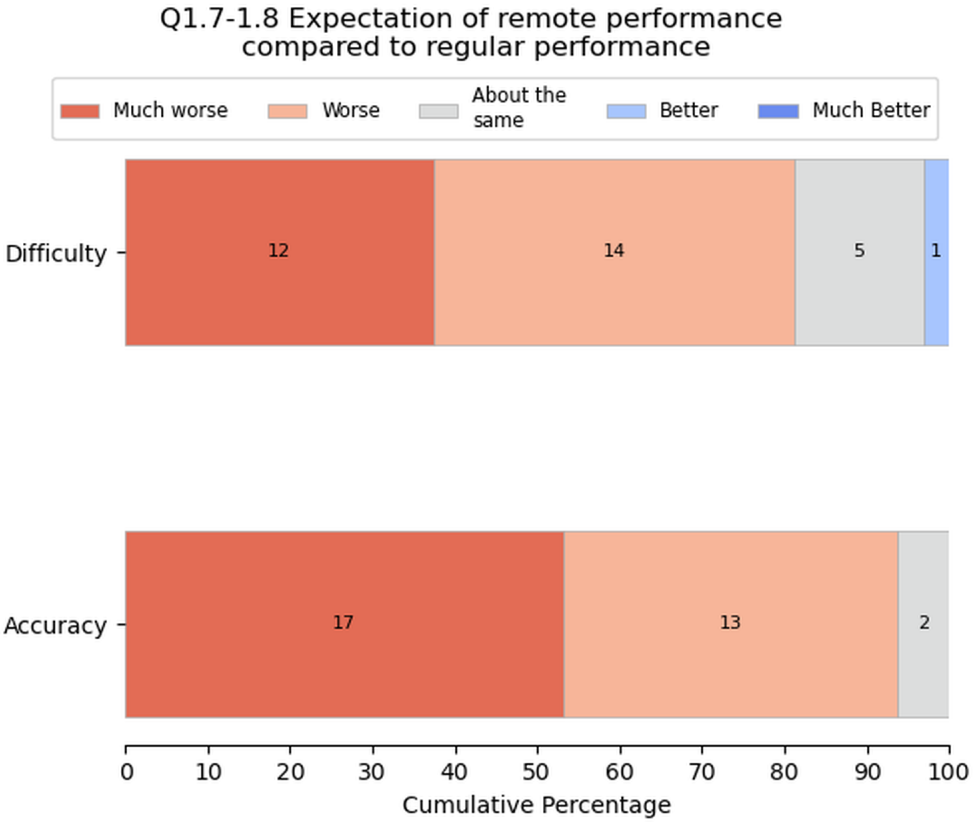


Figure 49: Responses for **Q1.7, Q1.8.** Proportionality-plot showing ratings for participants' expectation bias in regards to the difficulty and accuracy of remote performances as opposed to regular performances

Table 6: Demographic Questionnaire (Q1)

Q. ID	Type	Description
Descriptive		
0.1	Text	*"Please indicate your main study program or field of study"*
0.2	Numerical	*"How many years of musical experience do you have?"*
Familiarity		
1.1	Likert-5	*"Rate the level of musical familiarity with your co-performer"*
1.2	Likert-5	*"Rate your level of familiarity with "Clapping music" by Steve Reich."*
1.3	Likert-5	*"Rate the level of familiarity of performing "Clapping Music" with your experiment co-performer."*
NMP Experience		
1.4	Likert-5	*"Rate your level of experience with internet-based performance."*
1.5	Likert-5	*"Rate your level of experience with performing in the presence of signal latency."*
1.6	Likert-5	*"Rate your level of experience with performing in the absence of visual contact."*
Bias		
1.7	Likert-5	*"What is your expectation of difficulty when in a remote internet-based performance, compared to a regular performance? (less difficult - more difficult)"*
1.8	Likert-5	*"What is your expectation of accuracy when in a remote internet-based performance, compared to a regular performance? (less accurate - more accurate)"*

Table 6: Questions of the Demographic Questionnaire (**Q1**) completed once by participants before starting the experiment. The questionnaire was used for initial test candidate selection. Results were used later aggregated to create participant scores in each category.

2.2 Baseline Performance Phase

The first stage of the experiment was the *Baseline* data collection. This stage was carried out in *Dolan's Live Room* (see Table 2 and Fig. 35 for details) and is illustrated in Fig. 50. After live rehearsals, the musicians were placed at a distance of 8 ft (corresponding to the measured remote system latency) and miked using cardioid dynamic microphones. A total of four complete performance takes were recorded for each pair. The first two takes were recorded as in the setup described above. Two additional takes were recorded while the performers were asked to turn $180\circ$ at their spot to face opposite each other. The reason for this procedure was to attempt to "smooth out" the effect of taking away visual contact from baseline performance metrics. For

each take, a metronome count-in in quarter beats (85BPM) was played over a loudspeaker for four bars, cueing the start of the take. Due to varying levels of musical ability, some pairs were allowed to re-interpret the meter of the piece as a more comfortable simple-quadruple meter in $\frac{4}{4}$ rather than the original compound-quadruple meter in $\frac{12}{8}$. Later analysis was rendered agnostic to the type of meter chosen by each performer pair.

Having a co-located baseline stage served several purposes. First, it allowed participants, who in most cases met for the first time, to rehearse the musical piece as much as desired, in order to reach a mutually satisfactory level of performance. Secondly, it implicitly helped the pair participants form a personal construct of what "feels" like to play music within the same room (i.e., a traditional "regular" performance), just before being distributed across remote rooms. This helped to establish in the participants a recent inner reference construct against which to apply the rating of the auditory copresence questions of the trial questionnaire (Q2). Last but not least, this sub-phase allowed the recording of digital signal data from which the reference objective assessment metrics were extracted (see Sect. 4) and used to put in relative terms, for each pair, the change in performance metrics observed across the test conditions of the experiment.

2.3 Distributed Performance Phase

For the distributed sub-phase of the experiment, each participant was led to their assigned room and briefed on the operation of the headphone amplifier and the general flow of the experiment. For practical reasons, the room assignment was fixed to assign the Static-part performer to start in the Booth room (*Research Lab*, see Table 2), and the Shifting-part performer to start in the theater (*Frederick Loewe Theater*). The exact seating location of each participant corresponded to the location where the BRIR measurements were taken in that room (see Ch. V). The experimenter was located in the dedicated control room, able to communicate via a talkback channel to both participants either jointly or selectively.

2.3.1 Level Calibration and Familiarization

As a preparation step before the start of the distributed collection procedure, performers were tasked to participate in a level calibration sub-procedure. The purpose of this step was to calibrate the signal loudness to comparable subjective levels between participants, each of whom could

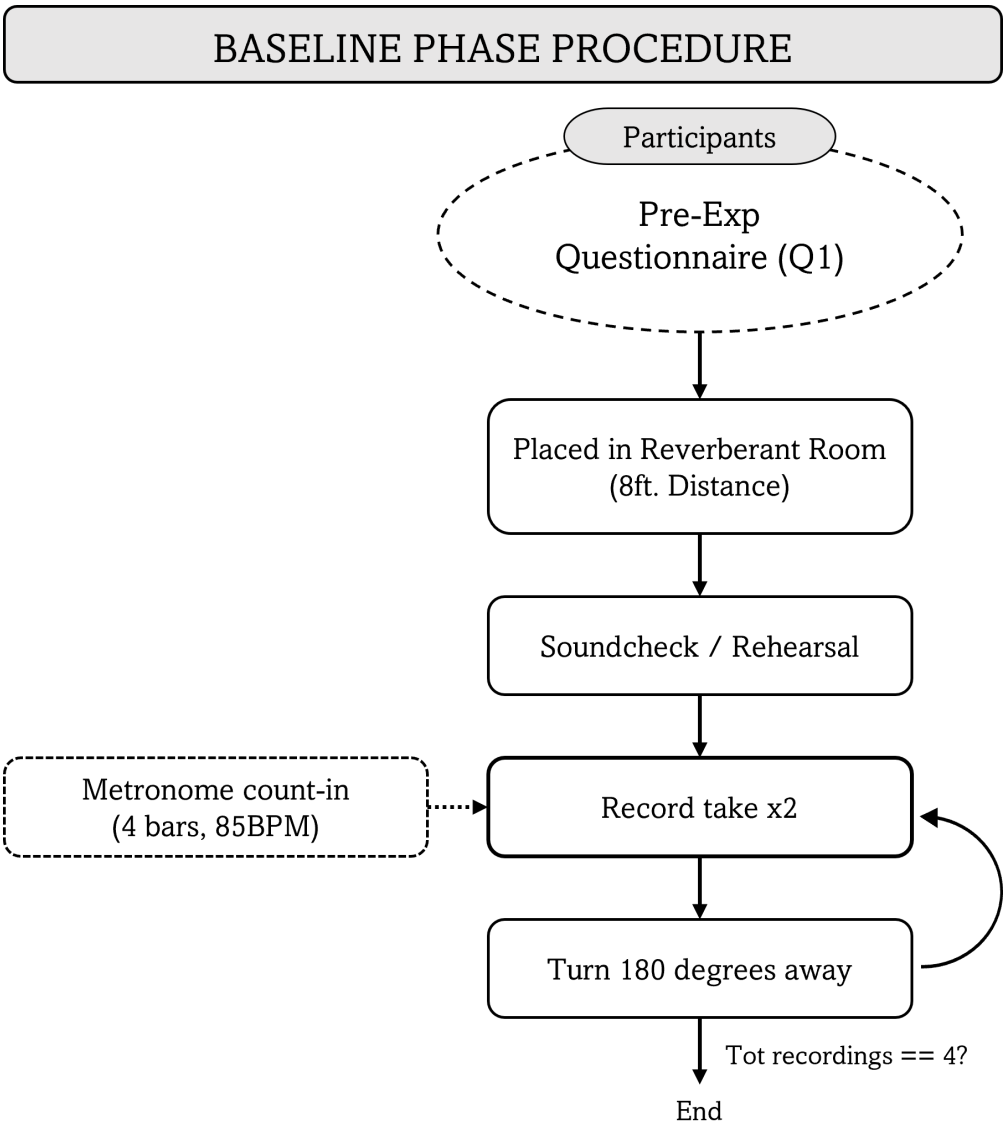


Figure 50: Flow diagram representing the procedure for the *baseline* sub-phase of the primary data collection. The two members of the participant ensemble are here located in the same room and are recorded playing the selected piece together. The first two takes were taken with the players facing each other, and the second two takes had them face against each other.

exhibit a unique clapping behavior producing a varying sound amplitude. First, each participant individually was asked to self-calibrate the output level of their headphone amplifier until the sound of their own average clap on the headphones and their own average clap in real life reached similar perceived loudness (musicians were instructed to wear headphones on one ear for easier assessment). Next, the raw co-performer "send" level was adjusted from the mixer in the control room, such that the unprocessed streamed signal level received on the headphone of the receiving party would also perceptually match the level of their own average clap. The loudness relationship of the "far" virtual position of the received co-performer level was later re-established by the amplitude levels embedded in the BRIR filters. For the "Raw (R)" acoustic mode, this was instead provided by a gain reduction of the send stream of -12dB to simulate approximate propagation loss.

The final step before commencing the trials consisted of a simple familiarization run covering the whole range of conditions. Participants were asked to clap quarter beats together at the minimum latency level. Every 20 seconds of clapping, the combination of each auralization mode with each latency level was progressed in real-time to the next condition in the sequence, until all 15 conditions were covered once. Participants had the opportunity to repeat the familiarization run if desired.

2.3.2 Trials

Trials were carried out as depicted in Fig. 51. The sequence of test conditions was activated through the DAW at the central node following a randomized order from a generated list, different for each pair, with one repetition per combination of latency level (namely $[7, 20, 40]$) and acoustic mode ($[R, AC, AD, SC, SD]$). At each iteration, a four-bar metronome count in quarter beats was sent to the Static player headphone mix (exclusively, to avoid early synch offsets in high-latency conditions) who was tasked to count together with the metronome beats for bars 3 and 4 of the count. The microphone signals were recorded raw at the experimenter station (devoid of room-acoustic processing and additional artificial latency). At the end of each trial, participants were tasked to complete a questionnaire (*"Trial Questionnaire (Q2)"*, details in Sect.3) which collected their subjective rating of the performance they just completed.

Once the set of 15 combinations of latency levels and auralization modes was completed, the participants took a resting break before switching rooms. To account for differences in players clapping strength, the headphone and microphone levels were recalibrated once more after switching rooms as described in Sect. 2.3.1. Thus, the entire set of trials was repeated as before, in a new randomized sequence order. Once the second set of trials was completed, the participants were gathered and asked to complete a post-experiment debrief questionnaire ("**Q3**: feedback questionnaire") before dismissal.

2.3.3 Data Summary

Excluding rehearsal takes, the collected primary data resulted in a total of 1088 individual recording takes, subdivided into 960 takes concerning the distributed sub-phase (30 repetitions per participant) and 128 individual takes concerning the baseline sub-phase (4 repetitions per participant). Given an average take length of one minute and fifteen seconds, the total material amounted approximately to 20 hours of recorded audio data (22h and 40m with the baseline).

2.4 Debrief Questionnaire

This section covers the responses collected from the *"Debrief Questionnaire (Q3)"*, taken by each participant right after their primary data collection session. For this questionnaire, the principal goal was to collect general high-level insights on secondary research questions and obtain qualitative feedback on experience impressions. As for Q1, responses were also used to obtain additional descriptive attributes of each participant for potential use in the analysis stage (for example, *fatigue*). Table 7 shows the questions present in this questionnaire. By this stage, the participants were already briefed on the meaning of terms such as *copresence* from the instructions of the trial questionnaire (Q2) discussed in Sect. 3. Fig: 52 shows the distribution of responses for Q3.1, placed in this section due to its descriptive purpose. Results for the other questions are instead located in Ch. VII since those responses are related to the high-level research questions rather than the classification of participants.

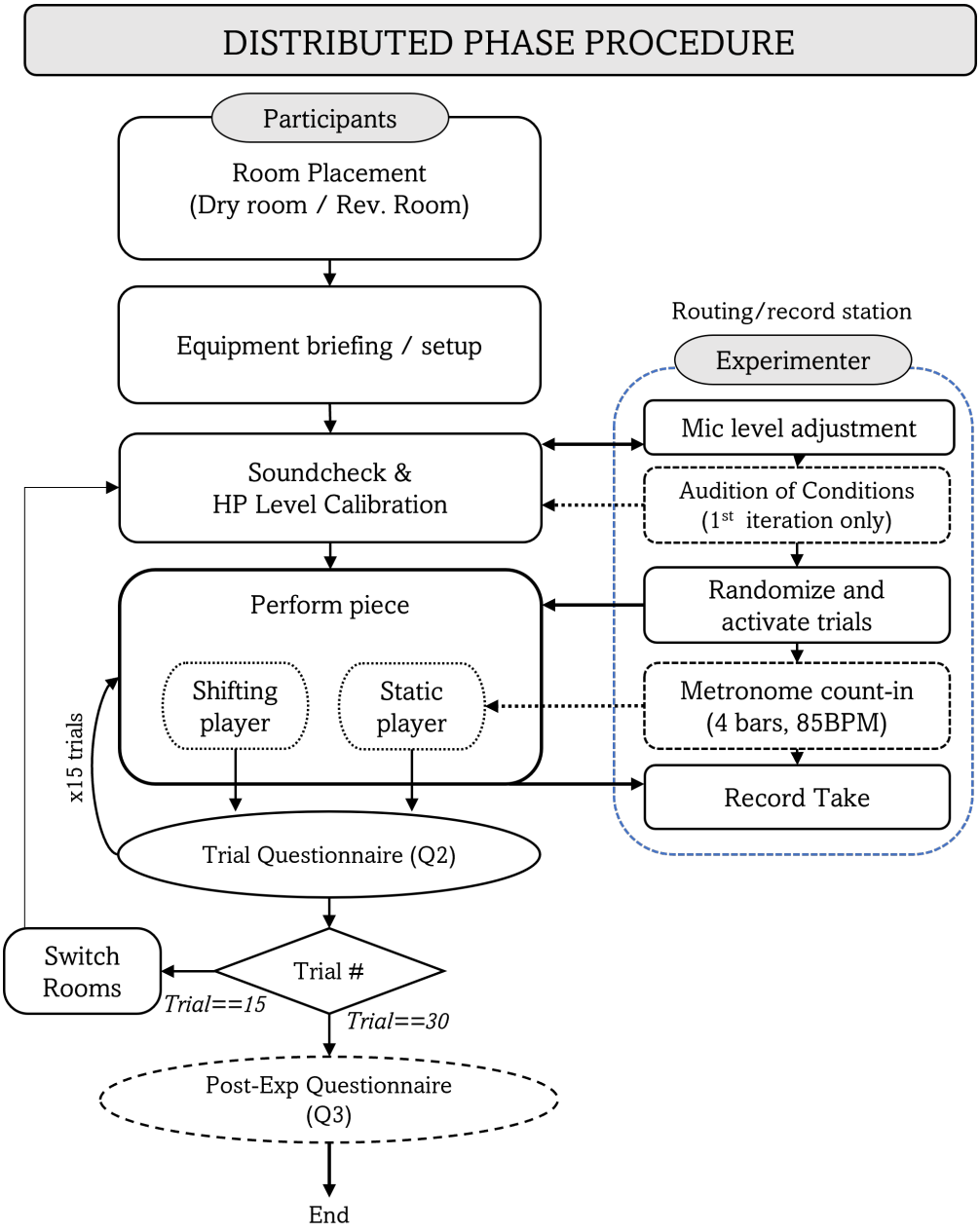


Figure 51: Flow diagram representing the procedure for the *distributed* sub-phase of the primary data collection. The two participants are conducted to the assigned rooms and directed through the various stages by the experimenter. After familiarization with the setup, the experiment proceeds with 15 randomized trials spanning 3 latency levels and 5 auralization conditions. Participants are then asked to switch rooms and repeat the experiment. Signals are recorded raw at the central node.

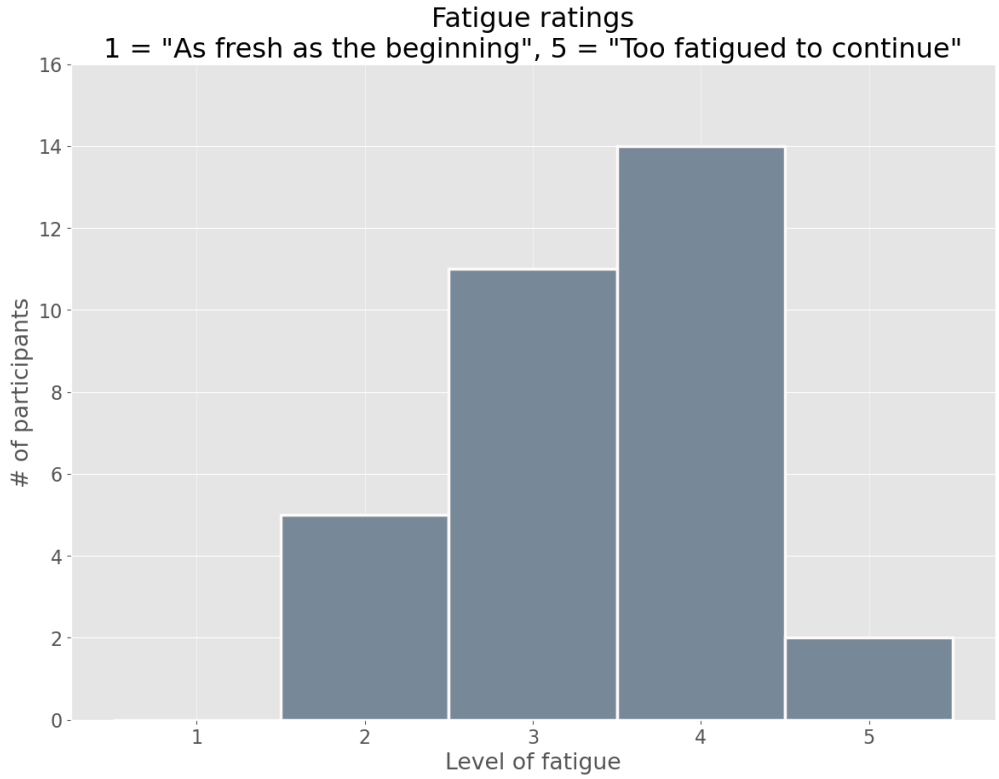


Figure 52: Responses for **Q3.1.** Level of fatigue experienced by participants at the end of the experiment, a possible factor in affecting performance quality over time.

Table 7: Debrief Questionnaire (Q3)

Q. ID	Type	Description
Fatigue		
Q3.1	Likert-5	*"Please rate your level of fatigue reached at the end of the experiment (1 - As fresh as the beginning, 5 - Too fatigued to continue)"*
Agreement ratings (1: Strongly Disagree, 7: Strongly Agree)		
Acoustic Mode Impressions		
Q3.2	Likert-7	*"It was generally easier to perform with stronger reverberant conditions rather than dryer conditions"*
Q3.3	Likert-7	*"Certain acoustic environment modes felt "more real" than others"*
Q3.4	Likert-7	*"Certain acoustic environments modes helped the performance more than others"*
Latency Impressions		
Q3.5	Likert-7	*"When latency was present, reverberant conditions made it easier to perform"*
Q3.6	Likert-7	*"In the presence of heavy latency, I felt disconnected from my coperformer"*
Copresence Impressions		
Q3.7	Likert-7	*"In the trials where I rated higher 'copresence', it was easier to perform"*
Q3.8	Likert-7	*"In the trials where I rated higher 'copresence' my performance was more accurate"*
Q3.9	Likert-7	*"In the trials where I rated higher 'copresence', I enjoyed myself better"*
Feedback and comments		
Q3.10	Text	*"Please provide some comments about how latency and acoustic reverberation in general affected your performance"*
Q3.11	Text	*"Has this experience, in any way, changed your expectations about distributed music, augmented acoustic environments or internet-based performance? If so, how?"*

Table 7: Questions of the Debrief Questionnaire (**Q3**) completed once by participants at the end of their experiment session. Results serve to provide additional high-level insights into the experiment effects. In the case of *Fatigue*, the responses were used to further categorize the participants' analysis groups by feeding in the analysis models as potential random effect

3 Subjective Trial-experience Questionnaire

This section covers the details related to the *"Trial Questionnaire (Q2)"*, one of the three layers of assessment data (or "secondary" data). This questionnaire was completed directly by the participants of the data collection phase. The responses are directly related to the primary data, as they were collected jointly during the distributed phase of the experiment, in between session

trials. Each participant individually completed the questionnaire 30 times during their session. A total of 960 entries of this questionnaire were obtained among all participants.

The purpose of this questionnaire was to collect subjective data related to the performance experience as it was completed under a given set of latency and acoustic mode conditions. The first set of questions regarded the perceived technical success of the musical performance task. This involved Likert-type ratings of *perceived accuracy* and *performance difficulty* in regards to the immediately prior experience. The second set of questions directly polled the psychological constructs of auditory *"copresence"* and *"Cohesion"* (see definitions on next paragraph) on a unidirectional general opinion score scale going from 'low' to 'high'. The 'high' level was explained as meaning "just as good as a regular co-located performance" as that stands as the target reference objective of immersive systems that aim for *realism*. This inner reference was provided to the participants at the baseline sub-phase when they first played in the same location. The third set of questions was portrayed as a series of statements on which to rate agreement (Likert, 7 points). These questions were principally designed to zoom in on the "direction" of copresence experienced by participants (e.g. "being there" vs. "being here"), while also providing alternative definitions to the constructs of interest and further validating the results of the previous set of questions. Besides being analyzed individually, the responses to these questions were also aggregated to get a "presence score" for each questionnaire entry. An additional set of questions concerning "involvement" and "engagement" was initially present but eventually scrapped as participants claimed to be confused by the definitions when asked for feedback. Table 8 provides details on each question on Q2.

The following definitions were provided to participants:

- **Acoustic Environment:** *"The natural "room tone" that you hear and feel around you as you create sound. An acoustic environment is composed by a unique pattern of sound reflections, reverberation, and resonances which color the sound around you and provide a sense of space (e.g. think about the difference between the sound of a choir in a church vs the same choir singing outside in a field)."*

- **Auditory Copresence:** *"The auditory sensation of being "together" with your co-performer in a similar way as you experience in a traditional col-located rehearsal or performance. Copresence*

can be felt in both directions: you can feel your co-performer "being here" with you, or you can feel "being there" with them. In other words, "copresence" is characterized by a sense of being in the same space as another human, virtual, or otherwise, as well as the perception of mutual awareness and attention from others" (Zhao 2003)

- **Auditory Cohesion:** *"The sensation that the sound you create and the sound you hear plausibly fit together in your current physical space / The feeling that your acoustic timbre and that of your co-performer belong together and are naturally fitting your local acoustic environment or your expectation of it"*

4 Objective Signal Metrics

The second layer of the evaluation consisted of objective performance metrics (described below in Sect. 4.4) extracted from the raw primary data. These metrics were introduced to provide insights and observations on both the influence of the main effects on the success of the musical task and to establish correlations between different realms of assessment (subjective experience quality vs task success). Ultimately, the objective layer was needed to help the experiment discussion move toward a better understanding of the effect of "immersive qualities" in NMPs.

The objective analysis relied on having reasonably "clean" signals to exercise algorithms upon, meaning that a pre-processing step was introduced to transform the signals into versions that could be better parsed by the tempo detection algorithms. In practice this meant the removal of "bleed" from the baseline recordings and of the room reflections from the distributed recordings collected from the Theater room to avoid having the algorithm consider those as beat onsets. A further compressor step was then used to reduce the dynamic range between low-amplitude claps and high-amplitude claps for more consistent signals. The robustness of this process was varying, as in occasional circumstances ambiguities between low-amplitude claps and signal bleed meant that actual pattern claps got removed from the signal. To answer this limitation, subsequent tempo/beat calculations and metrics were maintained at a high level, with low temporal resolution. This approach emphasized the analysis of more reliably derived overall performance patterns, rather than concentrating on local short-term fluctuations influenced by unstable estimations.

Table 8: Trial Questionnaire (Q2)

Q. ID	Type	Description
Performance Evaluation		
Q2.1	Likert-7	*"Rate your impression of performance accuracy for this trial. (1: Very poor, 7: Very accurate)"*
Q2.2	Likert-7	*"Rate the difficulty you experienced performing during this trial (1: Not at all difficult, 7: Very difficult.)"*
Copresence & Cohesion direct ratings		
Q2.3	Numerical (7-points)	*"Rate your general level of feeling auditory "copresence". (1: Not at all, 7: Perfect copresence)*
Q2.4	Numerical (7-points)	*"Rate your general level of feeling auditory "Cohesion". (1: Not cohesive at all, 7: Perfectly cohesive)*
Agreement ratings (1: Strongly Disagree, 7: Strongly Agree)		
Q2.5	Likert-7	*"It felt as if my co-performer and I were performing in the same room"*
Q2.6	Likert-7	*"It felt as if my co-performer was here with me, in my location"*
Q2.7	Likert-7	*"It felt as if I were transported to my co-performer's location"*
Q2.8	Likert-7	*"The acoustic timbre of my co-performer matched the acoustic environment of where I am now"*
Q2.9	Likert-7	*"I was able to clearly picture or imagine my co-performer being nearby me"*

Table 8: Questions of the Trial Questionnaire (**Q2**) completed in between each trial during the distributed phase of the primary data collection. The data gathered by this questionnaire formed one of the principal layers of analysis

The processed signals were fed into two pipelines. One extracting a "dynamic tempo curve" based on the onset envelope computed via spectral flux (Böck and Widmer 2013) to capture individual tempo variations, and one tracking the estimated beat intervals using dynamic programming (Ellis 2007). Both curves were obtained thanks to the *Librosa* Python3 library (McFee et al. 2015). These two curves served as the basis over which individual beat-related metrics and pair beat-synch metrics were extracted. This part of the process was first executed on the baseline primary data, with metrics being averaged across all the pair-related takes taken during collection. These metrics represented the reference level of each pair from which their distributed performance metrics were compared to. The relative metric allowed to understand the impact of the main effects on the distributed data in terms of performance improvement or degradation while being agnostic to the starting ability level of a pair of performers. The complete processing pipeline details are summarized in Figs. 53 for the baseline data and 54 for the distributed data.

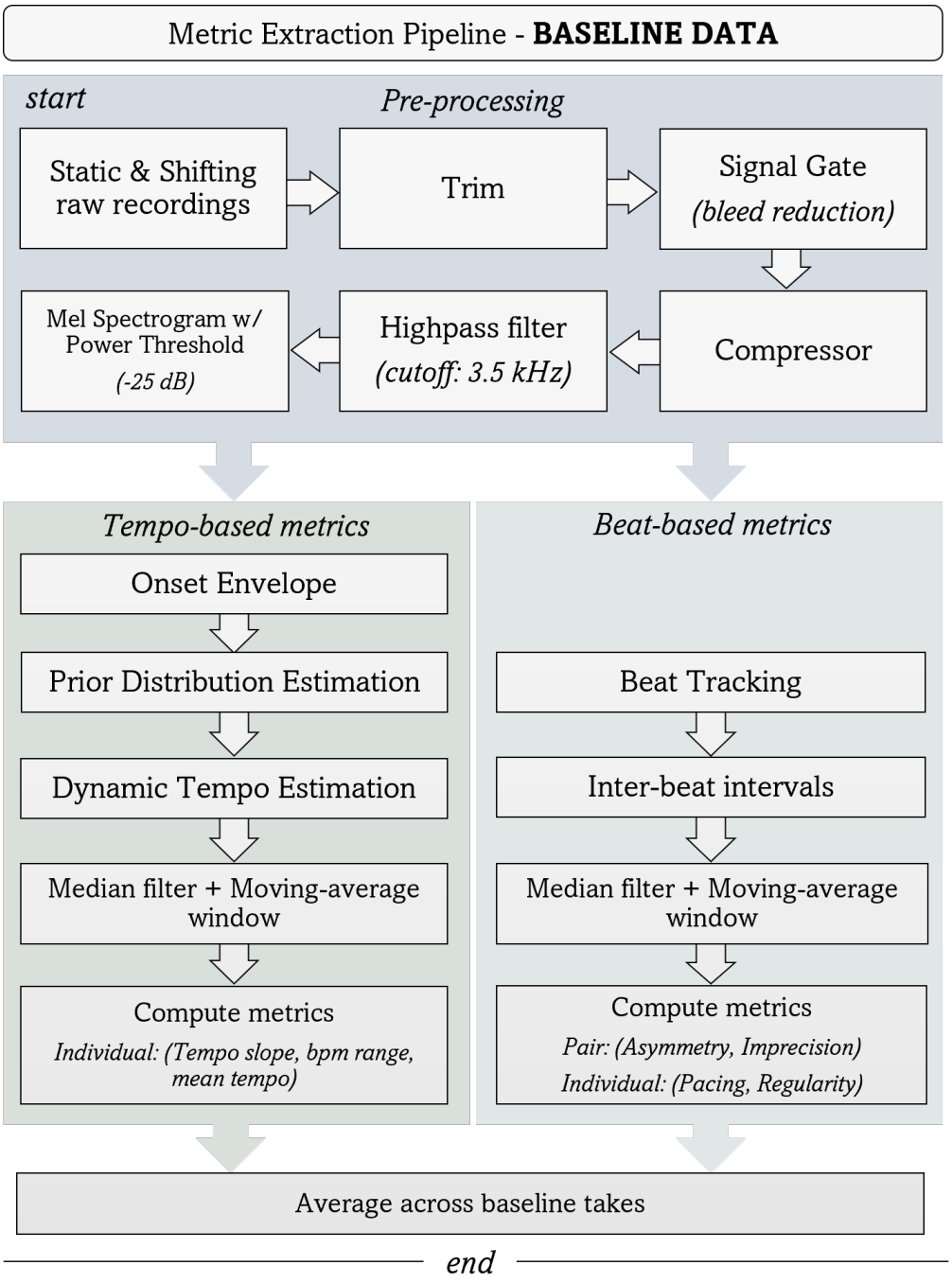


Figure 53: Signal processing flow for the extraction of objective performance metrics from the baseline primary data. Signals are first pre-processed to reduce signal bleed as much as possible and reduce dynamic variation across clap onset strengths. Individual performance metrics are extracted from the dynamic tempo curve and from the inter-beats intervals. Pair-related synchronization metrics are also extracted from the inter-beat intervals. Results are finally averaged across the baseline takes.

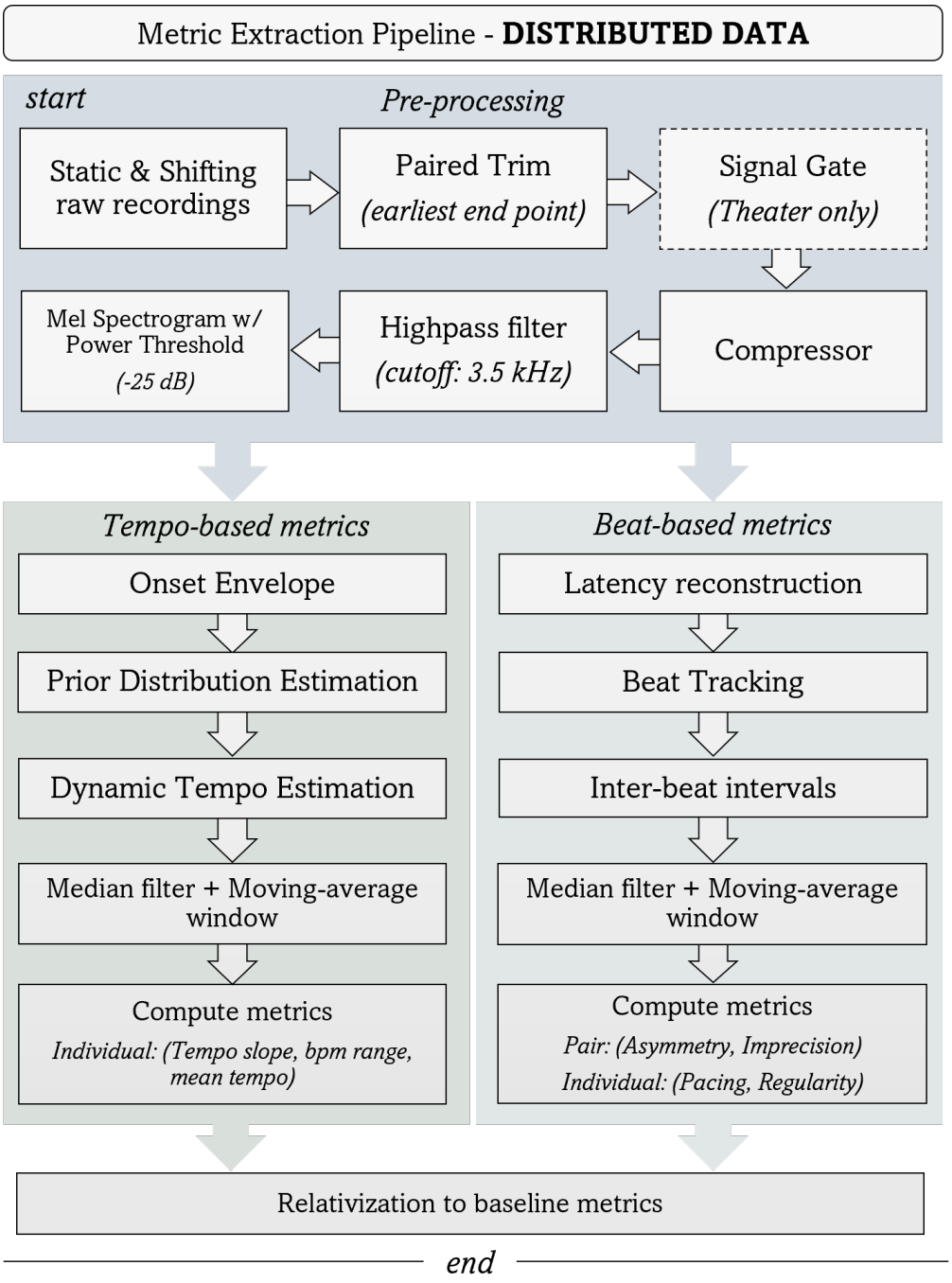


Figure 54: Signal processing flow for the extraction of objective performance metrics from the distributed primary data. Signals are first pre-processed to remove reflections and compressed to reduce dynamic range. The rest of the processing runs similarly to the baseline data, with the difference of a latency recreation step in order to capture the beat synchronization as experienced at each node by the performer. The final values are transformed in relative terms for each pair, using the pair's baseline metrics.

4.1 Pre-processing

The pre-processing step started with the trimming of the raw take recordings to eliminate any start or trailing silences. For each performance take, the static-player and shifting-player clips were trimmed to equal length using the earliest performance end-point. The average take length after trimming was between 60 to 80 seconds depending on the effective tempo. The trimming also removed the initial pre-clapping metronome cues that were counted in by the static player. The signal gating stage was tuned to a threshold of -38dB and applied to the baseline data (bleed reduction) and to the distributed data collected at the Theater (reflection reduction). A compressor stage (compression threshold of -20dB and compression ratio of 10) was subsequently applied to all signals to improve the consistency of the dynamic range. The gating and compressor stages were applied through the *Pedalboard* library for Python3 (Spotify 2021). A 3^{rd} order high-pass filter with a cutoff frequency of 3.5 kHz was then applied to sharpen the clap transients. Finally, the signal was converted to a Mel-scale spectrogram and thresholded at -25dB which was the signal fed to later stages. The *Librosa* Python3 library was used for this final step (McFee et al. 2015). To choose the filter and threshold parameters, a hyper-parameter optimization search was operated by looking at the stability output of the dynamic tempo curve extraction (Sect. 4.2). Fig. 55 shows a visualization of the result of the pre-processing steps.

4.2 Dynamic Tempo Curve

The first curve-extraction step involved the extraction of a dynamic-tempo curve to represent the change of performance BPM tempo over each recorded take. To begin this process, an onset envelope function using the spectral flux function was computed over the spectrogram output of the preprocessing stage (McFee et al. 2015; Böck and Widmer 2013). The parameters were set to use a "hop size" of 512 samples and a median band-wise aggregation function. The envelope was first passed to an autocorrelation function to provide a preliminary set of estimated overall-BPM levels across the entire envelope (Fig. 56. The nature of tempo extraction from an onset envelope is such that several harmonic levels of the target BPM can be detected (for example, a quarter-note beat interval can be confused by the presence of stronger eight-notes), leading to potential "BPM octave" latching errors. To address these errors, easily reproduced in such a rhythmically complex music piece, the autocorrelation output was used to find the location of the "most likely" BPM peak

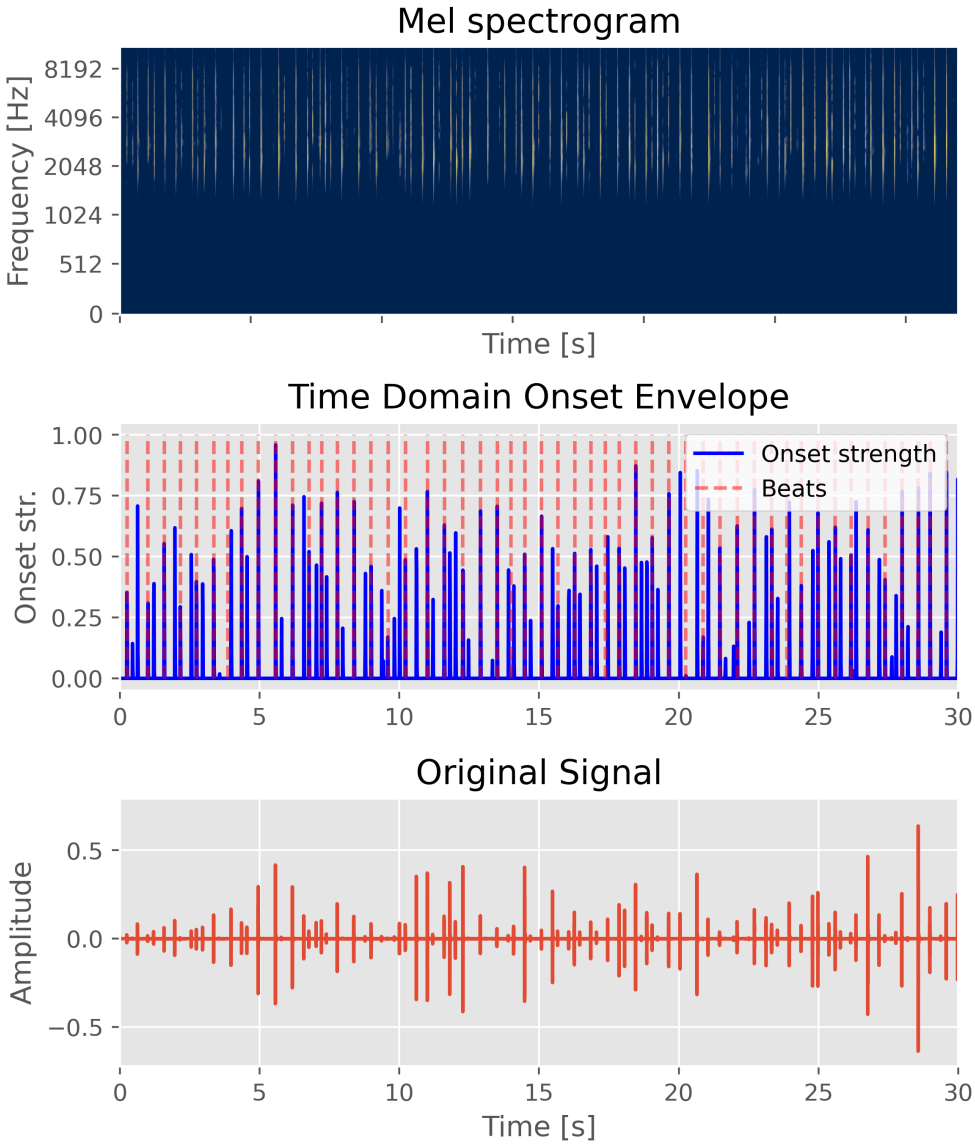


Figure 55: Post-processed signal shown in time-domain (bottom), and as a Mel-frequency spectrogram (top). This signal was fed to both the spectral flux onset envelope algorithm and the to the dynamic beat-tracking algorithm.

representing the average BPM, or the center of a distribution of BPM probability throughout the take ("prior"). As shown in Fig. 56 the most prominent peaks were first identified, out of which the strongest peak close to the 85BPM mark (the intended performance tempo) was considered to be the one associated to the "most likely" tempo. Therefore, the BPM value of the peak was used as the center of a "prior" uniform distribution function (ranging from $\pm 10BPM$ from the center value). This step allowed the later dynamic tempo extraction stages to latch to the right BPM harmonic for a continuous frame estimation.

Using the computed take-prior, a dynamic tempo extraction curve converted the onset envelope data into a continuous tempo estimation (using an autocorrelation window of four seconds, and "hop length" of 512 samples). To get more robust general trends, the obtained curve was then cleaned of spurious artifact using a four-second median filter and again smoothed with a two-second moving average window. The first and last 4 seconds of performance tempo were eventually removed from the curve due to being more susceptible to *window edge* artifacts. Fig. 57 shows an example of a dynamic tempo curve over a raw tempogram plot showing the BPM octaves. Getting the curve to latch to the right harmonic was highly dependent on having a correct prior estimation. To account for the possibility of wrong prior estimates, the final summarizing metrics focused on relative BPM line trends (such as "slope") rather than absolute BPM values.

4.3 Beat Extraction

A parallel pipeline process used the spectrogram output of the preprocessing stage to compute beat tracking curves (McFee et al. 2015; Ellis 2007). A quarter-beat tracking algorithm was operated with a "hop size" of 512 samples and a start BPM guess of 85BPM. The first and last 4 seconds of the performance were dropped from this estimation. The resulting beat locations were used to obtain the interbeat intervals (IBI), as well as the local BPM estimates shown in equation 8 (Rottondi et al. 2016). The BPM curve was then smoothed through a median filter and average window of a size representing eight quarter-beats (2 bars). The IBI/BPM curve directly fed the calculation of certain individual metrics (see Sect. 4.4).

$$IBI[n] = \frac{\texttt{beat}[n] - \texttt{beat}[n-1]}{\texttt{sample rate}} \tag{8}$$

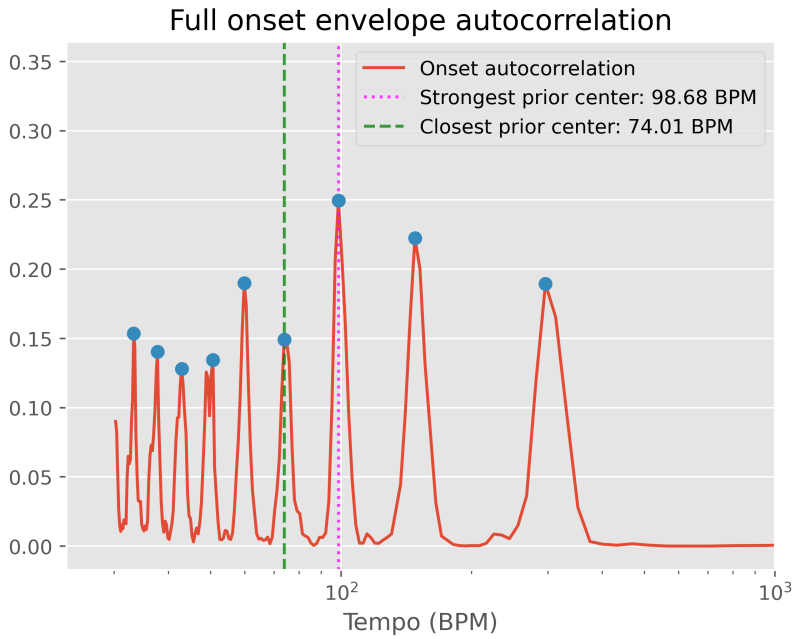


Figure 56: Static auto-correlation analysis was employed to derive estimates of BPM probability distribution centers ("priors"). The prior distribution was defined by selecting the most prominent near peak to the reference value of 85BPM, and use it as the center of a uniform distribution

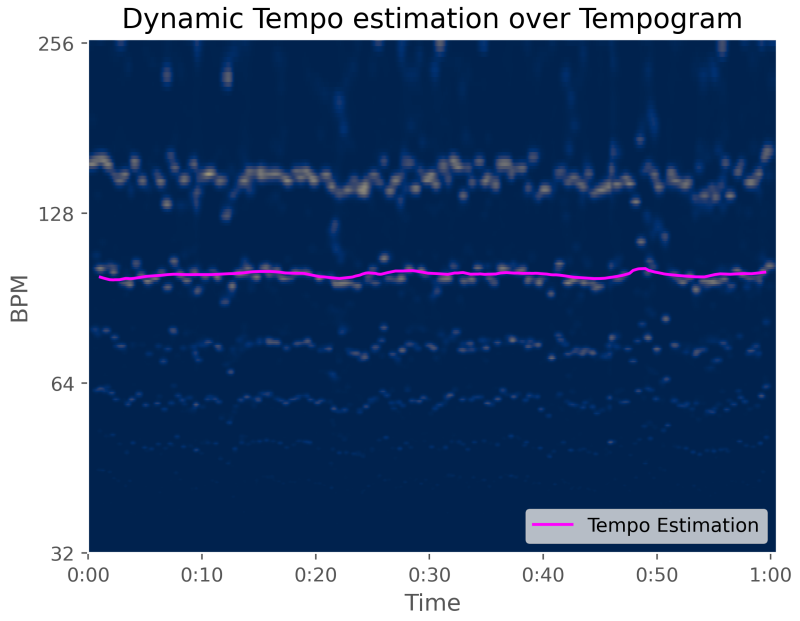


Figure 57: Example of a smoothed tempo-curve plotted over the raw tempogram plot, taken from the baseline data

$$BPM[n] = \frac{60 \times \texttt{sample rate}}{\texttt{beat}[n] - \texttt{beat}[n-1]} \tag{9}$$

To calculate inter-subject pair-synchronization metrics, reference "key" beats acting as synchronization checkpoints were identified as the beat locations occurring every eight beats (which correspond to two bars of score music). The determination of the locations of the checkpoints was performed on the static-player clip since it could more reliably provide beats corresponding to score measures. With the first bar being dropped for stabler metrics, the resulting amount of checkpoints amounted to twelve per take (Fig.: 58). The synchronization metrics were finally derived by determining the absolute time difference between the checkpoint timestamps from the static player, and the nearest available beat from the shifting-player curve. In the case of the distributed primary data, this step was preceded by a latency-reconstruction adjustment, designed to recreate the combination of signals as heard from the perspective of each performer during that trial. The adjustment was made by delaying the remote node signal (by either 7ms, 20ms, or 40ms according to the trial condition) with respect to the receiving node, in the direction according to the node to which the clip belonged.

4.4 Performance Evaluation Metrics

The final metrics extracted by the objective evaluation process were adapted from NMP evaluation metrics reported in previous literature (Rottondi et al. 2016; Chafe et al. 2010). Individual-level performance metrics were collected from both the tempo curves and the beat-interval curves. Pair-level metrics were instead collected exclusively from the re-sampled beat-interval curve (with recreated latency perspective from each node side). Baseline metrics were averaged across all takes per pair, with strong outliers removed, but with a minimum of two baseline takes considered per pair.

All the metrics relating to the distributed primary data were later transformed at the analysis stage to be in a relative form in relation to the average baseline metrics for each subject. Details of the relativization process are provided in Ch. VII, Sect. 2.2.2.

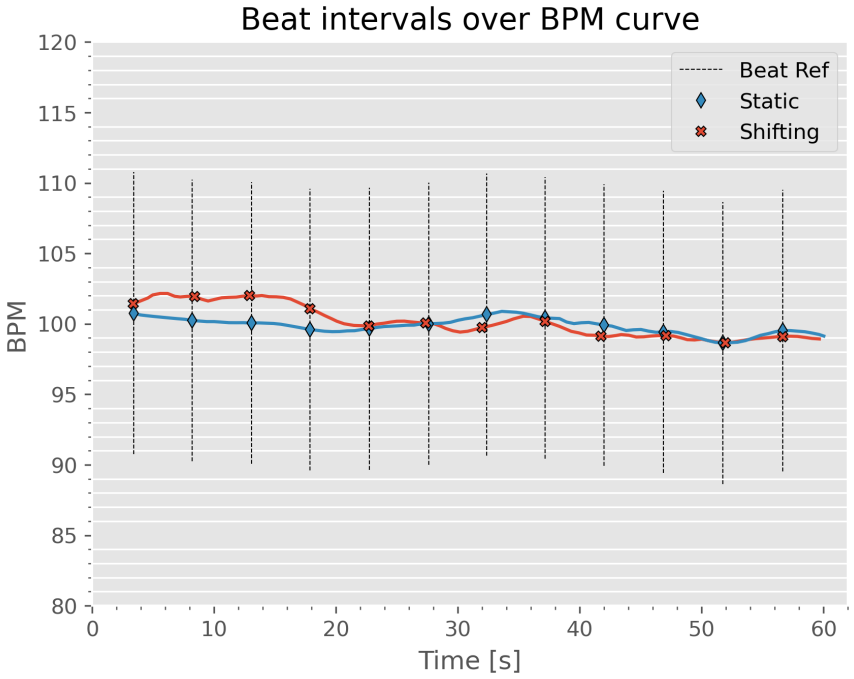


Figure 58: Beat synchronization analysis (baseline data example) sampled every 2 bars of performance using the Static beat as reference. The first two bars were dropped from the synch analysis

4.4.1 Individual Metrics

- ◇ ***Overall BPM*:** Overall static tempo estimate from the dynamic tempo curve[1]

- ◇ ***Tempo range*:** Difference between the minimum and maximum BPM values, taken from the dynamic tempo curve[1].

