Acoustic Information Retrieval for Interactive Sound Rendering in Virtual Environments

Mattia Colombo

A report submitted as part of the requirements
for the degree of Research to PhD in Computing

Faculty of Computing, Engineering, and the Built Environment

Birmingham City University

May 2025

Supervisors
Dr Carlo Harvey, Dr Maite Frutos-Pascual, Dr Alan Dołhasz

A Karensa, alla mia famiglia, e in memoria della nonna Giovanna

Abstract

The adoption of Extended Reality across industry and research domains has incentivised the development of Head-Mounted Display (HMD) technology, driving the field towards better and more optimal techniques for efficient and realistic rendering, offering sensing and capturing capabilities. Virtual entities visualised through HMDs can interact with spatial features of the users' surroundings, allowing for realistic context-aware interactions and improving task performance and the perceptual quality of the overall immersive experience. However, techniques for rendering realistic audio stimuli that respond to spatial features of the immersive environment are underrepresented, considering an extensive body of literature on Mixed Reality (MR) research domains. Perceptually valid sound rendering is key to realism in audio stimuli within immersive environments, as spatial features of the users' surroundings can be considered, approximating fundamental characteristics of sound propagation in environments. This enables listeners to use natural hearing abilities that interpret sound propagation effects to sense space and entities in their proximities, affecting interactions in immersive experiences. This thesis reviews the current state of sound rendering techniques and their application and feasibility across several use cases, proposing, as a novel contribution, a pipeline that can generate context-aware realistic audio for MR applications. The development of this pipeline involves adopting computer vision techniques in the process of decomposing complex scenes to recognise acoustic characteristics of space, determining physical and structural features of the environment surrounding HMD users, and allowing audio stimuli to respond to spatial characteristics of the immersive environment. The experiments presented demonstrate applications of scene understanding techniques applied to virtual environments and reconstructions of real space to determine acoustic properties of surfaces and entities for automating the application of sound rendering. This is done by identifying the current state of automatic acoustic material recognition for virtual environments and proposing novel evaluation methods that test the efficacy of automatic systems for tagging acoustic materials in virtual environments. Proof-of-concept systems have been tested on state-of-the-art acoustic renderers to demonstrate their efficiency in real-world scenes. Participant testing using a prototype deployment of the proposed pipeline that measures the performance of psychoacoustics-related tasks suggests that audio stimuli generated using the proposed pipeline have a significant effect on task performance within Mixed Reality applications. Current directions are aimed at designing end-to-end pipelines for interactive, real-time applications, with the ambition of adopting computer vision to understand the acoustic space, even in contexts of dynamic geometry typical of HMD technology, where the acoustic space is constantly updating based on the users' surroundings.

Preface

> That is praxis: it's in doing that you find out. In doing it, you find out why you did it.
>
> *Tony Wilson*

With this preface, I want to express my gratitude to Carlo Harvey, for his continued and ongoing support as director of studies and coach of my research career. Thanks, Alan Dołhasz, for beating an appropriate amount of work ethic into me as a directionless graduate. Thank you to Jason Hockman for inspiring me to pursue post-graduate studies since my first days of higher education, and guiding me through my undergraduates. I also want to thank Maite Frutos-Pascual for dedicating a lot of her time to reviewing much of this writing. I have been fortunate enough to have these figures provide me with the opportunity to study for a PhD, which materialised in three years of developing myself and putting together the work presented in these pages.

Circumstances over the last months of this work had myself at the front lines of the institution, as a lecturer, which marked the end of this magic with a busy schedule of teaching duties. Watching my work taking the back seat over the last months eroded my motivation and energy. I'm thankful for Vedad Hulusic playing a pivotal role here. He dedicated his time to untangle the technical threads of this work, creating a discussion that gave me fresh perspectives and reminded me what it was that I was doing.

Thanks a lot to Matthew Cheshire, Hamid "Nathan" Dewell, Richard Davies, Becky Spittle, Borislav Yordanov, Nick Jillings, Pinar Satilmis, and Ian Williams for contributing to the work and supporting me in many ways. It is thanks to this amazing bunch of people that I was able to work my way through this journey. Thanks to the Rabin Ezra Scholarship Fund too, for funding a significant chunk of this work.

Thanks to my family for continuosly supporting me, making this journey possible; I'll forever be grateful for this opportunity and freedom to study in the UK. Finally, thanks a lot to Karensa for continuosly and patiently putting up with my lows, highs, climbing obsession, and much more.

Declaration

I confirm that the work contained in this PhD project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged, and all verbatim extracts are distinguished by quotation marks.

Signed  Date: 6th May, 2025

Mattia Colombo

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

AABB	Axis-Aligned Bounding Box.
AAR	Audio Augmented Reality.
AR	Augmented Reality.
BRDF	Bidirectional Reflectance Distribution Function.
BRIR	Binarual Room Impulse Response.
BSP	Binary Space Partitioning.
BVH	Bounding Volume Hierarchy.
CNN	Convolutional Neural Network.
DFT	Discrete Fourier Transform.
DNN	Deep Neural Network.
DSP	Digital Signal Processing.
DTFT	Discrete-Time Fourier Transform.
ERB	Equivalent Rectangular Bandwidth.
FDTD	Finite-Difference Time-Domain.
FFT	Fast Fourier Transform.
FIR	Finite Impulse Response.
GA	Geometrical Acoustics.
GAN	Generative Adversarial Network.

HAS	Human Hearing System.
HMD	Head-Mounted Display.
HRIR	Head-Related Impulse Response.
HRTF	Head-Related Transfer Function.
ILD	Interaural Level Difference.
IR	Impulse Response.
ISM	Image-Source Model.
ITD	Interaural Time Difference.
JND	Just-Noticeable Difference.
LoD	Level of Detail.
MR	Mixed Reality.
MSE	Mean Squared Error.
NAF	Neural Acoustic Field.
NN	Neural Network.
RIR	Room Impulse Response.
RL	Reinforcement Learning.
STFT	Short-Time Fourier Transform.
TPU	Tensor Processing Unit.
VE	Virtual Environment.
VR	Virtual Reality.
XR	Extended Reality.

Chapter 1

Introduction

Since the first experiments by Sutherland (1968) in engineering systems to present users with stimuli to convey three-dimensional virtual objects, research domains around immersive virtual environments have only risen in popularity and become central topics in the major journals and scientific venues in computing.

Virtual Environments (VEs) constantly improve their ability to express virtual worlds in multimodal platforms (Rubio-Tamayo, Gertrudix Barrio and García García, 2017). Their applications range from computer games, including serious and educational games, to digital museums, cultural preservation, and architectural design, raising the need for realistic and compelling representations of VEs. Realism is achieved by emulating the physics of the natural world, improving user engagement and performance in tasks in immersive applications. Realism in the audiovisual domains are strongly related to the interaction with scene elements via perception of stimuli (Zimmons and Panter, 2003; Lokki and Grohn, 2005). Sound cues alone, for instance, are sufficient to enable users in VEs to pinpoint locations of sound-emitting entities in a scene by using auditory sound localisation, a natural ability associated with the human auditory system.

1.1 Motivations

1.1.1 Immersion and Realism in Immersive Applications

Creating compelling and realistic auditory displays is problematic, presenting many challenges in the acoustics and computer graphics research communities due to the complex nature of sound in the real world. As the acoustic principles that govern how sound propagates in space are difficult to reproduce in digital systems, many methods exist, providing variable orders of approximations, depending on the application. Such approaches emulate the wavefield of an environment, simulating how sound interacts with boundaries and scene objects. A subset of these can reproduce phenomena of sound, such as diffraction, reflection, and refraction, which are determinants of realism as they emulate how waves

bend around obstacles. Such phenomena make the simulated wavefield dependent on the accuracy of scene geometry and materials represented in a VE (Kuttruff, 2016).

Graphics and visual rendering systems advance towards better methods for representing geometry and objects of an environment, improving the relationship between virtual entities and their expression through multimodal stimuli. Such a relationship determines realism evoked in user interactions when rules of the physical world can be applied to multimodal displays to reason about these virtual entities. *Half Life: Alyx* is an example of a modern game that uses advances in multimodal rendering and interactions to create a complex physics system that allows users to manipulate and leverage the virtual environment reproducing natural, life-like physical interactions (Bollmer and Suddarth, 2022). Psychological factors relating to the human perception system determine the degree of realism evoked by multimodal stimuli produced by rendering systems. The perception of virtual entities in a complex virtual scene is ultimately subjective, requiring rendering methods to consider aspects of the human sensing system.

Auditory information is paramount to human perception in natural and virtual environments, helping in orientation and navigation, increasing immersion and aiding in task performance (Lokki and Grohn, 2005; Bork et al., 2015; Shivappa et al., 2016). The sound field a listener perceives is a function of the surrounding environment's shape, dimensions, boundaries, and transmission mediums. Even though the physics of sound propagation makes realistic audio rendering challenging, many proposed approaches allow realistic simulations of sound fields in VE. Computer games, compelling simulations, and digital tourism benefit from realistic audio rendering and improved auditory realism evoked in virtual environments (Lokki et al., 2002; Selmanović et al., 2020a).

The realism of digital media has increased in recent years thanks to recent advances in computer games technology (Rubio-Tamayo, Gertrudix Barrio and García García, 2017). As a listener in a sound transmission, the human auditory system is aware of acoustic characteristics manifesting in auditory cues, which enable spatial hearing abilities, such as sound localisation, aiding interaction tasks with objects in the world. Intrinsic acoustic characteristics are dependent on the sound transmission's wavefield, dictated by structural properties of the environment, associated with boundaries and materials, as they interact with sound propagating to the listener's ears. Acoustic rendering methods simulate real or virtual auditory environments by deriving from sound propagation algorithms that discretise the geometrical representation of an environment to synthesise a wavefield. They render spatialised sound adapting signal processing chains to reproduce realistic sound transmission in the simulated wavefield, considering the listener's position, orientation, and physical characteristics, described by Head-Related Transfer Functions (HRTFs) (Hulusic et al., 2012).

