

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Real-Time Vocal Auralisation in XR: from Multimodal Perception to Applications in Virtual Performance Spaces

Mauricio Flores Vargas, M. Phil.

Presented to the University of Dublin, Trinity College

in fulfilment of the requirements for the Degree of

Doctor of Philosophy

March 2026

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I acknowledge the use of GenAI, including OpenAI ChatGPT (versions 4o and 5) and Anthropic Claude (Sonnet 4.5), to assist with grammar, structural edits, and proofreading.

I agree to deposit this thesis in the University's open-access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Mauricio Flores Vargas

March 19, 2026

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Mauricio Flores Vargas

March 19, 2026

To Ollie

Acknowledgments

First, I want to thank my supervisors, Dr. Rachel McDonnell and Dr. Enda Bates, for their time, dedication, and invaluable guidance and support throughout my PhD. I am deeply grateful for giving me the opportunity to pursue this research journey.

In addition, I want to thank Dr. Gareth Young for generously sharing his knowledge and advice on both a professional and personal level. Similarly, I am grateful to Rose Connolly for her collaboration and hard work, which is reflected in this thesis and in the award we proudly share, and to Dr. Rebecca Fribourg for her collaboration and for sharing her expertise.

I must also thank all my friends in Stack B (Stephen, Yuan, Bharat, Priyansh, Theo, Sam, Alberto, Darragh, Caoimhe, Donal, and many more), my friends outside the lab (Carlos, Jesús, Ollie, Fionnuala, John, Katie, Justine, Cian), and everyone who, at some point during these four years, shared their ideas, thoughts, and time for the benefit of my research.

Finally, I want to thank Sarah, for her patience, love, and support during both the good and the tough times. This achievement is as much hers as it is mine. I also want to thank my parents for shaping who I am today; Pau, for being the best life accomplice I could have; and my Irish family for making me feel at home.

I am sure there are many others who also deserve mention, and I thank you, the reader, for your interest in my work.

Mauricio Flores Vargas

University of Dublin, Trinity College
March 2026

Abstract

This thesis explores the sensory modalities that contribute to the perception of acoustic environments in eXtended Reality (XR), focusing on visual input and the auditory acoustic response (auralisation). Using vocal interaction as first-person auditory feedback, it examines key factors of audio-visual multimodality in Immersive Virtual Environments (IVEs).

As multimodal research and applications grow with advances in immersive technologies, auditory feedback has been shown to be a key dimension in audio-visual contexts. Prior research has highlighted the role of sound auralisation in supporting perceptual factors in IVEs and shown that altering the quality of one modality can influence overall multimodal perception. Moreover, self-produced vocal interaction and real-time auralisation have also been found to affect musical performance. However, most research has focused on sound sources within the environment, perceived by the user from a third-person perspective. This viewpoint is also reflected in previous work on sound auralisation and audio-visual reconstruction of virtual spaces. To this end, this work investigates self-produced and embodied auditory feedback in multimodal IVEs, its effect on perceptual constructs and psychophysiological responses, and how visual and audio rendering fidelity influence perceived quality and audio-visual integration. In addition, it evaluates the plausibility and usability of a virtual performance space and how it shapes singers' overall experience.

Findings show a higher sense of presence and its related subscales when real-time vocal auralisation is present, and demonstrate that rendering fidelity in both the audio and visual domains influences overall perceived quality and audio-visual integration. Moreover, high-fidelity audio-visual representations can provide meaningful environments for vocal training and performance, enabling singers to understand spatial distribution and acoustics while supporting a psychological shift that enhances vocal training in IVEs. Thus, through perceptual, practical, and technical contributions, it highlights the role of real-time vocal auralisation in shaping audio-visual experiences in IVEs.

Contents

Acknowledgments **v**

Abstract **vi**

List of Figures **x**

List of Tables **xiii**

List of Publications **xiv**

Chapter 1 Introduction **1**
1.1 Motivation 2
1.2 Thesis Focus 3
1.3 Scope 5
1.4 Contributions 6
1.5 Summary of Chapters 7

Chapter 2 Background **8**
2.1 Introduction 8
2.2 Audio Rendering 9
2.2.1 Spatial Audio 9
2.2.2 Virtual Room Acoustics & Auralisation 11
2.3 Visual Rendering 13
2.3.1 3D Scene Representation 13
2.3.2 Visual Rendering Fidelity 15
2.4 Perceptual Constructs 16
2.4.1 Immersion 18
2.4.2 Presence 18
2.4.3 Realism 20
2.4.4 Embodiment 21
2.5 Multimodality 25
2.6 Audio-Visual Feedback in XR 27
2.6.1 User-Centred Considerations 28
2.6.2 Technical Considerations 30
2.7 Self-Produced Sounds & Embodied Audio Feedback 35
2.7.1 Vocal Perception & Production 36
2.7.2 Vocal Production in Musical Contexts 37
2.8 Summary 39

Chapter 3 Effect of Continuous Embodied Auditory Feedback on Animal Embodiment in IVEs **41**
3.1 Hypotheses . . . 43
3.2 Experimental Setup . . . 44
3.3 Measures . . . 46
3.4 Experiment 1 . . . 47
3.4.1 Stimuli . . . 48
3.4.2 Participants . . . 49
3.4.3 Procedure . . . 50
3.4.4 Results . . . 51
3.5 Experiment 2 . . . 54
3.5.1 Stimuli . . . 56
3.5.2 Participants . . . 56
3.5.3 Results . . . 57
3.6 Discussion . . . 58
3.7 Conclusion . . . 59

Chapter 4 VOREA: Design & Implementation of a Vocal Real-time Auralisation Pipeline **61**
4.1 SIR Production . . . 63
4.1.1 Pre-recording Setup . . . 64
4.1.2 Recording Process . . . 65
4.1.3 SIR Creation . . . 65
4.2 Auralisation Pipeline . . . 66
4.2.1 System Setup . . . 66
4.2.2 Real-time Signal Auralisation . . . 70
4.3 Insights . . . 71
4.4 Discussion . . . 73
4.5 Conclusion . . . 75

Chapter 5 Vocal Auralisation in XR & Its Influence on Perceived Presence & Audio-Visual Quality **76**
5.1 Hypotheses . . . 78
5.2 Experimental Setup . . . 79
5.3 Measures . . . 81
5.4 Experiment 1 . . . 81
5.4.1 Stimuli . . . 82
5.4.2 Participants . . . 84
5.4.3 Procedure . . . 85
5.4.4 Results . . . 86
5.5 Experiment 2 . . . 89
5.5.1 Stimuli . . . 89

5.5.2 Participants . . . 90
5.5.3 Procedure . . . 91
5.5.4 Results . . . 91
5.6 Discussion . . . 93
5.7 Conclusion . . . 96

Chapter 6 Evaluation of an Audio-Visual IVE for Vocal Rehearsal & Performance 97
6.1 Hypotheses . . . 100
6.2 Experimental Setup . . . 101
6.3 Measures . . . 102
6.4 Experimental Design . . . 103
6.4.1 Stimuli . . . 103
6.4.2 Participants . . . 106
6.4.3 Procedure . . . 106
6.5 Results . . . 107
6.5.1 Shift into a Performance Mode . . . 107
6.5.2 Auralisation Driving Exploration . . . 108
6.5.3 Immersion rooted in Realism . . . 109
6.5.4 System Usability Shaped Experience . . . 110
6.5.5 Benefits & Applications . . . 111
6.6 Discussion . . . 112
6.7 Conclusion . . . 115

Chapter 7 Conclusion and Future Work 116
7.1 Summary of Contributions . . . 116
7.1.1 Perceptual Contributions . . . 116
7.1.2 Practical Contributions . . . 118
7.1.3 Technical Contributions . . . 119
7.2 Limitations . . . 119
7.3 Future Work . . . 121
7.4 Closing Statement . . . 123

Bibliography 124

List of Figures

1.1 Diagram illustrating the conceptual framework for this thesis, showing the integration of multisensory stimuli and vocal interaction. 4

2.1 Conceptual representation of audio formats in terms of perceived dimensionality. 9
2.2 Examples of different explicit 3D representations. Figure from Gao et al. [1]. . 14
2.3 Reality-Virtuality (VR) Continuum framework proposed by Milgram and Kishino [2]. 17
2.4 Illustration of three virtual animals and their first-person control, as presented by Krekhov et al. [3]. 23
2.5 Illustration of the natural beaver, robot beaver, and amorphous body character used in the study by Sierra-Rativa et al. [4]. 24
2.6 Nature walk IVE with a gravel path employed in the experiments by Kern et al. [5]. 28
2.7 Modifying the self-perception over the embodied avatar in VR by Buetler et al. [6]. 29
2.8 Reverberation to compensate distance compression virtual environment employed by Huang et al. [7]. 30
2.9 Visual and Spatial Audio Rendering on VR Immersive Virtual Environment by Potter et al. [8]. 31
2.10 Bedroom virtual environment illustrating variations in materials and furniture from Burnett et al. [9]. 32
2.11 Illustration of the three visual representations used in the study by Fargeot et al. [10]. 33
2.12 Comparison of the real space and its 3D model representation from Bargum et al. [11]. 34
2.13 Comparison between the physical concert hall and the high-fidelity 3D model by Ppali et al. [12]. 38

3.1 VR scene from the experiment in which the participant is embodied in the dog avatar while a virtual character trims its fur. 41
3.2 Dog model used in the experiments for participants to embody. 45
3.3 General view of the virtual shelter, including other dogs in cages and the virtual mirror placed in front of the participant embodied as a dog. 46
3.4 Dog avatar being brushed by the virtual character. 48
3.5 The experimenter brushing the participant's back during tactile feedback conditions. 50
3.6 Threat used in Experiment 1: the dog avatar being administered the lethal injection. 52

3.7 Experiment 1 results for one Ownership question (OW1), Change dimension of the SoE and Empathy measure. 53
3.8 Threat used in Experiment 2: a ceiling light falling over the dog avatar. 55
3.9 Experiment 2 results for one Ownership question (OW2), Fur Ownership perception (F1) and Threat perception (TH1), depending on sensory stimuli and participants' gender. 57

4.1 Loudspeaker and Ambisonics microphone setup during the acoustic capture. . . 61
4.2 Left: *Audience perspective* where the source and receiver are spatially placed. Right: *Performer perspective* where source and receiver are co-located. 63
4.3 Grid of positions used for SIR capture in the Abbey Theatre. The front direction is relative to the performer's point of view. 64
4.4 Top: Directivity patterns of the Genelec 1029A loudspeaker at 1, 2, 4, and 8 kHz [13]. Bottom: Mean directivities and standard deviations in the horizontal plane for an *[a]* (a) and an *[f]* (b) at 1 kHz, 2 kHz, 4 kHz, and 8 kHz measured by Pörschmann et al. [14] 66
4.5 Directional interpolation (left) and Translational interpolation (right). 67
4.6 Signal flow setup for two of the twenty folder tracks in the Reaper project. . . . 68
4.7 Setup to measure the amplitude loss caused by the noise-cancelling headphones. 69
4.8 Trimmed section of the SIR showing the direct signal and the Initial Time Delay Gap (ITDG), up to the first reflection, highlighted in red. 69
4.9 Reaper tracks for a single position, with the gain faders of directional tracks in red, and the folder track gain fader in green. 71
4.10 Variation of acoustic characteristics across the stage commonly observed in performance venues. 72

5.1 Virtual room used for the study, with a participant singing while immersed in the experience. 76
5.2 Audio-enhanced HMD setup used in the study, with a lapel microphone attached and positioned close to the participant's mouth. 80
5.3 Real-time vocal auralisation system. 80
5.4 View of the environment and Tasks UI in the Full Visuals (FV) condition (top) and in the Low Visual (LV) condition (bottom). 82
5.5 Black space used in the NV condition, limiting participants' visual input to the instructions UI and their avatar. 83
5.6 Music stand displaying the UI of vocal tasks performed by participants in the IVE. 86
5.7 Experiment 1 mean scores across audio (MA, SA, NA) and visual conditions (FV, LV, NV). 88

5.8 Visual comparison between the virtual room (top) and the physical room (bottom) used in Experiment 2. 90
5.9 Experiment 2 mean scores across audio (MA, NA) and visual conditions (VR, MR). 92

6.1 Point cloud distribution of the GS rendering, with higher density at the stage and close surroundings (highlighted in green). 97
6.2 Locomotion system implemented in the IVEs, showing the ray interactor used to navigate within the defined stage area. 102
6.3 Visual comparison between the Gaussian Splatting (top) and the 3D Modelled (bottom) virtual environments used in the study. 104
6.4 Participant singing while immersed in the GSA environment, wearing the augmented HMD. 106

List of Tables

2.1 Previous studies on Animal Embodiment in VR. 25

3.1 Virtual body perception assessment questions 47

4.1 Comparison of spatial formats with corresponding channel counts and convolution requirements. 73

5.1 Audio & Visual rendering assessment questions 81
5.2 Comparison between a Measured & Synthesised SIR acoustic parameters. . . . 85
5.3 Experiment 1 Aligned Rank Transform (ART) ANOVA results with post-hoc comparisons. 87
5.4 Experiment 2 Aligned Rank Transform (ART) ANOVA results with post-hoc comparisons. 91

6.1 Participant experience, practice environment, and vocal range data. 105
6.2 Themes, subthemes, and key findings from participant responses. 108

List of Publications

1. Flores-Vargas, M., Conolly, R., Bates, E., & McDonnell, R. (2025). *Immersive Vocal Practice: Audio-Visual Evaluation of an Acoustic Virtual Environment for Vocal Rehearsal and Performance.* IEEE International Workshop on the Musical Metaverse.
   Best Student Paper Award

2. Flores-Vargas, M., Bates, E., & McDonnell, R. (2025). *First-Person Vocal Auralisation in XR and Its Influence on Perceived Presence and Audio-Visual Quality.* IEEE International Symposium on Mixed and Augmented Reality.

3. Flores-Vargas, M., Bates, V.E., & McDonnell, R. (2025). *Real-Time Auralization Pipeline for First-Person Vocal Interaction in Audio-Visual Virtual Environments.* Audio Engineering Society 158th Convention.

4. Flores-Vargas, M., Fribourg, R., Bates, E., & McDonnell, R. (2023). *Now I Wanna Be a Dog: Exploring the Impact of Audio and Tactile Feedback on Animal Embodiment.* IEEE International Symposium on Mixed and Augmented Reality.

5. Young, G.W., O'Dwyer, N., Vargas, M.F., Donnell, R.M., & Smolic, A. (2023). *Feel the Music!—Audience Experiences of Audio–Tactile Feedback in a Novel Virtual-Reality Volumetric Music Video.* Arts, 12(4):156.

6. Buck, L., Vargas, M.F., & McDonnell, R. (2022). *The Effect of Spatial Audio on the Virtual Representation of Personal Space.* IEEE Conference on Virtual Reality and 3D User Interfaces, Abstracts.

Chapter 1

Introduction

As Extended Reality (XR) technology evolves, its adoption has expanded across various domains and applications. Once primarily viewed as a platform for gaming and entertainment promoted by the Metaverse narrative, XR has grown beyond this, establishing its strength and core value in areas such as remote collaboration, skill development, and environmental simulations for training purposes. This shift highlights its current relevance, offering plausible, if not realistic, environments that are entirely virtual or that augment the physical world, which facilitates real-time interaction and meaningful engagement. Thus, XR applications and immersive experiences present a significant opportunity not only for technological innovation and for providing solutions across industry, education, and healthcare, but also for advancing the arts and creative industries.

Nowadays, XR systems can provide sensory information from a variety of modalities, particularly visual, via the Head Mounted Display (HMD) or devices' screen, aural, via integrated speakers or headphones, and haptic, using handheld controllers or the device itself in mobile contexts [15]. The multimodal nature of these systems supports their intrinsic capacity to provide coherent multisensory feedback, particularly audio-visual integration. This capability engages users at both sensory and cognitive levels, thus enhancing immersion, the feeling of being present in an environment, and the realism of virtual environments [5]. Consequently, XR allows for fully immersive experiences that combine visual, auditory, and additional sensory channels, shaping how users perceive, interact with, and respond cognitively and emotionally to virtual stimuli [15]. For this reason, it is timely to study the integration process of multimodal stimuli in order to gain a comprehensive understanding of sensory feedback within an Immersive Virtual Environment (IVE) and develop successful applications that benefit from this integration.

Auditory feedback is an essential component in the perception of our environment. It enables us to locate objects even when they are out of our field of vision, distinguish the material and size of objects by the sound they produce, recognise a person's emotional state, and identify environmental cues essential to determine whether sounds are produced in a small or large space [16, 17]. Auditory feedback enables us to decode our surroundings as well as the objects within them. Therefore, how we perceive and interact with the world is greatly influenced by aural information. The effectiveness of an immersive experience, whether Augmented Reality (AR), Virtual Reality (VR), or Mixed Reality (MR), relies on successfully replacing or augmenting

real-world sensory feedback with virtual elements [18] to induce a suspension of disbelief in which users perceive the virtual elements as derived from the real world. Thus, the objective of audio in immersive experiences is to replace or overlay sonic information, and how it propagates and interacts with the environment, with a plausible virtual representation [19].

Audio in XR has received considerable attention due to its potential in IVEs. Beyond replicating the real-world behaviour of auditory phenomena and its conventional uses in traditional media, it can serve several functions such as enabling localised and non-localised action-based feedback on interactions [20, 21], direct users' gaze and attention to elements in the environment even when they are out of the field of view [22], and provide contextual information about the virtual environment and its characteristics [23, 20]. Moreover, it can also help increase perceived realism, and the quality level of one modality can be influenced by cross-modal perception [15, 8]. Additionally, auditory feedback has been recognised for its ability to elicit physical and psychological responses, such as lowering heart and respiration rates and promoting relaxation, which further support meaningful immersive experiences [23, 24]. Beyond user-centred benefits, research has also shown that audio-visual stimuli can help overcome limitations of VR systems, such as visual distance compression [7, 25] and auditory source localisation [26].

Auditory feedback through real-time auralisation (the audible rendering of an acoustic simulation [27]) of one's voice has the potential to create multisensory environments and enable true audio-visual integration. Although some research has investigated the effects of real-time vocal auralisation in both audio-only and VR contexts, the primary focus has been on vocal adaptation, measuring vocal comfort and acoustic parameters such as mean fundamental frequency, Sound Pressure Level (SPL), and spectral moments [28, 29, 30]. Moreover, multimodal integration has demonstrated benefits such as increased presence, improved user experience, and greater ecological validity [31, 32]. In addition, the users' own voice has been employed to evaluate the accuracy of sound auralisation rendering for musical purposes [33, 34, 35]. The acoustic response of the space has been found to have a strong influence on musical performance, as singers rely on the timing and level of acoustic features, such as early reflections and reverberation, to hear their performance and gather information about the dimensions of the performance space [33].

1.1 Motivation

Perceptual constructs and illusions in XR, such as immersion and presence, are essential for creating successful immersive experiences [36]. Research has shown that multisensory feedback influences how users perceive these phenomena, and that the auditory feedback of an IVE, including sound spatialisation and auralisation, enhances these factors and contributes to place illusion, spatial perception, and the perceived realism of the scene [23]. Therefore, the more

accurate or plausible the auditory representation, the stronger the sense of presence, immersion, embodiment, and realism experienced by users [37, 8]. While auditory feedback has been shown to amplify physical and psychological effects, offering additional benefits to IVEs [15], most of this research has examined sounds produced by external sources within the environment, rather than self-produced sounds and those produced in relation to one's own body.

Prior studies suggest that auditory feedback can influence body representation and embodiment. However, in the specific context of embodying dysmorphic avatars, the impact of auditory cues on Sense of Embodiment (SoE) appears negligible [38, 39]. Whether this outcome stems from methodological limitations (e.g., a small sample size and a non-immersive setup [39]) or from the particular nature of the sounds employed (movement-generated audio such as wind during wing flapping [38] or hand claps [39]), audio feedback requires further exploration. Moreover, the lack of controlled investigations isolating auditory input from other modalities makes it difficult to disentangle its specific contribution to SoE. While previous work on audio-tactile stimulation has considered only punctual sounds [6], the role of continuous sounds remains unexplored in this context, and it is unclear whether such findings could extend to the perception of a whole altered avatar.

Vocal perception and production have been studied in both audio-only and VR contexts to examine how one's own voice is perceived within an acoustic environment and how this perception influences vocal production. However, real-time vocal auralisation remains underexplored in terms of its potential to shape the overall experience in IVEs, its effect on the sense of presence (the feeling of being in or interacting with the virtual environment [40]), and its impact on perceived audio-visual quality in VR and MR contexts. Similarly, while IVEs have been investigated in musical contexts as training methods that support singers in improving vocal technique and skills, as well as tools for practice and performance, further research is required to understand how high-fidelity audio-visual representations of spaces influence singers' experiences, and how these factors contribute to creating a realistic and usable environment for singing-related endeavours.

1.2 Thesis Focus

Based on the motivations outlined above, the main focus of this thesis is to evaluate the sensory modalities that contribute to the perception of an acoustic environment in XR, namely visual and auditory sensory feedback, using vocal interaction to examine factors of multimodality.

A preliminary study investigates the potential of auditory feedback, when paired with visual and tactile feedback, to promote and enhance the SoE. This study serves as an initial exploration of auditory feedback intrinsically associated with one's own body within a multimodal framework. As an early exploration, it helps frame a methodological and conceptual foundation for investigating sounds more closely connected to our body and intrinsically experienced from a first-person perspective, such as one's own voice.

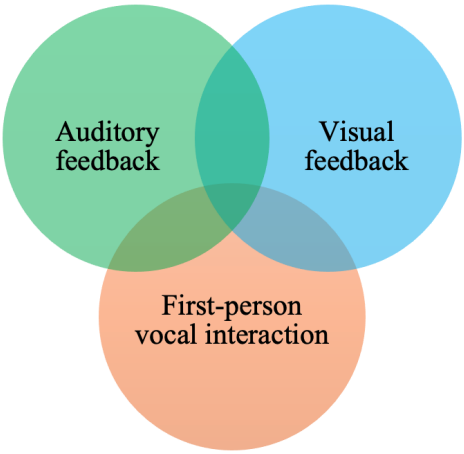


Figure 1.1: Diagram illustrating the conceptual framework for this thesis, showing the integration of multisensory stimuli and vocal interaction.

The main body of this thesis investigates one's own voice as the primary element of exploration and embodied interaction due to our heightened sensitivity to changes in its sound. In this context, the work examines the audio-visual perception of real-time vocal auralisation in IVEs, specifically its impact on the sense of presence and perceived audio-visual quality across VR and MR (see Figure 1.1). Moreover, since the perception of sensory information is strongly influenced by the attention involved in the task being executed [15], the use of one's own voice as a means to explore the IVE is particularly appropriate, as it can promote a sense of agency while providing a clear and well-defined task for perceptual evaluation.

Furthermore, it examines how singers experience and evaluate a multimodal VR-based theatre as a potential medium for vocal performance and practice, extending prior explorations of IVEs. We investigate how an audio-visual representation of a virtual performance space, produced by capturing photorealistic visuals paired with real-time vocal auralisation and the dynamic exploration of the acoustic response across the stage, influences singers' experiences, and whether it is perceived as a plausible and effective environment for vocal performance and rehearsal.

Throughout this thesis, we use the term *embodied auditory feedback* to refer to body-mediated sounds that arise through interactions with objects and the environment (e.g., brushing one's teeth), a notion further discussed in Section 2.7.1.

The main research questions addressed in this thesis are:

RQ1 How does self-produced and embodied auditory feedback affect perceptual constructs and psychological responses in multimodal IVEs?

Because the effectiveness of immersive experiences relies on eliciting psychological effects, does auditory feedback, congruent with other modalities, enhance the sense of immersion, presence, embodiment, and realism?

RQ2 How does rendering fidelity across visual and audio feedback influence perceived quality and audio-visual integration in IVES using real-time vocal interaction?

If a user experiences a virtual environment with varying levels of visual and auditory fidelity, these variations might alter the overall perceptual experience. Does adding auditory feedback enhance perceived quality by reinforcing audio-visual integration?

This question is especially pertinent given the growing popularity of standalone XR devices with constrained computational resources and the use of passthrough features in current HMDs, which allow MR experiences that vary the rendering level of virtual content.

RQ3 To what extent does a high-fidelity audio-visual representation of a virtual performance space provide a plausible environment, and a usable tool for vocal practice and performance?

Immersive technologies are increasingly used for audio-visual scenarios in skill development and training. Could a virtual representation of a space that enables users to hear themselves as if talking or singing within it through dynamic auralisation enhance plausibility (the overall credibility of the virtual scenario based on users' expectations towards an equivalent real event [41]) and thus serve as a useful tool for vocal-related purposes in musical contexts?

1.3 Scope

This thesis investigates the effect of auditory feedback within a multimodal context on the perceptual experience of IVEs. It should be noted that the studies presented concentrate on perceptual evaluation within audio-visual setups that envelop the user in a fully rendered environment. The work employs various rendering qualities and techniques in VR as well as passthrough rendering in MR. Therefore, it does not consider AR scenarios, where the real world is only partially overlaid with virtual elements and lacks a complete virtual depiction.

Chapters 5 and 6 focus on one's own voice as the primary element of exploration and interaction within the audio-visual environment, particularly through real-time vocal auralisation, to evaluate audio-visual quality, multimodal integration, and their influence on the sense of presence, immersion, and perceived realism in IVEs. However, they do not investigate vocal adaptations, such as spectral or SPL variations, or specific vocal features.

The definitions of immersion, presence and realism are complex due to their multidimensionality, causing them to adopt different interpretations depending on context and application. This thesis follows the updated definition of presence proposed by Chandio et al. [40], which encompasses the full spectrum of XR experiences, as outlined in Section 2.4.2. This approach

is adopted because the term establishes a clear connection to self-presence and plausibility, which is well-suited to the primary focus of this work. While presence can also relate to social presence, this dimension is not central to the studies presented here. Regarding immersion, the term is reserved to describe the immersive experience, subscribing to Agrawal and Bech's [42] notion of mental involvement, which causes dissociation from the physical world, as discussed in Section 2.4.1. As for realism, Section 2.4.3 presents an objective and a subjective perspective. Throughout this thesis, the term realism refers to the subjective experience, closely related to plausibility, understood as the extent to which sensory feedback matches users' expectations and evokes the psychological effects of real-world qualities.

Lastly, due to current real-time audio input constraints related to latency in XR audio rendering systems, we designed a real-time auralisation pipeline that leverages three-dimensional Spatial Impulse Responses (SIRs) for multimodal research applications in XR that require users to perceive the real-time vocal auralisation of their own voice within the IVE. However, this work does not introduce a novel method for creating SIRs and builds upon existing methods [43, 44].

Its novelty lies in supporting translational degrees of freedom using of SIRs captured from a performer's perspective, where the source and listener are co-located (see Section 4.1), to enable users to dynamically perceive the auralisation of their own voice within an area of the IVE. In addition, the system developed allows for the comparison of both measured and synthesised SIRs in multimodal IVEs.

1.4 Contributions

Key contributions have been made in the multimodal perception of self-produced sounds in IVEs. These contributions offer novel insights into auditory feedback and the relationship between audio-visual integration and perception, and also highlight practical implications explored in this dissertation.

Through **perceptual experimentation**, we examined how self-produced and embodied auditory feedback shape perceptual constructs within IVEs. The main contribution presents findings on the influence of real-time vocal auralisation on the perceived sense of presence and audio-visual quality across varying levels of rendering resolution in VR and MR. Our results offer insights into the relevance of auditory feedback as part of multimodal stimuli and highlight the importance of real-time auralisation for embodied vocal feedback as a means of interaction in immersive experiences. Accordingly, we identify key considerations regarding the necessary level of auditory feedback fidelity, depending on the type of experience and the degree of realism required. Additionally, we gained insights into continuous embodied auditory feedback when combined with visuo-motor or visuo-tactile feedback to support a SoE in multimodal IVEs.

To complement these findings, we present a **qualitative evaluation** of the usability and perceptual qualities of an audio-visual representation of a virtual performance space, generated by

capturing high-fidelity photorealistic visuals together with the acoustic response from a physical theatre. We also provide a comparison between this virtual representation and conventional 3D model-based virtual environments that lack vocal auralisation, commonly used in voice-related applications. Moreover, we highlight the effectiveness of the environment as a usable tool for vocal performance and rehearsal for performing arts professionals, as well as for applications that benefit from this approach. Finally, the results serve to validate our vocal auralisation method, which proves useful for designing high-fidelity audio-visual representations and for producing highly realistic multisensory IVEs.

Finally, to facilitate the perceptual and qualitative evaluations, a **real-time vocal auralisation pipeline** is proposed, which enables users to experience the acoustic response of a virtual environment within an interactive area, allowing exploration of stage acoustics at various positions and revealing distinct acoustic characteristics. Practical guidelines are also provided regarding impulse response creation, the rendering workflow, audio-visual integration, as well as latency and computational considerations.

1.5 Summary of Chapters

The remainder of this thesis has been structured into the following chapters:

Chapter 2 provides an overview of previous and related work on auditory feedback in XR, multimodality and perceptual illusions.

Chapter 3 presents a preliminary approach to multimodality and embodied auditory feedback, examining audio and audio-tactile feedback and their influence towards animal avatar embodiment.

Chapter 4 outlines in detail the design and development of VOREA, an audio auralisation pipeline that supports real-time first-person vocal auralisation using SIRs and gain interpolation. It also presents practical insights regarding capture, computational constraints, and the relevance of the system for real-world VR applications.

Chapter 5 reports on the perceptual evaluation of how real-time first-person vocal auralisation affects users' sense of presence and audio-visual quality across varying rendering resolutions and visual modalities, encompassing both VR and MR environments.

Chapter 6 examines the usability and audio-visual plausibility of a VR-based theatre, as assessed by professional singers. It grounds the perceptual evaluation of real-time first-person vocal auralisation in a practical scenario for vocal rehearsal and performance.

Chapter 7 summarises our contributions to the field and also discusses plans for future work.

Chapter 2

Background

2.1 Introduction

The main focus of this thesis is to investigate how auditory feedback, and its multimodal integration with visual and haptic feedback, can enhance psychological effects such as immersion, presence, realism, and embodiment. Moreover, it explores how self-produced sounds, or those arising from bodily interactions such as the voice, can serve as a means to assess high-fidelity audio-visual environments, evaluate their perceived quality, and examine their potential in musical contexts such as vocal rehearsal and performance. To this end, this chapter presents the relevant literature that supports this thesis, which is structured as follows:

First, we discuss the foundations of audio and visual rendering for XR environments. We explore the importance of auditory feedback, focusing on spatial audio and sound auralisation, and address the different considerations involved in recording and rendering techniques. In addition, we examine different 3D scene representations, their rendering methods and trade-offs, as well as considerations of rendering fidelity and its implications for XR applications.

Second, we address perceptual constructs in XR experiences. We examine their different definitions and concepts, discuss factors such as sensory stimuli that shape and promote them, and explore the resulting physical responses and perceptual illusions.

Third, we discuss multimodality, its relevance, and examine key factors for the multimodal integration of sensory feedback. We then address audio-visual integration in XR environments, exploring the physical and psychological effects that arise from this integration, as well as its potential for enabling new interactive mechanisms and overcoming technical limitations of XR systems.

Finally, we focus on the perception of self-produced audio feedback and its influence on self-perception and the environment. We examine one's voice as a key element in audio-visual IVEs, investigate how XR influences vocal perception and production, and explore its potential to support vocal practice and performance.

2.2 Audio Rendering

Auditory feedback is essential for perceiving, decoding, and developing a comprehensive representation of the environment. Thanks to the high degree of immersion, multisensory capabilities and dynamic possibilities offered by XR systems, such as head and position tracking, it is feasible to provide sensory information all around the user, just as it happens in the real world. Immersive or 3D audio formats extend beyond the horizontal plane by incorporating height information, enhancing spatial perception and contributing to a more immersive auditory experience (see Figure 2.1) [45]. To accomplish this, two audio rendering strategies are fundamental for creating a 360° soundscape: audio spatialisation and auralisation.

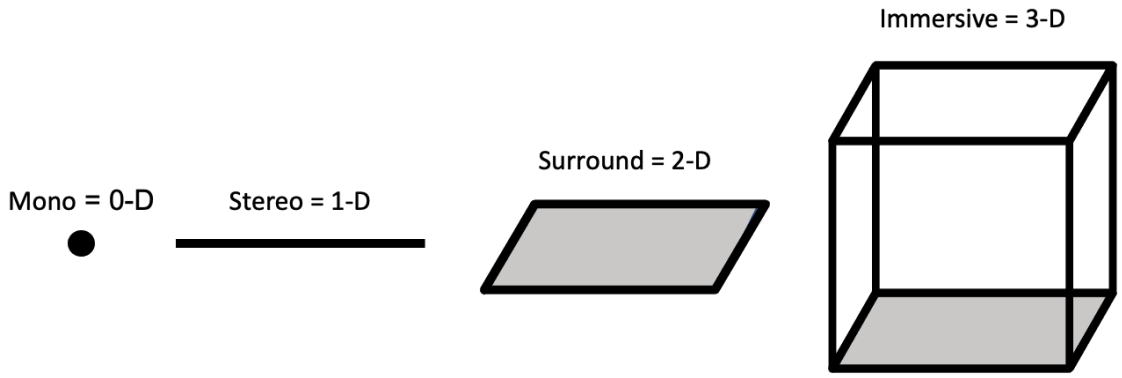


Figure 2.1: Conceptual representation of audio formats in terms of perceived dimensionality.

2.2.1 Spatial Audio

Spatial audio, often known as 3D audio, is any audio reproduction technique that allows sounds to be rendered with the perceptual qualities required to provide them with a distinct direction in the horizontal and sagittal planes, as well as distance from the listener [46, 47]. It refers to the artificial placement of a sound source at a particular location within a 360° space, enabling us to discern between sound sources in our surroundings, as well as to comprehend their distribution, behaviour, and interaction within the environment [48]. This is achieved through a collection of binaural perceptual features of audio signals, such as Interaural Time Differences (ITDs) (produced by the delayed arrival of the audio signal to the ear that is located further away from the source), Interaural Level Differences (ILDs) (caused by the decrease of sound intensity at the ear farther from the source due to head occlusion), spectral cues, and amplitude, that allow the auditory system to determine a sound source's specific angular position (azimuth and elevation) and distance from the listener [49]. These perceptual features are derived from the temporal and spectral filtering of the audio signals arriving at each ear, derived from one's head, pinnae (external part of the outer ear) and torso [50].

Spatial rendering can be achieved using loudspeaker arrays that incorporate both horizontal and vertical channels. However, most XR systems and immersive experiences rely on audio reproduction over headphones, which requires binaural rendering. This process filters the audio signal with Head-Related Transfer Functions (HRTFs), direction-dependent functions that describe the temporal and spectral cues of the sound arriving at each ear to simulate the effects of the listener's body. Thus, providing the auditory system with the spatial information needed for sound localisation and a natural acoustic perception [51, 26]. Consequently, sound spatialisation in XR contexts relies considerably on HRTFs to reproduce realistic auditory phenomena, as individual morphological variations are critical to localise audio sources accurately [50, 52]. Moreover, they are critical for the externalisation of auditory events, that is, the perception of sound sources as originating outside the head, in the way we perceive them in the real world [53]. This occurs when audio signals contain all the cues naturally picked up by the ears in a listening setting, incorporating properties of the source, the listener, and the environment [54].

Due to the current limitations, modelling individual, or personalised, HRTFs to reproduce the correct auditory phenomena for each listener is time-consuming and technically challenging. Therefore, most sound spatialisation systems in XR currently rely on generic HRTFs (non-individualised), measured with a binaural dummy head or averaged across a group of people [55], as it makes it easier to deliver spatialised sound across a wide range of applications and head-tracked devices [26]. However, using an HRTF different from the listener's can have a negative impact, leading to effects such as front-back confusion, localisation errors (particularly in the sagittal plane), and a lack of externalisation [52, 20]. Poor externalisation, in particular, can cause auditory events to collapse and be perceived inside the head rather than in the surrounding environment, directly influencing the auditory experience. While generic HRTFs may result in perceptual issues [56, 57], techniques such as pairing auditory feedback with visual stimuli and allowing a period of recalibration and training [26], or using active head movements enabled by head-tracking devices, which modify the source-listener relationship [58], can mitigate these limitations and improve sound localisation. It is important to note, however, that although the use of generic HRTFs remains standard practice, this may change as methods for personalisation continue to advance.

Spatial audio production for XR can be achieved through different techniques for capture and rendering. Capture methods consist of recording a 360° sound field, in which sounds are recorded using techniques that directly encode perceptual features between sources and the microphone (the listener) [59]. Binaural recordings, for example, capture sound using a dummy head with two microphones placed in the ears to record the temporal and spectral features that provide information about the position of sound sources. While binaural recordings provide an accurate spatial representation, they are not often used in XR, as they do not support listener individualisation or head tracking, which is key to immersive technology [60]. Similarly,

Ambisonic capture uses a spherical microphone array to record the 360° sound field, which is then encoded into spherical harmonics for subsequent rendering [61, 59]. Rendering consists of the virtual placement of sound sources, providing them with the perceptual features needed to create the illusion of a sound originating at a specific location around the listener [45, 62]. Different techniques and formats are available for encoding, including binaural and Ambisonic methods. However, additional algorithms such as Vector Base Amplitude Panning (VBAP), often leveraged in object-based rendering, can also be used [62]. Ultimately, due to its flexibility and dynamic spatial resolution, Ambisonics has become the standard approach for sound spatialisation in XR, particularly when combined with binaural decoding (HRTF filtering to generate left and right ear signals), as most XR systems rely on audio playback over headphones [60]

Previous research has recognised the benefits and functions of auditory feedback. Some of these stem directly from inherent perceptual qualities such as its omnidirectionality and continuous perception, which can be leveraged to provide users with information outside the field of view as well as offering a continuous stream of input without overloading the visual modality [63, 64, 23]. Moreover, the inclusion of audio feedback opens possibilities for interaction mechanisms and affordances within IVEs. Audio cues can be used to guide users' gaze to critical elements [65], provide constantly adapting feedback [23], deliver contextual sounds and soundscapes that convey information about the environment [66], and provide sounds that result from users' interactions, whether produced by objects themselves or as responses to user actions [21]. In addition, several studies have shown that auditory stimuli can enhance the perceived sense of presence and immersion [32, 67, 5], further underscoring the importance of audio feedback in shaping immersive experiences.

2.2.2 Virtual Room Acoustics & Auralisation

Auralisation is the process of virtually rendering, by physical or mathematical modelling, the sound field of a source. It simulates the behaviour of sound waves travelling through air and their interaction with boundaries and objects, resulting in a set of acoustic features inherent to the geometry of an environment [68]. By simulating acoustic phenomena such as early reflections and reverberation, it is possible to generate realistic acoustic responses that convey information about the size of the space in which the sounds occur [69], improve sound externalisation, and reduce front-back confusion [53]. In the context of XR, rendering the acoustic characteristics of a reverberant space has been shown to provide several perceptual benefits. It further supports the externalisation and localisation of sound sources [70, 71], and improves distance perception of sound sources in space [72]. However, its benefits expand beyond auditory perception. Sound auralisation contributes to IVEs by introducing an additional sensory dimension and increasing the user's spatial awareness, presence, and overall sense of immersion, which are key factors in XR experiences [73, 74].

A common technique for sound auralisation is the use of SIRs, which capture the acoustic characteristics of a specific environment, including its size, boundaries, objects, and materials. SIRs can be generated using two primary methods: measured (captured in real spaces) and synthesised (produced through computer-based models) [33, 34, 35]. SIRs, in contrast to mono room impulse responses (RIRs), capture the directional elements of a room's acoustic properties relative to the listener's orientation and position within the space, which is essential for spatial acoustic rendering in IVEs [43]. Measured SIRs are obtained by capturing the sound-field characteristics of a real-world room by exciting the acoustic environment with an audio signal containing constant energy across the audible frequency range, using techniques such as Maximum Length Sequence (MLS) and Exponential Sine Sweep (ESS), and recording the acoustic response binaurally or using Ambisonics [75]. MLS employs pseudo-random broadband noise to capture time-based room acoustic parameters. While this technique can provide a good signal-to-noise ratio, energy-based parameters, such as the Clarity index, cannot be calculated from SIRs produced using this method [76, 33]. ESS measures impulse responses using a logarithmic sine sweep covering the frequency range [76]. The recorded sweep undergoes a deconvolution process using a time-inversed filter to produce the impulse response of the measured room. ESS has become widely used since it offers several advantages, such as avoiding time-aliasing and providing robust non-linearity rejection, which allows the separation of harmonic distortion components [77]. Finally, SIRs are employed as the virtual acoustic representation of an environment, which is convolved with an audio signal to simulate its interaction with the acoustic space. This method delivers a realistic acoustic representation; nevertheless, it is limited to capturing real-world spaces and has constraints associated with dynamic acoustic rendering [33].

Synthesised SIRs may be obtained using different approaches to model the acoustic response of a virtual environment. Each of them offers distinct benefits and trade-offs in terms of rendering accuracy and computational resources. Geometric-based methods, such as the Image Source Method and Stochastic Ray tracing, estimate the acoustic characteristics of an environment by modelling sound propagation through reflections based on geometric primitives in 3D space [78, 51]. They take a simplified approach to sound propagation and reflection computation, ignoring several acoustic phenomena such as diffraction, interference, phase, and scattering [73]. Due to the reduced approximation, this approach is efficient in terms of computational power and time, allowing for real-time rendering and providing a good balance between plausibility and performance [79]. Wave-based methods involve subdividing the space into uniform discrete grids, which update pressure amplitude in each cell at each time step using numerical solvers [51]. This approach is more accurate in predicting both direct sound and reflections, as well as diffraction, occlusion, and scattering events, and it is useful for designing sound systems in complex environments where reflections and diffraction play a significant role in shaping the sound field. However, it is substantially more computationally expensive and does not allow for real-time

modifications [73]. Delay network-based models, such as Feedback Delay Networks (FDNs), have received considerable attention due to their perceptual performance and computational efficiency [80, 81]. These methods use networks of delay lines to model room acoustic properties by connecting delay lines within a feedback loop modulated by absorption coefficients, producing rich late reverberation [82]. Recent approaches have been developed to leverage the strengths of delay-based models and GA. For example, Scattering Delay Networks (SDNs) connect delay lines through scattering nodes, allowing first-order early reflections to be modelled accurately while reducing the accuracy and level of higher-order reflections that are perceptually less significant [8]. Similarly, models based on the Room Acoustics Rendering Equation (RARE) employ GA for early reflections and delay-based networks to generate late reverberation [81].

Sound auralisation and the tracking capabilities provided by current VR systems have led to the development of Acoustic Virtual Reality (AVR), an auralisation method in which the sound source and the receiver (most often the user) are not fixed and can move freely in position and rotation while interacting with the IVE [73]. This responsive framework introduces design and computational challenges, as the auralisation system must update the acoustic response in real time to reflect changes in the source-receiver relationship, including the user's position and head orientation within the virtual space (e.g., when users are in the centre or on the side of a room) [83]. AVR is commonly presented with a 3D visual representation of the architectural space, typically in VR, using HMD head tracking to update the acoustic response [73]

2.3 Visual Rendering

Visual feedback is the dominant and most stable sensory modality for most humans, serving as the primary channel for decoding the environment [15, 84]. XR systems are inherently visual, with perception and interaction largely centred on digital visual content [18]. This becomes particularly evident in immersive VR applications, where a complete three-dimensional representation of the environment is generated. Producing plausible visual feedback is essential for XR experiences to enhance spatial awareness and promote immersion and presence [36]. Therefore, to provide meaningful visual feedback in XR, two aspects are particularly critical: the technique of 3D scene representation and the level of visual fidelity with which it is rendered.