- ◇ ***Tempo slope*:** General slope trend of the linear interpolation fit computed on the dynamic tempo curve[1].

- ◇ ***Pacing (π)*:** Mean inter-beat interval computed over the whole clip (Bartlette et al. 2006b)

$$\pi = \frac{1}{N}\sum_{n=1}^{N} IBI_n \tag{10}$$

  Where N is the total number of IBI intervals computed on the curve.

- ◇ ***Regularity (ρ)*:** Coefficient of inter-beat interval variability (Bartlette et al. 2006b)

$$\rho = \left(\frac{\sqrt{\frac{\sum_{n=1}^{N}(IBI_n - \overline{IBI})^2}{N-1}}}{\frac{1}{N}\sum_{n=1}^{N} IBI_n} \right) \tag{11}$$

4.4.2 Pair Metrics

- ◇ ***Mean Lag* (α):** Mean of the inter-subject absolute time differences, calculated on checkpoints obtained from the downsampled beat-track curve. Adapted from (Chafe et al. 2010).

$$\alpha = \frac{1}{N}\sum_{n=1}^{N} |(t_A^n - t_B^n)| \tag{12}$$

  Where N is the number of synchronization "beat checkpoints", t_A^n is the time of the n^{th} checkpoint in the static player beat-track curve, and t_B^n is the closest beat time from the shifting player beat-track curve.

[1] In case the tempo range and slope metric computed from the dynamic-tempo curve were flagged as strong outliers, the metric was instead sourced from the BPM curve extracted from the inter-beat-intervals

◇ ***Imprecision*** **(μ):** Standard deviation of the inter-subject absolute time differences, calculated on checkpoints obtained from the downsampled beat-track curve. Adapted from (Farner et al. 2009).

$$\mu = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(t_A^n - t_B^n)^2} \tag{13}$$

5 Third-party Annotations and Metrics

The final layer of secondary evaluation data was designed as a "crowd-sourced" collection of annotations concerning the presence of pair-level performance inaccuracies in the primary data and related "opinion-score" ratings. The main underlying motivation behind this stage was to insert a subjective dimension of performance quality assessment into the list of explored evaluation scales, picking up on the "perceptually noticeable" inaccuracies and musical pattern mistakes. Furthermore, this data helped to address some of the technical uncertainties and lack of robustness observed in the objective algorithms tasked with detecting beat-pattern mistakes (eventually removed from the objective analysis pipeline). More specifically, certain encountered edge-case situations led the onset-detection algorithms to be easily skewed by occasional extra claps or to miss the tracking of relatively low-transient beats caused by variations in clap positioning, making a high-resolution beat analysis ambiguous. A non-optimal clap impact would in fact produce a transient level comparable to undesired reflection onsets, resulting indistinguishable from the onset detection algorithm, thus leading to a misclassification of whether pattern mistakes had occurred. Rather than manually checking the tuning of the onset-tracking thresholds, the delegation of the task to expert human listeners capable of understanding the musical context of encountered mistakes was deemed a less uncertain evaluation and annotation process.

Annotators were sourced among NYU Music and Performing Arts Professions students, including the same pool of subjects who participated in the experiment and were thus highly familiar with the piece and the nature of the experiment. Out of 16 total experiment sessions, 15 sessions were eventually evaluated by a total of 9 participating annotators who were paid for the effort. Ratings were effectuated once per recording (each recording comprising both static and shifting channels) meaning that each annotation set consisted of 30 stereo files. Some annotators

rated more than one session. If the annotator/rater was a previous participant in the primary data collection, they were excluded from rating their own session to avoid potential biases. Each annotator received a pre-formatted evaluation spreadsheet on which to record their ratings and annotate the presence of errors, an instruction document describing the meaning of each rating, a copy of the *Clapping Music* score, a reference "good" take extracted from the baseline recording of the most expert pair of subjects, and finally, the actual audio files to be rated. The files provided for each annotation set consisted of 30 anonymous processed stereo files pertaining to the takes of the assigned session. The audio files given to annotators were labeled with a number ranging from 1-30 in randomized order, thus not in the same order as the one in which the trials were executed (annotations were later cross-referenced with their corresponding entries in the data table). Annotators were instructed to audition the material in labeled order using stereo headphones for reproduction devices.

Table. 9 provides details on exact categories of performance inaccuracies to be annotated and rating scales. The description field illustrates the descriptions given to annotators. In the case of the beat pattern mistake categories, annotators were asked to focus on "major" mistakes, on the important beats, rather than minor infra-beat inaccuracies or occasional tempo deviations (expected to be omnipresent throughout the set of recordings). Annotators were also instructed to refrain from viewing varying meter interpretations (compound vs simple) as an "error" and instead, to base their assessments on this aspect. Some of the annotation categories were later aggregated for simpler analysis (see Ch. VII) and the rating scale results were *Z-scored* to account for annotator judgment biases.

5.1 Audio Material Processing

The annotators were blind to the conditions under which the files were recorded, as the audio provided did not include any of the auralization processing experienced by the performers. However, a one-way latency offset was injected in the stereo mix of the files in order to recreate the exact events of the performance as experienced by the player at the Static node. The delay offsets re-created in these signals corresponded to the latency level associated to each recording: 7 ms, 20 ms, or 40 ms (the system latency was already corrected in the recordings by the DAW

Table 9: List of annotation categories and rating scales

Name	Type	Description
Performance issues		
Silences	Binary checkbox	*"Check this box if one of the players suddenly stops clapping (more than 2/3 seconds) during the performance but then resumes to play."*
Unfinished	Binary checkbox	*"Check this box if one of the players (or both) stops clapping prematurely and do not resume to play."*
Tempo inaccuracies		
Unsynchronized beat	Binary checkbox	*"Check this box if you perceive players to be noticeably out of synch at any point in the take (for at least a sustained period of a few seconds)"*
Noticeable acceleration	Binary checkbox	*"Check this box if you can perceive the performers' tempo noticeably accelerating at any point in the take (for at least a sustained period of a few seconds)."*
Noticeable deceleration	Binary checkbox	*"Check this box if you can perceive the performers' tempo noticeably decelerating at any point in the take (for at least a sustained period of a few seconds)."*
Pattern mistakes		
Extra claps	Binary checkbox	*"Check this box if you heard that a player noticeably clapped more notes than indicated in the score pattern."*
Missed claps	Binary checkbox	*"Check this box if you notice that a player missed important claps in their pattern"*
Tot mistakes	Int number	*"Please indicate the total tally of "major" pattern mistakes you could detect in this take"*
Ratings		
Tempo rating	Scale (1-10)	*"Rate the stability of the players in keeping a consistent tempo across the performance (although not necessarily in synch or precise)."*
Synch rating	Scale (1-10)	*"Rate the synch of the players throughout the piece (although not necessarily precise or stable in terms of tempo)."*
Precision rating	Scale (1-10)	*"Rate the precision of the players in keeping to the score patterns (mostly focus on the shifting player)"*
Overall rating	Scale (1-10)	*"General technical and musical quality of the performance. From 1 (lowest) to 10 (highest). Consider overall synchronization, score precision, presence of mistakes, tempo stability and expressivity aspects."*
Other		
Additional comments	Text	*"Please indicate any other salient artifact or type of mistakes not covered by the previous fields. Please provide timestamps of your findings."*

Table 9: Complete list of annotation and rating instructions given to third-party expert annotators

so it did not have to be removed by the offset calculation). Fig. 59 illustrates the complete pre-processing pipeline for the audio material handed to annotators. After latency injection, the individual "Static part" and "Shifting part" channels were normalized (with signal gating applied on the signals coming from the "Theater" room in order to remove reflections) and passed through a compressor to attain comparable loudness levels to reduce the dynamic range between "loud" and "soft" claps due to strength variability. The *PedalBoard* python package was used for the gating and compression (Spotify 2021). Finally, the two channels were re-normalized and panned in stereo (70% mix ratio, Static-part on L channel, Shifting-part on R) to allow easier perceptual discrimination of the two players when auditioning the material.

[1]

[1] Examples of the listening material (reference baseline take, and example of rated take) can be found in the accessible dedicated media folder at the following link: https://drive.google.com/drive/folders/1i63EBwAcS0Jc7x2cR_7sKQ4riCTm7OnR?usp=sharing

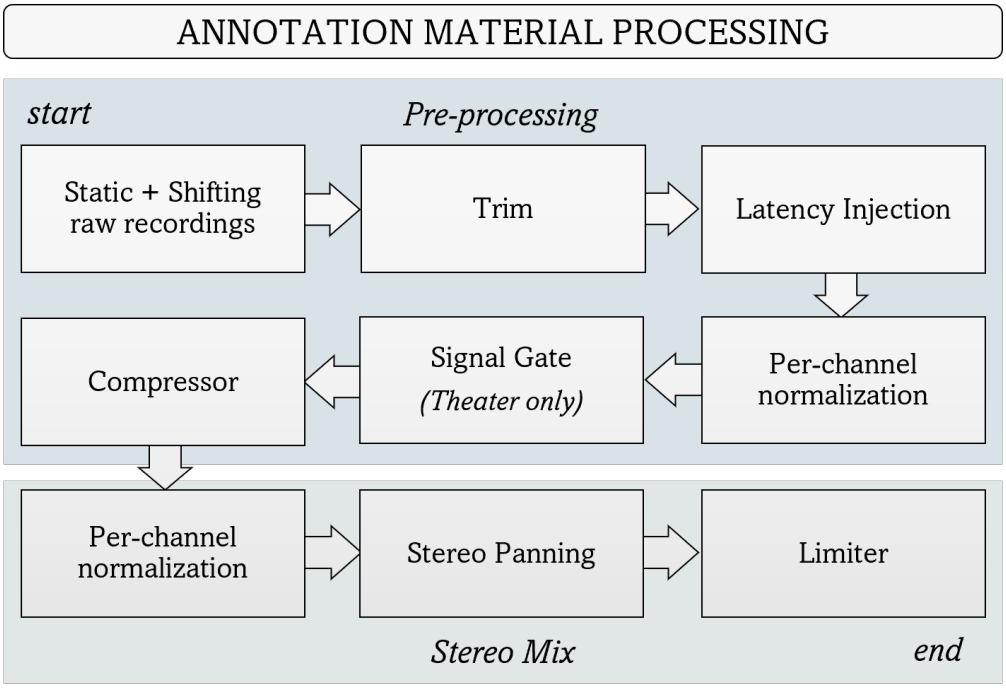


Figure 59: Signal processing flow for the creation of annotator's material from the recorded takes. The static and shifting recording are mixed together into a stereo mix as shown in the process above. The latency offset is injected in reference to the perspective experienced by the Static-part performer.

CHAPTER VII

ANALYSIS AND RESULTS

This chapter covers the statistical analysis performed on the collected secondary data consisting of three different layers of evaluations. The complete analysis framework is first described before presenting specific results. *Linear Mixed Effects Models* and *Generalized Linear Mixed Effects Models* were used as primary tools for statistical analysis of the main effects (latency and auralization mode), secondary effects, and random effects on the different types of observed variables and outcomes. Effect sizes and predictive relationships are reported for the models that exhibited the best fit. A correlation analysis is also performed on the data in order to capture the relationships between each layer of quality evaluation. The results are provided for all variables listed in Table 11, with deeper levels of detail for the responses on the evaluation of *Co-presence* from the trial questionnaires, since it is the principal component of the hypothesis space (and also to illustrate the statistical procedure).

To enhance the focus of this document and avoid overly verbose lengths, here is presented the complete set of analytic results, including diagnostics and validation graphs only for the principal variables investigated (*Copresence* and *Cohesion*). A reduced set of salient graphs, tables, and results, is provided for all other investigated variables on which significant results were observed. Nevertheless, the workflow of analysis described in Sect. 1 was conducted similarly, if not identically, for all observed dependent variables[1].

Note: Hereafter, in the document and in the presented plots, the auralization modes will be referred to by their acronym. Please refer to Table 1 for a quick overview of the reference codes used for each mode.

[1] Specific plots that did not appear in this manuscript are available on request.

1 Analysis Framework

The framework approach described in this section applied to the three layers of evaluation data variables: responses to the trial questionnaire, objective signal metrics indicative of the technical quality of musical performance, and third-party listening annotations and evaluations (see Sect. 2.2 for a summary list of all observed variables and Sect. 2.1 for a list of the considered fixed and random effects) for a full list of dependent variables and fixed effects). The main goals of the analysis stage regarded investigations of the impact of the declared effects on the observed evaluation scales and the performance of a correlation analysis to identify links and relationships between said scales. Furthermore, prediction estimations are provided using the output of regression-based mixed-effects models.

1.1 Mixed Effects Models

The complex and multilayered set of data types for observed variables and potentially influential effects called for a flexible approach capable of handling confounding factors and individual biases. The most appropriate way to handle the investigations in this landscape was the application of *Linear Mixed-effects Models* (LMM) and its extension *Generalized Linear Mixed-effects Models* (GLMM) to a regression-like framework, from which to extract effect-sizes and explanatory significance. Mixed-effects models are widely used in statistics because they provide a flexible and powerful framework for analyzing complex data structures and provide more accurate inferences compared to traditional linear models. In some cases, mixed-effects models can provide improved interpretability over traditional fixed-effects models. For example, by including random effects in the model, it is possible to separate subject-specific effects from population-level effects, which can make it easier to understand the results of the model and to make inferences about the population (Pinheiro and Bates 2000).

Fixed effects are a type of effect in a statistical model that are assumed to be constant across all observations in the study. In a mixed-effects model, fixed effects represent population-level trends of interest or relationships that are consistent across all subjects. Fixed effects are estimated using regression techniques and are typically represented as coefficients in the model which represent the magnitude of the effect of each predictor variable on the response variable,

and are estimated so as to minimize the residual sum of squares. Random effects, on the other hand, are effects that are considered to vary randomly across observations and are thus modeled as random variables that are drawn from a probability distribution. The advantage of this approach is that random effects can be used to account for individual response differences and biases by applying a random slope and random intercept to a set of non-independent observations (Clark and Linzer 2015).

The main difference between LMM and GLMM is the distributional assumption they make about the response variable. LMM assumes that the response variable follows a normal distribution and that the relationship between the predictors and the response is linear. This makes LMM appropriate for continuous response variables, such as height, weight, or temperature. In mathematical terms, the general formula for LMMs is given by Eq. 14, where y_{ij} is the response for the j^{th} measurement for the i^{th} individual, β_0 and β_1 are the fixed effects (in this example, two effects), x_{ij} is the predictor, b_i is the random effect for the i^{th} individual, and ϵ_{ij} is the residual error term.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij} \tag{14}$$

GLMMs, on the other hand, allow for non-normal response variables and non-linear relationships between predictors and response. GLMMs use a link function to relate the linear predictor to the response variable, and the response can follow a variety of distributions, such as the Bernoulli, Poisson, or exponential distribution. This makes GLMMs appropriate for count data, binary data, and other non-normally distributed response variables. The generic GLMM equation is given by Eq. 15, where $g^{-1}(\cdot)$ is the inverse link function, $E(y_{ij})$ is the expected value of the response for the j^{th} measurement for the i^{th} individual, β_0 and β_1 are the fixed effects, x_{ij} is the predictor, and b_i is the random effect for the i^{th} individual.

$$g^{-1}(E(y_{ij})) = \beta_0 + \beta_1 x_{ij} + b_i \tag{15}$$

The binomial logit link function is described in Eq. 16, where p is the probability of the event occurring (for example, the probability of a mistake to occur being equal to 1, or 100%).

$$g(p) = \log\left(\frac{p}{1-p}\right) \tag{16}$$

1.1.1 Model Selection

Several combinations of predictor effects within mixed models were tested using a "top-down" approach. This was done by progressively simplifying a saturated model by taking out fixed effects, which removal would increase the fitness metrics of the model and reduce complexity, but with consistent use of the main fixed effects (latency and mode) and of the same random effect (which in most cases was the ID of the participant or the ID of the pair of participants, according to the meaning of the metric).

BIC (*Bayesian Information Criterion*) and AIC (*Akaike Information Criterion*) were used to select the best model from a set of candidate models. When comparing candidate models, a lower BIC or AIC indicates a better fit of the model to the data, as they indicate an estimate of the prediction error (Kuha 2004; Neath and Cavanaugh 2012). Both measures indicate a trade-off between the goodness-of-fit and the simplicity of the model. The main difference between the two is the way they trade off the fit of the model to the data with the complexity of the model. BIC tends to prefer models with fewer parameters, whereas AIC tends to prefer models that fit the data well and pick up on subtler effects, even if they have more parameters. More precisely, to account for the limited pool size of subjects, here the AIC criterion is substituted with the AICc (*Corrected* AIC). AICc is derived from AIC by adding a correction factor that penalizes models with more parameters when the sample size is small. This correction factor helps to prevent overfitting, which can occur when the sample size is small and the model has many parameters. Mathematical formulas for AICc and BIC are given in the Appendix.

In most cases, the BIC and AICc rankings converged in pointing to the best candidate model to pass to the subsequent stages of analysis. As a further measure against overfitting and to focus the analysis on the variables with the highest explanatory power, the actual final model selected was the model that used the lowest number of parameters within 3 BIC / AICc points of the top-ranked candidate. In general, BIC is considered to be more conservative than AIC, as it places a stronger penalty on the number of parameters in the model, favoring candidates with lower dimensionality. Occasionally, the ranking distance between the BIC/AICc metrics was

higher than the tolerance given (i.e., within 3 ranking positions of each other), leading to further tests.

To determine which of the models provided the most explanatory power, an ANOVA *likelihood ratio test* (LRT) was executed between the BIC and AICc candidate models. The LRT test assesses whether a model is significantly a better fit to the data than another, indicating a significant p-value (Chi-sq statistic) if so. If no significant difference was detected, the simpler BIC model was ultimately selected, otherwise, the more complex AICc was chosen as able to provide a more appropriate fit to the data. The LRT test was also repeated with the selected model against the *null-model* (i.e., a mixed model calculated only with random effects and no fixed effects), in order to verify that the variables selected were able to provide a better model fit than just the random effect. If the chi-squared LRT test indicated a significant p-value, the fixed effects were influential in the rating of the variable outcome, otherwise, none of the studied fixed effect was explanatory towards the response variable.

As a side note, to allow model comparisons through BIC and AICc, LMMs were calculated using a *Maximum Likelihood* estimation method. After the selection stage, the chosen model was refitted using the *Restricted Maximum Likelihood* method, which more accurately estimates residual variance (Gilmour et al. 1995). The residual variance is used to estimate the uncertainty in the predictions made by the model.

1.1.2 Model Diagnostics

At this stage of the analysis, diagnostics plots were computed over the best candidate model to verify that the statistical assumptions were met and no strong outliers were influencing the regression fit. Linear mixed effects regression models are driven by the assumption of the dependent variable having a normal data distribution and that the residuals of the model are also roughly following a normal distribution. General mixed effects models instead assume other types of distributions such as *binomial* or *Poisson*.

One of the main diagnostics tools is the Q-Q plot. The Q-Q plot is a probability plot of the standardized residuals against the values that would be expected under normality (Marden 2004). If the model residuals are normally distributed then the points on this graph should fall on a straight line, if they don't, then the normality assumption is violated. Other types of diagnostics

involved distribution comparisons (normality of residuals), outlier detection, linearity, and homogeneity of variance. Multicollinearity was also tested using a *Variance Inflation Factor* (VIF) test, which points to issues of multicollinearity of the independent variables. A VIF value above 5 was deemed a sign of multicollinearity. In case the assumptions were not met, the model was deemed non-appropriate, and an alternative model or mixed-effects distribution family was selected.

1.2 Omnibus Test

The next step in the procedure was to run tests able to evaluate the overall significance of each one of the fixed effects within the validated model. Once having established the best candidate and verified that the assumptions are met, the model was refitted using a *Restricted maximum likelihood* (REML) to reduce bias, and a Chi-square *Likelihood Ratio Tests* (LRT) was then used to compare the full model with the fixed effect of interest to a reduced model without the fixed effect of interest. The LRT test is also used as *omnibus* test that informs on whether the main effects of the model are significant contributors to the model in general. This step serves to understand the importance of the contribution to the model that each independent variable has over the observations, and therefore - in addition to helping test the hypotheses of interest - it can help to keep or remove certain elements from the model with the knowledge that the removed variables are not significant predictors of the dependent variable. In other words, the LRT tests on the refitted model were used to test the null hypothesis that the reduced model, without a fixed effect, is as good of a fit as the model with the fixed effect.

If the normality assumptions were not inherent to the model (as the case for Binomial or Poisson distributions, where GLMMs were used instead of LMM), then a Type II or Type III *Wald's Chi-squared test* was used in the place of the LRT test for omnibus analysis. The difference between Type II and III is dictated by the presence of interaction terms. Type II is usually the preferred choice for understanding the impact of main effects, however, it does not assume the presence of interaction terms and it can provide ambiguous results when those are included. If an interaction is being investigated, Type III is the more appropriate kind of test as it allows for the evaluation of each predictor's unique contribution to the model after controlling for all other predictors in the model, including any interaction terms (Sahai and Ageel 2012).

1.3 Post-hoc Multiple Comparison Tests

Finally, a post hoc multiple comparison test was performed over each predicting variable in the model to observe the possibility of significant differences between each combination of categorical levels and gather estimates of the marginal means. The post hoc contrasts are obtained through pairwise multiple comparison tests using the estimated marginal means. In regression analysis, the marginal means represent the average value of the dependent variable for each level of the independent variable, after adjusting for the effects of any other predictor variables in the model, including random effects. These means can be used to make predictions about the dependent variable for a particular level of the independent variable. This is done either pairwise between all levels, or against a control condition (e.g., LAT=7ms or MODE=(R)). Significant differences between category pairs are taken from t-tests and would show a Tukey-adjusted p-value. The Tukey-adjusted p-value provides a more stringent test of significance than the unadjusted p-value, by taking into account the number of pairwise comparisons being conducted. The specific contrast of interest was adapted to the hypothesis of interest, the presence of interactions, and whether the contributing effects were main or secondary effects.

1.4 Correlation Analysis

A different type of analysis was conducted to find correlations among the full set of dependent variables. For this test, the *Pearson's r* coefficient was taken over the standardized continuous scales of response variables to create a correlation matrix. Correlation matrices were computed within layers (the "layers" being: questionnaire responses, objective metrics, and expert-listener annotation & ratings), and between layers. The purpose of this stage was to identify where relationships exist in the dataset and for consideration for future in-depth analysis. Categorical types of dependent variable responses were excluded from this analysis.

2 Data Formatting

Excluding rehearsals, the collected performance resulted in a total of 1088 individual recording takes, subdivided into 960 takes concerning the main phase (30 repetitions per 32 participant) and 128 individual takes concerning the baseline phase (four repetitions per participant). Given an

average take length of one minute and fifteen seconds, the total material amounted approximately to 20 hours of recorded audio data (22h and 40m with the baseline). In addition to the objective metric extraction of Ch. VI, the data formatting pre-analysis stage was also completed in *Python3*.

The full set of attributes and evaluations were formatted into a single dataframe source file. Each data entry in the source data represented a collected take for either the Static or Shifting musician (for a total of 960 entries analyzed). To each entry, its related subject/pair ID and condition effect attributes and evaluation results were appended. The observed variables described in Table 11 are related to each entry at either the subject level or the pair level. The values of the pair-level variables are identical across both player entries relating to the same trial take. Data from all 960 experiment entries were used for the questionnaire analysis. One session was fully discarded from the analysis of the objective metrics and third-party auditioning layers due to not meeting minimum quality levels (resulting in a total of 900 entries for those stages). The collection process and description of each outcome variable is described in the document section related to its evaluation layer of provenience in Ch. VI. Please, refer to Ch. VI, Sect. 3 for details about the trial questionnaire direct subjective evaluations, Sect. 4 for the objective metrics of technical performance extracted through beat-tracking algorithms, and Sect. 5 for an explanation of the third-party evaluation and rating process.

2.1 Fixed and Random Effects

This analysis considered as "main fixed effects" the fixed effects of *Latency* existing at three categorical levels labeled $[7ms, 20ms, 40ms]$ and *Auralization mode,* existing at five categorical levels labeled $[(R), (AC), (AD), (SC), (SD)]$. See Sect. 4.2 in Ch. IV or Table 1 for reference on the meaning of these label codes. These two elements are the most interesting effects to study for the discussion purposes set out by the hypotheses and research questions. The auralization effects are also explored for their grouping along the *symmetry* and *congruence* design axis (see Fig. 24).

Secondary effects were annexed to the main effects in the formulation of the model, with the expectation that they could provide a minor contribution to the explanatory power. At the individual participant level, the secondary fixed effects include the consideration of *Room location* and *Musical Part*. Other secondary fixed effects were drawn from the grouping of results from the pre- and post-experiment questionnaires (**Q1** and **Q3**). Examples are questionnaire derivations

Table 10: Summary of Effects

Name	Type	Range/Levels	Description
Random Effects			
Subject ID	Categorical	[1 to 32]	ID tag for the data entry participant
Pair ID	Categorical	[1 to 16]	ID tag for pair of assignment
Field ID	Categorical	*['MT', 'MJ', 'P', 'CS', 'MB']*	ID of subject's major subject of expertise (from **Q1**,Sect. 2.1.1).
Main Fixed Effects			
Latency	Categorical	$[7, 20, 40]$	Level of one-way latency applied in the trial.
Auralization mode	Categorical	*['R', 'AC', 'AD', 'SC', 'SD']*	Category of virtual acoustic intervention applied in the trial.
is Symmetric	Categorical	$[0, 1]$	Grouping of auralization mode category by Symmetry.
is Congruent	Categorical	$[0, 1]$	Grouping of auralization mode category by Congruence.
Secondary Fixed Effects			
Part	Categorical	*['Static', 'Shifting']*	Musical part related to trial data entry
Room	Categorical	*['Theater', 'Booth']*	Room location related to trial data entry
Trial Number*	Numerical	[1 to 30]	Trial chronological sequence number
Music YOE	Numerical	[4 to 25]	Years of music performing experience related to subj. of data entry
Familiarity	Numerical	[0 to 15]	Aggregated Familiarity score derived from pre-exp. questionnaire **Q1.1-1.3**).
NMP Experience	Numerical	[0 to 15]	Aggregated NMP-experience score derived from pre-exp. questionnaire **Q1.4-1.6.**
Bias	Numerical	[0 to 10]	Aggregated expectation-bias score derived from pre-exp. questionnaire **Q1.7-1.8** .
Fatigue	Categorical	[1 to 5]	Final fatigue level from questionnaire **Q3.**

Table 10: Fixed and Random effects used for the mixed-effects models estimations. **Note: Trial Number is tested in both fixed and random effect form.*

such as *level of familiarity* or parameters such as *years of musical experience* or *fatigue level*. Aggregate scores for *Familiarity, NMP-experience* and *Bias* were constructed by composing the ratings from the related question groups in Q1 (see Sect. 2.1.1, in Ch. VI). Questions Q1.1 to 1.3 summed to a "Familiarity score", Q1.4 to 1.6 formed a "NMP Experience score" and Q1.7 and 1.8 aggregated to a "Bias" score. These secondary effect scores were explored both in the form of continuous scales and as a median-split group division factor (e.g., "lower familiarity group" vs "higher familiarity group").

In regards to random effects, the ID code of individual participants - or the ID of the pair of participants - was always used as random intercept for the model (according to whether the outcome variable observed pertained to the individual or the pair). The application of a subject-level random effect was particularly necessary in the case of this analysis, as latent confounding variables lie not only in the performer's internal response biases but also in their experience as musicians (Carôt et al. 2009). A non-trivial concern was the handling of *Trial number* as a fixed or random effect. This measure reflected the progress in time of a subject throughout the experiment and could be used as a proxy to determine a general training effect (i.e., how much improvement a subject presents based on trial repetitions alone) or experience degradation due to accumulated fatigue. Within this analysis framework,*Trial Number* is used in most cases as a continuous secondary fixed effect, but its influence as factorial crossed random effects is also explored for observed variables where training effects are not expected to have an impact on the measure.

2.2 Summary of Variables

For each investigated dependent outcome variable, the specific mixed-effects model topology is adapted according to the nature of the observed outcome distribution. Most evaluation scales present a normal distribution, and thus *Linear Mixed-effects Models* (LMM) were used to construct a set of significant predictors for each investigation. All 7-point Likert-like scale ratings from the trial questionnaire **Q2** are treated as numeric variables with normal distribution, this decision is validated by Statistics literature (Harpe 2015). All numerical continuous scales were later standardized for comparable analysis and correlation metrics. The third-party binary annotations described in Sect. 5 are instead analyzed using *Generalized Linear Mixed-effects Models* (GLMM)

which can account for variables that possess a binomial or Poisson type distribution (Faraway 2016). Binomial GLMMs with logit link functions are used to model the relationship between linear predictors and the log odds of the binomial response being equal to one. The effect estimation translates to the log odds of increasing (or decreasing) the probability of each type of mistake occurring in response to a change in the fixed effect levels.

2.2.1 Trial Questionnaire Data

Results from the Trial Questionnaire (Q2) data were parsed as collected for questions Q2.1 to 2.4. A new variable called *Immersion Score* was calculated by summing the responses to the Likert agreement questions Q2.5 to 2.9. This score represented a generic quality of immersion representation, standardized as a continuous scale, which was tested for consistency using *Cronbach's alpha*. The distribution of the responses to the agreement questions was also explored individually to address some of the more qualitative explorations in regards to the auralization mode effectiveness at eliciting "presence".

2.2.2 Objective Metrics

All metrics of the distributed data were transformed into relative ratio terms. This allowed for comparative analysis across pairs by controlling for their relative technical ability. The approach was to take the log ratio of the distributed metric of each entry to the average baseline metric for the same subject ID. The log ratio transformation indicated the absolute level of change from a reference level. Log units make it easier to interpret decimal ratios as negative numbers, while the ratio indicates an absolute degree of change rather than the average metric level (which could present averaging fallacies).

$$\texttt{RELATIVE METRIC}_{\texttt{(trial;subj.)}} = \log_{10}\left(\sqrt{\left(\frac{\texttt{DISTRIBUTED METRIC}_{\texttt{(trial;subj.)}}}{\texttt{AVG. BASELINE METRIC}_{\texttt{(subj.)}}}\right)^2}\right) \tag{17}$$

It needs to be noted that the limitation of ratio results is that an assumption is made that musicians showed their best performance at the baseline primary-data capture and thus "less change" is "better", while this may not be necessarily the case if the player improved through repetitions. So the transform loses the ability to put into context whether the change was an

Table 11: Summary of observed outcome variables

Name	Distribution	Range/Levels	Description
Trial Questionnaire (Individual level)			
Accuracy	Normal	$[1 \texttt{ to } 7]$	Trial accuracy impression score **(Q2.1)**
Difficulty	Normal	$[1 \texttt{ to } 7]$	Trial difficulty impression score **(Q2.2)**
Copresence	Normal	$[1 \texttt{ to } 7]$	Auditory copresence score **(Q2.3)**
Cohesion	Normal	$[1 \texttt{ to } 7]$	Auditory cohesion score **(Q2.4)**
Immersion score	Normal	$[1 \texttt{ to } 7]$	Aggregate score from agreement scales (Q2.5 to 2.9)
Objective Metrics (Individual and Pair level)			
Static Tempo	Normal	$[0 \texttt{ to } \text{inf}]$	Overall BPM tempo estimated
Tempo Range	Normal	$[0 \texttt{ to } \text{inf}]$	Range between minimum and max BPM from the dynamic tempo curve
Tempo slope	Normal	$\pm \text{inf}$	Slope of the tempo trend across the take
Pacing (π)	Normal	$[0 \texttt{ to } \text{inf}]$	Mean IBI (quarter beats)
Regularity (ρ)	Normal	$[0 \texttt{ to } \text{inf}]$	Coefficient of IBI variability
Mean-lag (α)	Normal	$[0 \texttt{ to } \text{inf}]$	Mean of the inter-subject absolute beat differences
Imprecision (μ)	Normal	$[0 \texttt{ to } \text{inf}]$	Standard deviation of the inter-subject beat differences
Third-party annotations and subjective ratings (Pair level)			
Pattern Mistake	Binomial	$[0, 1]$	Evident mistake in the clapping pattern in the take, whether with missing or extra claps
Tempo Inaccuracy	Binomial	$[0, 1]$	Presence of acceleration or deceleration heard in take
Synch. Inaccuracy	Binomial	$[0, 1]$	Noticeable misalignment in synchronization heard in the take
Stopped Performance	Binomial	$[0, 1]$	Detection of performer stopping the performance for a few seconds, or not finishing up the end
Tot. Mistakes Count	Poisson	$[0 \texttt{ to } \text{inf}]$	Total amount of mistakes observed for all annotated categories
Tempo Rating	Normal	$[1 \texttt{ to } 10]$	Rating of tempo stability perceived throughout take

continues on next page

Third-party annotations and subjective ratings (Pair level) *(...cont.)*			
Accuracy Rating	Normal	[1 to 10]	Rating of clapping pattern accuracy heard throughout take
Synch Rating	Normal	[1 to 10]	Rating of synchronization perceived throughout take
Overall Rating	Normal	[1 to 10]	Overall rating of performance as a whole

Table 11: Complete list of the observed outcome-dependent variables pertaining to different layers of evaluations. These subjective and objective metrics come from direct evaluations from participants during the study, objective metrics from a beat-tracking signal analysis, and third-party annotator evaluations.

improvement or a degradation compared to what the baseline was. Instead, we can assess how similar the trial performances were to the baseline, in response to the studied factors.

The entirety of session 11 was discarded from the analysis due to non-satisfactory performance execution observed at the time of collection. In addition, due to the possible presence of algorithmic artifacts, caused by ambiguities in the beat pattern, strong outliers were removed from the data using a threshold of 1.5 times the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile. A total of 185 entries were removed from the original 960.

2.2.3 Annotation Data

Furthermore, to reduce the influence of annotators' bias in judging the musical quality of a take, the values of the continuous scale ratings were standardized using Z-score, calculated per ID of the evaluator. The formula for a z-score transformation is given by Eq. 18 where x is the original value, μ is the mean of the original values, and σ is the standard deviation of the original values. The transformed value, z, is the number of standard deviations away from the mean that the original value is. This process converts the set of values into a standard normal distribution with a mean of 0 and a standard deviation of 1.

$$z = \frac{x - \mu}{\sigma} \tag{18}$$

To reduce the number of dependent variables at hand, some of the binomial annotations were aggregated into compound annotations. Specifically, the *Acceleration* and *Deceleration*

annotations were aggregated into a single *Tempo Inaccuracy* category through an "OR" logic operation between the two columns. The same was applied to the *Extra Claps* and *Missed Claps* annotations, aggregated into a *Pattern Mistake* category, and also to the *Silences* and *Unfinished* annotations, aggregated into a *Stopped Performance* category. Strong outliers were removed from the analysis set due to the possibility of artifacts from the tempo extraction algorithms.

3 Results

This section illustrates key results derived from the analysis workflow explained in this chapter. Details are provided on the mixed effect model outputs in terms of the model-fit statistics, the effect estimate, the standard error, and the significant difference from the reference factor level. The reference levels for the main effects here are always the minimum latency (LAT$=7$) and the Raw auralization mode (MODE = R). The output of the post hoc pairwise comparisons is provided for those models that showed patterns of significance in the main effects. A larger focus and space are given to the subjective questionnaire variables forming the key components of the hypothesis space. To better understand the model output in the context of its scale, the model output results are not standardized. However, standardization was applied to the correlative analysis and to the models in which more than one continuous independent variable was applied. The *"R"* software tools were used for the application of the analysis framework.

More attention is given to the case of the *copresence* variable. For that case, and to further illustrate the workflow and methodology around the results, the document reports the full set of distribution and trend plots, the model selection tables (including null models), and model diagnostics plots, as well as detailed steps of the model validation process. These in-between steps are omitted from the results concerning other variables but are still consistent parts of the underlying analysis framework that was applied to every variable reported. To enhance focus, the document does not report on the models concerning secondary variables where no significant effect sizes were found.

Other support data results, such as the trends observed in the Debrief Questionnaire (Q3) and the individual copresence agreement scales from the Trial Questionnaire (Q2) are also reported to provide validatory context to the experimental design and to help discuss interpretable nuances of the statistical results.

3.1 Main Effects Overview

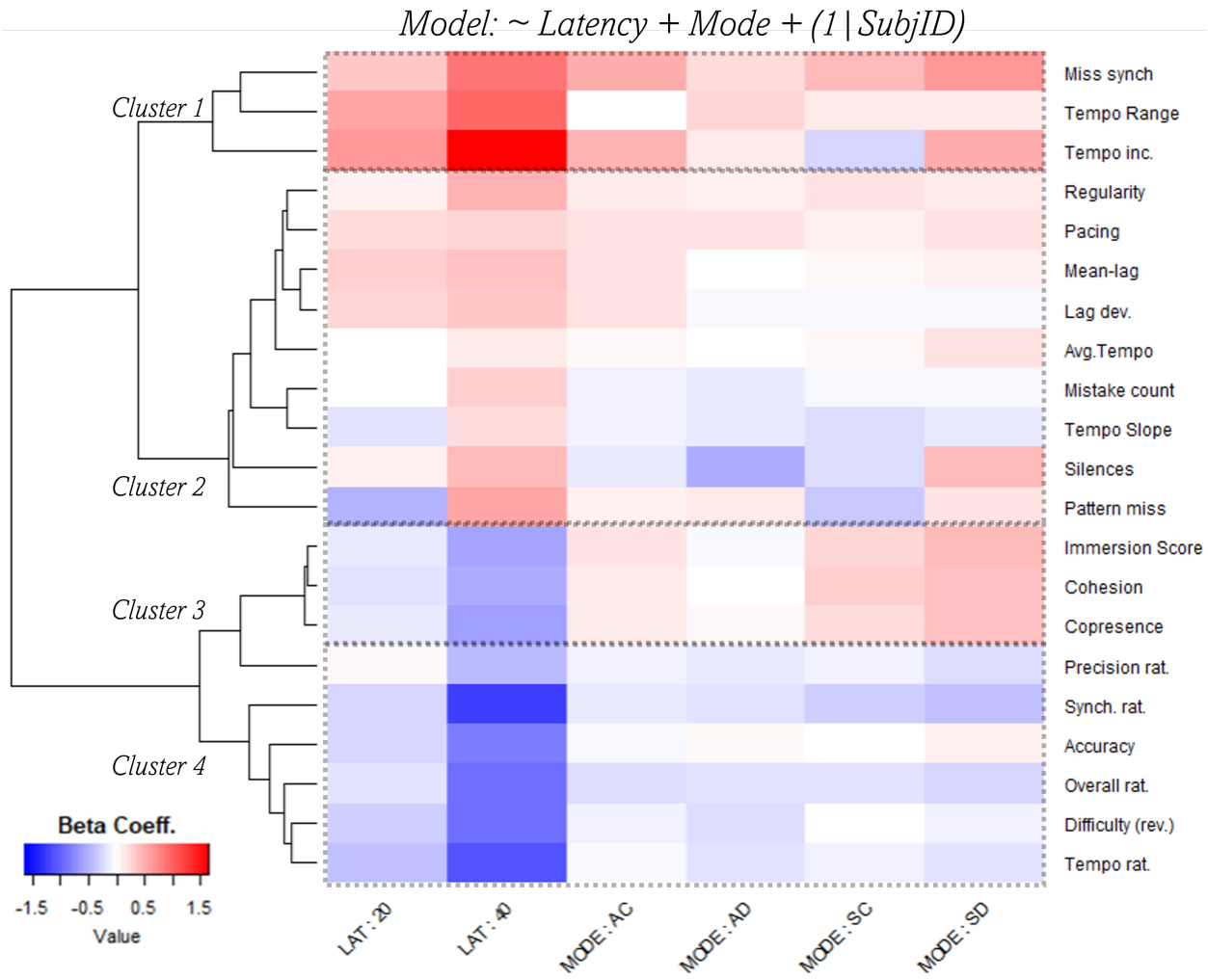


Figure 60: Overview of beta coefficient for latency levels and auralization modes for all models. Colors indicate the sign and magnitude of the model's beta coefficient in relation to their reference level (control group). Models are computed over scaled metrics. Rows are clustered per similarity. Cluster groups are overlaid on the heat-map plot.

A first overview of the impact of the main effects was explored by running the "base" mixed model over all the observed variables. The base model only included the main fixed effects under study (latency and auralization) and the random effect of subject's ID. In formulaic terms, the base model was defined as $\sim Latency + Mode + (1|SubjID)$. The goal of this step was to provide a first glance at how the main effects impacted the dependent variables, before going deeper into the model-selection phase to find better-fitting alternative models and significant sets of contributors.

The results of the base model over the full set of observed dependent variables are summarized in Fig. 60. The heat map shows the magnitude of the models' beta coefficients for each level of latency (in relationship to the lowest latency level) and each auralization mode (in relation to the "Raw" mode). The intensity of the color of each cell indicates the magnitude of the coefficient, with blue indicating a predicted decrease of the dependent variable, and red indicating a predicted increase. For comparable results, the base model coefficients were calculated (using *Restricted Maximum Likelihood*) over normalized versions of the variables (mean-centered, and divided by the standard deviation) making the reading in terms of multiples of standard deviations from the mean. Finally, the rows are reorganized based on similarity clusters in order to better visually highlight related groups of dependent variables. Since the base model was here not compared across different models, no particular fixed effect significance is implied by the plot. An alternative version of this plot, including the effect of "Trial number" is included in the Appendices.

The first evident trend shown in the plot is the dominant impact of latency, especially at its highest level (40ms). The auralization modes show a milder effect than the latency levels, with the possible exception of the *Symmetric Divergent* mode, showing some higher influence in two clusters of variables (clusters 1 and 3). The coefficient clusters form interesting groups (highlighted with overlaid dotted boxes) showing which variables are most closely impacted by the main effects. These clusters have been labeled from 1 to 4. Cluster 1 groups together annotations related to synchronization and tempo accuracy with the objective metric of tempo range (related to accelerations or decelerations in performance). In this cluster, almost every level of the main effects had some sort of impact on increasing the rate of tempo and synchrony inaccuracy and the range of the tempo-curve. Cluster 2 groups together other objective metrics and annotations of various kinds, where the effects are seen to have had a minor impact, besides perhaps the highest latency level. Sub-clusters in this group show expected relations between similar metrics. Cluster 3 is perhaps the most interesting grouping, showing that the first-hand "immersion-related" metrics are all related, negatively affected by latency, and positively affected by the auralization modes, especially by the SD mode, and with the exception of the AC mode. Cluster 4 groups together various kinds of third-party ratings with the performers' self-ratings (in the forms of "Perceived Accuracy" and "Perceived Difficulty"). Most effects led towards negative trends in

quality ratings within this group, at various degrees of magnitudes. Latency once again dominates the impact in decreasing the scores, while auralization modes had a minor contribution.

Although this first representation helped to set the expectations, more analysis was necessary to quantify, control, and compare these effects when accounting for other confounding variables. In many of these models, the auralization modes were eventually found not to be a significant contributor to the model fit (with other alternative fixed effects presenting themselves as significant). This "discovery" phase, described below, constituted the bulk of the analysis effort.

3.2 Full Step-by-step Case: *Copresence*

This step-by-step presentation of the analysis process for the *copresence* variable is portrayed in detail to both reinforce the validity of the applied statistical process and to allow stronger conversations around the main "protagonist" of the whole set of independent variables. Since the auralization modes are specifically designed to elicit copresence, a measure of their success can be extrapolated from their inferential power indicated by LMM models. The intermediate steps that lead to a model selection and validation are here shown just for the copresence case in order to better illustrate the analysis procedure. Similar steps were taken for all the other dependent variables to reach a final candidate of the regression model.

A preliminary distribution plot is shown to set the expectation of the Copresence response variable behavior in regard to the principal elements under investigation (Fig. 61). The image immediately shows that latency is expected to be a more influential group than the auralization mode, although some fluctuations are observed on the mode as well. Figures 62 and 63 anticipate the trends that turned out to produce the most meaningful model, by showing the detrimental effect of Latency over all modes and the interaction trends between the room type and the auralization mode.

The first step in the process was the establishment of an appropriate "null model" representing the random effects considered for the variable at hand. This was achieved by ranking a series of null models created with different potential random intercepts (always excluding the declared fixed effects). The ranking was done through AIC and BIC metrics (Tab. 12). In the vast majority of the cases, the "best" null model was the one using the ID of the subjects as random

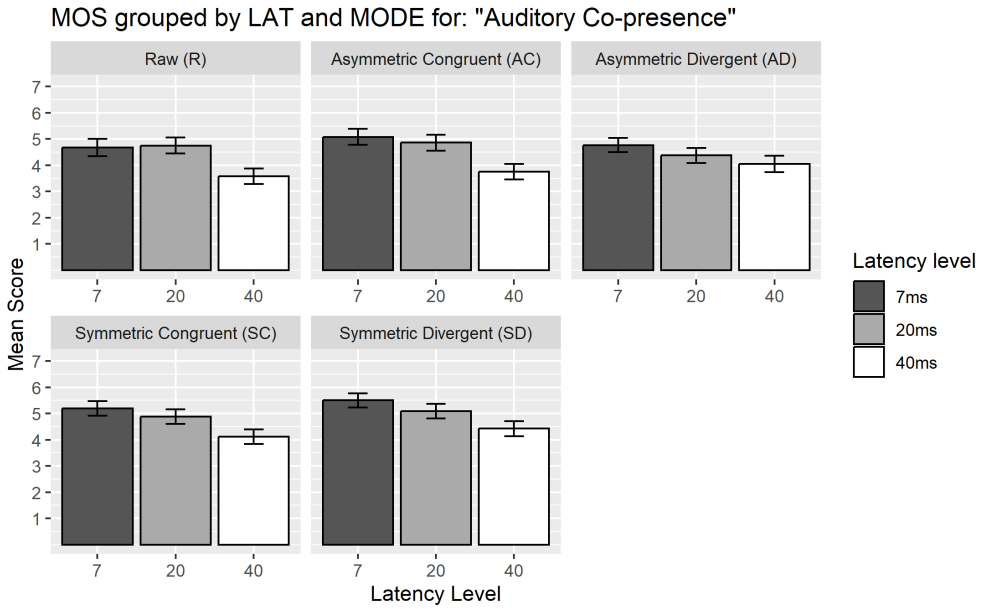


Figure 61: Mean opinion score distributions of *copresence* across auralization modes and latency levels

intercept (or the ID of the pair in case the dependent variable comprised a pair-level observation). Nested random effects were also considered in relation to subgroups of demographics attributes (e.g. "Familiarity group"), however, no null model with nested random effects was eventually selected. The details of the null model are given in Tab. 13, which shows the intercept effect estimate, the standard error, and whether a significant p-value was found. In this case, the intercept is significantly different from "zero".

Tab 12: Null-model selection *"Copresence"*

Model Names	K_i	BIC_i	Δ(BIC)	w_i(BIC)	$AICc_i$	Δ(AICc)	w_i(AICc)	log(L_i)
SubjID	3	3643.22	0.00	0.91	3628.65	0.00	0.46	-1811.31
SubjID + Trial#	4	3649.77	6.55	0.03	3630.34	1.70	0.20	-1811.15
SubjID/RoomID	4	3650.09	6.87	0.03	3630.66	2.02	0.17	-1811.31
SubjID/Part	4	3650.09	6.87	0.03	3630.66	2.02	0.17	-1811.31

Table 12: Null model selection showing the BIC/AIC ranking table among the top *"Co-presence"* models with only random effects present, with the best model at the top. Columns represent "number of factors", "BIC score", "BIC distance from best" and "selection weight".