1.1.2 Modern Immersive Human-Computer Interfaces

MR platforms are suited for a range of use cases within areas of training, digital tourism, or games, thanks to the enhanced interactions with the physical world surrounding the user. An MR platform, which throughout this work defines the set of technologies and hardware interfacing the user with virtual and physical environments via audio-visual and holographic displays, gesture, eye, and body tracking, and with the physical world via computer vision, scene understanding, space reconstruction, and spatial mapping techniques. This broad range of sensing technologies and multimodal interefaces open avenues for addressing tasks within research domains where users interact with entities in MR space. Accessibility, for instance, is a research domain benefitting from the development of this platform by helping users with disabilities navigate and understand their surrounding scene, leveraging the sensing data provided by the platform and performing speech recognition or scene understanding techniques (Mehra et al., 2020).

The increasing popularity of immersive media proliferating within an increasing number of industry and research domains (Park and Kim, 2022), enabling HMDs to support applications ranging from training scenarios, education, social learning, operation learning (Harris et al., 2020; Ahir et al., 2020) to entertainment or gaming (Yuen, Yaoyuneyong and Johnson, 2011; Ke and Moon, 2018) or to digital tourism and cultural preservation use cases (Schofield et al., 2018; Selmanović et al., 2020b). MR applications in modern HMDs, e.g. the Microsoft Hololens 2[1], can provide real-time space reconstruction, spatial mapping, and scene understanding features[2]. Such features have unlocked crucial potential in the sound rendering subdomain, as emerging novel pipelines allow acoustic simulation techniques to be deployed to MR HMDs, as discussed in Chapter 6. As a result, users experiencing holographic content projected onto their surroundings can perceive realistic sound propagation from holograms. Experiencing realistic sound rendering in MR directly impacts the above-mentioned industry domains and research as HMDs become more ubiquitous and accessible.

These example applications and use cases of MR with modern HMDs rely on auditory stimuli for conducting activities or completing tasks in VEs, such as navigation, locomotion, or entity localisation. As discussed in Chapter 2, the hearing sense relies on acoustic information encoded in auditory stimuli to predict or triangulate the position of sound-emitting objects within the hearing range.

1.1.3 Limiting Factors of Immersive Technology within Audio Interfaces

The current state of MR allows auditory interactions between virtual entities projected around the user. These are possible thanks to the advances in game engines, enabling the construction of complex immersive scenes that can be experienced through HMDs. The field of Audio Augmented Reality (AAR), as defined by Yang, Barde and Billinghurst

[1] https://www.microsoft.com/en-gb/hololens
[2] https://learn.microsoft.com/en-us/hololens/hololens2-hardware

(2022) and Yang and Mattern (2019) advances towards realistic auditory interactions, allowing for context-aware stimuli and reflecting the behaviour of sound transmissions in the physical world.

Current frameworks available to build complex augmented scenes, such as the Microsoft Mixed Reality Toolkit[3], provide a limited suite of tools to implement audio sources, beyond basic spatialisation effects. For instance, audio emitted from virtual sound sources projected onto the real world does not consider basic principles of acoustics.

One cause is due to the complexity around the process of simulating sound propagation effects in dynamic virtual environments. Despite, decades of research advancing the field of sound rendering, computing acoustic effects in real-time is a difficult and expensive task. The ever-changing surroundings and the limited onboard compute availability in HMDs pose additional challenges to the task. These limitations prevent users from experiencing realistic MR sounds as they cannot exploit natural phenomena to perform tasks associated with sound perception. Based on factors of sound transmissions from a sound-emitting entity to the two ears. Sound localisation, as an example, is a natural ability that humans perform with everyday sounds where they pinpoint the location of an object by its propagated audio. Realistic sound rendering techniques can allow listeners to apply psychoacoustic abilities like sound localisation in sound transmissions within virtual environments or augmented environments. These use representations of real or virtual space to create an approximated acoustic model where sound propagation can occur, taking into account the architectural and physical characteristics of the environment. Hence, the spatial mapping and scene understanding capabilities of MR HMD can provide sound rendering techniques with real-time geometry and acoustic features associated with the users' surroundings, which are needed for realistic sound propagation.

On the one hand, there is a large body of research in psychoacoustic abilities such as localisation or clustering (Lee, 2011) generating recommendations for audio reproduction apparatus; and on the other hand, there are emerging pipelines allowing customisable binaural rendering for immersive technology, making it possible to encode acoustic phenomena in audio stimuli to match visual stimuli (Plinge et al., 2018). There is a lack of research towards combining the two research domains in emerging areas of MR, exploring the potential in the relationship between the physical space surrounding the user and auditory interactions. An experiment conducted as part of the thesis work will ask human participants to perform tasks requiring them to experience sound transmissions in MR and conduct tasks dependent on psychoacoustic abilities. The study will measure and gather data on the performance of participants conducting tasks to compare realistic sound propagation against current sound rendering methods.

[3] https://learn.microsoft.com/en-us/windows/mixed-reality/develop/unity/unity-development-overview

1.1.4 Increasing Adoption of Realistic Audio Interaction Systems

One significant advantage of XR platforms is their ability to bridge the gap between academic study and professional work by simulating real-world applications of learned concepts in a virtual environment. This immersive approach enables students and trainees to practice and refine their skills in a controlled setting that mirrors actual work conditions. By providing a hands-on learning experience without the constraints of physical resources, XR technology prepares individuals more effectively for their professional roles; this has caused thousands of businesses in the manufacturing industry to adopt immersive platforms[4]. XR platforms are powerful tools for knowledge transfer: they can present complex information in an interactive and visually engaging manner, making it easier for users to understand and retain new concepts. For instance, in onboarding new employees, educating customers about products, or training staff on new processes, XR provides an effective medium for conveying information. Safety is another critical area where XR technology proves invaluable. By simulating dangerous scenarios in a virtual environment, users can learn how to handle emergencies and hazardous situations without any risk to their physical well-being. This capability is particularly beneficial in industries such as firefighting, military, aviation, and construction, where real-world training can be dangerous and costly. Cost reduction is a significant benefit of using XR for simulations in this area, traditional training methods often require substantial resources, including materials, equipment, and physical spaces. XR simulations eliminate these costs by creating virtual environments where users can practice and learn. Fundamental to all these applications are audio interactions. Sound is a key stimulus that conveys critical information about the environment, enhances realism, and provides feedback that guides user actions. In virtual environments, realistic audio cues can indicate the presence of hazards, signal the operation of machinery, or replicate the ambient sounds of a particular setting, all of which contribute to a more immersive and effective learning experience (de Giorgio et al., 2023; Jot et al., 2021). This has raised interest towards realistic audio, increasing the drive for developing frameworks for realistic sound rendering. Recently, the industry has seen the popularisation of sound rendering tools such as Microsoft Project Acoustics[5], Steam Audio (Audio, 2020), and even more recently, Meta XR Audio SDK[6].

1.2 Significance

A novel aspect of the proposed work is the application of computer vision methods to infer acoustic data associated with materials and boundaries in environments to enable sound rendering systems, such as acoustic simulation methods, to adapt to unseen complex scenes in materials.

[4] https://www.pwc.co.uk/issues/technology/immersive-technologies/how-uk-organisations-use-vr-ar/business-benefits-of-xr-are-all-around-us.html

[5] https://github.com/microsoft/ProjectAcoustics

[6] https://www.uploadvr.com/meta-xr-audio-sdk-acoustic-ray-tracing-simulation/

Contributions deriving from this work would feed into computer games technology applied to entertainment, serious games, architectural acoustics, and digital tourism. Architectural acoustic applications would benefit from automated reconstructions of auditory spaces from real environments, allowing designers to create and test acoustic models with MR applications or game engines (Berardi, Iannace and Ianniello, 2016). Digital tourism, besides, would benefit from improvements in sound rendering pipelines to let users experience and interact with reconstructions of places of cultural interest (Schofield et al., 2018).

The need for realistic audio in computer games is of growing importance as it allows players to immerse in virtual worlds. Spatialising sound in three-dimensional worlds, reflecting acoustic features of space, improves the performance of tasks associated with auditory cues, enhancing the overall experience and quality of games. Activities in virtual environments using auditory cues, such as navigation in space and sound localisation are fundamental in simulations and training applications (Lokki and Grohn, 2005). Hence, sound rendering systems are of crucial importance in these fields.

The implications of advanced sound rendering extend significantly into training and simulation scenarios. In these environments, the ability to discover visually hidden entities through auditory information is crucial. For instance, in industrial training simulations (Machała, Chamier-Gliszczyński and Królikowski, 2022), realistic audio cues can help trainees locate and identify instruments, alarms, or hidden sound-emitting entities that are not visible or in their line of sight. This enhances the users' ability to respond quickly and accurately to various situations, mirroring real-world conditions where sound plays a critical role in situational awareness.

Furthermore, several branches of the broader "Metaverse" technology rely on realistic auditory interactions to drive engagement and enable realistic XR experiences (Park and Kim, 2022; Al-Ghaili et al., 2022). Auditory interactions are central components of the immersive aspects of the metaverse and, as a result, influence key activities such as collaboration across users in remote geographical locations.