2.3.1 3D Scene Representation

Generating 3D scenes and objects can be achieved through a variety of techniques. These modelling approaches are grounded in different theoretical methods and provide distinct advantages and limitations, shaping their suitability for specific domains such as gaming, architectural visualisation, filmmaking, and immersive experiences [85]. In XR applications, scene representation is fundamental for depicting captured environments that can be processed, analysed, and manipulated. The choice of representation significantly influences both the fidelity and the computational

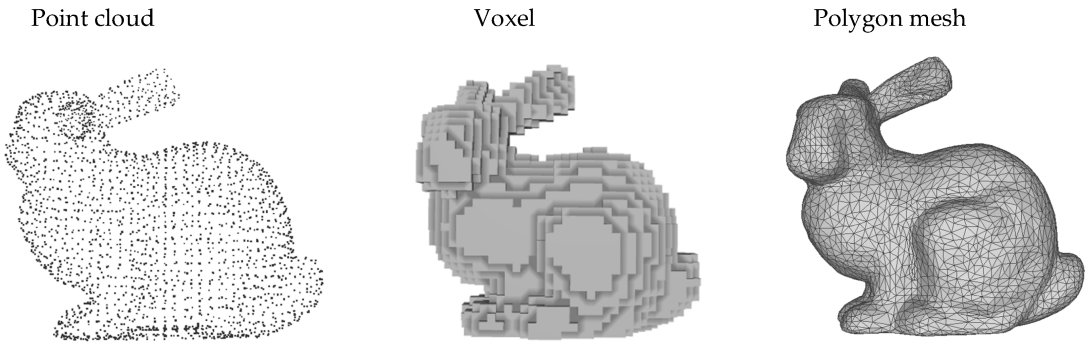


Figure 2.2: Examples of different explicit 3D representations.
Figure from Gao et al. [1].

efficiency of rendering, which are critical factors in XR systems. 3D representations can be categorised as either explicit or implicit, depending on the mechanisms used to encode and describe the characteristics of space and objects [86, 87].

Explicit representations describe the spatial structure of 3D objects using discrete geometric elements, allowing precise control over geometric detail. This level of control makes them widely used in tasks such as modelling, rendering, and simulation, as it enables efficient editing, optimisation, and real-time rendering [86]. Polygonal meshes define 3D shapes and surfaces by encoding the spatial coordinates of vertices along with their adjacency relationships through edges and faces [88] (see Figure 2.2, right). This technique is by far the most commonly adopted among explicit representations, owing to its rendering efficiency and support for geometric transformations, which make meshes particularly well suited for modelling the surfaces of 3D scenes [89]. Point clouds are also a common technique, referring to an unstructured set of points that approximate an object's surface [90] (see Figure 2.2, left). They are commonly generated using sensing technologies such as depth sensors, Light Detection and Ranging (LiDAR), or structure-from-motion to sample object surfaces in the physical world [90, 85]. Lastly, voxel grid is a technique that divides space into a uniform 3D array in which each voxel is assigned specific properties such as density and colour, providing a volumetric representation of the environment [87] (see Figure 2.2, centre).

Implicit representations encode the geometric characteristics of 3D space through mathematical functions [91, 86]. These techniques model a scene as a continuous function that represents features of the space, such as colour, light distribution and signed distance, without defining the scene using explicit geometric structures [86]. This allows for differentiable (the ability to quantify how changes in its parameters impact the output) and often compact representations of complex scenes [92]. Signed Distance Fields (SDF) map a spatial position to a signed distance, defining a surface as its zero-level set [93, 85], where the sign of the value (negative inside and positive outside) encodes the interior-exterior relationship of the geometry [93, 85, 86]. Building

upon this method, Neural Radiance Fields (NeRF) extend the representation by mapping spatial position and view direction to view-dependent colour (or radiance) and volume density. This is achieved by using a high-frequency positional encoding, which allows capturing high geometrical detail, including reflections and opacity, through differentiable volume rendering [91]. As a result, NeRF has gained significant popularity and relevance, particularly for modelling high-fidelity static scenes [94].

While NeRF representations set a new standard for photorealistic rendering, their inherent challenges, such as high computational demands that require dense sampling throughout the 3D space, have limited their use [95, 94]. In recent years, 3D Gaussian Splatting (GS) has emerged as an innovative, flexible, and expressive technique for modelling and rendering 3D scenes, producing results comparable to NeRFs while relying on existing graphics pipelines [94]. Its rendering method can be categorised as a hybrid representation, as it integrates explicit geometric primitives with implicit field functions [96]. GS represents the scene using a collection of millions of learnable 3D Gaussians as the point primitives distributed in space. This approach efficiently approximates both the geometry and appearance of a scene, providing volumetric information and explicit attributes including colour and opacity [94]. Such an approach enables efficient computation and real-time rendering, while offering significant advantages in training speed and visual quality compared to traditional implicit NeRF methods [96]. GS leverages explicit point-based representation and efficient rasterisation to approximate volumetric radiance fields, enabling high-fidelity scene renderings with smooth surfaces, realistic textures, and visual effects such as depth-of-field and soft shadows [95]. Several studies have shown the effectiveness of GS in modelling objects and environments for XR applications [95, 27, 97]. While there are different trade-offs compared to other methods (structured light scanning, LiDAR scanning, NeRFs), it is an effective 3D reconstruction model for creating intricate and detailed digital 3D models [98, 27]. In addition, its efficient computation, rendering, and optimisation make it suitable for environment and scene reconstruction, delivering visually compelling IVEs and enabling high-quality visualisation on hardware with limited capabilities, such as standalone XR devices [99, 94, 97]. Given these advantages, GS was chosen as the scene representation method for the study described in Chapter 6, which aims to produce a high-fidelity audio-visual environment.

2.3.2 Visual Rendering Fidelity

Visual fidelity refers to the quality and accuracy with which virtual content, whether a single object or an entire virtual scene, is represented, and how it compares to a ground truth or to its real-world counterpart [100]. Different levels of visual fidelity can be achieved by increasing the count or density depending on the underlying representation, such as polygons or point clouds [88, 95]. High-fidelity 3D scene generation ultimately involves a combination of geometry,

texture, lighting, and multi-view consistency [92, 85]. Thus, different methods address these attributes with inherent strengths and limitations, leading to trade-offs in rendering fidelity.

Trade-offs are inherent across representation methods, as each offers unique advantages and faces distinct challenges in scene visualisation. Neural representations based on volume rendering achieve high-fidelity and photorealistic visual quality but are limited by capture conditions such as lighting, computational efficiency, real-time performance, and compatibility with rasterisation-based graphics pipelines [91, 86]. In contrast, the flexibility of mesh representations allows for efficient computation of 3D models and scenes, thanks to their support for geometric transformations and real-time rendering [88, 89], offering a balance between visual plausibility and performance [88, 86]. However, they face challenges when representing real spaces at a photorealistic level. Therefore, the choice of representation directly impacts rendering fidelity, computational efficiency, and the range of applications for which the model can be employed. Furthermore, device-specific features and limitations, such as resolution, refresh frame rate, latency, and field of view, impose practical constraints on the possible level of rendering fidelity that can be achieved [92, 101].

Rendering fidelity is a crucial factor in XR applications, particularly in VR, due to the amount of visual content that must be rendered to create a 3D environment around the user. While the tendency is to pursue higher visual fidelity and more accurate representations, the computational limitations of HMDs and the risk of simulator sickness introduce a fundamental trade-off between refresh frame rate and visual fidelity [92]. High-resolution in VR can provide more detail of virtual environments, but they require a better hardware setup with higher performance [101]. At the same time, higher refresh rates, particularly 120 fps and above, significantly reduce symptoms of simulator sickness [102]. Overall, it is essential to examine in detail the benefits of different 3D scene representations with respect to fidelity and performance, while identifying an appropriate balance between resolution and refresh frame rate, as all of these factors directly influence and shape the user's experience in XR applications. This is particularly relevant as these factors contribute to perceptual illusions and constructs such as immersion and presence, which are central to how users perceive and engage with virtual environments.

2.4 Perceptual Constructs

The power of immersive experiences relies on tricking the brain into believing that virtual stimuli are real, creating a physical and psychological connection to them, and generating different perceptual experiences and responses [103]. Immersive technologies achieve this by replacing or augmenting the real world with virtual content. However, XR is an umbrella term that encompasses different forms of reality, distinguished by the level of virtualisation or how much the real world is replaced or augmented [40], and the level of interaction of the user with the virtual objects in the scene [41]. The level of virtualisation can be understood through taxonomies

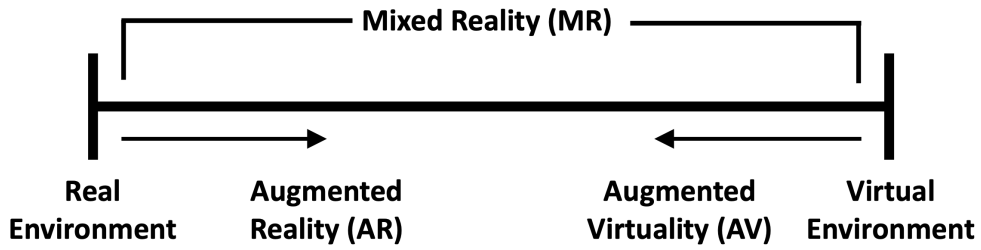


Figure 2.3: Reality-Virtuality (VR) Continuum framework proposed by Milgram and Kishino [2].

such as the Reality-Virtuality Continuum proposed by Milgram and Kishino [2], where mixed reality (MR) is defined as the blend of real and virtual elements determined by the amount of virtual content and real-world information, with the real and virtual environments situated at the extreme opposites, and all other forms falling within this continuum (see Figure 2.3). While this taxonomy provides a framework for defining MR, it offers a foundational understanding of how XR can be classified according to its level of virtualisation. This distinction results in three main categories:

- **Virtual Reality (VR)** replaces the physical world with a virtual environment, fully immersing the user in a digital space, enabling a sense of being inside and interacting with that environment.
- **Augmented Reality (AR)** overlays virtual elements onto the physical world, allowing users to see digital content while remaining aware of their surroundings.
- **Mixed Reality (MR)** represents a continuum where virtual and real elements coexist and interact in real time, allowing users to interact with digital content as if it were part of the physical world, unlike AR.

Technology-mediated stimuli provide users with a perception of their immediate surroundings or objects within the physical world by engaging the sensory system, such as vision, sound and touch [104]. It actively combines bottom-up processing of sensory inputs with the user's prior experience, expectations, and beliefs, drawing upon their previously existing model of the world [18]. The outcome is the construction of perceptual experiences characterised by immersion, sense of presence, realism and sense of embodiment, that promote physical responses and perceptual illusions, which are key to creating authentic and meaningful immersive experiences [40, 41]. These perceptual constructs frame the development of the research questions and constitute essential measures explored throughout the studies presented in this thesis.

2.4.1 Immersion

Immersion is a fundamental construct in XR and is widely applied when evaluating immersive experiences. However, there is a lack of consistency in how the term is used, as its definition varies across the literature [41, 42]. Although many definitions have been proposed that highlight different aspects, the technical and psychological perspectives are the two primary viewpoints.

The most generalised definition in the immersive experiences literature is that of Slater [105], who describes immersion as "what the technology delivers", or "the extent to which the computer displays are capable of delivering an inclusive, extensive, surrounding, and vivid illusion of reality to the senses of a human participant" [106]. This perspective focuses on the objective properties of the system, such as fidelity, interaction and multisensory capabilities, which in turn can elicit subjective experiences as a consequence. Other definitions emphasise the psychological experience, framing immersion primarily as a subjective mental state experienced by the individual. Witmer and Singer [107] defined it as "a psychological state characterised by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences". Similarly, Murray [108] described immersion as a "metaphorical term derived from the physical experience of being submerged in water. We seek the same feeling from a psychologically immersive experience that we do from a plunge in the ocean or swimming pool: the sensation of being surrounded by a completely other reality, as different as water is from air, that takes over all of our attention, our whole perceptual apparatus".

In recent years, research has underscored this lack of consensus, emphasising the difficulty of providing a clear and unified definition, as the construct is multidimensional and acquires a different meaning depending on the context and field of study [41]. Agrawal and Bech [42] proposed a definition built from descriptions within the literature: "Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world". While this definition provides a more rounded understanding of immersion as it highlights the cognitive processes that lead to the mental disassociation from the awareness of the physical world (perceptual illusion of non-mediation), the definition of the term remains inconsistent between the two viewpoints. Therefore, Lee [41] proposed standardising the terminology by differentiating Immersive System and Immersive Experience, which consequently clarifies the relationship between system (cause) and experience (effect).

2.4.2 Presence

Effective immersive experiences rely centrally on the concept of presence. Although its definition is more consistent throughout the literature, as with immersion, different definitions have been

adapted to different contexts and applications. It is commonly defined as the subjective feeling of "being there" or being fully immersed in the environment even when one is physically situated in another [106, 107]. This definition relies on the conceptualisation of presence as a human reaction to immersion [105], which is constituted by two orthogonal illusions: Place Illusion (PI), the illusion of being in the place despite knowing that one is not, and Plausibility Illusion (Psi), the illusion that what is being experienced is happening despite knowing the events are virtually generated [109].

However, this definition is debated in AR and MR contexts, where the sense of "being there" is often less pronounced due to the level of virtualisation provided by the immersive system. Therefore, Chandio et al. [40] define presence as "the subjective experience that a particular object exists in a user's environment, even when that object does not". This definition appears more suitable across AR and MR contexts, highlighting the sense of being in or interacting with the virtual environment, particularly when situated in a physical location that is contextually congruent with the virtual experience [110]. This definition emphasises plausibility as a central quality, which correlates to the level of congruence, or the objective match between system stimuli and user expectations across sensory, perceptual, and cognitive layers [111, 112].

While this definition is well-suited to cover XR experiences, Lee [41] recognises the need to distinguish presence in different contexts to conceptualise its relationship with immersion, building on Biocca's [113] work. Physical presence, analogous to PI, refers to the sense of being physically situated in the virtual environment and is mediated by the level of sensory feedback provided by the immersive system. Social presence, similar to co-presence, describes the sense of being together and interacting with another intelligence, mediated by the perceived level of intelligence, intentions, and sensory impressions of the other. Lee [41] extends this by defining it as "the sense of perceiving virtual intelligences in a virtual or real environment as perceptually and/or socially real." Finally, self-presence refers to the psychological and emotional experience of the user within the virtual environment, which Lee redefines as "the sense of being embodied into a virtual self in sensory and/or cognitive manners". Although Lee provides a clear distinction between different forms of presence, the definition proposed by Chandio et al. [40] is adopted because it offers a consistent conceptualisation across AR, MR, and VR contexts.

Presence, despite its varying definitions, has been extensively studied and is recognised as a fundamental construct shaping users' experiences in IVEs, promoting physiological and emotional responses. For example, Meehan et al. [114] investigated how the perceived sense of presence in IVEs evokes physiological responses, such as changes in heart rate and skin conductance level, as happens in real environments. By comparing participants' reactions to a non-threatening virtual room and a stressful virtual height situation, their results revealed that a greater level of presence elicits stronger physiological responses. Similarly, Riva et al. [115] investigated the relationship between presence and emotions in IVEs. Their findings highlighted

the effectiveness of VR in eliciting specific emotional states, such as anxiety and relaxation, depending on the design and content of the environment. Most importantly, they showed a bidirectional relationship, revealing that the level of presence was significantly higher in the emotional IVEs (anxious and relaxing parks) compared to a neutral one, and that presence served as a mediating variable between the IVE and the emotions it induced.

2.4.3 Realism

Realism is a construct commonly discussed in the context of immersive experiences and can be considered a factor closely related to the sense of presence, particularly PI, and immersion [116, 104, 103]. In its broadest sense, realism in XR experiences refers to the degree to which computer-generated sensory stimuli replicate those perceived in the real world. However, it is a multifaceted concept that can be approached from both objective and subjective perspectives [116]

From an objective standpoint, realism can be understood as the degree to which virtual content reproduces real-world conditions based on fidelity and resolution [103]. This can be quantified through rendering parameters, such as the polygon count in a mesh and point cloud density for 3D reconstruction, as detailed in Section 2.3.2, or the Ambisonic order and auralisation method used for audio rendering, as outlined in Section 2.2.2. This objective perception is further shaped by the number of sensory modalities incorporated, with multiple sensory stimuli contributing to a more convincing and immersive experience, as well as by technological factors such as field of view, refresh rate, and sample rate [104].

The subjective perspective frames realism as a perceptual experience, shaped by plausibility, coherence, and the extent to which the virtual environment aligns with expectations of the real world [116, 41, 111]. Lombard and Ditton [117] link realism to plausibility, identifying subcomponents such as social realism, which refers to the extent to which virtual stimuli could occur in real life, and perceptual realism, which concerns how closely the implementation of stimuli aligns with one's expectations if they were to exist in the real world [41]. While the definition is closely related to plausibility, the distinction lies in the fact that plausibility does not necessarily require a direct real-world representation to be perceived as convincing, whereas perceptual realism, also defined as psychological fidelity, focuses on how faithfully the sensory feedback matches users' expectations and the psychological effects of real-world qualities [116].

This distinction is illustrated by studies showing that plausibility can drive realistic behaviour even in the absence of photorealistic representation. For example, Slater et al. [118] examined bystander responses to a violent incident in a virtual bar and found that participants made physical and verbal interventions, demonstrating that a plausible virtual scene can elicit realistic social behaviour. Thus, an environment can be perceived as plausible without portraying the level of realism found in a photorealistic GS environment. This situates realism as closely connected to coherence and the perceived alignment of the virtual experience with its real-world

counterpart [111]. Skarbez et al. [119] investigated factors influencing the Psi, focusing on system coherence aspects such as the behaviour of the user's embodied avatar and virtual characters, the physical coherence of objects, and the contextual coherence of the environment's visual appearance. Their findings highlight that plausibility depends on credibility and consistency, primarily driven by the coherence of the virtual scenario, particularly the correspondence between users' movements and the behaviour of the embodied virtual body. Moreover, the contribution of auditory feedback to the illusion of plausibility in an IVE has also been investigated. Bergström et al. [120] examined how the quality of auditory rendering and other sensory features, including gaze, sound spatialisation, auralisation, and environmental sounds, contribute to the Psi. Using a virtual string quartet performance with varying auditory conditions, their study found that gaze and environmental sounds were the most important factors for Psi. Among the auditory factors, those contributing to contextual realism, such as environmental sounds and auralisation, were more important for Psi than spatialisation.

2.4.4 Embodiment

The experience of possessing, commanding, and inhabiting a virtual body is referred to as the Sense of Embodiment (SoE). This illusion results when users feel that the avatar's physical characteristics belong to their biological bodies [121]. The SoE is based upon three subcomponents: the sense of self-location (self occupies a specific volume of space), the sense of agency (having global motor control over the body), and the sense of body ownership (feeling of self-attribution towards a body) [121, 122].

These subcomponents derive, in turn, from a series of interrelated factors that influence the SoE [123]. Avatar appearance has been shown to play a central role, including aspects such as shape, characteristics, and realism [124, 125], with avatar personalisation further strengthening the sense of body ownership [123, 121]. The Point of View (PoV), or visuospatial perspective, from which users interact with and perceive the virtual environment, is another important factor [125, 126]. First-person perspective is generally preferred by users, as it induces a sense of self-location and can positively influence changes in one's perceived body schema [123, 127]. However, it does not necessarily yield higher levels of body ownership or agency compared to a third-person perspective [127]. A further factor is the level of control over the virtual avatar, which refers to the synchronisation between a user's movements and the avatar's behaviour, as delays in visuo-motor synchrony can negatively affect the sense of agency [128, 126]. While objective or system-level factors have been widely examined, subjective factors such as users' emotional state, immersive tendencies, and preferences have also been shown to influence the SoE [129]. Interestingly, body ownership has been demonstrated to exist even when the body features of the avatar, such as gender [126], race [130], and age [131] differ from those of our actual body.

Dysmorphic Avatar Embodiment Thanks to the wide range of possibilities offered by XR for creating IVEs, we might desire to embody dysmorphic avatars with distorted body shape and body schema, as well as non-humanoid forms [132, 133]. For instance, we could embody an avatar with limbs of unnatural dimensions [134], add extra limbs such as an additional arm [135] or even give the user's virtual avatar a tail-like bodily extension [136]. Embodying dysmorphic avatars in IVEs can serve diverse purposes across different contexts. For example, embodying a fantastical creature to enhance acting performance [137] or perceiving the environment through the eyes of an animal species for educational purposes [4, 138]. However, embodying dysmorphic avatars that differ from our own body presents challenges that require careful design of sensory and control mechanisms [135]. This involves providing rich and convincing sensory feedback to support body distortion [136, 135], accounting for potential proprioceptive drift (the perceived displacement of real body parts) [134], and designing meaningful and intuitive control metaphors to alter perceived body representation and successfully induce the illusion of ownership [139].

In line with previous SoE research, providing a sense of body ownership over such avatars to alter our perception of the body schema requires providing users with synchronous multisensory feedback to establish identification and ownership, including visuo-motor and visuo-tactile cues [134, 136]. For example, Kilteni et al. [134] conducted a study to examine the extent to which an asymmetrical virtual body with elongated arms could be experienced as one's own body. They compared regular arm length to different elongated arm lengths while providing visuo-motor feedback in combination with congruent and incongruent visuo-tactile feedback. Their findings provided evidence that body representation can be altered through spatio-temporally congruent multisensory and sensorimotor stimulation, and that ownership can be induced over body transformations even when the virtual limb is up to three times the length of the real arm. Moreover, their study revealed that visuo-motor feedback and a first-person perspective are the strongest factors in inducing ownership illusion and can override visuo-tactile inconsistencies.

Ownership over extra limbs requires developing appropriate mapping and control methods for avatar features that differ from our own body [136, 135], enabling users to perform movements and gestures that control such extra limbs as if they were their own. Steptoe et al. [136] experimented with an "extended" humanoid avatar that featured a tail-like appendage extending from its coccyx, which participants controlled in a synchronous condition by mapping lateral hip movements to the tail's posture. Their findings suggest that through synchronous visuo-motor feedback, participants could rapidly learn the tail-control method and perceive an enhanced sense of ownership and agency over the extended humanoid avatar despite its extra-human form. Additionally, the results showed that participants experiencing body ownership were more likely to experience emotional responses, such as anxiety, toward virtual threats to the tail and body. Similarly, Won et al. [135] experimented with the addition of an extra arm to examine whether users could control bodies different from their own by changing the relationship between tracked



Figure 2.4: Illustration of three virtual animals and their first-person control, as presented by Krekhov et al. [3].

and rendered motion. In their study, participants controlled an enhanced avatar with a longer third arm, which they manipulated through wrist-rotation mapping of their physical arms. Their results confirmed that through synchronous visuo-motor feedback, participants could rapidly learn to control such novel body configurations even when the body schema differs from the user's own body. Overall, across altered body schemas, eliciting a strong SoE relies on rich and synchronous sensory feedback, a first-person perspective, and intuitive control metaphors.

Animal Avatar Embodiment While embodying elongated or extra limbs represents one type of altered or different body schema, another involves the embodiment of animals in IVEs. Animal embodiment comes with additional challenges that must be considered to elicit an effective SoE, particularly in relation to the morphological diversity of animals, which differ significantly in skeleton, posture, and shape from our own morphology. Krehov et al. [3] investigated embodiment towards animals that varied in one of these three aspects with respect to a humanoid avatar: a tiger (posture), a spider (skeleton) and a bat (shape) (see Figure 2.4). Their findings suggest that, despite morphological differences, a SoE toward virtual animals is feasible. They also found that positions similar to those of humans are preferable, as animal poses can be physically demanding for participants. Moreover, the first-person perspective generally outperforms other points of view, suggesting that, as with humanoid embodiment, the perspective and control of animal avatars are crucial for evoking a SoE.

Providing rich sensory feedback was also shown to improve embodiment towards virtual animals [138, 4]. Ahn et al. [138] investigated visual, haptic and sound feedback while embodied in a shorthorn cow and a coral on a rocky reef. Their results suggest that increased feelings of embodiment, presence and a sense of connection between the self and nature were experienced as a result of including rich multi-sensory feedback. Likewise, Pimentel et al. [140] investigated visuo-motor and visuo-tactile stimulation immersing participants in the body of a loggerhead

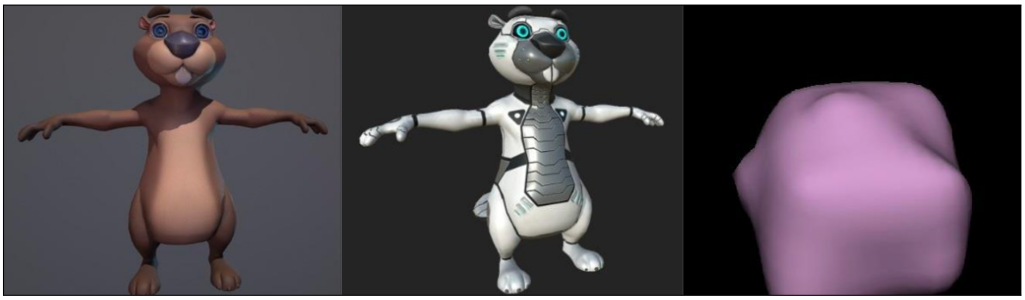


Figure 2.5: Illustration of the natural beaver, robot beaver, and amorphous body character used in the study by Sierra-Rativa et al. [4].

sea turtle. Their findings not only show that body transfer via non-humanoid avatars is feasible but also suggest that animal embodiment could be used for biodiversity conservation, as users perceived their virtual bodies as a conduit for the victims' lived experiences.

Building on this work, animal embodiment in IVEs has been examined for its impact on both empathy and self-related psychological processes. Researchers have used animal avatars to investigate whether embodying non-human perspectives can foster empathy and raise awareness of animal abuse, wildlife, and environmental issues [4, 140, 138]. For example, Sierra Rativa et al.[4] developed an IVE to examine whether empathy, immersion, and perceived pain changed when embodying different animal avatars. Participants experienced a distressing event with three avatars: a realistic beaver, a robotic beaver, and an amorphous body character (see Figure 2.5). The results revealed that avatar appearance substantially influenced immersion and pain perception, although not situational empathy or embodiment.

In addition, Oyanagi et al.[141, 142] investigated the Proteus effect, the idea that a user's behaviour, attitude, and mental state are shaped by their avatar's appearance [143], by embodying participants in a bird avatar. Their findings showed that a bird avatar significantly enhanced the psychological experience of flying and reduced fear of heights and falling. In addition, Li et al. [144] examined the sense of self-illusion (perceiving ourselves as the role played in the IVE) by embodying participants in a cat avatar. They discovered that a user's high level of self-illusion generally contributes to a greater connection to the virtual role and results in more coherent responses within the IVE.

Overall, research on the SoE with animal avatars has encompassed a wide range of species (tiger, bat, spider, scorpion, rhino, bird, beaver, loggerhead turtle, cat, shorthorn cow, and coral reef) and goals, from assessing the contribution of visuo-motor and visuo-tactile feedback to exploring outcomes such as empathy, species conservation, pain perception, or illusions like the Proteus effect. A summary of prior work on animal embodiment, including species, feedback, and measures, is presented in Table 2.1. As with altered body schemas, animal embodiment also relies on synchronous multisensory cues, a first-person perspective, and intuitive control to

Table 2.1: Previous studies on Animal Embodiment in VR.

Experiment	Animal	Feedback	Measures
Ahn et al. [138]	Shorthorn cow, coral	Visual, haptic, audio	Body transfer, spatial presence, inclusion of nature in self
Krehov et al. [145]	Scorpion, rhino, bird	Visuo-motor	Immersive tendencies, intrinsic motivation inventory, player experience of need satisfaction, presence, simulator sickness
Krehov et al. [3]	Bat, tiger, spider	Visuo-motor	Acceptance, control, change, fascination, ease of control, fatigue
Li et al. [144]	Cat	Visuo-motor, haptic	Body ownership, self-illusion
Oyanagi et al. [141]	Bird	Visuo-motor, audio	Body ownership, change, agency, feature
Oyanagi et al. [142]	Bird	Visuo-motor	Sense of flight, agency, ownership, immersion, reality, self-location, transformation, anxiety about falling, fear of heights
Pimentel et al. [140]	Loggerhead turtle	Visuo-motor, visuo-tactile	Environmental concern, environmentalism, spatial presence, copresence, donation intention, nature in self, threat perception, conservation behavioural intentions
Sierra Rativa et al. [4]	Beaver	Visual, haptic	Dispositional empathy, self-reported situational empathy, immersion, VR sickness symptoms, self-reported emotions, animal conservation tendency, perceived pain, embodiment

elicit a SoE. However, these studies have underscored the challenges of providing convincing visuo-motor and visuo-tactile feedback for embodying dysmorphic avatars due to discrepancies between the physical and virtual bodies. Thus, Chapter 3 examines continuous audio feedback and audio-tactile feedback as sensory dimensions through which to elicit a SoE over an animal avatar in IVEs, contributing to **RQ1**.

2.5 Multimodality

In the real world, we perceive and decode the environment through all our senses, and while visual information is the predominant stimulus for most humans, all the perceived sensory information allows us to expand and construct a unified experience of our surroundings, especially when one or more modalities are unreliable and non-existent [15, 20]. Therefore, it is important to consider the effects and consequences of multimodal and crossmodal integration to develop a comprehensive understanding of sensory feedback in an IVE [15, 146].

XR experiences are inherently multimodal, since they incorporate different sensory feedback in some way. Commonly, they entail complex scenarios and interactions resulting from a variety of visual, audio and haptic associations [84]. However, when it comes to research, a large number of perceptual studies are unimodal and investigate the influence of sensory stimuli in isolation. While understanding the influence of individual feedback is critical, it does not reflect the real perceptual experience of a multimodal ecosystem, creating a gap in understanding of the precise

function and impact of multimodal perception in XR applications [15]. As perceived immersion, presence, realism, and embodiment depend on the information perceived by our sensory system (vision, sound, touch, force, taste, and smell), the intended goal of XR is to effectively replace or augment real sensory perceptions with virtual ones [18].

XR ecosystems involving multimodal feedback have been shown to provide more comprehensive and immersive experiences, reducing the cognitive load that can arise from the overload of a single sensory input [147]. This enhances the realism and immersion of IVEs [84] and elicits stronger psychological illusion effects, such as a heightened sense of presence[5], SoE[121] and improved spatial perception [116]. Moreover, multimodal integration in XR opens up possibilities for new interactive mechanisms that shape how users engage with virtual environments, enhancing task performance and training [147], while also advancing communication, interaction, and simulation across diverse fields, including industry, healthcare, entertainment, design, and many others [84, 15].

Multimodality serves as a fundamental framework for this thesis. It provides a basis to explore how auditory feedback, whether self-produced or related to one's own body, when integrated with congruent visual and haptic feedback, may enhance psychological effects such as presence, realism, and embodiment, and how it might influence the overall perceived audio-visual quality within IVEs.

Multimodal Integration For multimodal effects to occur, multisensory stimuli must be properly integrated so that they are perceived as part of the same stimulus, event, or object [15], as mismatches can undermine the benefits of multimodal integration, such as increased presence, ecological validity, and overall user experience [32, 31]. Therefore, two critical factors must be considered: synchronisation and stimuli congruency. The former refers to the temporal synchrony between modalities, which needs to fall within a window of integration in order to be fused together [66]; for example, triggering the sound of footsteps of a virtual character at the moment their feet hit the floor. The latter implies that multimodal stimuli must be congruent with each other and coherent with the environment [148], ensuring that the feedback is credible and aligns with user expectations, while avoiding mismatch effects such as room divergence, which arises from inconsistencies between perceived auditory and visual feedback [149].

The impact of multimodal feedback is complex and extends beyond simple additive effects [116, 150]. While the integration of more modalities generally shows positive outcomes, improvements are not necessarily additive and vary significantly based on several factors, including the specific modalities involved, the fidelity of each modality, and the context of the interaction [147, 150]. Thus, it is vital to evaluate the modalities and tasks involved, as well as the associated cognitive processes required when designing a multimodal experience, so that each modality has a clear purpose without producing an undesired cognitive load that can impact users negatively [151].

Although it is evident that multimodality offers considerable advantages over unimodality, it is important to highlight the complexities involved when designing multimodal experiments and applications to ensure coherent matching between modalities [15, 152]. The first aspect to consider is hardware limitations. The design of multimodal applications increases the processing resources necessary to run the XR environment. Because XR systems, particularly standalone headsets, have limitations in terms of GPU and CPU power, running a computationally intensive application might result in refresh frame rate fluctuations and frame drops, as well as undesired audio artefacts. Furthermore, because multimodal integration relies on the temporal synchrony of the modalities involved within a limited integration window [152], it is critical to maintain low system latency to ensure that sensory feedback remains within the audio-visual synchrony window of approximately 25–50 ms [153], allowing sensory inputs to be fused into a single event within a virtual environment [154]. Secondly, each sensory input requires a level of involvement and concentration from the user, and with each added modality, the complexity and dimensionality increase [15]. In addition, studies have shown that the attention needed for task performance affects the perception of sensory inputs. For these reasons, the design of multimodal interactions presents a significant challenge and requires careful consideration to ensure that each modality serves a clear function and purpose [15, 84, 152].

Visual feedback is the primary modality in XR systems, and in most XR experiences, particularly in VR and MR, other modalities are integrated to support the visual stimuli [15]. However, audio feedback also plays a central role in multimodal interactions, contributing to a holistic and plausible virtual environment, and enhancing immersion, realism, and performance [84, 116]. While Audio-Visual-Haptic configurations are frequently explored for enhanced interaction and training, Audio-Visual integration is by far the most common and widely studied [147, 152].

2.6 Audio-Visual Feedback in XR

Given the intrinsic potential of XR systems to offer auditory and visual feedback, which is not necessarily true for haptics (e.g., VR on mobile devices), research on audio-visual multimodality is becoming more common. Many studies have investigated various aspects of XR, including immersion, presence, plausibility, and their impact on physical, mental, and emotional responses. Furthermore, studies have sought to understand how audio-visual multimodality can benefit different immersive experiences, including medical, well-being, virtual learning, and entertainment applications. Additionally, research has aimed to explore how multimodal pairing may account for various hardware and software limitations in XR technology, such as computational power, audio and visual resolution [15].



Figure 2.6: Nature walk IVE with a gravel path employed in the experiments by Kern et al. [5].

2.6.1 User-Centred Considerations

As outlined in Section 2.4.2, presence is essential for effective XR experiences and has received significant attention. Several studies have shown that the addition of auditory stimuli such as ambient and sounds related to one's own body increases the perceived sense of presence and immersion [67]. Larsson et al. [32] conducted a study to investigate the impact of audio-visual spatial consistency on the sense of presence in IVEs across four conditions with varying auditory feedback: no sound (visual only), non-auralised, mismatched auralisation, and matched auralisation. While the results among the sound conditions did not yield significant differences, the no-sound condition ranked significantly lower in presence ratings. Moreover, post-ranking of the conditions revealed a tendency towards a preference for the correct auralisation condition. Similarly, Kern et al. [5] conducted a study to evaluate the effect of auditory stimuli on presence and immersion employing two types of sounds: a nature soundscape comprised of localisable and non-localisable sounds and movement-triggered step sounds (see Figure 2.6). According to their findings and in line with previous research[155, 32, 67, 156, 157], auditory cues, including soundscape and step sounds, significantly affected participants' felt presence and realism of the visual environment, highlighting the importance of multimodal feedback in creating a sense of presence and immersion in VR.

Audio-visual feedback has an impact on users' emotional and physical responses, and research has shown that it has a greater impact than unimodal feedback, whether visual or aural. An experiment conducted by Naef et al. [24] demonstrated that audio-visual stimulation in IVEs

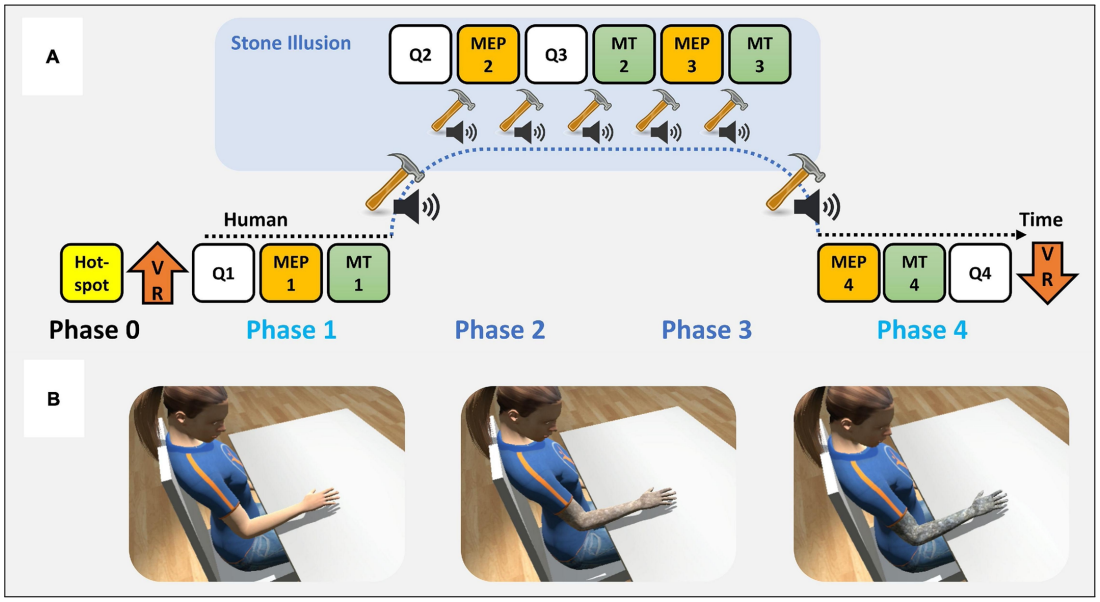


Figure 2.7: Modifying the self-perception over the embodied avatar in VR by Buetler et al. [6].

is more effective at inducing a relaxation response compared to no artificial sensory input or unimodal stimulation (visual or audio only). Participants showed a greater effect on lowering heart rate, respiration rate, and blood pressure levels, as well as improving participants' state of mind, which was reflected in their physiological responses. The study also reported a significant interaction between the auditory and visual cortices during audio-visual stimulation. Furthermore, Mahmud et al. [158] evaluated the effect of different types of audio feedback: spatialised, monoaural and rhythmic over gait disturbances in VR (e.g., balance, velocity, step length, cadence). Their findings revealed that audio-visual feedback enhanced gait performance while immersed in VR, with spatial audio being the most significant.

The brain continually integrates sensory input from many sources [6, 159, 160]. The perceived sensory information not only affects us physically but also modifies the mental representation of ourselves. According to auditory feedback research, altering the sounds emitted by our bodies in synchrony with our movements can influence our impression of our body size and weight [161, 159], as well as its perceived characteristics [160]. By modifying the physical properties of an embodied avatar, Buetler et al. [6] investigated the impact of multisensory input on motor brain networks and control in VR. Participants in this experiment observed the visual appearance of their avatar arm gradually change from human skin to stone while receiving concurrent audio-tactile input (see Figure 2.7). They found that participants experienced the "stone arm illusion" without affecting their sense of body ownership, experiencing their arm as colder, heavier, and stiffer, as well as physically compensating for the embodied bodily attributes of the stone avatar.

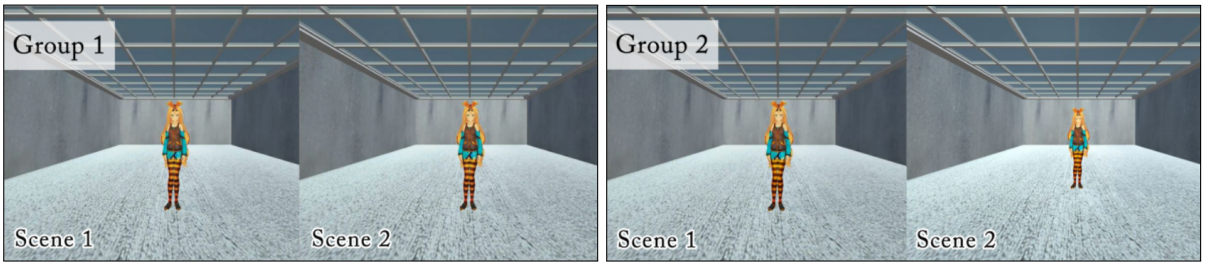


Figure 2.8: Reverberation to compensate distance compression virtual environment employed by Huang et al. [7].

The notion of modulating the mental representation and perceived characteristics of our body through sensory feedback, including auditory cues, relates directly to the concept of SoE, which has been shown to occur even when virtual avatars differ from the features of our physical body. Since perceived sound and the visual appearance of the virtual avatar can make us feel as if our body is made of stone, this principle could be extended to other physical attributes, such as having fur, to promote a SoE over an animal avatar, a topic further explored in Chapter 3.

2.6.2 Technical Considerations

Research has shown that audio-visual pairing may affect users' bodily responses, mental representations, and emotions. Other studies have examined the impact of multimodal feedback in accounting for the technical limitations of XR systems. Distance compression remains an issue in VR experiences, as users perceive distances to be closer in virtual environments [162, 163]. Despite improvements in HMD technology, egocentric distances are typically perceived at approximately 74% of the actual distances [163]. As sound reverberation is an essential cue for auditory distance perception, Huang et al. [7] reported that reverberation time has the potential to compensate for distance underestimation in users' perceived depth in VR, especially in near-field configurations (1-5 meters). However, an excessive increase in reverberation time can cause sensory segregation, in which the perceptual mismatch between modalities prevents users from perceiving the sensory stimuli as one whole unit (see Figure 2.8). Similarly, Lin et al. [25] examined the perceptual learning and carryover effects of reverberation time calibration related to the depth of a target in IVEs. Their results revealed that the reverberation time of an audio signal may be utilised to calibrate users' perception of depth, with the effect carrying over for an extended period of time.

As mentioned in Section 2.2.1, accurate sound localisation and externalisation using headphones depends largely on HRTF responses, and utilising an HRTF different from the listener's has a negative impact on the accuracy and realism of virtual sound perception [52, 20]. Multisensory research conducted outside XR has revealed how visual feedback may impact audio perception and can cause rapid changes in acoustic perception, such as the ventriloquist effect. Berger



Figure 2.9: Visual and Spatial Audio Rendering on VR Immersive Virtual Environment by Potter et al. [8].

et al. [26] investigated whether auditory source localisation could be improved for users of generic HRTFs through cross-modal learning by leveraging plasticity mechanisms of the brain. They found that auditory source localisation for users of generic HRTFs could be improved via cross-modal learning by exposing participants to synchronous audio-visual stimuli, implying that personalised HRTFs may not be required if users can calibrate their auditory perception.

Additionally, beyond sound spatialisation, studies have explored how immersive technologies can support speech communication. Luberadzka et al. [164] investigated how audio technology can promote social harmony within XR by addressing challenges in speech communication, including intelligibility and audio coherence in virtual interactions. A pilot study tested spatial audio under varying acoustic conditions with optional speech enhancement. Their results highlight the value of speech enhancement in XR and show its potential to evaluate audio processing algorithms to improve audio quality in multi-user XR environments.