Once the null model was finalized, the next step in the process was the creation of a series

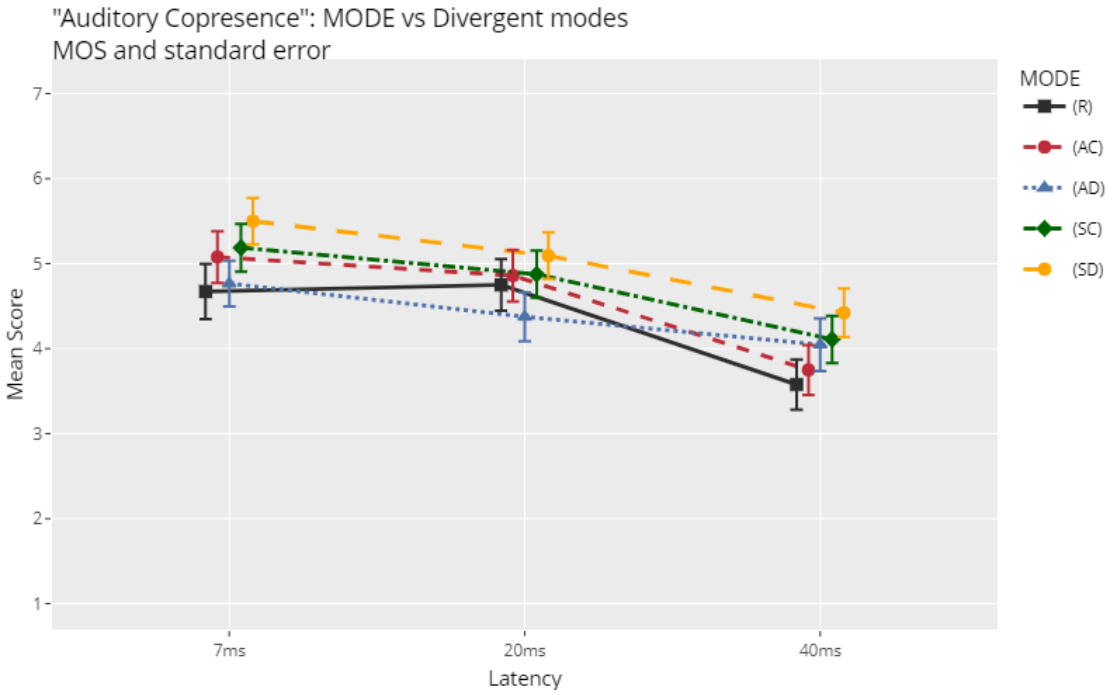


Figure 62: Copresence responses divided by auralization MODE and plotted over latency levels

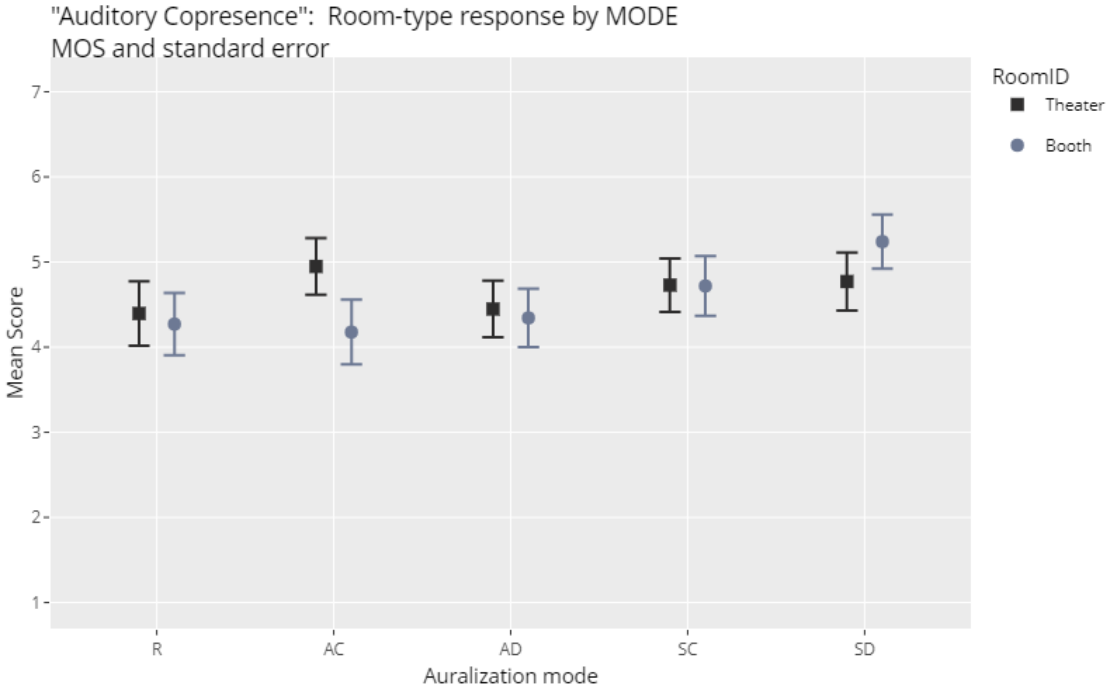


Figure 63: Copresence responses divided by room location ("Theater" vs "Booth") plotted over auralization modes

	Copresence - null model
Effect estimates	
Intercept	**4.60*****
- std. error	(.14)
Fitness Statistics	
AIC	3628.62
BIC	3643.22
Log Likelihood	−1811.31
Num. obs.	960
Num. groups: SubjID	32
Var: SubjID (Intercept)	.58
Var: Residual	2.37

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 13: Null model output statistics (random factors only) showing the effect estimate and model fitness parameters, for Copresence. Table shows effect estimate, (standard error). Asterisks denote significant p-values

of "candidate" full mixed effects models (by LMM or GLMM according to the distribution of the outcome data) ranked through the AICc and BIC fitness indicators and fitted using *maximum likelihood estimation*. The models always comprised the previously identified random effect pertaining to the null model, and the two main effects (auralization mode and latency). Secondary effects and interaction combinations were introduced using the variables listed in Tab. 10. The selection of the model to be fed to further stages of analysis was thus made through the ranking. Generally, the simpler BIC candidate model was favored over more-complex AIC candidate models, and the two metrics agreed in most cases. However, in situations such as this one, enough distance in the ranking was present such that the models needed to be compared a little more in depth. In the case of *Copresence* the BIC and AICc metrics favored different candidates, as shown by Tab. 14 (several non-influential model candidates are discarded from the table for clarity), the BIC favored a two-factor model $\sim LAT + isSymmetric + (1|SubjID)$ (*"isSymmetric"* indicating the pooling of the (SC) and (SD) modes into a single category) while the AICc metric favored a more complex model which included an interaction $\sim LAT + (MODE * RoomID) + (1|SubjID)$. The AICc model ultimately turned out to be chosen as the best representative model after an in-depth comparison of the models' inference output.

Tab 14: Full model selection *"Copresence"*

Model Names	K_i	BIC_i	Δ(BIC)	w_i(BIC)	$AICc_i$	Δ(AICc)	w_i(AICc)	$\log(L_i)$
$LAT + Symmetric$	7	3562.58	0.00	0.97	3528.62	9.93	0.01	-1757.25
LAT	5	3570.41	7.83	0.02	3546.14	27.44	0.00	-1768.04
$LAT + MODE$	9	3571.44	8.87	0.01	3527.83	9.13	0.01	-1754.82
$LAT + (MODE * RoomID)$	14	3586.39	23.81	0.00	3518.70	0.00	0.98	-1745.13
$LAT * Symmetric$	11	3586.95	24.37	0.00	3533.69	14.99	0.00	-1755.71
$LAT * MODE$	17	3618.01	55.44	0.00	3535.93	17.23	0.00	-1750.64
$MODE$	7	3646.64	84.06	0.00	3612.69	93.99	0.00	-1799.29

Table 14: Mixed effect model candidates for *copresence* ranked by BIC (reduced set). *K* is the number of model parameters. For this particular case the BIC and the AICc metrics disagreed, leading to further tests of selection over the two candidates with the highest inferential power.

Table 15 shows the best-performing BIC-selected model against the best-performing AICc model. The table indicates the fitness results of the mixed models, the regression coefficients, and whether significance is detected in the individual contrasts between each level of the independent variables with their reference group. The asterisk marks indicate significant p-values obtained through t-tests using a Type III *Satterthwaite's* method between each independent variable level and its relative reference level or category (for auralization that reference is always the MODE = $Raw(R)$ level and for latency is always the low latency level, LAT = $7ms$). For *copresence*, the BIC-candidate model output highlighted the significant effect of Symmetry in pooled auralization modes (in reference to the "Raw" condition), while the AICc metric favored a mode complex model that considered the individual auralization modes and their interaction with the Room in which the participant rated copresence (Theater or Booth). Both models showed a significant negative influence of latency levels (contrasted to the baseline level of 7ms), with the "highest-latency" level being indicated as the largest coefficient estimate across all models. The AIC-candidate model results suggested that the room location of a participant influenced the copresence ratings when interacting with certain modes, in particular when evaluating the effects of the (AC) and (SD) modes. This interaction is further supported by the plot of Fig. 63 that shows a noticeable increase in ratings (of about 0.65 score points on a scale of 7-points) in the AC mode for the participant in the Theatre room compared to the ratings that occurred at the Booth location. The trends are instead inverted for the SD condition, where the rating in the Theater was lower to the rating in

the Booth for the same auralization mode (by about 0.59 points on the 7-point rating scale). The BIC model instead pointed to the fact that *symmetric* modes had a statistically significant impact in the rating of copresence, leading towards higher scores compared to the reference level (the *Raw (R)* mode). This is instead not the case for the *asymmetric modes*, indicating no particular general difference from the (R) condition. However, the BIC candidate model failed to capture the observation that copresence response in the asymmetric (AC) mode in the Theater room was actually much higher than the response in the ISO booth room.

To further help in the decision of which model was to be considered the "best fit", an analysis of variance test between the two models was performed. This test helped to understand whether the more complex AIC model was able to capture the data significantly better than the simpler BIC model. As indicated by the resulting p-value of the test, shown in Tab. 16, the AIC model is a statistically significant better fit than the BIC model and is thus ultimately favored over it. For further validations, the two models were also tested against the null model (which only includes the random intercept), as shown in the same table. Both models were significantly more explanatory than the null model.

As shown in the graphs of Figs. 64 and 65, the assumptions were roughly met to a satisfactory level, and no outliers were detected (as expected since the scale of rating for copresence was bounded from 1 to 7). A VIF multicollinearity analysis was also computed, albeit without the interaction terms as those are expected to artificially inflate the VIF value (Allison 2012)). VIF values > 5 would affect the process by indicating the need to go back to the model selection stage and consider alternative models, but in this case, where all variables involved were categorical, no multicollinearities were found. The diagnostics showed a reasonable fit of the assumptions and no multicollinearity was detected.

Since the data followed a roughly normal distribution, a *Likelihood Ratio* test was used to determine the general contributions of each independent variable to the model and assess their significance. This gives an opportunity to understand the statistical importance of the identified categorical effect factors on the final model against their removal from a reduced model. Tab. 17 confirms that the impact of Latency on the model is the most significant, with a Chi-square related p-value $p < 0.001$, the factor of Mode also significant, but only at the $p < 0.05$ significance level. "RoomID" by itself was instead shown as not significant as a standalone component. However, the

Summaries of best models for *Copresence*

	LAT + Symmetric	**LAT + (MODE ∗ RoomID)**
Effect estimates		
Intercept	**4.77** (.19)***	**4.83** (.21)***
LAT(20ms)	−.25 (.12)*	−.25 (.11)*
LAT(40ms)	−**1.06** (.12)***	−**1.06** (.11)***
Symmetric(0)	.15 (.13)	
Symmetric(1)	.**53** (.13)***	
MODE(AC)		.**55** (.21)**
MODE(AD)		.05 (.21)
MODE(SC)		.33 (.21)
MODE(SD)		.37 (.21)
RoomID(Booth)		−.13 (.21)
MODE(AC):RoomID(Booth)		−.65 (.29)*
MODE(AD):RoomID(Booth)		.02 (.29)
MODE(SC):RoomID(Booth)		.11 (.29)
MODE(SD):RoomID(Booth)		.59 (.29)*
Fitness Statistics		
AIC	3540.91	3541.27
BIC	3574.98	3609.40
Log Likelihood	−1763.46	−1756.63
Num. obs.	960	960
Num. groups: SubjID	32	32
Var: SubjID (Intercept)	.61	.61
Var: Residual	2.12	2.08

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 15: Model results for the best BIC and AICc candidate models showing the coefficient estimates and fitness parameters, for the prediction of Copresence. *SubjID* is used as random effect in all models. Asterisks denote significant p-values. In this particular instance the two best candidate models are able to highlight different effects. The AIC model is eventually chosen as the best fit.

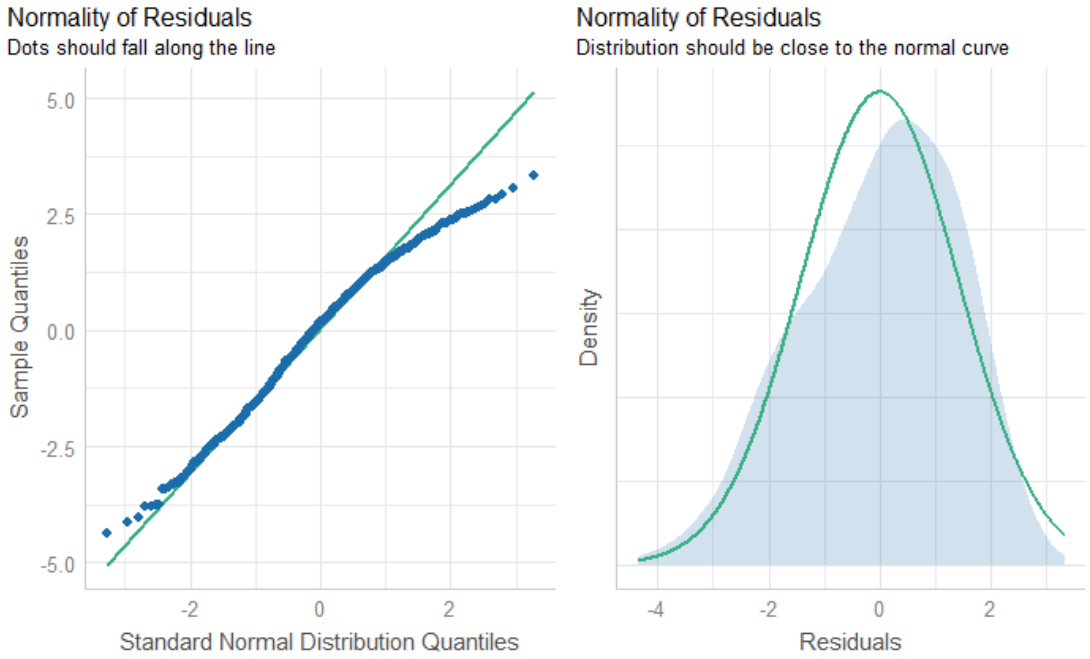


Figure 64: Diagnostics plots checking that normality assumptions are met. The first graph shows the Q-Q plot used to assess the normality of residuals.

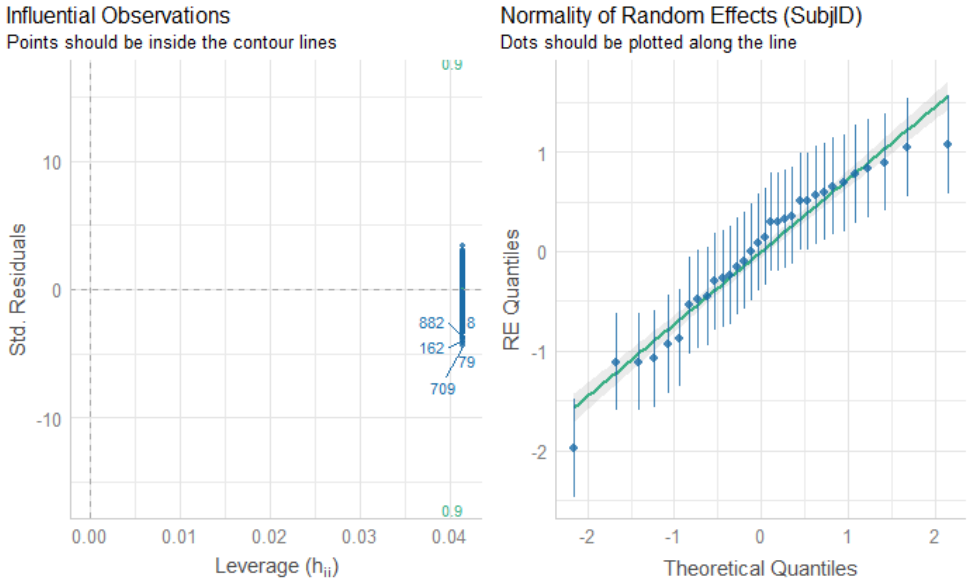


Figure 65: Diagnostics plots showing outlier detection plot and Q-Q plot for assessing the normality of the residuals of the random effects

Analysis of Variance

Models:

null: $Copresence \sim (1|SubjID)$

BEST_BIC: $Copresence \sim LAT + Symmetric + (1|SubjID)$

BEST_AIC: $Copresence \sim LAT + (MODE * RoomID) + (1|SubjID)$

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
null	3	3628.62	3643.22	-1811.31	3622.62			
BEST_BIC	7	3528.51	3562.58	-1757.25	3514.51	108.11	4	2.2e-16 ***
BEST_AIC	14	3518.25	3586.39	-1745.13	3490.25	24.25	7	0.001 **

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 16: ANOVA using Chi-Square test to see if proposed models are significantly different from each other and from the null model. Models were ranked for complexity and tested against the previous simpler model. In this case the best BIC model was significantly better in fitting the data than the null model, and the AIC model was in turn significantly better than the BIC model.

interaction of Mode and RoomID was instead significant at the $p < 0.01$ significance level. The intercept was separately also found to be significant by using a *Wald's Chi-sq* test.

Copresence: omnibus likelihood-ratio test

Model: $\sim LAT + (MODE * RoomID) + (1|SubjID)$

	Df	Chisq	Chi Df	Pr(>Chisq)
LAT	12	90.719	2	**2e-16*** **
MODE	10	9.956	4	**0.041***
RoomID	13	0.364	1	0.546
MODE:RoomID	10	18.047	4	**0.001****

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Df full model: 14

Table 17: Omnibus likelihood-ratio test used to identify the significance of the model factors. Results relate to the best model used to predict *Copresence*

The next step of the analysis consisted in analyzing the final model's predictive effects in order to report an overview of the trends and significant differences identified in the model. This allowed to enhance the resolution of the significance within each independent predictor and unpack the interaction terms. Figs. 66 and 67 show the standardized effect estimates extracted from the model. The first plot shows the effect estimates for each variable rated against its relative reference group (which are $LAT = 7ms$, $MODE = (R)$, and $RoomID = Theater$). Results show

a strong significant effect of Latency and significant interactions of the Room with the AC and SD modes. The effects of MODE and RoomID independently of the interaction are also portrayed but those are potentially misleading as there is no differentiation between the interacting terms. To address that, the bottom plot reveals how the (AC) mode performed better in the Theater room rather than the Booth, while the opposite is true for the (SD) mode. This point to the fact that the success of the auralization modes in eliciting Copresence in participants depended from where the mode was applied. The suggestion here is that the more reverberant Theater room benefited from the asymmetric congruent condition more than the dryer Booth room, while the opposite is true for the Symmetric Divergent mode.

Finally, a post hoc multiple comparisons test is used to see how the levels compared on average against each other. All levels of LAT were tested against each other. For the MODE:Room interaction, the levels were tested as grouped by room, against the control "Raw" condition. Tab. 18 shows the estimated marginal means, standard errors, and p-values for pairwise contrasts for the Latency variable, and "treatment vs control" for the auralization mode, grouped by room. The pairwise test indicated that the 20ms level of latency was not significantly different from the baseline of 7ms. This difference from the previous suggestions that indicated the 20ms level as significantly different from the 7ms level is motivated by the application of the Tukey adjustment, which is a more conservative in the estimation of the p-value and corrects it to account for Type I error inherent to multiple comparison tests. However, the 40ms level was significantly different from both the 7ms and the 20ms level, indicating that the degrading effect of latency on Copresence is non-linear and - by looking at the estimate - relatively strong. In regards to the mode and room interaction, within the Theater room, the (AC) condition proved to be significant at the $p < 0.05$ level against the raw condition, with a mean rating increase of 0.552 points. Other modes were not significantly different than the control mode. For the Booth room, the (SD) condition proved to be highly significant with $p < 0.001$ for a mean increase close to 1 point on the rating scale.

Figs. 68 and 69 show the model's estimated marginal means for latency levels and combinations of room and auralization modes. These plots visualize the magnitude of the difference in copresence prediction (on a scale of 1-7) as affected by the independent variables across the average of all other variables. The latency plot shows a sharp exponential decrease

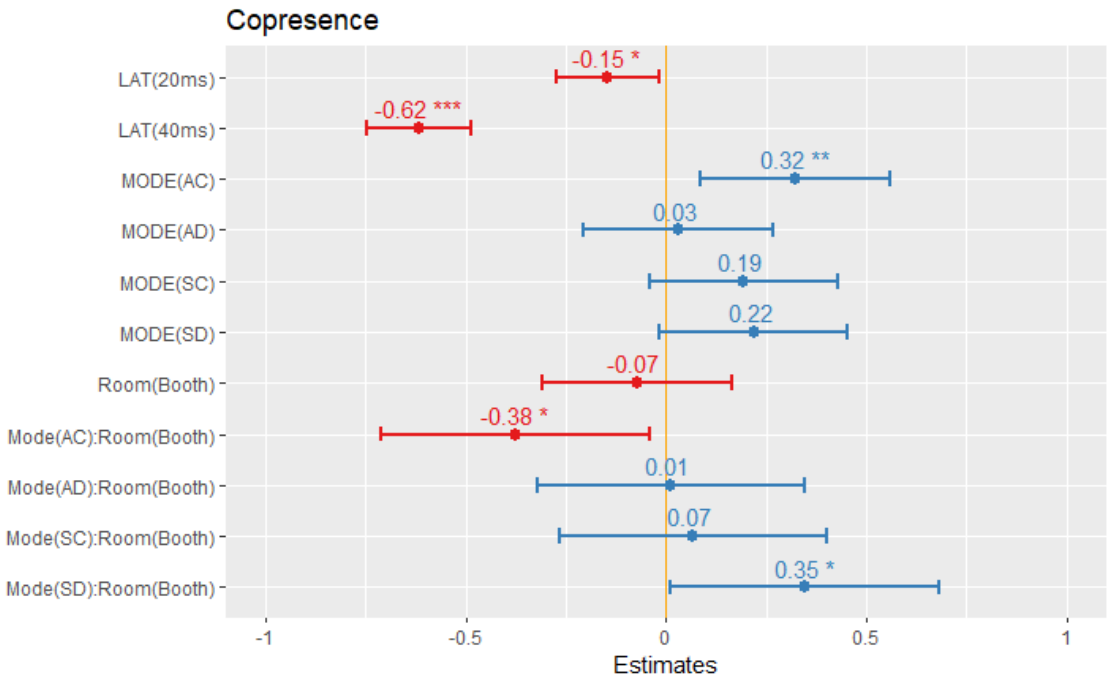


Figure 66: Standardized effect estimates for the independent variables, each estimate is rated against its relative reference group

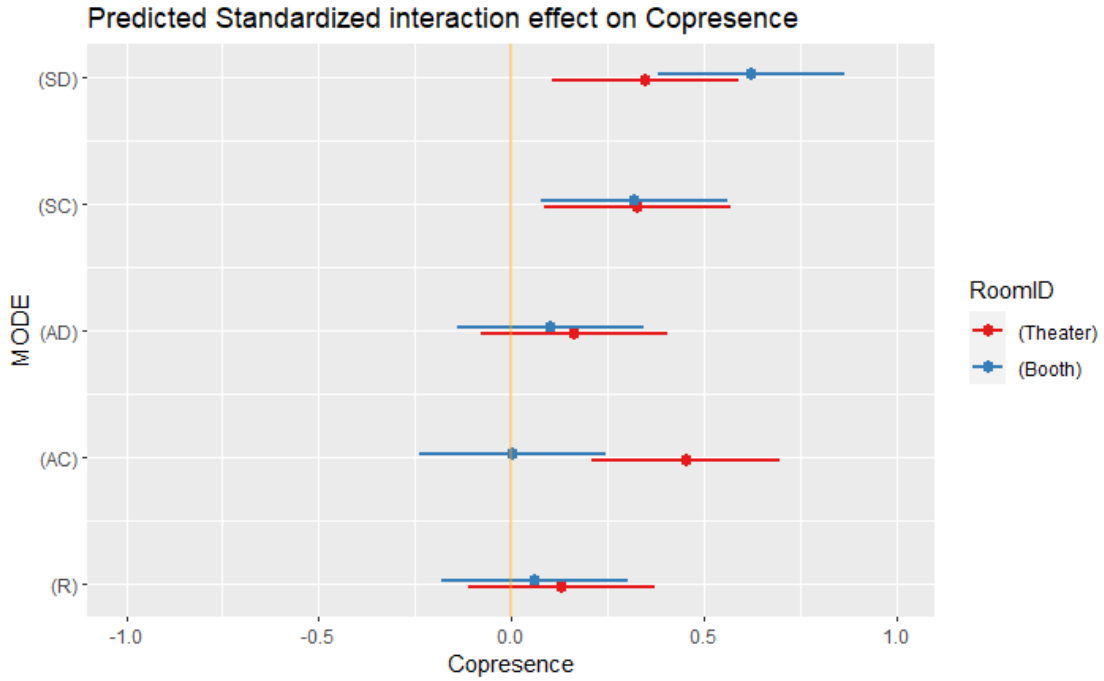


Figure 67: Standardized interactions effects visualized for the combinations of MODE with each node's room

Copresence: Multiple comparisons test

contrast	estimate	SE	df	t.ratio	p.value
Latency					
(7ms) - (20ms)	0.250	0.114	939	2.191	0.0733
(7ms) - (40ms)	1.059	0.114	939	9.282	<.0001***
(20ms) - (40ms)	0.809	0.114	939	7.092	<.0001***
Results are averaged over the levels of: MODE, RoomID					
P value adjustment: tukey method for 3 tests					
MODE / RoomID = (Theater)					
(AC) - (R)	0.552	0.208	939.13	2.650	0.0324*
(AD) - (R)	0.052	0.208	939.13	0.250	0.9985
(SC) - (R)	0.333	0.208	939.13	1.600	0.3726
(SD) - (R)	0.375	0.208	939.13	1.800	0.2591
MODE / RoomID = (Booth)					
(AC) - (R)	-0.094	0.208	939.13	-0.450	0.9855
(AD) - (R)	0.073	0.208	939.13	0.350	0.9944
(SC) - (R)	0.448	0.208	939.13	2.150	0.1214
(SD) - (R)	0.969	0.208	939.13	4.649	<.0001***
Results are averaged over the levels of: LAT					
P value adjustment: sidak method for 4 tests					

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 18: Results of the post hoc multiple comparison test, pairwise contrasts for LAT and contrast vs control for MODE

trend as the latency increases from the base level. As for the table, the room-mode interaction plot shows that in the Theater room, the (AC) mode performed about ~ 0.55 points higher ($\sim 9\%$ improvement) than the raw condition, while in the Booth room, the (SD) condition presents about ~ 1 point difference ($\sim 17\%$ improvement). Other conditions were not significantly different from the control, but none showed a degrading effect.

3.3 Summary of Results

For the rest of the response variables observed, results are summarized by showing the results of Wald's Chi-sq omnibus analysis of deviance test (or LRT test if interactions are present) and the pairwise comparisons for the best model concerning every variable. This section shows the results for the questionnaire responses, *Auditory Cohesion*, *Perceived Accuracy*, *Perceived Difficulty*, and *Immersion* (Immersion taken as a composite score of the agreement questions of Q1. Distribution plots and other relevant plots concerning each response are found in the appendix. Variables for which no better fit than the null model was found are omitted.

Results are here summarized per layer of category. Each table shows the best explanatory model using the previously shown procedure, summarized through an omnibus test of independent significance and through pairwise multiple comparison tests (with adjusted p-values) among the variables that were identified as the most explanatory, showing the estimated effect in relation to the control level of each categorical predictor. In the case of interactions, the pairwise comparisons are shown grouped by one of the interacting levels (in the case of interesting results, the interaction results are portrayed from both hierarchical directions). For continuous variables, the pairwise test is done between the first and last levels of the scale. Significant p-values are highlighted and marked with asterisks.

3.3.1 Trial Questionnaire

The other investigated response variables for the trial questionnaire involved "Auditory Cohesion", "Perceived Accuracy", "Perceived Difficulty" and "Immersion Score". The latter measure is a combination of the Likert-scale agreement questions posed in questionnaire Q2 (Ch. VI, Sect. 3). Results show how the impact of latency was the strongest factor in all regression models, with the highest significance level always showing between the 20ms-to-40ms

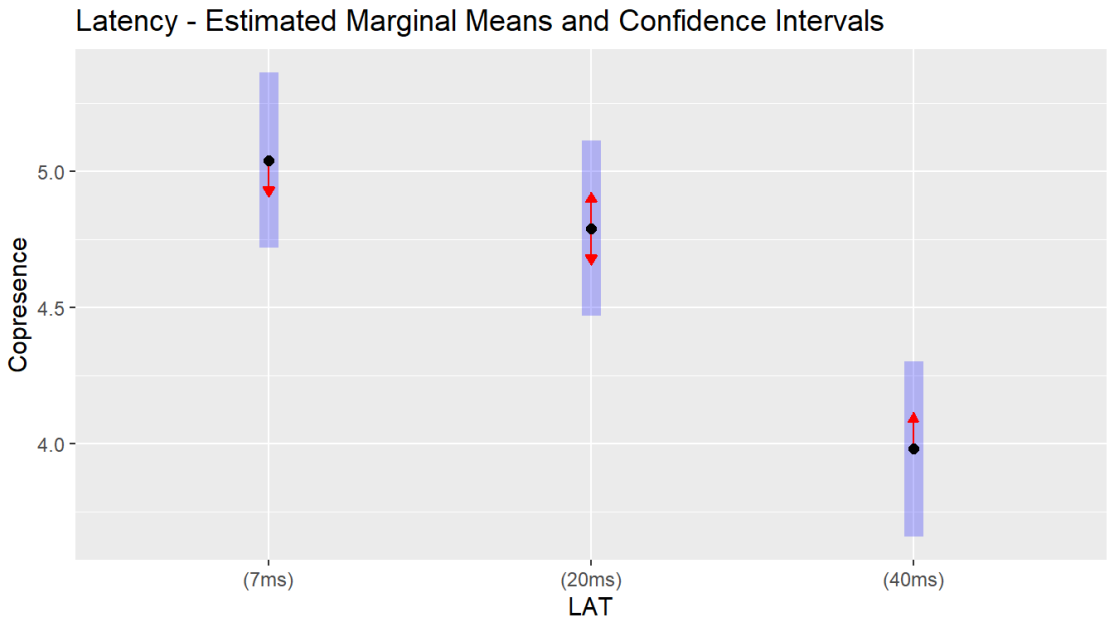


Figure 68: Estimated marginal means for Latency as predicted by the final model

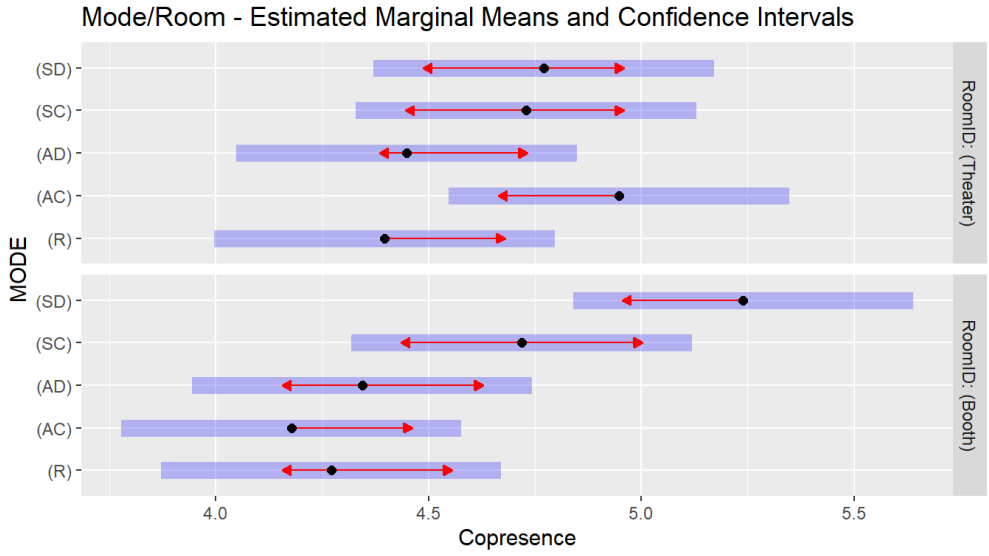


Figure 69: Comparisons between the different auralization modes grouped by room, showing confidence intervals. Comparisons are evaluated against the control group (R). If the red arrows are non-overlapping, the contrast is significant.

comparison. Besides the *immersion* response, the random intercept was highly significant in all cases, indicating high variability across subjects.

For *cohesion* (Tab. 19), the highest model fit included the group variable *isSymmetric* signifying a pooled group of three levels, "R", "AC/AD" (asymmetric) and "SC/SD" (symmetric). The pooled symmetric modes showed in average a significant improvement in the ratings when the auralization mode was symmetric rather than asymmetric or raw (the negative estimates are in the direction of the treatment so they need to be read as the difference of the first category towards the second one) consisting in about half a point of improvement on the 7-point rating scale ($\sim 8\%$) improvement, while the asymmetric treatments did not show a statistical difference from the raw condition and regardless of latency level. Latency is shown to bring the ratings down, with high-latency decreasing the average rating by $\sim 15\%$ from the base latency level on the response scale. Figure 70 shows the standardized effect sizes for latency and symmetry in regards to their associated reference level, Fig. 71, further zooms in on the data distributions grouped by Symmetry of the modes.

The results for perceived *accuracy* (Tab. 20) and *difficulty* (Tab. 21) aim to capture the factors that influenced the participant's self ratings of their own performance experience (in the attempt to understand if some modes conduct towards a facilitated musical performance). Once again, evidence points to Latency being the most influential factor in determining the degradation of the accuracy rating or the improvement of the difficulty rating (higher difficulty means a more difficult performance environment). However, an influence of the trial number (an indicator of time) was found, suggesting that through repetitions, the performance was perceived as more accurate or less difficult (as naturally expected through rehearsal). For *accuracy*, an interaction of trial number and latency was found, showing that the difference between low and high latency accuracy was perceived as diminishing over time. According to the model's estimates, through time, the difference between low and highest latency is expected to reduce from a 1.9 points average decrease ($\sim 31\%$) of perceived accuracy to 0.92 points ($\sim 15\%$). This trend was highly significant for the highest latency level and less for the lower latency levels, indicating that the effect of time is more about the reduction of performance sensitivity to latency than general improvement (Fig. 72). A portion of the model explanatory power was also found to rely on the base level of NMP-performance experience rated by musicians in questionnaire Q1. The question

Best model for: **Cohesion** $\sim LAT + Symmetric + (1|SubjID)$

Type III LRT Chi-sq Test					
	Chisq	**Df**	**Pr(>Chisq)**		
(Intercept)	616.209	1	**< 2.2e-16** ***		
LAT	63.523	2	**1.607e-14** ***		
Symmetric	27.210	2	**1.235e-06** ***		
Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
Latency					
(7ms) - (20ms)	0.278	0.114	932	2.442	**0.0393***
(7ms) - (40ms)	0.887	0.114	932	7.791	**<.0001*****
(20ms) - (40ms)	0.609	0.114	932	5.350	**<.0001*****
Results are averaged over the levels of: Symmetric					
Symmetric *levels* –>[0 = (AC/AD), 1=(SC/SD)]					
(R) - (0)	-0.107	0.127	932	-0.838	0.6792
(R) - (1)	-0.560	0.127	932	-4.396	**<.0001*****
(0) - (1)	-0.453	0.104	932	-4.358	**<.0001*****
Results are averaged over the levels of: LAT					

Original response scale: 1-7

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 19: Results of the omnibus test and post hoc multiple comparison test for “Auditory Cohesion”, pairwise contrasts for LAT and “Symmetric”. The “Symmetric” variable pools modes as: (SC/SD) vs (AC/AD) vs (R)

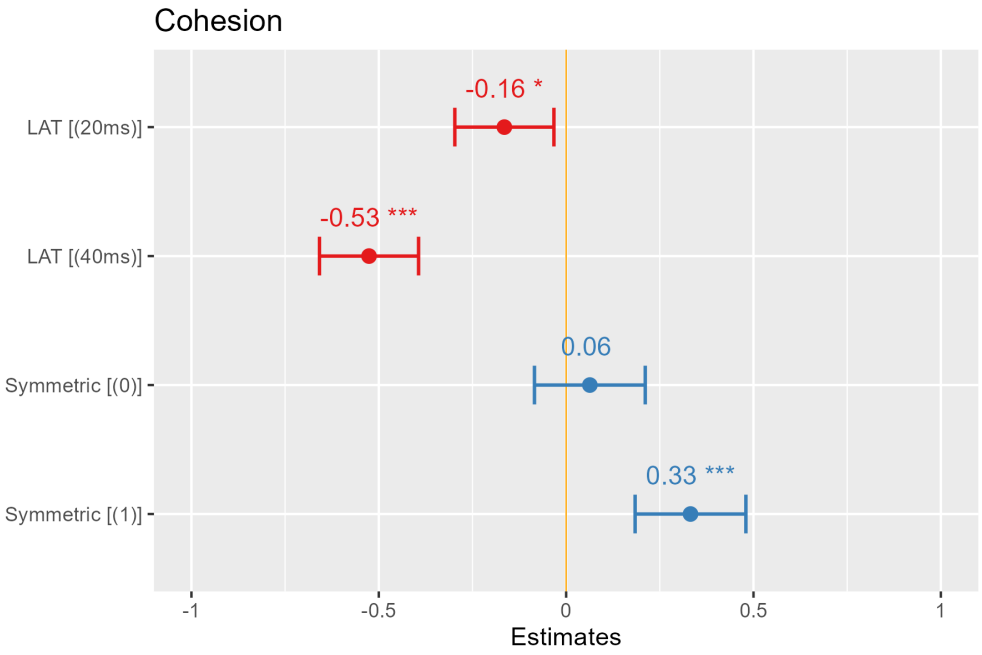


Figure 70: Standardized effect estimates for *Cohesion*

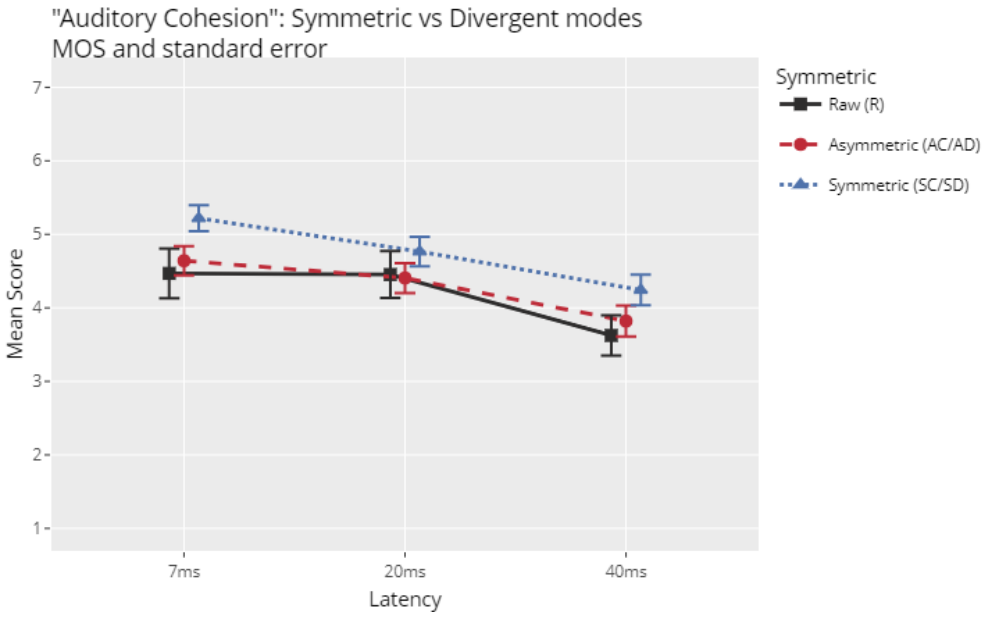


Figure 71: Standardized effect estimates for *Cohesion*

was used to divide the pool of subjects into a "lower experience" group vs "higher experience" group (median-split) showing that being more experience, led to an average higher rating of 0.75 points on the self-accuracy rating scale ($\sim 12.5\%$ improvement).

For the *difficulty* ratings, an interaction between the auralization mode and latency level was found meaning that difficulty was rated individually different per combination of mode and latency. However, most of the statistical weight is in this case carried by the latency differences rather than the MODE differences. The interaction results show that compared to the Raw mode, difficulty was in average significantly rated higher only for the AD mode 73 at the mid latency levels (difficulty increasing by about a full point in the rating scale, $\sim 16\%$ increase), while the other modes responded similarly to the increase of difficulty dictated by latency. This result is more of a comment to the specific AD mode than auralization in general. The factor of time was found to be independently significant in reducing the perceived grade of difficulty (another sign of adaptive behavior), with the difficulty between start and end of the experiment predicted to decrease by 0.66 points on the rating scale (- $\sim 11\%$ change).

Finally, the *immersion score* results are portrayed (Tab. 22). The score is taken as the sum of the five agreement questions of questionnaire Q2 (Q2.5 to Q2.9) and then standardized to interpret it as a distribution centered on its mean. Therefore, the interpretation of this table is not on a 1-7 scale (more interpretable for the other variables) but is in bidirectional units of standard deviations. To assess the reliability of the composite score, *Cronbach's alpha* was calculated over the set of standardized individual agreement-question responses, and found to be $\alpha = 0.875$ which indicates high internal consistency and reliability of the questionnaire scales (Lance et al. 2006).

In addition to the latency factor, the auralization mode and trial number (time) were independently significant as shown in the omnibus results (Fig. 74). The post hoc pairwise test shows that immersion degraded significantly with high latency (by 0.57 standard deviation units). The SD and SC modes were on average significantly better rated than the raw control condition, regardless of latency level. The AD mode was the only condition showing a negative trend of immersion compared to the raw condition, although this was not found to be statistically significant (Fig. 75). The trial was indicated as significantly improving the immersion scores through time overall conditions by an average of 0.35 standard deviations, indicating that

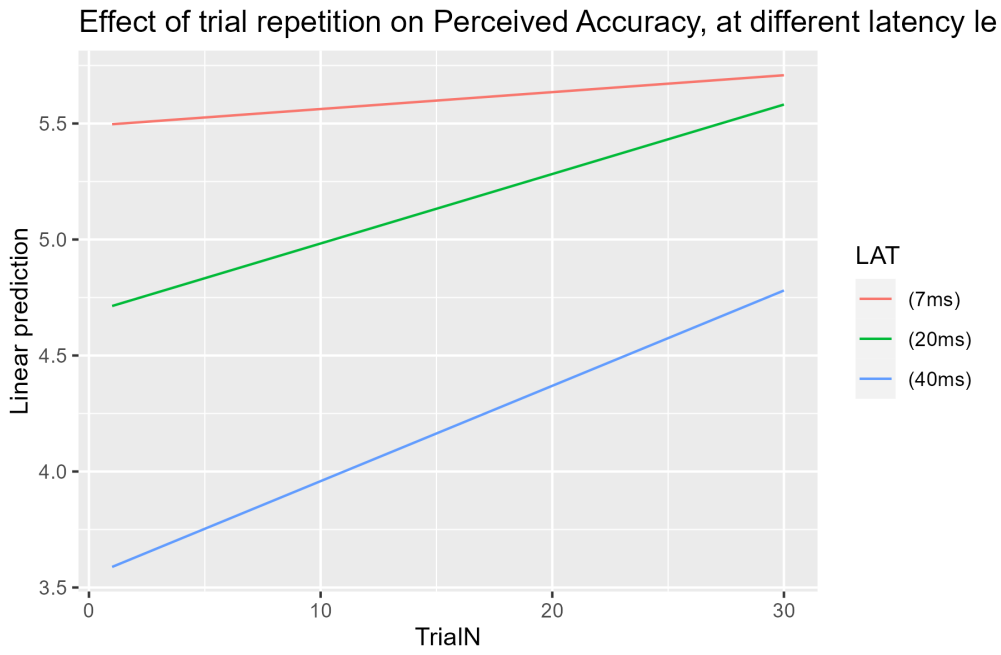


Figure 72: *Perceived Accuracy:* Effect of time over different latency levels (model prediction)

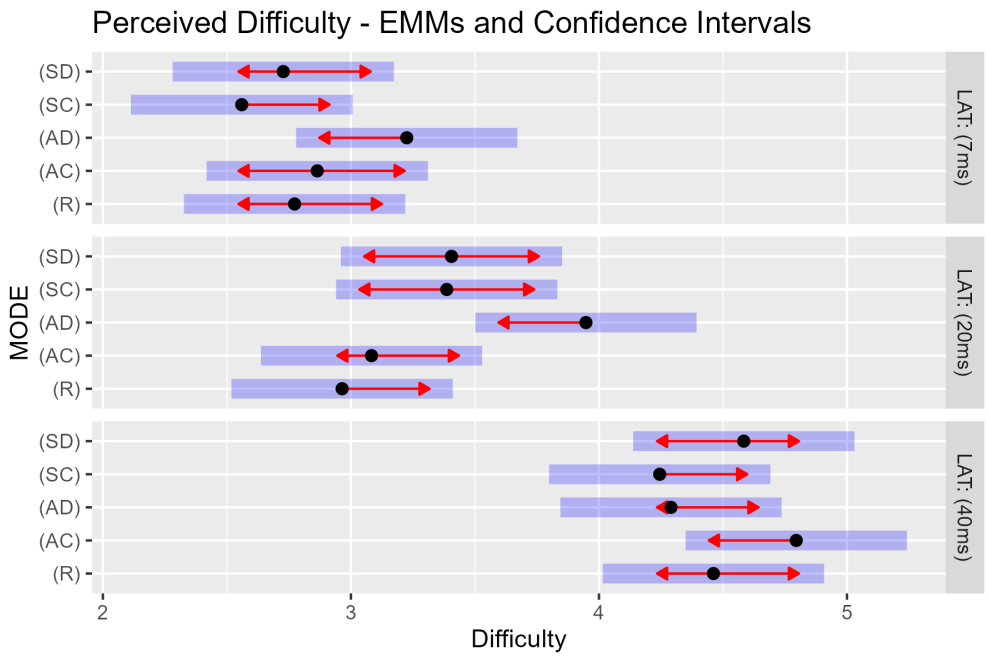


Figure 73: *Perceived Difficulty:* Estimated marginal means and confidence intervals for interactions between MODE and Latency levels.

Best model for: **Perceived Accuracy** $\sim (LAT * TrialN) + EXP_{\text{NMP}} + (1|SubjID)$

Type III LRT Chi-sq Test					
	Chisq	**Df**	**Pr(>Chisq)**		
(Intercept)	503.26	1	**< 2.2e-16** ***		
LAT	74.48	2	**< 2.2e-16** ***		
TrialN	0.60	1	0.4386		
ExpScoreGroup	10.37	1	**0.0013****		
LAT:TrialN	7.10	2	**0.0288***		
Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
Latency / Trial = 1					
(7ms) - (20ms)	0.783	0.220	934	3.564	**0.0011*****
(7ms) - (40ms)	1.908	0.215	935	8.857	**<.0001*****
(20ms) - (40ms)	1.125	0.216	934	5.196	**0001*****
Latency / Trial = 30					
(7ms) - (20ms)	0.126	0.222	934	0.570	0.8360
(7ms) - (40ms)	0.928	0.218	935	4.258	**0.0001*****
(20ms) - (40ms)	0.801	0.213	934	3.767	**0.0005*****
Results are averaged over the levels of: ExpScoreGroup					
Trial / Latency = 7ms					
TrialN30 - TrialN1	0.211	0.273	934	0.774	0.4391
Trial / Latency = 20ms					
TrialN30 - TrialN1	0.868	0.267	934	3.258	**0.0012****
Trial / Latency = 40ms					
TrialN30 - TrialN1	1.192	0.252	934	4.721	**<.0001*****
Results are averaged over the levels of: ExpScoreGroup					
NMP EXP *levels* –>[(<50%) = (“lower half”), (>50%)=(“upper half”)]					
(<50%) - (>50%)	-0.754	0.234	34.1	-3.220	**0.0028****
Results are averaged over the levels of: LAT					
Original response scale: 1-7					
Degrees-of-freedom method: kenward-roger					
***$p < 0.001$; **$p < 0.01$; *$p < 0.05$					

Table 20: Results of the post hoc multiple comparison test for “Perceived Accuracy”, pairwise contrasts for LAT and TrialN interaction, and NMP Experience Group (indicating if the subject was part of the lower or upper half of the percentile groups). Regardless of auralization mode, the results suggest that participant were able to slowly adapt to higher latencies

Best model for: **Perceived Difficulty** $\sim (LAT * MODE) + TrialN + (1|SubjID)$

Type III LRT Chi-sq Test				
	Chisq	**Df**	**Pr(>Chisq)**	
(Intercept)	168.1905	1	**< 2.2e-16**	***
LAT	51.2731	2	**7.348e-12**	***
MODE	7.3024	4	0.1207	
TrialN	17.1117	1	**3.524e-05**	***
LAT:MODE	19.9315	8	**0.0106**	*

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
MODE / Latency = 7ms					
(AC) - (R)	0.0915	0.258	943	0.354	0.9941
(AD) - (R)	0.4524	0.258	943	1.752	0.2839
(SC) - (R)	-0.2129	0.259	943	-0.823	0.8792
(SD) - (R)	-0.0455	0.258	943	-0.176	0.9996
MODE / Latency = 20ms					
(AC) - (R)	0.1186	0.258	943	0.459	0.9843
(AD) - (R)	0.9830	0.258	943	3.807	**0.0006*** **
(SC) - (R)	0.4219	0.258	943	1.634	0.3514
(SD) - (R)	0.4411	0.258	943	1.708	0.3079
MODE / Latency = 40ms					
(AC) - (R)	0.3339	0.258	943	1.293	0.5829
(AD) - (R)	-0.1711	0.258	943	-0.663	0.9412
(SC) - (R)	-0.2166	0.258	943	-0.839	0.8719
(SD) - (R)	0.1223	0.258	943	0.473	0.9825
TrialN *levels* –>[(1) = (“first trial”), (30)=(“last trial”)]					
TrialN30 - TrialN1	-0.657	0.159	943	-4.137	**<.0001*** **
Results are averaged over the levels of: LAT, MODE					

Original response scale: 1-7

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 21: Results of the post hoc multiple comparison test for “Perceived Difficulty”, pairwise contrasts for interactions of LAT and MODE, and Trial number. High latency proved significant across all contrasts, while mid-latency was not significant for all modes. Mode was significant only between (AD) and (R) at the 20ms level.

Best model for: **Immersion Score** $\sim LAT + MODE + TrialN + (1|SubjID)$

Type III LRT Chi-sq Test				
	Chisq	**Df**	**Pr(>Chisq)**	
(Intercept)	0.976	1	0.3232	
LAT	83.680	2	**< 2.2e-16**	***
MODE	40.863	4	**2.869e-08**	***
TrialN	16.127	1	**5.923e-05**	***

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
LAT					
(7ms) - (20ms)	0.130	0.0648	935	2.001	0.1125
(7ms) - (40ms)	0.566	0.0648	935	8.730	**<.0001****
(20ms) - (40ms)	0.436	0.0648	935	6.731	**<.0001****
Results are averaged over the levels of: MODE					
MODE					
(AC) - (R)	0.1565	0.0836	935	1.872	0.2244
(AD) - (R)	-0.0329	0.0836	935	-0.394	0.9912
(SC) - (R)	0.2586	0.0837	935	3.091	**0.0082****
(SD) - (R)	0.4254	0.0836	935	5.086	**<.0001****
Results are averaged over the levels of: LAT					
TrialN *levels* ->[(1) = ("first trial"), (30)=("last trial")]					
TrialN30 - TrialN1	0.356	0.0887	935	4.016	**0.0001****
Results are averaged over the levels of: LAT, MODE					
Response scale is standardized					
Degrees-of-freedom method: kenward-roger					
***$p < 0.001$; **$p < 0.01$; *$p < 0.05$					

Table 22: Results of the post hoc multiple comparison test for the composite "Immersion Score", pairwise contrasts for LAT and MODE, and Trial number.