1.3 Research Aim

The proposed work aims to explore rendering pipelines with respect to sound propagation in VEs, studying the relationship between the perceived wavefield and the environment's visual representation: materials, structural geometry, and physical characteristics of objects in complex scenes.

1.4 Objectives

The following is a list of objectives established as part of the thesis work.

1. To review the current state of audio rendering techniques in virtual environments,

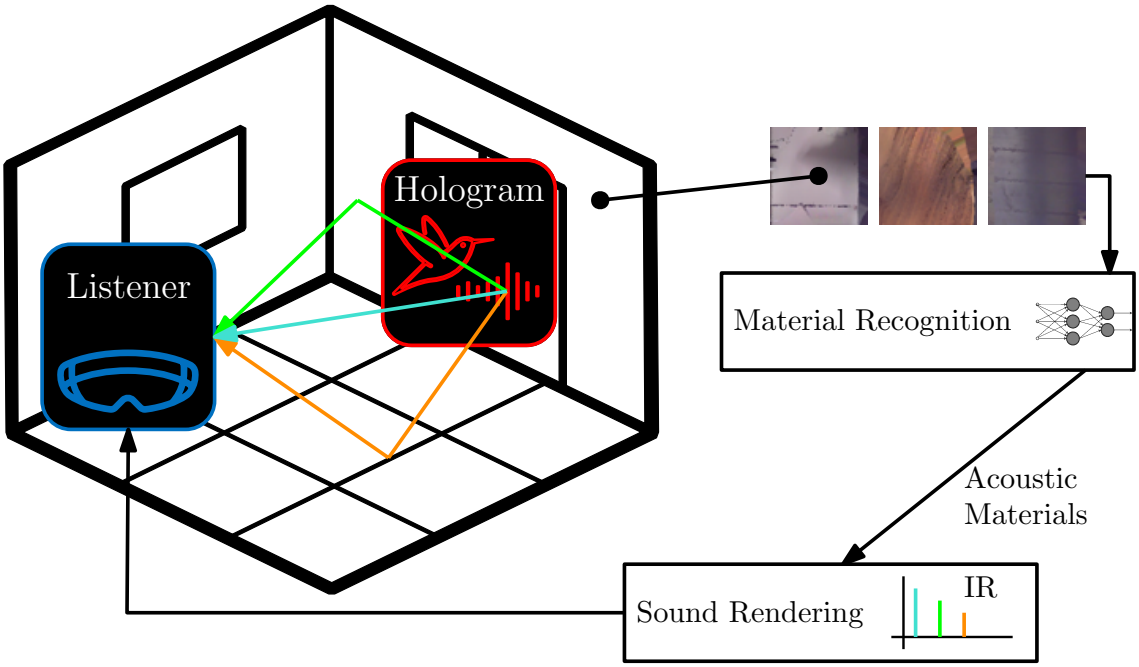


Figure 1.1: Overview of the interactive acoustic rendering system for MR proposed as part of the thesis work.

assessing their limitations, realism, computational requirements, and application to real-time acoustic simulations.

2. To explore how, in modern approaches to acoustic simulations, visual representations of materials relate to simulated soundfields.

3. To design and test systems to automatically attribute acoustic materials to scene geometry in virtual environments, recognising and distinguishing between materials in the acoustic and visual domains.

4. To study and evaluate the application of sound rendering methods with acoustic material recognition.

5. To design and propose a novel pipeline for acoustic rendering applied to augmented reality platforms.

6. To investigate psychoacoustic and human factors in auditory displays created by the proposed pipeline by testing an augmented reality prototype.

1.5 Thesis Structure

1.6 Chapter List

Chapter 1 *Introduction* presents the goal and objectives to the reader, starting from the rationale and factors motivating this research, providing background on research domains

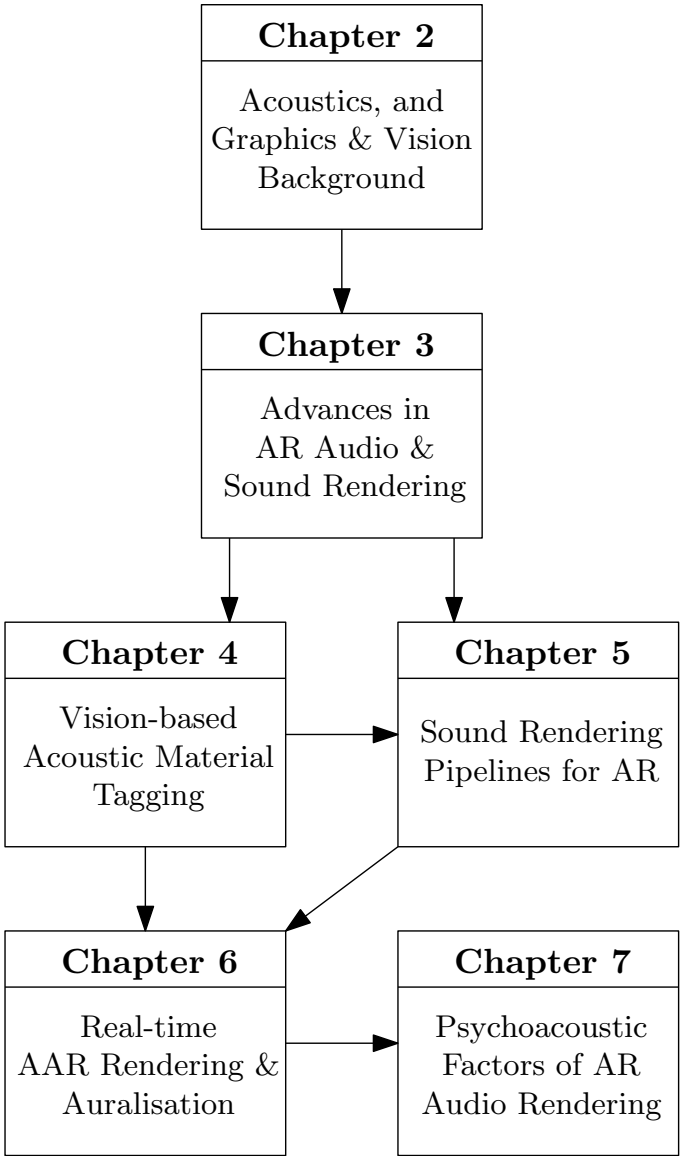


Figure 1.2: Structure and dependency graph of Chapters of this thesis. Introductory and conclusive Chapter 1 and 8, respectively, were omitted for simplicity.

of interest, and illustrating the contributions and how these are structured throughout the Chapters.

Chapter 2 *Background Research* provides the reader with foundations on wave physics, sound propagation, acoustic data in digital systems, as well as computer graphics in the context of handling and representing virtual environments. The reader is also introduced to elements of machine learning to provide grounding for vision-based scene understanding systems designed as part of the thesis aim.

Chapter 3 *Advances in Visual-Acoustic Mapping and Sound Rendering Pipelines* reports relevant bodies of literature relevant to research and industry domains overlapped by the proposed overarching system. Limitations of the current state of industry and research on these domains are reviewed alongside visions and research directions identified by authors, defining the contributions discussed across the thesis.

Chapter 4 *Methods for Acoustic Characteristics Retrieval from Complex Virtual Environments* introduces the design and application of computer vision methods for extracting acoustic characteristics from virtual environments, which constitute building blocks for the proposed pipeline and address acoustic material tagging, a crucial problem in the domain of sound propagation for virtual environments.

Chapter 5 *Geometrical Acoustics Rendering Pipelines for Augmented Acoustics* integrates acoustic material tagging systems with standard acoustic rendering techniques based on approximating acoustic waves with geometrical primitives. This Chapter aims to provide a baseline acoustic renderer for designing the interactive MR pipeline proposed in the following Chapter.

Chapter 6 *Towards Scene-Aware Acoustic Rendering Pipelines for Augmented Audio Reality* demonstrates the design of the proposed pipeline, expressed by the overarching research aim, as an end-to-end system for Augmented Reality Head-Mounted Displays, illustrating technical components and workflows for achieving real-time interactions. Considering related work reviewed by Chapter 3, visions, impact, and future research are discussed, indicating avenues for expansions.

Chapter 6 *Psychoacoustic Characterisation of Rendering Pipelines for Augmented Acoustics* deploys a prototype to an augmented reality platform in order to conduct a set of psychoacoustic tests on the pipeline illustrated by the previous Chapter. Here, a novel methodology and framework for testing the characterisation of sound rendering pipelines are evaluated and discussed.

Chapter 8 *Conclusions* provides a high-level discussion on the results gathered from objective and subjective experiments gathered by testing components of the proposed system, reflecting on broader impact, detailing potential use cases, and recommending future expansion avenues.

1.7 Contributions

The main contributions of this thesis are the proposal, design, prototyping, and testing of a novel pipeline for generating realistic auditory displays in augmented reality platforms, leveraging the potential that modern hardware for holographic rendering has unveiled over the last decades. These contributions represent a stepping stone towards realistic and physically-based auditory interactions in immersive applications, allowing users to understand, reason, and act within a virtual environment from acoustic information conveyed by the augmented reality platform. Throughout this thesis, the terms "pipeline" and "prototype" are used interchangeably, referring to the system, illustrated in Figure 1.2, composed of technical components that interact with a VE, working together to produce realistic audio to the user experiencing the environment.