Audio-visual fidelity relies on the perceptual integration of both visual and auditory modalities. Research has demonstrated that altering one modality can have detrimental or beneficial effects over the other [15]. Potter et al. [8] conducted a study to determine the impact of spatial audio and video resolution on the perceived audio-visual quality and immersion in an IVE. The experiment considered music and speech as the auditory feedback (in third-person perspective) and three levels of audio rendering and video resolution. Their results indicate that both video and spatial audio have considerable influence over the sense of immersion (see Figure 2.9). Interestingly, the auditory condition that considered the addition of room acoustic rendering to head-tracking



Figure 2.10: Bedroom virtual environment illustrating variations in materials and furniture from Burnett et al. [9].

binaural audio (full auralisation) with 20% video resolution had the same effect on immersion as increasing the video resolution five-fold with no room acoustics. Moreover, to produce an audio-visual quality decrease by altering only the auditory feedback compared to 20% video resolution, both sound auralisation and head-tracking had to be removed.

As mentioned previously, coherence is an important factor in multimodal integration, and the mismatch of audio-visual feedback can reduce perceptual effects or lead to negative outcomes, such as a diminished sense of realism and presence. Research has also investigated the perception of room acoustics as an integral component of audio-visual feedback in IVEs [165, 166, 9, 167, 10]. Matching the acoustic features to their visual counterparts in an IVE is crucial for creating a plausible environment and eliciting a higher sense of presence, highlighting the importance of accurate sound auralisation and dynamic sound spatialisation in immersive experiences [166].

Burnett et al. [9] examined how visual attributes, including room size, surface materials, and furnishings, influence users' expectations of reverberation time (T60) and acoustic room size. Using speech and percussive samples auralised using a Scattering Delay Network (SDN) reverberator, they tested nine virtual rooms with varied room size, materials, and furnishing (see Figure 2.10), where participants adjusted the reverberation time and acoustic room size to match their expectation of how a visually presented room (with varying size, materials, and furnishing) should sound. The study found that visual room size significantly influenced both expected T60 and acoustic room size. Visual surface materials also had a significant effect on T60, while furnishing showed a marginally significant effect on expected T60.

Similarly, Rolkowski et al. [167] investigated how the interaction between visual and auditory feedback affects the perception of IVEs. Using third-order Ambisonics SIRs of five different rooms and a 3D model of a virtual theatre, participants identified the sound scene that they felt was more in line with their subjective expectations of the room's characteristics for each pair. Their findings confirmed the role of spatial sound in creating immersive VR experiences and the importance of audio-visual matching in achieving full immersion. The results revealed varying levels of distinguishability between different impulse responses, with participants showing a preference for sound scenes with a reference reverberation time of about 2.5 seconds.



Figure 2.11: Illustration of the three visual representations used in the study by Fargeot et al. [10].

Fargeot et al. [10] explored audio-visual coherence in virtual environments by assessing the perceived visual-auditory (VA) coherence between acoustic environments and their visual representations, and whether the type of visual rendering affected this judgement. The study considered five acoustic environments with different reverberation times and volumes. Two audio stimuli, including a short speech excerpt and a classical guitar excerpt recorded in anechoic conditions, were auralised by convolution with the higher-order Ambisonics SIRs of the five environments, resulting in ten different auditory conditions. These were presented across three types of visual representations: 360° panoramic photos rendered by spherical projection, textured point clouds obtained through photogrammetric capture of the environments, and simple polygonal 3D models with uniform textures (see Figure 2.11). Their results indicate that participants were unable to associate the acoustic characteristics of studied environments with their visual representation, and neither the type of visual rendering nor the sound stimulus significantly influenced these judgments.

This study highlights the challenge of achieving perceived audio-visual coherence in virtual environments using static auralisation and conventional visual rendering, raising the possibility that high-fidelity visual representations as well as dynamic acoustic exploration may be necessary to strengthen audio-visual coherence in XR. These challenges are explored in Chapter 5 and 6, where we examine audio-visual integration (coherence) across different visual representations and levels of fidelity using real-time vocal auralisation and dynamic audio-visual exploration.

High-fidelity IVEs have been designed, combining real-time auralisations (through convolution of impulse responses from geometrical acoustics and acoustic measurements) and 3D-modelled visualisations (mostly polygon-based models) [168, 169, 170]. These representations have demonstrated effectiveness in producing accurate virtual models of complex architectural venues, providing perceptual accuracy in acoustic representation and closely matching the acoustic and spatial attributes of their real counterparts [171, 11, 169]. This enables the development of perceptual and archeoacoustic studies, architectural and acoustic design, immersive and interactive audio-visual experiences, and the virtual recreation of musical per-



Figure 2.12: Comparison of the real space and its 3D model representation from Bargum et al. [11].

formances [171, 168, 169, 11, 170]. Moreover, these studies have underscored the significant influence of visual rendering on auditory perception features such as loudness, distance, source width, and room size [11, 171]. Thus, these findings highlight the importance of visual realism in shaping user perception, a dimension that remains central to the design of effective IVEs.

Bargum et al. [11] conducted a study to validate an accurate audio-visual recreation of the Ambisonic concert hall at the Royal Danish Academy of Music (RDAM) (see Figure 2.12). The environment was created from a 3D polygonal model, and the audio was encoded into 5th-order Ambisonics and rendered binaurally over headphones. Subjective listening results showed the Ambisonic room to accurately match the acoustic and spatial attributes of the real space, as well as providing high accuracy in head tracking, acoustic reproduction, and localisation. Similarly, Katz et al. [169] created an audio-visual simulation of the current and former Théâtre de l'Athénée configurations, enabling historians and theatre researchers to realistically experience performances in past spaces. Impulse responses were generated using geometrical acoustics and integrated with a visual 3D model in VR. The project allowed for the comparison of acoustics across three historical configurations and successfully established a VR framework for audio-visual perceptual studies in historic theatres. Postma et al. [165] further examined the influence of visuals on the acoustic judgments of room acoustic auralisations, comparing audio-only tests with audio-visual tests using dynamic VR content. Using a virtual reality framework based on Katz et al. [169], the auralisations employed calibrated 2nd-order Ambisonics impulse responses. Their results showed that visuals influenced the acoustic judgments of several room acoustic attributes. In particular, distance perception was dependent on the visuals, especially for seating positions close to the stage.

Sensory feedback from any modality influences how users perceive the environment, and when properly integrated with one or more additional modalities, it can significantly amplify the impact of a sensory event. Studies have demonstrated that audio-visual multimodality enhances several aspects of immersive experiences, including perceptual illusions of immersion and presence [5], physical and psychological effects [158, 24], modifying the physical properties of an embodied

avatar [6], impact of spatial audio and video resolution [8], and the mitigation of technical limitations [26, 7]. However, prior research has primarily focused on the perception of external audio sources, leaving the role of audio-visual multimodality from a first-person perspective, such as sounds produced by one's own body and real-time first-person vocal auralisation within IVEs, largely unexplored. Some multisensory studies have investigated applications of this integration, including vocal adaptation in virtual room acoustics [35], a VR system for ensemble singing [34], and public speaking training in immersive environments [172], which will be discussed in the following section.

2.7 Self-Produced Sounds & Embodied Audio Feedback

Audio feedback plays a significant role in how we perceive ourselves, our environment, and how we relate to it. Because we engage with the environment through our bodies, self-produced sounds originating from the body itself (e.g., voice and breathing) or sounds perceived in relation to one's own body through interactions with objects and the surroundings (e.g., brushing one's teeth), can modulate the mental representation of the self [16] and affect action performance [160]. Tajadura et al. [16] explored the concept of the sonic self-avatar, a sound-based representation of one's own body [16]. While this term provides a definition analogous to a visual avatar, these sounds are intrinsically connected to our body and are perceived through bodily interactions. Therefore, they can be defined as *Embodied Auditory Feedback*, sound perceived from an embodied perspective and generated through the interaction of objects with the body. This notion is based upon the principles of body-centred interaction [173] and embodied interaction [174], both of which emphasise the central role of the body in perception and action within Human-Computer Interaction (HCI). For example, changing the frequency of perceived footstep sounds has been shown to make people feel lighter when listening to higher-frequency footsteps, while also enhancing their emotional state, or heavier when listening to lower-frequency footsteps [161, 175].

Multisensory feedback, such as audio-tactile feedback, has also been found to change our body perception. Senna et al. [160] examined whether the perception of the body's material properties could be altered through audio-tactile stimuli by hitting participants' hands and gradually substituting the sound of a hammer against the skin for that of a hammer against a stone. Their findings showed that when the audio-tactile feedback was synchronous, and the impact of the sounds provided clear and unambiguous cues about the object's material, participants' perception of the material properties of their hands was altered. As outlined in Section 2.6, Buetler et al. [6] extended this research by investigating the effects of multi-sensory feedback on motor brain networks and control in VR by altering an embodied avatar's physical characteristics.

In addition, some studies have examined the influence of self-produced audio feedback on the SoE in XR. Sikström et al.[38] conducted an experiment exploring the impact of audio stimuli on the sense of ownership of virtual wings. They examined how self-produced sounds affected

participants' perception of control and ownership over virtual wings in an IVE using sounds of wind, footsteps, and wings flapping. However, participants perceived little to no difference between conditions: no sound, body-only sounds, body and wings sound, and body sounds with asynchronous wings sound. Similarly, Lugrin et al.[39] investigated how audio feedback influenced the SoE in non-immersive MR environments. Their method involved a mixed-reality game with two factors: audio versus no audio feedback, and human versus robot avatars. Results indicated that audio feedback did not affect the SoE, suggesting virtual body sound cues are not essential. However, the small sample size limits the reliability of these findings.

2.7.1 Vocal Perception & Production

Our voice is an essential element of the perception of our environment, the mental representation of ourselves, our emotional state, and it also affects how we perform actions [68, 160, 17]. Since we are sensitive to changes in how it sounds, how we hear our voice has a significant impact on how we speak. Therefore, we modify our voice according to the environment and circumstances in which we find ourselves. For example, we adjust our voice level in response to the noise level in the environment [83]. Furthermore, our voice, as a self-produced sound, contributes to our perception of real and virtual environments [14], influences identity representation [176], and provides a sense of agency, or the impression that we are the ones producing an action [177]. XR ecosystems are characterised by enabling users to engage with the virtual environment and the elements within it. Research has shown that the sense of agency has a significant role in inducing a sense of presence and is an essential component of the SoE [178, 121].

When we speak, we hear the sound of our voice both directly, via the sound travelling through the air into our ear canal and through the bones and skull into our cochlea, and indirectly, by listening to the sound reflections (early reflections and reverberation) that result from our voice interacting with the acoustic response of the environment [14]. Consequently, room acoustics have a direct influence over the adjustment of vocal parameters concerning the acoustic characteristics perceived by our auditory system [33]. Moreover, because our voice propagates in the environment in which we find ourselves, the acoustic response of the space modifies the sound signal, colouring the voice sound and modulating features such as clarity, brilliance, and timbre, as well as providing us with information such as the size, distribution and materials of the environment [20].

First-person vocal production has been used to study sound auralisation techniques and the impact of room acoustics on vocal perception and production. Sierra-Polanco et al. [28] investigated vocal responses in artificial acoustic environments, altering reverberation time, background noise, and sidetone gain. They found that higher reverberation time and background noise led to greater vocal effort and SPL. In contrast, higher sidetone gain improved comfort and reduced effort, underscoring the perceptual impact of room acoustics. Similarly, Luizard

et al. [35] examined singers' vocal adaptation across eight physical acoustic environments and their virtual representation. Their study revealed that singers adapt vocal features like timbre and sound level to room acoustics using individual strategies, with stronger adaptations observed in physical rooms.

Within the scope of VR, Daşdöğen et al. [29] investigated the influence of audio, visual, and audio-visual feedback on the perception and production of one's voice within IVEs. They varied room volume and speaker-to-listener distance to measure self-reported vocal loudness, effort, comfort, SPL, and spectral moments. Their findings suggest that visual and audio-visual feedback significantly impact vocal perception and production, with room volume and speaker-to-listener distance having the most pronounced effect. Similarly, Rodríguez et al. [179] evaluated the reliability of a VR-simulated classroom using 360° video for assessing self-perceived voice quality and phonation. Using classroom noise delivered via headphones to simulate a realistic auditory environment, they demonstrated that VR can effectively replicate real-world vocal experiences and serves as a valid tool for assessing voice production.

In sum, studies have examined the perception of one's voice and its impact on vocal production and adaptation across both audio-only and VR contexts. Real-time vocal auralisation and environmental factors have been shown to significantly shape these processes within IVEs, particularly when integrated with visual feedback. However, the audio-visual perception of real-time vocal auralisation in IVEs and its effect on perceived audio-visual quality across different visual representations, and its influence on the sense of presence, immersion, and realism have not yet been fully explored. Therefore, Chapter 5 investigates this aspect, directly addressing **RQ1** and **RQ2**, which focus on how audio-visual feedback can enhance perceptual constructs such as the sense of presence and realism, as well as perceived quality and plausibility.

2.7.2 Vocal Production in Musical Contexts

Beyond vocal production and adaptation in VR, studies have also explored how IVEs can support vocal performance and training. Musical performance is an essential skill that requires sustained practice to develop and maintain. However, its practice is often challenging due to the difficulty of recreating the conditions to elicit the physical and mental states that arise during a performance, such as performing in a real venue and in front of an audience [180, 12]. Simulation training has been shown to support the development of these conditions, promoting performance awareness and reducing symptoms such as performance anxiety [180, 181]. In particular, musician-centred VR experiences can provide accurate audio-visual feedback and rehearsal conditions, enabling musicians to rehearse, improvise, and adapt their sound to the unique acoustic properties of a virtual space [12, 182, 183].

Ppali et al. [12] investigated how VR can support musicians' creative musical practice by immersing them in novel sonic and visual environments. While the study focused on musicians of



Figure 2.13: Comparison between the physical concert hall and the high-fidelity 3D model by Ppali et al. [12].

different instruments and backgrounds, it included several singers as part of their evaluation. The VR system adopted a music-centred or performer perspective design, enabling users to perform and create music. It incorporated the audio-visual representation of a music hall, produced from a 3D polygon model and Ambisonics impulse responses captured at various locations, to offer a concert hall-like immersive experience (see Figure 2.13). The environment offered customisable settings for the lighting and four auralisation options (three plausible and one authentic room impulse response). In addition, they included two audio-visual environments: a peaceful orchard (natural environment) and a magical flat (a space that allows users to change the time of day). Contextual background sounds were also incorporated to increase the plausibility of the environments. Results from a qualitative analysis showed that participants experienced a strong sense of presence, and the environments were considered suitable for musical practice and improvisation. The concert hall was generally perceived as realistic and was described as producing a sense of awe and magnetic pull, which encouraged musical performance. Moreover, it provided a safe space for practice and offered potential for easing performance anxiety. Besides, in the orchard, participants described a sense of relaxation, while in the Magic Flat, they felt comfortable with the space. This study provides valuable insights into the use of VR to support musicians, highlighting the potential of audio-visual IVEs in performers' practice. Their contributions relate directly to **RQ3** of this thesis and are further explored in Chapter 6, which examines the usability and audio-visual plausibility of a VR-based virtual theatre.

While VR has demonstrated value for musicians more generally, its application to singers has also been investigated. Research has explored the potential of VR as a training method for singers [184, 185]. VR-based training has been shown to improve breath control and vocal comfort [186], reduce stage fright and enhance performance quality [187], and support the development of emotional expressiveness [188]. In addition, audio-visual coherence plays a key role in shaping voice perception and production in IVEs [29]. Interactive systems have

been developed that enable singers to explore diverse virtual spaces [12], perform with a virtual ensemble and choir [34, 189], and rehearse in VR environments with realistic acoustic conditions using spatialised choral recordings [185]. These systems highlight the potential of multisensory feedback to support rehearsal, practice, and the development of performance and vocal adaptation strategies, and they demonstrate the value of VR as a training tool for singers.

Kearney et al.[34] developed a VR prototype for choral singing, allowing users to perform with a virtual quartet in a 360° video environment. The study addressed key audio and video reproduction challenges, including Ambisonics SIR modelling. Participant feedback was positive, describing the virtual representation as a convincing and useful tool for rehearsal, practice, and learning. Similarly, Di et al. [189] explored the use of VR to improve singers' remote choir rehearsals through virtual avatar representations in an IVE, examining their impact on presence and enjoyment. Their evaluation showed that singers reported high levels of enjoyment, focus, and a strong sense of co-presence with the conductor and other singers, suggesting that IVEs can enhance presence and non-verbal communication compared to traditional teleconferencing systems. Dedousis et al. [185] investigated the mismatch between choir rehearsal spaces and performance spaces and the vocal adaptation of singers to address these differences. They explored the use of XR music technology to assist singers in familiarising themselves with on-stage acoustic conditions. Their results confirmed the acoustic mismatch between performance and rehearsal spaces, requiring singers to modify their vocal approach, and highlighted that, by rehearsing within a virtual choir using auditory spatialisation, singers can adapt to varying acoustic environments and enhance their integration within a real choir.

Overall, IVEs have shown potential in musical contexts. VR has been explored as a training method that supports singers in improving vocal technique and skills, as well as a meaningful tool for practice and performance in safe spaces that foster creativity and help reduce performance anxiety. In addition, audio-visual systems have enabled choir singing and facilitated remote rehearsals. Together, these studies support the practice and development of performance and vocal adaptation strategies, underscoring the value of VR as a training tool for singers. However, further exploration is required to understand how high-fidelity audio-visual representations of spaces, such as GS paired with real-time vocal auralisation and the dynamic exploration of the acoustic response across the stage, influence vocal performance and singers' experiences, and how these contribute to creating a plausible and usable environment for singing-related endeavours. Chapter 6 addresses this gap by investigating these aspects through the development and evaluation of a VR-based virtual theatre, aiming to inform and **RQ3** of this thesis.

2.8 Summary

In this chapter, we reviewed the relevant literature that supports this thesis. We examined the foundations of audio and visual rendering for XR environments, covering essential aspects of

audio and visual rendering in IVEs. We also discussed perceptual constructs necessary for meaningful immersive experiences, considered the factors that enable perceptual illusions and elicit physical and psychological responses, and reviewed research on multimodality, with a particular focus on audio-visual integration in XR, addressing both user-centred and technical perspectives. Finally, we examined self-produced audio feedback, with particular emphasis on vocal feedback, covering aspects of its perception and production, and explored its potential to support vocal performance and practice. Taken together, we highlight the relevance of multimodality in shaping user perception and experience in IVEs, and how auditory feedback, as part of a multimodal framework, can provide essential sensory input in XR experiences.

In addition, this background reveals gaps in the literature, including the limitations of providing convincing visuo-motor feedback for embodying dysmorphic avatars, particularly animals; the lack of vocal-related studies that explore the influence of dynamic auralisation in audio-visual IVEs and its effect on immersion, presence, and realism; and the limited understanding of how audio-visual fidelity and dynamic acoustic exploration using one's voice contribute to creating a plausible and usable environment for singing-related endeavours. Based on the gaps identified in the reviewed literature, this thesis formulates research questions that focus on how sensory modalities contribute to the perception of acoustic environments in XR, how they may enhance psychological effects such as presence, realism, and embodiment, and how they influence the overall perceived audio-visual quality within IVEs, which together shape the overarching theme of this thesis.

Once the background and related work of this thesis have been established, the next chapter presents the first study, which investigates how continuous audio feedback and its integration with other sensory modalities contribute to the SoE towards an animal avatar.

Chapter 3

Effect of Continuous Embodied Auditory Feedback on Animal Embodiment in IVEs

Based on **"Now I Wanna Be a Dog: Exploring the impact of audio and tactile feedback on animal embodiment"**[1]
Article published at IEEE ISMAR 2023



Figure 3.1: VR scene from the experiment in which the participant is embodied in the dog avatar while a virtual character trims its fur.

[1]*The research described in this chapter was co-authored with Dr. Rebecca Fribourg (ENSA Nantes, École Centrale Nantes, CNRS, AAU-CRENAU, UMR 1563, Nantes Université, Nantes, France), who contributed to the study design, data analysis and manuscript write-up.*

As outlined in Chapters 1 and 2, the effectiveness of immersive content and experiences relies on eliciting perceptual illusions that foster a suspension of disbelief. When users perceive the sensory input as real, the experience becomes convincing. Therefore, the more senses engaged, the greater the possibilities of eliciting these perceptual illusions.

This work presents an initial exploration of embodied auditory feedback as part of multimodal stimuli to elicit psychological effects in IVEs. Specifically, we investigate its influence when presented in combination with visuo-motor or visuo-tactile feedback, or in isolation, on the SoE towards an animal avatar. By focusing on continuous embodied auditory feedback and its integration with other sensory modalities, this study contributes to the thesis's broader investigation into how it shapes perceptual illusions and psychological responses in multimodal IVEs.

With the growing adoption of Virtual Production (VP) and XR experiences across industries, particularly in the film and entertainment sector, on-set pre-visualization of computer-generated content have become standard practice. These methods, apart from allowing actors to place themselves virtually within a location, enable them to embody creatures or animals, which can improve their performance of such creatures [137]. To provide convincing experiences, however, actors need to achieve a SoE, the feeling of being inside, having, and controlling their virtual body in relation to the virtual representation [121]. SoE is typically achieved by delivering congruent sensory information, such as visuo-motor feedback (users moving and seeing their virtual body move congruently) and visuo-tactile feedback (an external effector providing tactile stimuli on the user's real body in synchrony with visual stimuli on the virtual body).

While previous studies have successfully examined embodiment in anthropomorphic avatars, research addressing dysmorphic avatars (i.e., virtual body types that differ from the user's real body in terms of morphology and structure, such as creatures or animals [138, 140, 3]) remains limited. A major challenge in providing convincing visuo-motor feedback for embodying dysmorphic avatars arises from the discrepancies between the physical and virtual bodies. As compared to anthropomorphic avatars, the virtual character can differ structurally and morphologically from the user's body. As for visuo-tactile feedback, it has the limitation of requiring external stimulation (e.g., delivered by either an experimenter or a robotic effector [190]). Therefore, exploring alternative sensory stimuli, such as sound, as an additional dimension of multisensory feedback is essential to provide a meaningful SoE towards dysmorphic avatars in IVEs.

Previous work has investigated the capacity of audio feedback to alter the mental representation of one's physical body (e.g., its size, weight, and material properties) and body-related emotional state [161, 16, 160]. When combined with synchronous tactile feedback in VR, it has been shown to alter the perceived material of an embodied avatar without breaking the SoE [6]. While studies on SoE with dysmorphic avatars report negligible results [38, 39], further research is needed, particularly to explore different types and natures of auditory feedback. These gaps highlight the

need for experimental evaluation to assess the role of audio feedback, both in isolation and in combination with tactile cues, in supporting embodiment in dysmorphic avatars. Moreover, the lack of controlled investigations isolating auditory input from other modalities makes it difficult to disentangle its specific contribution to SoE. While previous work on audio-tactile stimulation has considered only punctual sounds [6], the role of continuous sounds remains unexplored in this context, and it is unclear whether such findings could extend to the perception of a whole altered avatar.

To this end, this study investigates the influence of continuous audio feedback, presented in combination with visuo-motor or visuo-tactile feedback, or in isolation, on the SoE towards an animal avatar. Two experiments were conducted in which participants were immersed in a dog shelter environment, embodying in first-person a virtual dog whose head movements were mapped to the participants' head orientation while experiencing different types of sensory stimuli (see Figure 3.1). A dog was chosen as the avatar to embody due to its structural morphology, which differs from that of humans, making it challenging to feel embodied in and to assimilate this change in body space. Furthermore, rendering individual hairs in VR is limited by high computational demands, since most dynamic models rely on a physical per-hair approach that requires calculating factors such as force and air-resistance accelerations [191]. This issue is magnified by the large number of strands that must be rendered [192], which highlights the need to explore alternative methods, such as auditory and tactile cues, to enhance the perceptual feeling of having fur in immersive environments.

3.1 Hypotheses

Based on the context and research gaps outlined above, a number of hypotheses were developed to frame the design of this study.

Experiment 1 evaluated audio and tactile feedback and their influence over the SoE. In the IVE, a virtual human character was animated to brush the dog avatar while participants received synchronously either audio, tactile (from the experimenter), audio-tactile or no additional feedback (as a control condition). The virtual character's action was chosen to facilitate the perception of spatial audio features where the sound changed in both position and direction while providing continuous tactile feedback to participants.

We believe that exposing participants to specific audio feedback, such as fur brushing sounds, will increase their sense of owning a virtual fur, and by extension their SoE towards the virtual dog, especially when paired with additional stimuli such as tactile feedback. More precisely, our main hypotheses are:

H1 Visual with Audio feedback will provide a higher SoE compared to Visual only.

H2 Visual with Audio-Tactile feedback will have a more significant effect on the SoE than all other conditions.

Based on previous work showing that audio-tactile feedback towards a virtual marble arm could induce a change in temperature, making participants feel cold [160], we also expected audio feedback to modulate perceived temperature, having an impact on participants' feelings of warmth as a result of embodying an avatar covered in fur. In addition to these perceptual effects, we considered the potential impact from a perspective of empathy. Embodying animals has been shown to raise awareness and affective responses towards animal abuse, wildlife, and environmental issues [4, 140, 138]. Therefore, we assumed that an increase in SoE towards the dog avatar would lead to higher empathy towards dogs in shelters.

To further investigate the influence of audio feedback on the SoE, a second experiment was conducted. Experiment 2 focused on audio alone, without tactile feedback as a factor, providing a more direct examination. We introduced the sound of scissors cutting hair, which is more distinct than the brushing sound, while still allowing for an ecologically feasible scenario. We believe that a sound that provides clearer audio cues of the virtual character's actions will elicit a stronger SoE over the dog avatar. In addition, we propose that reducing the number of conditions will enable us to increase the embodiment duration of each condition and help to prevent repetition that could induce boredom in participants. Therefore, our hypothesis is:

H3 Visual with Audio feedback will induce a higher SoE compared to Visual only.

3.2 Experimental Setup

The experiments were designed to run on a tethered Oculus Quest 2 system, comprised of the HMD and two hand-held controllers. The HMD was connected using a Meta Link Cable to a PC equipped with an Intel i9-11900KF 3.5GHz CPU with 32 GB RAM and an Nvidia 3090 graphics card. A pair of Beyerdynamic DT-770 Pro 80 Ohm headphones was connected directly to the computer's audio output for audio playback.

Virtual dog In this study, we chose a dog as the animal to be embodied by participants as it differs from humans in its natural posture while maintaining a similar skeleton and number of limbs [3] (see Figure 3.2), but also considering the cultural relationship and familiarity with the animal. A first-person point of view was employed to embody the dog since research has shown that it increases the sense of virtual body ownership and sense of self-location [193]. Since we were mainly interested in studying audio feedback alone, the dog's body was fixed to restrict the visuo-motor feedback, and participants only controlled the motion of the virtual dog's head with the HMD. Giving participants control of the dog's head was necessary to ensure an efficient first-person point of view towards the virtual avatar.



Figure 3.2: Dog model used in the experiments for participants to embody.

To obtain natural and realistic head movement while maintaining the fixed position of the dog's body, we experimented with several Inverse Kinematics (IK) methods to control the head and neck bones of the dog's rig with the HMD. However, due to the differences in translation between the HMD and the avatar head, the VR camera popped out of the dog's head when directly mapping the HMD's position to the avatar. Therefore, we developed a system that only used the HMD's rotation and applied a multi-rotation constraint to the head and the neck's three bones. We also clipped out the rendering of the snout from the first-person view to avoid inconsistencies between the natural-looking rotations of the dog and the human's own neck rotations.

Virtual Environment A dog-shelter environment was selected as an ecologically valid setting that also offered the opportunity to induce empathy towards dogs. In addition, the presence of a virtual human character in the scene grooming the dog was plausible in this scenario. The IVE was built with the Unity Game Engine (version 2021.3.2f), and all scripts were written in C#. In addition, BiomotionLab Toolkit for Unity Experiments [194] was used to run the experiment and log all of the participants' answers, allowing for developing and performing experiments efficiently. The virtual room and 3D models were designed in Unity or obtained from CG Trader and TurboSquid. The virtual character in the scene was taken from the Microsoft Rocketbox collection, while the dog model was sourced from Unity's asset store (see Figure 3.3).

The sound design and audio spatialisation in the virtual environment were implemented using the Oculus Audio spatialiser (version 32.0.0). We employed the Audio Propagation feature of the spatialiser, which allows real-time reverberation and occlusion modelling from the scene geometry. All the custom audio samples in the experiment were recorded in an acoustically treated room using an AKG C414 large-diaphragm condenser microphone and a Chandler Limited TG2 Dual Mono Mic Preamp. In addition, several sound effects taken from the Sonniss GameAudioGDC library were placed in the scene to create a more plausible sonic environment.



Figure 3.3: General view of the virtual shelter, including other dogs in cages and the virtual mirror placed in front of the participant embodied as a dog.

3.3 Measures

The questionnaire was composed of 7-point Likert scale questions from *strongly agree* to *strongly disagree* related to embodiment. We used Roth et al.'s questionnaire [195] because it was previously used for animal embodiment measure [3] and suited our study as we consider 2 (ownership and change) of the 3 factors used in the questionnaire. This questionnaire comprises questions to assess the sense of Ownership, Agency and Change towards the virtual body. Because our experiment-controlled design did not allow participants to move the dog avatar's limbs, we did not include questions related to agency, which does not impact measures of the other embodiment dimensions. One question was adapted to the experiment ("human" replaced by "animal").

Additionally, the questionnaire included threat-related questions adopted from Peck and Gonzalez-Franco [196]. A threat was introduced because participants' responses can serve as a subjective measure of embodiment (self-location and body ownership) towards the dog and help evaluate the extent to which they consider the virtual body to be their own [197, 198].

Finally, because there is no standardised questionnaire related to animal embodiment in VR, we also added specific questions for this study. In particular, we added a question to assess if participants felt like they had fur, and another question on how warm they felt during the experiment on a scale from -3 (much too cold) to 3 (much too warm), to explore a potential behavioural effect of being embodied in fur through self-subjective ratings. Moreover, we were interested in measuring participants' empathy for the animal as it could be influenced by the

Table 3.1: Virtual body perception assessment questions

ID	Question
OW1	It felt like the virtual body was my body.
OW2	It felt like the virtual body parts were my body parts.
OW3	The virtual body felt like an **animal** body.
OW4	It felt like the virtual body belonged to me.
CH1	I felt like the form or appearance of my body had changed.
CH2	I felt like the weight of my own body had changed.
CH3	I felt like the size (height) of my own body had changed.
CH4	I felt like the width of my own body had changed.
T1	When the needle prick happened, I felt the instinct to move to avoid it.
T2	I had the feeling that the syringe might harm me.
F1	*I felt like I had a fur.*
W1	*How warm did you feel during the immersion?*
EM1	*How much would you consider donating to help the cause of abandoned dogs in shelters?*

Note: The element in bold has been adapted from Roth et al.'s questionnaire [195] to match the experiment's avatar, questions in italic are specific to this study, and T1, T2 are from Peck and Gonzalez-Franco's questionnaire [196].

SoE participants' experienced towards the dog [138]. To do so, we measured how much money participants would hypothetically donate to help abandoned dogs in shelters, and they responded using a 6-point scale divided into $10 increments ($0 to $50). In total, the questionnaire consisted of 13 questions (see Table 3.1).

3.4 Experiment 1

The experiment was conducted using a within-subjects design to reduce errors related to individual differences across conditions. The order between each stimulation was counterbalanced between participants to account for potential ordering effects; therefore, variables were presented in all possible permutations. Two independent variables were considered, whose levels depended on the different feedback provided to participants. **Tactile** feedback could either be Visuo-Tactile (VT): the experimenter brushed the back and arms of the participant with a real brush synchronously with the virtual brush movement, or Without Visuo-Tactile (WVT): no tactile feedback provided. **Audio** feedback could either be Visual-Audio (VA): spatialised pre-recorded brush sounds were played synchronously with the virtual brush movement, or Without Visual-Audio feedback (WVA). When VT and VA were combined, the participant was exposed to the tactile and audio stimuli synchronously with the virtual brush movement. When WVT and WVA were combined, there was no additional feedback, only visual feedback of the brushing.



Figure 3.4: Dog avatar being brushed by the virtual character.

3.4.1 Stimuli

Visual Feedback The virtual environment was built using Unity Game Engine (version 2021.3.2f). In the virtual scene, an animated human character brushed the dog avatar (see Figure 3.4). The overall brushing animation consisted of four stages, focused on different parts of the dog's body: the back's centre, the back's right side, the back's left side and the left forelimb. These animations were designed to loop seamlessly to their recurrence and sequencing. Every separate animation was played twice in a sequence that looped at least two times, providing over 1:40 minutes of brushing. In synchrony with this visual information, audio, tactile and audio-tactile feedback were provided.

Audio feedback The audio implementation was mainly designed to induce a SoE, but we additionally included additional audio sources to increase the plausibility of the scene. The brushing sounds (played in synchrony with the virtual character's animation) were newly created dry monophonic recordings of a deer skin rug being brushed. Recording a live dog was not feasible, as we needed to capture noise-free and consistent recordings of continuous brushing audio. Therefore, we examined different static pieces, including rugs and skins of various materials and lengths, to identify the best alternative. A deer skin was chosen to obtain high-quality and realistic audio samples of fur, given the similarity between the hair length compared to the dog avatar. Furthermore, to ensure synchrony between the audio and visual feedback, the audio clips were trimmed to the length of the animation's fur brush strokes.

Additionally, as ambient noises can enhance the immersive qualities of IVEs [199], we created an acoustic scene of contextual sounds in the virtual room to enhance its plausibility, including air vent humming, dripping water from a sink, panting and whining from the other virtual dogs, and soft breathing of the virtual character. Every sound in the scene was spatialised to fix it to an object's specific location in the virtual room and provide coherent and dynamic auditory cues. In addition, all audio sources included an array of samples that triggered randomly without consecutively playing the same one to avoid repetition. These ambient noises were present in all experimental conditions at a level which would not disrupt the audio stimuli to preserve a similar level of immersion between conditions.

Finally, we created a virtual acoustic response matching the room's characteristics to enhance the acoustic environment's realism and provide additional sound localisation cues. We used the Audio Propagation feature of the spatialiser plugin to select appropriate materials for the surfaces of the virtual room with different acoustic reflection and absorption qualities, such as ceramic tile for the walls and floor and acoustic tile for the roof. Additionally, we enabled the early reflections and reverberation settings in the spatialiser's reflection engine.

Tactile and audio-tactile feedback While participants were immersed in the virtual scene, the experimenter provided tactile feedback by physically brushing their backs and arms with a plastic bristle hair-brush (see Figure 3.5). The feedback was synchronous with the virtual human's animations and was applied to: the centre of the back, the back's right side, the back's left side and the left arm, matching the locations of the dog. Participants were asked to wear a long-sleeved top made of thin material throughout the duration of the experiment in order to maintain the consistency of the brushing feedback between their backs and arms.

In order to maintain a more straightforward experiment setting, we did not employ real-time motion capture of the experimenter to animate the virtual human and synchronise the tactile feedback. Instead, to ensure the synchrony between the physical and virtual brushing, the experimenter monitored a progress bar displaying the duration of when the brush came into touch with the dog's body in each individual animation. Following the progress bar and the animation itself allowed the experimenter to time the brushing motion and speed accurately.

3.4.2 Participants

An a priori power analysis was performed using G*Power [200, 201] to determine the correct sample size for the experiment. To accommodate a medium effect size, we chose an effect size of d = 0.25, an alpha error probability of 0.05, and a power of 0.8. As we planned to run a within-group repeated measures ANOVA, we set the number of groups to 1 and the number of measurements to 4 to account for the conditions in the experiment. The correlation between repeated measurements and the nonsphericity correction were set at the default value. Based on



Figure 3.5: The experimenter brushing the participant's back during tactile feedback conditions.

the power analysis, 24 participants, recruited through institutions' mailing lists, took part in the experiment over a period of two weeks.

Prior to the experiment, participants read and signed an informed consent form and completed a demographic questionnaire about age (between the ages of 22 and 53, mean/s.d. age: 33.5±7.7), gender (13 males, 10 females, 1 prefer not to say), ethnicity (5 Asian or Pacific Islanders, 2 Hispanic or Latino, 17 White Caucasians), VR experience (5 experts, 2 moderate users, 9 sporadic users and 8 never used) and video games experience (4 experts, 7 moderate users, 10 sporadic users and 3 never used).

3.4.3 Procedure

The experiment was divided into two parts: a screening test to assess the extent to which participants could perceive spatialised sound, and the main experiment, in which participants were immersed in VR and exposed to different sensory stimuli.

Screening Test Phase As discussed in Section 2.2.1, most VR systems employ generic HRTFs for binaural rendering, which can negatively affect sound localisation and reduce externalisation. To ensure that participants perceived the audio feedback as intended, we conducted a screening test to verify whether the majority could perceive spatialisation and externalisation with the HRTF used in the experiment.

Prior to the main experiment, participants were immersed in a simplified version of the main experiment's room to reduce the amount of visual input. They were asked to face forward, listen

to a series of finger clicks and decide whether the sounds were externalised (appearing to come from outside their head) or internalised (appearing to come from inside their head); then to select their answer using a hand-held controller. The test consisted of eight trials, six of which were spatialised at different angular positions and distances both in front of and behind the participant, and two of which were non-spatialised. Results showed that the majority of participants (22 out of 24) correctly perceived externalisation, so we deemed the generic HRTF appropriate for use.

Experiment Phase Participants were asked to sit on a stool in a neutral position and were immersed in VR with an HMD and wearing headphones (see Figure 3.5). They were not in a quadruped position, as we were interested in a more challenging embodiment (different posture between avatar and participant) and also deemed it more likely in the case of virtual experiences that the human would be in a bipedal pose, which is encouraged by Krehov et al. [3]. Participants were given a brief about the experiment, stating that they would be immersed in a dog shelter IVE embodied in a virtual dog and be able to see their reflection (as their virtual dog representation) in a virtual mirror see Figure 3.4). In addition, they were informed that during the immersion phase, they would see themselves being brushed by a virtual character while simultaneously receiving different kinds of sensory feedback and that following each of the four conditions, they would be asked to complete a questionnaire using a hand-held controller.

The scene represented a dog shelter and contained other virtual dogs in cages to increase its plausibility. The virtual dogs were animated with light and natural animations to provide realism to the scene without distracting participants. During the embodiment phase, participants could see themselves being brushed by a virtual character and were exposed to different kinds of sensory feedback synchronised with the visual movement of the hairbrush. After approximately two minutes of immersion, the experimenter triggered the threat using the left handheld controller, which involved the virtual avatar assaulting the dog with a syringe in a threatening and forceful manner, mimicking a lethal injection (see Figure 3.6). The threat was activated when participants were looking at themselves in the mirror, allowing them to watch the virtual character's actions clearly. Consistent with prior research [197, 198, 202], the user's avatar behaviour was not artificially modified during the threat, and participant's body was not really threatened. Following each of the experiment's four conditions, participants completed a questionnaire using a hand-held controller.

3.4.4 Results

The parametric analyses were performed using multiple two-way repeated measures ANOVAs, with **Tactile** and **Audio** feedback as within-subject factors. The normality assumption was tested using Shapiro-Wilk test, and when this assumption was not met, an Aligned Rank Transformation (ART) [203] was applied to the data. This statistical method is designed for analysing multifactor,



Figure 3.6: Threat used in Experiment 1: the dog avatar being administered the lethal injection.

nonparametric data, commonly observed in Likert-scale responses, and enables the analysis of interaction effects, which traditional nonparametric tests often fail to address. Tukey's Post-hoc tests ($\alpha = .05$) were conducted to check the significance for pairwise comparisons. In addition, post-hoc tests were corrected using Bonferroni correction. As for correlation analyses, Pearson's r (r) was used for parametric data and Spearman's r (r_s) was used for non-parametric data. Results of the questionnaire were split into the sense of Ownership (OW1-OW4) and Change perception (CH1-CH4) following Roth et al. [195] methodology, Threat perception (T1-T2), Fur Ownership perception (F1), Warmth perception (W1) and Empathy (EM1).

Sense of Ownership The analysis performed on the aggregated results of the four Ownership questions showed no effect of **Audio** but a significant effect of **Tactile** feedback ($F_{1,23} = 4.342$, $p = .048$, $\eta_p^2 = 0.16$). Post-hoc analysis showed that Ownership scores were higher in the condition with just tactile (VT) as compared to without tactile WVT ($p < .05$). A second analysis was performed on each Ownership item independently, since the aggregated score did not reveal an effect of Audio, and we sought to investigate further. OW1 showed a tendency toward an effect of Audio feedback ($F_{1,23} = 3.721$, $p = .06$, $\eta_p^2 = 0.139$), which meant that audio feedback increased the feeling that the virtual body was their body (see Figure 3.7, left).

Change The analysis performed on the aggregated results of the four Change questions showed a main effect of **Tactile** feedback($F_{1,23} = 9.289$, $p < .01$, $\eta_p^2 = 0.29$). Post-hoc tests showed significantly higher scores of **VT** compared to **WVT** ($p < .01$). Participants felt more *change*

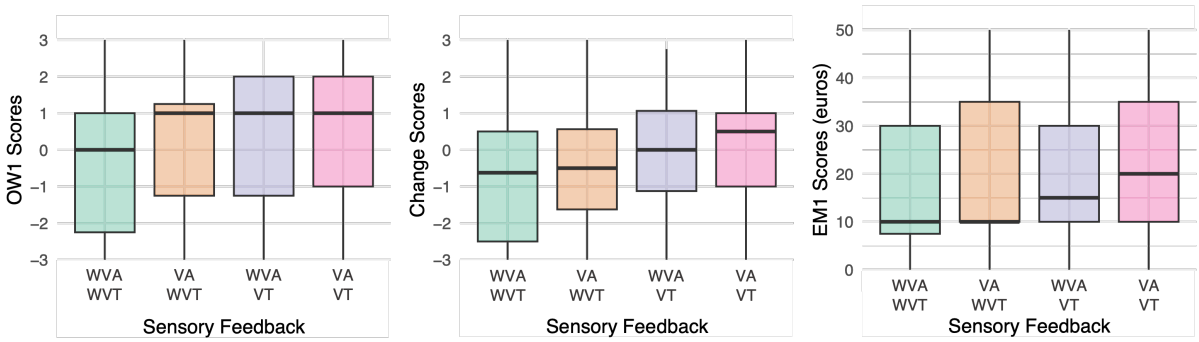


Figure 3.7: Experiment 1 results for one Ownership question (OW1), Change dimension of the SoE and Empathy measure.

towards their body when they received tactile feedback synchronised with the visual brushing compared to the other conditions (see Figure 3.7, centre).

Threat Perception The analysis performed on the aggregated results of the two Threat questions did not show any main effect of **Tactile** or **Audio** feedback, and no interaction effect was found. Participants did not seem to react to the threat over their virtual dog bodies, and the threat stimulus did not elicit a measurable response. This could be because the syringe animation was not sufficiently surprising or threatening, or because it targeted a body region where participants did not feel embodied.