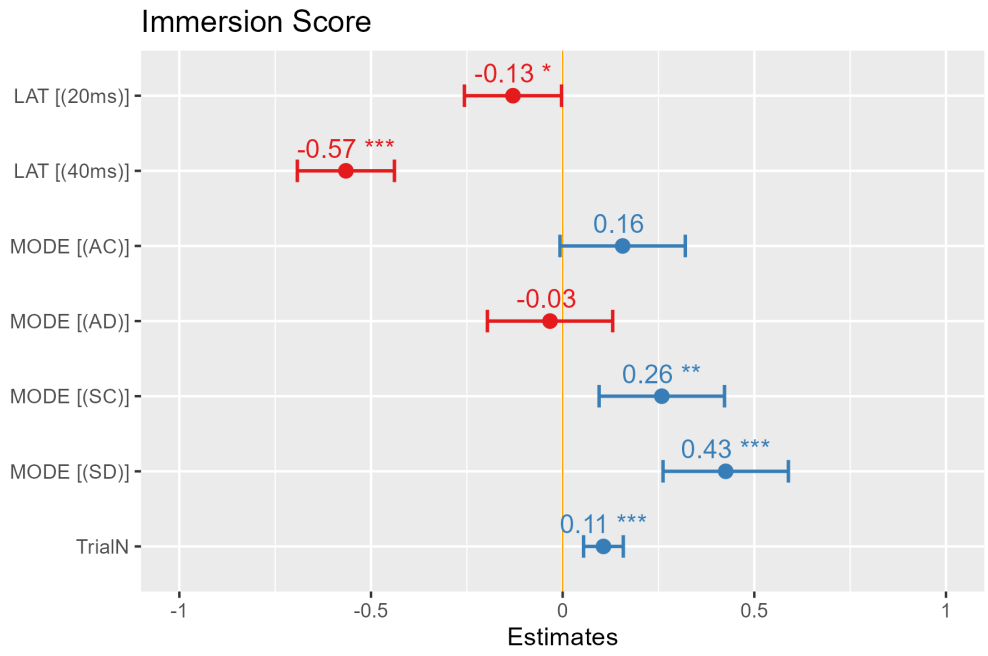


Figure 74: *Immersion Score:* Standardized effect estimates

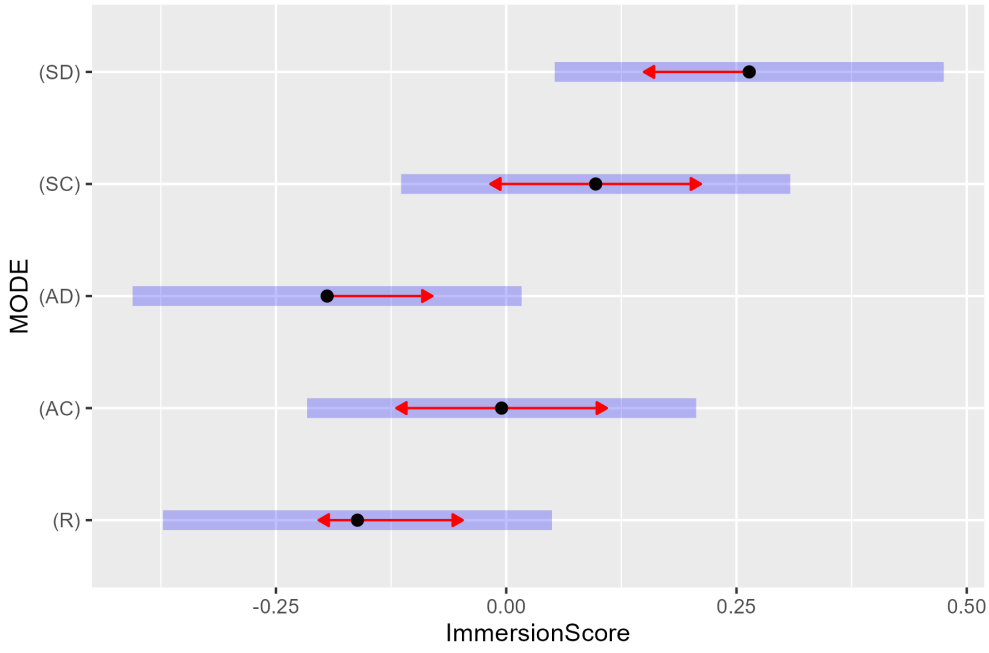


Figure 75: *Immersion Score:* Estimated marginal means and confidence intervals for MODE

immersion metrics (in the form of *copresence* and *cohesion* scales) are partially related to the time spent within using an immersive system.

3.3.2 Objective Metrics

The objective metrics were analyzed using a log ratio between the trial metric and baseline metric (see primary data collection in Ch. VI, sect. 2.2. This was because tracking the absolute level of the tempo and beat metrics was not the focus of the work, rather the relative level of change of performance metric in relation to a subject's base ability was of interest. In regards to the pairwise estimations, a positive pairwise estimate difference, between two levels of a factor, means that the observed level has a larger difference from the baseline compared to the reference level, while a negative estimate difference means that response gets closer to the baseline level. Instead, the effect sizes (and marginal means) need to be read as a multiplier factor of change from the baseline across levels (e.g. a positive unit means that a factor level leads to the response being "x times larger than the baseline metric" while a negative unit would mean the response is "x times smaller than the baseline metric").

The first objective metric analyzed, "tempo range" regards the distance between the lowest and highest tempo in the dynamic tempo curves. A stabler performance would present a smaller range of tempo. The model that best fit the data, indicated an interaction between latency and auralization mode, and a small impact of time. Generally, higher levels of latency led to higher increases in range (meaning less stability in the performance), with the higher latency showing an average estimated effect of $\sim 20\%$ increase from the baseline (Fig. 77) and carrying most of the impact on the metric. At low and mid latency, the raw condition showed the lowest average tempo range (even outperforming the baseline at the 7ms latency level, showing a mean $\sim 10\%$ smaller). Instead, the AD mode showed a significantly higher metric, $\sim 5\%$ higher than baseline. This difference across modes reduced at the 20ms level, and it disappeared at 40ms, with the exception of the AC mode showing less degradation than all other modes (mildly significant). Trial number (time) lead to an average very small but slightly significant increase of tempo range of about a factor of 0.03 between start and end trial.

The *Pacing* metric (Tab. 23) represents the mean time-interval between beats (therefore a measure of average tempo). Latency was found to have an impact, with higher latencies showing a

Best model for: **Pacing (log ratio to baseline)** $\sim LAT + TrialN + (1|SubjID)$

Type III LRT Chi-sq Test					
	Chisq	**Df**	**Pr(>Chisq)**		
(Intercept)	1.7306	1	0.1883		
LAT	25.9007	2	**2.375e-06** ***		
TrialN	22.2170	1	**2.435e-06** ***		
Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
Latency					
LAT7 - LAT20	-0.010575	0.00237	750	-4.457	**<.0001*** **
LAT7 - LAT40	-0.010686	0.00242	750	-4.423	**<.0001*** **
LAT20 - LAT40	-0.000111	0.00234	750	-0.048	0.9988
TrialN *levels* –>[(1) = (“first trial”), (30)=(“last trial”)]					
TrialN30 - TrialN1	0.0153	0.00324	749	4.713	**<.0001*** **
Results are averaged over the levels of: LAT					

Scale is the log ratio between response and pair’s baseline

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 23: Results of the objective tempo slope trends, estimate scale is a log-ratio between objective metrics of the trial performance and the baseline level measured at the primary-data baseline stage, for each pair. In the case of pacing, a higher value means a higher beat-interval and therefore a slower tempo compared to the baseline level. Latency had an effect, as well as repetition

larger difference than lower latencies in regards to the baseline levels (no difference between mid and high latency). Trial also had an impact, showing time leading to a higher change of pacing from the baseline (potentially a sign of playing fatigue). Effect sizes remain small, with about $\sim 1-2\%$ increase in mean pacing (Fig. 78). Auralization was shown not to have an impact on this measure.

The model for determining *"Mean Lag"* (mean time difference between the player's beat onsets) show that an interaction between latency and part ("static" vs "shifting") showing that the lag level changed the most from the shifting player's perspective than for the static player, more so as the latency increased (Tab. 25. The indication being that the variation of beat stability depends on the interaction of complexity of the musical material and on latency. Instead the simpler "static" beat was not significantly affected across latencies. Fig. 79 shows a summary of both the pairwise differences and the effect size towards the baseline (zero on the x-axis), showing that the mean lag experienced at the Shifting node augmented with latency, with an effect size around $\sim 7-8\%$.

3.3.3 Annotations and Ratings

The next layer of analysis regarded the annotations and ratings of the expert-listeners tasked to rate the recordings of the distributed experiment. This section covers the measures of *overall rating* (scale 1-10), and the results of the annotations of perceived *pattern inaccuracies* and perceived *tempo inaccuracies*. Results in regards to the annotations were computed through GLMMs for binomial distributions and are presented as log odds-ratios, indicating the change in probability of the performance inaccuracy being present from being absent. Since the annotations and ratings were performed over mixes of the complete performances, the random effect related to the pair ID rather than subject ID. Auralization effects were not present in the material given to annotators, but still investigated in the analysis in order to observe potential latent effects on performance.

Tab. 26 shows that the best model in regards to the overall tempo was fitted using Latency and Trial number. High latency was highly impactful in decreasing the rating, with the highest level being rated on average 9.2 points (on a scale 1-10) $\sim$ 10% lower than the low-latency trials. Trial number also had a positive effect on the ratings, showing that an average increase

Best model for: **Tempo Range** $\sim (LAT * MODE) + TrialN + (1|SubjID)$

Type III LRT Chi-sq Test				
	Chisq	**Df**	**Pr(>Chisq)**	
(Intercept)	13.5368	1	**0.0002339**	***
LAT	102.4079	2	**< 2.2e-16**	***
MODE	31.3455	4	**2.603e-06**	***
TrialN	4.3599	1	**0.0367946**	*
LAT:MODE	21.7019	8	**0.0054990**	**

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
MODE / Latency = 7ms					
(AC) - (R)	0.065833	0.0281	762	2.341	0.0757
(AD) - (R)	0.148125	0.0281	762	5.264	**<.0001*** **
(SC) - (R)	0.088033	0.0278	762	3.165	**0.0064** **
(SD) - (R)	0.113354	0.0276	762	4.102	**0.0002*** **
MODE / Latency = 20ms					
(AC) - (R)	0.029426	0.0262	761	1.123	0.7033
(AD) - (R)	0.068188	0.0266	762	2.565	**0.0413***
(SC) - (R)	0.053832	0.0262	762	2.053	0.1521
(SD) - (R)	0.059587	0.0262	762	2.272	0.0903
MODE / Latency = 40ms					
(AC) - (R)	-0.078371	0.0275	762	-2.849	**0.0179***
(AD) - (R)	0.000791	0.0262	762	0.030	1.0000
(SC) - (R)	-0.003897	0.0262	762	-0.148	0.9998
(SD) - (R)	-0.010468	0.0267	762	-0.392	0.9913
TrialN *levels* –>[(1) = (“first trial”), (30)=(“last trial”)]					
TrialN30 - TrialN1	0.0346	0.0166	762	2.088	**0.0371***
Results are averaged over the levels of: LAT, MODE					

Scale is the log ratio between response and pair’s baseline

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 24: Results of the objective tempo range calculated from the dynamic tempo-curve, estimate scale is a log-ratio between objective metrics of the trial performance and the baseline level measured at the primary-data baseline stage, for each pair

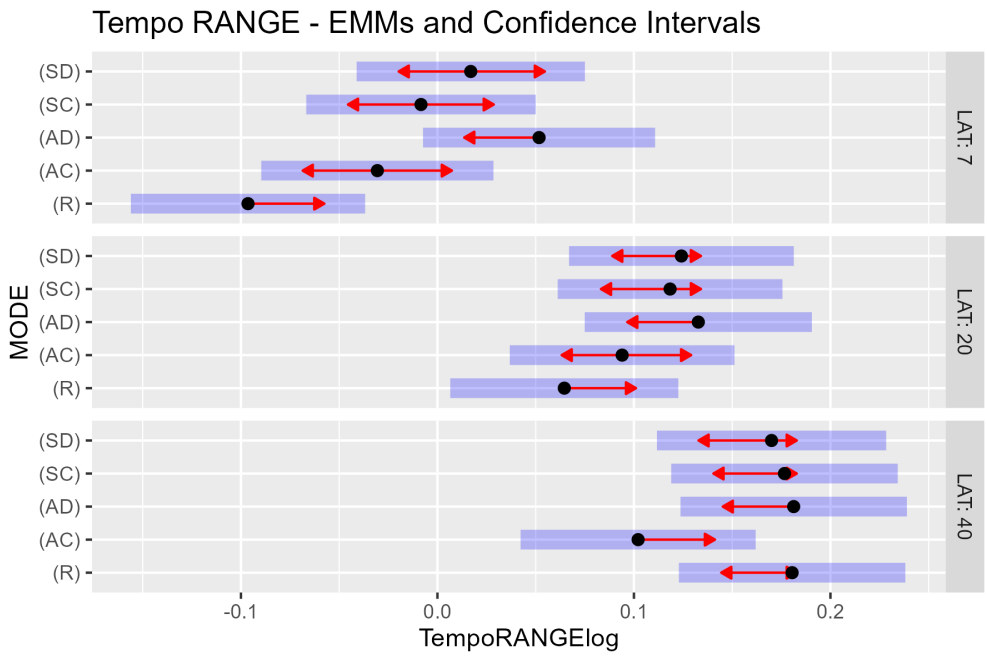


Figure 76: *Tempo Range:* Estimated marginal means across modes, grouped by latency

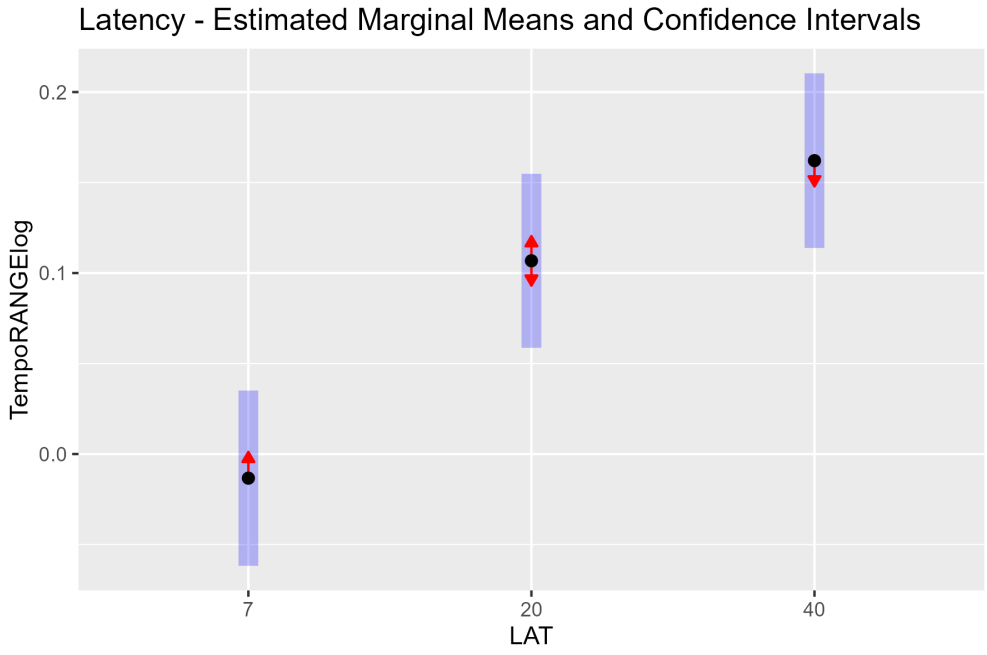


Figure 77: *Tempo Range:* Estimated marginal means across latency levels

Best model for: **Mean Lag (log ratio to baseline)** $\sim LAT * Part + (1|SubjID)$

Type III LRT Chi-sq Test				
	Chisq	**Df**	**Pr(>Chisq)**	
(Intercept)	2.7033	1	0.100141	
LAT	32.1031	2	**1.069e-07** ***	
Part	0.1065	1	0.744183	
LAT:Part	10.1933	2	**0.006117** **	

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
Latency / Part = *Static*					
LAT7 - LAT20	-0.0465	0.0257	753	-1.808	0.1676
LAT7 - LAT40	-0.0155	0.0262	758	-0.593	0.8241
LAT20 - LAT40	0.0310	0.0248	751	1.247	0.4261
Latency / Part = *Shifting*					
LAT7 - LAT20	-0.1205	0.0256	751	-4.708	**<.0001*** **
LAT7 - LAT40	-0.1319	0.0259	753	-5.084	**<.0001*** **
LAT20 - LAT40	-0.0114	0.0258	754	-0.441	0.8986
Part / Latency = (7ms)					
Static - Shifting	-0.0124	0.0382	73.4	-0.323	0.7474
Part / Latency = (20ms)					
Static - Shifting	-0.0864	0.0372	66.1	-2.324	**0.0232***
Part / Latency = (40ms)					
Static - Shifting	-0.1287	0.0377	69.5	-3.418	**0.0011****

Scale is the log ratio between response and pair's baseline

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 25: Results of the objective mean beat lag (measure of beat asynchrony), estimate scale is a log-ratio between objective metrics of the trial performance and the baseline level measured at the primary-data baseline stage, for each pair. We can observe that for the shifting player, latency had much more of a detrimental effect (increase in mean lag) than for the static player

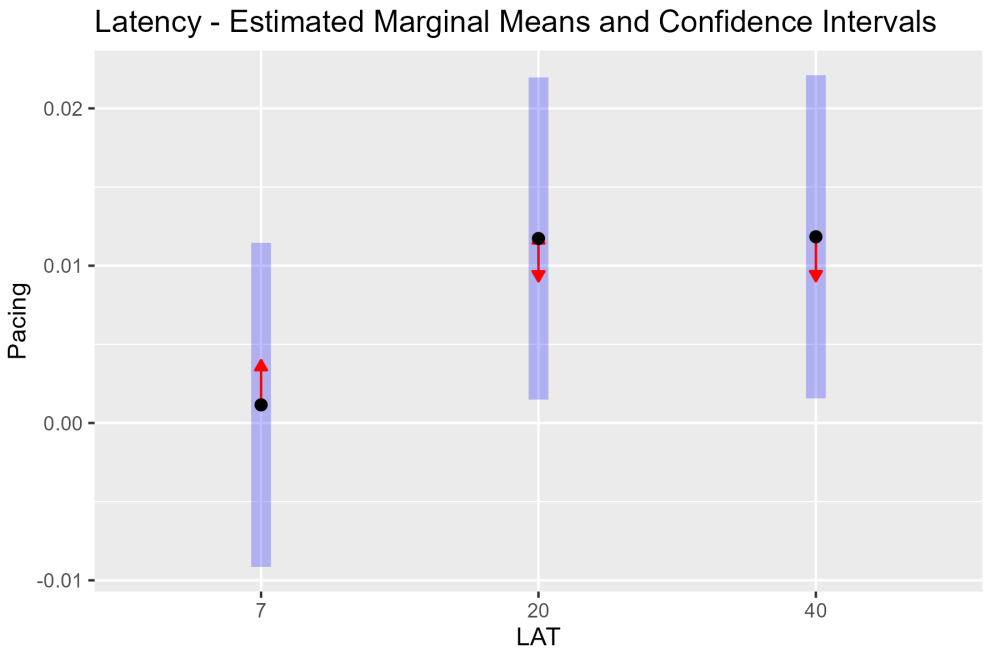


Figure 78: *Pacing:* Estimated marginal means across latencies

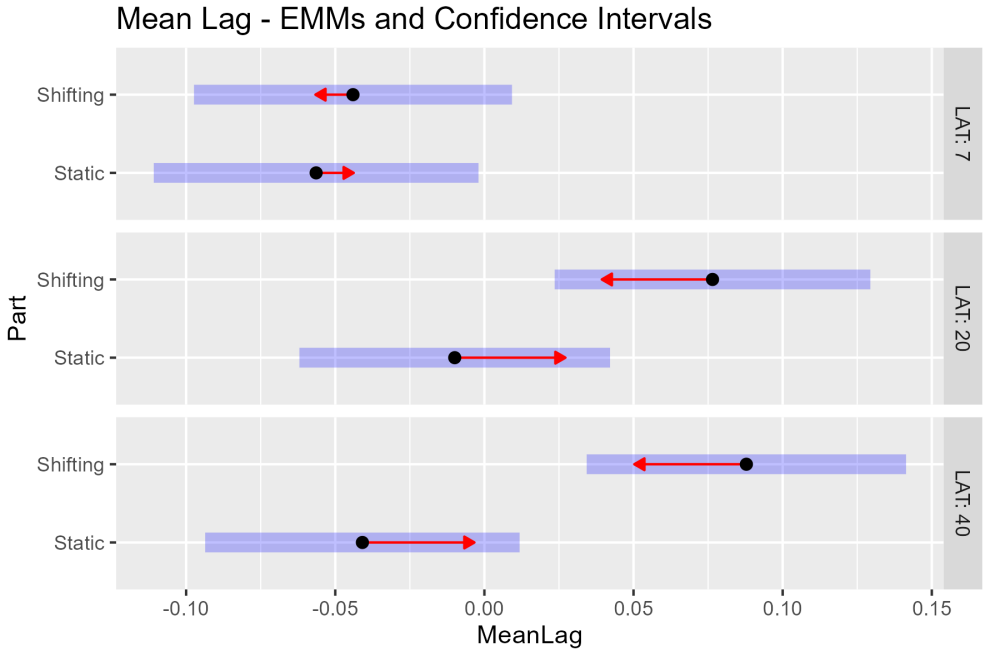


Figure 79: *Mean Lag:* Estimated marginal means showing interactions across latencies and parts.

of 0.42 points occurred for performances that were executed later in time during the experiment (suggesting that pairs got better over time at performing). The evaluations were done in a random order, different from that of performance, indicating that the detected difference was indeed due to the performance change over time rather than the evaluator's own ratings changing over time. As seen in Fig., the effect of time was quantified as an increase of 0.013 points of rating per unit of time (trial repetition).

Best model for: **Overall Rating (ext.)** $\sim LAT + TrialN + (1|PairID)$

Type III LRT Chi-sq Test			
	Chisq	**Df**	**Pr(>Chisq)**
(Intercept)	2.3432	1	0.1258
LAT	215.5275	2	**< 2.2e-16** ***
TrialN	21.1807	1	**4.179e-06** ***
TrialN coeff estimate (response scale) = 0.013			

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
Latency					
LAT7 - LAT20	0.200	0.0671	853	2.985	**0.0082****
LAT7 - LAT40	0.927	0.0665	853	13.951	**<.0001*****
LAT20 - LAT40	0.727	0.0669	854	10.869	**<.0001*****
TrialN *levels* –>[(1) = ("first trial"), (30)=("last trial")]					
TrialN30 - TrialN1	0.419	0.0911	853	4.602	**<.0001*****
Results are averaged over the levels of: LAT					
Original response scale: 1-10					
Degrees-of-freedom method: kenward-roger					
***$p < 0.001$; **$p < 0.01$; *$p < 0.05$					

Table 26: Results of the overall rating by external listeners. Ratings are standardized per evaluator.

Similarly, the detection of pattern inaccuracies (e.g. missed beats, evident extra claps) was statistically more likely to happen for higher latencies than lower latencies, with the high-level latency producing an increased chance of error detection being 0.56 times more likely than the low level (Tab. 27). The difference between the low and mid latency is hard to interpret due to its contradictory trend, but the estimate is only significant at the < 0.05 level so it may represent statistical noise. Trial number had a strong impact on the odds of mistakes being present, with

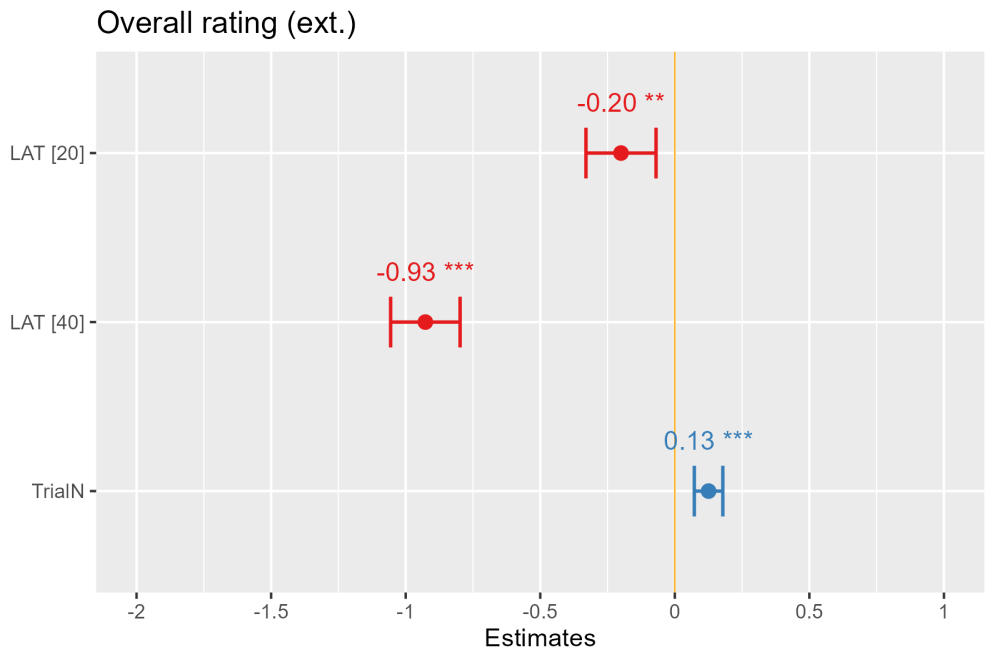


Figure 80: *Overall ratings*: Estimated effect sizes

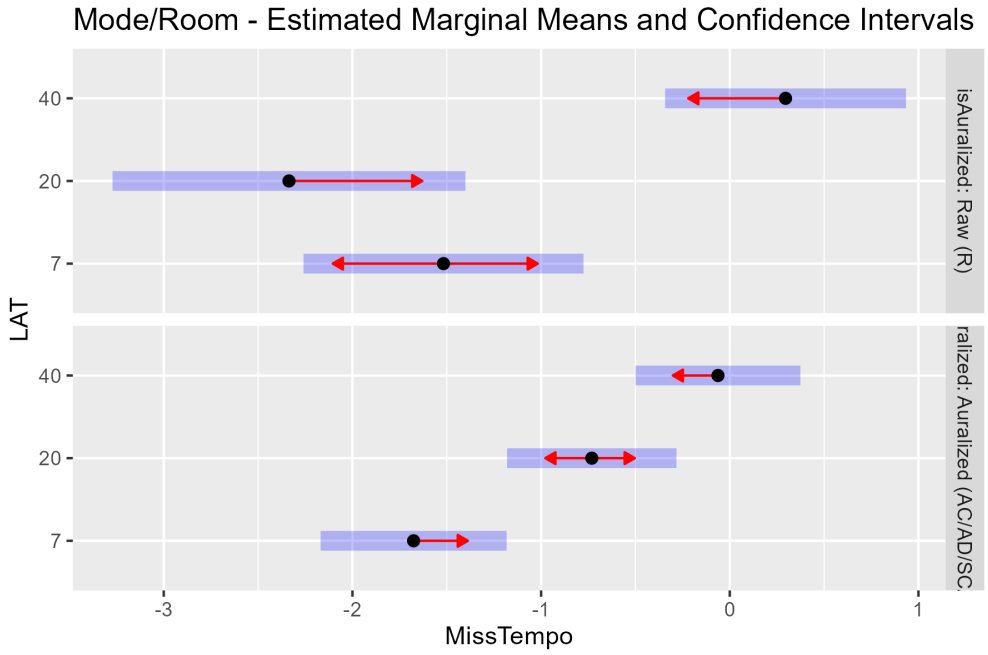


Figure 81: *Tempo Inaccuracies*: Marginal means over latency, grouped by auralization group

Best model for: **Pattern Inaccuracies** $\sim LAT + TrialN + (1|PairID)$

Type II Wald's Chi-sq Test			
	Chisq	**Df**	**Pr(>Chisq)**
(Intercept)	4.9678	1	**0.025823** *
LAT	27.4007	2	**1.122e-06** ***
TrialN	7.3229	1	**0.006808** **
TrialN coeff estimate (log odds) = -0.024			

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
Latency					
LAT7 - LAT20	0.468	0.193	Inf	2.429	0.0402*
LAT7 - LAT40	-0.566	0.194	Inf	-2.918	**0.0099****
LAT20 - LAT40	-1.034	0.198	Inf	-5.229	**<.0001*****
Results are given on the log odds ratio (not the response) scale.					
TrialN *levels* –>[(1) = ("first trial"), (30)=("last trial")]					
TrialN30 - TrialN1	-0.722	0.267	Inf	-2.706	**0.0068****
Results are averaged over the levels of: LAT					
Results are given on the log odds ratio (not the response) scale.					

Scale is in log-odds ratio

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 27: Results of the model used for assessing Pattern Inaccuracy responses (presence of mistakes in the musical beat pattern). Coefficients represent changes in the log odds of a pattern mistake probability to happen in response to changes to the independent variables.

Best model for: **Tempo Inaccuracies** $\sim LAT * isAuralized + (1|PairID)$

Type III Wald's Chi-sq Test			
	Chisq	**Df**	**Pr(>Chisq)**
(Intercept)	16.0675	1	**6.113e-05** ***
LAT	32.2963	2	**9.704e-08** ***
isAuralized	0.1769	1	**0.674067**
LAT:isAuralized	13.1825	2	**0.001372** **

Pairwise Comparison Tests					
contrast	**estimate**	**SE**	**df**	**t.ratio**	**p.value**
isAuralized *levels* –>[0 = (R), 1=(AC/AD/SC/SD)]					
LAT / isAuralized = 0					
LAT20 - LAT7	-0.819	0.553	Inf	-1.482	0.2577
LAT40 - LAT7	1.813	0.433	Inf	4.185	**0.0001*** **
LAT / isAuralized = 1					
LAT20 - LAT7	0.944	0.228	Inf	4.143	**0.0001*** **
LAT40 - LAT7	1.613	0.225	Inf	7.172	**<.0001*** **
Results are given on the log odds ratio (not the response) scale.					
isAuralized / LAT = 7					
Raw (R) - (Auralized)	0.158	0.377	Inf	0.421	0.6741
isAuralized / LAT = 20					
Raw (R) - (Auralized)	-1.605	0.464	Inf	-3.455	**0.0005*** **
isAuralized / LAT = 40					
Raw (R) - (Auralized)	0.358	0.306	Inf	1.171	0.2417
Results are given on the log odds ratio (not the response) scale.					

Scale is in log-odds ratio

Degrees-of-freedom method: kenward-roger

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 28: Results of the model used for assessing Tempo Inaccuracy responses (presence of strong perceived accelerations/decelerations). Coefficients represent changes in the log odds of the probability of a change in tempo to be perceived by a listener in response to changes to the independent variables.

trials occurring at the end of an experiment session being 0.722 times less likely to present a detectable mistake within it compared to trials occurring at the beginning (and effect size of $\sim 2.4\%$ decrease of chance per unit of time increase).

For the tempo-inaccuracy annotation (evident accelerations or decelerations in performance) the results shown in Tab. 28 depended on the interaction with latency and the presence of auralization (in practice all modes vs raw). The interaction shows that the combinations of high latency with every mode saw increases chances of tempo inaccuracies being heard, more so for the raw cases than the auralization cases, although this result needs to be looked at a bit skeptically since there are imbalances in the data distributions and contradictory trends at the mid level. High level of latencies increased the chances of error significantly in both auralization and raw cases.

3.4 Correlation Matrices

After having analyzed each response variable on its own, a correlation analysis looked for the existence of trend similarities across conditions. The correlation was performed within and between each layer of evaluation data. Only variables with a continuous response scale were fed to the correlation step, thus discarding the binomial annotations. Each variable standardized for equal range of distribution. For this phase, the original absolute-level objective metrics were used instead of the relative metrics. These were then standardized along the other metrics in the same manner. For the between-layers correlations, the same data polishing steps of removing entries with outliers and missing values from the individual analysis were applied, with the layer with less entries defining the total amount of selected entries. The correlations were computed with the *Pearson's r* correlation coefficient. Finally, correlation matrices are used for visualization. A full 3-layers correlation matrix is found in the appendix.

The within-layer correlations are shown in Figs. 82, 83 and 84. In the questionnaire layer, *copresence*, *cohesion* and *immersion* show high degree of positive correlation with each other, this aspect validates previous literature on the subject (and was expected from the way the agreement questions were formulated). A strong negative correlation was found between *accuracy* and *difficulty* as predictable. The most interesting result is the mild but substantial correlation between *accuracy* and the other ratings, suggesting that *immersion* and *presence* are subjectively

correlated to the impression of accurate performance, and negatively correlated to the perception of *difficulty*. For annotation/ratings layers, all variables were positively correlated with each other, with the overall rating mostly correlated to *beat precision* rather than *tempo*.

The within-layer correlation results of the objective metrics instead revealed that there is strong individuality of response across all metrics employed with little to no correlations found. Very reasonably, the *imprecision* (or Lag deviation) metric was highly correlated to the mean lag since they are similar metrics. The mean lag was mildly correlated to the pacing, suggesting that higher tempos were correlated to higher deviations in synchronization.

Figs. 85, 86 and 87 show the between-layers results. Although mild, interesting correlations were found across the participant's presence ratings and the third-person quality evaluation ratings (negative in the case of difficulty). Correlation was higher for perceived accuracy and difficulty than copresence and cohesion. In regards to correlations between objective and subjective ratings, the correlations were much milder or absent, with the exception of *tempo slope* and *tempo range* (lower numbers meaning a stabler performance, therefore a likely higher rating in the quality evaluations) being mildly correlated with *accuracy* and *difficulty* and some of the third-party ratings.

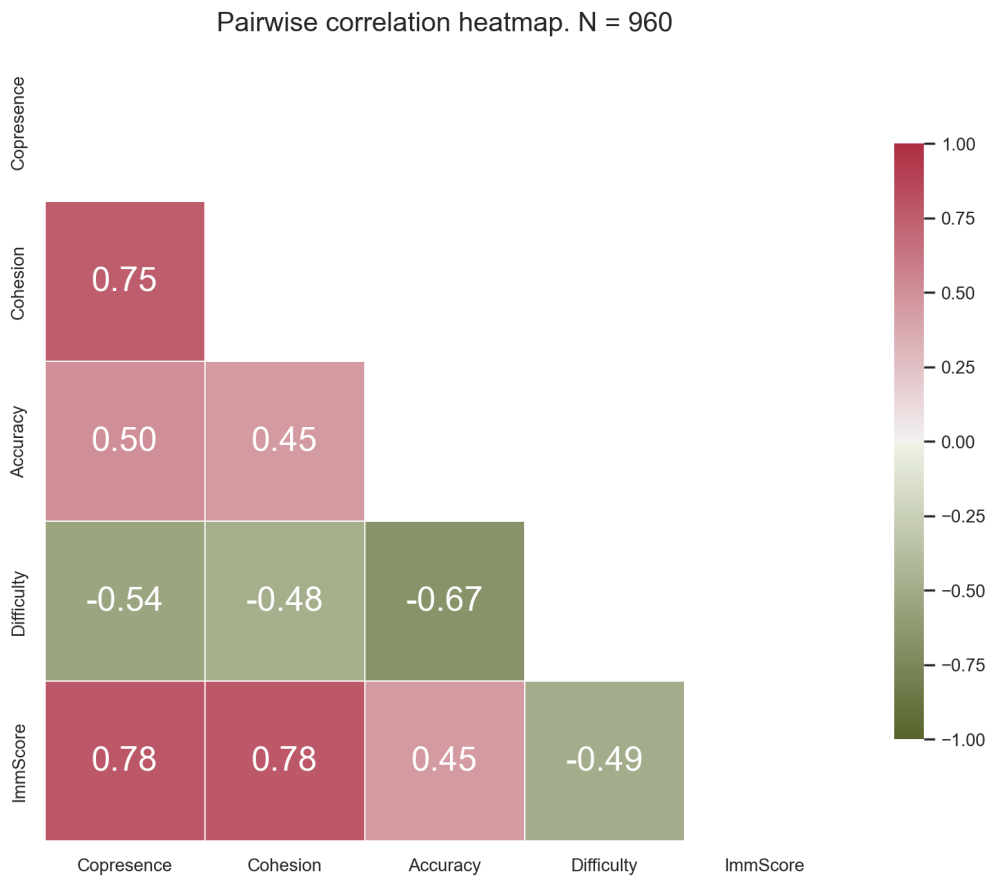


Figure 82: Correlation matrix between subjective responses to the trial questionnaire (Q2)

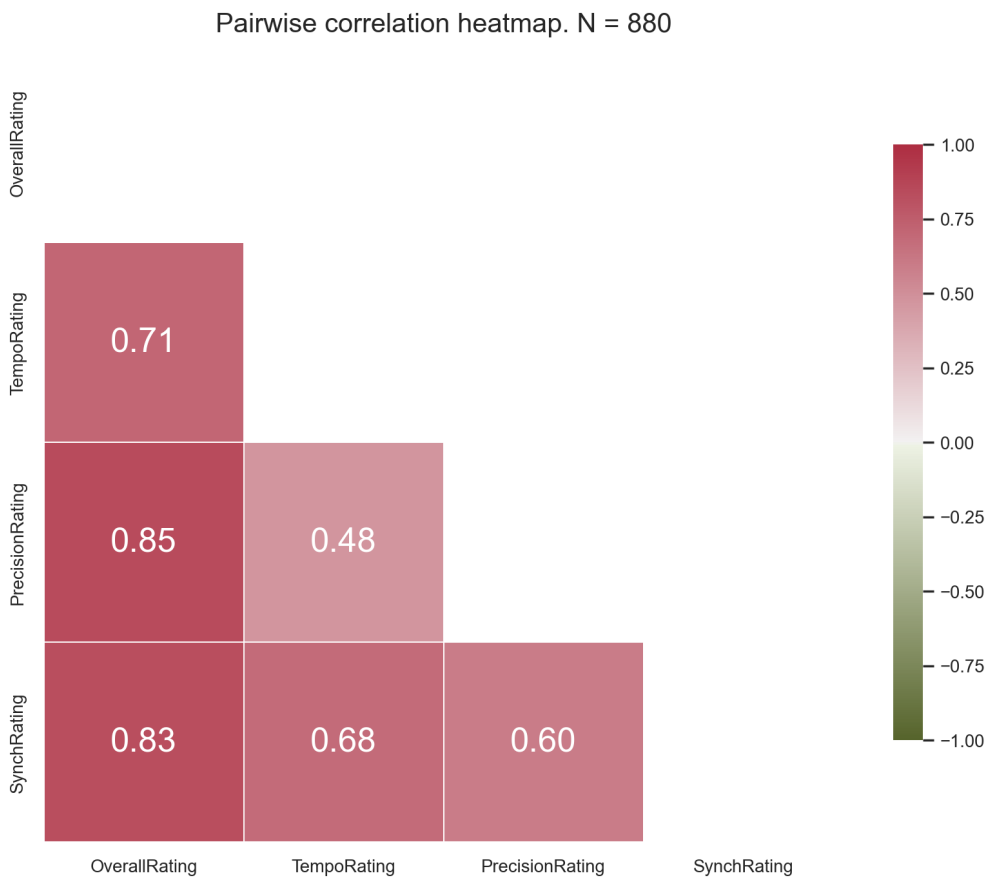


Figure 83: Correlation matrix between third-party ratings

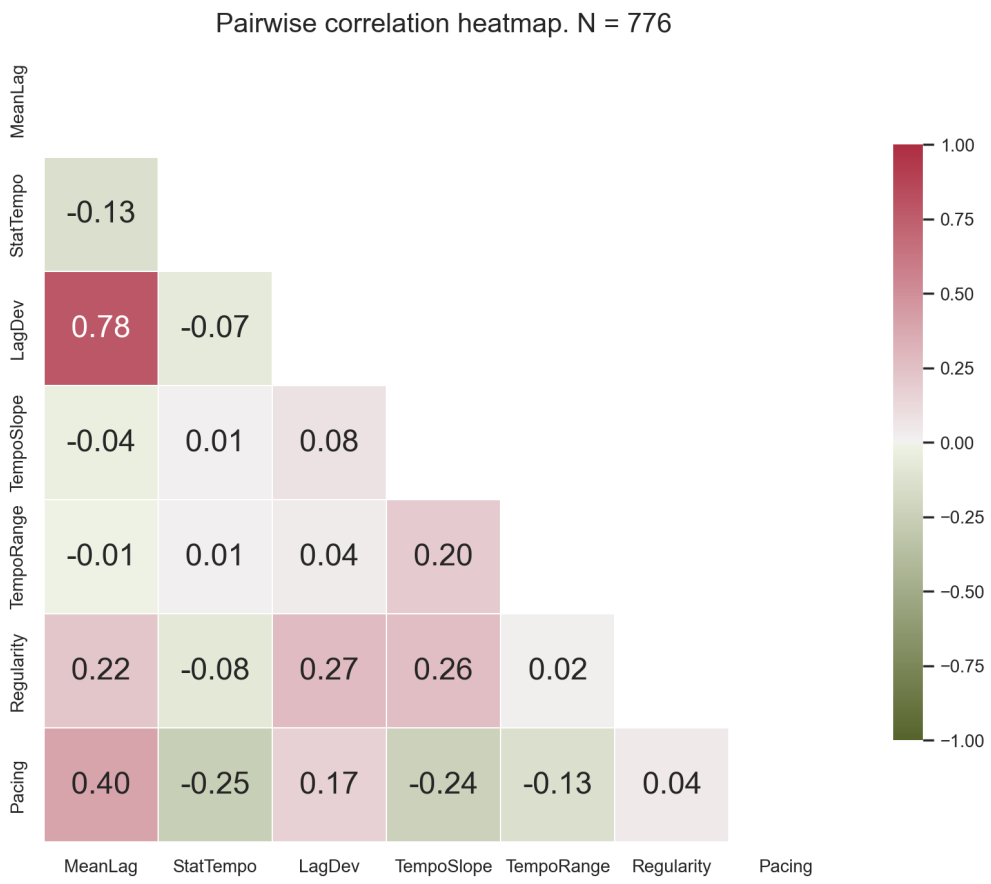


Figure 84: Correlation matrix between extracted objective metrics

Figure 85: Correlation matrix between trial questionnaire responses (Q2) and third party ratings

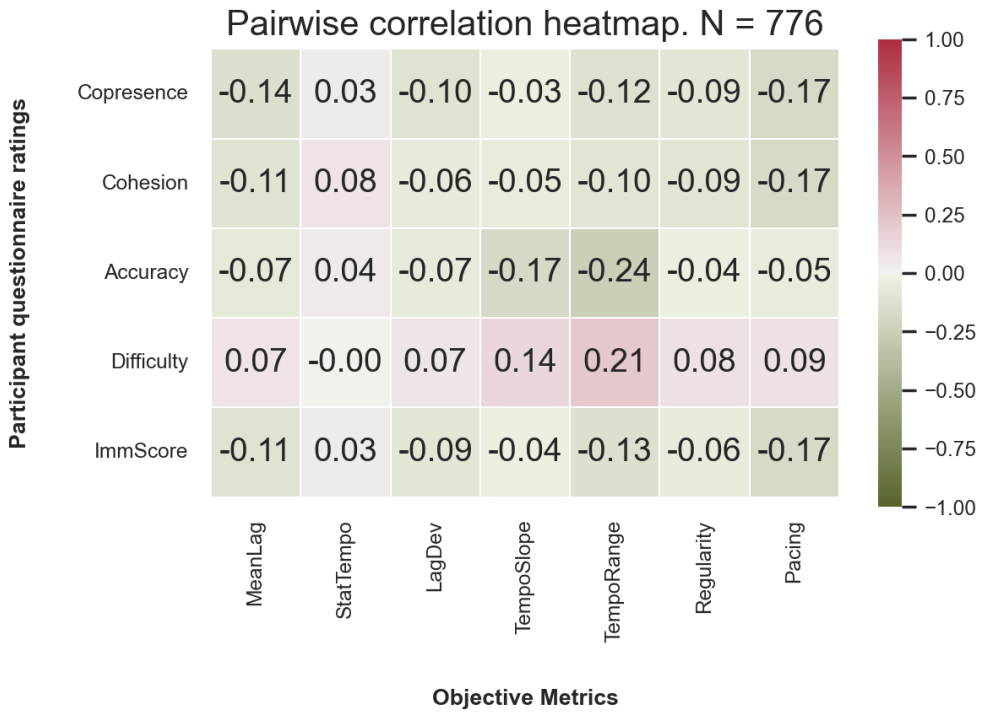


Figure 86: Correlation matrix between trial questionnaire responses (Q2) and objective metrics

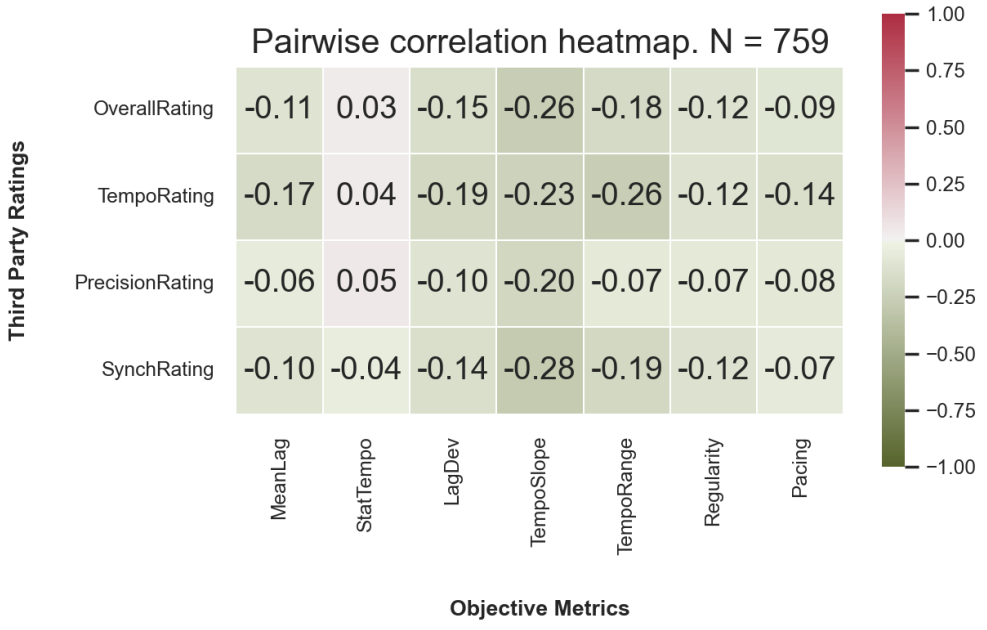


Figure 87: Correlation matrix between third-party ratings and objective metrics

CHAPTER VIII

DISCUSSION

This chapter discusses the results obtained by the LMM and GLMM analysis framework, placing the experimental results in the context of the theoretical framework and study design. The chapter assesses each previously formulated hypothesis, distinguishing between the null hypotheses that can be rejected and those for which not enough supporting evidence has been found. Secondary observations deriving from measured trends and contextualization of additional support data are also addressed, although without strong statistical claims, with the objective of pointing towards expansions of the problem statement for future studies. Finally, the chapter discusses the technical limitations of the study and to what extent the work generalizes in the larger context of immersive distributed networks.

1 Discussion of Results

The collection of results obtained over the course of this experiment is discussed here in terms of the objectives and hypotheses stated in Ch. IV. The format of this section is to restate the main research questions and hypotheses and provide a discussion beneath each one. Generally, different trends were observed across the evaluation layers, leading to a mixed set of rejections of null hypotheses and rejection failures. The random intercept (ID of a subject) was always highly significant and with a relatively strong effect size.

1.1 Impact of Auralization

H1. Auralization treatments, inspired by mixed and virtual reality systems, have a measurable positive impact on distributed music performance networks.

The investigation of this question yielded mixed results, and the null hypotheses can only be partially rejected in certain subsets of response variables. Focusing on the hypothesis related

to the questionnaire responses (H1.1), the auralization modes appeared to exert considerable influence on the *copresence* and *cohesion* latent constructs, indicating that a partial causal relationship exists. The interaction observed in the *copresence* responses between auralization and RoomID is intriguing, as it implies that the combination of dry/reverberant physical space potentially has a significant impact on copresence in general. It was observed that, compared to the raw condition that is devoid of any signal processing, the asymmetric congruent mode only improved the scores in the more reverberant Theater node, while the symmetric divergent mode led to a greater degree of improvement in the acoustically dry Booth room. One could deduce that a dry node would thus respond favorably to more reverberant auralizations; however, it remains challenging to determine whether this is a function of the specific BRIR characteristics, a function of the performance style being optimally consistent in situations with symmetric acoustic environments, or even a more complex amalgamation of the two. Given that the modes were explicitly designed to elicit copresence in distinct ways, this connection demonstrates at least a partial success of the implementations.

Regarding *cohesion*, the symmetric modes displayed an increase compared to all other modes. Notably, this was not observed for the AC condition, which involved the application of auralization specifically tailored for the room environment, and therefore it was anticipated that it would be rated relatively *cohesive* according to theory on room divergence effects (Werner et al. 2016). Similarly, the symmetric modes proved effective in improving *immersion score* by a relatively considerable amount, with the SD case performing optimally - although this may have been a function of the particular reverb rather than the mode. In terms of perceived *accuracy* and *difficulty*, the auralization mode was not identified to have a significant influence. The exception is the AD method that negatively interacts with the mid-level latency, but this result is approached with skepticism since the trend was not linearly observed at the higher latency level or across other modes.

Auralization exhibited minimal impact on the other evaluation layers, primarily because it did not emerge as a significant contributor to the mixed effects models. The sole objective evaluation metric that demonstrated an influence of the auralization mode was *tempo range*, where the modes were found to actually deteriorate the response (higher range of tempo curves) compared to the raw control condition; this finding is not corroborated by the other metrics. No

third-party annotation or rating exhibited an effect of the auralization mode. Consequently, we are unable to reject the complementary null hypotheses related to H1.2 and H1.3, which posited that meaningful patterns would be discerned across these layers. In summary, auralization modes were predominantly associated with subjective scores of *copresence* and *cohesion*, but did not show significant trends for the other responses. The statistical significance of the random intercept in all models also indicates that subject variation remains high and confirms the suggestion that "immersion" is also influenced by individual bias.

1.2 Impact of Latency

H2. Latency effects can degrade the quality of a distributed music experience.

The null hypotheses related to this question (i.e., *latency does not negatively impact the quality of a distributed experience*) can be unequivocally rejected across all the response layers examined (H2.1). The influence of latency was identified as a significant contributor to the model for every observed response, subjective, or objective. Notably, significant interaction terms in models still showed that latency serves as the primary contributor to degrading trends of "quality", particularly at the high level (40 ms), while the midlevel (20 ms) occasionally falls below the significance threshold, indicating that the effect size of latency is proportional to its amount. Although this outcome is not unexpected, the pervasive impact across all measured variables is of considerable importance to extend the existing literature to immersive NMP implementations. Arguably, the most innovative aspect of this result is that the degradation trends typically reported by the literature on objective metrics based on rhythmic characteristics (Rottondi et al. 2016; Chafe et al. 2004; Chafe et al. 2010) are also evident in the performers' subjective first-hand experiences when rating the immersive experience and in third-party listening evaluations. In conclusion, the null hypotheses associated with H2.1 and H2.2 can be confidently rejected.

1.3 Correlation of Metrics

H3. There exist positive correlations between *copresence* and other dependent variables.

The exploration related to this research question also yielded mixed results. The correlation analysis reveals that *copresence* exhibited a strong positive correlation with other subjective

metrics of experience and a moderate positive correlation with external listener ratings. Contrary to expectations, no discernible correlation was identified between the objective evaluation layer and copresence, nor between any other subjective metric and an objective one. In summary, there is minimal to no existing evidence to suggest that higher ratings of copresence correspond to a superior musical outcome, implying that the dimensions of "immersive" quality and the actual technical outcome of a musical interaction are distinct dimensions of evaluation, and improving one dimension does not necessarily count as a proxy for improving the other. The method used for exploring correlations only scratches the surface of the data, more advanced PCA and hierarchical clustering methods can serve to uncover lower-level dimensions of relationships across grouping factors or identify latent components not addressed by the data collection methodology.

Nevertheless, the more robust correlations observed within the subjective layers can contribute to the assertion that subjective quality attributes related to the "immersive experience", as rated first-handedly by a performer, were correlated to other subjective assessments related to the quality of interaction (e.g. *accuracy* and *difficulty*). Furthermore, these subjective assessments were moderately correlated with external listening evaluations of performances, tracing the fact that common effects (mainly latency) affected the rating of both layers.