These contributions stem from components of the pipeline developed and tested, addressing research questions associated with the design of an end-to-end system. Chapters 4, 5, and 7 present novel systems to address such research questions adopting bespoke methodologies and evaluations. The following list briefly summarises the contributions of this work:

- a novel pipeline for realistic acoustic displays for augmented reality platforms;
- two novel systems for retrieving acoustic characteristics from virtual environments;
- novel testing methodologies for evaluating system for acoustic characteristics retrieval, comparing simulated against real soundfields;
- two sound rendering systems integrating methods for acoustic characteristics retrieval;
- a study evaluating the proposed sound rendering methods;
- a perceptual evaluation investigating psychoacoustic factors of a prototype of the proposed system.

The following papers have been authored as part of the thesis work:

1. Colombo, M., Dolhasz, A. and Harvey, C., 2020, August. A computer vision-inspired automatic acoustic material tagging system for virtual environments. In 2020 IEEE Conference on Games (CoG) (pp. 736–739). IEEE.
2. Colombo, M., Dolhasz, A. and Harvey, C., 2021, May. A texture superpixel approach to semantic material classification for acoustic geometry tagging. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–7).
3. Colombo, M., 2021, October. Vision-based acoustic information retrieval for interactive sound rendering. In 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (pp. 487–490). IEEE.

4. Colombo, M., Dolhasz, A., Hockman, J. and Harvey, C., 2022, August. Acoustic rendering based on geometry reduction and acoustic material classification. In 2022 IEEE Conference on Games (CoG) (pp. 409–416). IEEE.

5. Colombo, M., Frutos M., Harvey, C., 2024. Psychoacoustic Characterisation of Geometrical Acoustics-based Rendering Pipelines in Augmented Reality. In Review.

6. Harvey, C., Colombo, M., West, B., Happa, J., and Ali-MacLachlan, I., 2024, October. Immersive Acoustics in Cultural Heritage Applications in Interactive Media for Cultural Heritage. Nicosia, Cyprus: Springer Cham. Book Chapter. In Press.

Chapter 2

Background

The following Sections introduce a body of knowledge from domains intersecting the overarching aim of this work, providing the reader with the necessary tools to dissect the components presented throughout the thesis. This Chapter begins by introducing basic sound physics, wave theory, and human perception as they define basic concepts of auditory interactions in the physical and virtual worlds.

Since a human listener is a link in the chain of the system proposed as part of this work, this Chapter will introduce basic psychoacoustic concepts linking the objective and technical aspects of sound rendering systems to human factors and subjective perception.

2.1 Background on Human Hearing

2.1.1 Characteristics of the Human Hearing System

The Human Hearing System (HAS) generally comprises two ears on either side of the human head, and each ear is a system that can be divided into three main parts, visualised in Figure 2.1: the outer ear, commonly referred to as the ear; the middle ear; and the inner ear. The outer ear, also referred to as the pinna, is shaped like a shell, is made of cartilage and skin, and serves both as protection for the system of receptors, ossicles, and nerves within the middle and the inner ear and a funnel that collects sound energy and transmits it to the middle ear via the outer ear canal. The outer ear canals are largely responsible for the frequency response of the HAS. Auditory stimuli are sent to the brain via sensory cells that are surrounded by fluids displacing according to the received sound pressure. The middle ear converts sound pressure from the ear canal to displacement to these fluids, which send signals to the brain. This part of the ear is able to withstand variations of air pressure arriving at the outer ear and is responsible for matching different impedance magnitudes between the air or the medium in which sound is arriving at the apparatus and the impedance of the fluids in the inner ear. The inner ear comprehends the vestibular system, a sensory system that allows humans to sense their spatial position, perceive rotation or

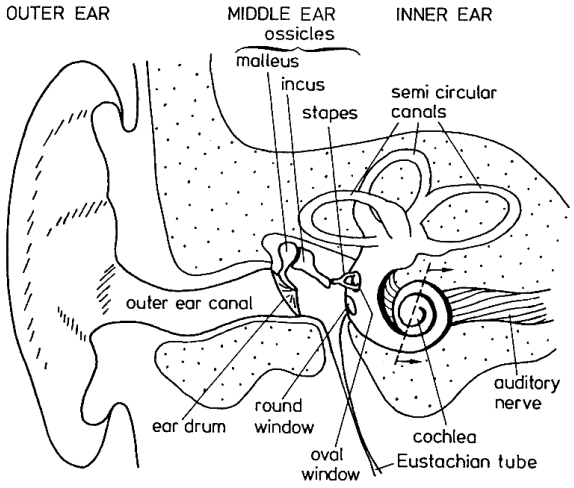


Figure 2.1: Overview of the human hearing system (Zwicker and Fastl, 2013).

displacement, and achieve balance and the cochlea. The cochlea, shaped like a snail, is embedded in the hard temporal bone, part of the skull (Zwicker and Fastl, 2013). Understanding basic functionalities of the HAS is fundamental to reasoning auditory perception within virtual environments. When designing or evaluating sound rendering pipelines for immersive applications, factors influencing the perception of sound need to be considered. Anthropometric features of humans affect how propagating sound waves are interpreted by listeners. Designing systems to simulate auditory perception in virtual environments requires considerations of key features of the HAS. Section 2.1.2 will demonstrate how existing methods can simulate aspects of human listeners in digital systems. Hearing, as a human ability, does not work in isolation and it is influenced by vision and other functions of human perception. The interaction of different senses, particularly how sound influences visual perception in virtual reality scenarios, affects how environments are perceived. Auditory stimuli can affect visual perception and vice-versa, adding dimensions of complexity to the process of designing an audio apparatus for immersive technology (Malpica et al., 2020). For instance, the act of observing in a large hall or church can inform a listener about sound sources being affected by a significant amount of reverberation and echo, making it harder to understand speech in such a space. Conversely, hearing reverberant audio without vision information can provide some assumptions and visual cues of the space being of large dimensions. The feature mismatch between visual and auditory cues can affect the externalisation and perception of sounds. For instance, differences in quality between acoustic and visual stimuli can hinder the ability of a listener or observer to understand their surroundings from perceived stimuli, demonstrating the importance and validity of congruent sensory cues for effective auditory perception in virtual environments (Bonneel et al., 2010). With matching and coherent audio-visual stimuli, the human perception system can infer and "fill in the blanks", understanding spatial or semantic information of a sound-emitting object when visual information is scarce or missing. Replicating such ability in digital systems can be crucial for accessibility applications, allowing users with vision impairments to understand their surroundings by understanding acoustic characteristics.

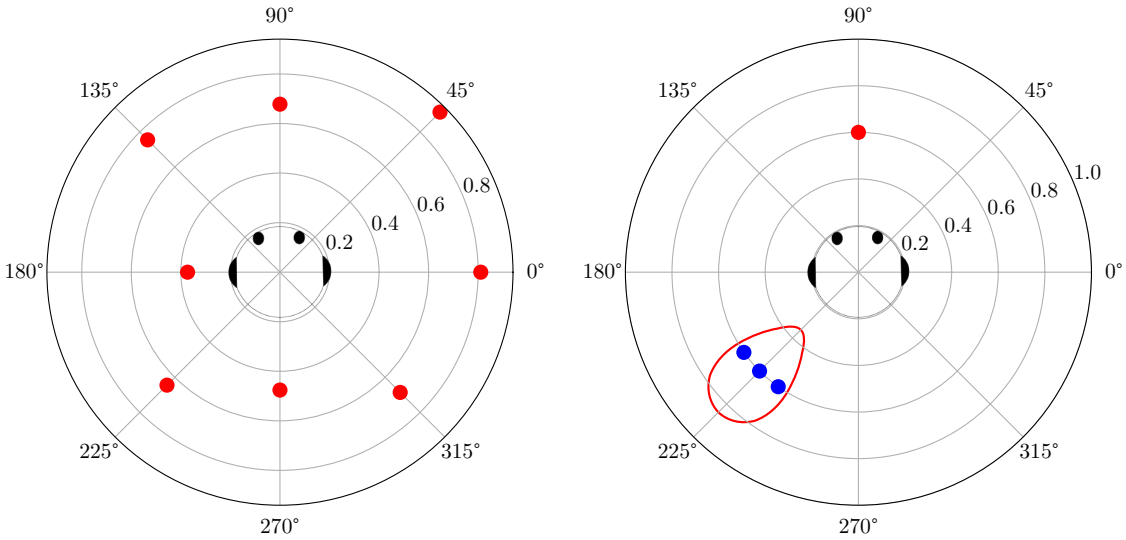
2.1.2 Introduction to Psychoacoustics

The HAS enables one of the fundamental functions of perception of surrounding space. In humans and species of the animal kingdom, hearing is the basis of many mechanisms, such as communication or survival instincts. Such mechanisms are neural processing applied to auditory stimuli arriving at the hearing system in order to compute tasks or solve problems, such as communicating using acoustical data or pinpointing the location of a sound-emitting entity relying on auditory stimuli. These are example applications or problems that can be solved by processing acoustical data interpreted by the HAS. Psychoacoustics investigates how the HAS responds to auditory stimuli and investigates applications like loudness perception, localisation, lateralisation, or room volume estimation. The understanding of psychophysical responses of the HAS to acoustical data in environments influences everyday activities that involve communicating, listening to musical instruments, or delivering messages to an audience of multiple listeners. The design process of built environments, infrastructure, or concert halls takes into account psychoacoustic factors to facilitate or improve human perception in specific environments during specific activities. Psychoacoustic models significantly enhance the design and functionality of auditory displays by ensuring that sound reproduction systems are more aligned with the natural processing capabilities of the HAS. Over the course of this thesis, several designs of audio rendering apparatuses will be discussed, comparing their effectiveness and efficiency. Psychoacoustic metrics provide an essential benchmark to test these systems, as they can indicate how effective a method is towards allowing users to perform natural abilities on a given stimulus. Studies such as Rungta et al. (2016)'s indicate how well listeners can apply psychoacoustic abilities, like localisation or room volume estimation, on stimuli generated by experimental sound rendering pipelines.