Fur Perception For F1, Post-hoc tests showed significantly higher scores for **VT** compared to **WVT** ($p < .01$). This implied that participants felt more that they had a fur with tactile feedback. Although not statistically significant, a higher **Audio** feedback rating was identified ($F_{1,23} = 3.282$, $p = .08$, $\eta_p^2 = 0.12$). No interaction effect was found.

Warmth Perception The analysis performed on W1 did not show a main effect of **Audio** feedback, only had a higher rating of **Tactile** influence, though not significant ($F_{1,23} = 3.44$, $p = .07$, $\eta_p^2 = 0.13$). No effect of **Tactile** feedback or interaction was found.

Empathy The analysis performed on EM1 showed a significant main effect of **Tactile** feedback ($F_{1,23} = 7.042$, $p < .05$, $\eta_p^2 = 0.234$). Post-hoc tests showed significantly higher scores in **VT** condition than in **WVT** condition ($p < .05$). A tendency of **Audio** influence was found ($F_{1,23} = 3.27$, $p = .08$, $\eta_p^2 = 0.12$) (see Figure 3.7, right). Participants considered donating more money towards the cause of abandoned dogs in shelters when they had tactile feedback of the brushing compared to all other conditions. As expected, a positive correlation was found between Ownership scores and Empathy towards dogs in shelters ($r_s = 0.28$, $p < .01$).

Game and VR experience A ranked score was attributed to Game and VR experience self-reported by participants, in order to perform correlation analysis with other perceptual items. A positive correlation was found between Ownership scores and Game Experience ($r_s = 0.30$, $p < .01$), but not with the other SoE dimensions (Change and Threat) or sense of warmth. No correlation was found between VR experience and the other components. Overall, this analysis highlighted that participants with more gaming experience were more likely to feel a strong sense of ownership towards the dog avatar.

Referring back to our hypotheses, we found that **H1** was not validated, meaning that the addition of our naturally-recorded spatial audio feedback to the visual stimulus did not improve the SoE. While this result contrasts with works showing that audio feedback contributes towards changing one's mental body representation [161], it seems to be in line with the few other previous works investigating audio on embodiment [38, 39]. A likely reason is that the recording did not match participants' expectations of what brushing fur should sound like, or that the audio did not provide sufficiently distinct cues about the action being evoked. Another possible reason is that, despite the externalisation check, the use of generic HRTFs may have negatively influenced participants' perception.

Similarly, **H2** was not validated, as audio feedback did not enhance tactile feedback. We suspect this is because the audio feedback was not a strong enough stimulus to provide an added effect over tactile feedback, or due to the incongruency between the brushing sound and the actual tactile sensation of the brush on participants' clothes. Notably, this result was not exclusive to tactile feedback, as audio feedback did not significantly improve any of the other conditions either. However, several trends of audio influence were observed on one ownership item, fur perception and empathy, underscoring the need to deeper explore their perception with audio feedback.

3.5 Experiment 2

Since Experiment 1 did not yield statistically significant results in the audio-only condition and suggested that tactile feedback was required to elicit SoE, we conducted a second experiment. The motivation arose from the limitations of Experiment 1. In particular, the brushing audio did not improve embodiment when presented alone. Since the sound of brushing is essentially noise, it may have been difficult for participants to recognise it as fur-related. Nonetheless, non-significant trends in Ownership, Fur Perception, and Empathy suggested that auditory feedback could still play a role, motivating refinement of the auditory stimulus. To address this, we implemented a new scenario in which the virtual character performed a hair-trimming action. The sound of scissors cutting hair is more distinct and easily recognisable than the predominantly white-noise brushing sound, while still allowing the virtual scissors animation to remain in contact with the fur of the virtual dog.

Furthermore, given the lack of significant effect shown with the lethal injection threat and the



Figure 3.8: Threat used in Experiment 2: a ceiling light falling over the dog avatar.

potential ambiguity of the related action, we introduced a new threat in this experiment. To make the danger more prominent and perceptually evident, we employed a ceiling light falling from above for participants to react to (see Figure 3.8). In addition, to account for individual differences in empathy, participants completed an Interpersonal Reactivity Index (IRI) questionnaire [204] which measures four subscales of empathy (perspective taking, fantasy, empathetic concern, and personal distress) and their correlations. We chose to employ this questionnaire as it was found to be correlated with empathetic behaviour in animal embodiment studies [4].

Finally, since we intended for participants to have a natural experience and be fully immersed in the shelter scene to elicit empathy towards the dog, in Experiment 1, they were not explicitly instructed to look at their virtual reflection in the mirror and focus on the brushing. As a result, many explored the scene during exposure, which may have reduced attention to the brushing stimuli and weakened the tactile and audio effects. Therefore, in Experiment 2, participants were further instructed to look at themselves in a mirror during the exposure, allowing them to explore the scene beforehand and extending the embodiment time.

In Experiment 2, an animated virtual human character trimmed the hair of the dog avatar while participants synchronously received either audio or no additional feedback (control condition). The experiment followed a within-subjects and counterbalanced design with **Sensory Feedback** as the main factor with two levels depending of the sensory feedback provided to participants:

- Audio (VA): spatialised pre-recorded scissors sounds played synchronously with the movement of the virtual scissors opening and closing;

- Control condition (V): no additional feedback, only visual feedback of the hair trimming, using a particle system to generate hair particles flying through the air away from the body.

As this experiment builds upon the first one, all components, including the IVE, virtual dog, apparatus, screening test, experiment procedure, and measures, were retained to ensure methodological consistency.

3.5.1 Stimuli

In the virtual scene, an animated human character trimmed the hair of the dog avatar (see Figure 3.1). The trimming animation covered the same areas of the dog's body as the brushing animation of Experiment 1: the centre of the back, the back's right side, the back's left side and the left forelimb. However, to ensure that participants had enough embodiment time, the number of loops was increased so that participants were immersed in the IVE for 2:30 minutes for each condition. Of the total immersion time, participants were asked to look around and experience their surroundings for the first 30 seconds. Afterwards, they were instructed to look at the mirror for the rest of the immersion to ensure participants focused on the mirror and their reflection for the remaining 2 minutes, as done in previous studies [145, 142].

Audio feedback Similar to Experiment 1, the auditory scene included sounds to induce a SoE and to increase the plausibility of the scene. The scissor sounds were spatialised audio samples of scissors cutting hair, edited to trigger in synchrony with the animation of the scissors. Besides, to provide the same coherent and dynamic auditory cues and preserve a similar level of immersion as in Experiment 1, the auditory scene, including all the contextual sounds and the virtual acoustic response of the IVE, remained the same.

3.5.2 Participants

We performed an a priori power analysis using the same parameters as in Experiment 1. However, the number of measurements was reduced to two to account for the experimental conditions. Over the course of one week, 34 participants took part in this experiment. Participants were recruited in the same way as for Experiment 1. Prior to the experiment, participants read and signed an informed consent form and completed a demographic questionnaire about age (between the ages of 18 and 51, mean/s.d. age: 23.6±7.2), gender (17 males, 17 females), ethnicity (13 Asian or Pacific Islanders, 3 Black or African Americans, 18 White Caucasians), experience with VR (1 expert, 7 moderate users, 10 sporadic users and 15 never used) and experience with video games (5 experts, 17 moderate users, 11 sporadic users and 1 never used). The screening test results showed that the majority of participants (32) could correctly distinguish between internalised and externalised sources.

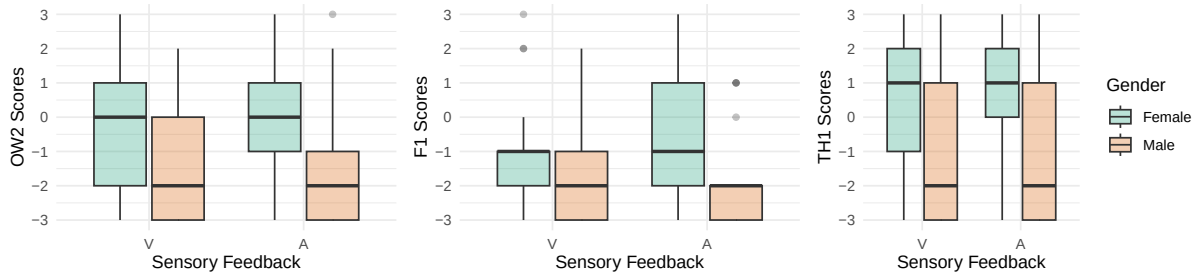


Figure 3.9: Experiment 2 results for one Ownership question (OW2), Fur Ownership perception (F1) and Threat perception (TH1), depending on sensory stimuli and participants' gender.

3.5.3 Results

The same measures from Experiment 1 were considered, with the addition of the IRI questionnaire [204], and the same analysis methods were applied (see Section 3.4.4). However, this time multiple one-way repeated measures ANOVAs were performed in order to search for effects of **Sensory Feedback** on the dependent variables. No significant impact of **Sensory Feedback** was found on any of the dependent variables (Ownership, Change, Threat, Fur and Warmth Perception, and Empathy), nor on individual items (e.g., OW1, OW2, etc.). **H3** was not validated as our audio feedback alone did not elicit a better SoE towards the dog avatar than without it, despite the efforts in improving its quality compared to experiment 1. Our result is in line with another study that explored audio feedback in VR to elicit embodiment towards a virtual body [38], yet their experiment combined audio with visuo-motor feedback, which we avoided in our study to isolate the impact of audio only. Our study, therefore, seems to indicate that inducing a SoE towards a virtual body using audio feedback alone is potentially more complex to achieve than the methods used in this study. This is further discussed in Section 3.6.

In order to assess the potential influence of gender on the results, we additionally conducted two-way ANOVAs considering Gender as a between-groups factor (see Figure 3.9). A main effect of Gender was found on OW2 ($F_{1,32} = 5.284$, $p < .05$, $\eta_p^2 = 0.141$), F1 ($F_{1,32} = 5.057$, $p < .05$, $\eta_p^2 = 0.136$) and TH1 ($F_{1,32} = 6.900$, $p < .05$, $\eta_p^2 = 0.177$). Post-hoc tests show higher scores for female compared to male participants for OW2, F1 and TH1 ($p < .05$ each). We searched for the influence of gender because previous work already highlighted its potential influence on the SoE [198] and because our sample size was big enough to conduct such an analysis. Again, to our knowledge, the influence of gender was never before observed in embodiment studies towards dysmorphic avatars. While our results tend to show that female participants felt more embodied in the animal avatar, we cannot know if this result is really due to gender differences or other biases between the two samples.

No correlation was found between IRI scores and empathy (EM1) scores, nor with any of our

dependent variables except Threat ($r_s = 0.38$, $p < .001$). Interestingly, compared to Experiment 1, no correlation was found between Ownership and empathy (EM1) scores.

We included the questionnaire IRI, which aims to measure subscales of empathy, and was found to be correlated with empathetic behaviour for participants embodied in virtual animals [4]. Interestingly, in our study, no correlation was found between IRI scores and empathetic behaviour. However, the measure of empathetic behaviour differs from theirs (hypothetical amount of money to give to shelter associations compared to perceived pain related to the distress event in the game), which might explain the difference.

3.6 Discussion

In this chapter, we addressed the relative influence of audio and tactile feedback on SoE towards a dysmorphic avatar, in particular, towards a virtual dog. We did not validate our hypotheses regarding the influence of audio feedback alone to elicit a stronger SoE compared to visual feedback alone. Although this was achieved in previous work when audio was combined with visuo-motor feedback [38], our results seem to consolidate the fact that this combination is necessary to observe a beneficial influence over the SoE. Yet, we believe it is important to note that some tendencies of audio influence were observed and that subjective scores were consequently spread, highlighting strong variability in participants' responses.

Because the differences between humans and animals are important (morphology, fur, etc.), it might be that, independently of the stimulation (audio, tactile), removing visuomotor feedback makes it too challenging to feel embodied in such nonhuman characters. Therefore, it would be interesting to see how our results might transfer to a human character, for instance, by inducing a feeling of owning a beard or specific clothing. Besides, it should be noted that the experiment only provided control over the animal's head to preserve a first-person point of view, without full-body tracking, to minimise additional factors influencing the SoE towards the dog. However, it would be interesting to explore if adding visuo-motor feedback interferes with the impact of audio-tactile feedback on the SoE.

None of our subjective measures of embodiment were impacted by audio feedback only, even the one related to threat perception. This raises the question of possible ambiguity regarding whether the threat was perceived from a first-person or third-person perspective (fear of one's own body being hurt vs. fear of an animal being hurt), but also the limitations of subjective questionnaires. We believe that additional objective measures of embodiment should also be explored in this context, such as actual body temperature changes in combination with subjective warmth questions. Moreover, previous studies on animal embodiment use different measures and questionnaires, making it challenging to compare results across studies. Therefore, in order to establish a clearer definition of animal embodiment and its underlying mechanisms, it would be valuable to build a standardised questionnaire with measures that take into account key features

and components of animal embodiment, such as posture, shape, control, agency, empathy, etc.

Although empathy was not the main goal of this study, the results of the Empathy item in Experiment 1 suggested increased perspective-taking with higher embodiment. However, in Experiment 2, no correlation was found between empathy and ownership. These effects are important to enhance pro-social behaviours, with Slater and Banakou [205] showing that a dual VR exposure model can enhance this further, suggesting that embodying the dog and the virtual character that threatens it could be a promising avenue to investigate. Moreover, while animal embodiment is expected to induce empathy and have positive implications towards animal welfare, the long-term effects of embodying an animal on the user are yet to be explored.

Previous work showed that personality traits could influence one's SoE towards a virtual anthropomorphic avatar [198]. Yet, whether individual differences such as personality traits could influence embodiment towards non-anthropomorphic avatars remains unclear. In our second study, we measured another personality trait - an empathy trait (IRI), which was correlated with our threat dimension of Peck and Gonzalez-Franco's embodiment questionnaire [196]. This reinvigorates the need to explore further how other personality traits may influence the challenging SoE towards a dysmorphic avatar.

While this study explored improving embodiment towards a virtual dog by enhancing the sensation of having fur, future work could investigate the use of other avatar-related material properties to strengthen embodiment, for example, clothing sounds for human avatars or feathers for birds. In addition, other body-related sounds, such as footsteps or barking, may be easier to identify from a first-person perspective and could further support embodiment.

3.7 Conclusion

This chapter examined the influence of auditory feedback as an alternative factor to visuo-motor and visuo-tactile feedback in enhancing the SoE towards a non-human avatar. This perceptual illusion is fundamental when embodying virtual characters, as it enables users and actors to experience the sensation of owning and controlling the virtual body.

We introduced a novel audio-tactile feedback through continuous brushing or cutting of the virtual fur to enhance the feeling that the human's skin was covered in fur. It was the first time that audio-only feedback (without visuo-motor or visuo-tactile feedback) was studied to elicit a SoE towards a non-human avatar. Our results indicate improved ownership with combined audio-tactile feedback, but not with audio or tactile feedback alone. This underscores the potential of auditory feedback in multimodal contexts for eliciting psychological effects in IVEs, supporting **RQ1**. In addition, we believe our research provides interesting insights into this much under-explored topic.

As a first exploration of embodied auditory feedback, the findings highlight both the limitations and opportunities of audio feedback as a sensory modality for eliciting psychological effects in

immersive experiences. These insights help frame broader research on sounds generated through the interaction of objects with one's own body and identify promising research avenues for future exploration. Moreover, this work helps frame a methodological and conceptual framework for investigating the perception of sounds that are more closely connected to our body and intrinsically experienced from an embodied perspective, such as one's own voice. It also extends towards examining different dimensions of auditory phenomena, including acoustic features, their role in creating more plausible environments, and their influence on perceptual effects such as the sense of presence. This framework provides the basis and motivation to explore the perception of audio-visual environments through vocal auralisation, which is examined in detail in Chapter 5.

Finally, this study identified the features and limitations of current real-time audio rendering tools and spatialisers, particularly regarding latency and the lack of control over direct and processed signals, which limit their usability. These constraints highlight the need to develop an audio spatialisation pipeline that enables real-time auralisation of user-driven input signals within IVEs, providing high-fidelity audio feedback and supporting embodied exploration of virtual spaces. The following chapter addresses this need by presenting the design and implementation of VOREA, a pipeline for vocal auralisation in IVEs.

Chapter 4

VOREA: Design & Implementation of a Vocal Real-time Auralisation Pipeline

Based on **“Real-Time Auralization for First-Person Vocal Interaction in Immersive Virtual Environments”**
Article presented at the AES 158th Convention 2025



Figure 4.1: Loudspeaker and Ambisonics microphone setup during the acoustic capture.

This chapter presents the design, development, and implementation of the **Vocal Realtime Auralisation (VOREA)** pipeline, created to enable real-time first-person vocal auralisation for multimodal immersive experiences using SIRs interpolation. By convolving the user's voice with SIRs corresponding to different positions and orientations within the IVE, VOREA allows users to perceive in real-time the acoustic features of a space. This provides them with the agency to explore and interact with the space visually and acoustically through dynamic auralisation rendering that reflects their movements. The approach ensures a meaningful representation of how their voice interacts with the environment, enriching the multimodal experience by integrating audio and visual feedback into a cohesive, immersive experience.

Commercial audio spatialisers used in immersive sound design, such as Meta XR Audio, Steam Audio, Unity DearVR, and Google Resonance, provide a broad set of features, including sound spatialisation, source directivity, sound propagation, and Ambisonics support. A common capability among these tools is the simulation of room acoustics, which renders early reflections, reverberation and other acoustic phenomena that align with the visual representation of the virtual environment, enhancing perceptual plausibility and realism by embedding sounds within an acoustic context. However, the motivation behind the development of this pipeline stems from the limitations of existing auralisation tools, which make them unsuitable for scenarios requiring real-time, realistic auditory feedback of self-produced sounds, such as vocal input. First, they exhibit significant signal latency when processing real-time audio input. A test was conducted to evaluate the audio latency of spatialisers within Unity, revealing an approximate latency of 400 ms between the microphone input and the processed signal. The latency was calculated by recording both the microphone input and the headphone output after real-time processing and measuring the time difference between the two signals. Second, the evaluated spatialisers lack independent level control of the direct and processed signals (the auralised output). Since users hear their voice directly while speaking, the system only needs to reproduce the acoustic response, as the direct input can be routed back through the direct monitoring features of audio interfaces to minimise latency. However, due to the absence of independent control of audio signals or a dry-wet mixing feature, the output always includes both the direct signal and the acoustic response, exacerbating the latency issue. Finally, the auralisation techniques employed by these systems are mostly based on synthesised models, either precomputed static responses or simplified raycasting methods (typically shoebox approximations), which are inadequate in scenarios where acoustic fidelity and real-time rendering are essential.

The VOREA pipeline has been divided into two distinct systems to address the challenges mentioned above. The first focuses on the considerations for measuring and generating SIRs from a performer's perspective in the physical environment (see Figure 4.2). It accounts for essential factors such as loudspeaker and microphone placement within the stage, as well as methods for effective and streamlined recording. The second is dedicated to the real-time

auralisation of the user's voice. This involves signal processing to convolve the input signal with the relevant SIRs and accurately interpolating the SIRs' output according to the user's position and orientation within the virtual environment. It describes key elements such as the calibration of direct input and process output, data flow between software platforms (game engine and digital audio workstation), and Virtual Studio Technology (VST) plugin handling for effective resource management. By splitting the pipeline, it was possible to address these challenges more effectively, since this modular approach not only improves the overall effectiveness of the system but also simplifies debugging and facilitates future enhancements.

4.1 SIR Production

The creation of 360° SIRs for this pipeline is based on the methods outlined by Kearney et al. [43] and Mroz et al. [44], which present a setup and acoustic measurement procedure for producing SIRs. These SIRs are widely used to provide accurate acoustic representations of physical spaces and have been effectively applied in vocal auralisation research [35, 34, 206, 28].

A key consideration in the production of SIR datasets for vocal auralisation arises from the captured perspective. Most available datasets are produced using a configuration where the source and receiver are placed at different positions. The source (the loudspeaker) is placed on the stage to represent the performer, and the receiver (the microphone) is placed in the seating area to represent the listener, referred to here as the *audience perspective* (see Figure 4.2, left). This setup is effective in cases where the user represents the audience. However, for first-person vocal interaction, the configuration must follow a *performer perspective*, as it will be referred to in this thesis, in which the source and receiver are co-located to simulate the mouth and ears of a person (see Figure 4.2, right). This setup is essential for representing auditory experiences in which the user explores the acoustic space using their voice.

Although the dataset by Kearney et al. [43] includes SIRs captured at different positions



Figure 4.2: Left: *Audience perspective* where the source and receiver are spatially placed. Right: *Performer perspective* where source and receiver are co-located.

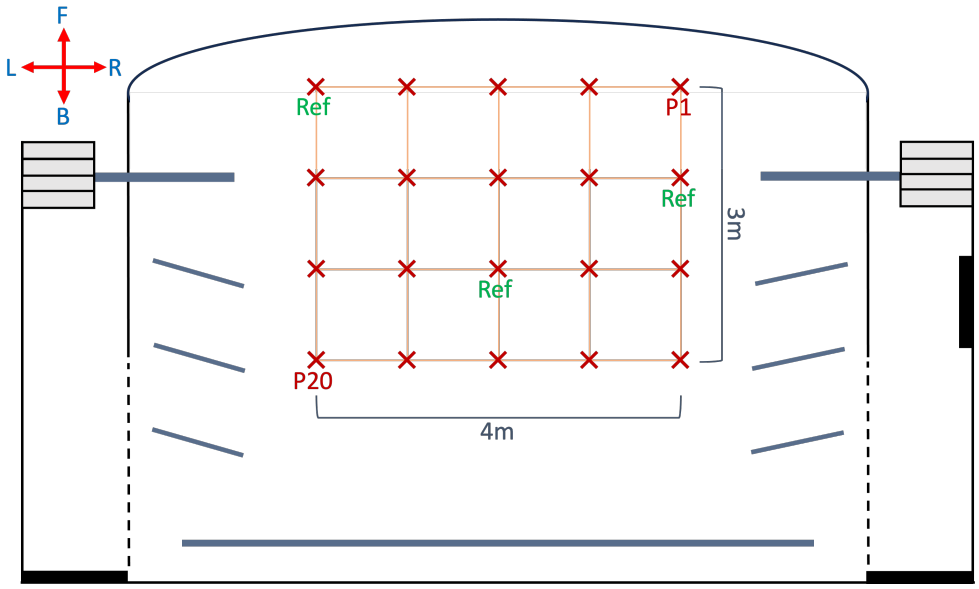


Figure 4.3: Grid of positions used for SIR capture in the Abbey Theatre. The front direction is relative to the performer's point of view.

and orientations, the performer-perspective SIRs included in their dataset were obtained from non-adjacent positions, restricting their application to directional degrees of freedom. Thus, the proposed system is an adaptation aimed at capturing SIRs from a range of adjacent positions and orientations. This enables translational degrees of freedom, providing the user with spatial flexibility to explore the space freely. It is important to note, however, that VOREA is limited to degrees of freedom in the horizontal plane (x- and y-axis) as the elevation remained fixed in all acoustic measurements. Including the vertical dimension (z-axis) would involve recording layers of impulse responses at different heights [207]. Given that singers typically maintain a stable vertical position during performances, we deemed the inclusion of elevation unnecessary due to its significant complexity and cost implications, and because this approach has proven effective in previous studies with comparable scenarios [34, 44].

4.1.1 Pre-recording Setup

To allow for translational DoF, a grid comprising 20 recording positions spaced 1 m apart was defined, providing a total area of 12 m^2 (4 x 3 m) for user exploration (see Figure 4.3). Initially, a triangular distribution, as used by Mroz et al. [44], was considered since it is less computationally intensive due to the number of SIRs involved at the interpolation stage. However, this arrangement results in a smaller area for the same number of recorded positions (approximately 9 m^2 for 20 positions spaced 1 m apart) with an irregular shape.

To enable rotational DoF, four directions were considered for each position: front, right, back, and left [43]. This approach was necessary because the speaker used (Genelec 1029A) to excite the environment has a somewhat subcardioid directivity pattern [208], comparable to

the directivity of the human voice. As a result, the acoustic response varies depending on the speaker's orientation within the acoustic space (see Figure 4.4).

A Digital Audio Workstation (DAW) project was set up to streamline the recording process. The project included a mono audio track set at -20 dBFS containing 80 sine sweeps, each 20 seconds long, corresponding to 20 positions with 4 directions each. Preceding each sine sweep, the project included a voice identifier that specified the respective position and direction, followed by a sync-pop to ensure synchronisation during subsequent processing. Additionally, three reference recordings were included for comparison against the convolved signal outputs. A 5-second interval followed each sine sweep to allow the reverberation to decay before the next recording. In addition, the project included a 20-channel track to record the input signal from the 19 omnidirectional condenser microphone capsules of the Zylia ZM-1 microphone [209]. These recordings were subsequently encoded into third-order Ambisonics (3OA) using the Zylia Ambisonics Converter plugin. The total length of the DAW project was about 45 minutes in duration for 80 impulse response measurements.

4.1.2 Recording Process

Following the approach presented by Kearney et al. [43], acoustic recordings were performed with a 3OA microphone (Zylia ZM1) to capture acoustic responses within a 360° sound field. The speaker was set at a fixed height of 1.5 m, with the microphone placed 5 cm above, simulating the relative position of the mouth and ears of a person and representing a first-person perspective (see Figure 4.1). The speaker output was calibrated to 85 dBA an approximate level of singing or theatrical speech [210, 211] using an SPL meter placed 1 m away. An SPL reading was also taken at the microphone position as a reference. The microphone gain was then adjusted to peak around -3 dBFS to prevent clipping caused by resonances in the acoustic space. During the recording process, both the speaker and microphone were placed at each designated position. The microphone was consistently oriented to the front of the stage, while the speaker was rotated sequentially to face each of the four directions: front, right, back, and left.

4.1.3 SIR Creation

After completing the measurement recordings, a single 16-channel WAV file (3OA) was exported that included all 80 measurements. A MATLAB script was developed to segment the file into 80 individual measurements, using the sync-pops for precise segmentation. Each channel was exported separately, resulting in 16 mono audio files per measurement. These mono files were deconvolved using Voxengo's Deconvolver [212], chosen for its reversed test tone deconvolution technique, which effectively mitigates artifacts and improves the quality of impulse responses. As the software supports a maximum of eight channels, processing was performed on the mono files, which were then recombined into 16-channel SIRs after deconvolution.

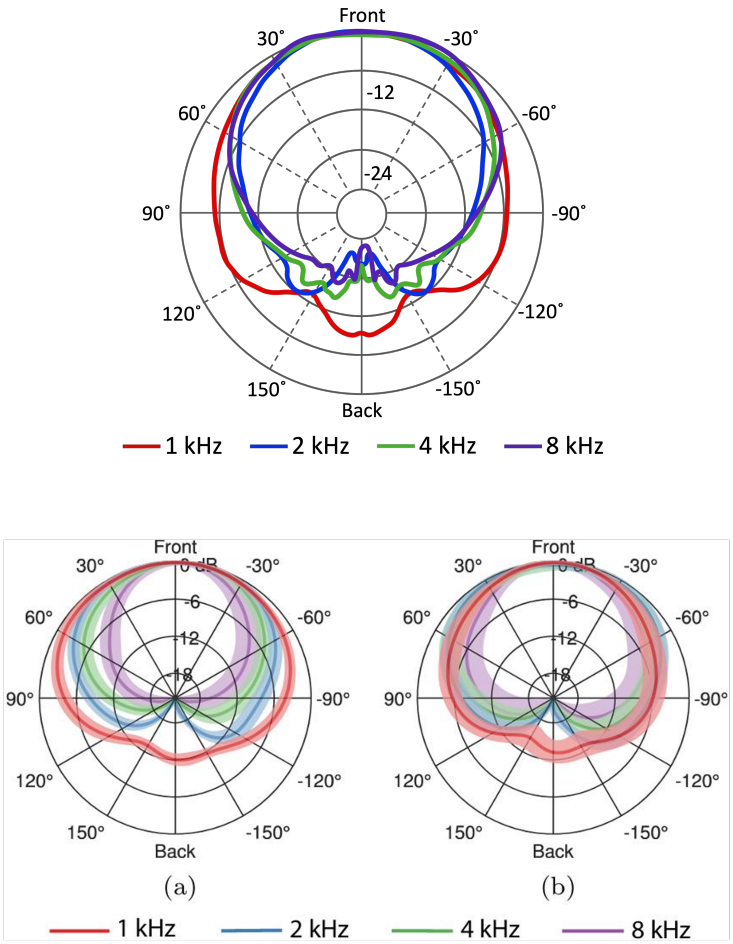


Figure 4.4: Top: Directivity patterns of the Genelec 1029A loudspeaker at 1, 2, 4, and 8 kHz [13]. Bottom: Mean directivities and standard deviations in the horizontal plane for an *[a]* (a) and an *[f]* (b) at 1 kHz, 2 kHz, 4 kHz, and 8 kHz measured by Pörschmann et al. [14]

4.2 Auralisation Pipeline

This part of the pipeline addresses the real-time auralisation of the convolved input signal, specifically the performer's voice, based on the user's position and rotation within the virtual environment. To achieve the necessary level of freedom, the system interpolates in real-time the output gain of the microphone input convolved with each of the SIRs. Therefore, the interpolation happens across different positions and four different directions (see Figure 4.5).

4.2.1 System Setup

This section outlines the system setup essential for real-time vocal auralisation. It encompasses the signal flow configuration, the amplitude compensation and gain calibration procedures, and the processing and alignment of SIRs for effective latency management.

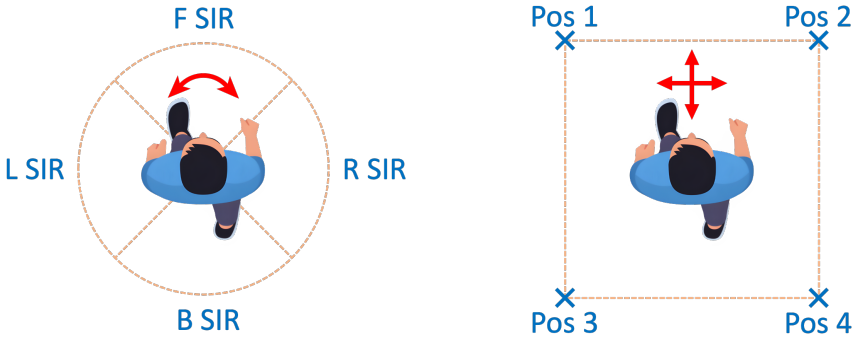


Figure 4.5: Directional interpolation (left) and Translational interpolation (right).

DAW Project Setup for Auralisation The audio signal processing was handled within the Reaper DAW [213]. The project was organised using folder tracks. Each folder track corresponded to a position on the area grid and contained four 16-channel audio tracks representing each direction (front, right, back, left), with their output routed directly to the parent folder track. Each audio track corresponding to a direction included a JS Time Adjustment Delay VST [213] and an X-MCFX multichannel convolver VST [214], loaded with the corresponding SIRs as inserts. The microphone input was assigned to a single 16-channel track containing an Ambix Encoder VST [215] to encode the mono signal into 3OA, which was then routed to the directional tracks of all positions for convolution. Finally, all position folder tracks were routed to the Master track, which contained an IEM Binaural Decoder VST [216] to decode the Ambisonic signal into a binaural headphone signal (see Figure 4.6).

Direct Signal Amplitude Compensation Noise-cancelling headphones were used to block the existing acoustic response of the physical environment. While some studies propose using open-back headphones to allow direct sound to reach the musicians ears [206], we opted for noise-cancelling headphones based on the fact that, unless users were in an acoustically transparent space such as an anechoic chamber or heavily treated room, open-back headphones would result in hearing a mix of both the virtual auralisation and the physical room's acoustic response. However, using headphones, especially noise-cancelling ones, results in a perceived amplitude loss of the user's voice. To account for this, the amplitude loss was measured by comparing white noise recordings made with a Sennheiser KU100 dummy head microphone, with and without the noise-cancelling headphones placed on its ears, and a speaker positioned 1 m away (see Figure 4.7). The input of the microphone was then fed directly to the headphones via the audio interface's direct monitoring feature, which allows listening to the input signal with near-zero latency. The monitoring level was adjusted based on the measured amplitude loss to compensate for the attenuation and ensure that the direct voice signal reproduced over the headphones matched the perceived level without headphones.

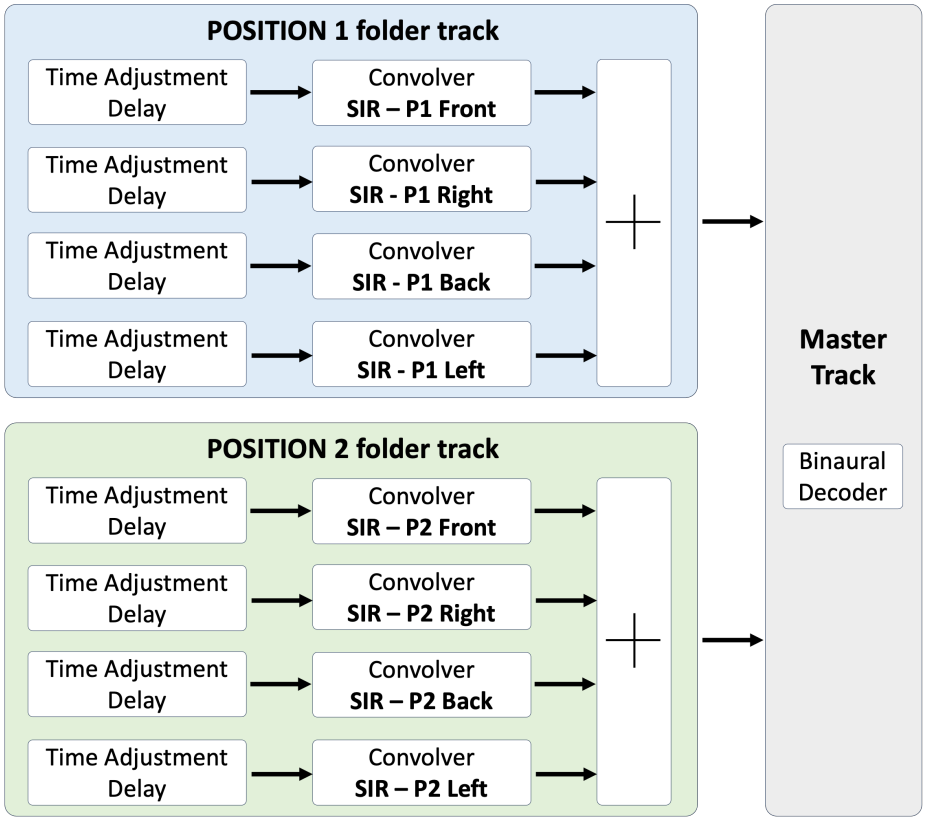


Figure 4.6: Signal flow setup for two of the twenty folder tracks in the Reaper project.

Latency Compensation One of the biggest challenges in self-produced sound processing in real time is the inherent latency between producing a sound and hearing its processed output. This latency is not fixed and arises from multiple factors, including the audio interface and the system's digital processing and buffering before it reaches the headphones' output. Therefore, we aimed for the lowest buffer size possible without performance glitches or dropouts.

Since the direct sound is fed directly to the user via direct monitoring, the SIRs were trimmed up to the first reflection, removing both the direct sound and the Initial Time Delay Gap (ITDG) (see Figure 4.8). The trimmed SIRs were then aligned within the Reaper project using a delay plugin, with the delay time compensation determined by subtracting the round-trip latency from the removed portion. The ITDG to the nearest surface (the ceiling, in this case) was calculated to accurately estimate the arrival time of the first reflection, using the following expression:

$$\Delta t = t_1 - t_0 = \frac{(L_1 + L_2) - L}{c}$$

where t_0 and t_1 are the arrival times of the direct and reflected paths respectively, L is the direct path length, $L_1 + L_2$ is the reflected path length, and $c \approx 344\,\mathrm{m/s}$ corresponds to the speed of sound [217].

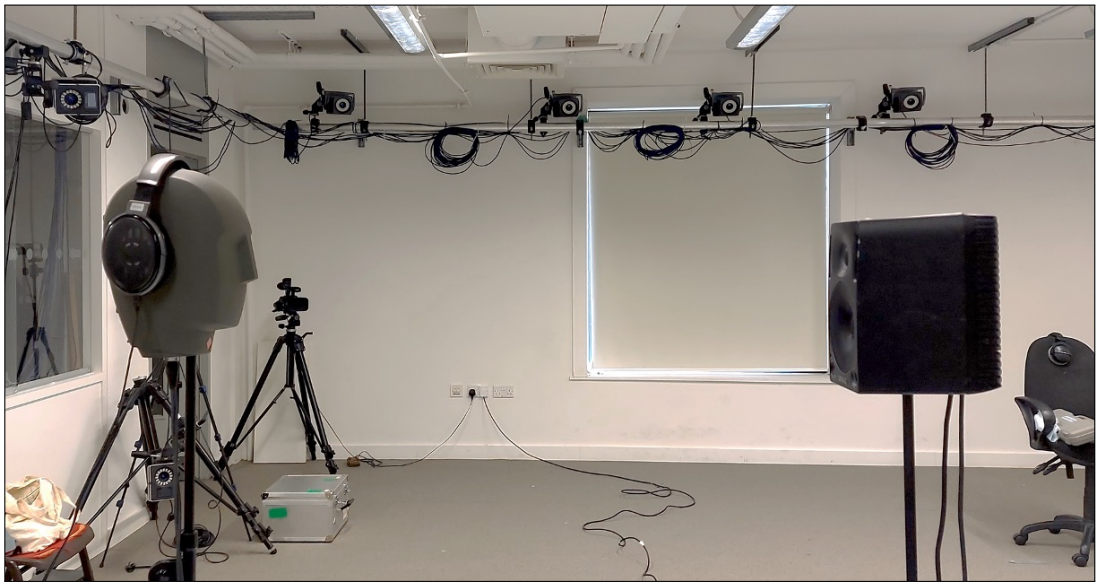


Figure 4.7: Setup to measure the amplitude loss caused by the noise-cancelling headphones.

Most commonly, the first arriving reflection comes from the floor as it is the closest surface to the performer. However, similar to Kearney et al. [34], we did not consider this reflection due to its negligible significance, attributed to the shadowing effect of the singer's torso and the voice's directivity. Nonetheless, we acknowledge that different scenarios may require estimating the ITDG based on the floor reflection.

Input Signal Calibration Accurately reproducing the auditory experience in IVEs requires matching the loudness of the auralised signal reproduced over headphones to the perceived amplitude of the physical acoustic space. Thus, calibrating the input level ensures that the convolved signal results in a Direct-to-Reverberant Ratio (DRR) consistent with the acoustic response of the physical space. To achieve this, white noise was played through the speaker at 85 dBA, measured at a distance of 1 m, matching the output level and setup used during the SIRs

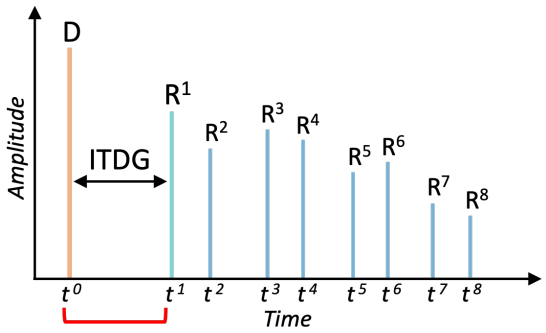


Figure 4.8: Trimmed section of the SIR showing the direct signal and the Initial Time Delay Gap (ITDG), up to the first reflection, highlighted in red.

production. With the microphone positioned 5 cm from the speaker, corresponding to its distance from the user's mouth during auralisation, the signal level was calibrated within Reaper to match the RMS amplitude of a recorded sine sweep from the SIR production. This approach allowed the convolved signal to be reproduced at the intended direct-to-reverberant balance between the direct and auralised signals reproduced over headphones.

4.2.2 Real-time Signal Auralisation

Following the system setup and calibration for direct monitoring and auralisation, we now describe the real-time interpolation achieved by adjusting the output levels of the convolved signal with different SIRs across two independent stages: Directional and Translational interpolation.

Directional Interpolation Interpolation between the four directional SIRs (front, right, back, left) of each position was performed using 4-way Constant Power Panning (CPP), which adjusted the gain of each corresponding audio track in Reaper based on the HMD orientation relative to the IVE (see Figure 4.9).

The VR experience was rendered using the Unity game engine [218]. A C# script was developed to manage the interpolation, continuously tracking the HMD's rotation each frame, wrapping the values to a 0-360° range to avoid negative angles, and applying 4-way constant-power panning based on the yaw angle. Afterwards, the resulting four gain values were transmitted to Reaper via the OSC protocol, controlling the directional tracks. The gain G_d for each direction d was calculated using cosine-based constant power panning. Since power is proportional to the squared amplitude, and $\cos^2(\theta) + \sin^2(\theta) = 1$, assigning gains as $L(\theta) = \cos(\theta)$ and $R(\theta) = \sin(\theta)$ provides constant power panning [219].

Translational Interpolation Previous work has explored different interpolation strategies for 6DoF audio rendering [207, 220, 221]. In the translational interpolation stage of this pipeline, we opted to implement Inverse Distance Weighting (IDW) due to its flexibility with node distribution, perceptual robustness, and computational efficiency. IDW computes a weighted average of surrounding node values, where each weight is inversely proportional to the node's distance from the listener [220]. The gain interpolation was defined as:

$$w_i = \frac{\frac{1}{d_i}}{\sum_{i=1}^{M} \frac{1}{d_i}}$$

where d_i is the distance from the performer to the ith impulse response position, and M is the number of contributing nodes.

A second C# script was developed to perform IDW-based interpolation between positions,

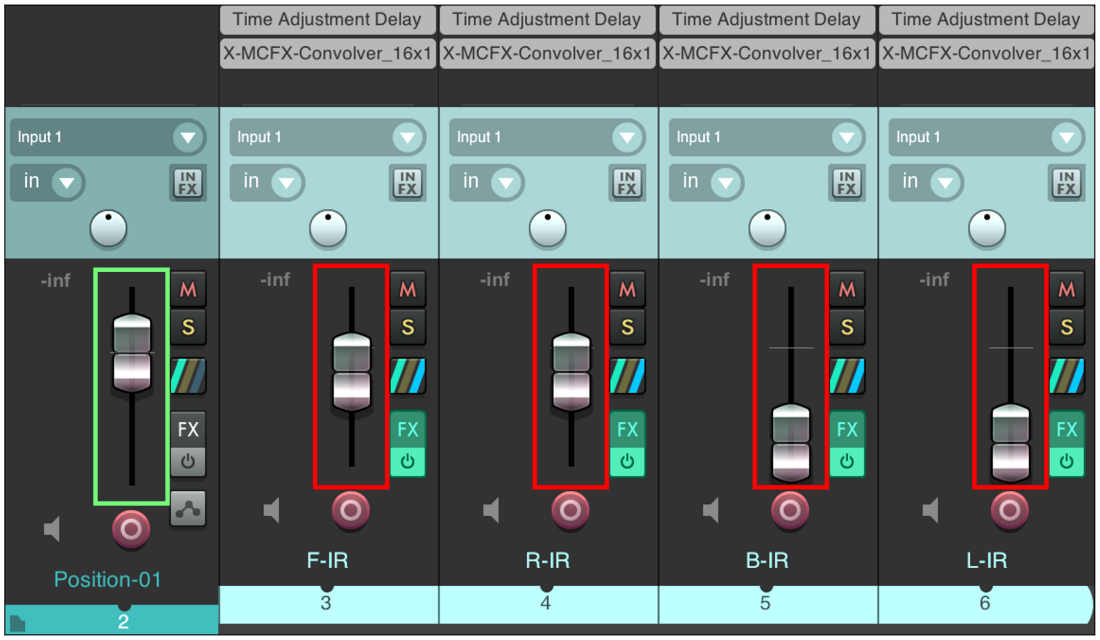


Figure 4.9: Reaper tracks for a single position, with the gain faders of directional tracks in red, and the folder track gain fader in green.

using the HMD's location within the IVE. The script dynamically tracked the user's position and computed normalised weights based on the inverse distance to active IR nodes. These values were then transmitted to Reaper via OSC to control the amplitude of the corresponding folder tracks (see Figure 4.9). To improve functionality and system performance, the script also limited the number of active nodes and disabled audio plugins in the Reaper project when a node was inactive, using hysteresis thresholds to prevent unintended on/off toggling.