1.4 Secondary Effects

The most prominent secondary factor manifested in the results was the *trial number* effect, which corresponds to the "time" or the number of repetitions within an experiment session. Generally, the impact of time exhibited a positive trend toward the improvement of metrics; nevertheless, conflicting tendencies emerged for certain responses. The probable reasons for the effect of time serving as a significant predictor on a response variable include "improvement through rehearsal" and "degradation through fatigue", which represent contrasting nonlinear confounding factors at play. However, the data do not adjust for this effect, complicating the differentiation between the two.

The random intercept relating to individual subjects was found to be significant for all metrics involved, indicating that individual biases and differences account for a large portion of the observed variability. This could be explained by the inherent differences in musical ability across subjects, intersecting with prior experience in performing in the presence of

latency, familiarity with the piece, and internal reference biases affecting ratings of presence and immersion. Although it was attempted to identify some grouping factors through the demographic questionnaire (Q1), no improvements in model fitness emerged from using those groupings as alternative random effects (tested individually and as nested random effects) thus making it difficult to point to specific attributes causing differences among the subject pool.

Other secondary effects that occasionally showed an influence were the interaction of the auralization mode with Room ID (*copresence*), the interaction of the musical part with latency (for the *mean lag* response), and the previous experience reported in NMP systems (for *perceived accuracy*). To avoid overfitting and focus on the larger predictors, the models were selected in a way that penalized the presence of a high number of factors and categories. Possible alternative models with higher fit potentially exist and can be found by favoring the AIC metric in all situations while adding effects to the model equations in order to capture the smaller contributions that might have been missed. However, this process can lead to noisy trends being mistaken for effects (overfitting) and to less generalizable claims.

2 Discussion of Supporting Data

This section reports the distribution of the answers to the agreement questions of the Debrief Questionnaire (Q3) (see Ch. VI, Sect.2.4) and the individual questions on the agreement on the direction of copresence ("here" vs "there") of the "Trial questionnaire" (Q2) (see Ch. VI, Sect.3). The answers to these questions were not analyzed through the mixed-models analysis framework, but are shown here because of their additional value in validating and expanding the arguments stated.

2.0.1 Q2 Agreement Scale Results

The agreement scales of Q2 were collected during the primary data collection phase. Each participant completed 30 trials of the experiment, thus 960 entries were available for this data. The purpose of this set of sub-question was to poll the understanding of the direction of copresence (e.g., participant feeling transported "there", or participant feeling coperformer "here" if the stream signal has a congruent local fit) with scores summing together to an "Immersive Score".

The set of scales was found to be internally consistent and reliable (*Cronbach's alpha* = 0.875) thus allowing their aggregation for the composite score. Here, those sub-questions are explored individually.

The figures below show the general trends of these questions over groups of auralization mode and latency. Since the agreement scale is bidirectional, the data is observed in standardized form (z-scored), where the value zero means a neutral opinion on whether copresence or cohesion was felt. Fig. 88 shows results for a generic reformulation of the main copresence question, both to serve independently as a copresence score and also to provide participants with a second definition of what it was seeking to capture. The initial concern was that participants may not easily connect with the definition of copresence, thus these questions were partially phrased with the intent of providing alternative definitions in the form of agreement questions. This ultimately proved unnecessary since the trends of Q2.5 (Fig. 88) and Q2.9 (Fig. 92) are very similar to those of Q2.3 (Fig. 61) which suggest that the direct rating of Copresence was largely negatively affected by latency, with symmetric auralizations having an effect in improving the ratings. Generally, responses are highly correlated with each other, favoring Symmetric auralization environments over asymmetric ones. As found in some of the mixed models' outputs, the (AD) auralization case shows the lowest ratings across all conditions including the raw condition.

A notable exception was the performance of the (AC) condition on Q2.8, the question was more related to "cohesion" rather than "copresence". Data shows that the asymmetric congruent auralization scheme (following the rendering principle of "local adaptation") was fairly successful in creating a cohesive environment for the listener and eliciting a larger sense of "receiver presence" rather than "transportation presence". This trend somewhat contradicts the responses to the direct rating of *cohesion* (Q2.4), thus requiring further investigations on whether specific formulations of the question were interpreted differently.

2.0.2 Q3 Questionnaire Results

The Debrief questionnaire was completed once per performer (total of 32 data points) at the end of the primary data collection process. Figures 94 to 96 report the answers provided by participants as they were asked to agree or disagree with several statements. The goal was to gather participants' impressions about auralization effects, latency effects, and subjective

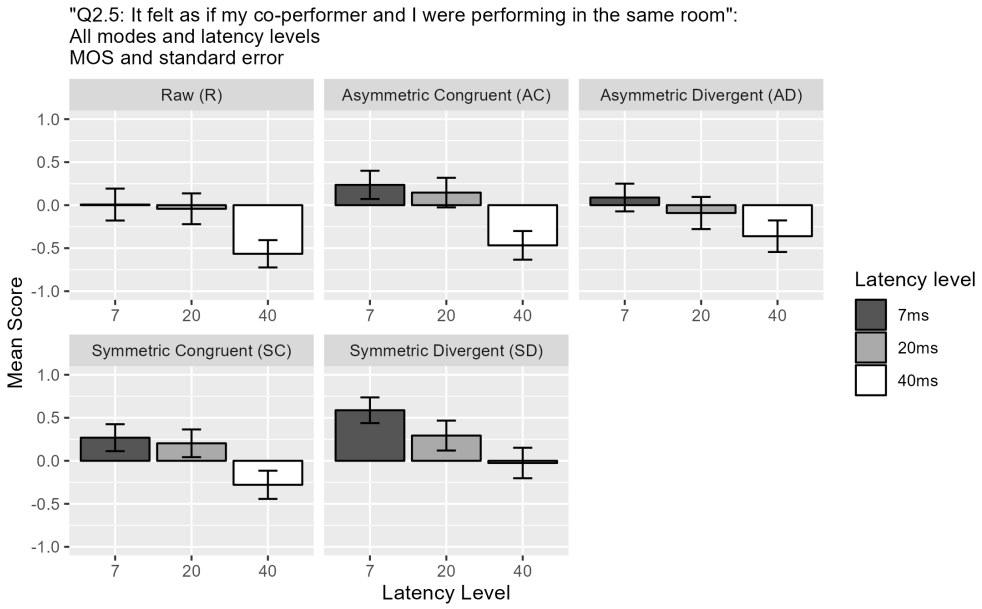


Figure 88: Trends over latency and auralization mode for question Q2.5, polling general feeling of copresence.

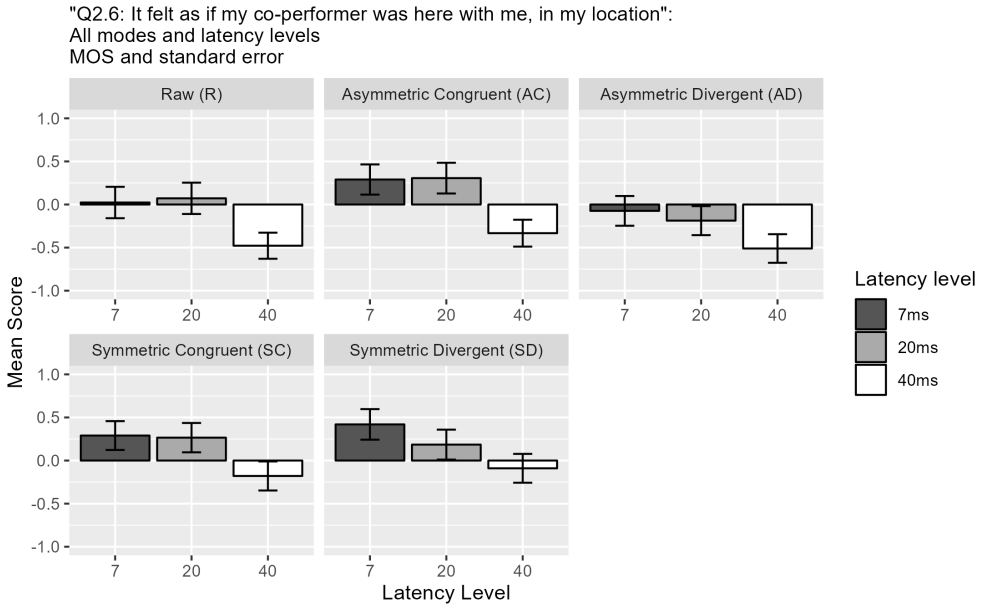


Figure 89: Trends over latency and auralization mode for question Q2.6, polling "local" copresence. The feeling that a connected used is present in the room of the listener.

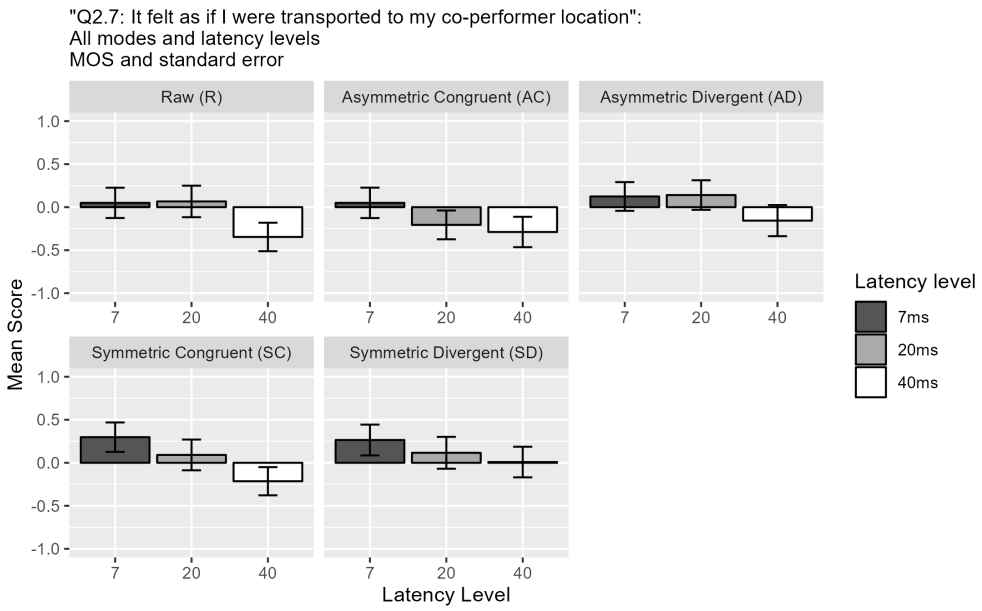


Figure 90: Trends over latency and auralization mode for question Q2.7, polling "remote" copresence. The sensation felt by a listener in being transported to a different location where a connected user is preset.

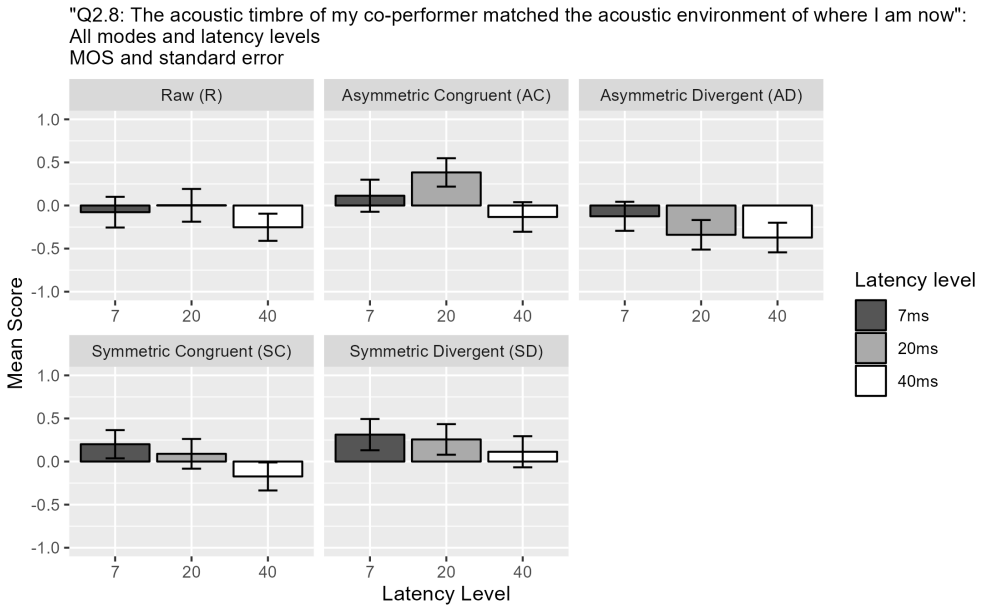


Figure 91: Trends over latency and auralization mode for question Q2.8, polling "acoustic cohesion", or "plausibility" .

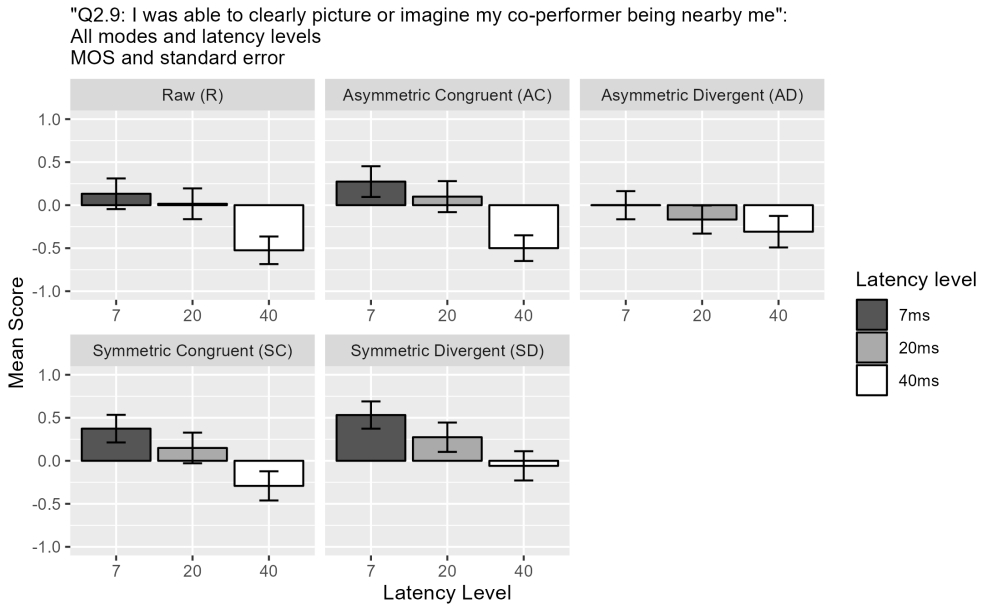


Figure 92: Trends over latency and auralization mode for question Q2.9 with another definition concerning auditory virtual presence.

impressions on copresence as a measure of self-evaluation and quality. The results point to the fact that some auralization modes were more conducive to a smoother performance as opposed to others (Q3.1). However, the modes were judged differently by subjects (Q3.2), implying that personal preferences could have played a factor in the judgment of environments. All participants agreed that a difference between modes was perceived in terms of "realism" (Q3.2), supporting the observations pointing to auralization having an impact on auditory copresence and cohesion.

With regard to latency, the results validate how much salient latency levels left an impression compared to everything else, validating the idea that high latency leads to a deterioration of immersive presence (Q3.4). The inquiry of the possibility of high-reverb effects leading to latency masking was not validated by the results of (Q3.5). Finally, it can be observed that copresence was for the large majority rated positively and subjectively correlated with smoother performance and enjoyment of personal experience (Q3.7 to 3.9). These ratings show how the general goal of reaching for high copresence is a valid one for augmented music experience; in other words, the factors that can improve copresence can likely improve the subjective quality of experience of a musician user.

Fig. 97 shows a qualitative summary of the responses to questions Q3.11 (*"Has this*

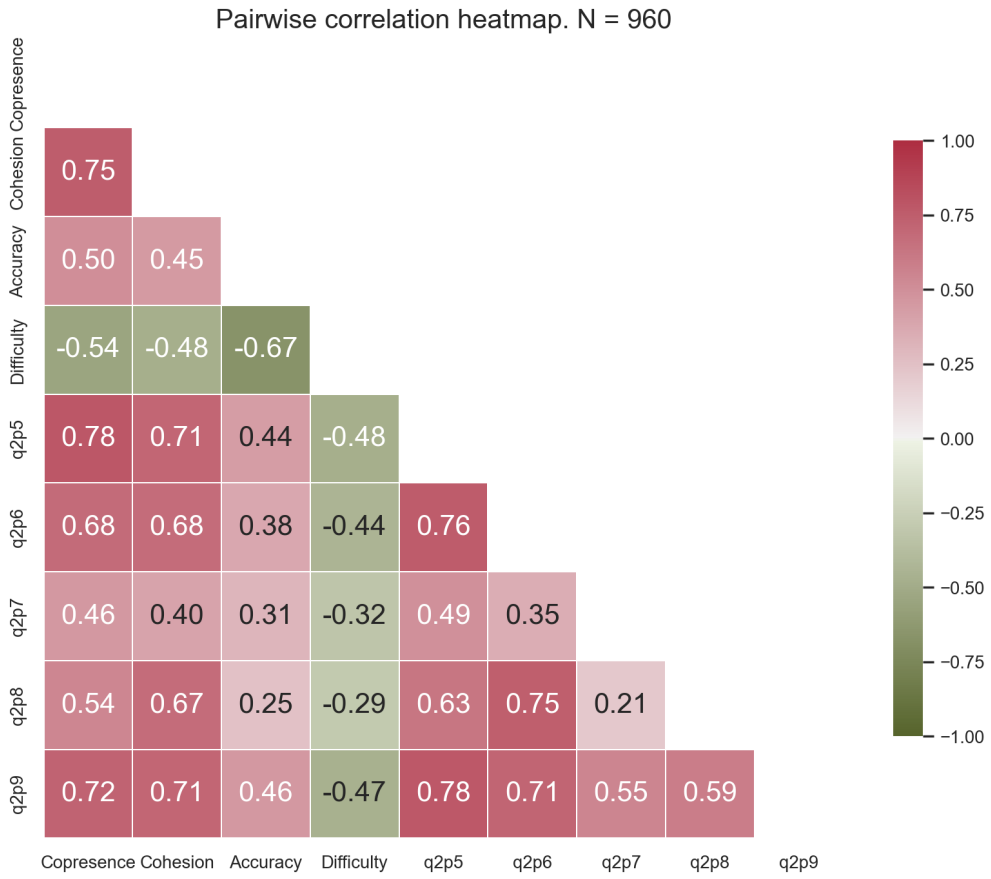


Figure 93: Correlation matrix for the expanded trial questionnaire results. Please refer to Ch. VI, Sect. 3 for the specific question wording.

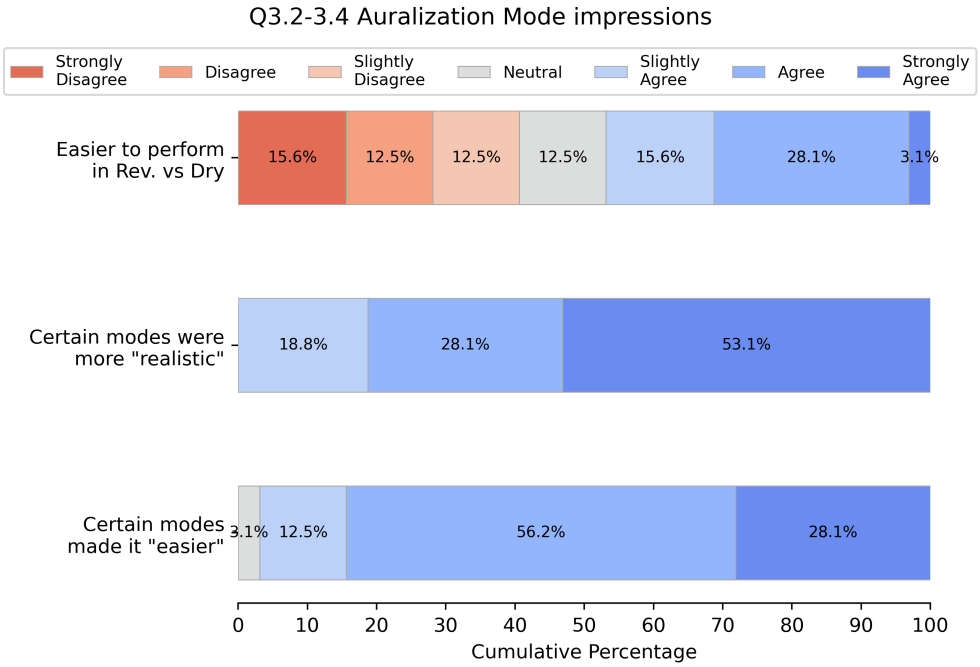


Figure 94: Answers to Likert-scale agreement questions, Q3.2 to Q3.4. "Auralization impressions"

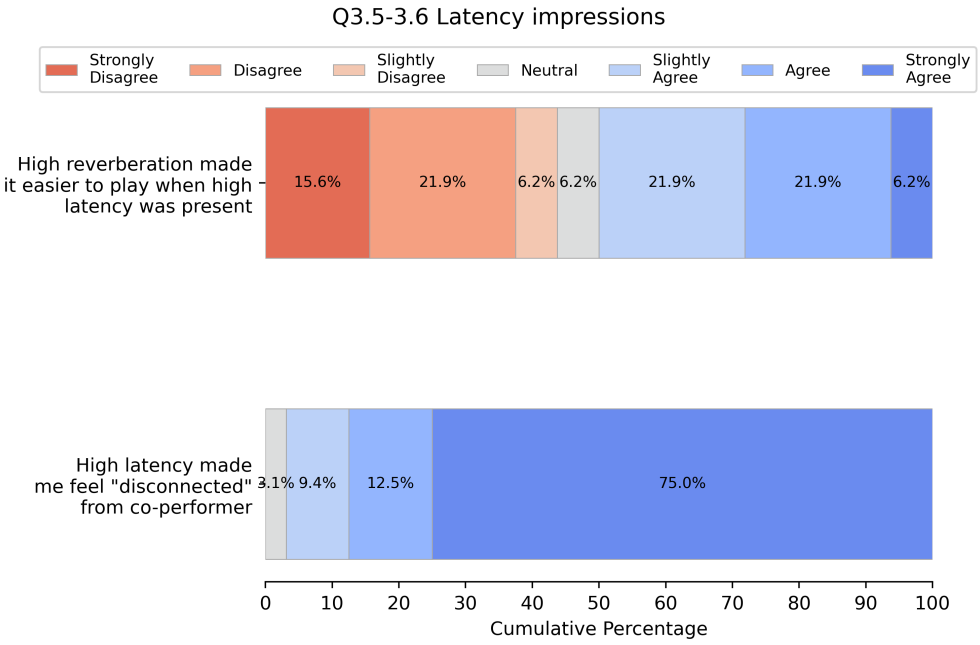


Figure 95: Answers to Likert-scale agreement questions, Q3.5 to Q3.6. "Latency impressions"

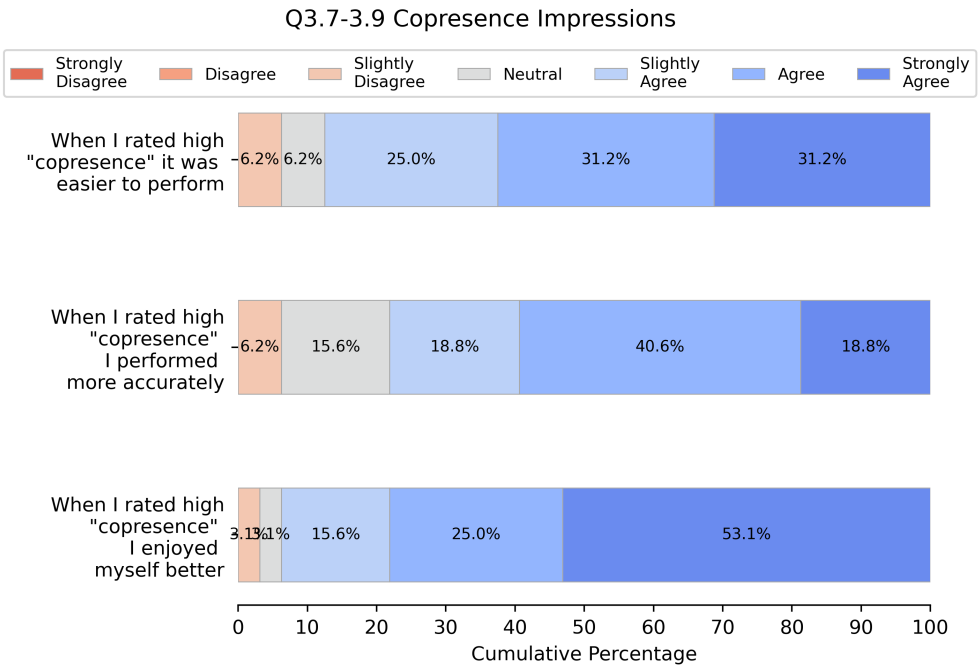


Figure 96: Answers to Likert-scale agreement questions, Q3.7 to Q3.9. “Copresence impressions”

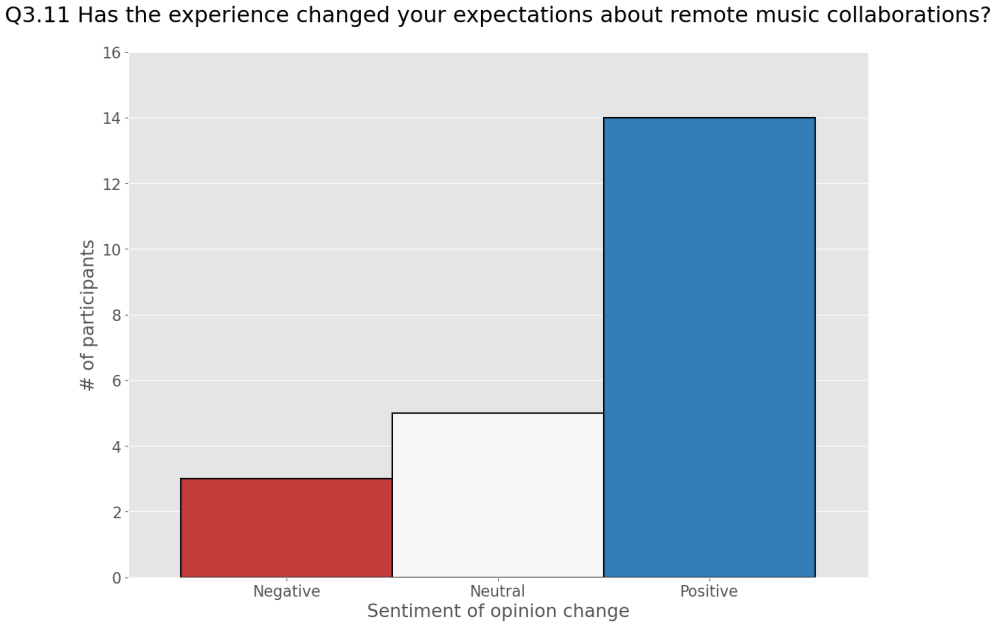


Figure 97: Sentiment analysis results on the responses provided for Q3.11, indicating the valence of reported changes of opinions about network music performances

experience, in any way, changed your expectations about distributed music, augmented acoustic environments, or internet-based performance? If so, how?". The collected text prompts (when provided as this question was optional, n=22) were fed into a sentiment analyzer tool (*Free online sentiment analysis tool* n.d.) that returned a label of "Negative", "Neutral" or "Positive" to the text extract. The label indicated the valence nature of the comment, with "Neutral" signifying that no change of opinion was reported by the participant. According to the tool, the majority of respondents reported a change of opinion towards a positive valence, as a result of being exposed to the immersive NMP system. Although these pieces of data remain high-level, they suggest that introducing auralization interventions in a distributed system can make the experience more attractive to a musician.

3 Contextualization of Findings

The effect of auralization was found to be impactful in the direct ratings of "immersive quality" produced by the performers, who did experience the effects firsthand, but not in regards to third-party listeners (who did not experience the effects in their evaluations) and objective metrics looking at tempo stability and beat synchronization. Generally, the present findings indicate a greater impact of "symmetric" modes than "asymmetric". Theory on the "room divergence" effect (Werner et al. 2016) led to the hypothesis that higher copresence would be experienced in congruent modes equally at each node. However, the "congruent" modes showed a room-dependent impact on copresence. Looking closer at this interaction, when the AC *asymmetric congruent* mode was applied, it was found that the adaptation of the signal character to the less reverberant "studio booth" room (which in this case was particularly "dry") produced lower scores than the same effect applied to the more reverberant "theater" room. The opposite trend was identified for the *symmetric divergent* condition, in which the response of the booth room was higher than that of the theater (both of which had a positive impact on the ratings). The suggestion presented by this finding is that the general effect of reverb is likely more impactful than the application of "congruency" at least within the applied evaluation scales, and dry acoustic conditions are generally rated lower than reverberant conditions. As a result, the implication is that the design of an immersive NMP might not necessarily need to invest in creating asymmetrically congruent auralizations (a high-effort process) in order to raise

subjective quality factors, especially in rooms that do not possess compelling acoustic characters to start with. Instead, an "optimal" reverberation curve may be found as a function of the more reverberant node, as applied in the SC case. There is no evidence that these findings extend to the assessment of virtual "plausibility" or sound "externalization" since literature suggests otherwise (Lindau and Weinzierl 2012; Pike et al. 2014; Klein et al. 2017; Werner et al. 2016), but perhaps in a musical interaction, the cognitive level of engagement required for the successful completion of the task may be drawing the focus of attention away from plausibility and more towards the timbral qualities of an auralized environment. "Plausibility" and "Externalization" were also not directly rated by the participants, so no strong conclusion can be made in terms of "realism" of the auralization mode.

The case of the *asymmetric divergent* mode is perhaps an interesting exception as it occasionally showed a reduced improvement effect, or a degradation effect compared to the other modes (e.g. for *immersion score*, or *tempo range*). From the design stage, this mode was considered the "least preferable" to apply, since the acoustic environments at each end node are asymmetrically non-congruent at each node. Looking at the model results, this non-optimal mode may have had a latent impact on the performances. Although participants were not able to hear the auralizations applied at the opposite node, the performance playing styles at each node might have responded differently with respect to the virtual acoustic character, potentially leading the musically educated performers to detect an inconsistency between their own expressivity and that of the coperformer. We know from the literature that musicians react differently to reverberation (Ueno et al. 2010) and listeners experience room divergence effects differently over time (Klein et al. 2017) so analogous mechanisms could be at play in this scenario. However, this remains to be investigated.

There is not much extensive literature on the impact of reverberation effects in tempo-based objective metrics within NMPs, meaning that there are not many comparison points available. Generally, the finding of this study agrees with (Carôt et al. 2009) in the fact that no discernible significant improvements, nor degradations, were recorded in response to auralization. The illustrated findings disagree with (Jung et al. 2000) which states that reverberation effects were detrimental and not preferred by performers, and also disagree with (Farner et al. 2009) where improvements in beat-synchronization precision were recorded as a result of auralization.

Post-experiment support data show that participants "liked" playing under auralized conditions, and no systematic degradations of either listening ratings or performance quality derive from their introductions.

With reference to the immersive model of experience (Lee 2020) it is observed that "copresence" and immersion (and to some extent "cohesion") are indeed related dimensions; however, there is no indication, within the context of NMPs, that these metrics can be used as a proxy to determine objective task success as hypothesized in (Zahorik and Jenison 1998; Mantovani and Riva 1999), suggesting a revision of the hypothesis space. Similar findings showing that subjective and objective evaluations can be in disagreement were also found in previous co-authored work (Hupke et al. 2020). Future application designs concerning XR and NMPs can be informed by these findings when assessing the balance of cost vs. quality in accordance with their target user and requirements. It is yet unknown whether the results would repeat when different latency-coping strategies are applied (e.g., leader-follower organizations such as a "laid-back" approach or "delayed feedback approach") (Carôt et al. 2007). There is no "one size fits all" encompassing answer but similar trends are expected across systems when traditional music playing is concerned.

In terms of latency, it is found to be a major factor in determining the success of a distributed network under all applied lenses of evaluation. The degradations that occur at a mild one-way latency (20ms) were not as impactful as the higher amount (40ms), indicating that the degradation is proportional to the amount. This is extensively corroborated by previous literature (Carôt et al. 2006; Rottondi et al. 2016; Chafe et al. 2004; Hupke et al. 2020; Farner et al. 2009) that also indicate the threshold of 20ms as the point from which measurable degradation occurs. One difference was that no strong tempo-deceleration trends in the clapping patterns were found. However, compared to other studies, the maximum latency levels were less severe and tempo variability was still similarly affected, so this is not deemed to be a significant deviation in objective results. What transpires is that distributed networks that aim to create an immersive experience need to pay great attention to the issue of latency, in order to ensure the right conditions for "immersion" and "presence" to occur. It was also found that auralization methods were not particularly effective in mitigating the effect of latency, a slight deviation from (Farner et al. 2009) who suggested that rhythmic precision was observed to be moderately improved when reverberation was applied

under latency conditions. Consistent with (Carôt et al. 2009) it is found here that auralization was not consequential in mitigating latency effects.

This study partially disagrees with the findings of (Olmos et al. 2009) which indicated that presence ratings were not affected by latency (although that experiment included visual elements), on the other hand, it agrees on the identification of "rehearsal time" and "increases in familiarity" being a contributor towards immersive qualities as shown by the "immersive score" results. The impact of "time" has been consistently found to bring improvements to several of the observed metrics. This finding is in agreement with previous work that points to the effects of "rehearsal" within a distributed system to be a factor of improvement both in terms of musical outcome (Chew et al. 2005) and spatial plausibility (Klein et al. 2017). Previous experience on the "Holodeck" concert events also supports this. An interesting question that derives from this finding is whether rehearsal within a distributed system using a particular topology would transfer to other topologies. To the knowledge of the author, no studies have looked into this possible question. However, general "exposure" to the technology and music material is likely to positively contribute towards the rate of performance improvement (Olmos et al. 2009).

A final word goes to the impact of the individual person. In every explored model, the effect of the random intercept is fairly high, meaning that subject variation is a factor in the ratings of immersive quality and performance quality. This is likely a function of training time, base ability, familiarity, and engagement levels, which bias the internal references that apply when evaluating an immersive system (Lee 2020) or reacting to latency factors (Farner et al. 2009; Carôt et al. 2009). While no data is available in regards to different rates of improvement, the trends here hypothesize to work similarly across subjects, albeit with different offsets in metric value (thus treated as random intercept rather than random slope). Further work assessing improvement variation may shed light on the magnitude of variability and help to adjust future statistical models accordingly.

4 Study Limitations

The results presented in this dissertation relate to a two-node rhythm-based distributed music interaction in which the two connected nodes present a divergent acoustic character (a theater and a studio booth). The generalization of the results needs to be verified, but when confronted

with other literature, it is deemed reasonable to state that the effects of latency and training are inherent to any setup topology, perhaps with more nodes or with equivalent rooms. Future studies may be interested in dissecting the auralization conditions in order to magnify the investigative lens over which aspects of it actually impact distributed performance. However, the interactions or reverb with latency and training effects are hypothesized to generalize regardless of the network topology, the auralization method, and the acoustic character due to the existence of similar findings in the literature.

Regarding auralization, it is unclear how well the results would generalize to other network topologies since there are setup-specific confounding factors such as reverberation time, visual field, and other parametric descriptions of the reverberation field, for which no representative sample distribution is available. Although this was not directly measured, the lack of visual anchor might also have had a negative impact on externalization as we know from the literature that the visual perception of a "likely" sound source can help to perceive a sound as externalized in the frontal hemisphere (Lindau and Weinzierl 2012). A possible variant of this experiment may organize the performers in a more orchestra-like seating arrangement, standing side-by-side and applying spatialization from a lateral position rather than the median plane. This scenario may play out better in the absence of visual cues, as the chances for localization confusion diminish and externalization may be facilitated (Reardon et al. 2018b). There is not a demonstrated link so far between externalization and copresence, but we know that copresence is also a function of the degree of plausibility of an immersive system (Lee 2020), itself driven by externalization of sound sources.

The other main constraint applied to this study is that of the music material studied. The piece represents a dynamic interchange of a complex rhythmic piece representative of an ecologically viable "realistic interaction". This choice steers away from other experimental approaches that adapt the musical interaction style to cope with latency (Carôt et al. 2009) and away from "clinical" repetitive clapping patterns used in previous similar studies. In other words, the chosen music piece represents an actual existing piece of music that could be realistically chosen in a traditional music interaction between performers. While this choice creates an added layer of complexity to the study and less robustness to quantitative metric extractions, it also lets the observations apply to a more common use case.

It was not the scope of this dissertation to examine the impact of acoustic processing methods on different genres and styles of musical expression. However, to fully understand the impact of the proposed fixed effects and generalize their impact, it is important to add comparative findings regarding the various choices of music material. Possible differences in results may arise from the choice of music score and its annexed family of instruments. Musical instrumentation would also affect the way performances should be evaluated. A hand-clapping rhythmic piece is appropriately evaluated by looking at tempo and beat metrics, as they are good characterizations of this category of music. Different objective metrics would need to apply for music pieces that make heavier usage of harmony or melody, with either score or improvisational structure.

4.1 Limitations of Evaluation Scales

Quantifying "presence" presents a difficult challenge, as it encapsulates a highly subjective and personal sense of engagement mixed with the illusion of "plausibility" and is influenced by various environmental factors and internal biases, as stated in the literature on the subject (Lombard et al. 2009; Lee 2020). The existing literature on "presence" does not offer a comprehensive formalization for the social auditory domain, and even less so for the realm of music. Consequently, this gap in established methodology leaves the responsibility of designing and adapting questionnaires to the specific application in large part in the hands of experimenters. Although this is beneficial in tailoring the questions to the application at hand, it makes research less comparable between studies. Validating measurement scales involves estimating reliability within and between studies. From what was possible to observe within this study, the internal consistency of the agreement scales used to measure "Immersion" was found to be highly reliable (*Cronbach's alpha* = 0.875). However, the particular formulation used relies on the assumption that "presence" and "immersion" are associated by a causal relationship (as indicated in the literature), but there is no strong evidence that this would necessarily be the structure for NMP contexts where several factors of engagement and sensorial inputs are at play. Future work on *confirmatory factor analysis* would have to be used to validate that the assumed association of *copresence* with *immersion quality* does indeed exist in NMP studies. Alternatively, different methods of evaluation related to sound "naturalness", "realism" (Rumsey 2002) or "plausibility" (Lindau and Weinzierl

2012) can be applied, although a different interaction paradigm would have to be designed to provide an immediate comparative reference of what a "natural" or "real" performance sound is to a performer.

The objective metrics used for this study are not immune to noise. The computation of the tempo curves on the shifting beat of the chosen piece proved to be unstable due to the sudden change in the beat that happens at the circling of the pattern. This measure is also highly sensitive to the shifting-part performer's individual mistakes (which were sampled from a population of music students rather than professionals), where a missed or extra clap in the score progression would cause a beat ambiguity and tempo "octave-error" (Schreiber and Müller 2014). To respond to these artifacts and improve robustness, a high degree of smoothing and regularization was applied to the extraction of dynamic tempo curves and beat patterns. This came at the cost of losing nuance and detail of the measure, possibly "drowning" out some of the more high-resolution effects that influence the final metric value. Strong effects, such as the impact of latency, are still captured by the objective metrics, but detailed differences are lost through the applied processing pipeline.

The relativization and contextualization of the objective metrics are also a point of debate. The extracted metrics on the distributed primary data are portrayed as degrees of increase of decrease from the baseline co-located performance metric. The baseline acts as a reference point under the presumption that a colocated performance consistently outperforms a distributed one. Consequently, the findings are examined in relation to the measurements acquired at the baseline stage point and interpreted in relation to how similar does a distributed performance gets to it. The advantage of doing this is that the base musical ability of each pair can be controlled for. However, in light of the observed time or "training" effects, this might be a flawed transformation of the metrics. Considering that the baseline was captured early in the experiment, the performers had limited opportunities to rehearse the piece and were less familiar with it compared to the later stages of the experiment. Reevaluation of the models allowing for an additional random slope effect could potentially uncover additional underlying patterns in how the objective metrics are affected by the auralization mode; on the other hand, these effects would be less impactful than the removed fix effect of time, potentially negligible. Another way would be to reinterpret the results with a different baseline (e.g. using the pair's "best" or "last" performance rather than

the baseline) or by absolute terms, at the cost of introducing confounding effects relating to the performer's ability.

4.2 Technical Limitations

As previously mentioned in Ch. IV, technical compromises had to be reached to balance the elements of immersive technology with the requirement of real-time signal processing. In terms of technical implementation, the auralization methods used in this study, which are based on static generalized BRIRs, do not exemplify the most advanced immersive systems available but still provide a reasonably high degree of fidelity. Potential improvements to the current auralization methods involve the use of individually measured BRIRs for a personalized fit and head-tracked dynamic 3DOF soundfield rendering. These are important components that can enrich the immersive experience (Roginska and Geluso 2017). Despite these limitations, the proposed lower complexity study platform still serves as a valid study model since the performers were not given the incentive to move around the sound field environment or rotate their heads (mostly they were looking at the music score).

An individualized measurement process was determined to be impractical due to the considerable engineering costs involved in individually acquiring the BRIRs of the participants in various rooms. Furthermore, although desirable, accurate localization of sound was not a crucial part of the experience as long as a sense of "externalization" and the "sense of space" were induced, something that static BRIRs are still able to provide if the recorded soundfield is sufficiently reverberant and decorrelated. BRIRs embed a spatial soundfield representation that includes the reverberation character of a room; thus, it is not possible through these results to separate the two effects with respect to the observed variables. Other studies (Farner et al. 2009; Cairns 2021) found that reverb insertion was generally a positive contributor to NMP. Instead, two- and three-dimensional source panning as a method to allow perceptual source segregation was not substantially found to improve a distributed music network (Jung et al. 2000; Hupke et al. 2020), although it is possible that this impact may vary in scenarios where additional nodes participate in the musical exchange. Other works point to spatialization in the virtual environment as a positive contributor to "presence" but not "realism" (Hendrix and Barfield 1996). Future potential work can

look deeper into the share of contributions brought by spatialization vs. auralization processes on the distributed music task.

In the context of this study, wired head-tracking systems would require a local rendering machine (Mania et al. 2004) to handle soundfield rotation, requiring an alternative optimized network topology that enables the rendering process to take place at each end-node instead of a central node making it difficult to remotely control the auditory virtual environments and the synchronization of the recording of the performance signals. Alternatively, the head-tracking rotational data can be streamed to the central node, where the binaural rendering is computed and routed back to the originating node. However, this would likely result in higher resource demand and the possible introduction of extra latency, exceeding the maximum viable levels outlined by the study. One possible alternative experimental setup is to establish a split-rendering system, where the sound spatialization is rendered locally with dynamic response, and the diffuse sound is rendered by the central node, where recording happens. This organization should be tested for synchronization concerns. Sect 5.0.1 provides a deeper discussion on the viability of head-tracking for real-time NMPs.

Another area of technical limitation relates to the absence of head-mounted-display (HMD) technology, an element of interest for future applications of distributed performance. Previous work on the "Holodeck" created the right study framework to introduce the use of HMDs in live performances (**Andrea Genovese** et al. 2019b). However, this element was removed from the study for the purposes of limiting the experiment variables and also because synchronization of the rendered video display to the audio streams could not be guaranteed without introducing additional buffering or without removing the live collaboration component. Nevertheless, it is worth considering HMDs in future applications because of their high potential to increase the immersive qualities of a multimedia system. As a note, prior work on acoustic calibration measurements conducted on HMDs shows that a mixed-reality application involving concurrent real and virtual sound sources may need additional equalization corrections to seamlessly blend the acoustic character of the virtual display with the soundfield surrounding a listener wearing an HMD (**Andrea Genovese** et al. 2018; **Andrea Genovese** and Roginska 2019).

5 Future Expansions of Study

In the short term, the current work can be bolstered by some additional analysis efforts. The most immediate improvement involves performing a more comprehensive and advanced correlation analysis, examining the size of the correlation difference between various grouping variables, and including the binomial responses. Due to time constraints, the present correlation analysis evaluated general trends without differentiating by grouping level. This can be achieved in the short term and could lead to a better understanding of relationships across evaluation layers for specific modes or latency levels. PCA analysis can also be applied to explore the data space and provide correlation insights on the response variables in relation to abstract principal components. One other low-hanging fruit concerns a fixed effect that was left out of the analysis, the RT60 level of the virtual rooms. This effect can be easily explored through the existing models and reveal another dimension of the acoustic influence on preference and performance, as an alternative to the "Auralization Mode" framework. In addition, a deeper look at the "double-slope-decay" (Boren and **Andrea Genovese** 2018), which concerns the interaction of two joined reverberant environments, may cast light on why certain reverberant environments were rated higher than others (for example in the divergent modes, local reflections might be heard through the open-back headphones, but the late reverb would be perceived from the more reverberant room between local and virtual environments), especially in the dry room. These results would only be exploratory since there is not a continuous sample of RT60 parameters available within this data.

Another aspect to consider is to review some of the objective metrics that demonstrated lower robustness in the primary data. Some signal processing steps taken to remove artifacts may have inadvertently removed partially meaningful data, resulting in the absence of significant effects for some proposed metrics (e.g., "tempo slope"). One approach to analyzing these metrics involves adding a finer resolution of the metric by sampling the tempo curve at regular intervals or at salient performance points. Additionally, additional LMM and GLMM models can be explored by looking deeper into the variables of questionnaire Q2, by introducing a random slope effect related to "Trial number", or by incorporating other available data as potential predictors.

There are several possible areas of exploration for the methodology as applied to new

variations of this experiment. Further validations of the conclusions could be assessed by separating this study into broken-down iterations focusing individually on each of the key effects with higher resolution. One possible investigation is the study of the impact of specific reverberation parameters on distributed musicians. A major candidate for attention is the "reverberation decay time" and whether it alone can serve as a meaningful predictor over the observed scales. Although we have acoustic parametric data available for the rooms used in this study, the sampled collection is too sparse for being representative, as the sampled set of rooms used for auralization was driven by practical accessibility rather than for being a representative linear space of variation across acoustic characters. Furthermore, the space of acoustic variation is multidimensional, and different subband decay times and relative energy can be found in the frequency domain of an acoustic response. This is best explored through synthetic room reverberation models, as they allow flexible parametric control. Such an investigation might determine what is an "optimal" virtual room for distributed performance and whether the outcome would land differently in relation to the physical location of the performers.

As pointed out in the discussion of the limitations, the experiment results would benefit from the contextualization of ulterior data points considering different network topologies, score features, instrument-specific experience, and instrumentation, in order to identify variations and validate the effect sizes in a more generalizable way. At a design level, the choice of instrumentation and piece (or different tempo, beat pattern, etc.) or hierarchical organization of the ensemble is likely to require alternative sets of metrics more capable of capturing response variations as needed. The role of different musical genres and styles in shaping the NMP experience should not be underestimated. For example, the complexity of rhythmic performance has been reported to lead to different latency tolerances (Rottondi et al. 2015). Another component of variation is the "node-ensemble" definition; different case studies may involve additional nodes, hierarchical relationships, and various combinations of physical environments from which the interaction occurs. Gathering more data points across network variations could provide valuable insights into the generalizability of the findings and offer opportunities to optimize the NMP experience for diverse musical contexts. Moreover, the use of complementary measures, such as physiological indicators or behavioral data, could help provide a more comprehensive

understanding of the factors that influence presence and validate the self-report measures used in the study.

5.0.1 Improving the immersive system

Another important aspect to explore is the use of a dynamically rendered spatialization environment; the technology exists to introduce these elements on connected nodes, provided that the latency requirements can be slightly relaxed. This can be achieved by applying efficient soundfield rotation methods and spatialized dynamic auralizations by capturing diffuse soundfields through multichannel microphones (Merimaa and Pulkki 2005). This is particularly desirable for large-scale virtual ensembles, such as orchestras, where the spatial audio cues are part of the regular experience of their members who in real life may rely on the directional soundfield environment to synchronize with co-performers during key score events. In general, the importance of an immersive display is likely to scale up with the size of a virtual ensemble given that more perceptual discriminatory information is provided by the auditory display, potentially enabling a "cocktail-party effect" (Cherry 1953), or spatial-release from masking (Litovsky 2012). Additionally, a visual reproduction system, such as a VR or AR headset, would incorporate a multimodal interaction component. However, the processing latency and computational resources of this process are likely beyond acceptable levels for traditional music performance, necessitating the introduction of alternative musical interactions based on virtual music instruments and musical playing styles based on latency-coping strategies.

As for head-tracked spatial audio rendering, it is a very desirable and feasible feature to implement to achieve dynamic spatial audio rendering, but certain precautions need to be taken. Wireless head tracking would be the preferred avenue in order to avoid cabling hurdles. However wireless latency is prone to variability, jitter, clock synchronization issues, and sensitivity to interferences, reaching latencies ranging from $10ms$ to $140 + ms$ (McPherson et al. 2016). Holistically speaking, the wireless data transmission latency would be on top of head-tracker internal processing update latency and Ambisonics rendering for 3DoF sound (which itself scales up by a few milliseconds with increasing order/channel count). Wired approaches can greatly reduce the stages of signal processing involved, reducing the taxing of the "latency budget". In fact, recent wired implementations have placed the additional latency of wired dynamic rendering

at $30ms$ (Cairns et al. 2020), which is high for rhythmic interactions, but still within tolerable general performance range, making the system more usable. Ultimately the overall latency is a function of many system-specific attributes and parameters, such as equipment clock speed, buffer size, sample rate, streaming medium, and more, making the choice of head tracking scheme something to balance between available resources and latency headroom. Part of this tradeoff conversation also regards the capacity of Ambisonics orders implementable for dynamic rendering. Processing latency scales linearly with the Ambisonics encoding order (thus with the number of channels) and soundfield rotation computation expense follows that. The higher the Ambisonics order, the better the frequency balance and the spatial resolution, but the higher the processing cost. The choice of Ambisonics order and head-tracking system are thus two sides of the same coin.

A way to possibly decrease the latency of a system is to introduce parametric instruments and local synthesis instead of routing processed audio buffers. Transmitting metadata can be fairly more efficient, especially when dealing with low-bandwidth scenarios (e.g. wireless multi-user mobile systems). In the case of claps, perhaps an onset detector and re-synthesis processor could turn out more efficient in terms of packetization latency and bit rate and be a possible solution for clap inconsistencies across performers. However, some synthesis/rendering latency would be introduced at each receiving node and the naturalness of the sound character would be lost. Convolution itself is an expensive process. The computational cost of convolution methods, while reasonably efficient in certain partition-convolution implementations (Torger and Farina 2001), linearly scales up with the number of channels involved (i.e. two convolutions per stream in the case of stereo displays, more in the case of FOA/HOA displays). The implementation of parametric reverberation systems, such as a *Feedback Delay Network* (FDN) (Jot and Chaigne 1991) can be a great cost improvement for reasonably good-sounding reverb. This usually comes at the cost of fidelity to the measured space (lower than BRIRs), so auditory cohesion is expected to suffer. Some more recent FDN versions could be tested to verify fidelity to a measured space (Ibnyahya and Reiss 2022) and simulation of reflections directionality (Alary et al. 2019). To make accurate decisions, there are too many factors at play to precisely point to exact comparative latency metrics, but elements such as device clock speed, codec, transmission protocol, buffer size, and sample rate, all have an impact on the end-to-end system and are part

of the latency-reduction technical considerations that take place in the establishment of an NMP application. Smaller targeted technical experiments should be able to test some model systems and provide a systematic table of expected cost differences and scalability.