2.1.3 Sound Localisation

Sound localisation is a natural ability of the HAS that allows the determination of the direction and distance of a sound-emitting entity. This ability is essential for humans, animals, and autonomous agents, enabling sensing and discovering environments. In VEs, accurate sound localisation can greatly enhance immersion, making the experience more realistic and interactive by replicating a natural phenomenon in a virtual space. Figure 2.2 shows a top-down representation of a listener performing localisation and resolving auditory masking.

The HAS localises sound based on the acoustic cues that reach the ears; they can be binaural or monoaural. Binaural cues simultaneously involve the two ears (or receiving points), while monoaural cues consider mechanisms relating to individual receiving points. Binaural cues are the most significant for performing sound localisation as they consider differences in the arrival of sound at the two ears, revealing distance information between the source and each receiving point. These cues are often subdivided into Interaural Time Difference (ITD) and Interaural Level Difference (ILD), referring to the differences in time



(a) Sound localisation is the ability of the HAS to pinpoint the position and direction of sound sources, red dots, at different directions and distances around the listener, centre.

(b) Masking is a phenomenon that occurs when the presence of one sound makes it difficult or impossible to hear another sound at the same time.

Figure 2.2: Visualisation of sound localisation and masking.

and sound pressure level, respectively, at the time of arrival at the two receiving points.

Monoaural cues consider aspects relating to the physical and anatomical characteristics of the listener affecting propagating sound waves being perceived by the pinna, middle, and inner ear. Physical characteristics can include the shape of the pinna or the ear canal and other parts of the hearing system involved in the perception of sound pressure. These effects can transform or alter the spectrum of the perceived auditory stimuli (Blauert, 1997; Howard and Angus, 2013).

2.1.4 Psychoacoustics in Auditory Stimuli

In the realm of digital audio technology, enabling listeners to apply psychoacoustic abilities to perceived auditory stimuli is an ongoing research area. Monoaural or two-channel audio formats, despite being among the most popular audio reproduction formats, have limited potential in expressing the directionality of sound sources and the sense of space around the entities existing in the audio scene being reproduced.

Ambisonics is a system engineered to overcome the issue and allow the reproduction of a surround image that expresses the direction, height, and distance of audio sources (Frank, Zotter and Sontacchi, 2015).

Ambisonics is implemented through several formats; one of the most popular implementations, the B-Format uses spherical harmonics to represent the sound field. A special ambisonics microphone captures sound from all directions by adopting multiple capsules

(often four) arranged to capture not just the intensity but also the direction of incoming sound waves. Sound is captured and reproduced using a set of audio channels that represent the sound field in terms of these spherical harmonics, which can describe sound coming from any direction around the listener. The audio from the microphone capsules is encoded into a multi-channel format. The basic form of ambisonics, known as first-order ambisonics, uses four channels: one for the overall sound pressure (omnidirectional) and three for directional information along the three spatial axes (X, Y, and Z) (Zotter and Frank, 2019).

Sound rendering pipelines often use Head-Related Transfer Functions (HRTFs) to model monoaural and binaural cues in virtual environments, to express directionality through stimuli more effectively and precisely. They describe how sound is affected by the listener's head, ears, and torso before reaching the ears. HRTFs are often individualised, i.e. modelled after an individual hearing system, and can be used to simulate how the listener perceives sounds from different directions. Modelling and creation of these functions often involve measurement systems to capture sound from varying distances and directions arriving at the hearing system (Zotkin et al., 2003). However, sound localisation can also be affected by cognitive and psychological factors such as experience, expectation, and attention, as the brain can use past experiences and contextual information to make educated guesses about the location of sound sources.

In immersive applications, HRTFs have become fundamental in allowing sound localisation of virtual sound-emitting entities, though they can vary significantly between individuals, and creating a one-size-fits-all model for all listeners is still an open research question (Schäfer et al., 2024). There are research trends that focus on systems for individualising HRTFs automatically by predicting functions based on features of the target listener. Modern techniques make use of computer vision systems to infer HRTF data from a virtual representation of a human head; i.e., a 3D scan of the listener's head and upper body (Zotkin et al., 2003).

2.1.5 Masking

Masking occurs when the presence of one sound makes it difficult or impossible to hear another sound at the same time. This effect can significantly influence how sounds are perceived in everyday environments and is critical in the design of sound rendering pipelines for several reasons, including emulating how hearing systems operate or managing computational resources by avoiding rendering sounds that listeners cannot hear. Masking can be classified into several types based on the characteristics of the sound responsible for masking other sounds (masker) or the sound being masked (maskee). Simultaneous Masking occurs when the masker and maskee are present at the same time. High-intensity frequencies can mask nearby lower-intensity frequencies, affecting the ability to discern sounds that are close in frequency range. Temporal Masking occurs when a masker precedes the maskee. Spectral Masking occurs when masking sounds across different frequency

bands, where a strong presence in one band can affect the perception of sounds in another (Howard and Angus, 2013). Masking varies by individual and is influenced by cognitive factors, context, environment, and content or nature of the auditory stimuli; due to this, challenges remain in fully understanding and predicting masking effects.

2.1.6 Just-Noticeable Differences

The Just-Noticeable Difference (JND) is the smallest change in a stimulus that can be detected by the sensory system. In the realm of hearing, it applies to various acoustic parameters, such as frequency (pitch), intensity (loudness), and duration (length of sound). The JND is not a fixed quantity but varies depending on the baseline intensity and frequency of the sound, as well as the listener's sensory acuity and environmental factors (Dolhasz, 2021) Psychoacoustic models, which predict human auditory perception, incorporate JNDs to simulate how various sounds are processed and understood. The concept of just-noticeable differences is vital for understanding human perception and forms the basis of numerous applications in psychoacoustics and to replicate phenomena of the HAS in virtual environments. As research advances, our grasp of JNDs continues to refine the development and optimisation of sound rendering pipelines.

2.2 Sound Propagation Background

2.2.1 Sound Propagation in Real and Virtual Environments

Sound propagation is a transmission of energy in a sound field, which can be thought of as a superposition of sound waves travelling in a medium. In this work, we consider air as the sound propagation medium, which is assumed to be homogeneous, i.e., determining a constant velocity of sound c expressed as:

$$c = (331.4 + 0.6\Theta)\ \frac{m}{s} \tag{2.1}$$

where Θ is the temperature in centigrade. A vibrating object in a sound field causes air particles to move, initiating the transmission of energy in the field. Such an object is defined as a sound source, and if the intensity and frequency of the vibrations are within the perceptible range of the human hearing system, a listener may experience sound emitted by the said sound source. In everyday sound transmissions, the air within sound fields is not at rest and features many inhomogeneities caused by external factors affecting the state of its particles, such as windows or air conditioning systems. However, according to (Kuttruff, 2016), such inhomogeneities are imperceptible, and generally, the air temperature has a perceptual effect on sound transmissions, especially in large concert halls and open spaces. Air temperature effects can be neglected in indoor sound propagation.

2.2.2 Metrics and Descriptors of Real and Virtual Soundfields

Standard acoustic parameters are crucial for measuring and evaluating soundfields, particularly in diverse environments like concert halls, studios, or public spaces. These parameters help in quantifying aspects such as sound quality, clarity, and diffusion, and they can express high-level characteristics of the behaviour of propagating sound waves in a given environment.

One of the most common acoustic parameters, reverberation time is essential for understanding one of the most characteristic aspects of an environment. It allows listeners to infer information about the size or basic architectural features of the space. T_{60} reverberation time is defined as the time it takes for acoustic energy levels to drop by 6.0×10^1 dB after a sound source has stopped emitting (Eckhardt, 1923).

Similar to reverberation time but focused on the early part of the decay, providing a better description of the acoustic environment in terms of initial sound fading, Early Decay Time is another metric that provides insights into how energy behaves in environments. It is often calculated by fitting a curve on energy levels registered over time by a receiver (Jordan, 1970). The clarity index, also a commonly adopted metric, measures the clarity of sound in terms of its impact or sharpness. It quantifies the ratio of early (within 8.0×10^1 ms or 5.0×10^1 ms) to late reflections (Reichardt, Alim and Schmidt, 1975). Equation 2.2 shows how the ratio between early and late reflections is calculated: p denotes energy from a source, registered at the listener point at the time t.

$$C_{80} = 10 \log \frac{\int_0^{80ms} \mathrm{p}^2(\mathrm{t})\mathrm{d}t}{\int_{80ms}^{\infty} \mathrm{p}^2(\mathrm{t})\mathrm{d}t} \tag{2.2}$$

Similarly to clarity, the definition metric D_{50} analyses early and late reflections, expressing sharpness and definition of propagating sound waves. The metric differs from clarity by quantifying the ratio between early reflections and the total aggregate reflections, see Equation 2.3.

$$D_{50} = 10 \log \frac{\int_0^{50ms} \mathrm{p}^2(\mathrm{t})\mathrm{d}t}{\int_0^{\infty} \mathrm{p}^2(\mathrm{t})\mathrm{d}t} \tag{2.3}$$

These metrics have strong relationships with subjective factors, like perceived quality or perceived resolution. As humans are often the target of digital systems that simulate soundfields, subjective factors need to be considered. This is crucial in the context of generating and validating acoustic simulations: although objective metrics can quantify and evaluate acoustic characteristics of soundfields, they require validation against human perception to assess psychoacoustic factors.