4.3 Insights

Several practical considerations have been identified throughout the development and implementation of VOREA.

Time constraints and limited venue access are significant considerations when conducting acoustic recordings, as many venues operate under tight schedules and impose restricted access. Therefore, minimising setup time is essential to ensure the efficient use of available time. To achieve this, a recording grid can be pre-traced onto an acoustically transparent fabric in advance, allowing for efficient setup and requiring only the placement of the recording area on stage.

In addition to setup efficiency, the placement of the recording grid is a key consideration, as the perceived acoustic response within an acoustic space varies significantly depending on the performer's location (see Figure 4.10). For example, on a performance stage, the centre tends to provide stronger early reflections from overhead canopies and walls, improving clarity and vocal support acoustic parameters [222, 223]. In contrast, positions toward the front of the stage present

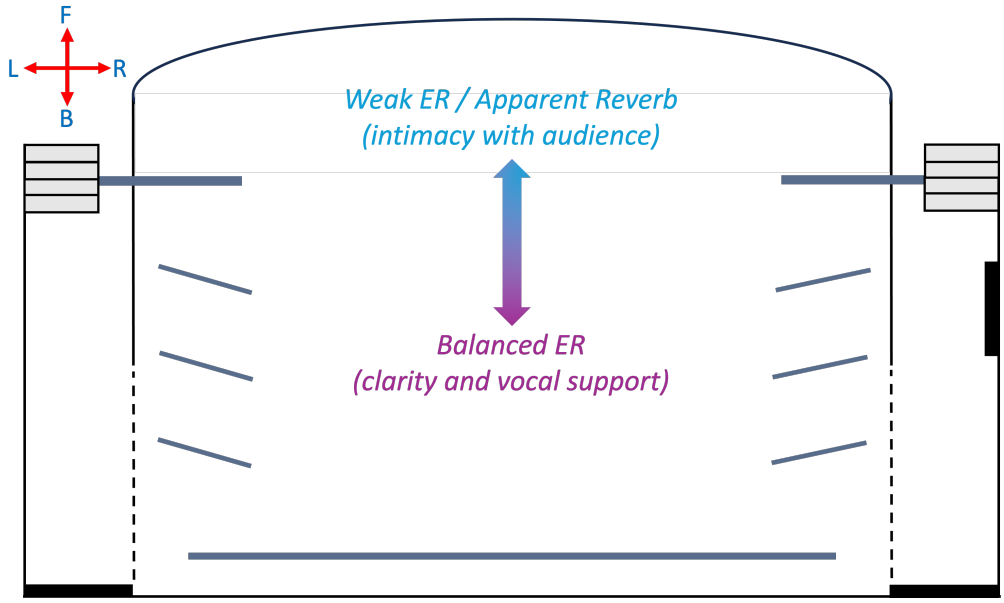


Figure 4.10: Variation of acoustic characteristics across the stage commonly observed in performance venues.

lower early reflection energy, while diffuse reverberation becomes more apparent, reducing vocal feedback but enhancing projection toward the audience [224]. Given the translational DoF allowed by the system, a grid placement that captures acoustic variation makes the most practical sense. However, the optimal placement may vary depending on the required acoustic features and intended objectives.

Expanding the spatial scope of VOREA presents practical challenges, as it is inherently constrained by the number of acoustic measurements required to capture the defined exploration area, making measurements in larger spaces potentially unfeasible. Alternative strategies may accommodate larger spaces or those with different spatial and architectural characteristics. For instance, designing the recording grid with greater spacing or irregular distribution could reduce the total number of required nodes or allow for a larger exploration area, as reasonable increments might not be perceptually salient. This approach is particularly suited to IDW interpolation, which performs well with sparse and non-uniform grids [220]. Furthermore, incorporating satellite positions or grids recorded at key spatial locations could further extend the exploration capabilities of the system, enabling direct investigation through teleportation-based locomotion within the IVE. This method could enable the exploration of acoustic variation derived from specific architectural features or spatial attributes, such as ceiling height and geometry variations.

Lastly, higher-order Ambisonic SIRs provide greater rendering accuracy but come at the cost of significantly increased computational load due to the number of impulse response channels required [220]. Although 3OA offers a balanced compromise between perceptual fidelity and computational efficiency [225], depending on system resources, alternative formats such as

Table 4.1: Comparison of spatial formats with corresponding channel counts and convolution requirements.

Parameter	3OA	FOA	Binaural
Channels	16	4	2
Convolution per position	64	16	8
Convolution per grid (20 positions)	1280	320	160


(ZM-1)


(NT-SF1)


(KU100)

first-order Ambisonics or binaural SIRs may provide a suitable trade-off, delivering plausible results at a fraction of the computational cost required for convolution (see Table 4.1).

4.4 Discussion

While this work did not include a formal user evaluation, VOREA was informally tested by audio researchers using SIRs captured at the Abbey Theatre (employed in the study presented in Chapter 6), and they reported that the system delivered a convincing spatial impression and that the interpolated auralisation felt responsive and perceptually plausible for real-time vocal feedback in IVEs. With directional interpolation, the system delivers natural and seamless auralisation, while the addition of translational rendering enables spatial continuity, granting users greater agency to explore acoustic virtual spaces in a more immersive and dynamic way. Moreover it has been successfully incorporated in perceptual evaluations on audio-visual perception and rendering quality, enabling the integration of real-time first-person auralisation into multimodal VR research, as detailed in Chapter 5. Feedback from users with professional singing experience was positive, commenting on the enjoyment of the immersive singing experience and recognising its potential value for musical applications such as rehearsing and performing in virtual venues, an application further explored in Chapter 6.

The proposed system, VOREA, supports real-time auralisation through convolution by leveraging the ITDG to account for inherent system latency. Although latency constrains the minimum distance to the nearest reflective surface, it still enables accurate acoustic reproduction of realistic environments. In our case, system latency of approximately 13.1 ms remained below the ITDG threshold for performance stages, achieving perceptually coherent results. In a room with a more standard ceiling height, the ITDG was slightly shorter than the system's latency; however, this difference remained below the perceptual just noticeable difference (JND) for

ITDG [217, 226], which is unlikely to affect the auditory experience. Latency can be further improved through hardware optimisation, using audio interfaces with lower latency performance and faster data transfer protocols than the Focusrite Scarlett 2i2 [227] used in this case study.

While audio latency was a primary concern in the system design, motion-to-data latency also needed to be addressed. This refers to the delay between the tracking data from the HMD and the corresponding audio update within the rendering software. In our setup, directional and translational motion-to-data latency with a tethered Meta Quest 3 [228] was approximately 43 ms, remaining below the 60 ms threshold identified as perceptually acceptable in previous research [229], thus suitable for immersive vocal auralisation.

Some areas for optimisation were identified during implementation, highlighting opportunities for signal interpolation and performance enhancement. CPP and IDW were suitable for achieving our research aims due to their flexibility and ability to perform gain-based interpolation without frequency-dependent features. However, alternative interpolation methods, such as dynamic SIR selection across transitions [230] or object-based rendering approaches [221], may offer better perceptual realism and computational performance in more complex or dynamic environments. In addition, the system used directional SIRs for rotational interpolation, which significantly increased the number of impulse responses required at each grid position. This number could be reduced by producing 360° SIRs using an omnidirectional speaker instead. During auralisation, the system could implement a directivity shaper to approximate the natural directivity of the human voice, oriented in real-time based on the user's head direction.

Beyond the technical design and the benefits of richer, congruent multisensory feedback, VOREA demonstrates clear relevance for practical scenarios and applications within XR that leverage audio-visual integration. Among these, the development of audio-visual digital twins, which are high-fidelity virtual replicas of physical spaces [231], is a significant area of focus due to its broad range of potential applications. For instance, acoustical heritage has gained recognition as an essential component of cultural identity and memory, promoting interest in preserving the unique acoustic qualities of historical buildings and concert halls [232, 233]. Emerging digital tools and methods, such as VR and AR, enable the integration of coherent, immersive auditory scenes with the visual presentation of historical sites [233]. Therefore, the production of audio-visual digital twins, particularly of performance venues that allow for first-person vocal interaction, contributes to the preservation of these acoustic sites, capturing how performers perceived the stage's acoustic response during performances. This is explored in Chapter 6, which focuses on the evaluation of a high-fidelity virtual theatre produced using Gaussian Splatting and VOREA.

In addition to cultural heritage, audio-visual IVEs could also enable real-time vocal auralisation for actors and singers, allowing for live performances within the metaverse. Since the rendering of acoustic features enhances both the sense of presence and the quality of vocal performance,

providing performers with a coherent acoustic representation can potentially support their performance and overall experience. Acoustic parameters, such as room supportiveness and noiselessness, have been shown to influence vocal performance and preference [234]. Yet, research has found that music practice rooms often lack adequate acoustic conditions, forcing musicians to adapt their playing according to their present practice room [235]. In this context, the production of virtual audio-visual rehearsal rooms offers the potential for more acoustically appropriate rehearsal conditions, closely aligned with those of performance venues. Furthermore, similar to the effects highlighted in public speaking research, such environments may help performers reduce anxiety and better adapt to the acoustics of the actual venue.

4.5 Conclusion

Overall, VOREA demonstrates strong potential for real-time vocal auralisation in IVEs, supporting both directional and translational interpolation. Its relevance and main contribution lie in the design and implementation of an auralisation pipeline that provides users with the agency to explore acoustic spaces with spatial freedom, enabling them to experience the acoustic characteristics of a defined area within IVEs using their voice and to become fully immersed in the environment.

This chapter discussed in depth the design and implementation of the pipeline, from the creation and capture of adjacent SIRs from a *performer perspective* to the interpolation approach using gain-based methods. Moreover, it presented practical insights and considerations identified during the development process, including grid placement for acoustic variation, methods to expand the spatial scope, and alternative rendering formats. The implementation also revealed certain shortcomings, highlighting opportunities for refinement in hardware optimisation and computational efficiency, while opening pathways for practical deployment and future research. Future work will focus on alternative translational interpolation methods, formal user evaluation, and broader integration into interactive and performance-based VR applications.

Although the main focus and research questions of this thesis concern audio-visual perception and perceptual illusions in IVEs, the development of VOREA arises from the need for a real-time vocal auralisation system to conduct perceptual studies on audio-visual perception and rendering fidelity and address the relevant research questions. By allowing users to hear a coherent acoustic representation of a visual space, VOREA supports the creation of more plausible audio-visual environments, with the potential to promote psychophysical effects such as the sense of presence in VR and MR. These aspects are further examined in the next chapter.

Chapter 5

Vocal Auralisation in XR & Its Influence on Perceived Presence & Audio-Visual Quality

Based on **"First-Person Vocal Auralisation in XR and Its Influence on Perceived Presence and Audio-Visual Quality"**
Article published at IEEE ISMAR 2025

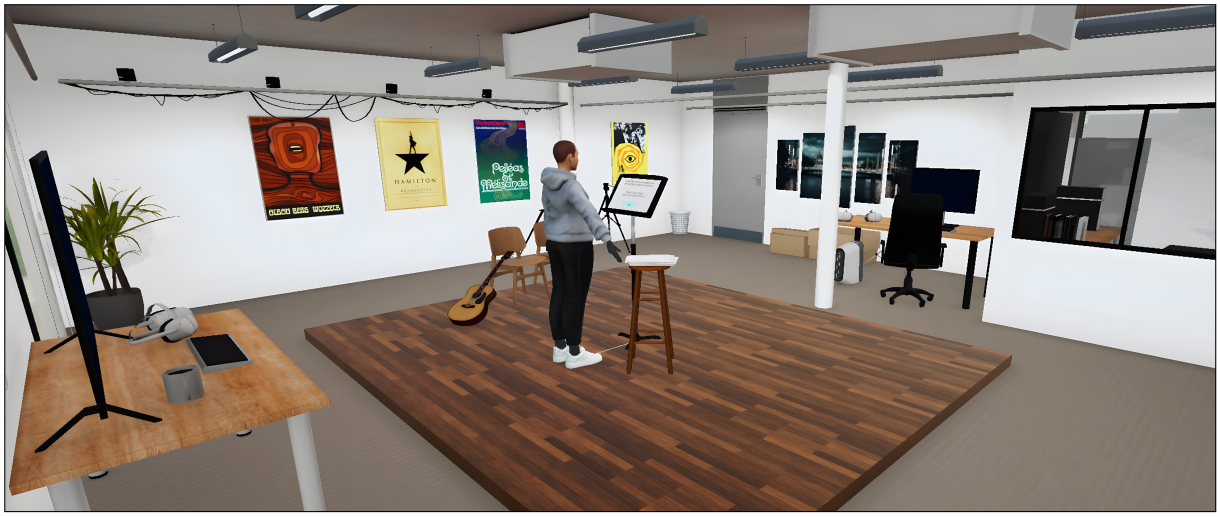


Figure 5.1: Virtual room used for the study, with a participant singing while immersed in the experience.

We perceive the world around us through our bodies, and it is via our senses that we construct a mental representation of the environment. However, perception is inherently multimodal, relying on the integration of different sensory stimuli to decode complex environmental information [84]. Chapter 3 examined the influence of continuous auditory feedback on enhancing the SoE towards a virtual avatar, presenting an initial exploration of self-produced sounds as part of multimodal stimuli to elicit psychological effects. Chapter 4 then focused on the development of a real-time vocal auralisation pipeline to enable the integration of audio-visual feedback, addressing the limitations of existing tools and spatialisers in terms of latency and independent signal control within IVEs. This chapter builds upon the work described above by examining how real-time first-person vocal auralisation influences the perceived sense of presence and audio-visual quality across different rendering resolutions and visual representations in both VR and MR.

Prior studies have investigated sound perception, spatialisation, and auralisation in IVEs [68, 37, 24, 158, 5, 8, 168, 236]. However, most studies have focused on auditory feedback produced by external sound sources within the environment, rather than on sounds produced by the user, particularly their own voice, as a means to explore the environment acoustically. Due to the degree of immersion and dynamic possibilities offered by VR systems, such as head tracking, it is feasible to provide real-time sound auralisation that responds to the user's head direction and position within the IVE. However, doing so with the user's own voice is particularly challenging due to the synchronisation required between the direct sound and its processed auralisation, which can be affected by system latency and computational limitations.

Visual feedback has been shown to influence the room-acoustic experience of auralisations, highlighting the importance of testing how one modality can affect the other in cross-modal perception [237, 165]. Although some studies have investigated the effects of real-time vocal auralisation in both audio-only and VR contexts [69, 70, 72, 238, 9], work on vocal perception has mainly focused on the perception and adaptation of one's own voice, measuring vocal comfort and acoustic parameters such as mean fundamental frequency, Sound Pressure Level (SPL), and spectral moments [28, 29, 30].

However, the audio-visual perception of real-time vocal auralisation in IVEs, particularly its effect on the sense of presence (the sense of being in or interacting with the virtual environment [40]) and on perceived audio-visual quality in VR and MR contexts, remains largely unexplored. This is particularly relevant given the growing prevalence of MR experiences enabled by the passthrough feature of current HMDs, which support both VR and MR by varying the level of digital content rendered. To address this, we compared three levels of voice auralisation and visual quality in XR. This allowed us to assess the perceived quality and integration of audio-visual feedback, as well as its influence on the sense of presence. In doing so, we evaluated the relevance of real-time vocal auralisation as a component for enhancing presence. We then extended our investigation to examine how first-person voice auralisation is perceived across

different visual representations, specifically between VR and MR, and whether it affects perceived presence and audio-visual integration.

5.1 Hypotheses

The objective of the first experiment was to investigate how the quality of acoustic and visual rendering influences the sense of presence and the perceived multimodal integration of audio-visual feedback in VR. We believe that providing participants with a coherent acoustic response will create more plausible and realistic environments, allow them to become more immersed in the virtual space, and thus promote a higher sense of presence. This leads to our main hypothesis:

H1 Perceived presence and audio-visual integration will be greater under conditions where audio feedback is present.

Sound auralisation using measured SIRs provides a more accurate acoustic representation of reverberant spaces, as they capture more detailed acoustic features compared to synthetic modelling, which relies on simplified geometric approximations [20]. However, their production is more challenging, as it requires more time and effort and is constrained to existing physical spaces. Therefore, we aim to determine whether the perceived quality of measured auralisation yields higher ratings than synthesised auralisation, justifying the need for producing high-fidelity captured SIRs. This rationale forms the basis of our second hypothesis:

H2 Measured Auralisation will be rated higher than Synthesised Auralisation and No Auralisation across measures.

A previous study has shown that both visual and spatial audio fidelity are important for achieving a strong sense of immersion in VR, with visual resolution having the larger effect [8]. The same study suggested that music and speech produced from a sound source within the virtual environment can influence perceived visual feedback in audio-visual setups. However, these findings were based on sounds perceived from a third-person perspective, where the sound source is external to the listener (see Figure 4.2). In contrast, we aim to assess this from a first-person *performer perspective*, which informs our third and fourth hypotheses:

H3 Audio Naturalness and Reverberance scores will be greater in full-resolution visual conditions.

H4 Perceived visual quality will be rated higher in conditions where sound auralisation is present compared to when it is absent.

In Experiment 2, we build upon the findings of Experiment 1 to further examine how sound auralisation contributes to the sense of presence and perceived audio-visual quality in MR.

Therefore, we aim to assess how MR compares to VR, given that it provides a more accurate representation of the physical environment. We hypothesise that:

H5 Perceived presence will be greater under conditions where vocal auralisation is present.

H6 Perceived visual quality and audio-visual integration will be rated higher in MR than in VR.

H7 Audio Naturalness and Reverberance will be rated higher in MR than in VR.

H8 Measured Auralisation will be rated higher than No Auralisation across measures within MR.

The hypotheses presented here are derived from the research questions of this thesis and are informed by similar works with comparable research objectives and methodologies. Hypotheses **H1** and **H5** arise from **RQ1**, which seeks to determine whether the perception of one's own voice within an IVE can affect psychological responses by measuring the sense of presence, an essential aspect of immersive experiences. We believe that providing participants with a coherent acoustic response will create more plausible and realistic environments, allow them to become more immersed in the virtual space, and thus promote a higher sense of presence. In addition, hypotheses **H2** to **H4** of Experiment 1 and **H6** to **H8** of Experiment 2 are related to **RQ2**, and they are geared towards exploring and understanding how rendering fidelity affects the perception of visual and auditory feedback quality and their multimodal integration.

5.2 Experimental Setup

The experiments were conducted using a tethered Meta Quest 3 HMD, connected via Meta Link Cable to a PC equipped with an Intel i9-11900KF 3.50GHz CPU, 64GB RAM, and an Nvidia GeForce RTX 3080 Ti graphics card. Audio playback and capture were handled using a pair of Bose QuietComfort 25 noise-cancelling over-ear headphones and a DPA SCO60F00-H omnidirectional microphone connected to a Scarlett 2i2 mk4 audio interface. The microphone was fixed to the HMD using a wire rig, ensuring consistent placement throughout the study (see Figure 5.2), forming an audio-enhanced HMD setup for immersive vocal interaction.

Virtual Environment A small-to-medium-sized laboratory space (8.85 × 10.16 × 2.55 m) was selected as the physical location, as it provided a clear, open area suitable for experimentation. The room featured plain plaster walls with minimal furnishings, a carpeted floor, and ventilation elements distributed across the ceiling, resulting in a moderately reflective acoustic environment. The acoustic response of the space (mid-frequency average of the 500 Hz and 1 kHz bands) presents a moderate reverberation time $T_{30} \approx 723$ ms, with an $EDT \approx 702$ ms, suggesting relatively uniform decay behaviour. It exhibits clarity indices of $C_{50} \approx 3.1$ dB and $C_{80} \approx 6.4$ dB,



Figure 5.2: Audio-enhanced HMD setup used in the study, with a lapel microphone attached and positioned close to the participant's mouth.

indicating reasonably strong early reflection support. The lateral fraction coefficient $JLF_c \approx 0.40$ suggests a moderate spatial impression.

During the study, participants were immersed in a virtual space at a fixed location, with their movement restricted to a stationary position. The virtual environment corresponded to a virtual representation of the physical space. The visual feedback, rendered in Unity, was a faithful replica of the room, accurately matching its dimensions, layout, and objects. For auditory feedback, the acoustic response was obtained by capturing an SIR at the defined location within the space. The audio input was processed in the Reaper DAW, where participants' voices were auralised in real time. To provide coherent audio-visual stimuli, VOREA transferred the gain values from the interpolation according to the HMD's position and rotation to Reaper, via OSC, ensuring that the auralised output dynamically reflected users' movements (see Figure 5.3).

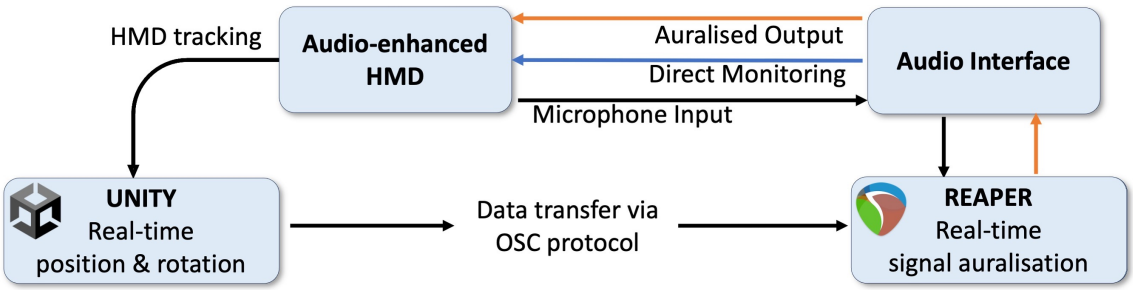


Figure 5.3: Real-time vocal auralisation system.

Table 5.1: Audio & Visual rendering assessment questions

ID	Question
ANAT	How realistic and natural the sound is; how well the sound conforms to what you would expect the sound in the room to be like.
AREV	How the room's reverberation creates a sense of space and atmosphere, contributing to an engaging acoustic environment.
VRES	The sharpness and detail of visual elements; how well-defined objects appear without blurring, distortion, or pixelation.
AVI	The quality level between sound and visuals, resulting in a plausible and coherent environment.

5.3 Measures

Presence Questionnaire The sense of presence and realism were assessed using the Igroup Presence Questionnaire (IPQ). It comprises 14 items rated on a 7-point Likert scale (−3 and +3), designed to measure overall presence and three subscales: Spatial Presence, Realism, and Involvement. The IPQ was chosen for its brevity, which allowed us to employ the full validated questionnaire while maintaining a reasonable overall length, as well as for its prior use in audio-related VR studies [239, 5].

Audio-Visual Questionnaire Four items related to audio naturalness and reverberance, visual quality and audio-visual integration were included to assess the specific qualities of the stimuli provided to participants (See Table 5.1). We designed these questions following an Absolute Category Rating (ACR) test method from Recommendation ITU-T P.910 [240]. The items used a 7-point rating scale to maintain consistency with the IPQ. The final questionnaire was composed of 18 items divided into eight measures: Audio Naturalness (ANAT), Audio Reverberance (AREV), Visual Resolution (VRES), Audio-visual Integration (AVI), Presence (PRES), Spatial Presence (SP), Realism (REAL) and Involvement (INV).

5.4 Experiment 1

The experimental design employed a two-factorial within-subjects design to control for individual differences across conditions. Furthermore, the order of conditions was randomised for each participant to minimise ordering effects. Two independent variables, audio and visual, were examined in this experiment, each with three levels of rendering quality. For auditory feedback, the levels included Measured Auralisation (MA) captured from a physical space, Synthesised Auralisation (SA) created using VST tools, and No Auralisation (NA). For visual feedback, the levels included Full-Resolution Visuals (FV) rendered at 100% resolution, Low-Resolution Visuals (LV) at 22%, and No Visuals (NV).

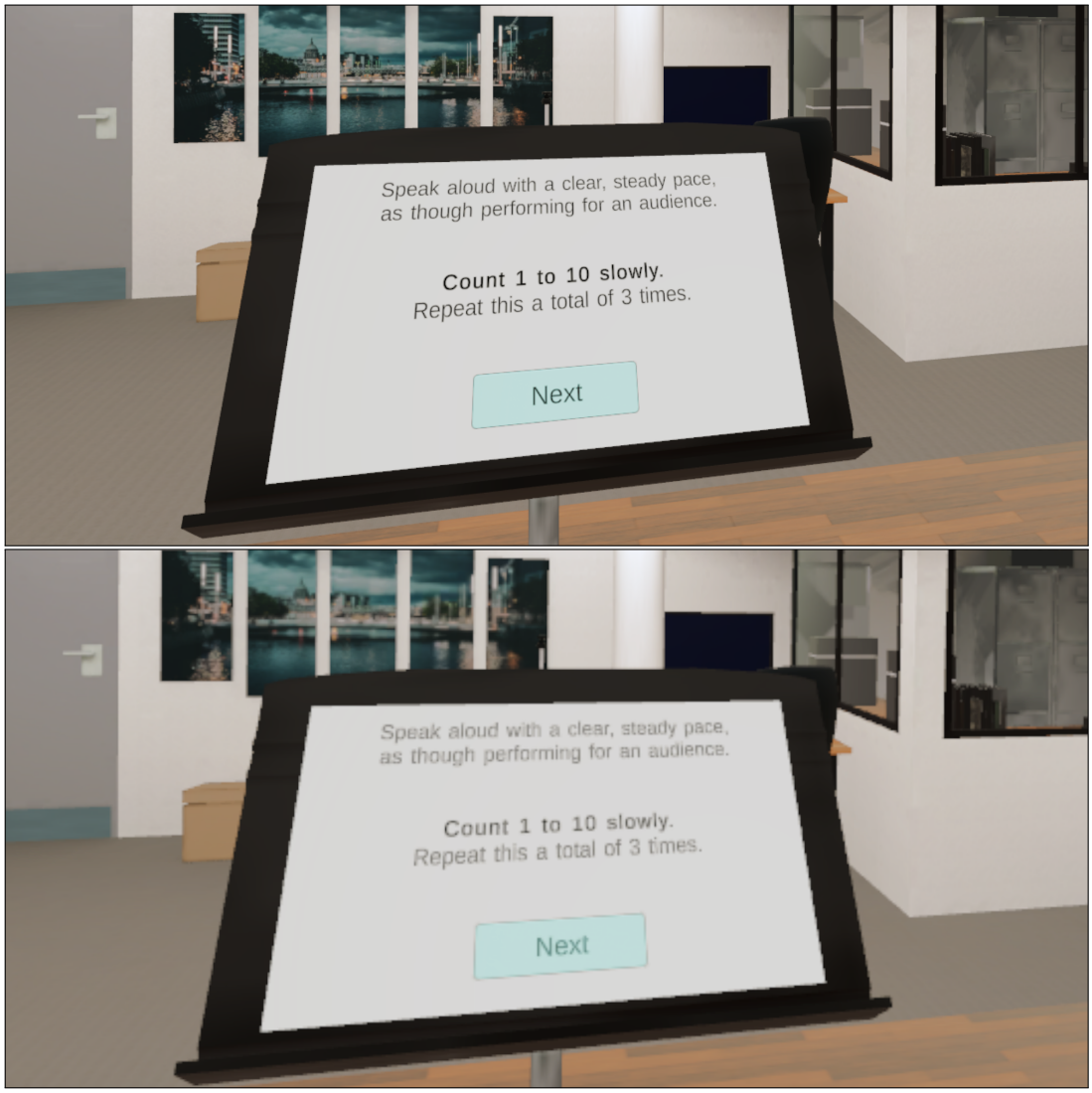


Figure 5.4: View of the environment and Tasks UI in the Full Visuals (FV) condition (top) and in the Low Visual (LV) condition (bottom).

5.4.1 Stimuli

Visual Feedback The virtual environment was built using Unity Game Engine (version 2022.13.91f1). The virtual room was modelled using ProBuilder (version 5.2.2) in combination with 3D models sourced from the Unity Asset Store and CG Trader to closely match the physical room. However, additional objects, such as musical instruments and posters, were added to the scene to provide participants with additional elements to assess the visual rendering quality, as well as to provide further size and depth cues [241] (see Figure 5.1).

The experiment presented three visual rendering conditions. Full Visuals (FV) were rendered at the highest possible resolution on the HMD, at 2064 × 2208 pixels per eye, approximately 4.6 megapixels (see Figure 5.4, top). Low Visuals (LV) at 968 × 1036 pixels per eye, approximately

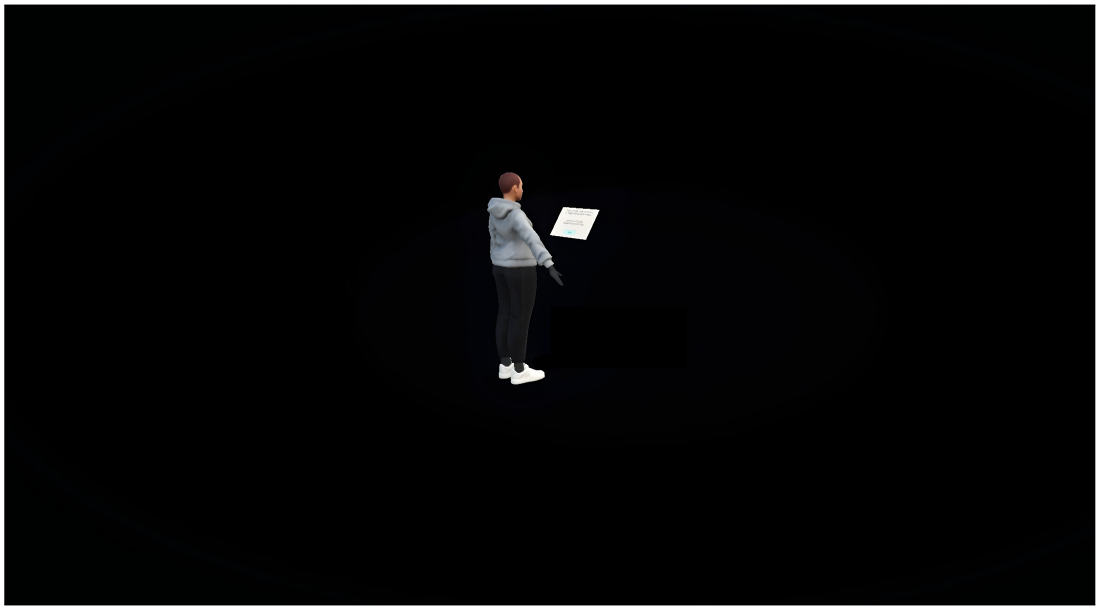


Figure 5.5: Black space used in the NV condition, limiting participants' visual input to the instructions UI and their avatar.

1.0 megapixels, representing 22% of the full-resolution total pixel count (see Figure 5.4, bottom). For the No Visuals (NV) condition, the virtual room was disabled, leaving participants immersed in a black space only with the task instructions UI and virtual avatar visible (see Figure 5.5). Additionally, participants were immersed in an empty grey scene while answering a questionnaire, which also served as a transition between experimental blocks.

Participants were embodied in a human avatar with gender-neutral clothing and hair designed and sourced from Ready Player Me. The avatar's size and position were estimated using the Meta Movement SDK for Unity, which uses body tracking to match the height and movement of the virtual avatar to the participant's real-world movements. The avatar's size and position were estimated using the Meta Movement SDK for Unity, which leverages body tracking to align the virtual avatar's movement and height with the participant's real-world posture and proportions.

Audio Feedback The audio auralisation was handled using VOREA, which auralised participants' vocal input in real time using 3OA SIRs to provide accurate acoustic representations of physical spaces. SIRs have been successfully used in vocal auralisation studies that capture them from an egocentric perspective [34, 35, 28, 206]. In this similarly egocentric configuration, which we refer to as the *performer perspective*, the source and receiver are co-located to simulate the position of the mouth and ears of a person (see Figure 4.2), which is essential for representing auditory experiences where the user explores the acoustic space using their own voice.

The study considered two types of SIRs, measured and geometrically synthesised, which are the most common techniques and offer different trade-offs in realism, acoustic characteristics,

and computational efficiency. Both sets of SIRs were not intended to be directly equivalent due to fundamental production differences and acoustic features of each method. However, synthesised SIRs were generated using parameters derived from the measured SIRs to allow an objective comparison and ensure they resembled the physical room's acoustic properties.

Measured SIRs (MA) were obtained from acoustic measurements of the physical room on which the virtual environment was based. They were captured at a single predefined position near the centre of the space, which was also where participants stood during the experiment. Following the method described in Section 4.1, measurements were taken for each direction (front, right, back, and left).

For the Synthesised SIRs (SA) condition, SIRs were created in Reaper (v7.16) using the IEM Plug-in Suite (v1.14.1) in 3OA format to match the rendering format of the measured SIRs. A sine sweep was first processed with IEM's DirectivityShaper plugin to simulate the directivity pattern of the physical loudspeaker. Then, the sine-sweep was positioned in a 3D virtual space using the IEM RoomEncoder plugin to render early reflections. The dimensions of the room were configured to match the real space (width, length, and height). Finally, the source and listener positions were set to coincide with the physical speaker and microphone setup. To match the acoustic characteristics of the physical room, additional spatial and temporal diffusion was added using the FDNReverb plugin. Afterwards, four directional recordings (front, right, back, and left) were captured by adjusting the relative position and orientation of the source and listener within the virtual space. Table 5.2 presents a comparison of measured and synthesised SIR acoustic parameters. These values are reported solely for comparative purposes and to estimate potential perceptual differences. They should not be interpreted as valid ISO 3382 acoustic parameters [242, 243], as the source–receiver configuration does not follow the conditions defined by the standard and the analysed SIRs were trimmed up to the first arriving reflection.

Both measured and synthesised recordings were deconvolved using Voxengo Deconvolver (v1.9.3), resulting in four directional SIRs per condition. As described in Section 4.2, the direct signal (i.e., the participant's own voice) is delivered via the headphones direct monitoring and does not require processing, so the SIRs were trimmed up to the first reflection.

5.4.2 Participants

An a priori power analysis was performed using G*Power [200, 201] to determine the experiment sample size. The effect size was set to 0.17 with an alpha error probability of .05 and a power of 0.8 to accommodate a small-medium effect size. The statistical analysis was conducted using within-group repeated measures ANOVA; therefore, the number of groups was set to 1 and the measurements to 9 to account for experimental conditions. Our analysis indicated that a sample size of thirty participants was required for the study.

Prior to the experiment, participants completed a demographic questionnaire. Thirty

Table 5.2: Comparison between a Measured & Synthesised SIR acoustic parameters.

Parameter	Measured SIR	Synthesised SIR	JND	Perceptibility
C50	-0.63 dB	-2.82 dB	2.19	Perceptible
C80	3.05 dB	2.25 dB	0.79	Not perceptible
EDT	880 ms	791 ms	2.02	Perceptible
T30	814 ms	810 ms	0.10	Not perceptible
JLFc	0.22	0.24	0.62	Not perceptible

Note: Values represent the mid-frequency average of the 500 Hz and 1 kHz bands [242].

participants took part in the experiment (14 males, 15 females, and 1 genderfluid) with ages ranging from 21 to 46 years (mean ± SD: 30 ± 6.44). Among participants, 12 reported audio training, 9 vocal training, and 18 musical training. Furthermore, they reported on their VR experience, where 6 had never used VR, 12 had played one or two VR demos, 4 had used VR several times, 4 owned a VR headset, and 4 were VR developers/researchers.

5.4.3 Procedure

Upon arrival, participants were given an information sheet, read and signed an informed consent form, and completed a demographic questionnaire. Participants were given a verbal explanation of the vocal tasks they had to perform during each block's assessment phase to explore the virtual space acoustically. They were instructed to speak aloud with a clear, steady pace, as if performing for an audience, to ensure the vocal level was adequate to perceive the sound auralisation.

Based on previous research[29, 30], the vocal tasks consisted of counting from 1 to 10 slowly three times, sustaining the vowel */a/* phonetically for 3 seconds three times, reciting the CAPE-V phrase "The blue spot is on the key again" 3 times, reciting the phrase "When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow" and singing "Happy Birthday" (see Figure 5.6).

While in VR and wearing noise-cancelling headphones, participants stood at a fixed position in the room. They were first immersed in a reference scene featuring the highest visual and audio quality (FV-MA) and were instructed to observe the visual elements and count aloud from 1 to 10 to assess the audio quality. Then, participants completed nine blocks, one for each experimental condition. Each block began with an assessment phase, during which participants were immersed in the virtual room for a minimum of 2 minutes and explored the room visually and acoustically by completing vocal tasks. Afterwards, they were encouraged to examine further the audio-visual qualities of the space using their voice and by observing virtual objects within the virtual space. Following this, participants completed the questionnaire designed for this study, rating the audio-visual resolution and presence before proceeding to the next block. Since the total duration of VR immersion was approximately 45 minutes, a five-minute break was provided after the fifth block to help mitigate fatigue.



Figure 5.6: Music stand displaying the UI of vocal tasks performed by participants in the IVE.

5.4.4 Results

A two-way within-subjects analysis of variance (ANOVA) was conducted in RStudio to examine the impact of audio and visual feedback. Data normality was assessed using the Shapiro-Wilk test; however, since most data violated the normality assumption, an Aligned Rank Transformation (ART) model [203] was applied. Post-hoc pairwise contrasts ($\alpha = .05$) were then computed using Aligned Rank Transformation Contrasts (ART-C) to evaluate significant differences. Bonferroni correction was applied to control for errors across multiple comparisons.

No significant interaction effects were found between **Audio** and **Visual** factors across any measures. A summary of the main effects is provided in Table 5.3, and relevant findings are discussed below.

Presence and Audio-Visual Integration As expected, the presence of audio feedback led to significantly higher scores in **Presence** and **Audio-Visual Integration** (see Figures 5.7a and 5.7b). Post-hoc comparisons revealed that for **Presence**, both MA and SA scored significantly higher than NA ($p < .001$, $p = .001$), with no difference between MA and SA ($p = 1.000$). For the **Visual** factor, FV was significantly higher than LV and NV ($p < .001$, $p < .001$), but LV and NV did not differ ($p = 1.000$). In **Audio-Visual Integration**, similar effects were found: for the **Audio** factor, MA and SA > NA ($p < .001$, $p = .001$); no difference between MA and SA ($p = .088$). For Visuals, FV > LV, NV ($p < .001$, $p < .001$); LV = NV ($p = 1.000$). These findings support **H1**, demonstrating that appropriate real-time vocal auralisation, combined with coherent visual

Table 5.3: Experiment 1 Aligned Rank Transform (ART) ANOVA results with post-hoc comparisons.

Measure	Factor	F-statistics	p-value	η^2	Post-hoc
Audio Naturalness	Audio	$F^*(2, 232) = 56.89$	$p^* < .001$	0.329	MA > SA > NA
	Visual	$F^*(2, 232) = 6.98$	$p^* = .001$	0.057	FV > LV, NV
Audio Reverberance	Audio	$F^*(2, 232) = 29.83$	$p^* < .001$	0.205	MA > SA > NA
	Visual	$F^*(2, 232) = 5.64$	$p^* = .004$	0.046	FV > LV, NV
Visual Resolution	Audio	$F^*(2, 232) = 32.58$	$p^* < .001$	0.220	MA, SA > NA
	Visual	$F^*(2, 232) = 41.76$	$p^* < .001$	0.264	FV > LV, NV
Audio-Visual Integration	Audio	$F^*(2, 232) = 21.69$	$p^* < .001$	0.158	MA, SA > NA
	Visual	$F^*(2, 232) = 19.92$	$p^* < .001$	0.147	FV > LV, NV
Presence	Audio	$F^*(2, 232) = 10.95$	$p^* < .001$	0.086	MA, SA > NA
	Visual	$F^*(2, 232) = 15.72$	$p^* < .001$	0.119	FV > LV, NV
Spatial Presence	Audio	$F^*(2, 232) = 8.14$	$p^* < .001$	0.066	MA, SA > NA
	Visual	$F^*(2, 232) = 23.94$	$p^* < .001$	0.171	FV > LV, NV
Realism	Audio	$F^*(2, 232) = 4.33$	$p^* = .014$	0.035	MA, SA > NA
	Visual	$F^*(2, 232) = 38.77$	$p^* < .001$	0.250	FV > LV, NV
Involvement	Audio	$F^*(2, 232) = 5.98$	$p^* = .003$	0.049	MA, SA > NA
	Visual	$F^*(2, 232) = 23.95$	$p^* < .001$	0.171	FV > LV, NV

rendering, positively influences perceived **Presence** and the coherence between audio and visual modalities.

Measured vs. Synthesised & No Auralisation When participants ranked audio quality-related measures such as **Audio Naturalness** and **Audio Reverberance**, they could discriminate between the three audio conditions, ranking MA the highest and NA the lowest (see Figures 5.7c and 5.7d). For **Audio Naturalness**, MA was rated significantly higher than SA and NA ($p < .001$, $p < .001$), and SA was higher than NA ($p < .001$). Similarly, for **Audio Reverberance**, MA outperformed SA and NA ($p = .024$, $p < .001$), and SA also outperformed NA ($p < .001$). These results could be related to inherent acoustic differences between the measured and synthesised SIRs. While some parameters, such as reverberation time (T30), remained below the JND threshold, others differed more substantially. As shown in Table 5.2, the largest difference occurred in Early Decay Time (EDT). This is particularly relevant since EDT has been shown to correlate more strongly with perceived reverberance than other acoustic metrics [244, 242]. The difference derives from the method used to create the synthesised SIRs, despite being configured to match the real space as closely as possible.

However, this was not true for **Spatial Presence**, **Realism**, and **Involvement**. For all three subscales, post-hoc comparisons for **Audio** showed no significant difference between MA and SA ($p = 1.000$ for all), but MA and SA were significantly higher than NA (SP: $p = .001$, $p < .046$; REAL: $p = .014$, $p < .037$; INV: $p = .016$, $p < .043$). Therefore, **H2** was partially validated, as



Figure 5.7: Experiment 1 mean scores across audio (MA, SA, NA) and visual conditions (FV, LV, NV).

MA was not consistently rated higher than SA across all measures. This suggests that although participants could recognise their vocal auralisation when evaluating audio-specific qualities, its fidelity did not significantly enhance the sense of **Presence** and its subscales.