5.0.2 Theoretical validation

Alternative assessment scales can be developed to capture different dimensions of "immersion". "Engagement" and "Involvement" are two dimensions of immersion that have been excluded from this study but that contribute to the high-level construct of "immersive experience" (Lee 2020). Although this study points out that immersive quality does not translate to musical quality as observed externally, this has been assessed from the point of view of *copresence* and auditory *cohesion*. Other forms of assessment or a composite comprehensive questionnaire may be able to re-evaluate this conclusion.

Larger considerations must also be addressed when examining the underlying theoretical framework under test. A *confirmatory factor analysis* validation study with a large sample size would be necessary to establish the relationships to the latent construct in a more formal manner and adapt informed hypotheses accordingly. This is partially addressed in previous literature on presence (Witmer and Singer 1998) and immersion, but not in the same context as this work. No specifically established questionnaire was available for application in the NMP field, leaving the formulation of questions primarily to the discretion of the authors and experimenters by adapting methodologies used in other fields. However, very recent work on the topic is moving towards the exploration of validated scales for immersive audio listening in XR (Toet et al. 2021; Lee 2020; Wycisk et al. 2021; Tsioutas et al. 2020; Cairns et al. 2020), showing promising developments towards establishing research resources that can allow comparative studies across systems.

CHAPTER IX

CONCLUSIONS

This final chapter summarizes the conclusions reached by the study and the stream of projects that originated it. In addition, the chapter summarizes the value of the work towards its contribution to the field and towards future directions in immersive distributed music applications.

1 Summary

This dissertation has portrayed a series of projects and studies focused on the exploration of immersive distributed networks dedicated to music performance. The works are linked by a common interest in introducing virtual and augmented immersive technologies in different modalities, with the intention of creating a high-quality, plausible, and "realistic", immersive experience for musicians and audiences. Previous work on the "Holodeck", a multi-room experiential research platform, led to considerations on how auralization techniques can be used to provide an illusion of copresence to remotely placed musicians, and whether the immersive experience of a musician is conducive towards a higher quality of performance. These early projects involved the implementation of multimodal distributed performance networks, the use of motion capture technology, and pilot experiments that sought to explore the relationship between objective and subjective evaluation layers with respect to the impact of immersive audio techniques.

The experience gained through the engineering challenges and observations encountered led to the formulation of the hypotheses that were used to drive an empirical study on immersive NMP. The study focused on the interactions of network latency with auralization treatments, which were motivated by design principles that aimed to improve the sense of "immersion" and "presence" in participants. The results of this experiment can help to contextualize the

complexity of an immersive system with an expected outcome of quality from the perspective of musicians and audiences, showing that there is a distinction between interventions that improve the musician's experience from those that improve the audience judgment or measurable performance metrics. This finding is useful for guiding future application designs in the balancing of complexity according to the desired principal target user (e.g. audience or musician). Furthermore, the observations obtained are useful for identifying areas of agreement with previous literature on NMPs, psychoacoustics, and virtual copresence, and areas that require further studies.

2 Experiment on Immersive NMPs

The empirical study conducted for this dissertation work has been designed to study real potential scenarios of distributed music collaborations that may arise in performance networks based on the Internet. The primary objective of this work was to explore a multilayered evaluation framework and to offer insights into how subjective and objective measures relate to each other in enhanced NMP, with a particular focus on the aspect of *auditory copresence*.

The selected case study composed of two acoustically divergent nodes connected through a central distribution node represented a typical situation incurred during previous projects in immersive distributed music. A set of auralization modes was specifically assembled for this study using design principles derived from collaborative interactive VR and AR experiences involving different combinations of *congruence* and *symmetry*. The auralization modes were realized through BRIR measurements and implemented over a local analog infrastructure model network composed of two interacting nodes and a central distribution node, from which signal latency levels could be controlled and performance data recorded.

Primary data was collected from a total of 32 musicians organized in pairs who performed a relatively complex clapping music piece, both colocated and distributed over the network nodes. Three layers of evaluation response data were collected through the primary data and fed to an analysis framework based on mixed-effects regression models. The evaluation layers consisted of a participant questionnaire that polled quality ratings related to the "immersive experience", objective metrics based on tempo and beat synchronization, and third-party listening evaluations consisting of quality ratings and annotations of performance inaccuracies. It was

hypothesized that auralization interventions would produce positive improvements on all scales of evaluation explored, that latency would degrade observed responses with potential interactions with auralization modes, and that correlations would be found across layers showing a connection between "copresence" as an indicator of "immersion", and the other dependent variables.

2.1 Latency as Dominant Effect

The results show that the latency factor is the dominant effect in producing quality evaluation degradations in all data layers collected, subjective, or objective. The amount of observed degradation is also proportional to the amount of latency within the network. Generally, latency affected all observed response variables, embodying the principal contributing factor affecting the success of a distributed interaction as evaluated by all lenses applied.

2.2 Auralization, a Room-dependent Contributor to "Immersion"

Introducing different types of auralization processes in an NMP interaction yielded partially successful results toward the objective of improving evaluation metrics. The auralizations had a positive impact on the improvement of subjective experience ratings, such as *copresence* and *cohesion*, and little impact on the other evaluation layers. While the *symmetric* modes were generally more successful, the *congruent* modes showed a higher degree of variation in rooms, pointing to higher success in more reverberant spaces than acoustically dryer spaces. Objective metrics and third-party ("audience") evaluations were not affected by the auralization mode.

The implication derived from the results achieved is that auralization is effective in improving subjective participants' scores in the context of "immersive quality" but this effect degrades sharply with higher levels of latency. In general, auralization was not effective in improving third-party ratings or objective metrics, highlighting that its introduction did not significantly affect the musical outcome, neither positively nor negatively. Since applying auralizations usually entails an increase in complexity costs and engineering efforts, application designs may consider whether an NMP is tailored for delivering a quality experience to a musician or to an audience, and choose to introduce immersive audio techniques according to where the computational "budget" should be allocated.

2.3 Learning Effect or Familiarization

The trial number, a time indicator, was found to have a consistent impact on most response variables, although with a minor contribution compared to the latency factor. In most cases, the effect of time brought about improvements in all the rated levels of quality, both objective and subjective. Consistent with similar studies, this finding suggests that over time, both the perceived immersive qualities of a system and the objective outcome of performance can improve. The attribution of this effect is not definitive but is hypothesized to be caused by "learning effects" over two dimensions, as players became more comfortable with both the performance material and the distributed system.

2.4 Copresence as Quality Indicator

A correlation study showed that the *copresence* response, which was the most direct target of the auralization process, is highly correlated with other subjective metrics of experience rated by performers. Moreover, *copresence* is found to be mildly correlated with third-party ratings. This finding supports the theoretical association of "copresence" with "immersion" as evaluated by the performers and "task success" as evaluated by an external audience, and these associations are found to be primarily affected by the level of signal latency. There is no evidence available to determine the possibility of a causal relationship between these mildly correlated dimensions, so further work on factor analysis is needed to help validate and deepen the conceptual models from which the hypotheses stemmed. No significant correlations were found between *copresence* and rhythm-based objective metrics, nor between third-party listener evaluation ratings and objective metrics.

3 Value of Work

Facing the question originating the author's interest in the topic *"Does immersive audio technology improve the quality of a distributed network performance?"* The answer seems to depend on the method of evaluation. There are significant indications that the subjective immersive experience does indeed improve with the introduction of spatial auralization methods, and applications tailored for musicians and performers can well benefit from these processes. However, there

is no evidence to state that these methods translate to improvements in the musical output of the collaborative interaction, meaning that the engineering cost of creating an immersive experience for musicians may not be a worthwhile contributor to the audience experience.

In summary, current work provides valuable insights into the impact of auralization modes and latency on networked music performance. The main contribution of this work lies in the improved understanding of the factors affecting "quality" evaluation across different realms of judgment and insights into the interaction of auralization designs in the subjective experience of collaborating musicians involved in the system and the final musical outcome. Several layers of methodology have been proposed to capture different types of variables, providing several points of comparison with similar experiments found in the literature. There is also a significant discovery value that is useful in identifying the essential factors on which future study designs can focus and the relationships that exist between different response variables, expanding the conversation around "immersive experience" and "presence" and their relationship with interactive virtual collaborations. Moreover, the dissertation presents literature validations in which the trends published in the literature regarding the impact of latency have been confirmed and expanded. Immersive network music performance applications exist in many different forms; this work helps to identify the understanding of the challenges involved when traditional music performance goals are set for the application.

The data collected for this dissertation contain a wealth of information that has yet to be unpacked and are a valuable resource for several possible new studies that examine NMP behavior or "copresence" effects. Going beyond the main questions posed here, new study directions are possible by delving deeper into the primary data, for example, by examining different types of performance evaluation metrics, increasing the time resolution of the proposed metrics, reevaluating the data under different categories, or exploring different types of statistical models. Another intriguing aspect that was not studied here is the potential effect of the acoustical parameters on the responses, although this is not optimally sampled in the set of auralization filters used for this study. Therefore, the collected primary data set (and the evaluation data annexed) is planned to be published and made available to the public, with the desired result of sparking more interest in the subject. The establishment of new journals with interest in the "Internet of sounds" (Turchet et al. 2020) or "Internet of musical things" (Turchet et al. 2018)

are signs of a growing field. To the author's knowledge, no public dataset of distributed music performances is currently available to the wider community for study, making the data attractive also for subset studies or meta-analysis studies.

4 Future Directions in Immersive NMPs

Although some of the hypothesized associations did not show in the results, pursuing optimal immersive distributed music networks is still a worthwhile endeavor. This is primarily supported by the improved quality of immersive experience felt by participants within the system, where we see some of the auralization interventions consistently having a measurable positive impact on the provided scales, while no consistent detrimental impacts on performance were found. Remarkably, when looking at the results of post-experiment questionnaires, there are encouraging responses from the participants indicating positive opinions of auralization and a raised interest in the distributed music experience.

It is ultimately largely up to the application goals to determine how the tradeoff between computational costs and subjective immersive experience quality should be balanced. Provided that an audience experience seems to be agnostic to the experience felt by distributed musicians, an immersive experience implementation may not be a priority when the computational "budget" or the "rendering-latency budget" is limited. Being that latency is such a determining factor, keeping the one-way delay within an optimal range (up to a value between 20-35 ms according to genre, tempo, and beat complexity) is crucial for traditional or tempo-critical musical interactions. While this thesis had access to high-tier professional equipment which kept rendering latency within <3ms, it is expected that lower-tier audio equipment or mobile systems would present much higher rendering latencies when applying artificial reverb and spatialization, making the music interaction very sensitive to virtual acoustic environments. On the other hand, high-latency networks which already do not permit a traditional real-time musical approach, might as well consider adding a few extra steps to make the experience immersive, considering that once the real-time latency interaction threshold is broken the addition of extra rendering processes is relatively less impactful to performers. An interesting middle ground might present itself in hybrid low- and high-latency scenarios, where parts of an ensemble are able to produce music with a real-time interaction approach while more distant nodes are not capable of doing

so. In this case, certain hierarchical or asymmetric system designs can be applied, according to what node is considered the most "critical" for an audience or for recording, or according to the musical organization of a piece (for example a percussive or melodic section of an ensemble is more sensitive to latency than a harmonic section), with the intent of optimizing the latency experienced across the most sensitive connection paths and allow an immersive experience whenever possible. Overall cost also scales up with the number of streams to spatialize per processor node, prompting the consideration that perhaps certain kinds of instrumentation (e.g. low-frequency bass) may be excluded from the spatialization process in order to save computational budget.

Another important implication stems from the observation that auralization "congruency" does not seem to positively affect the copresence experience, nor the musical outcome. Given the fact that accurate local acoustic characterizations are hard and costly to obtain, this is a positive remark for XR application development. In the context of musical interaction, there is likely a relationship to uncover between the acoustic character of the local listening space and the optimal artificial reverb that can be applied at the rendering stage. However, this study brings no evidence that the accuracy of a virtual acoustic environment to the local listening space is a significant factor for perceptual quality, or that this accuracy is even a desirable feature when applied to excessively dry or reverberant spaces. This finding significantly simplifies the auralization problem as it indicates that the investment in local acoustic adaptation is not a determining factor for a music performer within a virtual environment, suggesting instead that the timbral quality of the experienced reverberation is instead more conducive to an enjoyable music performance.

The road ahead is still very open for further explorations that can help to generalize the results to different interaction paradigms. There is ample opportunity for further exploration and development in the field. An important question to answer more in-depth is the investigation of the impact of linear changes of reverberation parameters and their interaction with the physical room environment where each performer is located. By addressing the larger considerations and expanding the scope of research, future studies can contribute to establishing a more comprehensive and nuanced understanding of the factors that influence the quality of experience in immersive distributed music performances, and how results generalize across different interaction paradigms.

A future where compelling distributed performances are an everyday occurrence can be enabled by low-latency XR technology, which may well be capable of creating a mobile adaptive plausible virtual experience backed by fast transmission networks and IoT support data. A promising outlook for the future of music performances may encompass the potential for musicians to practice and execute any kind of music genre with instantaneously generated audio-visual holograms or avatars of other artists. This setup could offer an experience akin to a conventional in-person performance, maintaining both the depth of immersive presence and the standard of musical quality. While these challenges are being solved by the engineering field, a platform such as the "Holodeck" can uncover application-related challenges and serve as a model system upon which to study future-oriented augmented collaborative interaction in laboratory settings. By building a controlled prototype system, we allow research for virtual collaborations to happen several years in advance of their possible future widespread usage. Some example areas of study concern 6DOF audio applications, soundfield rendering, data sonification displays, and of course distributed music performances. The study of complex distributed performance topologies and varying degrees of technological complexity, for example involving headsets and avatar generation, can form the next generation of research in the field.

BIBLIOGRAPHY

AES (2018). *AES New York 2018*. URL: https://www.aes.org/events/145/spatialaudio/?ID=6351 (cit. on p. 40).

Alary, Benoit, Archontis Politis, Sebastian Schlecht, and Vesa Välimäki (2019). "Directional feedback delay network". In: *Journal of the Audio Engineering Society* 67.10, pp. 752–762 (cit. on p. 237).

Algazi, V Ralph, Richard O Duda, Ramani Duraiswami, Nail A Gumerov, and Zhihui Tang (2002). "Approximating the head-related transfer function using simple geometric models of the head and torso". In: *The Journal of the Acoustical Society of America* 112.5, pp. 2053–2064 (cit. on p. 18).

Allen, Jont B and David A Berkley (1979). "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4, pp. 943–950 (cit. on p. 20).

Allison, Paul (2012). "When can you safely ignore multicollinearity". In: *Statistical horizons* 5.1, pp. 1–2 (cit. on p. 177).

Baratè, Adriano, Goffredo Haus, Luca Andrea Ludovico, Elena Pagani, Nello Scarabottolo, et al. (2019). "5G technology for augmented and virtual reality in education". In: *Proceedings of the international conference on education and new developments*. Vol. 2019, pp. 512–516 (cit. on p. 2).

Barbosa, Álvaro (2003). "Displaced soundscapes: A survey of network systems for music and sonic art creation". In: *Leonardo Music Journal*, pp. 53–59 (cit. on p. 24).

Bartlette, Christopher, Dave Headlam, Mark Bocko, and Gordana Velikic (2006a). "Effect of Network Latency on Interactive Musical Performance". In: *Music Perception: An Interdisciplinary Journal* 24.1, pp. 49–62. ISSN: 07307829, 15338312. DOI: 10.1525/mp.2006.24.1.49 (cit. on p. 26).

Bartlette, Christopher, Dave Headlam, Mark Bocko, and Gordana Velikic (2006b). "Effect of network latency on interactive musical performance". In: *Music Perception* 24.1, pp. 49–62 (cit. on p. 150).

Begault, Durand R and Leonard J Trejo (2000). "3-D sound for virtual reality and multimedia". In: (cit. on p. 31).

Begault, Durand R, Elizabeth M Wenzel, and Mark R Anderson (2001). “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source”. In: *Journal of the Audio Engineering Society* 49.10, pp. 904–916 (cit. on p. 22).

Beyer, Tom (2016). “Collaborative Projects in the Performing Arts”. In: (cit. on p. 39).

Biocca, Frank, Chad Harms, and Judee K Burgoon (2003). “Toward a more robust theory and measure of social presence: Review and suggested criteria”. In: *Presence: Teleoperators & virtual environments* 12.5, pp. 456–480 (cit. on pp. 4, 14).

Bishop, Laura and Werner Goebl (2015). “When they listen and when they watch: Pianists’ use of nonverbal audio and visual cues during duet performance”. In: *Musicae Scientiae* 19.1, pp. 84–110 (cit. on p. 31).

Bissonnette, Josiane, Francis Dubé, Martin D Provencher, and Maria T Moreno Sala (2016). “Evolution of music performance anxiety and quality of performance during virtual reality exposure training”. In: *Virtual Reality* 20.1, pp. 71–81 (cit. on p. 27).

Blauert, Jens (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT press (cit. on pp. 18, 31).

Böck, Sebastian and Gerhard Widmer (2013). “Maximum filter vibrato suppression for onset detection”. In: *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*. Vol. 7, p. 4 (cit. on pp. 141, 144).

Boerner, Sabine, Diana Krause, and Diether Gebert (2004). “Leadership and co-operation in orchestras”. In: *Human Resource Development International* 7.4, pp. 465–479 (cit. on p. 87).

Boren, Braxton and **Andrea Genovese** (2018). “Acoustics of virtually coupled performance spaces”. In: *International Conference on Auditory Displays, ICAD*. Georgia Institute of Technology (cit. on pp. 10, 90, 234).

Bregman, Albert S (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press (cit. on pp. 15, 57).

Bui, Cindy, **Andrea Genovese**, Trey Bradley, and Agnieszka Roginska (2020). “Multimodal Immersive Motion Capture (MIMiC): A workflow for musical performance”. In: *Audio Engineering Society Convention 149*. Audio Engineering Society (cit. on pp. 9, 50).

Cáceres, Juan-Pablo and Chris Chafe (2010). “JackTrip: Under the hood of an engine for network audio”. In: *Journal of New Music Research* 39.3, pp. 183–187 (cit. on pp. 24, 25, 42).

Cairns, Patrick (2021). “VIIVA-NMP Audio System: The design of a low latency and naturally interactive Ambisonic audio system for Immersive Network Music Performance”. PhD thesis. University of York (cit. on p. 232).

Cairns, Patrick, Helena Daffern, and Gavin Kearney (2020). “Immersive network music performance: Design and practical deployment of a system for immersive vocal performance”. In: *Audio Engineering Society Convention 149*. Audio Engineering Society (cit. on pp. 237, 238).

Cairns, Patrick, Anthony Hunt, Jacob Cooper, Daniel Johnston, Ben Lee, Helena Daffern, and Gavin Kearney (2022). “Recording music in the metaverse: a case study of XR BBC Maida Vale Recording Studios”. In: *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*. Audio Engineering Society (cit. on p. 2).

Campos-Castillo, Celeste (2012). “Copresence in virtual environments”. In: *Sociology Compass* 6.5, pp. 425–433 (cit. on p. 14).

Carôt, Alexander, Ulrich Krämer, and Gerald Schuller (2006). “Network music performance (NMP) in narrow band networks”. In: *Audio Engineering Society Convention 120*. Audio Engineering Society (cit. on p. 227).

Carôt, Alexander, Pedro Rebelo, and Alain Renaud (2007). “Networked music performance: State of the art”. In: *Audio engineering society conference: 30th international conference: intelligent audio environments*. Audio Engineering Society (cit. on pp. 24, 227).

Carôt, Alexander and Christian Werner (2008). “Distributed network music workshop with soundjack”. In: *Proceedings of the 25th Tonmeistertagung, Leipzig, Germany* (cit. on p. 25).

Carôt, Alexander and Christian Werner (2009). “Fundamentals and Principles of Musical telepresence”. In: *Journal of Science and Technology of the Arts* 1.1, pp. 26–37 (cit. on pp. xxxi, 24, 25, 44, 57, 65, 86).

Carôt, Alexander, Christian Werner, and Timo Fischinger (2009). “Towards a comprehensive cognitive analysis of delay-influenced rhythmical interaction”. In: *ICMC* (cit. on pp. 27, 33, 165, 226, 228, 229).

Chafe, Chris, Juan-Pablo Caceres, and Michael Gurevich (2010). “Effect of temporal separation on synchronization in rhythmic performance”. In: *Perception* 39.7, pp. 982–992 (cit. on pp. 25, 86, 148, 150, 215).

Chafe, Chris, Michael Gurevich, Grace Leslie, and Sean Tyan (2004). “Effect of time delay on ensemble accuracy”. In: *Proceedings of the International Symposium on Musical Acoustics*. Vol. 31, p. 46 (cit. on pp. 25–27, 84, 86, 215, 227).

Chafe, Chris, Scott Wilson, Randal Leistikow, Dave Chisholm, and Gary Scavone (2000). "A simplified approach to high quality music and sound over IP". In: *COST-G6 conference on digital audio effects*. Citeseer, pp. 159–164 (cit. on pp. 24, 28, 29).

Chan, Ian H (2010). "Swept sine chirps for measuring impulse response". In: *Power (dBVrms)* 50.40, p. 30 (cit. on pp. 103, 295).

Charron, Jean-Philippe (2017). "Music audiences 3.0: Concert-goers' psychological motivations at the dawn of virtual reality". In: *Frontiers in psychology* 8, p. 800 (cit. on p. 2).

Cherry, E Colin (1953). "Some experiments on the recognition of speech, with one and with two ears". In: *The Journal of the acoustical society of America* 25.5, pp. 975–979 (cit. on p. 236).

Chew, Elaine, Alexander Sawchuk, Carley Tanoue, and Roger Zimmermann (2005). "Segmental tempo analysis of performances in user-centered experiments in the distributed immersive performance project". In: *Proceedings of the Sound and Music Computing Conference, Salerno, Italy* (cit. on pp. 25, 228).

Clark, Tom S and Drew A Linzer (2015). "Should I use fixed or random effects?" In: *Political science research and methods* 3.2, pp. 399–408 (cit. on p. 158).

Cooley, James, Peter Lewis, and Peter Welch (1967). "Application of the fast Fourier transform to computation of Fourier integrals, Fourier series, and convolution integrals". In: *IEEE Transactions on Audio and Electroacoustics* 15.2, pp. 79–84 (cit. on p. 17).

Dance, SM and BM Shield (1997). "The complete image-source method for the prediction of sound distribution in non-diffuse enclosed spaces". In: *Journal of Sound and Vibration* 201.4, pp. 473–489 (cit. on pp. xxix, 20).

Daniel, Jérôme, Sebastien Moreau, and Rozenn Nicol (2003). "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging". In: *Audio Engineering Society Convention 114*. Audio Engineering Society (cit. on p. 37).

Delle Monache, Stefano, Luca Comanducci, Michele Buccoli, Massimiliano Zanoni, Augusto Sarti, Enrico Pietrocola, Filippo Berbenni, Giovanni Cospito, and Michele Geronazzo (2019). "A presence- and performance-driven framework to investigate interactive networked music learning scenarios". In: *Wireless Communications and Mobile Computing* 2019. ISSN: 15308677. DOI: 10.1155/2019/4593853 (cit. on p. 26).

Earthworks (2022). *Earthworks M30*. URL: https : / / earthworksaudio . com / measurement - microphones/m30/ (cit. on pp. xx, 118).

Eaton, Callum and Hyunkook Lee (2019). “Quantifying factors of auditory immersion in virtual reality”. In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society (cit. on p. 30).

Eaton, James, Nikolay D Gaubitch, Alastair H Moore, and Patrick A Naylor (2015). “The ACE challenge—Corpus description and performance evaluation”. In: IEEE, pp. 1–5 (cit. on p. 34).

Ellis, Daniel PW (2007). “Beat tracking by dynamic programming”. In: *Journal of New Music Research* 36.1, pp. 51–60 (cit. on pp. 141, 146).

Faraway, Julian J (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC (cit. on p. 166).

Farina, Angelo (2000). “Simultaneous measurement of impulse response and distortion with a swept-sine technique”. In: *Audio Engineering Society Convention 108*. Audio Engineering Society (cit. on pp. 17, 103, 106).

Farner, Snorre, Audun Solvang, Asbjørn Sæbo, and U Peter Svensson (2009). “Ensemble hand-clapping experiments under the influence of delay and various acoustic environments”. In: *Journal of the Audio Engineering Society* 57.12, pp. 1028–1041 (cit. on pp. 25, 32, 151, 226–228, 232).

Floridi, Luciano (2005). “The philosophy of presence: From epistemic failure to successful observation”. In: *Presence: Teleoperators & Virtual Environments* 14.6, pp. 656–667 (cit. on p. 14).

Francis, Geoffrey (n.d.). *The Reaper Cockos effects summary guide*. URL: https://www.cockos.com/reaper/guides/ReaEffectsGuide.pdf (cit. on p. 113).

Free online sentiment analysis tool (n.d.). URL: https://monkeylearn.com/sentiment-analysis-online/ (cit. on p. 225).

Friberg, Johnny and Dan Gärdenfors (2004). “Audio games: new perspectives on game audio”. In: *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*. ACM, pp. 148–154 (cit. on p. 16).

Gamper, Hannes and Ivan J. Tashev (2018). “Blind Reverberation Time Estimation Using A Convolutional Neural Network”. In: *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)* (cit. on p. 34).

Andrea Genovese, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J Tashev (2019a). “Blind room volume estimation from single-channel noisy speech”. In: *ICASSP 2019-2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 231–235 (cit. on p. 9).

Andrea Genovese, Marta Gospodarek, and Agnieszka Roginska (2019b). "Mixed realities: a live collaborative musical performance". In: *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019; 5th International Conference on Spatial Audio; September 26th to 28th, 2019, Ilmenau, Germany*, pp. 159–164 (cit. on pp. 8, 50, 233).

Andrea Genovese and Agnieszka Roginska (2019). "Hmdir: An hrtf dataset measured on a mannequin wearing xr devices". In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society (cit. on pp. 9, 233).

Andrea Genovese, Gabriel Zalles, Gregory Reardon, and Agnieszka Roginska (2018). "Acoustic perturbations in HRTFs measured on Mixed Reality Headsets". In: *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society (cit. on pp. 9, 34, 233).

Ghezzo, D, J Gilbert, A Smith, and S Jacobson (n.d.). "The Cassandra Project. 1996". In: *For more information, see http://www. nyu. edu/pages/ngc/ipg/cassandra* () (cit. on p. 39).

Gilmour, Arthur R, Robin Thompson, and Brian R Cullis (1995). "Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models". In: *Biometrics*, pp. 1440–1450 (cit. on p. 160).

Gochfeld, David, Corinne Brenner, Kris Layng, Sebastian Herscher, Connor DeFanti, Marta Olko, David Shinn, Stephanie Riggs, Clara Fernández-Vara, and Ken Perlin (2018). "Holojam in Wonderland: Immersive Mixed Reality Theater". In: *Leonardo* 51.4, pp. 362–367. DOI: 10.1162/leon_a_01644. URL: https://doi.org/10.1162/leon_a_01644 (cit. on p. 13).

Gospodarek, Marta, **Andrea Genovese**, Dennis Dembeck, Corinne Brenner, Agnieszka Roginska, and Ken Perlin (2019). "Sound design and reproduction techniques for co-located narrative VR experiences". In: *Audio Engineering Society Convention 147*. Audio Engineering Society (cit. on p. 9).

Grosche, Peter, Meinard Müller, and Craig Stuart Sapp (2010). "What Makes Beat Tracking Difficult? A Case Study on Chopin Mazurkas." In: *ISMIR*, pp. 649–654 (cit. on p. 87).

Gurevich, Michael, Dónal Donohoe, and Stéphanie Bertet (2011). "Ambisonic Spatialization for Networked Music Performance". In: International Community for Auditory Display (cit. on p. 29).

Habets, Emanuel AP (2006). “Room impulse response generator”. In: *Technische Universiteit Eindhoven, Tech. Rep* 2.2.4, p. 1 (cit. on p. 21).

Harpe, Spencer E (2015). “How to analyze Likert and other rating scale data”. In: *Currents in pharmacy teaching and learning* 7.6, pp. 836–850 (cit. on p. 165).

Heeter, Carrie (1992). “Being there: The subjective experience of presence”. In: *Presence: Teleoperators & Virtual Environments* 1.2, pp. 262–271 (cit. on pp. xxix, 14).

Hendrix, Claudia and Woodrow Barfield (1996). “The sense of presence within auditory virtual environments”. In: *Presence: Teleoperators & Virtual Environments* 5.3, pp. 290–301 (cit. on p. 232).

Herre, Jürgen, Johannes Hilpert, Achim Kuntz, and Jan Plogsties (2015). “MPEG-H 3D audio—The new standard for coding of immersive spatial audio”. In: *IEEE Journal of selected topics in signal processing* 9.5, pp. 770–779 (cit. on pp. 22, 37).

Hirsh, Ira J (1959). “Auditory perception of temporal order”. In: *The Journal of the Acoustical Society of America* 31.6, pp. 759–767 (cit. on p. 25).

Holodeck - Experential Supercomputer (2017). URL: https://holodeck.nyu.edu/ (cit. on pp. 36, 38).

HOLOJAM (2014). URL: https://frl.nyu.edu/wonderland/ (cit. on p. 13).

Howard, David M and Jamie Angus (2017). *Acoustics and psychoacoustics*. Focal press (cit. on p. 21).

Hummersone, Chris (2017). *IOSR+ MATLAB Toolbox*. URL: https://github.com/IoSR-Surrey/MatlabToolbox (cit. on pp. xi, 104).

Hupke, Robert, Lucas Beyer, Marcel Nophut, Stephan Preihs, and Jürgen Peissig (2019a). “Effect of a global metronome on ensemble accuracy in networked music performance”. In: *Audio Engineering Society Convention 147*. Audio Engineering Society (cit. on pp. 54, 86).

Hupke, Robert, **Andrea Genovese**, Sripathi Sridhar, Jürgen Peissig, and Agnieszka Roginska (2020). “Impact of Source Panning on a Global Metronome in Rhythmic Networked Music Performance”. In: *2020 27th Conference of Open Innovations Association (FRUCT)*. IEEE, pp. 73–83 (cit. on pp. 9, 55, 58, 86, 227, 232).

Hupke, Robert, Stephan Preihs, and Juergen Peissig (2022). “Immersive Room Extension Environment for Networked Music Performance”. In: *Audio Engineering Society Convention 153*. Audio Engineering Society (cit. on p. 32).

Hupke, Robert, Sripathi Sridhar, **Andrea Genovese**, Marcel Nophut, Stephan Preihs, Tom Beyer, Agnieszka Roginska, and Jürgen Peissig (2019b). "A latency measurement method for networked music performances". In: *Audio Engineering Society Convention 147*. Audio Engineering Society (cit. on pp. 9, 55).

Ibnyahya, Ilias and Joshua D Reiss (2022). "A Method for matching room impulse responses with feedback delay networks". In: *Audio Engineering Society Convention 153*. Audio Engineering Society (cit. on p. 237).

Jot, Jean-Marc (1997). "Efficient models for reverberation and distance rendering in computer music and virtual audio reality." In: *ICMC* (cit. on p. 65).

Jot, Jean-Marc, Laurent Cerveau, and Olivier Warusfel (1997). "Analysis and synthesis of room reverberation based on a statistical time-frequency model". In: *Audio Engineering Society Convention 103*. Audio Engineering Society (cit. on p. 21).

Jot, Jean-Marc and Antoine Chaigne (1991). "Digital delay networks for designing artificial reverberators". In: *Audio Engineering Society Convention 90*. Audio Engineering Society (cit. on p. 237).

Jot, Jean-Marc, Véronique Larcher, and Jean-Marie Pernaux (1999). "A comparative study of 3-D audio encoding and rendering techniques". In: *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Audio Engineering Society (cit. on p. 20).

Jot, Jean-Marc and Keun Sup Lee (2016). "Augmented Reality Headphone Environment Rendering". In: *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society (cit. on p. 51).

Jung, Byungdae, Jaein Hwang, Sangyoon Lee, Gerard Jounghyun Kim, and Hyunbin Kim (2000). "Incorporating co-presence in distributed virtual music environment". In: *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 206–211 (cit. on pp. 57, 226, 232).

Katz, Brian FG (2001). "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation". In: *The Journal of the Acoustical Society of America* 110.5, pp. 2440–2448 (cit. on p. 21).

Kendall, Gary S (1995). "A 3-D sound primer: directional hearing and stereo reproduction". In: *Computer music journal* 19.4, pp. 23–46 (cit. on p. 15).

Klein, Florian, Stephan Werner, and Thomas Mayenfels (2017). "Influences of training on externalization of binaural synthesis in situations of room divergence". In: *Journal of the Audio Engineering Society* 65.3, pp. 178–187 (cit. on pp. 33, 66, 226, 228).

Kleiner, Mendel, Bengt-Inge Dalenbäck, and Peter Svensson (1993). “Auralization-an overview”. In: *Journal of the Audio Engineering Society* 41.11, pp. 861–875 (cit. on p. 16).

Kuha, Jouni (2004). “AIC and BIC: Comparisons of assumptions and performance”. In: *Sociological methods & research* 33.2, pp. 188–229 (cit. on p. 159).

Kuttruff, Heinrich (2014). *Room acoustics*. Crc Press (cit. on pp. 20, 32).

Kyriakakis, Chris (1998). “Fundamental and technological limitations of immersive audio systems”. In: *Proceedings of the IEEE* 86.5, pp. 941–951 (cit. on p. 35).

Lance, Charles E, Marcus M Butts, and Lawrence C Michels (2006). “The sources of four commonly reported cutoff criteria: What did they really say?” In: *Organizational research methods* 9.2, pp. 202–220 (cit. on p. 189).

Lavoie, Michel C, Scott G Norcross, and Gilbert A Soulodre (2004). “Distortion Audibility in Inverse Filtering”. In: *Audio Engineering Society Convention 117*. Audio Engineering Society (cit. on p. 106).

Lee, Hyunkook (2020). “A conceptual model of immersive experience in extended reality”. In: (cit. on pp. 4, 14, 23, 29, 30, 68, 123, 227–230, 238).

Lessiter, Jane, Jonathan Freeman, Edmund Keogh, and Jules Davidoff (2001). “A cross-media presence questionnaire: The ITC-Sense of Presence Inventory”. In: *Presence: Teleoperators & Virtual Environments* 10.3, pp. 282–297 (cit. on p. 14).

Lindau, Alexander and Stefan Weinzierl (2012). “Assessing the plausibility of virtual acoustic environments”. In: *Acta Acustica united with Acustica* 98.5, pp. 804–810 (cit. on pp. xxix, 23, 31, 35, 226, 229, 230).

Litovsky, Ruth Y (2012). “Spatial release from masking”. In: *Acoust. Today* 8.2, pp. 18–25 (cit. on p. 236).

Lobser, David, Ken Perlin, Lily Fang, and Christopher Romero (2017). “FLOCK: a location-based, multi-user VR experience”. In: *ACM SIGGRAPH 2017 VR Village*. ACM, p. 6 (cit. on p. 13).

Lombard, Matthew and Theresa Ditton (1997). “At the heart of it all: The concept of presence”. In: *Journal of Computer-Mediated Communication* 3.2 (cit. on pp. xxix, 14).

Lombard, Matthew, Theresa B Ditton, and Lisa Weinstein (2009). “Measuring presence: the temple presence inventory”. In: *Proceedings of the 12th annual international workshop on presence*, pp. 1–15 (cit. on pp. 14, 31, 230).

Loveridge, Ben (2020). “Networked music performance in virtual reality: current perspectives”. In: *Journal of Network Music and Arts* 2.1, p. 2 (cit. on p. 2).

Mania, Katerina, Bernard D Adelstein, Stephen R Ellis, and Michael I Hill (2004). “Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity”. In: *Proceedings of the 1st Symposium on Applied Perception in Graphics and Visualization*, pp. 39–47 (cit. on p. 233).

Mantovani, Giuseppe and Giuseppe Riva (1999). ““Real” presence: how different ontologies generate different criteria for presence, telepresence, and virtual presence”. In: *Presence* 8.5, pp. 540–550 (cit. on pp. 15, 227).

Marden, John I (2004). “Positions and QQ plots”. In: *Statistical Science*, pp. 606–614 (cit. on p. 160).

Masiero, Bruno and Janina Fels (2011). “Perceptually robust headphone equalization for binaural reproduction”. In: *Audio Engineering Society Convention 130*. Audio Engineering Society (cit. on p. 106).

Mason, Robin (1994). *Using communications media in open and flexible learning*. Psychology Press (cit. on p. 4).

McFee, Brian, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto (2015). “librosa: Audio and music signal analysis in python”. In: *Proceedings of the 14th python in science conference*. Vol. 8, pp. 18–25 (cit. on pp. 141, 144, 146).

McPherson, Andrew P, Robert H Jack, Giulio Moro, et al. (2016). “Action-sound latency: Are our tools fast enough?” In: (cit. on pp. 28, 236).

Mehra, Ravish, Nikunj Raghuvanshi, Lauri Savioja, Ming C Lin, and Dinesh Manocha (2012). “An efficient GPU-based time domain solver for the acoustic wave equation”. In: *Applied Acoustics* 73.2, pp. 83–94 (cit. on p. 21).

Merimaa, Juha and Ville Pulkki (2005). “Spatial impulse response rendering I: Analysis and synthesis”. In: *Journal of the Audio Engineering Society* 53.12, pp. 1115–1127 (cit. on p. 236).

Milgram, Paul and Fumio Kishino (1994). “A taxonomy of mixed reality visual displays”. In: *IEICE Transactions on Information and Systems* 77.12, pp. 1321–1329 (cit. on pp. xxix, 12, 34, 52).

Milgram, Paul, Haruo Takemura, Akira Utsumi, and Fumio Kishino (1995). “Augmented reality: A class of displays on the reality-virtuality continuum”. In: *Telemanipulator and telepresence technologies*. Vol. 2351. International Society for Optics and Photonics, pp. 282–293 (cit. on pp. xv, 13).

Minsky, Marvin (1980). “Telepresence”. In: (cit. on p. 11).

Mulcahy, John (2022). *REW room acoustics and Audio Device Measurement and Analysis Software*. URL: https://www.roomeqwizard.com/ (cit. on p. 116).

Murray, Janet Horowitz (2017). *Hamlet on the holodeck: The future of narrative in cyberspace*. MIT press (cit. on p. 12).

Neath, Andrew A and Joseph E Cavanaugh (2012). “The Bayesian information criterion: background, derivation, and applications”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2, pp. 199–203 (cit. on p. 159).

Nowak, Kristine (2001). “Defining and differentiating copresence, social presence and presence as transportation”. In: *Presence 2001 Conference, Philadelphia, PA*. Vol. 2. Citeseer, pp. 686–710 (cit. on p. 4).

NYU Corelink | Homepage (n.d.). Accessed: 2023-03-14. URL: https://corelink.hsrn.nyu.edu/ (cit. on pp. 13, 37, 50).

Ohta, Yuichi and Hideyuki Tamura (2014). *Mixed reality: merging real and virtual worlds*. Springer Publishing Company, Incorporated (cit. on p. 11).

Olko, Marta, Dennis Dembeck, Yun-Han Wu, **Andrea Genovese**, and Agnieszka Roginska (2017). “Identification of Perceived Sound Quality Attributes of 360° Audiovisual Recordings in VR Using a Free Verbalization Method”. In: *Audio Engineering Society Convention 143*. Audio Engineering Society (cit. on p. 23).

Olmos, Adriana, Mathieu Brulé, Nicolas Bouillot, Mitchel Benovoy, Jeff Blum, Haijian Sun, Niels Windfeld Lund, and Jeremy R Cooperstock (2009). “Exploring the role of latency and orchestra placement on the networked performance of a distributed opera”. In: *12th annual international workshop on presence, Los Angeles* (cit. on pp. 27, 31, 228).

Ozawa, Kenji, Yoshihiro Chujo, Yôiti Suzuki, and Toshio Sone (2003a). “Psychological factors involved in auditory presence”. In: *Acoustical Science and Technology* 24.1, pp. 42–44 (cit. on p. 24).

Ozawa, Kenji, Satoshi Ohtake, Yôiti Suzuki, and Toshio Sone (2003b). “Effects of visual information on auditory presence”. In: *Acoustical Science and Technology* 24.2, pp. 97–99 (cit. on p. 24).

Paradiso, Joseph and Flavia Sparacino (1997). “Optical tracking for music and dance performance”. In: *Optical 3-D Measurement Techniques IV*, pp. 11–18 (cit. on p. 31).

Pasanen, Jaakko (2020). *AutoEQ - Github Repository*. URL: https://github.com/jaakkopasanen/AutoEq (cit. on p. 106).

Perlin, Ken (2016). "Future Reality: How emerging technologies will change language itself". In: *IEEE computer graphics and applications* 36.3, pp. 84–89 (cit. on p. 13).

Pike, Chris, Frank Melchior, and Tony Tew (2014). "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room". In: *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society (cit. on pp. 23, 226).

Pinheiro, José C and Douglas M Bates (2000). "Linear mixed-effects models: basic concepts and examples". In: *Mixed-effects models in S and S-Plus*, pp. 3–56 (cit. on p. 157).

Plass, Jan L, Ken Perlin, Agnieszka Roginska, Chris Hovey, Fabian Fröhlich, Aniol Saurina Maso, Alvaro Olsen, Zhenyi He, Robert Pahle, and Sounak Ghosh (2022). "Designing Effective Playful Collaborative Science Learning in VR". In: *Serious Games: Joint International Conference, JCSG 2022, Weimar, Germany, September 22–23, 2022, Proceedings*. Springer, pp. 30–35 (cit. on pp. 13, 36).

Pralong, Danièle and Simon Carlile (1996). "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space". In: *The Journal of the Acoustical Society of America* 100.6, pp. 3785–3793 (cit. on p. 105).

Pulkki, Ville (2007). "Spatial sound reproduction with directional audio coding". In: *Journal of the Audio Engineering Society* 55.6, pp. 503–516 (cit. on p. 15).

Raghuvanshi, Nikunj, Rahul Narain, and Ming C. Lin (2009). "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition". In: *IEEE Transactions on Visualization and Computer Graphics* 15.5, pp. 789–801. ISSN: 1077-2626. DOI: 10.1109/tvcg.2009.28. URL: http://dx.doi.org/10.1109/tvcg.2009.28 (cit. on p. 21).

Rämö, Jussi and Vesa Välimäki (2012). "Signal processing framework for virtual headphone listening tests in a noisy environment". In: *Audio Engineering Society Convention 132*. Audio Engineering Society (cit. on p. 105).

REAPER, Digital Audio Workstation (n.d.). URL: https://www.reaper.fm/ (cit. on p. 113).

Reardon, Gregory, **Andrea Genovese**, Gabriel Zalles, Patrick Flanagan, and Agnieszka Roginska (2018a). "Evaluation of Binaural Renderers: Multidimensional Sound Quality Assessment". In: *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society (cit. on p. 23).

Reardon, Gregory, Gabriel Zalles, Andrea Genovese, Patrick Flanagan, and Agnieszka Roginska (2018b). “Evaluation of Binaural Renderers: Externalization”. In: *Audio Engineering Society Convention 144*. Audio Engineering Society (cit. on p. 229).

Reardon, Gregory, Gabriel Zalles, **Andrea Genovese**, Patrick Flanagan, and Agnieszka Roginska (2018c). “Evaluation of Binaural Renderers: Externalization”. In: *Audio Engineering Society Convention 144*. Audio Engineering Society (cit. on p. 23).

Reich, Steve and Russ Hartenberger (1980). *Clapping music*. Vol. 1. Universal Edition London (cit. on p. 87).

Reichinger, Andreas, Piotr Majdak, Robert Sablatnig, and Stefan Maierhofer (2013). “Evaluation of methods for optical 3-D scanning of human pinnas”. In: *2013 international conference on 3D vision-3DV 2013*. IEEE, pp. 390–397 (cit. on p. 28).

Riva, Giuseppe, F Davide, and WA IJsselsteijn (2003). “Being there: The experience of presence in mediated environments”. In: *Being there: Concepts, effects and measurement of user presence in synthetic environments* 5 (cit. on pp. xxix, 14, 34, 67).

Robillard, G, S Bouchard, P Renaud, and LG Cournoyer (2002). “Validation canadienne-française de deux mesures importantes en réalité virtuelle: l’Immersive Tendencies Questionnaire et le Presence Questionnaire”. In: *Poster presented at the 25e congrès annuel de la Société Québécoise pour la Recherche en Psychologie (SQRP), Trois-Rivières* (cit. on p. 27).

Roginska, Agnieszka and Paul Geluso (2017). *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*. Taylor & Francis (cit. on pp. 15, 89, 232).

Rottondi, Cristina, Michele Buccoli, Massimiliano Zanoni, Dario Garao, Giacomo Verticale, and Augusto Sarti (2015). “Feature-Based Analysis of the Effects of Packet Delay on Networked Musical Interactions”. In: *J. Audio Eng. Soc* 63.11, pp. 864–875 (cit. on pp. 25, 235).

Rottondi, Cristina, Chris Chafe, Claudio Allocchio, and Augusto Sarti (2016). “An overview on networked music performance technologies”. In: *IEEE Access* 4, pp. 8823–8843 (cit. on pp. 26, 57, 65, 84, 146, 148, 215, 227).

Rumsey, Francis (2002). “Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm”. In: *Journal of the Audio Engineering Society* 50.9, pp. 651–666 (cit. on pp. 23, 35, 65, 230).

Sahai, Hardeo and Mohammed I Ageel (2012). *The analysis of variance: fixed, random and mixed models*. Springer Science & Business Media (cit. on p. 161).

Sawchuk, Alexander A., Elaine Chew, Roger Zimmermann, Christos Papadopoulos, and Chris Kyriakakis (2003). “From remote media immersion to distributed immersive performance”. In: *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence*, pp. 110–120 (cit. on p. 26).

Schärer, Zora and Alexander Lindau (2009). “Evaluation of equalization methods for binaural signals”. In: *Audio Engineering Society Convention 126*. Audio Engineering Society (cit. on pp. 105, 106).

Scheirer, Eric and Malcolm Slaney (1997). “Construction and evaluation of a robust multifeature speech/music discriminator”. In: *1997 IEEE international conference on acoustics, speech, and signal processing*. Vol. 2. IEEE, pp. 1331–1334 (cit. on p. 296).

Schimmel, Steven M, Martin F Muller, and Norbert Dillier (2009). “A fast and accurate “shoebox” room acoustics simulator”. In: *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, pp. 241–244 (cit. on pp. xv, 17).

Schlagowski, Ruben, Kunal Gupta, Silvan Mertes, Mark Billinghurst, Susanne Metzner, and Elisabeth André (2022). “Jamming in MR: towards real-time music collaboration in mixed reality”. In: *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, pp. 854–855 (cit. on p. 2).

Schreiber, Hendrik and Meinard Müller (2014). “Exploiting global features for tempo octave correction”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 639–643 (cit. on p. 231).

Schröder, Manfred R (1975). “Diffuse sound reflection by maximum- length sequences”. In: *The Journal of the Acoustical Society of America* 57.1, pp. 149–150 (cit. on p. 103).

Schroeder, Franziska, Alain B Renaud, Pedro Rebelo, and Fernando Gualda (2007). “Addressing the network: Performative strategies for playing apart”. In: *In International Computer Music Conference ICMC*. Citeseer (cit. on p. 32).

Schroeder, Manfred R (1962). “Natural sounding artificial reverberation”. In: *Journal of the Audio Engineering Society* 10.3, pp. 219–223 (cit. on p. 20).

Schroeder, Manfred R (1979). “Integrated-impulse method measuring sound decay without using impulses”. In: *The Journal of the Acoustical Society of America* 66.2, pp. 497–500 (cit. on p. 104).

Shahab, Mojtaba, Alireza Taheri, Mohammad Mokhtari, Azadeh Shariati, Rozita Heidari, Ali Meghdari, and Minoo Alemi (2022). “Utilizing social virtual reality robot (V2R) for

music education to children with high-functioning autism". In: *Education and Information Technologies*, pp. 1–25 (cit. on p. 2).

Shu, Yu, Yen-Zhang Huang, Shu-Hsuan Chang, and Mu-Yen Chen (2019). "Do virtual reality head-mounted displays make a difference? A comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments". In: *Virtual Reality* 23, pp. 437–446 (cit. on p. 32).

Slater, Mel and Anthony Steed (2000). "A virtual presence counter". In: *Presence: Teleoperators & Virtual Environments* 9.5, pp. 413–434 (cit. on p. 14).

Smith, Lindsay I (2002). *A tutorial on principal components analysis*. Tech. rep. (cit. on p. xxxi).

Sondhi, M Mohan, Dennis R Morgan, and Joseph L Hall (1995). "Stereophonic acoustic echo cancellation-an overview of the fundamental problem". In: *IEEE Signal processing letters* 2.8, pp. 148–151 (cit. on p. 65).

Spotify (2021). *Spotify/Pedalboard: a python library for working with audio*. URL: https://github.com/spotify/pedalboard (cit. on pp. 144, 154).

Steuer, Jonathan (1992). "Defining virtual reality: Dimensions determining telepresence". In: *Journal of communication* 42.4, pp. 73–93 (cit. on pp. xxix, 34).

Thery, David, David Poirier-Quinot, Barteld NJ Postma, and Brian FG Katz (2017). "Impact of the visual rendering system on subjective auralization assessment in VR". In: *Virtual Reality and Augmented Reality: 14th EuroVR International Conference, EuroVR 2017, Laval, France, December 12–14, 2017, Proceedings 14*. Springer, pp. 105–118 (cit. on p. 32).

Toet, Alexander, Tina Mioch, Simon NB Gunkel, Omar Niamut, and Jan BF van Erp (2021). "Holistic Framework for Quality Assessment of Mediated Social Communication". In: (cit. on p. 238).

Torger, Anders and Angelo Farina (2001). "Real-time partitioned convolution for Ambiophonics surround sound". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, pp. 195–198 (cit. on p. 237).

Tosi, Jacopo, Fabrizio Taffoni, Marco Santacatterina, Roberto Sannino, and Domenico Formica (2017). "Performance evaluation of bluetooth low energy: A systematic review". In: *Sensors* 17.12, p. 2898 (cit. on p. 28).

Treurniet, Jan Jaap, Chayan Sarkar, R Venkatesha Prasad, and Willem De Boer (2015). "Energy consumption and latency in BLE devices under mutual interference: An experimental study".

In: *2015 3rd International Conference on Future Internet of Things and Cloud*. IEEE, pp. 333–340 (cit. on p. 28).

Tsingos, Nicolas (2017). "Object-Based Audio". In: *Immersive Sound*. Focal Press, pp. 258–289 (cit. on p. 22).

Tsioutas, Konstantinos, George Xylomenos, Ioannis Doumanis, and Christos Angelou (2020). "Quality of musicians' experience in network music performance: A subjective evaluation". In: *Audio Engineering Society Convention 148*. Audio Engineering Society (cit. on p. 238).

Turchet, Luca, György Fazekas, Mathieu Lagrange, Hossein S Ghadikolaei, and Carlo Fischione (2020). "The internet of audio things: State of the art, vision, and challenges". In: *IEEE internet of things journal* 7.10, pp. 10233–10249 (cit. on pp. 32, 243).