MUlti-Stimulus with Hidden Reference tests (MUSHRA) can be used to compare auditory stimuli subjectively. These tests are typically employed to assess digital signal processing algorithms, such as compression, and involve subjective testing. Subjects are often asked to rank a set of stimuli, expressing the perceptual distance between each stimulus and a reference. Such testing methodologies can be adapted effectively to compare different soundfield reproductions, particularly in evaluating spatial audio and soundfield reproduction systems within immersive environments to assess perceived audio quality, as demonstrated by Rummukainen et al. (2018).

2.2.3 Digital Representation of Audiovisual Information

The following Sections will introduce background knowledge on Digital Signal Processing relevant to the representation of acoustic signals in digital systems and the manipulation of auditory stimuli in virtual environments. Digital Signal Processing methods and techniques provide building blocks for the construction of realistic 3D auditory displays in immersive technology.

Digital Signal Processing (DSP) is the science of analysing time-dependent physical processes. The acoustics realm deals with analogue signals and digital signals, terms used to indicate a continuous variation of amplitude values in a physical process. Electricity utilised to drive loudspeakers is an example of an analogue signal, expressing continuous changes in voltage applied to magnets to displace the position of a cone. The cone displacement causes pressure differences in air particles, transforming such changes in voltage to changes in air pressure, which the human auditory system interprets as sound. Acoustic signals consist of one or multiple sound waves oscillating, where each wave is an oscillation of energy at regular intervals; the duration of each interval determines the wavelength λ, and oscillations are measured as frequency in Hz.

On the other hand, a digital signal is a discrete representation of a continuous physical process, resulting in a sequence of measurement samples of an analogue signal expressed as amplitude values over time. Figure 2.3 shows the difference between a continuous signal and a discrete signal: digital signal is represented with stems to indicate its nature of quantised measurements over time, abscissa, as opposed to a continuous change in amplitude, ordinate. The discrete nature of a digital signal has inherent problems and advantages that relate to the time interval between measurements: a digital signal representing an analogue one will always be an approximation of the continuous process as the system may change its state between measurement intervals. The approximated nature of digital signals causes information loss, which is counteracted by theories shown later, but allows digital systems to store and process acoustical data efficiently.

DSP applies to both, but in this chapter, we will only focus on the branch of DSP that deals with digital signals. Digital systems like computers are used to process stored acoustical signals for several reasons, such as storing recordings of anechoic acoustic signals that

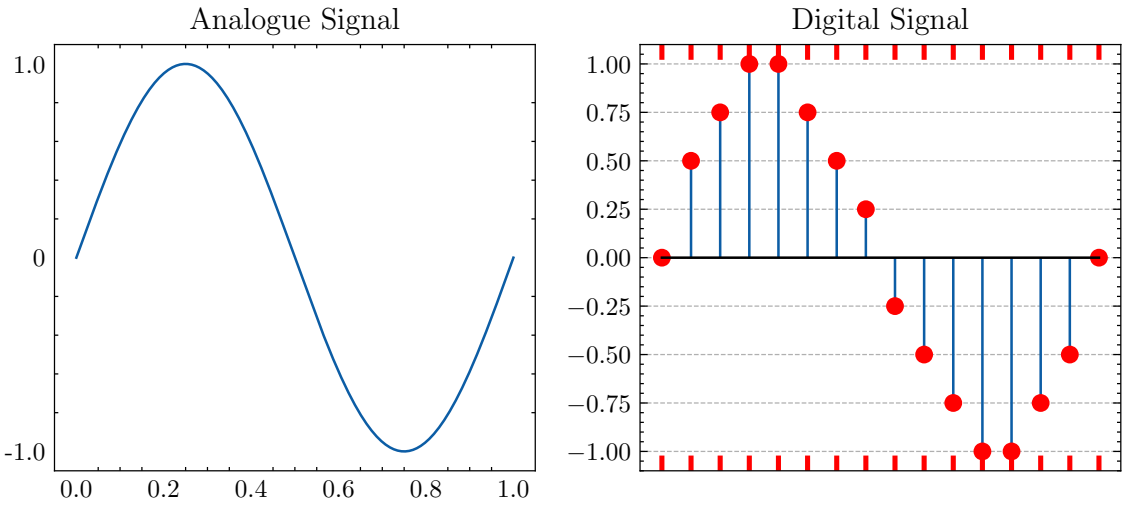


Figure 2.3: Analogue and digital signal: the left axes show a continuous signal and the right axes show a discrete sampled representation.

simulation software can then process to generate realistic acoustic simulations, expressed as a processed digital signal.

The process of converting continuous signals to digital information involves taking measurements of the amplitude of a continuous signal at regular time intervals. There are two dimensions in which the analogue signal is measured during this process, the amplitude and time, respectively, the abscissa and the ordinate of Figure 2.3. Due to the physical limitations of digital technology, A/D converters can only take a finite number of measurements between time intervals, and they have limited accuracy in representing amplitude levels. In Figure 2.3-b, it is possible to see how an A/D converter sees analogue signals: given an acoustic continuous signal as input, it takes amplitude samples at every time step, marked with by red ticks, and measures using the available amplitude levels (the dotted horizontal lines). As a result, the process outputs a series of data points, the red dots, approximating the input, and the resolution and fidelity of the approximation depend on the time elapsed between time steps and the available amplitude level points. There are standards to ensure the reproduction and manipulation of acoustical signals in digital systems with an appropriate fidelity, such as the "Red Book" IEC 60908 standard, adopted for the Compact Disc music format, determining that digital signals must be represented by 44.100 measurement samples per second, at 16bit amplitude resolution. 16-bit refers to the binary representation adopted by digital systems to store amplitude values, allowing $2^{16} = 65,535$ possible amplitude levels. The sampling frequency, the number of measurements per second, is calculated in Hertz (Hz), and it is a fundamental property of digital signals that must be taken into account for almost all types of audio manipulation and analysis involved in acoustical applications and it paramount to correct reconstructions of any acoustic information in any digital system.

The Nyquist-Shannon sampling theorem is used to ensure a digital system reconstructs an

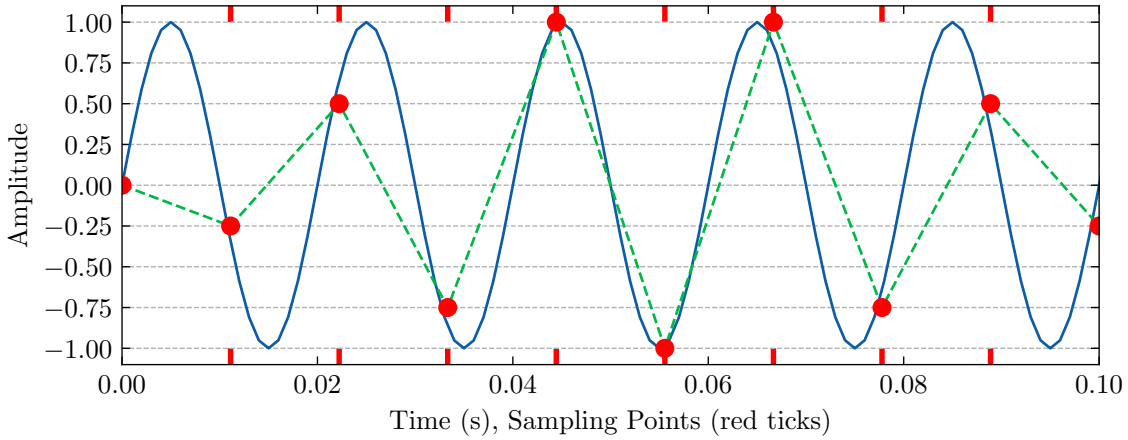


Figure 2.4: The blue signal is being sampled at a sampling frequency lower than the Nyquist frequency, causing aliasing, an incorrect reconstruction of the digital signal, as opposed to Figure 2.3 that shows a correctly reconstructed signal. As a result, the dotted green signal is created instead, having a frequency between 0 Hz and the sampling frequency.

analogue signal correctly. The theorem proves that a wave must be sampled at least twice during each oscillation period. A periodic wave oscillating at $20kHz$, which is around the maximum perceivable frequency in the human hearing range, would need to be sampled at least $40,000$ per second; hence, the standard $44.1kHz$ sampling rate. In Figure 2.4, for instance, a 5.0×10^1 Hz signal is sampled at 9.0×10^1 Hz, below the 1.00×10^2 Hz Nyquist sampling frequency, causing aliasing, an incorrectly reconstructed signal that will be able to oscillate at a maximum frequency of 4.5×10^1 Hz.

Analysis of Digital Signals

Acoustical signals are often analysed in the time domain, as varying sound pressure levels over time, or in the frequency domain. By considering acoustical signals as a Fourier series, a function composed of sine or cosine primitives, the frequency domain representation determines how the power of an acoustical signal is distributed in a range of sine and cosine functions with wavelengths usually ranging from the minimum to the maximum perceivable frequencies of the human auditory system — low to high frequencies. Time- and frequency-domain representations are often used for both analysis and manipulation of acoustical signals, often adopted in tasks like determining the effects of an environment in the perception of sound emitted by an object and arriving to a listener in said environment Ballou (2013).

In acoustics for interactive applications, engineers often adopt the Discrete-Time Fourier Transform (DTFT), a Fourier series for digital signals, which is one of the fundamental concepts in DSP. It takes a sequence, such as the signal represented in Figure2.3-b and generates N complex numbers, representing power across N sinusoids. The DTFT, visualised in Figure 2.5, a core principle of DSP theory Shenoi (2005), transforms a signal x_n containing samples $x_0, x_1, \ldots, x_{N-1}$ into a series X_k of complex numbers $X_0, X_1, \ldots, X_{N-1}$.