Visual Influence in Audio Quality Perception **H3** was corroborated, as FV consistently outperformed the other visual conditions in **Audio Naturalness** and **Audio Reverberance**. In both measures, significantly higher scores were reported for FV than for LV and NV (ANAT: $p = 0.013$, $p = 0.011$; AREV: $p = .040$, $p = .028$), while no differences were found between LV and NV ($p = 1.000$ in both). These findings illustrate the relevance of visual fidelity as a dominant perceptual cue in multisensory experiences. However, due to the lack of an interaction effect, we cannot conclude that auditory feedback was perceived as higher when paired with high-resolution visual feedback. Similar effects were observed in other measures. It is also important to note that, due to rendering constraints in Unity, the task instruction UI in the NV condition was displayed at full resolution, unlike in the LV condition. This may have led participants to infer the visual resolution of the environment based on the UI, potentially influencing their ratings of the NV condition. In addition, the presence of the UI, positioned similarly to FV and LV, may have

provided an additional sense of spatial awareness, making participants feel as though they were in a virtual room despite the absence of most visual content.

Similar effects were also seen in **Spatial Presence**, **Realism**, and **Involvement**, where FV outperformed LV ($p < .001$ for all) and NV ($p < .001$ for all) consistently, but LV and NV were not statistically different (SP & REAL: $p = 1.000$ for both; INV: $p = .114$).

Audio Influence in Visual Quality Perception Referring back to our hypotheses, **H4** was not validated, as the two-way ANOVA did not reveal significant interaction effects between conditions. While **Visual Resolution** (see Figure 5.7e) scores were higher for MA and SA compared to NA ($p < .001$, $p < .001$), there was no significant difference between MA and SA ($p = 0.971$). For the **Visual** factor, significant differences were found between FV and both LV and NV ($p < .001$, $p < .001$), but no significant difference was observed between LV and NV ($p = 1.000$). These findings suggest that although **Visual Resolution** scored higher in MA and SA, it was not significantly influenced by the combined effect of audio and visual feedback.

5.5 Experiment 2

Experiment 1 showed no statistical significance between MA and SA, except in ANAT and AREV measures. Similarly, LV and NV did not show significant differences, except for INV. Therefore, in Experiment 2 we decided to examine only two representation levels for each factor. For auditory feedback, the conditions included Measured Auralisation (MA) and No Auralisation (NA). The visual conditions comprised MR and VR at full resolution.

To maintain methodological consistency with Experiment 1, the equipment, measures, audio rendering, and procedure remained unchanged. Since MR involves real-time reconstruction and stereoscopic view synthesis of the real-world environment [245] rather than a direct video feed, users perceive a virtual representation of their surroundings. Therefore, the IPQ questionnaire was retained to assess Presence and maintain statistical consistency.

5.5.1 Stimuli

Visual Rendering This experiment presented two visual conditions. The VR condition was a detailed virtual reconstruction of the physical room. Additional matching posters were placed in the physical and virtual rooms to help assess the visual quality (see Figure 5.8). MR was a real-time video feed of the physical environment using the passthrough feature of the HMD. Passthrough does not simply provide a raw video feed from the cameras. Instead, it delivers a visual reprojection derived from the external-facing cameras and the depth sensor. This reprojection is adjusted to match the correct perspective for each eye and create a spatially accurate video feed. However, this process provides a moderate visual quality performance [246] and can introduce visual artefacts, and errors in depth estimation within the video [247].

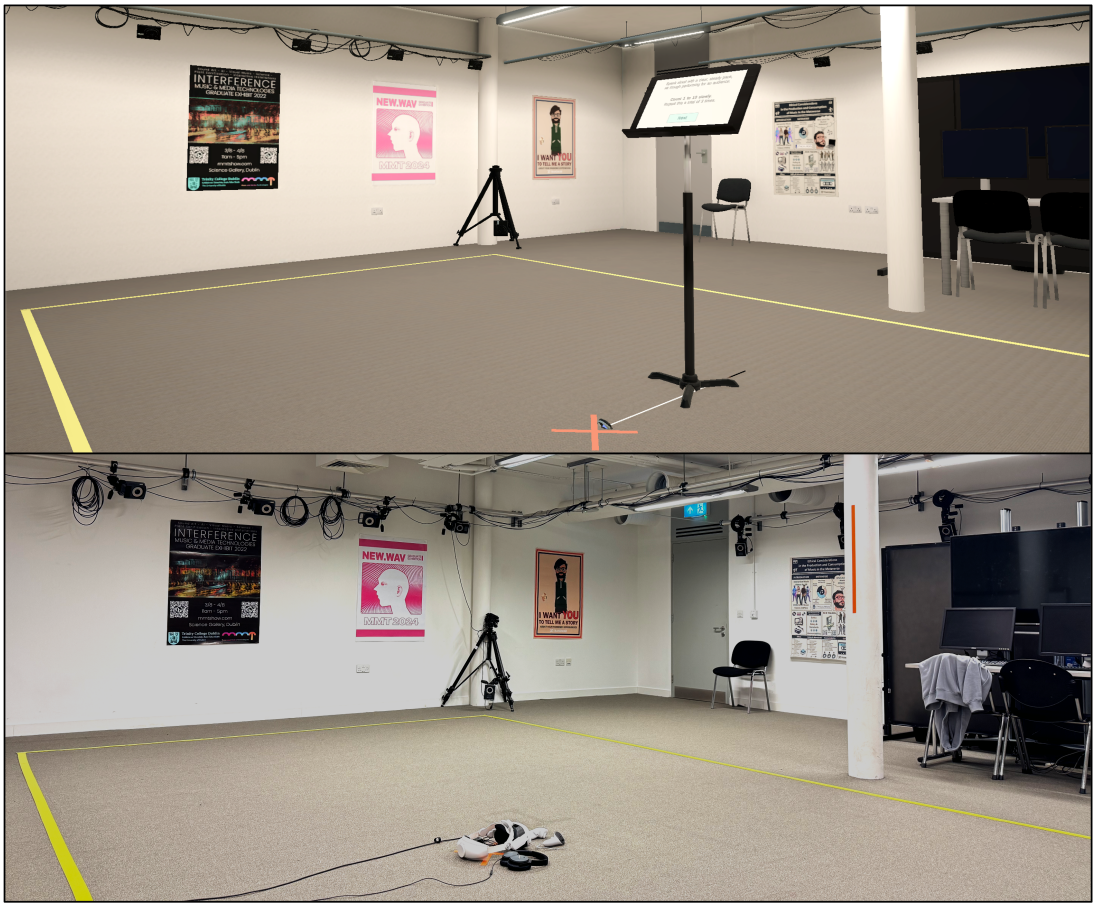


Figure 5.8: Visual comparison between the virtual room (top) and the physical room (bottom) used in Experiment 2.

Participants embodied a human avatar in the VR condition to promote a sense of embodiment. The avatar was customised to match their skin tone since they could see their bodies in the MR condition. In addition, participants were asked to wear a grey sweatshirt similar to that of the virtual avatar.

5.5.2 Participants

Thirty-two participants were recruited for the experiment. Of these, 21 identified as male, 10 as female, and 1 as non-binary. Participants ranged in age from 21 to 50 years (mean ± SD: 31 ± 8.29). 7 participants reported audio training, 7 reported vocal training, and 15 reported musical training. Among those with training, 12 had six or more years of audio, musical or vocal experience. Furthermore, participants reported on their VR experience, where 5 had never used VR, 8 had played one or two VR demos, 6 had used VR several times, 2 owned a VR headset, and 11 were VR developers/researchers.

Table 5.4: Experiment 2 Aligned Rank Transform (ART) ANOVA results with post-hoc comparisons.

Measure	Factor	F-statistics	p-value	Effect-Size	Post-hoc
Audio Naturalness	Audio	$F(1,90) = 42.36$	$p < .001$	0.320	MA > NA
	Visual	$F(1,90) = 4.26$	$p = .042$	0.045	VR > MR
Audio Reverberance	Audio	$F(1,90) = 52.11$	$p < .001$	0.366	MA > NA
	Visual	$F(1,90) = 1.33$	$p = .252$	0.015	ns
Visual Resolution	Audio	$F(1,90) = 12.62$	$p < .001$	0.123	MA > NA
	Visual	$F(1,90) = 5.94$	$p = .017$	0.062	VR > MR
Audio-Visual Integration	Audio	$F(1,90) = 34.58$	$p < .001$	0.278	MA > NA
	Visual	$F(1,90) = 4.06$	$p = .047$	0.043	VR > MS
Presence	Audio	$F(1,90) = 5.59$	$p = .020$	0.058	MA > NA
	Visual	$F(1,90) = 1.01$	$p = .318$	0.011	ns
Spatial Presence	Audio	$F(1,90) = 12.94$	$p < .001$	0.126	MA > NA
	Visual	$F(1,90) = 5.02$	$p = .028$	0.053	VR > MR
Realism	Audio	$F(1,90) = 10.71$	$p = .002$	0.106	MA > NA
	Visual	$F(1,90) = 13.11$	$p < .001$	0.127	VR < MR
Involvement	Audio	$F(1,90) = 1.64$	$p = .203$	0.018	ns
	Visual	$F(1,90) = 17.98$	$p < .001$	0.166	VR > MR

5.5.3 Procedure

The experimental procedure closely followed that of Experiment 1; however, the physical room was the audio-visual reference, as both VR and MR were representations of the physical room, providing a common ground for comparison between the two. Participants stood at a set position in the room to observe visual elements and count aloud from 1 to 10 to assess audio quality. Afterwards, participants wore the HMD and headphones and calibrated the virtual room to align with their physical position and orientation. They then completed four blocks, one for each experimental condition. Due to the experiment's short duration, no breaks were provided.

5.5.4 Results

As in Experiment 1, a two-way within-subjects Aligned Rank Transformation (ART) analysis of variance (ANOVA) [203] was conducted to assess the impact of **audio** and **visual** feedback, as most data violated the normality assumption. Post-hoc pairwise comparisons ($\alpha = .05$) were again conducted using Aligned Rank Transformation Contrasts (ART-C) to evaluate significant differences. Bonferroni correction was applied to account for errors across multiple comparisons.

Similarly to Experiment 1, no significant interaction effects were found between the **Audio** and **Visual** factors across all measures. Table 5.4 summarises the main effects for each factor and the relevant findings discussed below.

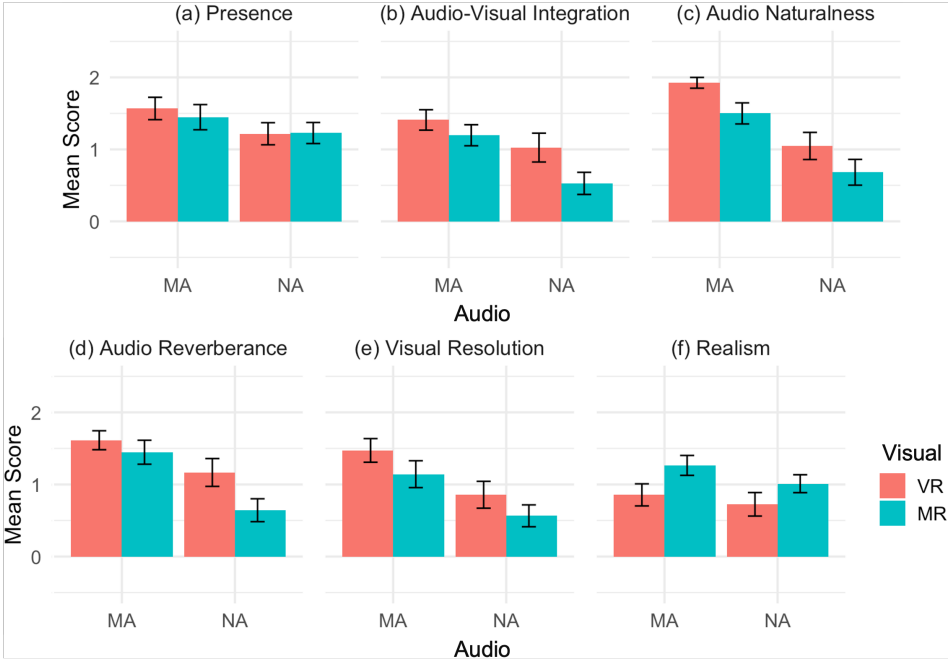


Figure 5.9: Experiment 2 mean scores across audio (MA, NA) and visual conditions (VR, MR).

Presence In line with the results of Experiment 1, **H5** was validated, as the perceived sense of presence was significantly higher in conditions where vocal auralisation was present. For the **Presence** measure (see Figure 5.9a), MA was rated significantly higher than NA ($p = .020$), while no significant difference was observed between VR and MR ($p = .318$). These findings underscore the relevance of real-time vocal auralisation not only in VR but also in MR experiences, while also suggesting that participants perceived a similar sense of presence across both conditions.

Audio-Visual Integration and Visual Resolution Contrary to our expectations, the results revealed significantly lower scores for MR for the **Visual** factor in both **Audio-Visual Integration** ($p = .047$) and **Visual Resolution** ($p = .017$), leading to the rejection of **H6** (see Figures 5.9b and 5.9e). However, MA was again rated significantly higher than NA across both measures (VRES: $p = .001$, AVI: $p < .001$). Similar results were obtained in all other measures, except for **Realism**. These results suggest that MR and VR were perceived similarly, possibly due to passthrough limitations of the Quest 3, such as low rendering quality and visual artefacts [247], or shortcomings of the IPQ in capturing MR experiences.

Audio Naturalness and Audio Reverberance Since the MR offers higher visual fidelity and a more accurate depiction than the 3D model used in VR, we expected **Audio Naturalness** and **Audio Reverberance** to be rated higher in MR. Surprisingly, the results indicated the opposite, with VR scoring higher than MR in the **Visual** factor (ANAT: $p = .042$, AREV: $p = .252$), thus

rejecting **H7**. Regarding the **Audio** factor, in **Audio Naturalness** (see Figure 5.9c), MA was rated significantly higher than NA ($p < .001$). A similar trend was observed in **Audio Reverberance** (see Figure 5.9d), with MA significantly outperforming NA ($p < .001$), although no significant difference was found between VR and MR conditions ($p = .252$). This suggests that participants perceived a more natural vocal environment with auralisation and surprisingly found VR more convincing than MR. However, this unexpected outcome could have arisen from a mismatch between the auralised acoustic response and the actual room or participants' higher expectations due to the more faithful MR representation.

Spatial Presence, Realism, and Involvement **Spatial Presence** and **Realism** were significantly influenced by the **Audio** factor (SP: $p = .001$, REAL: $p = .002$), and by the **Visual** condition (SP: $p = .028$, REAL: $p = .001$). Interestingly, **Realism** scores favoured MR, rating higher than VR, unlike the other subscales (see Figure 5.9f). Similar to Experiment 1, participants clearly distinguished between MA and NA, with MA yielding significantly higher scores across measures, especially in audio-related measures. This was not true for **Involvement**, in which no significant effects were found for the **Audio** factor ($p = .203$), but VR showed higher **Involvement** scores than MR ($p < .001$), which may indicate that **Involvement** is less susceptible to audio feedback. For this reason, **H8** was only partially supported. This suggests that, in both MR and VR, sound auralisation plays a key role in enhancing the overall perceived quality of IVE experiences [8] and that its inclusion can lead to a more effective audio-visual experience.

5.6 Discussion

Audio Naturalness and Reverberance Results from experiment 1 show that measured auralisation yielded higher scores than synthesised auralisation in audio-related questions, specifically audio naturalness and audio reverberance. However, this was not true for presence and its related subscales, where no significant differences were found between measured and synthesised audio conditions. These findings may have practical implications, informing the development of audiovisual experiences in VR and MR. They suggest that creating the best synthesised auralisation possible for vocal auralisation may be sufficient for enhancing perceptual effects such as presence, potentially avoiding the need to produce measured SIRs, which require considerable resources in terms of equipment, capture time, and system implementation. However, using synthesised over measured SIRs may depend on the type of experience and the level of auditory fidelity required. Measured auralisation may still offer perceptual advantages, particularly in audio-focused applications such as high-fidelity simulations, and singing practice and performance, which will be examined further in Chapter 6.

Audio-Visual Integration Multisensory systems in VR ecosystems that integrate different senses enhance the user experience, as congruent sensory feedback is perceived as unified, creating a coherent experience of the environment [15, 23]. Real-time vocal auralisation within this multisensory context reinforces this effect, yielding higher scores for audio-visual integration in both VR and MR conditions. This underscores the importance of rendering the acoustic features of the environment as a core component of multisensory feedback in immersive applications, particularly those involving self-generated sounds such as vocal production.

It should be noted that although participants wore noise-cancelling headphones to block the acoustic response of the physical room, some ambient sound may still have been audible, which may have influenced the results, particularly in the No Auralisation conditions. In most cases, this can be mitigated by conducting the experiment in an acoustically transparent space, such as an anechoic chamber or an acoustically treated room. However, this was not feasible due to the nature of Experiment 2, as participants needed to be physically present in the room to match the visual representation of MR and VR. Regardless, the consistently higher scores for audio-visual integration in both VR and MR indicate that the effect was salient, despite this constraint.

Presence Both visual and spatial audio fidelity are important for achieving a high sense of immersion in VR applications involving sounds produced by external sources in the environment [8]. Building on this, our study explored first-person vocal auralisation in an audiovisual context. Across two experiments, we found that the fidelity of audio and visual feedback contributes to the overall sense of presence in both VR and MR environments. However, since no interaction effect was observed between the audio and visual factors, it remains unclear whether their combination produces a greater impact on presence than each factor alone. Moreover, if we consider the excited acoustic response of the environment as part of the perceived ambient soundfield, and given that participants used their voice as self-produced audio feedback, our results align closely with previous research [67, 32], reinforcing the notion that vocal auralisation further enhances the perceived sense of presence through the inclusion of both ambient and self-generated sound.

Nonetheless, it is worth noting that the main effect for the visual factor was not observed in Experiment 2, as VR and MR were not statistically different. Participants seemed to judge MR more critically, as it is a direct reprojection of the real world rather than a virtual representation. This might be due to several factors. First, the 3D model used in VR was detailed and spatially accurate, closely resembling the physical space, which might have made it as convincing as the MR representation (see Figure 5.8). A second factor may have been the limitations with the HMD passthrough in terms of visual quality, artefacts, and errors in depth estimation [247]. For example, participants noticed a lack of clarity in the posters' text within the physical room compared to the VR counterpart, along with noticeable rendering fluctuations. The IPQ questionnaire might have also been a limitation. It was selected due to its relatively short length and prior use in

audio-related VR research [5], and we retained it for Experiment 2 to ensure methodological consistency. Although it has been used in MR contexts, it is not without shortcomings, as it was not specifically designed for this purpose, and the notion of 'being there' may not fully align with MR [248]. However, to our knowledge, there are currently no validated questionnaires available to measure Presence in MR.

In addition, these results may have been influenced by the Uncanny Valley effect, where users become more critical as depictions approach photorealism [249]. MR environments, particularly those using passthrough technology, are more closely tied to the real-world than VR, potentially raising user expectations and leading to harsher evaluations. In contrast, VR environments are fully synthetic, and users have a clear understanding that what they are perceiving is entirely virtual, allowing them to adopt more forgiving standards for visual realism and multimodal integration. Further research on this outcome is required to better understand how MR experiences are perceptually evaluated.

Beyond presence, the study highlights the importance of first-person vocal interaction and auralisation in multimodal perception within XR. Moreover, it provides a methodological framework that opens avenues to investigate key perceptual aspects in XR, such as the sense of embodiment and self-presence. Research suggests that auditory feedback, particularly the perception of one's own voice, contributes to the sense of agency, can influence body ownership, and alter mental self-representation [177, 250, 16]. Providing users with altered vocal feedback through artificial manipulation, simulating environmental conditions or vocal changes due to factors such as vocal health, may influence self-perception. Therefore, vocal perception, as a self-generated sound, opens new directions for exploring multisensory self-representation and the influence of vocal changes on self-identity.

Spatial Presence, Realism and Involvement Our results showed that vocal auralisation in VR led to higher scores across all three IPQ subscales, suggesting that first-person vocal feedback enhances presence and also other perceptual dimensions. This aligns with prior research suggesting that sound auralisation and self-produced audio cues can increase spatial awareness and the sense of agency, and the impression of being the source of one's actions [73, 177].

As mentioned previously, real-time vocal auralisation presents several implementation constraints, particularly regarding latency. Since the participant's voice was produced within a reverberant space, the microphone captured some of the acoustic features of the environment. Although they could have been removed using audio processing effects, it would have incurred unwanted additional signal latency. To minimise this, the microphone was placed very close to the participant's mouth, keeping the gain level to a minimum to reduce the amount of ambient sound captured. This placement resulted in audible plosives and sibilance, which could have potentially impacted the quality of the auditory feedback and, thus, the experimental results.

However, the results suggest that these artefacts were not noticeably perceived by participants and did not appear to affect their experience.

5.7 Conclusion

We have presented two experiments in which we examined how real-time first-person vocal auralisation influences the perceived sense of presence and audio-visual quality across different rendering resolutions and visual representations, including VR and MR.

Overall, the results consistently revealed higher presence ratings in conditions where real-time vocal auralisation was included, with similar effects observed across the IPQ subscales of Spatial Presence, Realism, and Involvement. These findings support **RQ1** of this thesis, indicating that the integration of both ambient and self-generated auditory feedback can promote perceptual illusions and elicit psychological responses in multimodal IVEs. Although an additive effect of audio and visual feedback cannot be confirmed, as no interaction effect was observed, the results suggest that congruent multimodal feedback enhances the perceptual experience without necessarily producing a compounded effect.

Regarding **RQ2**, these findings highlight that rendering fidelity in both audio and visual domains affects perceived quality and audio-visual integration. While Measured Auralisation offered perceptual benefits in terms of Audio Naturalness and Reverberance, Synthesised Auralisation proved sufficient to elicit a sense of presence, suggesting a practical compromise, particularly given the limitations associated with capturing SIRs. Moreover, the results revealed higher scores for audio-visual integration in both VR and MR, suggesting that rendering fidelity across both modalities strengthens perceived perceptual quality. While visual fidelity was a more prominent factor, the integration of congruent self-produced auditory feedback highlights the importance of multimodal rendering consistency.

These results underscore the central role of self-produced auditory feedback in shaping perceptual quality and realism in multimodal immersive environments, while also pointing to technical and design considerations for future XR systems. In addition, they provided the motivation for further explorations of vocal auralisation in multisensory applications. The following chapter presents an evaluation of first-person vocal auralisation in practical contexts such as vocal rehearsal and performance.

Chapter 6

Evaluation of an Audio-Visual IVE for Vocal Rehearsal & Performance

Based on **"Immersive Vocal Practice: Audio-Visual Evaluation of an Acoustic Virtual Environment for Vocal Rehearsal and Performance"**[1]
Article published at IEEE IWMM 2025

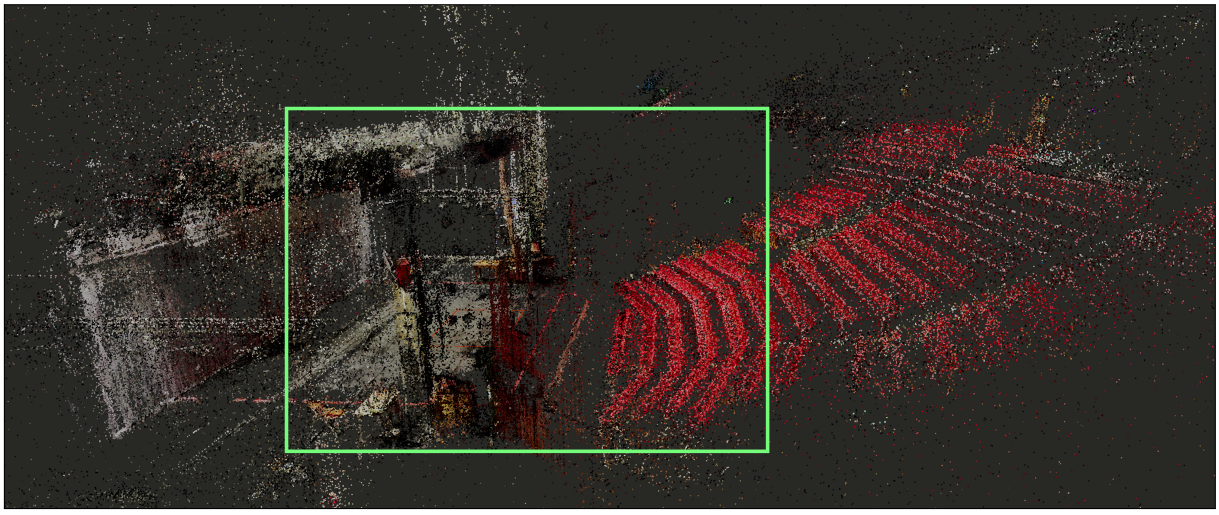


Figure 6.1: Point cloud distribution of the GS rendering, with higher density at the stage and close surroundings (highlighted in green).

[1]*The research described in this chapter was co-authored with Rose Connolly (Trinity College Dublin), who contributed to the data analysis and manuscript write-up.*

The perception of auditory feedback in a multisensory context plays a crucial role in shaping user experience in IVEs.As XR technology evolves, the ability of these systems to provide high-fidelity and coherent multisensory feedback highlights their intrinsic potential for multimodal integration. This enhances users' felt presence and the realism of virtual environments, making XR an effective platform for training and virtual simulations. Chapter 5 examined the influence of real-time vocal auralisation on the perceived sense of presence and audio-visual quality at different rendering resolutions and visual representations. It also compared VR and MR, revealing higher ratings of presence and perceived audio-visual quality in conditions where real-time vocal auralisation was present. In addition, it employed VOREA, which enabled real-time first-person auralisation using directional interpolation at a fixed location, as participants remained at a defined position within the IVE.

In this chapter, extending prior explorations of IVEs, we present an audio-visual evaluation of a VR-based virtual theatre designed for vocal performance and practice. This virtual theatre was created using audio-visual capturing techniques, including GS and 3OA SIRs, to provide participants with a high-fidelity and accurate multisensory representation of a real performance space. The setup allows singers to dynamically interact with and explore the acoustic response of the space using their voice, across different positions and directions within the stage, providing directional and translational DoF. Their vocal input is auralised in real time to reflect the acoustics of the simulated venue, enabling immersive, audio-visual exploration. Following the objective measures presented in Chapters 3 and 5, this chapter shifts the focus towards a practical, qualitative evaluation. By conducting a thematic analysis of singers' experiences, we aim to uncover perceptual and practice-related insights into the usability, realism, and potential of audio-visual IVEs for vocal performance. Therefore, this chapter moves from perceptual validation towards applied evaluation, exploring the potential of audio-visual IVEs as a platform for vocal rehearsal and performance.

Singers' vocal production is directly influenced by the inherent acoustic characteristics of the room in which they perform. The perception of their voice is shaped by interactions with the space's acoustic features, such as early decay time, early reflections, and reverberation. These interactions dynamically affect various aspects of their vocal production, such as vibrato extent, pitch accuracy, and fundamental frequency [251, 252], and also inform their musical approach, leading to performance variations in response to perceived room acoustics [234].

However, singers often rehearse in spaces that differ significantly from performance venues in terms of dimensions and acoustic characteristics [185]. While these spaces may be suitable for certain aspects of vocal and musical practice, they often lack adequate audio-visual conditions, forcing singers to adapt to the practice environment [235]. This leaves them without the ability to develop the necessary skills and strategies essential for musical performance in actual venues.

Studies have shown that vocal adaptations occur in artificial acoustic environments similarly

to physical spaces, affecting singers' vocal production by increasing vocal effort, sound level, and fundamental frequency [28, 238]. Such adaptations also prompt the use of individual strategies to adapt their performance to various concert venues [35]. These findings make virtual auralisation a valuable tool for accounting for acoustic limitations and providing users with acoustic feedback environments suitable for vocal performances.

While VR has demonstrated value for musicians more generally, its application to singers has also been investigated. Research has explored the potential of VR as a training method for singers [184, 185]. VR-based training has been shown to improve breath control and vocal comfort [186], reduce stage fright and enhance performance quality [187], and support the development of emotional expressiveness [188]. In parallel with these findings, recent work highlights the importance of multimodal integration, with audio-visual coherence playing a key role in shaping voice perception and production in VR [29]. Interactive systems have been developed that enable singers to explore diverse virtual spaces [12], perform with a virtual ensemble and choir [34], and rehearse in VR environments with realistic acoustic conditions using spatialised choral recordings [185]. These systems highlight the potential of multisensory feedback to support rehearsal and practice, the development of performance and vocal adaptation strategies, and they demonstrate the value of VR as a training tool for singers.

Overall, research has explored the use of virtual acoustics and VR for their influence on singers' vocal production and training. While several high-fidelity audio-visual environments have been designed, much of this work has been conducted from an audience perspective rather than focusing on performers' experiences[11, 171, 168, 170]. In addition, few singing-centred VR applications and studies have provided singers with auditory feedback through real-time auralisation of their voice to support meaningful audio-visual integration. Despite the demonstrated benefits of multimodal integration, such as increased presence, improved user experience and greater ecological validity [31, 32], real-time vocal auralisation and its potential in shaping singers' overall experience in IVEs has received limited attention.

Simulation training has been widely applied in musical performance to recreate the physical and psychological conditions of performing. VR-based simulations extend this by offering the possibility to rehearse in audio-visual reproductions of performance spaces, showing that VR can elicit physical and emotional responses and reduce anxiety related to public vocal performance. Although studies have recreated audio-visual conditions successfully, the exploration of singers' experiences in IVEs remains limited, particularly in relation to how they engage with the acoustics of a stage. Existing approaches often focus on a fixed position within the virtual environment, rather than giving singers the agency to move across the stage. Allowing such vocal exploration across different positions provides a useful dimension for understanding performance in IVEs.

At the same time, the technical approaches of these environments remain an important consideration. Singing-related VR research measuring different aspects of vocal production and

training typically relies on 3D-modelled IVEs. While most commercial applications in the area also follow this approach, they often provide limited visual detail of real performance spaces. Visual feedback has been shown to influence the perception of auralisation [9], emphasising the need for higher-fidelity visual representations. GS has potential for high-fidelity and photorealistic rendering of complex environments, which, when combined with spatially accurate auralisation, presents a promising direction for developing multimodal IVEs of performance spaces.

Building on these observations, we designed a study in which singers evaluated three environments: a GS environment with real-time auralisation of their voice using high-fidelity acoustic impulse responses, a small physical room resembling typical rehearsal spaces, and a 3D-modelled scene without vocal auralisation. We investigate how these environments are perceived in terms of plausibility and effectiveness for vocal performance and rehearsal, focusing on singers' qualitative experiences within a multimodal VR-based setting.

The motivation for using a GS capture arises from its growing popularity as a technique for recreating physical spaces with photorealistic fidelity, and its conceptual analogy to SIRs as a physically grounded capture of an environment. In contrast, the 3D model without vocal auralisation reflects the commonly used practice in many VR applications, which enabled us to evaluate a higher-fidelity multisensory environment against a conventional VR scene. Furthermore, allowing singers to move around the stage enabled us to evaluate the perception of real-time vocal auralisation from a first-person perspective and its implications as a method for acoustic exploration in immersive performance environments.

6.1 Hypotheses

Based on our objectives and given that self-produced auditory feedback shapes perceptual quality and presence in multimodal immersive environments, we propose the following main hypotheses:

H1 Singers will report a more positive response and greater control over their vocal performance and emotional expression in the virtual theatre space with real-time vocal auralisation.

H2 The virtual environment will be perceived as a plausible performance venue, helping singers bridge practice and performance experiences.

As previously established in chapter 2, most virtual representations are designed from the perspective of an audience, and only a few studies or prototypes have created environments specifically for singing rehearsal and training. However, these environments are generally based on 3D models and rarely incorporate vocal auralisation of the user's voice. In addition, prior research has identified the significant influence of visual rendering on auditory perception. We hypothesise that:

H3 The GS scene with vocal auralisation will provide a greater sense of audio-visual realism and stronger cognitive and emotional responses compared to the 3D model condition without auralisation.

The perceived acoustic characteristics of a performance space vary depending on the singer's position on stage, including aspects such as clarity, early reflections, and diffuse reverberation. These variations directly affect how singers perceive their voice, often making them adjust their voice production and musical delivery accordingly. Therefore, the IVEs in the study allow singers to move freely around a stage area, with the expectation that:

H4 Singers will actively use movement as a meaningful method of acoustic exploration across the stage.

Based on the motivation, research aims, and the results discussed in Chapter 5, the hypotheses outlined above stem from **RQ3** of this thesis, with the goal of fully exploring how the audio-visual representation of a virtual performance space can contribute to vocal-related practices within the performing arts, as well as broader applications such as public speaking. In particular, they address the extent to which XR can help overcome current limitations and serve as a useful tool for vocal rehearsal and performance.

6.2 Experimental Setup

The evaluation was carried out using a Meta Quest 3 Head-Mounted Display (HMD) tethered to a PC equipped with an Intel i9-11900KF 3.50GHz CPU, 64GB RAM, and an Nvidia GeForce RTX 3080 Ti graphics card. For participants' vocal input, a DPA SCO60F00-H omnidirectional microphone was fixed to the front of the HMD to ensure consistent placement, while audio playback was delivered using a pair of Bose QuietComfort 25 noise-cancelling over-ear headphones (see Figure 5.2). Both the microphone and headphones were connected to a Scarlett 2i2 mk4 audio interface.

Interaction System The interaction system was implemented using Unity's XR Interaction Toolkit, enabling spatial tracking, user input, and locomotion. Spatial tracking provided the user's position and rotation data, which was used within VOREA for real-time vocal auralisation. During immersion in the IVEs, participants could explore the virtual spaces from an area on the stage. This area, delimited by a visual rectangle, was predefined during the audio-visual capture of the physical space and matched its position within the IVEs. The locomotion system provided teleportation and discrete rotation using the thumbsticks on the HMD's controllers, allowing participants to move around within the area (see Figure 6.2). While the system supported free walking, we chose to conduct the study in a small room due to its dimensions and lack of

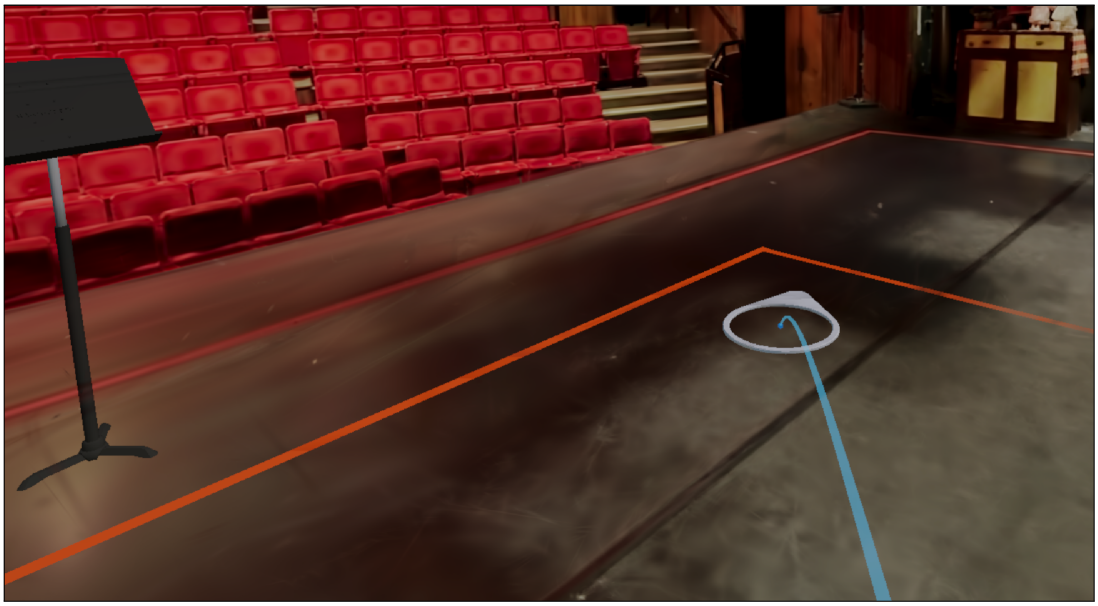


Figure 6.2: Locomotion system implemented in the IVEs, showing the ray interactor used to navigate within the defined stage area.

acoustic suitability, reflecting the type of spaces in which singers commonly practice. Therefore, teleportation was used due to the limited physical space available in the experimental room.

6.3 Measures

This study examined subjective measures based on participants' evaluations and reflections. Qualitative data were collected using semi-structured interviews, which allowed researchers to obtain rich and detailed insights into participants' perspectives on their experiences. This approach facilitated the subjective evaluation, enabling the identification of patterns and meaningful interpretations through thematic analysis. The interview considered human, system, and context factors exhibited by the immersive experience, which contribute to its quality and usability [253]. Aspects explored during the interview included:

- Audio-Visual Quality – Naturalness of the audio feedback, clarity or sharpness of the visual feedback, and multisensory coherence between both stimuli.
- Vocal Performance – Vocal production and performance changes resulting from the audio-visual experience.
- Cognitive and Interaction Factors – Level of perceived presence, involvement, and focus during immersion.
- Emotional Experience - Emotional changes experienced during immersion.

- Usability and Adoption – Ease of use of the system, usefulness and likelihood of adopting the VR application for vocal-related applications.
- Experience Consequences – Physical discomfort or negative effects experienced during or after immersion.

A thematic analysis was performed to identify relevant themes and patterns in our data [254]. Transcripts were coded to capture key concepts, with codes grouped into potential themes and sub-themes, which were iteratively reviewed and refined. While the interview was shaped by specific research interests and theoretical considerations, providing some deductive grounding, the analysis followed a primarily inductive approach, allowing themes to emerge from the data itself. Rather than coding for pre-defined categories, we allowed patterns and themes to emerge from the data itself, based on participants' language and framing. The analysis focused on the explicit content of the interviews, rather than underlying assumptions, and was therefore conducted from a semantic and largely inductive standpoint [255].

6.4 Experimental Design

To investigate the hypotheses established for this study, singers were asked to experience and visually and vocally assess three spaces:

- A small physical office (3.66 x 3.16 x 2.62 m) with no acoustic treatment and no additional real-time vocal rendering over headphones.
- A GS scene with real-time vocal auralisation (GSA).
- A 3D Modelled (3DM) scene with no real-time vocal rendering over headphones.

Participants evaluated the physical space before the virtual scenes, using it both as the ground truth for comparison with the virtual conditions and as a reference for the visual and acoustic characteristics of typical rehearsal and practice rooms. Afterwards, participants assessed the two virtual environments, which were presented in a counterbalanced order and equally distributed across participants to account for ordering effects. Finally, interviews were conducted after participants had assessed the three spaces to allow them to compare their experiences across conditions and provide more in-depth insight into their perceptions and impressions.

6.4.1 Stimuli

Visual feedback For this study, two virtual environments were designed using the Unity Game Engine (version 6000.1.1f1). The GSA scene was a captured representation of the Abbey Theatre, a prominent theatre in Ireland (see Figure 6.3 top). To produce the GSA environment, several



Figure 6.3: Visual comparison between the Gaussian Splatting (top) and the 3D Modelled (bottom) virtual environments used in the study.

videos were captured in the physical space using an iPhone 15 Pro and subsequently processed with Varjo's Teleport platform [256], which employs a cloud-based service for GS. The output was then imported into Unity and rendered using a Gaussian Splatting Virtual Reality rendering package [27]. The environment was rendered using different quality levels based on the number of splats, similar to Level of Detail (LOD) techniques, to optimise computational performance. To ensure the experience ran at a minimum of 90 FPS in the headset, the number of splats was reduced from over 4.5 million to under 2 million, without significant impact on visual quality, by strategically reducing splats in areas of lower visibility and importance, such as the stage ceiling (see Figure 6.1).

The 3D Model (3DM) scene was a theatre sourced from the Unity Asset Store [257], which was modified to resemble the Abbey Theatre and match its main features, including dimensions,

Table 6.1: Participant experience, practice environment, and vocal range data.

Participant	Vocal Range	Professional Experience	Virtual Reality Experience	Practice Room Size	Practice Room Type
P1	Soprano	5 years	I have played one or two virtual reality demos	Similar	Bedroom
P2	Tenor	6 years	I have played one or two virtual reality demos	Similar	Dedicated room at home
P3	Baritone	7 years	I have played one or two virtual reality demos	Larger	Bedroom & occasional theatre
P4	Mezzo-soprano	5 years	I have never used virtual reality	Similar	Bedroom & Music rehearsal room
P5	Mezzo-soprano	10 years	I have played one or two virtual reality demos	Similar	Music rehearsal room
P6	Soprano	12 years	I have used virtual reality several times	Similar	Bedroom
P7	Soprano	18 years	I have never used virtual reality	Smaller	Dedicated room at home
P8	Tenor	14 years	I have played one or two virtual reality demos	Slightly Larger	Dedicated room at home
P9	Soprano	32 years	I am a virtual reality developer/researcher	Slightly Larger	Dedicated room at home
P10	Baritone	35 years	I am a virtual reality developer/researcher	Larger	Acoustically treated studio
Practice room size in reference to the physical experiment room					

geometry, object distribution, and materials (see Figure 6.3 bottom). Moreover, to provide further depth and visual rendering quality cues [241], 3D objects were included around the stage wings.

Audio feedback The GSA scene provided real-time auralisation of the Abbey Theatre. The acoustic response of the theatre (average of the 500 Hz and 1 kHz bands) exhibits a moderate reverberation time $T_{30} \approx 795$ ms, with an $EDT \approx 789$ ms, reflecting a relatively low and controlled reverberant response appropriate for a theatrical space. It presents clarity values of $C_{50} \approx 2.48$ dB and $C_{80} \approx 5.74$ dB, pointing to moderate early reflection support. The lateral fraction coefficient $JLF_c \approx 0.07$ indicates relatively low lateral reflection energy. However, this low value may be influenced by the loudspeaker directivity used during the SIR capture.

The real-time auralisation was handled using VOREA, convolving participants' voices with SIRs generated from acoustic measurements of a predefined area of the Abbey Theatre stage (see Section 4.1). As the study allowed users to navigate within this area, interpolation was performed at both the directional level (4-way CPP) and the translational level (IDW). The user's orientation and position in the virtual environment were transferred from Unity to Reaper, where the gain of each convolution track was adjusted accordingly, and the final auralised signal was produced by summing the weighted SIR contributions, dynamically interpolated based on the user's movement within the IVE (see Section 4.2).



Figure 6.4: Participant singing while immersed in the GSA environment, wearing the augmented HMD.

6.4.2 Participants

Ten professional singers participated in the evaluation (4 males and 6 females) between the ages of 24 and 52 (mean ± SD: 32 ± 9.14). The participants had more than 5 years of professional musical experience, with a mean experience of 14 years, 7 years of vocal training, and 10 years of musical training. Additionally, they reported their vocal range, previous experience with VR, as well as the size and type of room in which they usually conduct vocal practice, as summarised in Table 6.1. All participants reported experience across multiple musical genres, including classical, choral, musical theatre, traditional, pop, soul, R&B, and rock, with most of them involved in a variety of music ensembles throughout their professional practice.

6.4.3 Procedure

Upon arrival, participants read and signed an informed consent form and were provided with detailed information about the evaluation procedure. First, participants vocally assessed the small physical office, which served as a reference for the typical types of spaces commonly used for practice, such as practice rooms, non-acoustically treated studio spaces, living rooms, and bedrooms. During this phase, participants wore headphones with noise cancellation disabled, allowing them to hear themselves and the natural acoustic response of the physical room. No audio feedback was provided through the headphones.

Afterwards, participants then wore the HMD to assess the GSA and 3DM IVEs. Once immersed in the IVE and before the vocal assessment, participants were asked to move around the virtual space to familiarise themselves with the locomotion method and settle into the environment.