Turchet, Luca, Carlo Fischione, Georg Essl, Damián Keller, and Mathieu Barthet (2018). "Internet of musical things: Vision and challenges". In: *Ieee access* 6, pp. 61994–62017 (cit. on pp. 32, 243).

Ueno, Kanako, Kosuke Kato, and Keiji Kawai (2010). "Effect of room acoustics on musicians' performance. Part I: Experimental investigation with a conceptual model". In: *Acta Acustica united with Acustica* 96.3, pp. 505–515 (cit. on p. 226).

Valente, Daniel L and Jonas Braasch (2010). "Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues". In: *The Journal of the Acoustical Society of America* 128.4, pp. 1952–1964 (cit. on p. 31).

Väljamäe, Er, Pontus Larsson, Daniel Västfjäll, and Mendel Kleiner (2004). "Auditory presence, individualized head-related transfer functions, and illusory ego-motion in virtual environments". In: *in in Proc. of Seventh Annual Workshop Presence 2004*. Citeseer (cit. on p. 24).

Vanasse, Julian, **Andrea Genovese**, and Agnieszka Roginska (2019). "Multichannel impulse response measurements in MATLAB: An update on scanIR". In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society (cit. on pp. 10, 103, 116).

Wagner, Ina, Wolfgang Broll, Giulio Jacucci, Kari Kuutii, Rod McCall, Ann Morrison, Dieter Schmalstieg, and Jean-Jacques Terrin (2009). "On the role of presence in mixed reality". In: *Presence: Teleoperators and Virtual Environments* 18.4, pp. 249–276 (cit. on pp. 12, 15, 35).

Werner, Stephan, Florian Klein, Thomas Mayenfels, and Karlheinz Brandenburg (2016). "A summary on acoustic room divergence and its effect on externalization of auditory events". In: *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, pp. 1–6 (cit. on pp. 32, 33, 66, 67, 74, 214, 225, 226).

Witmer, Bob G and Michael J Singer (1998). “Measuring presence in virtual environments: A presence questionnaire”. In: *Presence* 7.3, pp. 225–240 (cit. on pp. 14, 238).

Wycisk, Yves, Kilian Sander, Reinhard Kopiez, Friedrich Platz, Jakob Bergner, Stephan Preihs, and Jürgen Peissig (2021). “Wrapped Into Sound: Development of the Immersive Audio Quality Inventory (IAQI)”. In: (cit. on p. 238).

Xia, Haijun, Sebastian Herscher, Ken Perlin, and Daniel Wigdor (2018). “Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality”. In: *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, pp. 853–866 (cit. on p. 13).

Xu, Aoxiang, Wieslaw Woszczyk, Zack Settel, Bruce Pennycook, Robert Rowe, Philip Galanter, and Jeffreyn Bary (2000). “Real-time streaming of multichannel audio data over Internet”. In: *Journal of the Audio Engineering Society* 48.7/8, pp. 627–641 (cit. on p. 25).

Zahorik, Pavel and Rick L Jenison (1998). “Presence as being-in-the-world”. In: *Presence* 7.1, pp. 78–89 (cit. on pp. 15, 31, 227).

Zea, Elías (2012). “Binaural In-Ear Monitoring of acoustic instruments in live music performance”. In: *15th International Conference on Digital Audio Effects, DAFx 2012, 17 September 2012 through 21 September 2012, York*, pp. 1–8 (cit. on p. 28).

Zhao, Shanyang (2003). “Toward a taxonomy of copresence”. In: *Presence* 12.5, pp. 445–455 (cit. on pp. 4, 140).

APPENDICES

1 Holodeck Concert - Second Pilot Questionnaires

1.1 Audience Questionnaire

What is your profession (i.e., title at work, or student)?

_____

How many years of professional experience with Audio, Sound, Music and Tech Engineering do you have? _____

To what extent did you feel as though **the choir was physically present in the auditorium?**

Not at all **1 2 3 4 5 6 7** I felt exactly like the choir was present

To what extent were the performances of the **remote musicians and musicians on stage cohesive (like they were playing together)?**

Not at all cohesive **1 2 3 4 5 6 7** Completely cohesive

To what extent was the performance of the **remote dancers, and dancers and musicians on stage, cohesive?**

Not at all cohesive **1 2 3 4 5 6 7** Completely cohesive

If you had to choose, which piece was your favorite?

1. "Ecotone"
2. "Dancer in the Dark"
3. "No Comment"
4. "I've Always Hated Watching You Leave" *(with choir)*
5. "Wir Schaffen das" *(with dancers)*
6. "Fernweh" *(dancers and remote musicians)*
7. "Walking the Dead" *(choir, dancers, distributed musicians)*

Why was this piece your favorite?

_____

_____

In what way, if any, did the distributed technology impact the performance?

Very Negatively	1	2	3	4 Not at all	5	6	7	Very Positively

Overall, what was your impression of this concert in terms of…

	Very Negative						Very Positive
All **audio** components	1	2	3	4	5	6	7
All **visual** components	1	2	3	4	5	6	7
Audience Experience	1	2	3	4	5	6	7

Are there any other comments you would like to share?

1.2 Performer's Questionnaire

PDF printed from Google Forms

Performer Questionnaire

If you participated in the HoloDeck concert we would love to hear about your experience as a performer. Please fill the questions below to help us understand the impact of the technology on the performance.

* Required

1. **Email address** *

2. **Name** *

3. **What was your role in the October 18th, 2018 concert?** *
   Check all that apply.
   - [ ] Live Musician (on stage at Loewe Theatre)
   - [ ] Remote Musician
   - [ ] Live Dancer (on stage at Loewe Theatre)
   - [ ] Remote Dancer
   - [ ] Another role

4. **Where did you perform?** *
   Check all that apply.
   - [ ] Theatre
   - [ ] Studio
   - [ ] Dance Studio
   - [ ] Other: _____

5. **How many years of professional experience with Audio, Sound, Dance, Music and Tech Engineering do you have ?** *
   Please reply with a single number.

6. **Have you ever participated in a distributed concert before (performers in multiple remote locations)?** *
   Mark only one oval.
   - ( ) Yes
   - ( ) No

7. **For how many hours, if any, did you rehearse before the concert?** *
   Please reply with the number of hours you spent rehearsing. If you didn't rehearse, enter "0".

8. **For how many hours, if any, did you rehearse using a remote connection?** *

9. **How much did latency (signal delay) impact the quality of your performance?** *
   Mark only one oval.

	1	2	3	4	5	6	7	
Had no impact at all	◯	◯	◯	◯	◯	◯	◯	Had a major impact on my performance

10. **If latency affected your performance at all, (technically, creatively, emotionally) please describe your experience:**

11. **Which piece(s) did you participate in? Which one did you feel most comfortable participating in?**
    Check all that apply.

	Participated	Most Comfortable
Ecotone	☐	☐
Dancer in the Dark	☐	☐
No Comment	☐	☐
I have always hated watching you leave	☐	☐
Wir Schaffen Das	☐	☐
Fernweh	☐	☐
Walking the dead (Finale)	☐	☐

12. **To what extent did you feel that you and the other performers were connected and performing in the same space?** *
    Mark only one oval.

	1	2	3	4	5	6	7	
Felt like completely different spaces	◯	◯	◯	◯	◯	◯	◯	Felt like perfect integration in the same space

13. **What type of monitoring system did you use for this performance? (Select all that apply)** *
Check all that apply.

- [ ] Earpiece Monitor (Audio)
- [ ] Headphones (Audio)
- [ ] Loudspeakers (Audio)
- [ ] Digital rendering (Visual)
- [ ] Regular Video (Visual)
- [ ] Other: _____

14. **If the monitoring system impacted your ability to feel immersed in the performance, what was the impact? If the monitoring system had no impact, please choose '4'.** *
Mark only one oval.

	1	2	3	4	5	6	7	
Negative impact	◯	◯	◯	◯	◯	◯	◯	Positive impact

15. **As an artist, how enjoyable was the overall experience to you?** *
Mark only one oval.

	1	2	3	4	5	6	7	
Not enjoyable at all	◯	◯	◯	◯	◯	◯	◯	Incredibly enjoyable

16. **Please share any other comment, sensation or additional statements here:**

_____

_____

_____

_____

_____

2 Room Measurement Details

2.1 Baseline Live Room

The following plots concern the time-frequency fingerprint of the room used during the Baseline phase of the main co-presence study (“Dolan’s Live Room”).

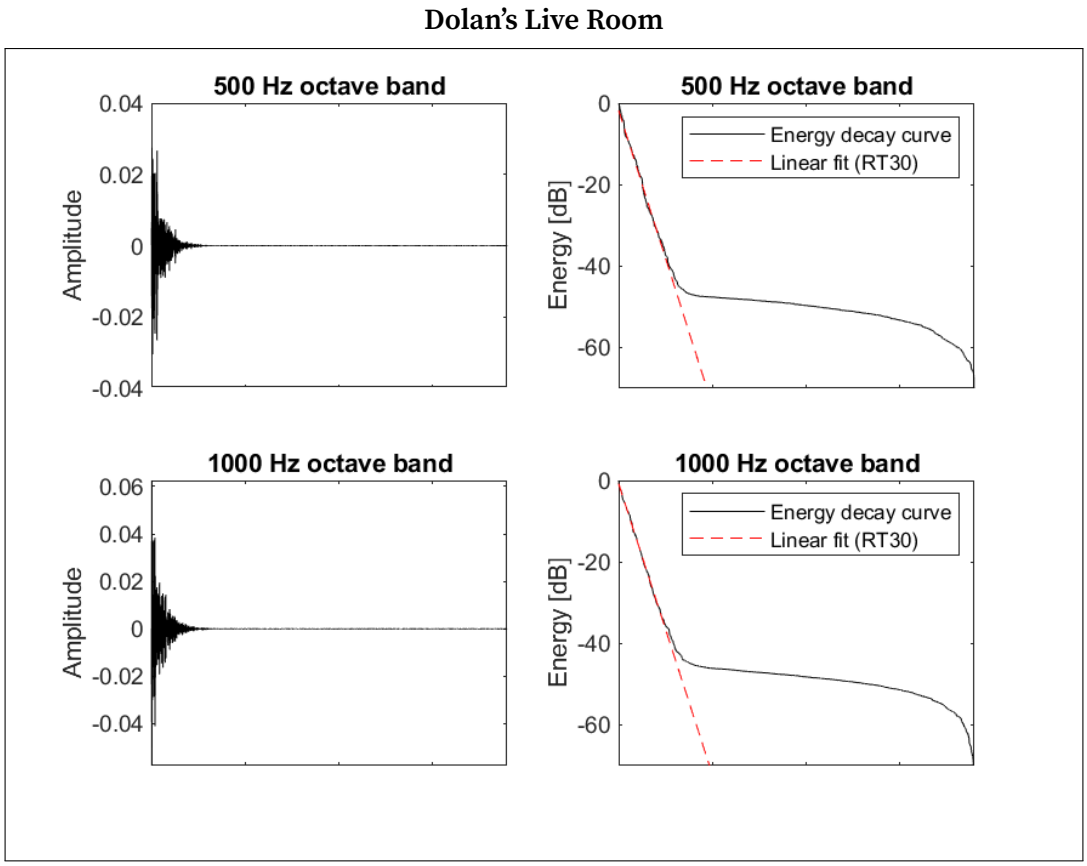


Figure 98: RT30 fit of the Live Room (“Dolan”) at 500 Hz and 1 kHz. Used for the baseline study of the co-presence experiment. Taken from the omnidirectional measurement of an impulse response from a source at 8ft.

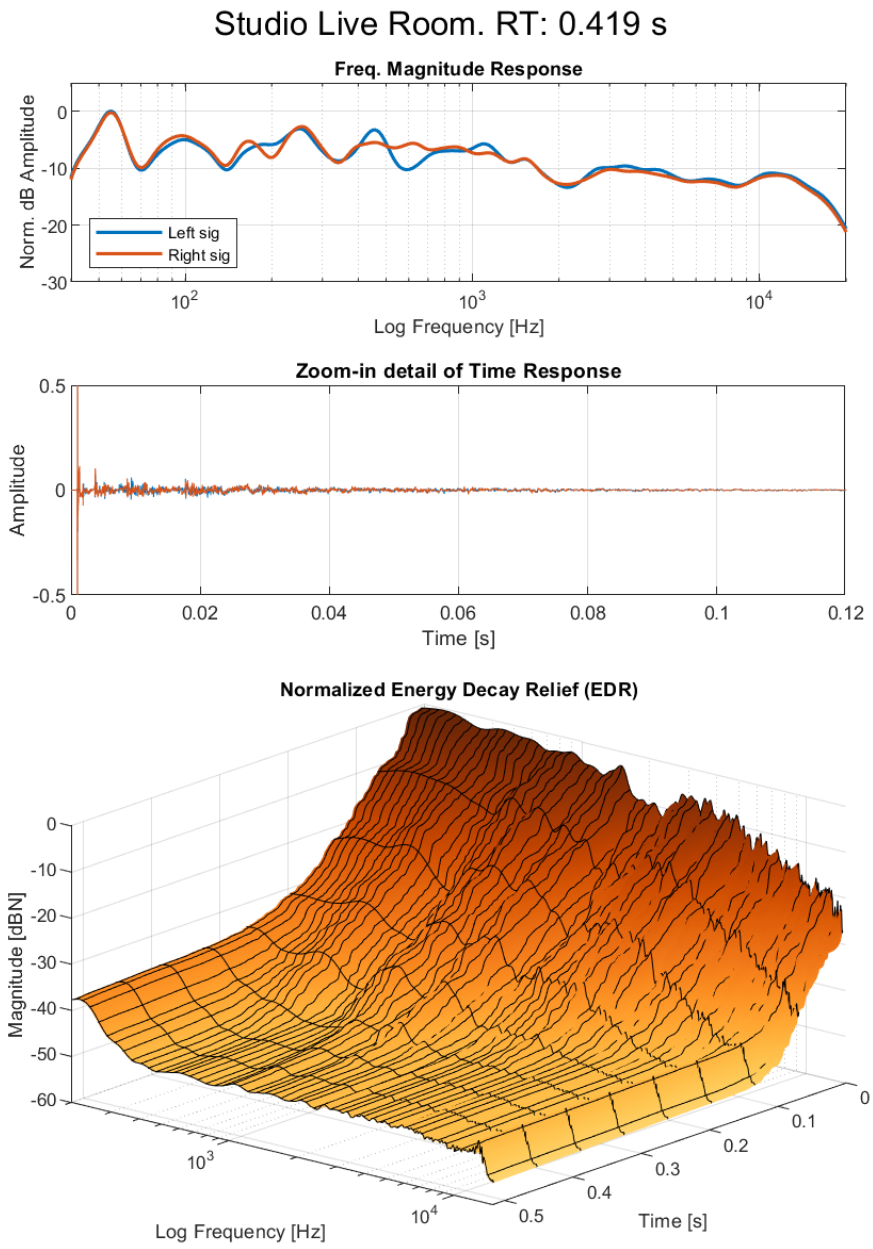


Figure 99: Frequency and Time behavior of the Live Room used for the baseline study of the co-presence experiment. Stereo omni pair measurement of an impulse response from a source at 8ft.

2.2 Remote Performance Rooms

The following plots concern the time-frequency fingerprint of the two remote performance rooms used during the main phase of the main copresence study (“Research Lab” and “Loewe Theater”). These are the rooms used for the virtual room acoustics processing of the *acoustically-congruent* conditions (AC) and (SC) illustrated in table 1. Plots are included for the BRIR recordings at far (8 ft) and near (1 ft) positions.

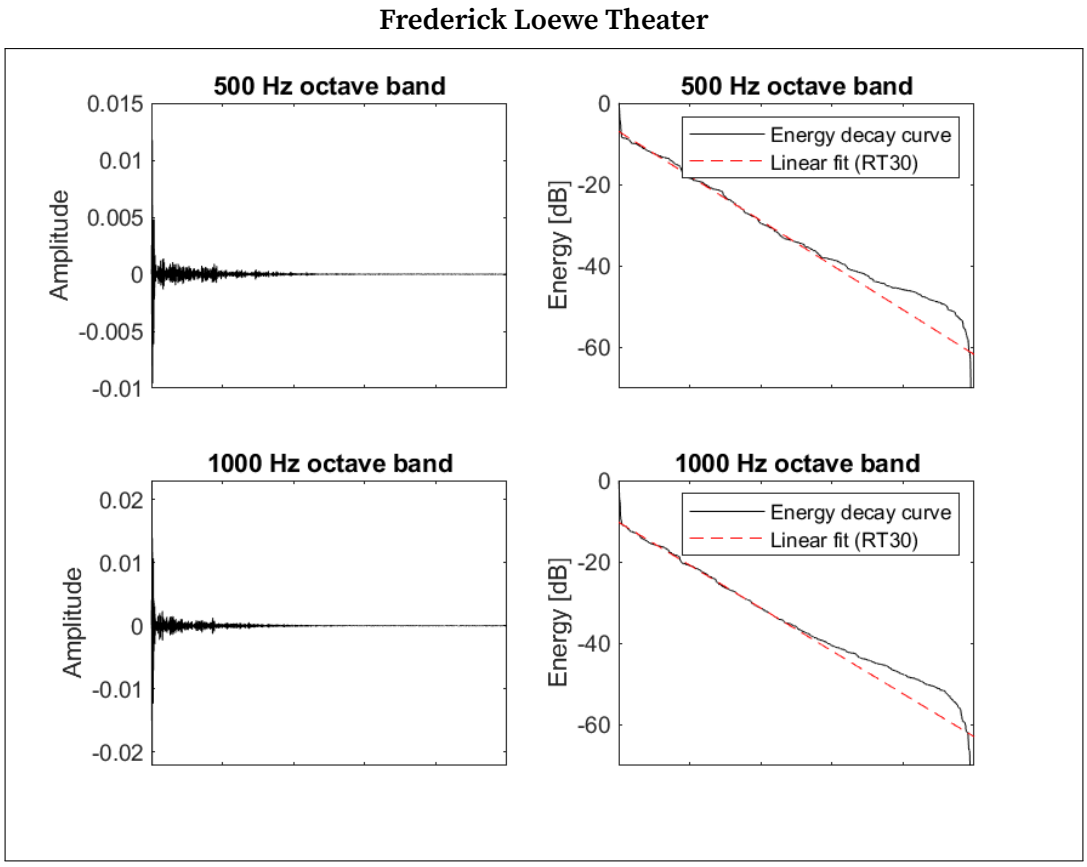


Figure 100: RT30 fit of the Theater location (“F. Loewe Theater”) at 500 Hz and 1 kHz. Used as one of the locations of remote performance and for conditions (AC) and (SC) of the main experiment phase. Taken from the omnidirectional measurement of an impulse response from a source at 8ft.

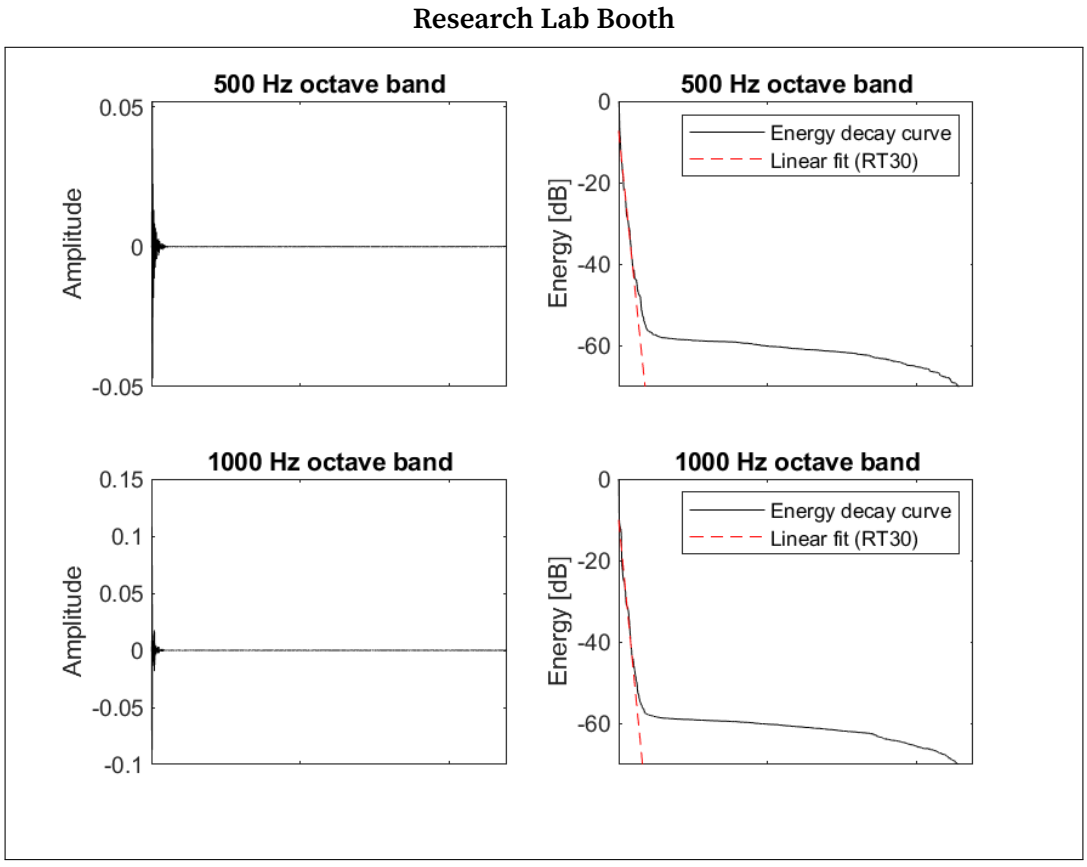


Figure 101: RT30 fit of the ISO Booth location (“Research Lab’) at 500 Hz and 1 kHz. Used as one of the locations of remote performance and for condition (SC) of the main experiment phase. Taken from the omnidirectional measurement of an impulse response from a source at 8ft.

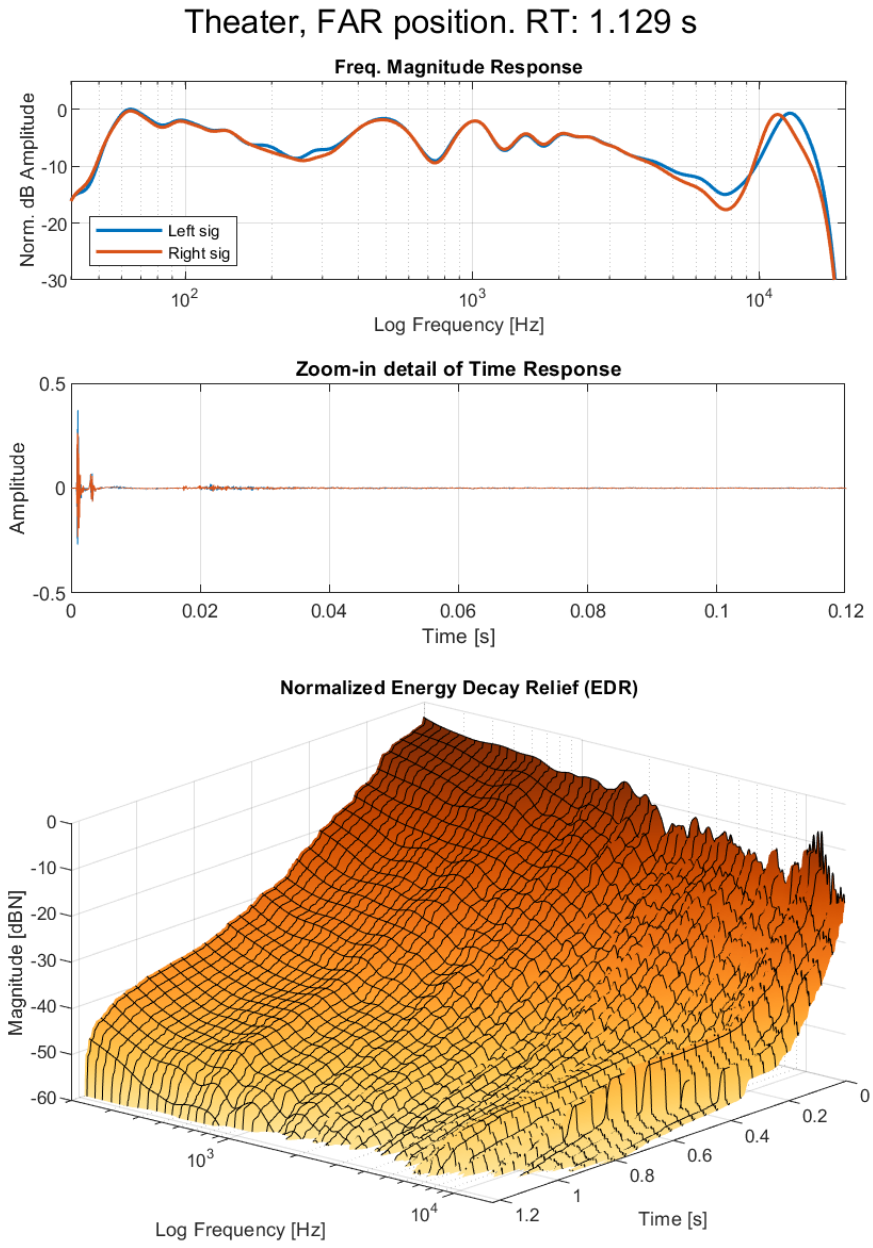


Figure 102: Frequency and Time behavior of the Theater location ("F. Loewe Theater") used as one of the locations of remote performance and for the measurement of processing filters used in conditions (AC) and (SC) of the main experiment phase. Measurement taken with a Binaural microphone at 8ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

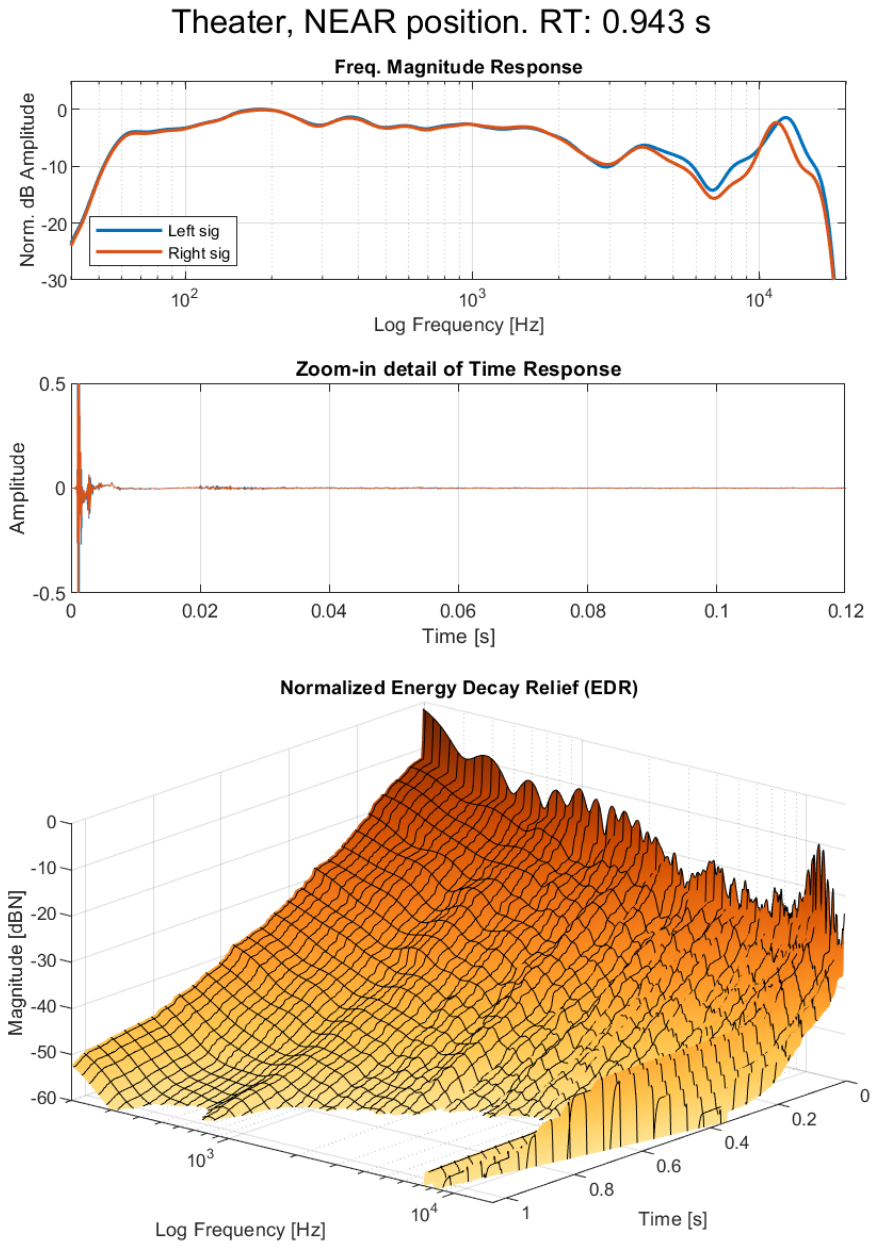


Figure 103: Frequency and Time behavior of the Theater location ("F. Loewe Theater") used as one of the locations of remote performance and for the measurement of processing filters used in conditions (AC) and (SC) of the main experiment phase. Measurement taken with a Binaural microphone at 1ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

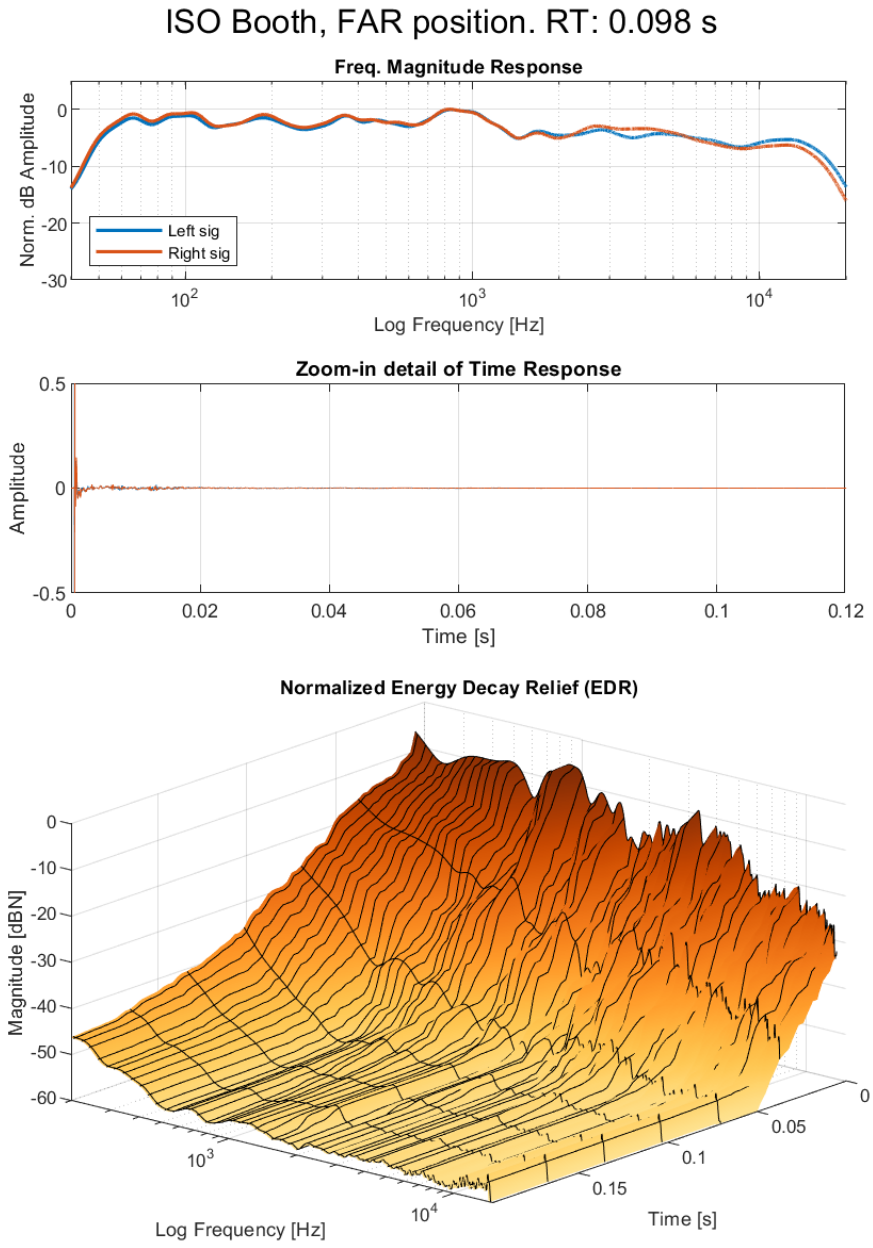


Figure 104: Frequency and Time behavior of the ISO Booth location ("Research Lab') used as one of the locations of remote performance and for the measurement of processing filters used in condition (SC) of the main experiment phase. Measurement taken with a Binaural microphone at 8ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

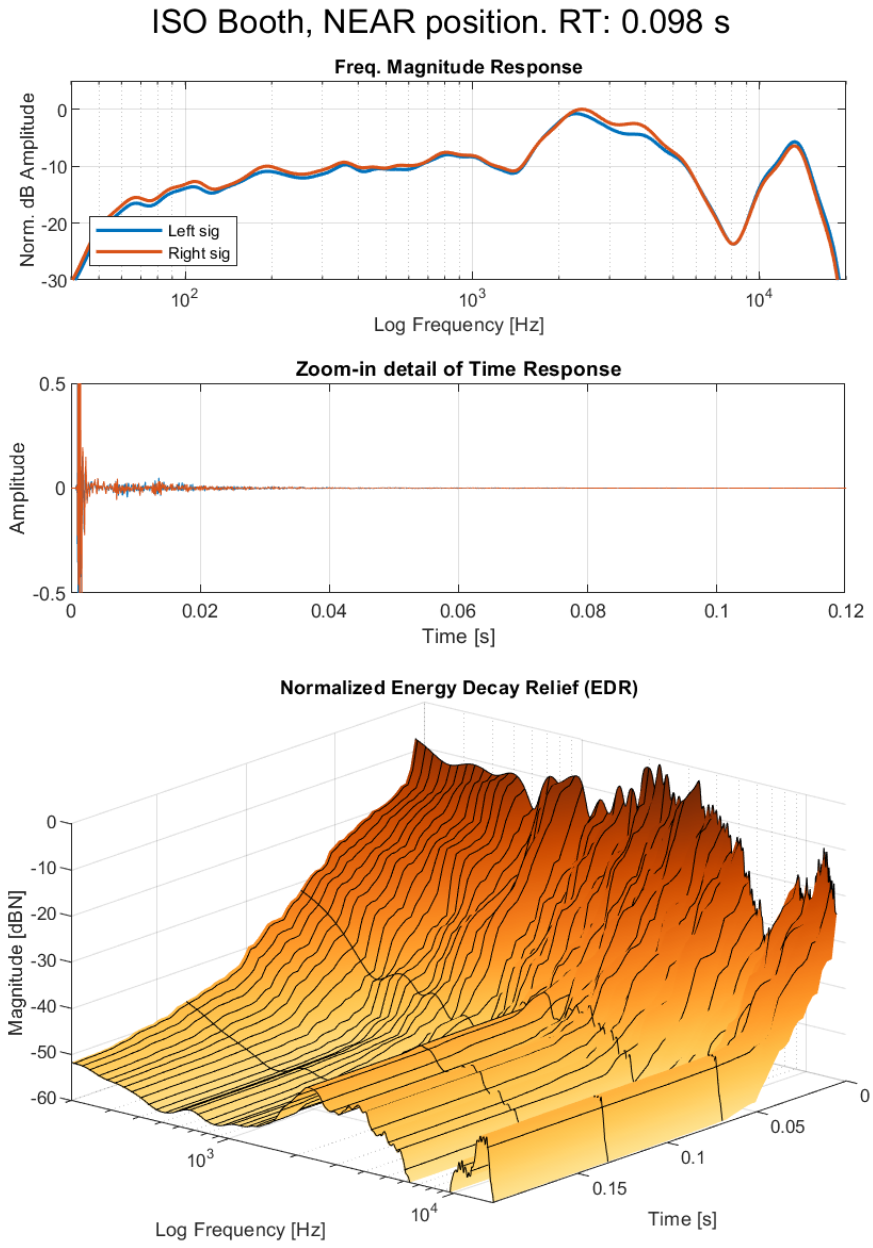


Figure 105: Frequency and Time behavior of the ISO Booth location ("Research Lab') used as one of the locations of remote performance and for the measurement of processing filters used in condition (SC) of the main experiment phase. Measurement taken with a Binaural microphone at 1ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

2.3 Other Measured Rooms

The following plots concern the time-frequency fingerprint of the two rooms used for the *acoustically-divergent* conditions (AD) and (SD) illustrated in Table 1. These consisted of two lecture halls ("303" and "Conference Room"). Plots are included for the BRIR recordings at far (8 ft) and near (1 ft) positions.

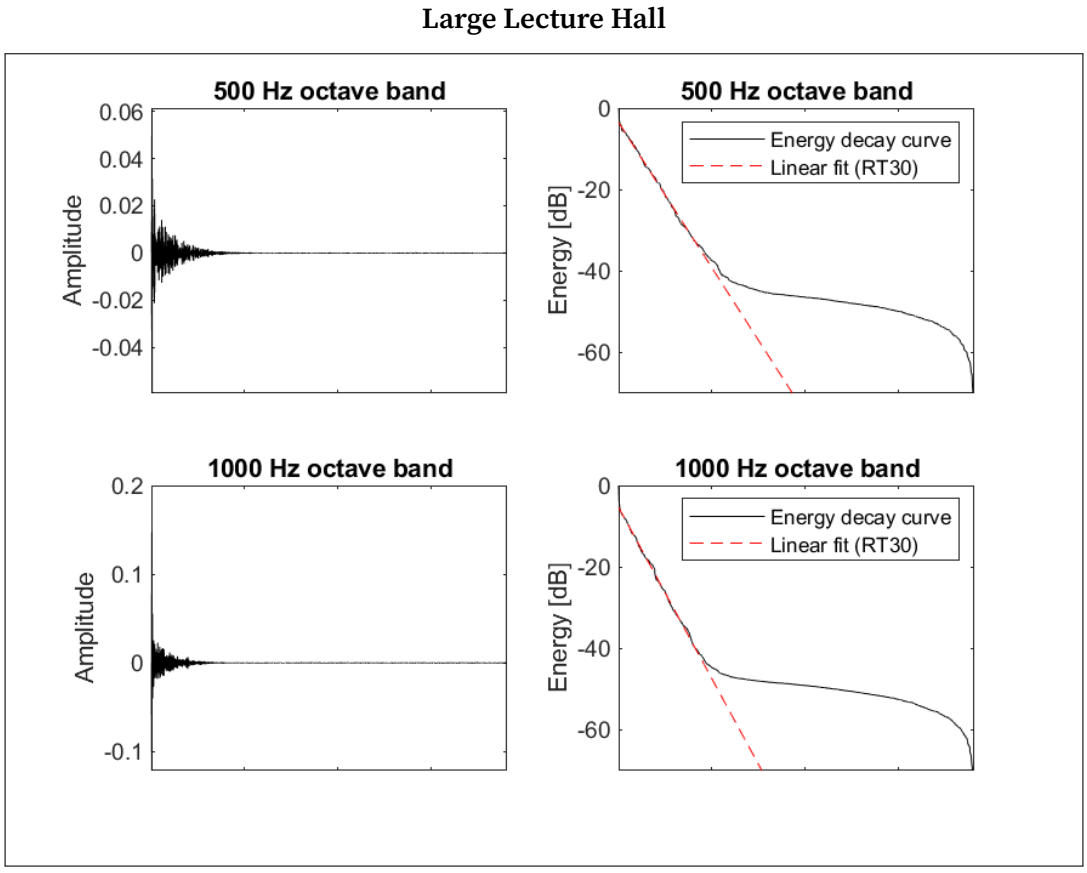


Figure 106: RT30 fit of the Large lecture hall ("Room 303") at 500 Hz and 1 kHz. Used for conditions (AD) and (SD) of the main experiment phase. Taken from the omnidirectional measurement of an impulse response from a source at 8ft.

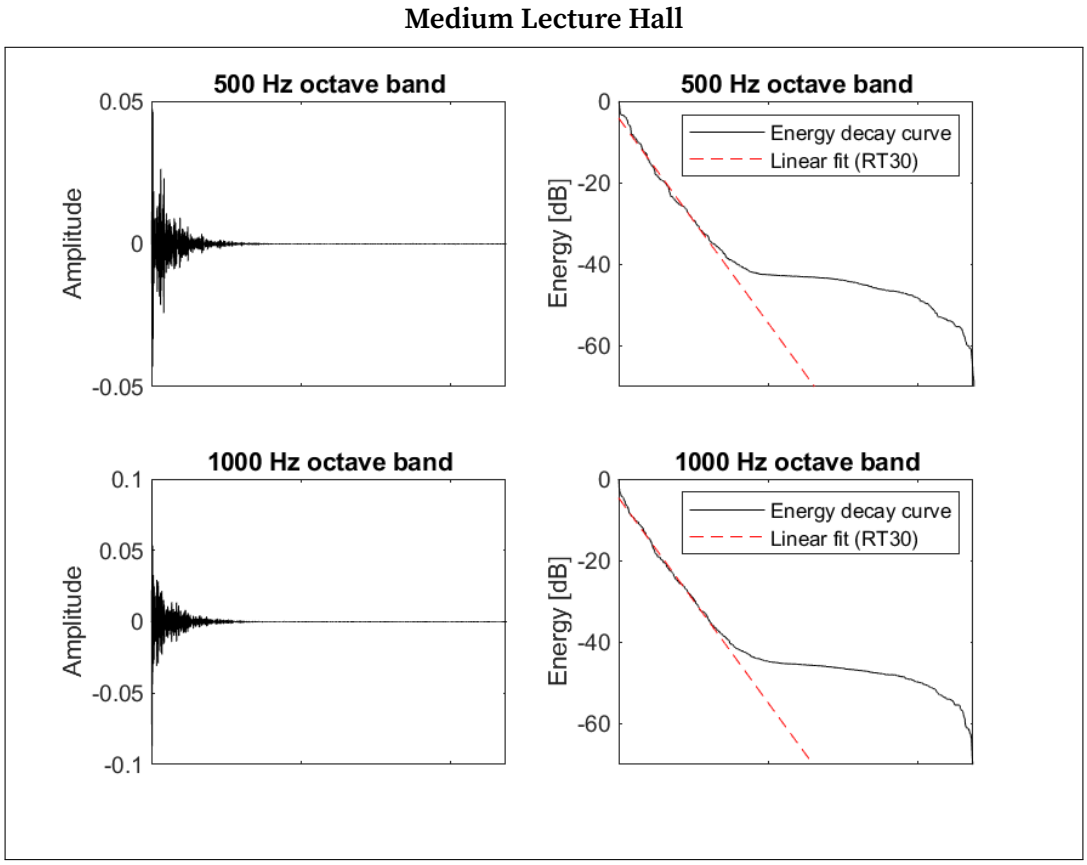


Figure 107: RT30 fit of the Medium lecture hall (“Conference Room”) at 500 Hz and 1 kHz. Used for conditions (AD) of the main experiment phase. Taken from the omnidirectional measurement of an impulse response from a source at 8ft.

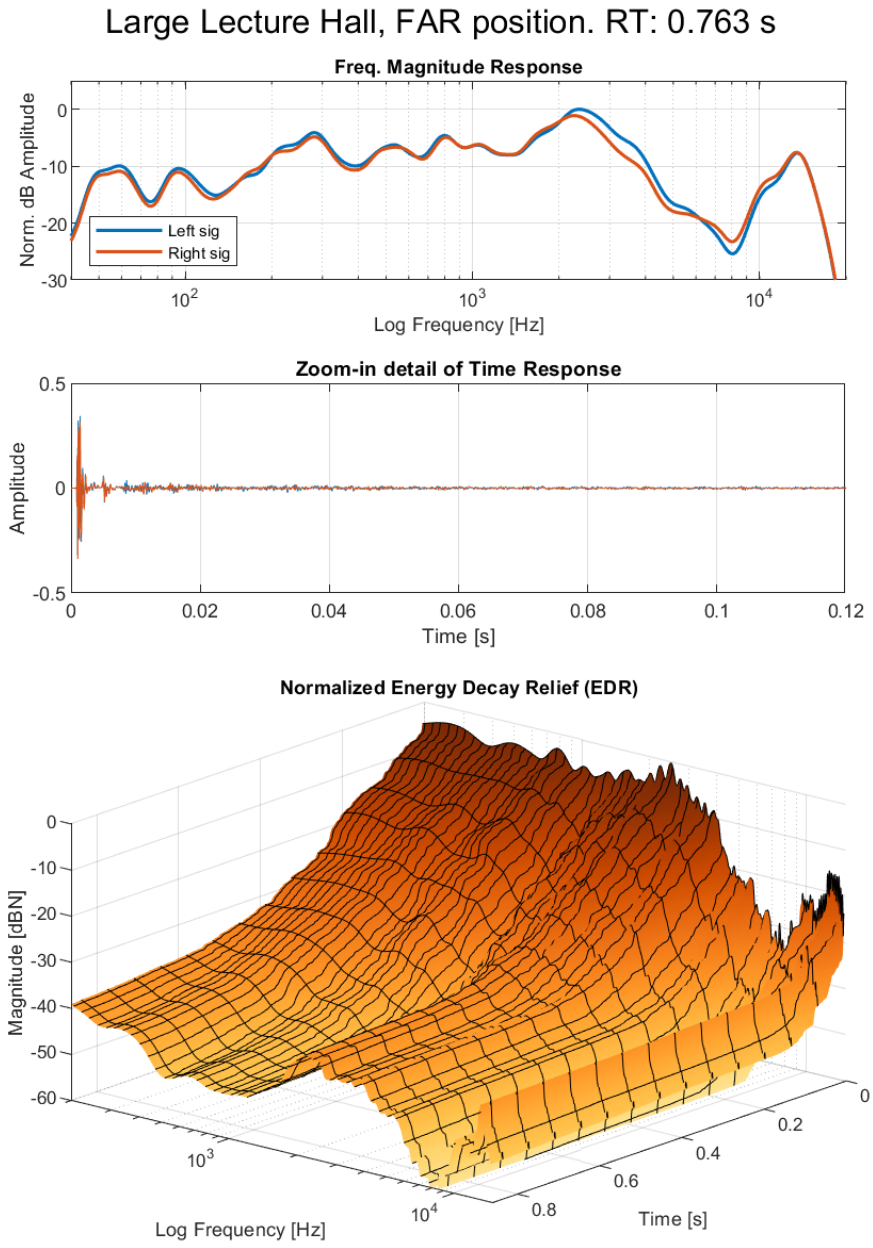


Figure 108: Frequency and Time behaviour of the Large Lecture Room used for the measurement of processing filters used in conditions (AD) and (SD) of the main experiment phase. Measurement taken with a Binaural microphone at 8ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

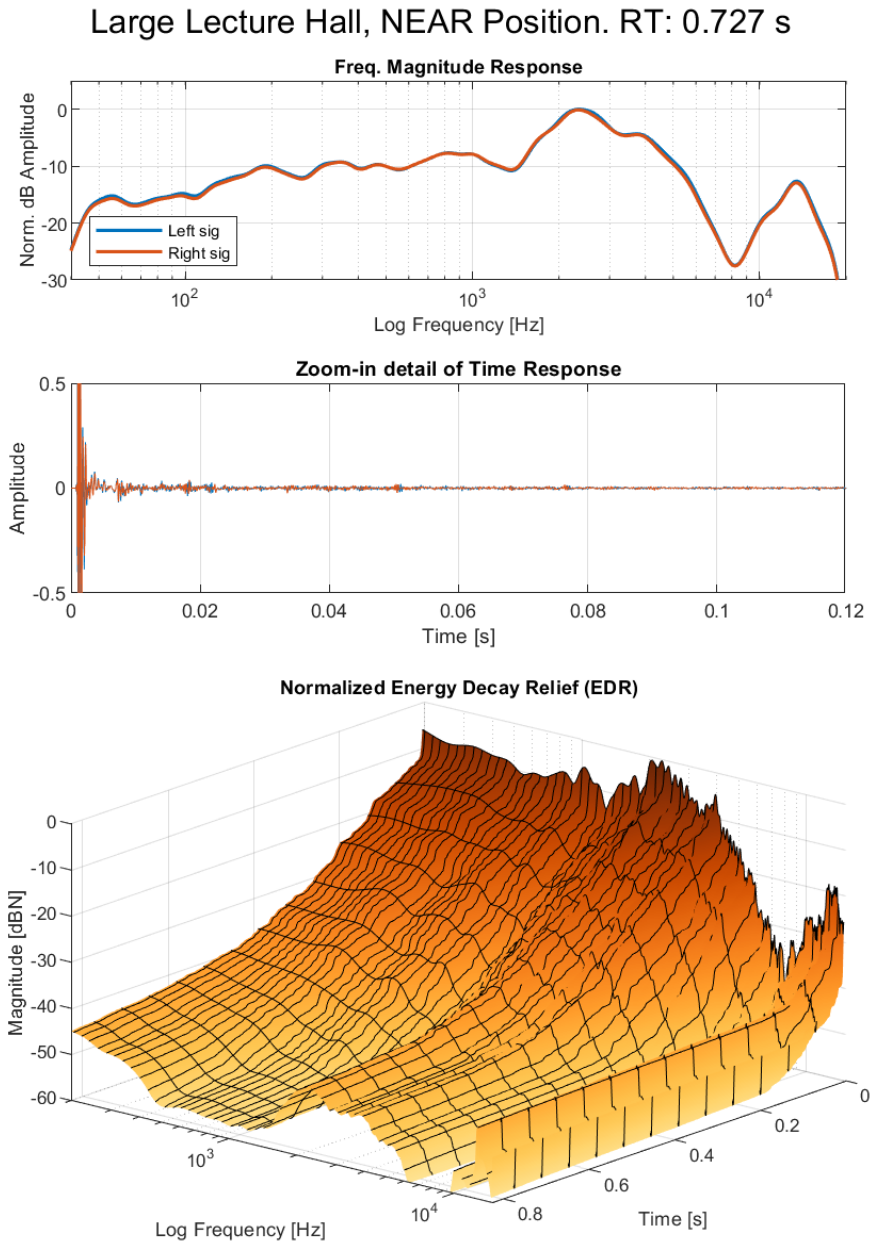


Figure 109: Frequency and Time behaviour of the Large Lecture Room used for the measurement of processing filters used inconditions (AD) and (SD) of the main experiment phase. Measurement taken with a Binaural microphone at 1ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

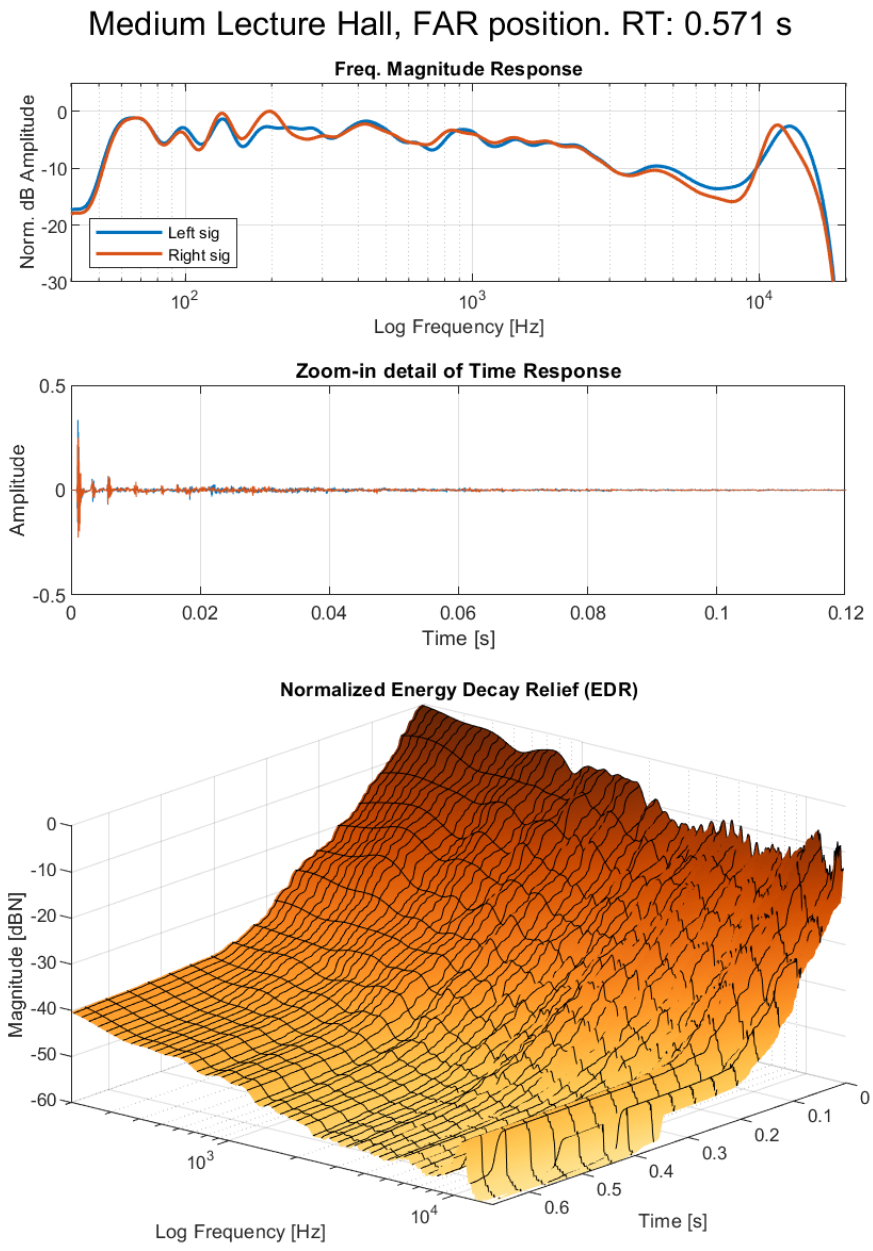


Figure 110: Frequency and Time behaviour of the Large Lecture Room used for the measurement of processing filters used in conditions (AD) of the main experiment phase. Measurement taken with a Binaural microphone at 8ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

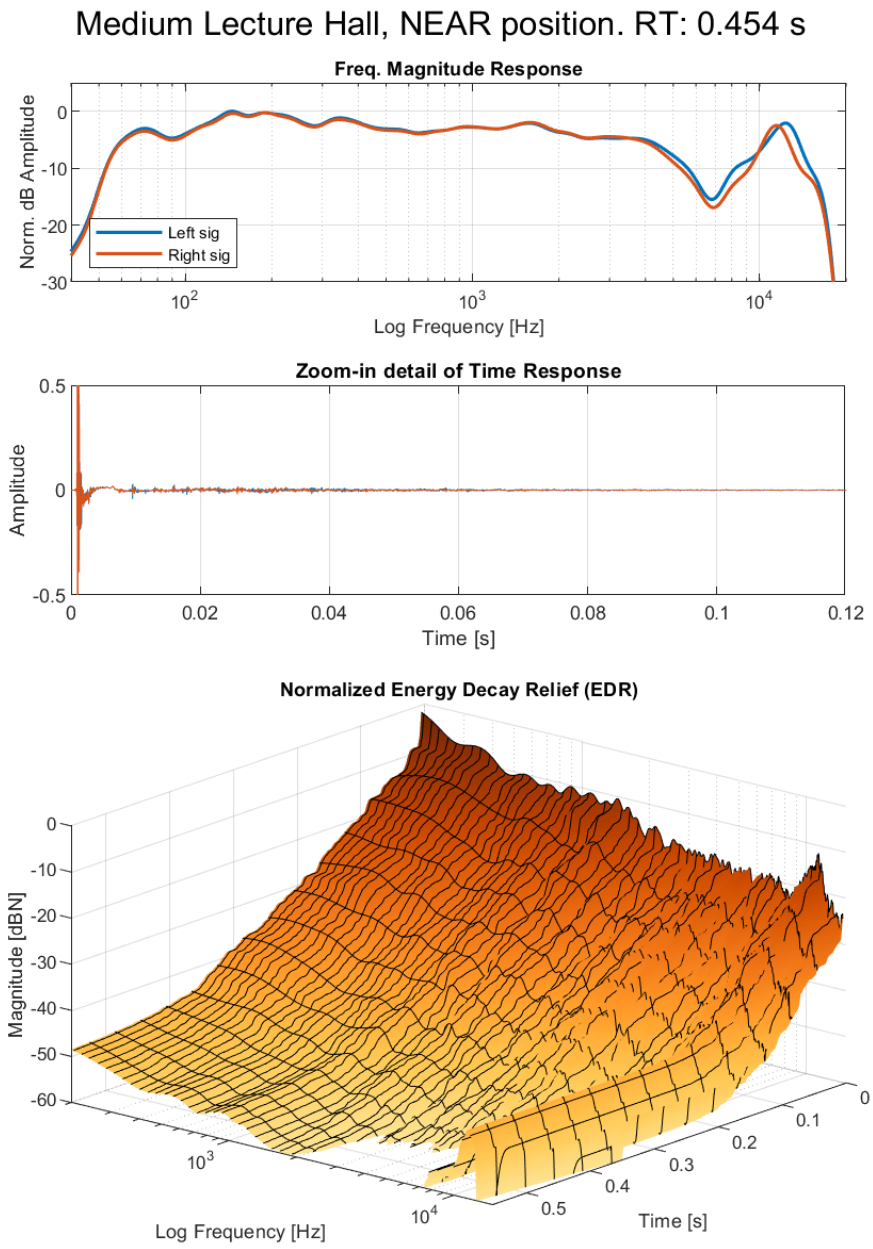


Figure 111: Frequency and Time behaviour of the Medium Lecture Room used for the measurement of processing filters used in conditions (AD) of the main experiment phase. Measurement taken with a Binaural microphone at 1ft distance. Frequency response is shown smoothed over $1/4$ octave bands.