X_k is defined by:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn} \tag{2.4}$$

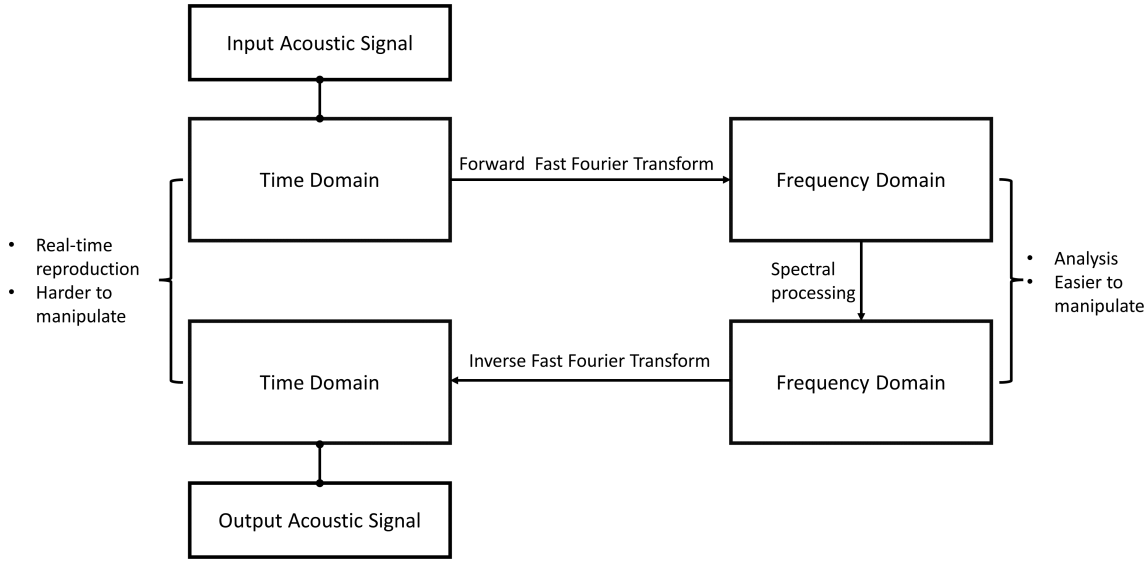


Figure 2.5: A basic chain for signal processing aimed at analysing or manipulating signal in auralisation, visualisation, or interactive applications. Analysis and processing of digital signals in the time domain is generally a hard task due to the complex nature of the function representing audio. The frequency-domain representation eases analysis and manipulation problems even with the added computational load of transforming between the two domains.

2.2.4 Binaural Room Impulse Responses

Common approaches to acoustic simulations involve the approximation of acoustic phenomena affecting a sound transmission occurring within a given environment between a sound source and a listener. To represent the result of such a simulation as a measurable process, where the environment is thought of as a dynamic system, Impulse Response (IR) are used. IRs describe the effect that a system has on a sound transmission as a function of time. From DSP theory, there are several variations of IRs that the fields of immersive acoustics borrow to model several dynamic systems that affect how the human auditory system perceives soundscapes. Figures 2.9 and 2.8 shows how the auditory display is affected by interconnected systems associated with aspects of the soundscape. Time invariance is the fundamental property of these systems, making it possible to model their effect as an IR by observing their response to a Dirac-Delta function, which is a function whose value is zero except at the origin, where it is infinite. In practical terms, the Diract-Delta function is an infinitely narrow energy spike often used to excite the system and obtain a response across the frequency spectrum over time. In DSP terms, the function is simply represented as a finite sequence of numbers, the Finite Impulse Response (FIR), representing amplitude levels of the output of the system over time, given an input, commonly used to measure the effects of time-invariant linear systems like amplifiers or loudspeakers.

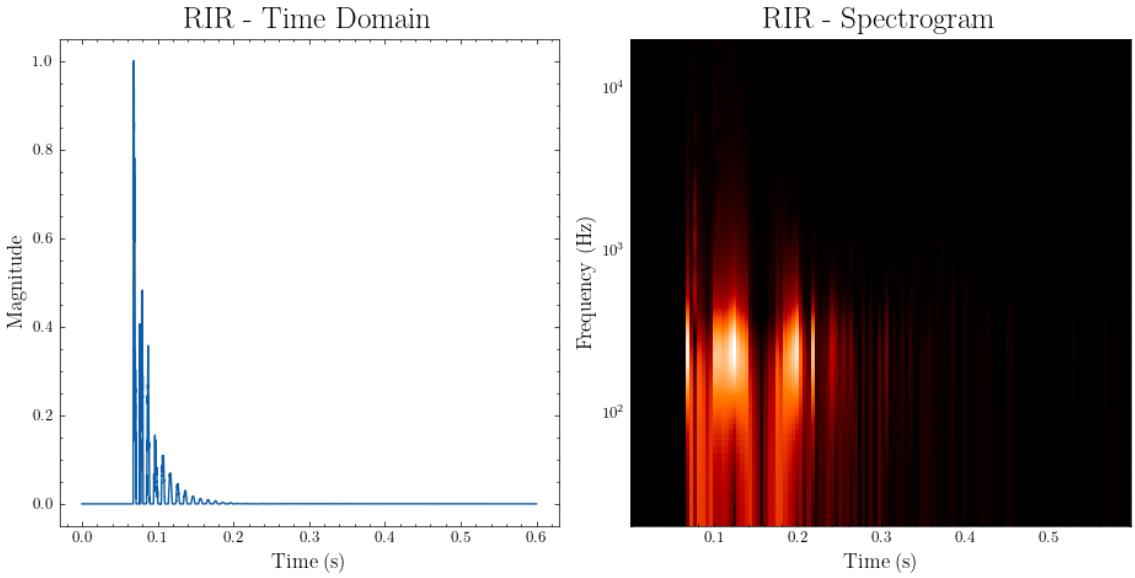


Figure 2.6: Both time and frequency-domain, left and right respectively, representations of a Room Impulse Response (RIR). The time domain representation shows the magnitude of sound paths from a sound source to a receiver over time. The frequency domain representation shows how the energy of sound paths is distributed across the frequency spectrum over time, which is visualised with an infrared colour map.

In the acoustic domain, the FIR adapts to several tasks, like modelling the acoustic fingerprint of a space with respect to a source and listener by observing, at the listener, a Dirac-Delta-like signal being emitted by the source. Such IR is differentiated from standard IRs and referred to as a Room Impulse Response (RIR); such distinction has emerged from the ongoing research in techniques and methods for measuring responses from real spaces, also due to the chaotic nature of room acoustics and real soundfields (Farina, 2007). IRs, as well as measuring the acoustic fingerprint of spaces, can extend as far as measuring the effect of the human auditory system on the perception of the soundscape, and there are methods for modelling how anthropometric characteristics of the human body affect sounds arriving at both ears. Such RIRs are defined as Binarual Room Impulse Response (BRIR): they extend IRs by providing individual responses for both ears. BRIRs are a representation of HRTF, a function that describes how the anatomic features, rotation, and position with respect to a sound source affect the arrival of sound to the ears. Figure 2.6 is an example of a monaural RIR, shown both in the time and frequency domain.

2.2.5 Measuring Real Soundfields

A fundamental process in the field of acoustic is capturing the acoustic characteristics of physical spaces. Over the last decades, both research and industry have developed techniques for capturing soundscapes using RIRs. Captured responses can encode fingerprints of unique soundfields, which are relevant to various applications like audio engineering or audio production, architectural acoustics, and virtual reality.

By means of the convolution operation, acoustic characteristics expressed by these responses can be applied to unpropagated audio signals. Commonly referred to as "auralisation", it allows rendering audio to a listener, giving the impression that it propagated within the response recording apparatus. Section 2.3.2 will discuss technical aspects and implementations of the operation.

The recorded impulse response can be analysed to determine acoustic parameters such as reverberation time, early decay time, and clarity (Farina, 2007). These parameters are crucial for acoustical analysis and for simulating the space's acoustics in audio production. Impulse responses allow engineers to apply the acoustic characteristics of actual spaces to studio recordings, creating authentic aural experiences in a controlled environment (de Lima et al., 2009).

Over time the process of measuring soundfields with speakers and microphones has been standardised by international standards such as (Liebetrau et al., 2014). Figure 2.7 shows an example recording apparatus for capturing acoustic responses from environments. The standard provides recommendations to maximise the representation quality of the acoustics features captured by the response. To record a response the technical apparatus may consist of a sound source that can produce a broadband audio signal, a microphone, and A/D and D/A converters to emit and capture signals from the sound system. A logarithmically swept sine is then emitted by the speaker and captured by the microphone. Ideally, the sound system should not introduce distortion into the captured signal and emitter and received should have minimal impact on the spectrum.

Farina (2007) presented techniques for recovering RIRs from the signal captured by the measurement microphone. The most notable technique is the convolution of the captured swept sine by the time reversal of the unpropagated sine sweep, resulting in the recovered RIR, Figure 2.6 shown an example recovered response.

This process helps audio professionals and acousticians accurately reproduce and study the acoustic behaviours of different environments, enhancing audio productions and architectural designs (Holters, Corbach and Zölzer, 2009).