Then, they were instructed to visually and vocally assess the spaces for a minimum of 3 minutes, as they would typically conduct their vocal practice and when evaluating a performance venue during rehearsals. Participants were then given as much time as needed to assess the space freely, allowing them to explore the environment vocally at their discretion (see Figure 6.4). While moving around the virtual stage, participants assessed the audio-visual environment by performing different vocal tasks, ranging from warm-up drills to singing entire pieces.

After experiencing the three conditions of the study, the semi-structured interview took place to allow participants to compare their experiences across conditions. They took approximately 10 minutes to complete, during which the experimenter guided participants with questions to systematically explore relevant factors, encouraging them to reflect on their experience and elaborate on aspects they found significant during the session. Afterwards, participants were given the opportunity to ask further questions about the study and were compensated for their time.

6.5 Results

Thematic Analysis resulted in the generation of 5 themes, each with 2 or 3 sub-themes, summarised in Table 6.2.

6.5.1 Shift into a Performance Mode

All participants indicated that the GSA condition elicited behaviour similar to that seen in real-life performance situations. This theme encompasses both emotional and vocal changes.

Nine participants described emotional responses typically associated with live performance. These included excitement (six participants) and confidence (three participants). Four participants experienced stage fright or anxiety. Other effects included a sense of focus or slipping into a performance 'mindset' (two participants), as well as the evocation of performance memories (two participants).

> *"Yeah, I felt very confident singing in the second one. I felt very like excited" (P6)*

> *"I did feel like I was getting ready for a performance. Yeah, so I did actually go into a different kind of mindset" (P7)*

Vocally, participants altered their singing techniques in response to the auralisation in the GSA. The most common alteration to vocal performance was projection (six participants). Other techniques included sustaining notes for longer, or adjusting their delivery to examine the acoustic response.

> *"I could decide how big the room was in the real kind of version and really decide how much projection I needed to use because that's kind of a big thing." (P5)*

Table 6.2: Themes, subthemes, and key findings from participant responses.

Themes & Subthemes	Description
Shift into Performance Mode	
↳ Emotional Changes	– The GSA evoked excitement, confidence, stage fright, and memories of past performances.
↳ Vocal Changes	– Participants adjusted vocal technique and delivery in response to the GSA.
Auralization Driving Exploration	
↳ Vocal Exploration	– Participants vocally experimented in the GSA environment to explore changes in acoustic response.
↳ Exploration Through Navigation	– Participants actively navigated the GSA environment to explore changes in acoustic response.
Immersion Rooted in Realism	
↳ Visual Realism	– Visual realism affected immersion, with focus on environmental detail.
↳ Audio Realism	– Auralization affected immersion and conveyed spatial authenticity.
↳ Audio-Visual Integration	– Synchrony between visual and audio was an important factor in realism and immersion.
System Usability Shaping Experience	
↳ Locomotion	– Teleportation was functional, but natural walking could improve immersion.
↳ Physical Components	– Some of the VR hardware slightly interfered with immersion and vocal performance.
Application and Utility	
↳ Practice	– The GSA supports rehearsal with acoustic feedback to practice staging, projection, and ease anxiety.
↳ Accessibility	– Enabling remote rehearsal, the GSA enables practice even when physical venues are unavailable.
↳ Learning & Engagement	– By building understanding of acoustic properties, the GSA serves as an educational, but fun tool.

In comparison, for the 3DM condition, participants did not report such vocal adjustments, with one participant describing the audio as *'nasty' (P8).* The 3DM condition also evoked fewer emotions, which were not linked to their performance. Two participants reported negative emotional reactions (fear and frustration), both of which were linked to the visual realism of the environment. Only one vocal shift was mentioned, describing a worsening vocal performance.

> *"I guess just like a bit more shaky because I felt like it wasn't getting back what the room should be giving me." (P6)*

6.5.2 Auralisation Driving Exploration

Eight participants engaged in spatial or vocal exploration to explore the GSA's acoustic feedback.

Five participants experimented vocally to test how their sound interacted with the environment. This included adjusting pitch, range, and pauses to observe the acoustic response.

"I did play around with the sound a bit more... I kind of let it fade out because of the reverb" (P7)

Eight participants discussed exploring changes in the acoustic response by navigating the environment. All of these eight noticed changes in orientation influenced the acoustic response. Most focused on the contrast between facing forward (toward the audience) and backwards (toward the curtain), while a few also noted differences when facing the wings, side walls, or other areas of the theatre. The remaining two participants did not comment on perceiving any acoustic variation in the audio feedback.

"I was moving around and hearing the difference in the reverb when you're out of projecting it to the audience versus projecting into the back side room or whatever. And you feel a big difference" (P2)

"I was just exploring every inch of the place and seeing how my voice reacted to it."

Participants could move using teleportation, though two particpants expressed a desire to explore further than the designated teleportation area.

"I wanted to get closer to the sides of the stage and stuff." (P1)

Such exploratory interactions were only reported in the GSA condition. In contrast, participants in the 3DM condition focused primarily on the singing.

"it kind of felt like, you know, you're just there to sing." (P7)

6.5.3 Immersion rooted in Realism

Immersion was strongly influenced by the perceived realism of the virtual environment. This sense of realism emerged across three areas — visual, auditory, and their integration, forming three sub-themes.

Overall, the GSA environment was perceived as visually realistic by six participants. Eight participants described the graphics as detailed and convincing, with one noting that *"because it was more detailed, it felt kind of more realistic" (P7).* Two participants also noted the addition of stage props in the GSA added to realism. Minor graphical imperfections (ceiling patches) were noticed in the GSA by three participants but described as unimportant, and *"not enough to break the illusion in any meaningful way." (P2)*

"It had a depth about it because of the detail... the devil is in the detail for stuff like that." (P10)

The 3DM's visual realism was described as poor by seven of the participants, likened to a *'mockup' (P1)*, *'game' (P2)*, and *"like if you asked AI 'make a theatre'"*. For four of these participants, this perceived lack of realism negatively affected their sense of immersion in the 3DM condition, making it more difficult to engage fully with the experience.

> *"the visuals didn't look that real, so it took a moment to kind of get into it" (P3)*

The audio in the GSA was described as realistic by seven participants.

> *"I felt more like I was in real performance room because of the reverb" (P9)*

On the contrary, the lack of auralisation audio in the 3DM model was described as dampening immersion or realism for seven participants.

> *"The first one was okay, but since I didn't hear the sound difference, it didn't really feel its real." (P7)*

Nine participants noted that audio-visual synchrony for GSA was a factor in improving their experience, particularly for immersion. However, two participants noted that audio realism was slightly more important to them than visual realism, as one stated they were *"more of an acoustic type of person" (P1)*.

> *"having a soundscape that moves with you and is, you know, acoustically correct to the size of the space you're in... ... it makes it a heck of a lot more real." (P10)*

6.5.4 System Usability Shaped Experience

Participants' experience in the VR environment was shaped by the usability of the system, specifically the physical characteristics of the hardware and the method of navigation.

Four participants commented on the headset's weight. Of these, two mentioned interference with vocal warm-up techniques. Two participants noted the light at the bottom of the HMD at the start, but upon entering the virtual world, failed to notice more. In addition, two participants noted that the tethering cable interferred with immersion.

> *"One thing I will say was.. ....I got used to it but.. ...when you first put on the headset, it kind of it pushes down your cheeks, it makes it harder to do like the vocal technique that is helpful for warming up." (P6)*

> *"I suppose with the headset, I found it heavy in some senses where like, going to sing, I thought it was like, okay, this is unusual. I've never done this before, with like pressure around my head." (P5)*

Seven participants commented on teleportation being intuitive or easy, but two participants noted that it reduced immersion or engagement. One noted in particular that it made precise positioning difficult.

> *"like the teleporting... ...it's functional... but it breaks the engagement a little bit, and especially because sometimes when you're teleporting you're not quite going to the same angle that you're expecting to. So then you're readjusting." (P3)*

Two participants expressed that physically walking would enhance immersion - one described it would be *"the ultimate way of feeling completely there." (P10)*

> *"I noticed myself wanting to walk, you know? I thought it was a really realistic capture of the space that it was so well done that I felt I needed to actually physically walk around and I had to stop myself." (P9)*

6.5.5 Benefits & Applications

All participants recognised the GSA as a valuable practice tool, especially for rehearsal and training. Three sub-themes emerged: practice, accessibility, and learning.

All participants mentioned the GSA environment would serve as a valuable rehearsal tool, helping to prepare them both vocally and psychologically. Many appreciated the ability to rehearse both vocally and spatially in a simulated venue, noting that it allowed them to experiment with projection, adjust to acoustic conditions, and rehearse stage positioning. Four participants also cited its potential to reduce performance anxiety or build confidence. However, two participants suggested that a virtual audience would further enhance realism and help with stage fright as you could *"get used to the eyes on you" (P7).*

> *"You can rehearse all you want outside of a venue and then you go and perform and you might have everything down, but then all of a sudden you realise that your balancing is off. It really makes a huge difference. (P3)"*

> *"for people who are more prone to the nerves and really want to get comfortable with performance anxiety, I think it could really set the level a lot to experience that." (P2)*

Two participants highlighted the GSA's value in supporting remote rehearsal, especially when venue access is restricted due to scheduling, cost, or location.

> *"Performers don't have very much time beforehand to be able to assess a space... ...it would it would be a great heads up to a lot of performers to know what they're coming into and to know the acoustic feel as well as the visual feel. (P10)"*

Three participants emphasised its potential for training or learning purposes, including teaching users about acoustic response or staging. Finally, three participants noted that the GSA could be used simply for opportunities to perform in larger venues for enjoyment.

> *"I think it actually could be really good for training people how to use the stage as well" (P5)*

> *"I think it could be cool for people for fun as well that might never have a chance to do a performance like that to see what it's like" (P7)*

6.6 Discussion

Our system for capturing and reproducing a real theatre's appearance and acoustics in VR proved to be both compelling and beneficial for singers in their vocal practice. Our main hypothesis that participants would report a more positive response and enhanced vocal performance when performing with real-time vocal auralisation **(H1)** was supported by the thematic analysis of participants' interviews. In particular, by Theme A, where participants consistently described a positive mindset shift when immersed in the GSA scene with real-time vocal auralisation. This shift was supported by emotional responses and vocal production adjustments resulting from the perceived acoustic quality of their voice, as occurs in physical spaces [234, 252]. These findings, consistent with previous work on vocal auralisation [12, 34], highlight the importance of multisensory feedback in IVEs and its relevance for designing immersive experiences that support singing and vocal performance.

Simulation training provides singers with a safe environment to rehearse, supporting skill development and reducing related symptoms [180, 181]. Therefore, we also hypothesised that the GSA virtual environment would be perceived as a plausible performance venue, helping singers connect their practice and performance experiences **(H2)**. This hypothesis was supported by Theme C, as participants perceived the GSA environment as a plausible representation of a theatre space. This was attributed to the realistic visual and auditory stimuli, as well as coherent multimodal integration, which allowed participants to feel highly immersed, thus experiencing both emotional and vocal changes. Furthermore, Theme E identified the experience as a valuable tool for rehearsal and training, not only supporting vocal-related practice but also contributing to the development of additional performance skills such as understanding the acoustics of a performance venue, stage management and helping reduce stage fright. These findings reinforce the value of VR applications in music education to help musicians overcome limitations of venue access and performance context, and provide adequate audio-visual conditions and opportunities for both performance training and creative music practice [183, 187].

We further hypothesised that the GSA would provide greater audio-visual realism, owing to its photorealistic rendering capabilities [96], and would evoke higher cognitive and emotional

responses compared to the 3D modelled scene without auralisation. Building upon the positive feedback of participants, **(H3)** was supported by Themes A and C. The GSA environment elicited a range of emotional responses, including excitement, confidence, and performance-related anxiety. Some participants also reported experiencing a heightened sense of focus or a shift into a performance mindset within the GSA. Such shifts could have arisen from the heightened levels of immersion and presence [5], as well as the cognitive and psychological changes enabled by simulated performance environments [181]. Moreover, as outlined in Theme C, participants described the GSA environment as realistic, both in its visual detail and its acoustic qualities. In contrast, the 3DM scene was perceived as a less convincing representation, making it difficult for participants to engage. This is particularly relevant since acoustic cues are critical in vocal performance, and singers are sensitive to room responses [251, 252].

Crucially, coherent audiovisual integration was seen as a key factor in enhancing immersion and overall engagement with the experience. Although participants commented on the visual quality of the 3DM scene, the spatial mismatch between the perceived visuals and the real-space acoustics, due to the lack of vocal auralisation, may have hindered their immersion and engagement. This aligns with broader XR research showing that mismatched modalities reduce presence and user engagement [32, 31].

The motivation for creating a 3D-modelled environment without vocal auralisation was to replicate typical environments used in XR applications where the user's voice is not auralised, even when other sound sources within the scene are. This design choice reflects the common visual-auditory mismatch that can occur under such conditions. However, we acknowledge that including a condition comparing the Gaussian and 3D modelled scenes with and without vocal auralisation could have provided more insights, particularly regarding the 3DM scene. Moreover, it would have allowed for a more focused examination of the role of visual realism, separate from audio realism. Therefore, future work will address this limitation by comparing the perceived levels of immersion, realism, and usability between photorealistic and modelled environments, both with and without vocal auralisation, to offer a more comprehensive understanding of the importance of vocal auralisation in IVEs for vocal-related applications.

The acoustic response of a stage varies depending on the performer's location within the stage [222]. Our final hypothesis was that participants would use movement as a meaningful method for acoustic exploration across the virtual stage **(H4)**. Theme B revealed that most participants assessed the acoustic qualities at different positions and directions by experimenting with different vocal techniques. However, this hypothesis was only partially validated, as participants reported noticeable acoustic variation when changing their orientation within the environment, but none commented on variation when facing the same direction at different stage positions (e.g., facing the audience from downstage vs. from centre stage). These results may be due to the minimal acoustic variation across positions within the exploration area caused by

the limited size of the captured physical space. A larger area could have provided more salient variation and enabled a broader audio-visual exploration of the stage, an observation noted by some participants.

We conducted the evaluation in a small room that resembled the dimensions and acoustic response of spaces where singers typically practice [235]. However, selecting a room that allowed participants to physically walk instead of relying on teleportation could have offered a more natural and gradual acoustic interpolation. Although Theme D identified that most participants found teleportation intuitive and easy to use, some expressed a desire to walk, noting that it would have been a more meaningful way to explore the space compared to the sudden jumps of teleportation.

While the equipment setup used in the evaluation proved reliable, allowing real-time visual rendering and vocal auralisation (see Figure 5.2), it also presented challenges during vocal performance, particularly for participants with less VR experience and those wearing glasses. The weight of the headset resting on the user's cheekbones made it more difficult for participants to perform vocal warm-ups like lip trills and to open their mouths while aiming to increase vocal projection, requiring additional physical effort.

During the interview, some participants mentioned their desire to experience spaces with different acoustic characteristics. Allowing participants to explore virtual spaces of differing size and acoustic features could offer a broader understanding of how environments shape vocal behaviour. For example, experiencing concert halls with varying acoustic signatures, small recital rooms, or churches may reveal how acoustic profiles influence singers' vocal perception, production, and adaptation strategies. Additionally, including non-musical spaces with unique acoustic features, like a large train station, could help expand our understanding of real-time vocal auralisation in diverse contexts.

In addition to the findings drawn from the thematic analysis, this study offers insights into the potential of multimodal captures of historical performance venues for heritage preservation. GS is gaining recognition for creating IVEs in architectural heritage contexts, offering photorealistic captures of visual aesthetics and atmosphere [258]. Combining GS photorealism with real-time auralisation using SIRs represents a novel approach to audio-visual archaeoacoustics and the reconstruction of performance venues, supporting the recreation of both their visual and acoustic identity, and enabling the exploration of acoustic conditions across historical spaces [169, 171, 233]. Based on participants' positive experiences, this work highlights the value of capturing not only the visual character and surrounding soundscape of such spaces, but also their acoustic identity, an essential element of cultural heritage [232, 233].

6.7 Conclusion

The focus of the study was to evaluate the use of audio and visual capture techniques, GS and SIRs, to design a realistic virtual environment that could serve as an effective tool for vocal performance and practice. We compared this environment to a small physical room resembling typical rehearsal spaces, as well as to a 3D-modelled scene without vocal auralisation, reflecting common practice in VR applications where the user's voice is not auralised.

Results from the thematic analysis revealed that participants felt present in the GSA environment, which allowed them to sing and explore the stage both acoustically and visually. They acknowledged the realism of the multimodal rendering and the potential practical uses of real-time vocal auralisation in IVEs as a tool for practice and rehearsal, providing a realistic and plausible representation of performance venues. These findings suggest that audio-visual IVEs can support performers in vocal practice, understanding the spatial distribution and acoustics of real venues, and in addressing other performance-related factors such as stage fright.

The hypotheses and analysis presented here aim to create a broad perspective on participants' experiences and evaluations, providing insights into **RQ3** of this thesis and examining the extent to which realistic, multimodal IVEs and high-fidelity audio-visual representations can serve as meaningful environments for vocal training and performance, as well as tools to address gaps in vocal-related endeavours. Moreover, it highlights the importance of audio-visual integration when designing IVEs to avoid the audio-visual mismatch and room divergence that can occur without coherent auditory feedback. In particular, it underscores the importance of auralising the user's own voice, leveraging it as a self-produced sound and a natural way to perceive the characteristics of an IVE, such as its dimensions and materials, while enhancing immersion and psychological effects and providing the user with a stronger sense of agency and control over their interaction within the environment.

The implications behind this study span beyond singing, placing audio-visual coherence and the auralisation of self-produced sound as key principles of the auditory experience within immersive experiences, and suggesting a framework for designing IVEs that support performative and embodied experiences.

Chapter 7

Conclusion and Future Work

This thesis has presented a series of studies conducted to evaluate the sensory modalities that contribute to the perception of acoustic environments in XR. These experiments provided insights into the role of audio-visual feedback in shaping immersive experiences and generated novel contributions that can guide future research in this area. This chapter provides a summary of the contributions of this thesis, reflections on technical and experimental design, and explores avenues for future work.

7.1 Summary of Contributions

In this thesis, several perceptual, qualitative, and practical contributions were identified across audio-visual perception, real-time auralisation, and self-produced and embodied auditory feedback in IVEs. Together, they provide researchers and developers with useful guidance on integrating auditory feedback, such as users' own voices, into research studies and XR applications.

7.1.1 Perceptual Contributions

Based on the experiment presented in Chapter 5, a primary contribution of this work is the set of results showing the influence of real-time vocal auralisation on the perceived sense of presence and audio-visual quality across different levels of rendering resolution and visual representation in VR and MR. We found that the addition of real-time vocal auralisation in IVEs positively influenced the perceived sense of presence and its subscales (spatial presence, realism and involvement) as well as all other measured outcomes related to audio-visual fidelity and integration. Although the visual and audio feedback did not appear to influence each other, the results highlight the importance of coherent multisensory feedback and the relevance of real-time vocal auralisation in IVEs. Consequently, these findings inform **RQ1**, which examined *how self-produced and embodied auditory feedback affects perceptual constructs and psychological responses in multimodal IVEs, such as immersion and presence*. Thus, they align with the broader understanding that auditory feedback is a key element in how users perceive and respond within immersive experiences.

Interestingly, it was found that the perceptual difference between measured and synthesised auralisation may depend on how critical auditory cues are for the task being performed or on the

level of fidelity required to produce a convincing environment. Results from our experiment imply that participants could notice differences in subjective quality judgments between measured and synthesised auralisation when responding to audio-related questions, with measured auralisation yielding higher ratings for audio naturalness and reverberance. However, this was not the case for the sense of presence and its subscales, where no difference was observed between measured and synthesised auralisation conditions. These findings directly support **RQ2**, concerning *how rendering fidelity across visual and audio feedback influence perceived quality and audio-visual integration in IVES using real-time vocal interaction*. This is consistent with the notion that multimodal perception depends highly on the context [15]. Therefore, we propose that when creating applications where auditory feedback is used to create a realistic environment, such as educational simulations, architectural walkthroughs, or narrative experiences, synthesised SIRs may be sufficient for meaningful multimodal integration. In contrast, for applications where auditory feedback is a fundamental element, such as musical practice, performance, or acoustic training, measured SIRs may be preferable, despite the time and effort required to produce them, as they deliver greater auditory fidelity.

Additionally, results suggest that participants did not perceive MR as having better visual quality than VR, contrary to the expectation that MR would be experienced as having higher rendering fidelity. While Section 5.6 offers possible explanations for these outcomes, the results indicate that VR-based IVEs remain an effective approach for high-fidelity visual representations and may be particularly suited for immersive simulation and training contexts. MR, on the other hand, can serve as a more direct means of interacting with the physical world and could be reserved for scenarios that leverage virtual content to enhance real-world interaction and affordances.

Overall, real-time vocal auralisation has been shown to be a core component of multisensory feedback in IVEs, and the integration of self-generated auditory feedback can promote a sense of presence and elicit psychological responses in multimodal contexts. These findings provide practical implications for the development of audio-visual VR and MR experiences, such as simulation and training applications where voice interaction is central, open up avenues for exploring other perceptual constructs, and provide important considerations for the design of future XR systems.

In addition, the experiment presented in Chapter 3 provided insights into continuous embodied auditory feedback when combined with visuo-motor or visuo-tactile feedback to support a SoE, particularly when embodying challenging avatars such as a virtual animal. Results showed improved ownership when audio and tactile feedback were combined, but not when presented independently. Consistent with previous research, our results showed that audio alone was not sufficient to elicit a SoE. However, the observed audio trends on one ownership measure, together with the positive effect seen in the audio-tactile condition, suggest a promising direction.

These novel findings support **RQ1**, further highlighting the potential of auditory feedback in a multimodal setup for eliciting psychological effects in IVEs. Moreover, they underscore the value of exploring alternative multisensory combinations to support the embodiment of avatars that differ from the user's body schema.

7.1.2 Practical Contributions

Building upon the perceptual contributions, we conducted a qualitative analysis to ground the results and provide valuable insight and practical contributions. The experiment described in Chapter 6 examined how singers experienced the usability and perceptual qualities of an audio-visual representation of a virtual performance space, generated by capturing high-fidelity photorealistic visuals together with the acoustic response of a physical theatre. This was then compared to a commonly used 3D model without real-time auralisation, as well as to a small physical room representing typical rehearsal spaces for singers to assess participants' experience across different environments.

Thematic analysis revealed that participants experienced a positive mindset shift when singing in the GS scene with matching real-time vocal auralisation. This elicited emotional responses such as excitement, confidence, and stage fright, which contributed to greater cognitive involvement in the experience. The shift also led singers to report vocal changes, adjusting their vocal techniques and musical delivery in response to the perceived acoustic quality of their voice. These findings were also driven by the opportunity to explore the acoustic characteristics of the stage from different positions and orientations, and to experiment with vocal techniques while moving around the space to examine acoustic variations. In addition, the photorealistic fidelity of the GSA scene provided a detailed representation of the environment, strongly influencing the perception of realism, while vocal auralisation added spatial authenticity and reinforced immersion. The integration of these modalities was recognised as a key factor in shaping perceived realism, immersion, and overall engagement with the experience. Since MR was also found to be a realistic representation, a direct evaluation between GS and MR would be valuable for examining differences in perceived realism and visual rendering fidelity, to determine how the two representation approaches compare, particularly in terms of audio-visual integration.

These findings contribute to understanding the importance of multimodal integration in IVEs, demonstrating that the combination of real-time auditory feedback and high-fidelity visuals can promote immersion, presence, and plausibility. Moreover, real-time vocal auralisation emerges as an effective means for active exploration and perception within the environment, allowing participants to navigate and interact with the acoustic properties of the space. VOREA builds upon prior system designs that employed measured SIRs for audio-visual reconstruction by introducing translational degrees of freedom within the captured area and adopting a performer-centred perspective that enables first-person vocal interaction, capabilities not supported by

previous systems. As a result, real-time vocal auralisation in IVEs creates opportunities for simulation-based training, enabling singers to develop performance-related skills such as vocal projection, stage management, and reducing performance anxiety. Consequently, it supports both stage performance and music training by connecting practice with performance experiences, highlighting its value for performing arts professionals and educational opportunities. The insights gained from this contribution directly address **RQ3**, which explored *to what extent a high-fidelity audio-visual representation of a virtual performance space provides a plausible and usable environment, as well as a useful tool for vocal practice and performance*. Moreover, they highlight the practical value of incorporating spatial variation in real-time auralisation systems to strengthen agency, engagement, and the overall plausibility of the XR experience.

7.1.3 Technical Contributions

To support the vocal interaction required in the perceptual studies described in Chapters 5 and 6, we designed VOREA, an audio auralisation pipeline for real-time first-person vocal auralisation in multimodal immersive experiences. The system leverages SIRs captured from a performer's perspective to convolve the user's voice with responses corresponding to different positions and orientations within the IVE. VOREA enables users to experience the acoustic response of a virtual environment within an interactive area, supporting both directional and translational degrees of freedom. This provides the users with the agency to explore and interact with the space visually and acoustically through dynamic rendering that reflects their movements. The effect is achieved by dynamically interpolating the convolved output of the input signal and the SIRs, using real-time amplitude interpolation across both rotational and translational dimensions, using 4-way CPP and IWD.

VOREA successfully addresses the limitations of the Unity audio engine and alternative spatialisers, including the lack of independent level control over the direct signal and reverberation, as well as latency. It enables the exploration of stage acoustics from different positions, allowing users to experience distinct acoustic characteristics and supporting audio-visual perception in both multimodal research and creative applications within IVEs. In addition, the development of the pipeline provided practical insights and considerations, such as grid placement for capturing acoustic variation, methods to expand the spatial scope, and alternative rendering formats.

7.2 Limitations

We chose to conduct evaluations using subjective metrics, specifically self-reported questionnaires, to measure perceptual aspects related to SoE, presence and immersion, audio quality, and audio-visual integration. However, additional objective and physiological measures, which have been adopted in XR research [259], such as heart rate, skin conductance, galvanic skin response, and body temperature, would have strengthened the perceptual evaluations and supported correlations

between subjective reports and objective effect measures. Similarly, while vocal adaptations were not within the scope of this thesis, integrating measures of vocal production, such as vocal loudness, SPL, and spectral moments [29], into the real-time vocal evaluations would have complemented the qualitative assessment with quantitative insights into vocal perception and production.

The focus of the study presented in Chapter 6 was on the use of audio and visual capture techniques to design a realistic virtual environment that could serve as an effective tool for vocal performance and practice. We created a 3D modelled scene without vocal auralisation, as this reflects common practice in VR applications where the user's voice is not auralised, resulting in a mismatch between visual and auditory stimuli. However, we acknowledge that including a condition comparing the GS and 3D modelled scenes with and without vocal auralisation could have provided more insights, particularly regarding the 3DM scene. Moreover, it would have allowed for a more focused examination of the role of visual realism.

In addition, technical limitations were identified in the study. Teleportation was found to be functional for navigating the environment, particularly since the experiment included a small physical room as one of the conditions tested, which made natural walking unfeasible. However, participants noted that walking would have offered them a more gradual and natural acoustic exploration without the sudden jumps of teleportation. Moreover, the weight of the HMD made it challenging for participants to perform certain vocal warm-ups, such as lip trills, and to comfortably open their mouths to increase vocal projection, which demanded additional physical effort. Finally, acoustic differences were only noticeable with changes in orientation, not from moving between the front and back of the stage while facing the same direction. This may have been due to the minimal acoustic variation in the area explored and requires further exploration. Nonetheless, the overall outcome was positive, offering participants meaningful involvement despite the noted technical constraints.

The development of VOREA arises from the limitations of existing audio spatialisers, which restricted their use in conducting perceptual studies that provide real-time vocal auralisation as an integral element of the perceptual experience. VOREA successfully enabled directional and translational interpolation, allowing perceptual studies on audio-visual perception and rendering fidelity, as well as evaluations of audio and visual realism and the usability of a VR-based virtual theatre designed for vocal performance and practice. While the system was tested and positively assessed by several audio researchers, a formal evaluation would have provided deeper insights into the perceptual validity and quality of the interpolated signal, as well as an objective comparison between measured and synthesised auralisation. Besides, the computational load resulting from convolving the input signal with a large number of SIRs proved challenging, particularly when running the VR application simultaneously. Conducting a comparative analysis of alternative formats such as first-order Ambisonics or binaural SIRs, together with different

interpolation techniques, may have helped identify an approach that better balances perceptual quality with computational performance.

7.3 Future Work

Rich and congruent multisensory feedback has been shown to be essential for eliciting perceptual constructs and illusions, such as the sense of embodiment within IVEs. Building on this, Chapter 3 examined whether continuous audio feedback could support the SoE when embodying a virtual dog. The study used embodied auditory cues as an alternative to the visuo-motor and visuo-tactile feedback typically used for avatar embodiment, and explored their potential to elicit the sensation of having fur. While the hypotheses were not validated, tendencies of audio influence were observed, indicating a promising direction for further research. Subsequently, Chapter 5 revealed higher presence ratings in conditions where real-time vocal auralisation was included, underscoring the relevance of self-produced auditory feedback in shaping perceptual quality and realism in multimodal IVEs, which could also extend to other perceptual effects. Additionally, since our voice provides a sense of agency and influences identity representation, it constitutes a promising form of auditory feedback for supporting the embodiment of dysmorphic avatars, including animals. This includes investigating the extent to which our voice, and its interaction with the environment, promotes the illusion of possessing, controlling, and inhabiting a virtual body, and whether vocal manipulations can elicit perceived changes in body representation in XR experiences.

Results outlined in Chapter 5 showed no significant differences between measured and synthesised SIRs, suggesting that using well-crafted synthesised SIRs may serve as an effective alternative depending on the type of experience required and the level of auditory realism, while also circumventing the limitations associated with producing measured SIRs. Moreover, VOREA was designed to auralise the input signal using convolution with SIRs. These SIRs, whether measured or synthesised, provide an accurate static auralisation. This means that even though interpolation produced perceptual variation in the output, the acoustic characteristics of the SIRs could not be modified, as they were set during their creation. Based on the above, two avenues of investigation could be examined in future work. First, since our study only compared two sound auralisation techniques, future research could assess additional approaches to real-time vocal interaction, such as Scattering Delay Networks (SDNs) or acoustic ray tracing for irregular meshes in VR. Second, with the advent of procedural and generative tools and advances in XR technology, such as mesh generation from spatial captures that represent the physical environment in MR, new opportunities for acoustic modelling are emerging. Therefore, future work could conduct perceptual evaluations of virtual auralisation generated from the estimation of room acoustics directly from 3D mesh data [260], or generative approaches such as the Multimodal Neural Acoustic Fields (MNAF) framework [261].

GS is becoming a commonplace technique for capturing photorealistic physical spaces and objects. However, high-fidelity and large-scale captures demand significant computational resources due to the large number of splats required for rendering. In the GSA scene produced for the study presented in Chapter 6, the high-quality capture of the physical theatre generated over 4.5 million splats, resulting in substantial GPU memory usage. This required us to compromise and find a balance between visual quality and computational load that would allow the environment to be rendered at a consistent 90 fps in the HMD. Recently, alternative high-fidelity radiance field rendering methods and algorithms have emerged, such as Triangle Splatting [262] and Sparse Voxel Rasterization [263], aiming to optimise computational performance without compromising visual fidelity. Perceptual studies that compare novel rendering methods in terms of visual fidelity, perceived immersion, realism, and usability could help identify those that provide the best visual results while requiring fewer computational resources, making this a valuable avenue for future research. Moreover, these studies could be extended to multimodal evaluations that combine the auralisation techniques discussed above, in order to better understand their combined impact in XR.

The findings on real-time vocal auralisation identified in this work can be interpreted through action–perception conceptual frameworks [264]. Real-time vocal auralisation enables users to actively interact and perceive the acoustic response of the IVE using their voice, producing an action–perception coupling [264]. This framework proposes a bidirectional signalling cascade, where top-down predictions of what the user expects based on their vocal production interact with bottom-up prediction errors generated by the perceived acoustic response [265]. Therefore, vocal auralisation could be considered effective when the acoustic response of the virtual environment aligns with the user's internal reference, formed from expectations based on previous experiences of the acoustic response associated with a given environment and vocal effort, thus reducing the cognitive effort required to interpret the environment [41]. This interaction also provides users with a sense of agency, as it creates clear sensorimotor contingencies in which vocal production acts as the cause and the auditory feedback of the environment the effect [177, 264]. Such sensorimotor engagement may contribute to the experience of physical presence, as users are able to move and interact with the environment and perceive the resulting sensory changes [41]. When acoustic variations in the environment respond coherently to the user's vocal actions, the space could feel more plausible and perceptually realistic, as the auditory behaviour of the environment aligns with the user's expectations of how sound should behave in a comparable real-world space [177, 41]. Future studies could validate the plausibility of vocal auralisation systems to evaluate how closely the virtual acoustic stimuli align with users' internal references and how predictive processing mechanisms influence the perception of plausibility and realism. In addition, they could examine how variations in the contingency between vocal actions and acoustic feedback influence perceptual constructs in immersive virtual environments.

Finally, the findings reported in Chapter 6 suggest the potential of multimodal captures of historical performance venues in the context of heritage preservation. Based on participants' positive experiences, the study highlights the value of capturing not only the visual character and surrounding soundscape of such spaces, but also their acoustic identity, an essential element of cultural heritage [232, 233]. Combining GS photorealism with real-time auralisation using SIRs represents a novel approach to audio-visual archaeoacoustics and the reconstruction of performance venues, supporting the recreation of both their visual and acoustic features, and enabling the exploration of acoustic conditions across historical spaces [169, 171, 233]. Future work could explore the production of audio-visual IVEs, particularly of performance venues that allow for first-person vocal interaction. Such environments would contribute to the preservation of acoustic sites by capturing how performers perceive the stage's acoustic response during performances. Moreover, they could open the possibility for singers to access spaces of restricted access and provide opportunities for performance training, creative music practice, and collaborative scenarios for ensemble experiences in XR.

7.4 Closing Statement

This thesis investigated the effect of auditory feedback within a multimodal context on the perceptual experience of IVEs. The results highlight auditory feedback, including self-produced sounds and embodied auditory feedback, as an essential component in shaping perceptual constructs and psychological responses such as presence, which support meaningful immersive experiences (RQ1). Moreover, using the voice as a primary means of acoustic exploration allowed participants to perceive differences in rendering fidelity, directly influencing judgments of audio and visual quality as well as audio-visual integration in IVEs (RQ2). Thus, realistic audio-visual IVEs demonstrate their potential as valuable tools for vocal-related purposes in musical contexts, supporting skill development, rehearsal, and performance (RQ3). While methodological and technical limitations were identified, this work provided valuable insights into audio-visual multimodality in XR and underscored the importance of real-time auralisation for enhancing the design of IVEs. Our contributions highlight implications for the design of multimodal experiences and provide valuable perspectives for XR, HCI, performing arts, and multimodal interaction research, offering opportunities for future developments in performance, education, and interaction within XR.

Bibliography

[1] M. Gao, N. Ruan, J. Shi, and W. Zhou, "Deep neural network for 3D shape classification based on mesh feature," *Sensors*, vol. 22, no. 18, 2022.

[2] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Transactions on Information and Systems*, vol. E77-D, no. 12, pp. 1321–1329, 1994.

[3] A. Krekhov, S. Cmentowski, and J. Krüger, "The illusion of animal body ownership and its potential for virtual reality games," in *2019 IEEE Conference on Games (CoG)*, 2019, pp. 1–8.

[4] A. Sierra Rativa, M. Postma, and M. van Zaanen, "Can virtual reality act as an affective machine? the wild animal embodiment experience and the importance of appearance," in *MIT LINC 2019*, vol. 3, 2020, pp. 214–223. [Online]. Available: https://jwel.mit.edu/linc-2019

[5] A. C. Kern and W. Ellermeier, "Audio in VR: Effects of a soundscape and movement-triggered step sounds on presence," *Frontiers in Robotics and AI*, vol. 7, p. 20, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.00020/full

[6] K. A. Buetler, J. Penalver-Andres, O. Ozen, L. Ferriroli, R. M. Müri, D. Cazzoli, and L. Marchal-Crespo, ""tricking the brain" using immersive virtual reality: Modifying the self-perception over embodied avatar influences motor cortical excitability and action initiation," *Frontiers in Human Neuroscience*, 2022.

[7] Y.-H. Huang, R. Venkatakrishnan, R. Venkatakrishnan, S. V. Babu, and W.-C. Lin, "Using audio reverberation to compensate distance compression in virtual reality," in *ACM Symposium on Applied Perception 2021*. Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3474451.3476236

[8] T. Potter, Z. Cvetković, and E. D. Sena, "On the relative importance of visual and spatial audio rendering on VR immersion," *Frontiers in Signal Processing*, 2022.

[9] B. Burnett, A. Neidhardt, Z. Cvetković, H. Hacıhabiboğlu, and E. De Sena, "User expectation of room acoustic parameters in virtual reality environments," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2023.

[10] S. Fargeot, M. Aramaki, and R. Kronland-Martinet, "Exploring visual–auditory coherence in virtual environments: A perceptual study," in *Proceedings of the AES 5th International Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA: Audio Engineering Society, August 19–21 2024.

[11] A. R. Bargum, D. Kandpal, O. I. Kristjansson, S. Rostami Mosen, J. Andersen, and S. Serafin, "Virtual reconstruction of a the ambisonic concert hall of the royal danish academy of music," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, 2021, pp. 99–102.

[12] S. Ppali, V. Lalioti, B. Branch, C. S. Ang, A. Thomas, B. S. Wohl, and A. Covaci, "Keep the VRhythm going: A musician-centred study investigating how Virtual Reality can support creative musical practice," in *Conference on Human Factors in Computing Systems*. ACM, 2022.

[13] Genelec, "Simulation data files," 2025, accessed: 2025-08-19. [Online]. Available: https://www.genelec.com/simulation-files

[14] C. Pörschmann, "One's own voice in auditory virtual environments," *Acta Acustica united with Acustica*, vol. 87, 2001.

[15] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano, "Multimodality in VR: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, 2022. [Online]. Available: https://doi.org/10.1145/3508361

[16] A. Tajadura-Jiménez, A. Väljamäe, F. Bevilacqua, and N. Bianchi-Berthouze, "Principles for Designing Body-Centered Auditory Feedback," in *The Wiley Handbook of Human Computer Interaction*. Wiley, 2017, pp. 371–403.

[17] A. Tajadura-Jiménez, P. Larsson, A. Väljamäe, D. Västfjäll, and M. Kleiner, "When room size matters: Acoustic influences on emotional responses to sounds," *Emotion (Washington, D.C.)*, vol. 10, pp. 416–22, 2010.

[18] M. Slater and M. V. Sanchez-Vives, "Enhancing our lives with immersive virtual reality," *Front. Robot. AI*, vol. 3, Dec 2016.

[19] S. Schütze and I. Schütze, *New Realities in Audio: A Practical Guide for VR, AR, MR and 360 Video.*, 1st ed. Boca Raton: CRC Press, 2018.

[20] M. Geronazzo and S. Serafin, *Sonic Interactions in Virtual Environments*. Springer VS, 2022.

[21] F. Kiss, S. Mayer, and V. Schwind, "Audio VR: Did video kill the radio star?" *Interactions*, vol. 27, no. 3, 2020.

[22] R. N. Meghanathan, P. Ruediger-Flore, F. Hekele, J. Spilski, A. Ebert, and T. Lachmann, "Spatial sound in a 3D virtual environment: All bark and no bite?" *Big Data and Cognitive Computing*, vol. 5, 2021.

[23] R. Nordahl and N. C. Nilsson, "The sound of being there: Presence and interactive audio in immersive virtual reality," in *The Oxford Handbook of Interactive Audio*. Oxford University Press, 05 2014.

[24] A. C. Naef, M. M. Jeitziner, S. E. Knobel, M. T. Exl, R. M. Müri, S. M. Jakob, T. Nef, and S. M. Gerber, "Investigating the role of auditory and visual sensory inputs for inducing relaxation during virtual reality stimulation," *Scientific Reports*, vol. 12, 2022.

[25] W. Y. Lin, Y. C. Wang, D. R. Wu, R. Venkatakrishnan, R. Venkatakrishnan, E. Ebrahimi, C. Pagano, S. V. Babu, and W. C. Lin, "Empirical evaluation of calibration and long-term carryover effects of reverberation on egocentric auditory depth perception in VR," *Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2022*, 2022.

[26] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang, "Generic hrtfs may be good enough in virtual reality. improving source localization through cross-modal plasticity," *Frontiers in Neuroscience*, vol. 12, 2018.

[27] C. Kleinbeck, H. Schieber, K. Engel, R. Gutjahr, and D. Roth, "Multi-layer gaussian splatting for immersive anatomy visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 5, pp. 2353–2363, 2025.

[28] T. Sierra-Polanco, L. C. Cantor-Cutiva, E. J. Hunter, and P. Bottalico, "Changes of voice production in artificial acoustic environments," *Frontiers in Built Environment*, vol. 7, 2021.

[29] U. Daşdöğen, S. N. Awan, P. Bottalico, A. Iglesias, N. Getchell, and K. V. Abbott, "The influence of multisensory input on voice perception and production using immersive virtual reality," *Journal of Voice*, p. S0892199723002357, 2023.

[30] A. Arabi, M. Tarameshlu, R. Behroozmand, and L. Ghelichi, "Correlation between auditory-perceptual parameters and acoustic characteristics of voice in theater actors," *Middle East Journal of Rehabilitation and Health Studies*, vol. 10, no. 1, 2023.

[31] N. Khenak, J.-M. Vézien, D. Thery, and P. Bourdot, "Spatial presence in real and remote immersive environments and the effect of multisensory stimulation," *PRESENCE: Virtual and Augmented Reality*, vol. 27, pp. 287–308, 2018.

[32] P. Larsson and D. Västfjäll, "When what you hear is what you see: Presence and auditory-visual integration in virtual environments," *Proceedings of the 10th Annual International Workshop on Presence*, 2007.

[33] J. Brereton, "Singing in space(s): Singing performance in real and virtual acoustic environments: Singers' evaluation, performance analysis and listeners' perception," Ph.D. dissertation, University of York, 2014.

[34] G. Kearney, H. Daffern, L. Thresh, H. Omodudu, C. Armstrong, and J. Brereton, "Design of an interactive virtual reality system for ensemble singing," *Proceedings of the Interactive Audio Systems Symposium*, 2016.

[35] P. Luizard, J. Steffens, and S. Weinzierl, "Adaptation of singers to physical and virtual room acoustics," in *Proceedings of the International Symposium on Room Acoustics*, 09 2019.

[36] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, p. 3549–3557, 2009.

[37] F. Avanzini, "Procedural modeling of interactive sound sources in virtual reality," in *Sonic Interactions in Virtual Environments*, M. Geronazzo and S. Serafin, Eds. Springer VS, 2022, ch. 2, pp. 51–80.

[38] E. Sikström, A. de Götzen, and S. Serafin, "The role of sound in the sensation of ownership of a pair of virtual wings in immersive VR," in *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*, 2014.

[39] J.-L. Lugrin, D. Obremski, D. Roth, and M. E. Latoschik, "Audio feedback and illusion of virtual body ownership in mixed reality," in *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, 2016, p. 309–310.