3 Additional Results

3.1 Alternative Heat-map Overview

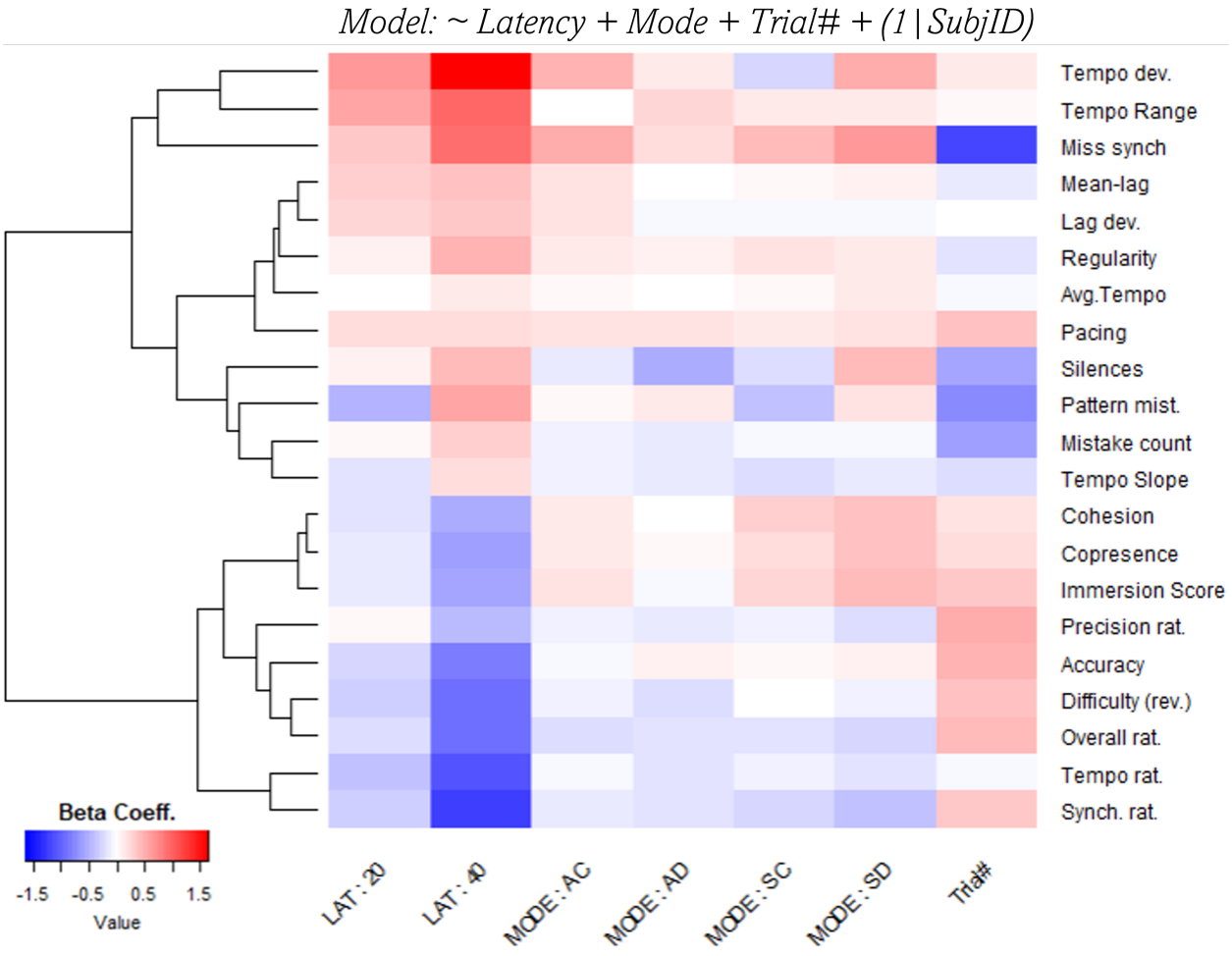


Figure 112: Overview of beta coefficient magnitudes for all models (including latency, auralization mode, and trial). Colors indicate the sign and magnitude of the model's beta coefficient. The coefficient for Trial is calculated as the step-size effect multiplied by the total number of trials. Models are computed over scaled metrics. Rows are clustered per similarity.

3.2 Response Distribution Means and Standard Error

3.2.1 Trial Questionnaire

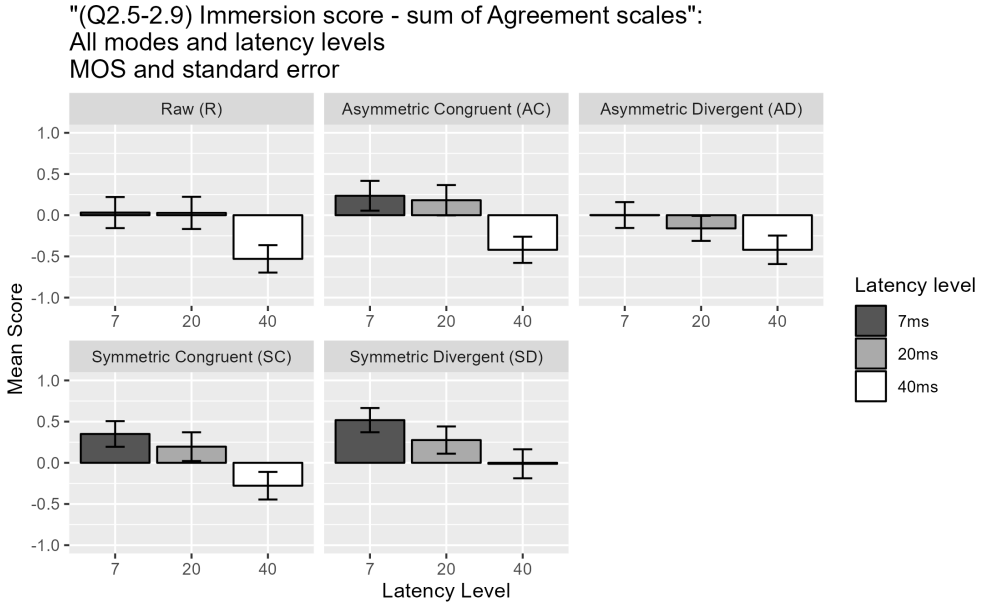


Figure 113: *Immersion score* evaluation results grouped by latency and auralization mode

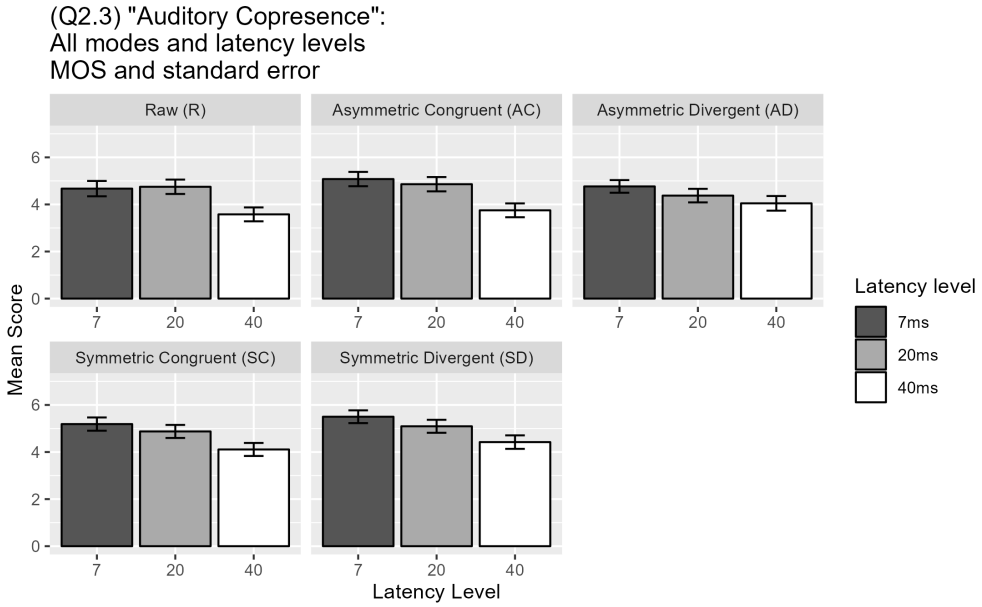


Figure 114: *Auditory copresence* evaluation results grouped by latency and auralization mode

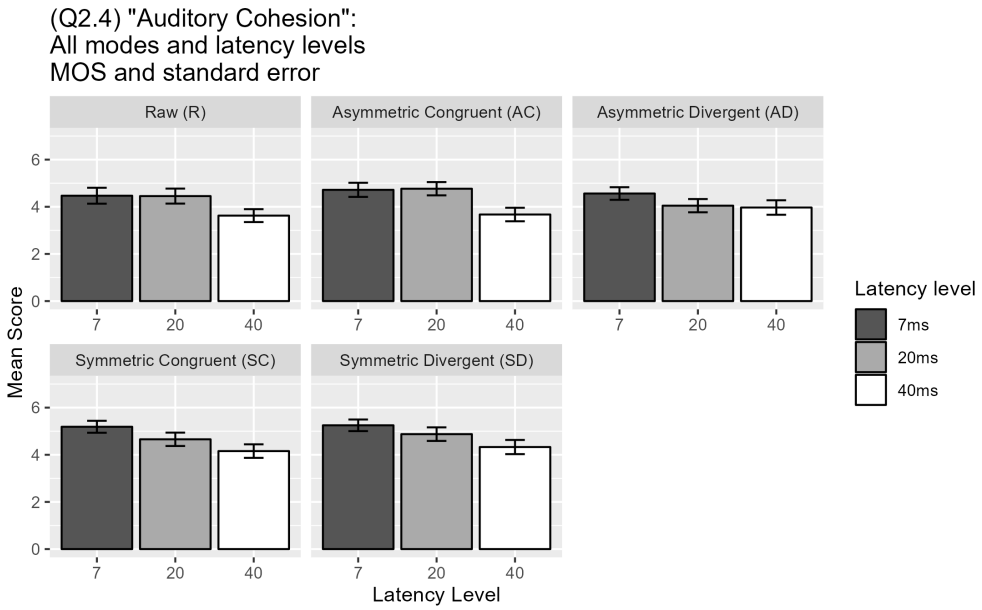


Figure 115: *Auditory cohesion* evaluation results grouped by latency and auralization mode

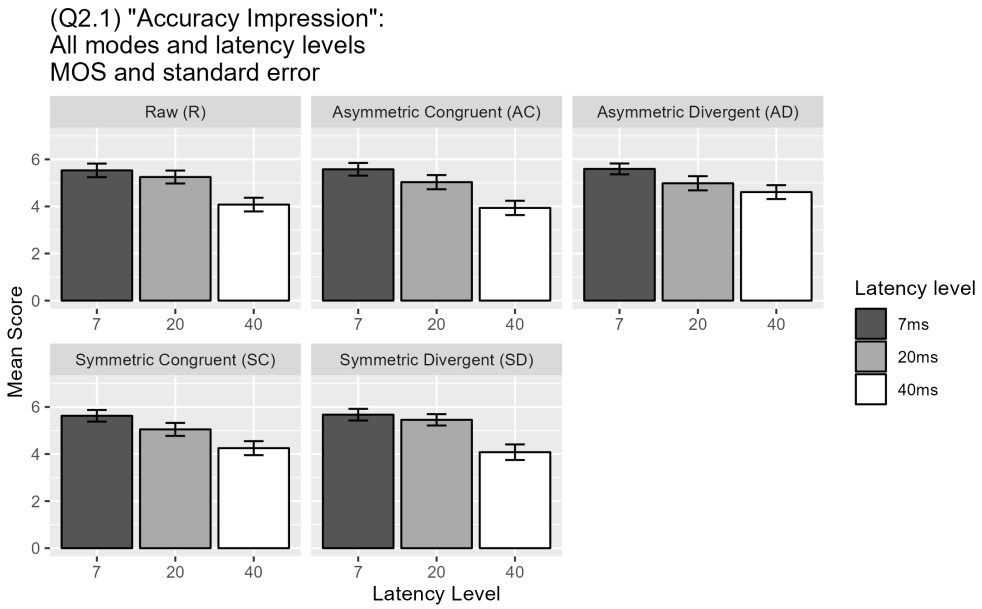


Figure 116: *Perceived Accuracy* evaluation results grouped by latency and auralization mode

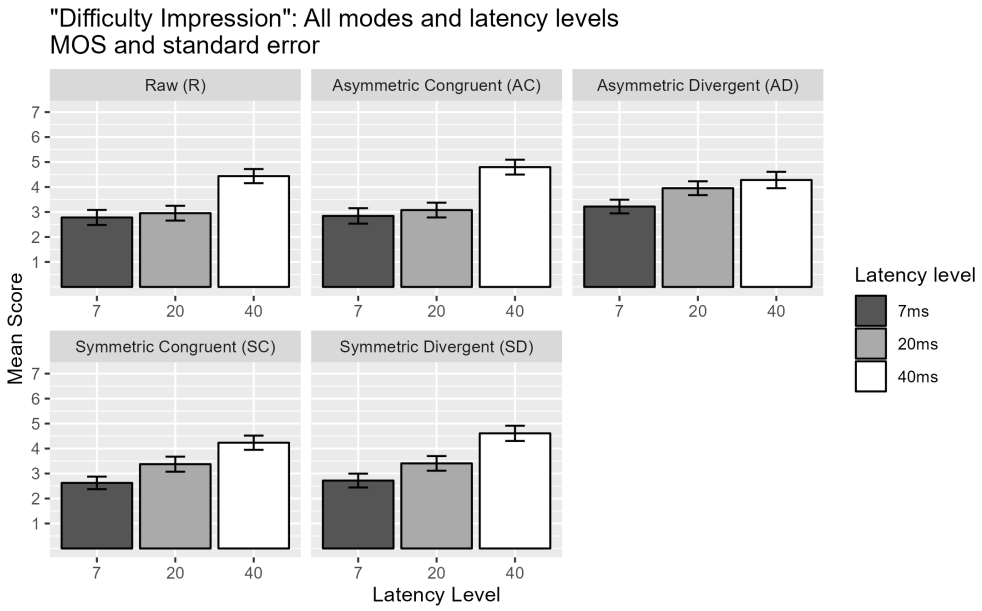


Figure 117: *Perceived Difficulty* evaluation results grouped by latency and auralization mode

3.2.2 Third Party Ratings and Annotations

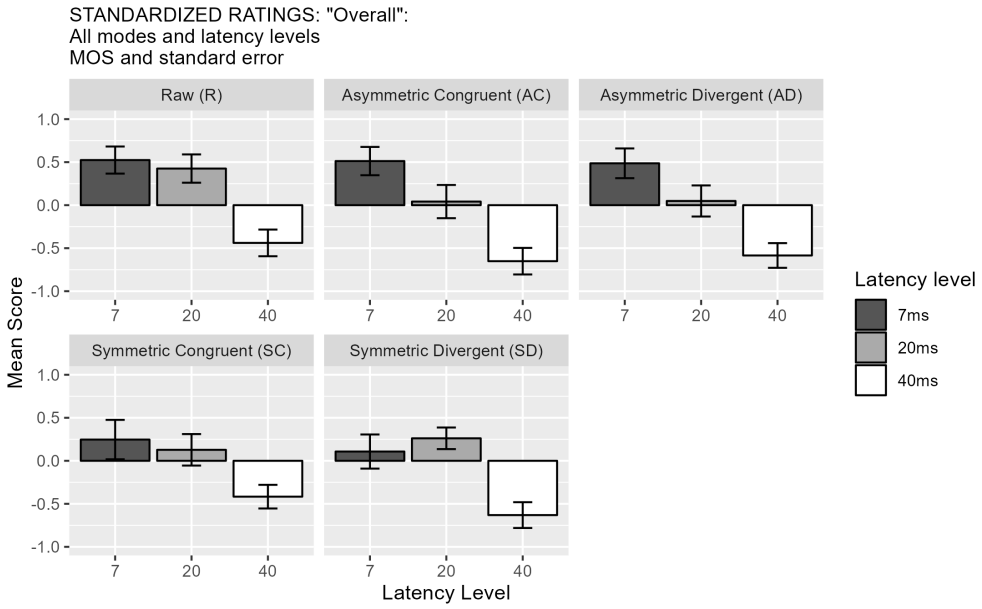


Figure 118: *Overall rating* evaluation results grouped by latency and auralization mode

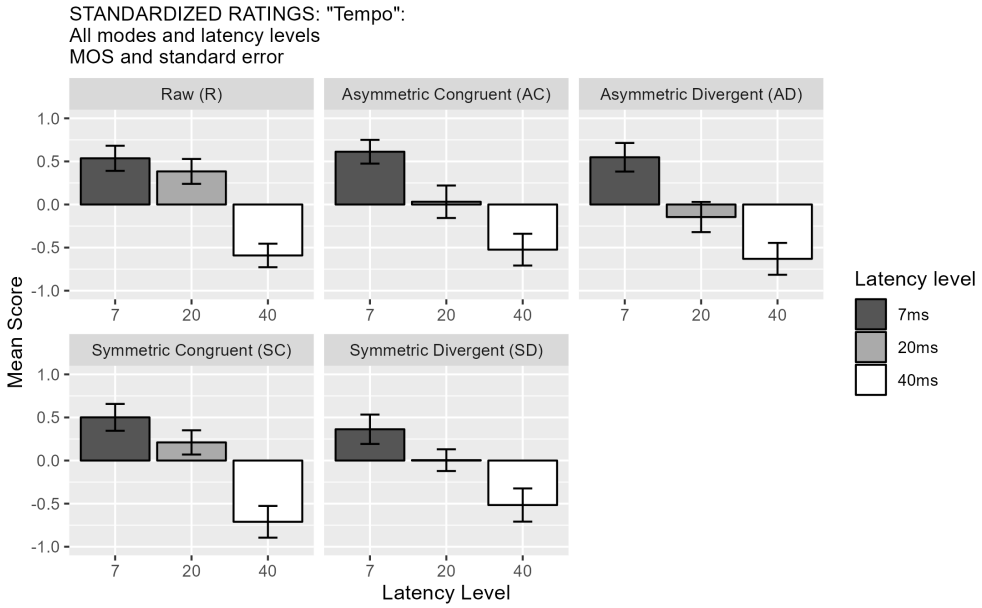


Figure 119: *Tempo rating* evaluation results grouped by latency and auralization mode

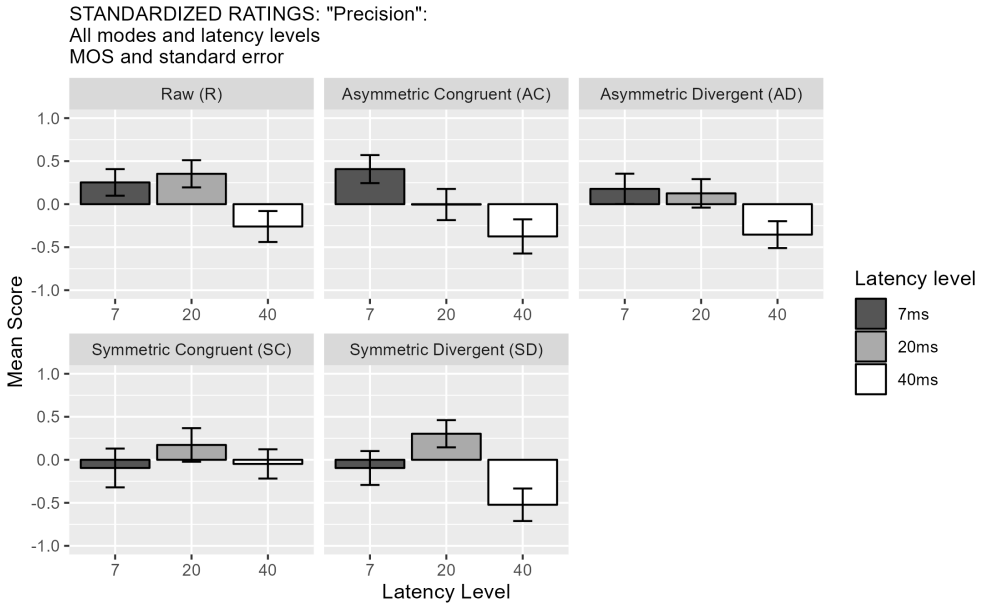


Figure 120: *Precision rating* evaluation results grouped by latency and auralization mode

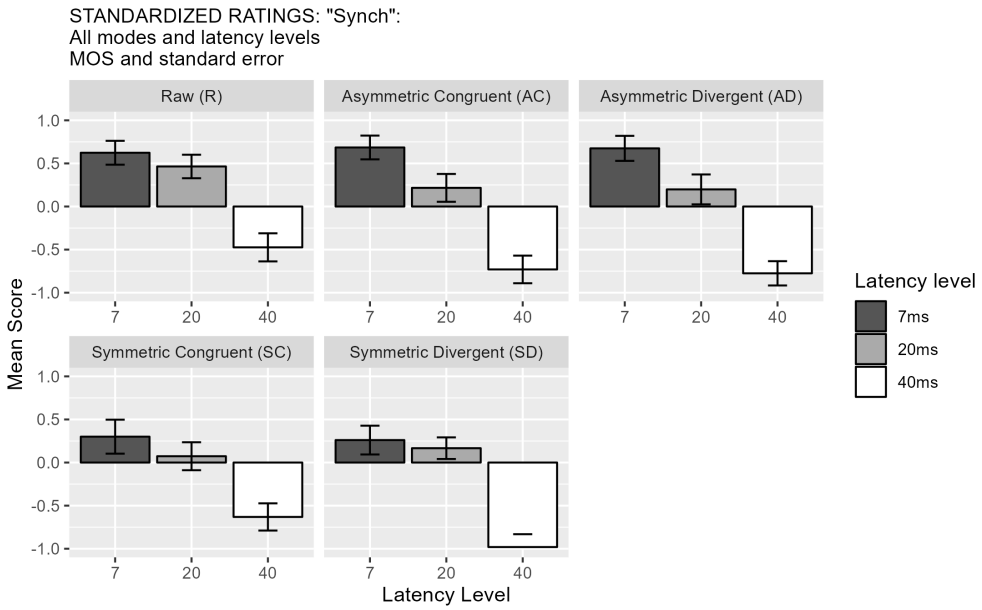


Figure 121: *Synchronization rating* evaluation results grouped by latency and auralization mode

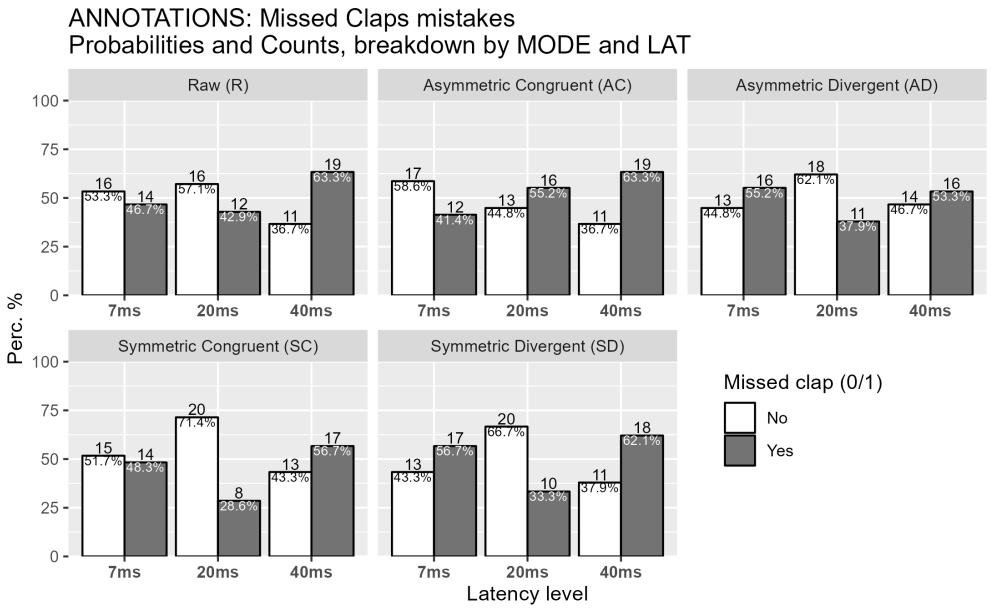


Figure 122: Rate of *Pattern Mistake* identifications grouped by latency and auralization mode

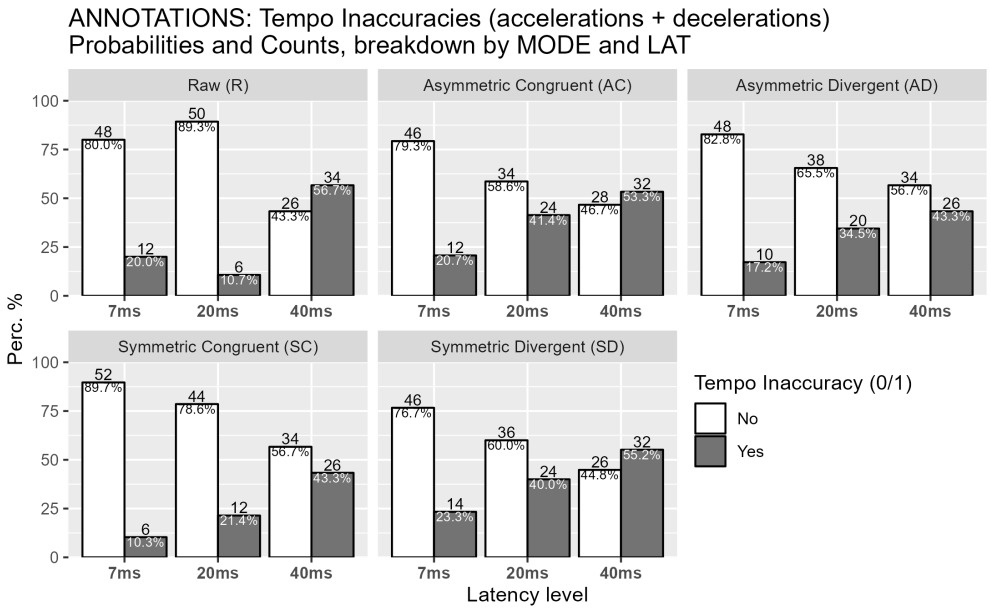


Figure 123: Rate of *Tempo inaccuracies* identifications grouped by latency and auralization mode

3.2.3 Objective Metrics

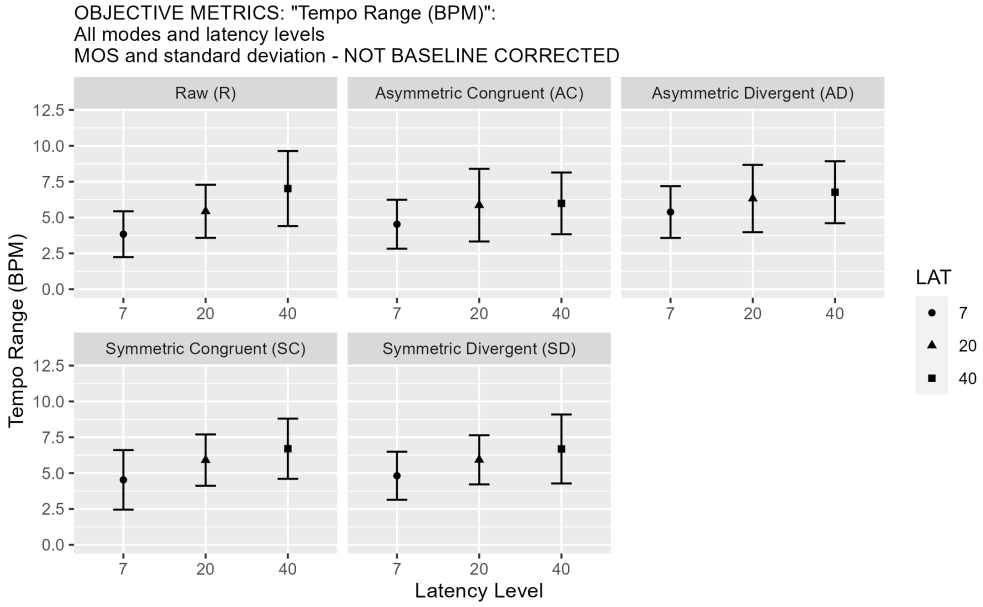


Figure 124: Mean and standard error of observed *Tempo Range*, grouped by latency and auralization mode

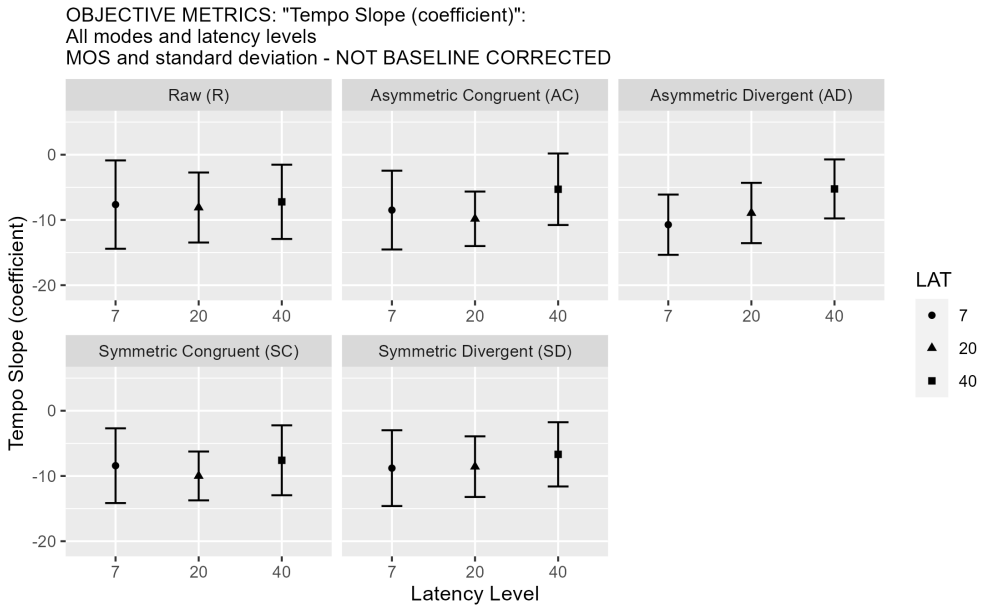


Figure 125: Mean and standard error of observed *Tempo Slope*, grouped by latency and auralization mode

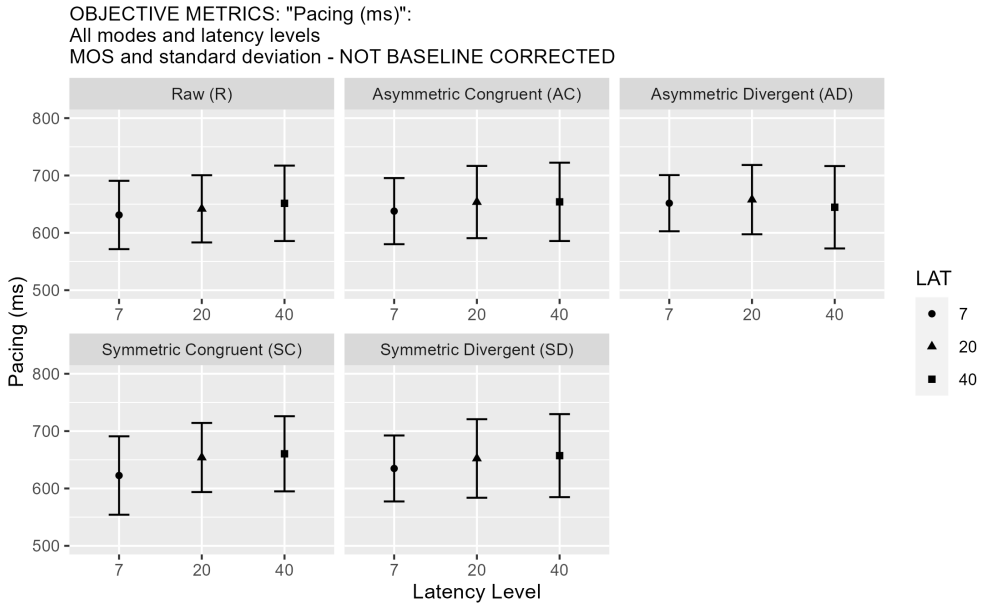


Figure 126: Mean and standard error of observed *Pacing*, grouped by latency and auralization mode

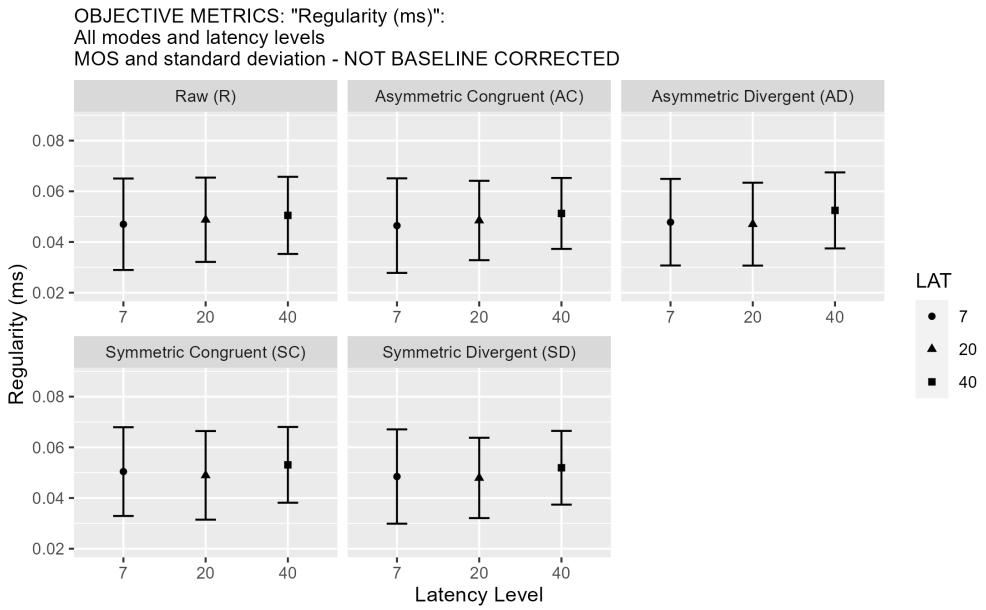


Figure 127: Mean and standard error of observed *Regularity*, grouped by latency and auralization mode

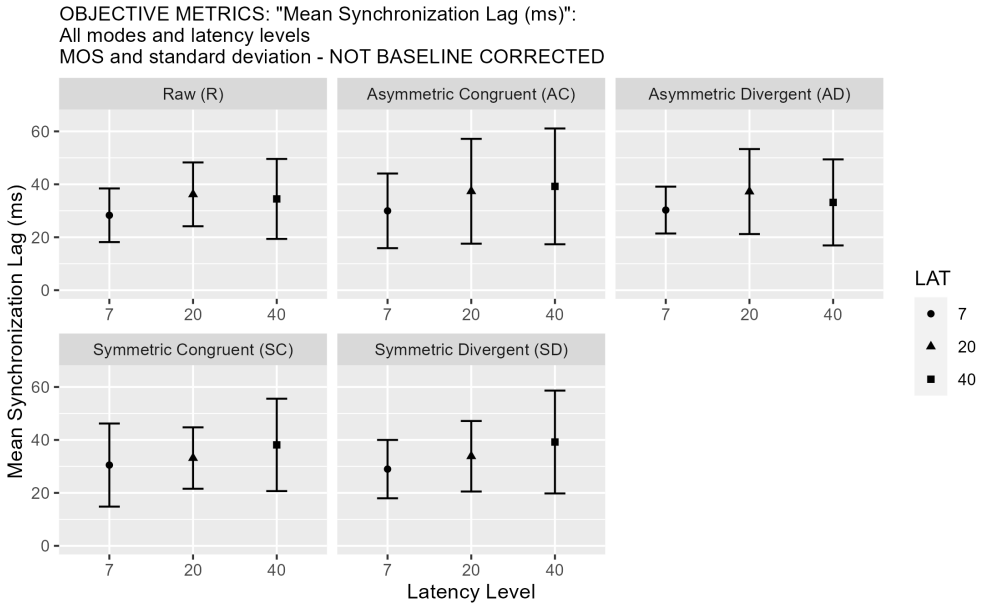


Figure 128: Mean and standard error of observed *Mean Lag*, grouped by latency and auralization mode

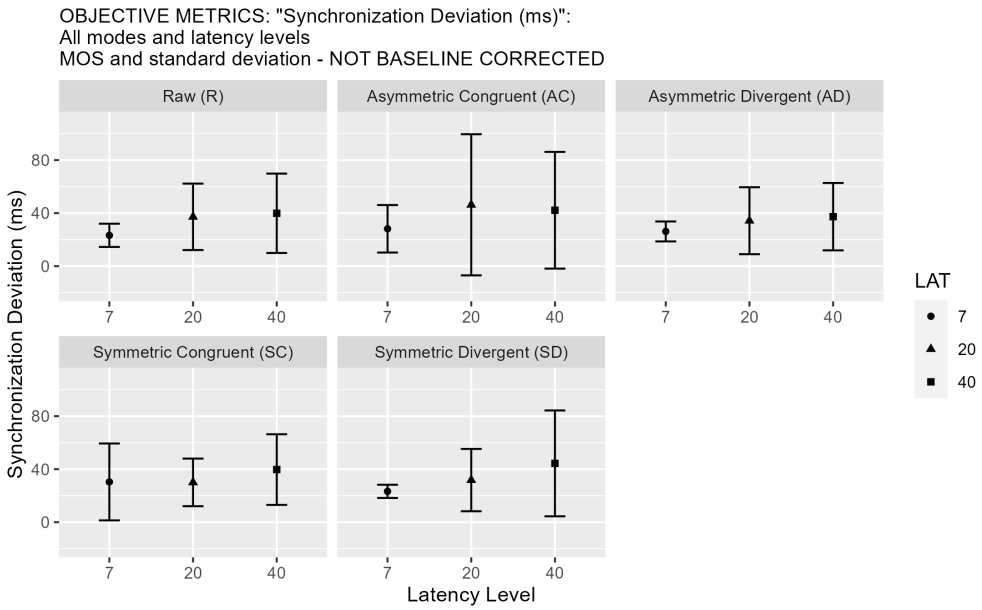


Figure 129: Mean and standard error of observed *Synch deviation*, grouped by latency and auralization mode

3.3 Full Correlation Matrix

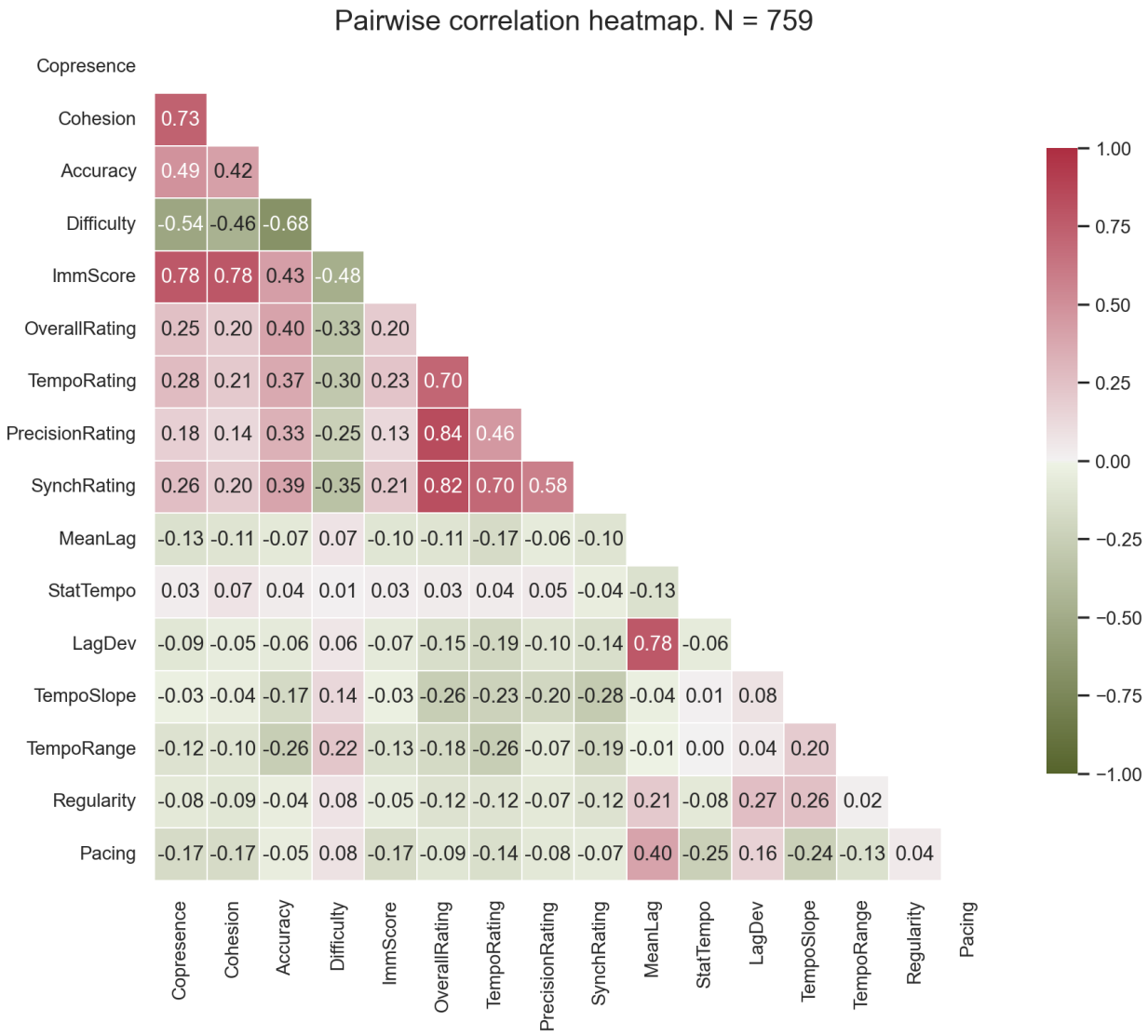


Figure 130: Full correlation matrix including subjective responses, third-party ratings and objective scales. All scales were standardized and outliers were removed using the objective metric data

4 Additional Equations

The formulas in this appendix are not included in the main body, but pertain to relative metrics used in the presented work.

4.1 Logarithmic Sine-sweep Equation

From (Chan 2010).

$$x(t) = \sin\left(\frac{2\pi f_1 T}{\ln \frac{f_2}{f_1}}\left[\exp\left(\frac{\ln \frac{f_2}{f_1} t}{T}\right) - 1\right]\right) \tag{19}$$

Where f_1 is the starting frequency, f_2 is the ending frequency, and T the duration of the chirp

4.2 Reverberation Time Calculation

The RT30 metric for calculating reverberation time is used when the dynamic range of the microphone employed for measuring the impulse response does not reach 60dB, or the noise floor is too high. In those cases the RT30 or RT20 can be employed by extrapolating from the available range.

$$\texttt{RT30} = 2 * (t_{EDC=-35dB} - t_{EDC=-5dB}) \tag{20}$$

Where EDC is the log energy decay curve obtained from the time-domain impulse response with the Schroeder Integration method.

4.3 Signal Cross-correlation Formula

This formula was used to calculate the round-trip system latency level in the distributed performance network for the co-presence study.

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t+\tau)dt$$

where τ is the time delay, $R_{xy}(\tau)$ is the correlation between the two signals at that time delay, $x(t)$ is the first signal, and $y(t)$ is the second signal. Over a series of time displacements, the τ level producing the highest correlation metric represents the round-trip latency amount in samples.

4.4 Spectral Flux

From (Scheirer and Slaney 1997)

$$\texttt{flux}(t) = \left(\sum_{k=b_1}^{b_2} |s_k(t) - s_k(t-1)|^P \right)^{\frac{1}{P}}$$

Where s_k is the spectral value at bin k. b_1 and b_2 are the band edges, in bins, over which to calculate the spectral flux. P is the norm type.

4.5 BIC and AICc

Formula for the *Bayesian Information Criterion*:

$$\texttt{BIC} = -2LL + k \ln n \tag{21}$$

Where k is the number of parameters in the model, LL is the log-likelihood, and n is the number of observations.

Formula for *Corrected Akaike Information Criterion*

$$\texttt{AICc} = -2LL + 2k + \frac{2k(k+1)}{n-k-1} \tag{22}$$

Where k is the number of parameters in the model, LL is the log-likelihood, and n is the number of observations.