[40] Y. Chandio, N. Bashir, V. Interrante, and F. M. Anwar, "Investigating the correlation between presence and reaction time in mixed reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 9, pp. 5976–5992, 2024.

[41] H. Lee, "A unified conceptual model of immersive experience in extended reality," *Computers in Human Behavior Reports*, vol. 18, p. 100663, 2025.

[42] S. R. Agrawal and S. Bech, *Immersion in Audiovisual Experiences*. Springer International Publishing, 2023, ch. 11.

[43] G. Kearney, H. Daffern, P. Cairns, A. Hunt, B. Lee, J. Cooper, P. Tsagkarakis, T. Rudzki, and D. Johnston, "Measuring the acoustical properties of the bbc maida vale recording studios for virtual reality," *Acoustics*, vol. 4, no. 3, pp. 783–799, 2022.

[44] B. Mróz, M. Kabaciński, T. Ciotucha, A. Rumiński, and T. Żernicki, "Production of six-degrees-of-freedom (6dof) navigable audio using 30 ambisonic microphones," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021.

[45] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio—the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, 2015.

[46] D. Begault, *3-D Sound For Virtual Reality and Multimedia*. Academic Press Professional, Inc., 09 2001.

[47] J. Yang, Y. Frank, and G. Sörös, "Hearing is believing: Synthesizing spatial audio from everyday objects to users," in *Proceedings of the 10th Augmented Human International Conference 2019*. Association for Computing Machinery, 2019.

[48] T. Letowski and S. Letowski, "Auditory spatial perception: Auditory localization," 05 2012.

[49] E. Bates, B. Bridges, and A. Melvin, "Sound spatialization," in *Foundations in Sound Design for Interactive Media*. Routledge, Jun. 2019, pp. 141–160.

[50] L. Picinali and B. F. G. Katz, *System-to-User and User-to-System Adaptations in Binaural Audio*. Springer International Publishing, 2023, pp. 115–143.

[51] N. Raghuvanshi and H. Gamper, *Interactive and Immersive Auralization*. Springer International Publishing, 2023, pp. 77–113.

[52] J. Broderick, J. Duggan, and S. Redfern, "The importance of spatial audio in modern games and virtual environments," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*, 2018.

[53] M. Geronazzo, J. Y. Tissieres, and S. Serafin, "A minimal personalization of dynamic binaural synthesis with mixed structural modeling and scattering delay networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[54] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound externalization: A review of recent research," *Trends in Hearing*, vol. 24, p. 2331216520948390, 2020.

[55] T. Rummukainen, Olli S.and Robotham, A. Plinge, F. Wefers, J. Herre, and E. A. Habets, "Listening tests with individual versus generic head-related transfer functions in six-degrees-of-freedom virtual reality," *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019 ; 5th International Conference on Spatial Audio*, 2019.

[56] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, 1993.

[57] C. Guezenoc and R. Séguier, "Hrtf individualization: A survey," *Preprint/Survey*, 2020.

[58] O. S. Rummukainen, T. Robotham, and E. A. P. Habets, "Head-related transfer functions for dynamic listeners in virtual reality," *Applied Sciences*, vol. 11, no. 14, 2021.

[59] J. Y. Hong, J. He, B. Lam, R. Gupta, and W.-S. Gan, "Spatial audio for soundscape design: Recording and reproduction," *Applied Sciences*, vol. 7, no. 6, 2017.

[60] Rafaely, Boaz, Tourbabin, Vladimir, Habets, Emanuel, Ben-Hur, Zamir, Lee, Hyunkook, Gamper, Hannes, Arbel, Lior, Birnie, Lachlan, Abhayapala, Thushara, and Samarasinghe, Prasanga, "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges," *Acta Acust.*, vol. 6, 2022.

[61] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Applied Sciences*, vol. 7, no. 5, 2017.

[62] K. Brandenburg, B. Fiedler, G. Fischer, F. Klein, A. Neidhardt, C. Schneiderwind, U. Sloma, C. Stirnat, and S. Werner, "Perceptual aspects in spatial audio processing," in *23rd International Congress on Acoustics, Aachen 2019*, 09 2019.

[63] D. R. Begault, S. R. Ellis, and E. M. Wenzel, "Headphone and head-mounted visual displays for virtual environments," *Journal of the Audio Engineering Society*, no. 15-019, October 1998.

[64] *Listening, Corporeality and Presence*. Centre for Interaction Design, School of Computing, Napier University, Edinburgh, 10 2007.

[65] P. Ulsamer, K. Pfeffel, and N. H. Müller, "Brain activation in virtual reality for attention guidance," in *Learning and Collaboration Technologies: Human and Technology Ecosystems*, P. Zaphiris and A. Ioannou, Eds. Cham: Springer International Publishing, 2020, pp. 190–200.

[66] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, "Sonic interactions in virtual reality: State of the art, current challenges, and future directions," *IEEE Computer Graphics and Applications*, vol. 38, no. 2, pp. 31–43, 2018.

[67] R. Nordahl, "Self-induced footsteps sounds in virtual reality: Latency, recognition, quality and presence," in *The 8th Annual International Workshop on Presence, PRESENCE 2005, Conference Proceedings, 21-23 September 2005, London, United Kingdom*. Department of Computer Science, Aalborg University, 2005, pp. 353–355.

[68] M. Vorländer, "Auralization. fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality by michael vorländer," *Acta Acustica united with Acustica*, vol. 94, 2009.

[69] A. J. Kolarik, B. C. J. Moore, S. Cirstea, E. Aggius-Vella, M. Gori, C. Campus, and S. Pardhan, "Factors affecting auditory estimates of virtual room size: Effects of stimulus, level, and reverberation," *Perception*, vol. 50, no. 7, pp. 646–663, 2021.

[70] H. M. Hoppe, S. van de Par, V. L. Flanagin, and S. D. Ewert, "Combined assessment of auditory distance perception and externalization," *The Journal of the Acoustical Society of America*, vol. 157, no. 6, pp. 4649–4666, 06 2025.

[71] S. Li, R. Schlieper, and J. Peissig, "The role of reverberation and magnitude spectra of direct parts in contralateral and ipsilateral ear signals on perceived externalization," *Applied Sciences*, vol. 9, no. 3, 2019.

[72] A. Rungta, N. Rewkowski, R. Klatzky, M. Lin, and D. Manocha, "Effects of virtual acoustics on dynamic auditory distance perception," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, 04 2017.

[73] F. Pind Jörgensson, C.-H. Jeong, H. Llopis, K. Kosikowski, and J. Strømann-Andersen, "Acoustic virtual reality – methods and challenges," in *Proceedings of BNAM 2018*, 2018.

[74] F. N. Khairul Anuar, R. Sulaiman, N. Che Din, and A. S. Abdul Razak, "A conceptual framework for immersive acoustic auralisation: Investigating the key attributes," *Journal of Physics: Conference Series*, vol. 2721, no. 1, mar 2024.

[75] L. Aspöck and M. Vorländer, "Differences between measured and simulated room impulse responses," *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 265, no. 4, 2023. [Online]. Available: http://dx.doi.org/10.3397/IN_2022_0451

[76] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[77] P. Guidorzi, L. Barbaresi, D. D'Orazio, and M. Garai, "Impulse responses measured with mls or swept-sine signals applied to architectural acoustics: An in-depth analysis of the two methods and some case studies of measurements inside theaters," *Energy Procedia*, vol. 78, pp. 1611–1616, 2015, 6th International Building Physics Conference, IBPC 2015.

[78] A. Oliveira, G. Campos, P. Dias, D. T. Murphy, J. Viera, C. Mendonça, and J. Santos, "Real-time dynamic image-source implementation for auralisation," in *Proceedings of the 16th International Conference on Digital Audio Effects*, Sep. 2013, pp. 368–372.

[79] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 08 2015.

[80] V. Välimäki, J. Parker, L. Savioja, J. O. Smith, and J. Abel, "More Than 50 Years of Artificial Reverberation," *Journal of the Audio Engineering Society*, jan 2016.

[81] M. Scerbo, L. Savioja, and E. De Sena, "Room acoustic rendering networks with control of scattering and early reflections," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3745–3758, 2024.

[82] H. Bai, G. Richard, and L. Daudet, "Late reverberation synthesis: from radiance transfer to feedback delay networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, p. 2260–2271, 2015.

[83] J. Sauer, "Acoustic virtual reality: An introductory research for application," Ph.D. dissertation, Delft University of Technology, 2020.

[84] I. Rakkolainen, A. Farooq, J. Kangas, J. Hakulinen, J. Rantala, M. Turunen, and R. Raisamo, "Technologies for multimodal interaction in extended reality—a scoping review," *Multimodal Technologies and Interaction*, vol. 5, no. 12, 2021.

[85] B. Wen, H. Xie, Z. Chen, F. Hong, and Z. Liu, "3D scene generation: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2505.05474

[86] X. Tang, R. Li, and X. Fan, "Recent advance in 3D object and scene generation: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2504.11734

[87] Z. Wang, "3D representation methods: A survey," *arXiv preprint arXiv:2410.06475*, 2024. [Online]. Available: https://arxiv.org/abs/2410.06475

[88] M. Botsch, M. Pauly, L. Kobbelt, P. Alliez, and B. Levy, "Geometric Modeling Based on Polygonal Meshes," in *Eurographics 2008 - Tutorials*. The Eurographics Association, 2008.

[89] Y. Yu, F. Yang, H. Liu, and W. Zhang, "Perceptual quality and visual experience analysis for polygon mesh on different display devices," *IEEE Access*, vol. 6, pp. 42 941–42 949, 2018.

[90] M. Berger, A. Tagliasacchi, L. M. Seversky, P. Alliez, G. Guennebaud, J. A. Levine, A. Sharf, and C. T. Silva, "A survey of surface reconstruction from point clouds," *Computer Graphics Forum*, vol. 36, no. 1, pp. 301–329, 2017.

[91] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[92] K. Li, M. Masuda, S. Schmidt, and S. Mori, "Radiance fields in xr: A survey on how radiance fields are envisioned and addressed for xr research," 2025. [Online]. Available: https://arxiv.org/abs/2508.04326

[93] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 165–174.

[94] G. Chen and W. Wang, "A survey on 3D gaussian splatting," 2025. [Online]. Available: https://arxiv.org/abs/2401.03890

[95] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3D gaussian splatting as a new era: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 8, pp. 4429–4449, 2025.

[96] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[97] Z. Cai, J. Yang, T. Wang, H. Huang, and Y. Guo, "3D reconstruction of buildings based on 3D gaussian splatting," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-4/W10-2024, pp. 37–43, 2024.

[98] P. Cascarano, J. Meglioraldi, G. Vallasciani, V. Armandi, G. Augello, S. Carradori, S. Hajahmadi, and G. Marfia, "A comparative analysis of 3D modeling methods for integration into an extended reality platform," in *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, 2025, pp. 213–217.

[99] S. Qiu, B. Xie, Q. Liu, and P. Heng, "Creating virtual environments with 3D gaussian splatting: A comparative study," in *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2025, pp. 1332–1333.

[100] A. L. Alexander, T. Brunyé, J. Sidman, and S. A. Weil, "From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games," in *The Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. NTSA, 2005.

[101] J. Wang, R. Shi, Z. Xiao, X. Qin, and H.-N. Liang, "Resolution tradeoff in gameplay experience, performance, and simulator sickness in virtual reality games," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022, pp. 542–543.

[102] J. Wang, R. Shi, W. Zheng, W. Xie, D. Kao, and H.-N. Liang, "Effect of frame rate on user experience, performance, and simulator sickness in virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2478–2488, 2023.

[103] M. Slater, P. Khanna, J. Mortensen, and I. Yu, "Visual realism enhances realistic response in an immersive virtual environment," *IEEE Computer Graphics and Applications*, vol. 29, no. 3, pp. 76–84, 2009.

[104] G. S. R. Reddy, "Captivating the senses: Crafting a multisensory virtual experience for enhanced realism," in *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (Ubi-Comp/ISWC '23 Adjunct)*. ACM, October 2023.

[105] M. Slater, "A note on presence terminology," *Presence Connect*, vol. 3, 01 2003.

[106] M. Slater and S. Wilbur, "A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, 12 1997.

[107] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 3, pp. 225–240, 06 1998.

[108] J. H. Murray, *Hamlet on the Holodeck: The Future of Narrative in Cyberspace (Updated Edition)*, 2017th ed. Cambridge, MA: The MIT Press, 2017.

[109] M. Slater, D. Banakou, A. Beacco, J. Gallego, F. Macia-Varela, and R. Oliva, "A separate reality: An update on place illusion and plausibility in virtual reality," *Frontiers in Virtual Reality*, vol. 3, 2022.

[110] K. Shapiro and A. Steed, "Context-relevant locations as an alternative to the place illusion in augmented reality," in *Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology*. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3641825.3687745

[111] M. E. Latoschik and C. Wienrich, "Congruence and plausibility, not presence: Pivotal conditions for xr experiences and effects, a novel approach," *Frontiers in Virtual Reality*, vol. 3, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2022.694433

[112] F. Westermeier, L. Brübach, M. E. Latoschik, and C. Wienrich, "Exploring plausibility and presence in mixed reality experiences," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2680–2689, 2023.

[113] F. Biocca, "The cyborg's dilemma: Progressive embodiment in virtual environments [1]," *Journal of Computer-Mediated Communication*, vol. 3, no. 2, p. JCMC324, 09 1997.

[114] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks, "Physiological measures of presence in stressful virtual environments," *ACM Trans. Graph.*, vol. 21, no. 3, p. 645–652, 2002.

[115] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, "Affective interactions using virtual reality: The link between presence and emotions," *CyberPsychology & Behavior*, vol. 10, no. 1, pp. 45–56, 2007.

[116] S. Jung and R. W. Lindeman, "Perspective: Does realism improve presence in VR? suggesting a model and metric for VR experience evaluation," *Frontiers in Virtual Reality*, vol. 2, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2021.693327

[117] M. Lombard and T. Ditton, "At the heart of it all: The concept of presence," *Journal of Computer-Mediated Communication*, vol. 3, no. 2, p. JCMC321, 09 1997.

[118] M. Slater, A. Rovira, R. Southern, D. Swapp, J. J. Zhang, C. Campbell, and M. Levine, "Bystander responses to a violent incident in an immersive virtual environment," *PLOS ONE*, vol. 8, no. 1, pp. 1–13, 01 2013.

[119] R. Skarbez, S. Neyret, F. P. Brooks, M. Slater, and M. C. Whitton, "A psychophysical experiment regarding components of the plausibility illusion," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 4, pp. 1369–1378, 2017.

[120] I. Bergström, S. Azevedo, P. Papiotis, N. Saldanha, and M. Slater, "The plausibility of a string quartet performance in virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 4, pp. 1352–1359, 2017.

[121] K. Kilteni, R. Groten, and M. Slater, "The sense of embodiment in virtual reality," *Presence*, vol. 21, no. 4, pp. 373–387, 2012.

[122] N. Braun, S. Debener, N. Spychala, E. Bongartz, P. Sörös, H. H. O. Müller, and A. Philipsen, "The senses of agency and ownership: A review," *Frontiers in Psychology*, vol. 9, 2018.

[123] R. Fribourg, F. Argelaguet, A. Lécuyer, and L. Hoyet, "Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 2062–2072, 2020.

[124] L. Lin and S. Jörg, "Need a hand? how appearance affects the virtual hand illusion," in *Proceedings of the ACM Symposium on Applied Perception*, 2016, p. 69–76.

[125] A. Maselli and M. Slater, "The building blocks of the full body ownership illusion," *Frontiers in Human Neuroscience*, vol. 7, 2013.

[126] M. Slater, B. Spanlang, a. V. Sanchez-Vives, and O. Blanke, "First person experience of body transfer in virtual reality," *PLoS ONE*, vol. 5, no. 5, 2010.

[127] I. de Villiers Bosman, M. Bujić, L. D. Cosio, O. Buruk, K. Jørgensen, and J. Hamari, "Embodiment in virtual reality: an experiment on how visual and aural first- and third-person modes affect embodiment and mindfulness," *Virtual Reality*, vol. 29, no. 49, mar 2025.

[128] M. V. Sanchez-Vives, B. Spanlang, A. Frisoli, M. Bergamasco, and M. Slater, "Virtual hand illusion induced by visuomotor correlations," *PLoS ONE*, vol. 5, no. 4, 2010.

[129] M. Guy, J.-M. Normand, C. Jeunet-Kelway, and G. Moreau, "The sense of embodiment in virtual reality and its assessment methods," *Frontiers in Virtual Reality*, vol. 4, dec 2023. [Online]. Available: https://doi.org/10.3389/frvir.2023.1141683

[130] T. C. Peck, S. Seinfeld, S. M. Aglioti, and M. Slater, "Putting yourself in the skin of a black avatar reduces implicit racial bias," *Consciousness and Cognition*, vol. 22, no. 3, pp. 779–787, 2013.

[131] D. Banakou, R. Groten, and M. Slater, "Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 31, pp. 12 846–12 851, 2013.

[132] A. Cheymol, R. Fribourg, N. Ogawa, A. Lécuyer, Y. Hirao, T. Narumi, F. Argelaguet, and J.-M. Normand, "Studying the role of self and external touch in the appropriation of dysmorphic hands," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2022, pp. 102–111.

[133] J.-L. Lugrin, J. Latt, and M. E. Latoschik, "Avatar anthropomorphism and illusion of body ownership in VR," in *2015 IEEE Virtual Reality (VR)*, 2015, pp. 229–230.

[134] K. Kilteni, J.-M. Normand, M. V. Sanchez-Vives, and M. Slater, "Extending body space in immersive virtual reality: A very long arm illusion," *PLOS ONE*, vol. 7, no. 7, pp. 1–15, 07 2012.

[135] A. S. Won, J. Bailenson, J. Lee, and J. Lanier, "Homuncular flexibility in virtual reality," *Journal of Computer-Mediated Communication*, vol. 20, no. 3, pp. 241–259, 01 2015.

[136] W. Steptoe, A. Steed, and M. Slater, "Human tails: Ownership and control of extended humanoid avatars," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 4, pp. 583–590, 2013.

[137] R. K. Kammerlander, A. Pereira, and S. Alexanderson, "Using virtual reality to support acting in motion capture with differently scaled characters," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 402–410.

[138] S. J. G. Ahn, J. Bostick, E. Ogle, K. L. Nowak, K. T. McGillicuddy, and J. N. Bailenson, "Experiencing Nature: Embodying Animals in Immersive Virtual Environments Increases Inclusion of Nature in Self and Involvement with Nature," *Journal of Computer-Mediated Communication*, vol. 21, no. 6, pp. 399–419, 2016.

[139] B. Laha, J. N. Bailenson, A. S. Won, and J. O. Bailey, "Evaluating control schemes for the third arm of an avatar," *Presence: Teleoperators and Virtual Environments*, vol. 25, no. 2, pp. 129–147, 11 2016.

[140] D. Pimentel and S. Kalyanaraman, "The effects of embodying wildlife in virtual reality on conservation behaviors," *Scientific Reports*, vol. 12, no. 1, 2022.

[141] A. Oyanagi and R. Ohmura, "Conditions for inducing sense of body ownership to bird avatar in virtual environment," *Journal of Computers*, pp. 596–602, 2018.

[142] ——, "Transformation to a bird: Overcoming the height of fear by inducing the proteus effect of the bird avatar," in *Proceedings of the 2nd International Conference on Image and Graphics Processing*, 2019, p. 145–149.

[143] N. Yee, J. N. Bailenson, and N. Ducheneaut, "The proteus effect: Implications of transformed digital self-representation on online and offline behavior," *Communication Research*, vol. 36, no. 2, pp. 285–312, 2009.

[144] S. Li, X. Gu, K. Yi, Y. Yang, G. Wang, and D. Manocha, "Self-illusion: A study on cognition of role-playing in immersive virtual environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 8, pp. 3035–3049, 2022.

[145] A. Krekhov, S. Cmentowski, K. Emmerich, and J. Krüger, "Beyond human: Animals as an escape from stereotype avatars in virtual reality games," in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2019, p. 439–451.

[146] S. Serafin, *Audio in Multisensory Interactions: From Experiments to Experiences*. Cham: Springer International Publishing, 2023, pp. 305–318.

[147] Y. Zhao and Z. Cen, "Exploring multimodal feedback mechanisms for improving user interaction in virtual reality environments," *Journal of Industrial Engineering and Applied Science*, vol. 2, no. 6, pp. 35–41, 2024.

[148] F. Fontana, H. Järveläinen, and S. Papetti, *Augmenting Sonic Experiences Through Haptic Feedback*. Cham: Springer International Publishing, 2023, pp. 353–381.

[149] F. Klein, S. V. A. Gari, J. M. Arend, and P. W. Robinson, "Towards determining thresholds for room divergence: A pilot study on detection thresholds," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021, pp. 1–7.

[150] N. D. Chandrashekar, S. Safford, M. Muniyandi, and D. Gračanin, "An extended reality simulator for pulse palpation training," in *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, IEEE. IEEE, 2023, pp. 178–182.

[151] V. Andonova, F. Reinoso-Carvalho, M. A. Jimenez Ramirez, and D. Carrasquilla, "Does multisensory stimulation with virtual reality (vr) and smell improve learning? an educational experience in recall and creativity," *Frontiers in Psychology*, vol. 14, 2023.

[152] S. Malpica, A. Serrano, M. Allue, M. G. Bedia, and B. Masia, "Crossmodal perception in virtual reality," *Multimedia Tools and Applications*, vol. 79, no. 5, feb 2020.

[153] J. Vroomen and M. Keetels, "Perception of intersensory synchrony: a tutorial review," *Attention, Perception, and Psychophysics*, vol. 72, no. 4, pp. 871–884, 2010.

[154] J. Liu, V. Drga, and I. Yasin, "Optimal time window for the integration of spatial audio-visual information in virtual environments," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 723–728.

[155] M. P. Snow and R. C. Williges, "Empirical models based on free-modulus magnitude estimation of perceived presence in virtual environments," *Human Factors*, vol. 40, no. 3, pp. 386–402, 1998.

[156] H. Dinh, N. Walker, L. Hodges, C. Song, and A. Kobayashi, "Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments," in *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, 1999, pp. 222–228.

[157] C. Hendrix and W. Barfield, "The sense of presence within auditory virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 5, no. 3, pp. 290–301, 08 1996.

[158] M. R. Mahmud, M. Stewart, A. Cordova, and J. Quarles, "Auditory feedback to make walking in virtual reality more accessible," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2022, pp. 847–856.

[159] A. Tajadura-Jiménez, O. Deroy, T. Marquardt, N. Bianchi-Berthouze, T. Asai, T. Kimura, and N. Kitagawa, "Audio-tactile cues from an object's fall change estimates of one's body height," *PLOS ONE*, p. e0199354, 2018.

[160] I. Senna, A. Maravita, N. Bolognini, and C. V. Parise, "The marble-hand illusion," *PLoS ONE*, vol. 9, no. 3, 2014.

[161] A. Tajadura-Jiménez, M. Basia, O. Deroy, M. Fairhurst, N. Marquardt, and N. Bianchi-Berthouze, "As light as your footsteps: Altering walking sounds to change perceived body weight, emotional state and gait," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, p. 2943–2952.

[162] L. Buck, R. Paris, and B. Bodenheimer, "Distance compression in the htc vive pro: A quick revisitation of resolution," *Frontiers in Virtual Reality*, vol. 2, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2021.728667

[163] J. W. Kelly, T. A. Doty, M. Ambourn, and L. A. Cherep, "Distance perception in the oculus quest and oculus quest 2," *Frontiers in Virtual Reality*, vol. 3, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2022.850471

[164] J. Luberadzka, E. Gusó Muñoz, U. Sayin, and A. Garriga, "Audio technology for improving social interaction in extended reality," *Frontiers in Virtual Reality*, vol. 5, Jan. 2025. [Online]. Available: http://dx.doi.org/10.3389/frvir.2024.1442774

[165] B. N. J. Postma and B. F. G. Katz, "Influence of visual rendering on the acoustic judgements of a theater auralization," *Proceedings of Meetings on Acoustics*, vol. 30, no. 1, p. 015008, 09 2017.

[166] M. Frank and D. Perinovic, "Matching auditory and visual room size, distance, and source orientation in virtual reality," in *Proceedings of the 17th International Audio Mostly Conference*. Association for Computing Machinery, 2022, p. 80–83.

[167] P. Rolkowski, P. Odya, and B. Mróz, "Reverberation divergence in VR applications," *International Journal of Electronics and Telecommunications*, vol. 70, no. 2, pp. 373–380, 2024.

[168] D. Poirier-Quinot, B. N. Postma, and B. F. Katz, "Augmented auralization: Complementing auralizations with immersive virtual reality technologies," in *International Symposium on Music and Room Acoustics*, 2016, pp. 1–10.

[169] B. F. Katz, D. Poirier-Quinot, and B. N. Postma, "Virtual reconstructions of the théâtre de l'athénée for archeoacoustic study," in *Conference Paper*, September 2019.

[170] A. Gozzi and A. Guazzini, "Exploring acoustics perception through xr spatial audio experiences: Experiments and data collection for the 'listen to the theatre' project," in *XR Salento 2024*. Springer Nature Switzerland AG, 2024, pp. 185–208.

[171] B. N. Postma, D. Poirier-Quinot, J. Meyer, and B. F. Katz, "Virtual reality performance auralization in a calibrated model of notre-dame cathedral," in *EuroRegio2016*, Porto, Portugal, 2016, pp. 1–10.

[172] S. Poeschl, "Virtual reality training for public speaking-A QUESTVR framework validation," *Frontiers in ICT*, vol. 4, 2017.

[173] M. Slater and M. Usoh, "Body centred interaction in immersive virtual environments," in *Artificial Life and Virtual Reality*, N. M. Thalmann and D. Thalmann, Eds. John Wiley and Sons, 1994, pp. 125–148.

[174] P. Dourish, *Chapter 4: Embodied Interaction*. Cambridge, MA: MIT Press, 2001.

[175] A. Tajadura-Jiménez, J. Newbold, L. Zhang, P. Rick, and N. Bianchi-Berthouze, "As light as you aspire to be: Changing body perception with sound to support physical

activity," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019, p. 1–14. [Online]. Available: https://doi.org/10.1145/3290605.3300888

[176] N. Lavan, A. M. Burton, S. K. Scott, and C. McGettigan, "Flexible voices: Identity perception from variable vocal signals," *Psychonomic Bulletin & Review*, vol. 2, no. 1, pp. 90–102, 02 2019.

[177] R. Ohata, T. Asai, S. Imaizumi, and H. Imamizu, "I hear my voice; therefore i spoke: The sense of agency over speech is enhanced by hearing one's own voice," *Psychological Science*, vol. 33, 2022.

[178] C. Jicol, C. H. Wan, B. Doling, C. H. Illingworth, J. Yoon, C. Headey, C. Lutteroth, M. J. Proulx, K. Petrini, and E. O. Neill, "Effects of emotion and agency on presence in virtual reality," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[179] D. Rodríguez, A. Borrego, M. Guzmán, and R. Llorens, "Acoustic characteristics of voice production in virtual reality-simulated and physical environments: a comparative study in university professors," *Virtual Reality*, vol. 28, no. 2, Mar. 2024.

[180] L. Aufegger, R. Perkins, D. Wasley, and A. Williamon, "Musicians' perceptions and experiences of using simulation training to develop performance skills," *Psychology of Music*, vol. 45, pp. 417–431, 2017.

[181] L. Aufegger and D. Wasley, "Virtual reality feedback influences musicians' physical responses and mental attitude towards performing," *Music and Medicine*, vol. 12, no. 3, pp. 157–166, 2020.

[182] V. Lalioti, S. Ppali, A. Thomas, R. Hrafnkelsson, M. Grierson, C. S. Ang, B. S. Wohl, and A. Covaci, "VR Rehearse & Perform - A platform for rehearsing in Virtual Reality," in *27th ACM Symposium on Virtual Reality Software and Technology*. ACM, 2021.

[183] H. Yi, "Research on the teaching effect of virtual reality technology in simulated vocal performance environment," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, 2024.

[184] L. Zong, "Evaluation on the effect of enhancing vocal music training experience with virtual reality technology," *International Journal of Web-Based Learning and Teaching Technologies*, vol. 20, no. 1, 2025.

[185] G. Dedousis, K. Bakogiannis, A. Andreopoulou, and A. Georgaki, "Room acoustics mismatches of rehearsal spaces and concert halls and their impact on music performance," in *Proceedings of the Institute of Acoustics*, vol. 45. Institute of Acoustics, 2023.

[186] S. Doganyigit and O. F. Islim, "Virtual reality in vocal training: a case study," *Music Education Research*, vol. 23, no. 3, pp. 391–401, 2021.

[187] A. Wu and A. Park, "A novel vocal training application using a virtual reality concert stage and voice scoring algorithm," pp. 93–102, 02 2025.

[188] M. Zhao, "The influence of virtual reality on improving emotional expressiveness in vocal performance," *Acta Psychologica*, vol. 255, p. 104969, April 2025.

[189] T. Di, D. Medeiros, M. Sousa, and T. Grossman, "VRChoir: Exploring remote choir rehearsals via virtual reality," in *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2023, pp. 895–896.

[190] M. Hara, P. Pozeg, G. Rognini, T. Higuchi, K. Fukuhara, A. Yamamoto, T. Higuchi, O. Blanke, and R. Salomon, "Voluntary self-touch increases body ownership," *Frontiers in Psychology*, vol. 6, p. 1509, 2015. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2015.01509

[191] A. Angelidis and B. McCane, "Fur simulation with spring continuum," *The Visual Computer*, vol. 25, no. 3, pp. 255–265, 2009.

[192] T. G. Andersen, V. Falster, J. R. Frisvad, and N. J. Christensen, "Hybrid fur rendering: combining volumetric fur with explicit hair strands," *The Visual Computer*, 2016.

[193] G. Gorisse, O. Christmann, E. A. Amato, and S. Richir, "First- and third-person perspectives in immersive virtual environments: Presence and performance analysis of embodied users," *Frontiers in Robotics and AI*, vol. 4, 2017.

[194] A. O. Bebko and N. F. Troje, "bmltux: Design and control of experiments in virtual reality and beyond," May 2020. [Online]. Available: psyarxiv.com/arvkf

[195] D. Roth and M. E. Latoschik, "Construction of the virtual embodiment questionnaire (veq)," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 12, pp. 3546–3556, 2020.

[196] T. C. Peck and M. Gonzalez-Franco, "Avatar embodiment. a standardized questionnaire," *Frontiers in Virtual Reality*, vol. 1, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2020.575943

[197] R. Fribourg, E. Blanpied, L. Hoyet, A. Lécuyer, and F. Argelaguet, "Does virtual threat harm VR experience?: Impact of threat occurrence and repeatability on virtual embodiment and threat response," *Computers & Graphics*, vol. 100, pp. 125–136, 2021.

[198] D. Dewez, R. Fribourg, F. Argelaguet, L. Hoyet, D. Mestre, M. Slater, and A. Lécuyer, "Influence of personality traits and body awareness on the sense of embodiment in virtual reality," in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2019, pp. 123–134.

[199] K. Rogers, G. Ribeiro, R. R. Wehbe, M. Weber, and L. E. Nacke, "Vanishing importance: Studying immersive effects of game audio perception on player experiences in virtual reality," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2018, p. 1–13.

[200] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175–191, May 2007.

[201] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyzes using G*Power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, pp. 1149–1160, 2009.

[202] J. Zhang and B. Hommel, "Body ownership and response to threat," *Psychological Research*, vol. 80, no. 6, pp. 1020–1029, Aug. 2015. [Online]. Available: https://doi.org/10.1007/s00426-015-0698-1

[203] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 143–146.

[204] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach." *Journal of Personality and Social Psychology*, vol. 44, no. 1, pp. 113–126, Jan. 1983. [Online]. Available: https://doi.org/10.1037/0022-3514.44.1.113

[205] M. Slater and D. Banakou, "The golden rule as a paradigm for fostering prosocial behavior with virtual reality," *Current Directions in Psychological Science*, vol. 30, no. 6, pp. 503–509, 2021.

[206] E. Accolti, L. Aspöck, M. Yadav, and M. Vorländer, "Real-time auralization for performers on virtual stages," 2023.

[207] E. Patricio, "Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields," *AES 146th Convention*, 03 2019.

[208] A. Canclini, L. Mucci, F. Antonacci, A. Sarti, and S. Tubaro, "A methodology for estimating the radiation pattern of a violin during the performance," in *23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1286–1290.

[209] Z. sp. z o.o., *Zylia ZM-1 Microphone Reference Manual v.1.5.0*, n.d., accessed: 2025-02-31. [Online]. Available: https://blog.zylia.co/uploads/8/3/7/1/83717266/zylia_studio_user_manual_v1.5.pdf

[210] B. Braxton, B. Agnieszka, and G. Brian, "Maximum averaged and peak levels of vocal sound pressure," *Journal of the Audio Engineering Society*, no. 8958, October 2013.

[211] P. Bottalico, S. Graetzer, and E. J. Hunter, "Effect of training and level of external auditory feedback on the singing voice: Volume and quality," *Journal of Voice*, vol. 30, no. 4, pp. 434–442, jul 2016.

[212] Voxengo, "Voxengo deconvolver manual," n.d. [Online]. Available: https://www.voxengo.com/doc/deconvolver/

[213] Cockos Incorporated, "Reaper digital audio workstation," https://www.reaper.fm/, 2025, accessed: 2025-02-26.

[214] JB-Luke, "X-mcfx convolver," n.d. [Online]. Available: https://github.com/JB-Luke/X-MCFX

[215] M. Kronlachner, "Ambisonics plug-in suite for production and performance usage," in *Proceedings of the Linux Audio Conference*, Graz, Austria, 2013.

[216] IEM, "Iem plug-in suite," n.d. [Online]. Available: https://plugins.iem.at/

[217] E. Epifani, "Perception of initial time delay gap, reverberation time, and density of impulse response of a room," Master's thesis, Scuola di Ingegneria dell'Informazione, Politecnico Di Milano, 2021.

[218] Unity Technologies, "Unity real-time development platform," https://unity.com/, 2025, accessed: 2025-02-20.

[219] A. Øland, "Loudness concepts & pan laws," 2019, accessed: 2024-06-30.

[220] B. Alary and V. Valimaki, "A method for capturing and reproducing directional reverberation in six degrees of freedom," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021.

[221] L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, "Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple ambisonic receivers," *Journal of the Audio Engineering Society*, 2022.

[222] J. Dammerud and M. Barron, "Concert hall stage acoustics from the perspective of performers and physical reality," *Auditorium Acoustics 2008*, 01 2008.

[223] J. Meyer, *Introduction to Acoustics*. Springer New York, 2009, p. 1–22.

[224] K. Ueno and H. Tachibana, "A consideration on acoustic properties on concert-hall stages," in *Proceedings of the International Symposium on Room Acoustics (ISRA 2010)*, Melbourne, Australia, 2010.

[225] K. Enge, M. Frank, and R. Höldrich, "Listening experiment on the plausibility of acoustic modeling in virtual reality," 4 2020.

[226] C.-Y. Chen, "Delving into hearing threshold of the delay gap of initial reflection in a room by using the response of cortical brainwaves," *Applied Sciences*, vol. 13, no. 21, 2023.

[227] Focusrite Audio Engineering Ltd., "Scarlett 2i2," 2024, accessed: 2025-02-26. [Online]. Available: https://focusrite.com/products/scarlett-2i2

[228] Meta Platforms, Inc., "Meta quest 3," 2024, accessed: 2025-02-26. [Online]. Available: https://www.meta.com/ie/quest/quest-3/

[229] M.-K. Nils, K. Miranda, S. S. J., and L. Tapio, "measuring motion-to-sound latency in virtual acoustic rendering systems," *journal of the audio engineering society*, vol. 71, june 2023.

[230] M. Thomas, M.-K. Nils, H. Christoph, S. S. J., and P. Ville, "Auralization of measured room transitions in virtual reality," *Journal of the Audio Engineering Society*, vol. 71, pp. 326–337, 2023.

[231] C. Zou, S.-Y. Rhee, L. He, D. Chen, and X. Yang, "Sounds of history: A digital twin approach to musical heritage preservation in virtual museums," *Electronics*, vol. 13, no. 12, 2024.

[232] X. Zhu, T. Oberman, and F. Aletta, "Defining acoustical heritage: A qualitative approach based on expert interviews," *Applied Acoustics*, vol. 216, 2024.

[233] B. F. G. Katz, D. Murphy, and A. Farina, "The past has ears (phe): Xr explorations of acoustic spaces as cultural heritage," in *Augmented Reality, Virtual Reality, and Computer Graphics*, L. T. De Paolis and P. Bourdot, Eds., 2020.

[234] Y. G. Redman, J. D. Glasner, D. D'Orazio, and P. Bottalico, "Singing in different performance spaces: The effect of room acoustics on singers' perception," *The Journal of the Acoustical Society of America*, vol. 154, no. 4, pp. 2256–2264, 2023.

[235] B. Knöfel, J. Troge, and L. Weisheit, "Musicians and their practice rooms: What do they think about present room acoustics and what would they prefer?" in *Euronoise 2018: Proceedings of the 11th European Congress and Exposition on Noise Control Engineering*, 2018.

[236] R. Warp, M. Zhu, I. Kiprijanovska, J. Wiesler, S. Stafford, and I. Mavridou, "Moved by sound: How head-tracked spatial audio affects autonomic emotional state and immersion-driven auditory orienting response in VR environments," *AES Europe Spring 2022 - 152nd Audio Engineering Society Convention 2022*, 2022.

[237] P. Larsson, D. Västfjäll, and M. Kleiner, "Ecological acoustics and the multi-modal perception of rooms: real and unreal experiences of auditory-visual virtual environments," *Proceedings of the 2001 International Conference on Auditory Display*, 01 2001.

[238] P. Luizard, N. Henrich Bernardoni, and C. Böhm, "Using virtual acoustics and electroglottography to study the adaptation of singing voice production," in *e-Forum Acusticum,* , Dec. 2020.

[239] J. Hedges, R. Sazdov, and A. Johnston, "Measuring the influence of audio on immersive experience in extended reality and digital games: a systematic review," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2023, pp. 1–13. [Online]. Available: https://ieeexplore.ieee.org/document/10289267/

[240] NA, "Recommendation ITU-t p.910 (10/2023) subjective video quality assessment methods for multimedia applications."

[241] R. L. Hornsey and P. B. Hibbard, "Contributions of pictorial and binocular cues to the perception of distance in virtual reality," *Virtual Reality*, vol. 25, no. 4, p. 1087–1103, Mar. 2021. [Online]. Available: http://dx.doi.org/10.1007/s10055-021-00500-x

[242] *Acoustics -Measurement of room acoustic parameters, Part 1: Performance spaces*, International Organization for Standardization Std. ISO 3382-1:2009(E).

[243] *Acoustics - Measurement of room acoustic parameters, Part 2: Reverberation time in ordinary rooms*, International Organization for Standardization Std. ISO 3382-2:2008(E).

[244] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, pp. 713–720, 2011.

[245] G. Chaurasia, A. Nieuwoudt, A.-E. Ichim, R. Szeliski, and A. Sorkine-Hornung, "Passthrough+: Real-time stereoscopic view synthesis for mobile mixed reality," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 3, no. 1, pp. 1–17, 2020.

[246] F. Vona, J. Schorlemmer, M. Stern, N. Ashrafi, M. Vergari, T. Kojic, and J.-N. Voigt-Antons, "Comparing pass-through quality of mixed reality devices: A user experience study during real-world tasks," 2025. [Online]. Available: https://arxiv.org/abs/2502.06382

[247] G. Kuo, E. Penner, S. Moczydlowski, A. Ching, D. Lanman, and N. Matsuda, "Perspective-correct VR passthrough without reprojection," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3588432.3591534

[248] T. Q. Tran, T. Langlotz, and H. Regenbrecht, "A survey on measuring presence in mixed reality," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. Association for Computing Machinery, 2024.

[249] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[250] D. Banakou and M. Slater, "Embodiment in a virtual body that speaks produces agency over the speaking but does not necessarily influence subsequent real speaking," *Scientific Reports*, vol. 7, Oct 2017.

[251] P. Bottalico, N. Lastowiecka, J. D. Glasner, and Y. G. Redman, "Singing in different performance spaces: The effect of room acoustics on vibrato and pitch inaccuracy," *The Journal of the Acoustical Society of America*, June 2022.

[252] P. Luizard and N. Henrich Bernardoni, "Changes in the voice production of solo singers across concert halls," *J. Acoust. Soc. Am.*, vol. 148, no. 1, p. EL33, 2020.

[253] S. Vlahovic, M. Suznjevic, and L. Skorin-Kapov, "A survey of challenges and methods for quality of experience assessment of interactive VR applications," *Journal on Multimodal User Interfaces*, vol. 16, no. 3, pp. 257–291, Sep. 2022.

[254] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[255] M. Naeem, W. Ozuem, K. Howell, and S. Ranfagni, "A step-by-step process of thematic analysis to develop a conceptual model in qualitative research," *International Journal of Qualitative Methods*, vol. 22, p. 16094069231205789, 2023.

[256] Varjo, "Teleport 3D Scanning: Turn reality into 3D with any camera — teleport.varjo.com," https://teleport.varjo.com/, 2025, accessed: 2025-06-20.

[257] Unity Technologies, "Unity asset store," https://assetstore.unity.com/, 2025, accessed: 2025-04-12.

[258] O. Jamil and A. Brennan, "Immersive heritage through gaussian splatting: a new visual aesthetic for reality capture," *Frontiers in Computer Science*, vol. 7, 2025.

[259] I. de Villiers Bosman, O. Buruk, K. Jørgensen, and J. Hamari, "The effect of audio on the experience in virtual reality: a scoping review," *Behaviour & Information Technology*, vol. 43, no. 1, pp. 165–199, 2024.

[260] L. Kelley, D. Di Carlo, A. A. Nugraha, M. Fontaine, Y. Bando, and K. Yoshii, "RIR-in-a-Box: Estimating Room Acoustics from 3D Mesh Data through Shoebox Approximation," in *Interspeech 2024*, 2024, pp. 3255–3259.

[261] G. Tong, J. C.-H. Leung, X. Peng, H. Shi, L. Zheng, S. Wang, A. C. O'Brien, A. P.-A. Neall, G. Fei, M. Gaspar, and P. Chakravarthula, "Multimodal neural acoustic fields for immersive mixed reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 5, pp. 3397–3407, 2025.

[262] J. Held, R. Vandeghen, A. Deliege, A. Hamdi, A. Cioppa, S. Giancola, A. Vedaldi, B. Ghanem, A. Tagliasacchi, and M. Van Droogenbroeck, "Triangle splatting for real-time radiance field rendering," *arXiv*, 2025.

[263] C. Sun, J. Choe, C. Loop, W.-C. Ma, and Y.-C. F. Wang, "Sparse voxels rasterization: Real-time high-fidelity radiance field rendering," *ArXiv*, vol. abs/2412.04459, 2024.

[264] S. L. Denham and I. Winkler, "Predictive coding in auditory perception: challenges and unresolved questions," *European Journal of Neuroscience*, vol. 51, pp. 1151–1160, 2020.

[265] Y. T. Huang, C.-T. Wu, Y.-X. M. Fang, C.-K. Fu, S. Koike, and Z. C. Chao, "Cross-modal hierarchical predictive coding for audiovisual sequences in the human brain," *Communications Biology*, vol. 7, 2024.