2.3 Common Approaches to Immersive Acoustics

Achieving a convincing immersive acoustic experience is no trivial task, and various techniques and methodologies have been developed to address this complex challenge. These approaches must consider sound's spatial, temporal, and perceptual aspects and the HAS's intricate response to auditory stimuli. Realistic audio plays a crucial role in immersive entertainment, regardless of how a virtual environment is experienced, as it allows users to perceive the location and movement of objects and characters in a three-dimensional space, making the virtual environment more believable and engaging. In games or VR environments where visual cues are limited or absent, high-quality spatialised audio can

(a) RIR measurement in a large space

(b) RIR measurement in a small space

Figure 2.7: Photographs showing an apparatus for recording Room Impulse Responses in real spaces, adopting the swept sine method. The method involves using a speaker to emit a logarithmically-swept sine to excite, generating a test signal that is then captured by the microphone. Through the process of deconvolution, the recorded signal can be used to produce an impulse response of the system composed by the speaker, microphone, and environment.

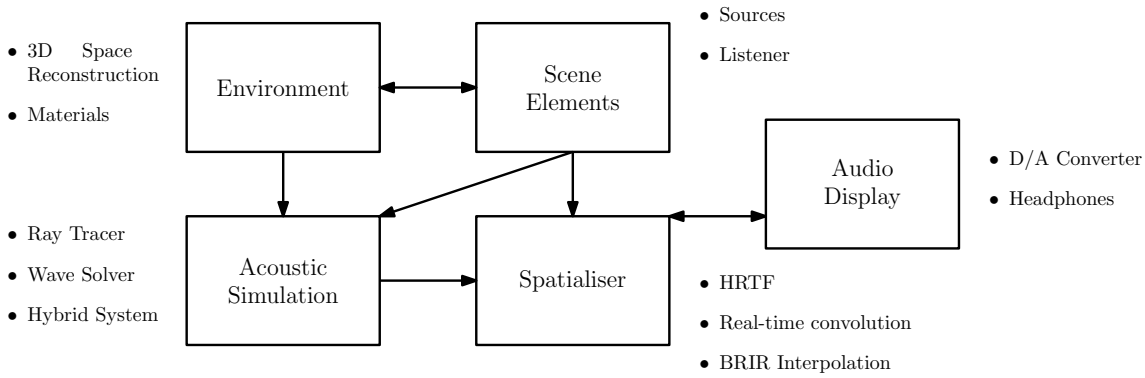


Figure 2.8: An example chain of 3D reproduction based on acoustic simulators: an environment is fed into acoustic simulators to produce BRIRs. A spatialiser system consumes these to generate an audio signal considering the listener and scene elements.

significantly enhance immersion and realism (Rubio-Tamayo, Gertrudix Barrio and García García, 2017). Complex sound models that include realistic sound effects such as echoes and reverberations can provide users with more information about their environment, aiding in navigation and interaction (Lokki and Grohn, 2005). In virtual reality gaming, the integration of realistic audio significantly influences player experience by enhancing immersion and engagement. This effect is crucial in genres like horror or adventure, where sound contributes significantly to the atmosphere and tension of the game (Poeschl, Wall and Doering, 2013).

2.3.1 3D Sound Reproduction Techniques

Sound reproduction for immersive acoustics can be defined as a rendering problem concerning providing a listener with synthetic believable auditory stimuli perceived as belonging to a specific space. The rendering is often engineered by adopting a system that has complex scenes and scene elements as input and an acoustic signal as output. Figure 2.8 shows a typical audio system for spatial sound reproduction in VEs.

2.3.2 Spatial Audio Rendering Algorithms

Audio rendering refers to the process of affecting anechoic audio with acoustic phenomena approximated by simulations. It utilises generated IRs and takes into account the characteristics of a human listener from the perspectives of a sound-perceiving object in a virtual environment and a human receiver with psychoacoustic abilities.

One of the fundamental operations in audio rendering is the manipulation of anechoic audio with a filter encapsulating the acoustic effect of an environment to a sound transmission, expressed as an IR. Given anechoic audio expressed as a digital sequence x containing $x_0, x_1, \ldots, x_n$ elements, and an IR expressed as a digital sequence h containing $h_0, h_1, \ldots, h_n$, through the convolution operation $*$ we can obtain the resulting sequence y containing $y_0, y_1, \ldots, y_n$ samples, expressing the resulting signal with the applied IR. The following mathematical notation shows how a new function is created as a result of

the convolution operation:

$$y[n] = x[n] * h[n]. \tag{2.5}$$

In audio rendering terms, these functions will often represent an anechoic acoustic signal that is convolved with an IR to apply to create an auralised resulting signal. Given a signal x as a sample sequence of N points and a filter h as a sample sequence of M points, the resulting full convolution y will be a sample sequence of $N + M - 1$ points. Each sample of the resulting y sequence is the sum of the products of both sequences:

$$(x * h)[n] = \sum_{m=0}^{M} x[n-m]h[m]. \tag{2.6}$$

As shown in Figure 2.5, frequency-domain representation makes specific problems easier to solve compared to the time-domain, and convolution is one example because of the summation required in the convolution process. This summation determines the computational complexity of the operation and grows with increasing M filter lengths. One key property of the convolution is that the product of the frequency-domain representation of a signal with the frequency-domain representation of a filter is the frequency-domain of their convolution. Essentially, the added complexity of summation is removed in the frequency domain at the cost of transforming the signals using the DTFT. Hence, audio rendering algorithms use the much faster Fast Fourier Transform (FFT) Convolution, commonly defined as:

$$(x * h)[n] = IDTFT_N(DTFT_N(x[n]) \cdot DTFT_N(h[n])), \tag{2.7}$$

where $DTFT_N$ and $IDTFT_N$ are, respectively, the DTFT and the inverse DTFT of both the signal and the filter calculated over N frequency points. The Overlap-Add or the Overlap-Save are examples of real-time convolution algorithms often adopted in DSP to implement a wide range of audio effects. They are solutions to the problem of applying BRIR to long signals or to implementing interactive systems, where the listener is displayed rendered audio from a dynamic virtual environment. Thanks to the advances in such algorithms, it is now computationally feasible to manipulate anechoic audio signals with simulated acoustics on the fly, evoking a sense of immersion in the listener due to the auditory stimuli responding to changes in the dynamic environment at interactive rates.

The Overlap-Add algorithm adopts the divide-and-conquer approach towards an acoustic signal by segmenting an input digital sequence into multiple parts, processing the individual parts, and assembling the resulting sequence to produce a whole manipulated sequence. The goal is to evaluate Equation 2.7 over small chunks of audio, storing resulting convolved chunks into a queue from which samples are summed together into an output sequence.

Interactive audio rendering algorithms benefit from such systems as they enable on-the-fly convolution with live audio streams, which are generally implemented as a circular audio buffer. In the case of an immersive application, such audio buffers may be storing audio propagating from sound-emitting objects that interact with the user. As illustrated in

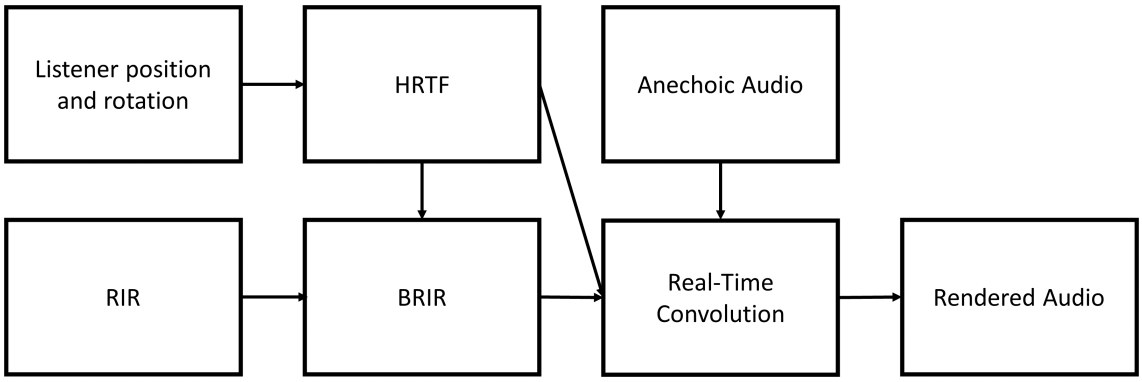


Figure 2.9: Common spatial audio pipeline: the listener's position and rotation in the scene is used to sample an interpolated HRTF from the currently loaded bank, combined with the generated IR, the real-time convolution algorithm applies the BRIR to an anechoic audio signal. The result is an audio signal arriving at the listener, taking into account their position and rotation with respect to the sound-emitting object in the virtual environment.

Figure 2.9, spatialiser systems or acoustic simulation systems provide filters in the forms of IRs that can be applied to audio chunks from the audio buffer.

2.3.3 Auralisations

The ability to auralising anechoic acoustic signals is one of the fundamental objectives in the domains of acoustics for surveying techniques, acoustics for interactive applications, and acoustics in extended reality. As seen in Section 2.2.3, there are DSP techniques that allow the application of acoustic fingerprints onto audio recordings by treating the acoustics phenomena as measurable functions that can be convolved to digital signals, see Equation 2.5. In higher-level terms, auralisation is the process of experiencing audio stimuli in a simulated soundscape, which can be perceiving an orchestra in a digital representation of a church, approximating how room acoustics affect the sound transmission between the orchestra and the listener in the virtual space. There are factors associated with this process that determine how well the resulting signal is able to fool the listener's auditory system into believing that the auralisation is real. Realism and presence are often a function of the performance of the components in the chain of the 3D audio reproduction system; see Figure 2.9.

2.3.4 Common methods for Auralisations

Methods for producing auralisation start from the creation of an environment, which is the first component of the system in Figure 2.9 hosting the virtual sound-emitting objects, e.g. an orchestra, and the virtual sound-receiving objects, e.g. the listener. In computers, environments are generally represented using a broad range of computer graphics techniques, from simplistic Computer-Aided Design (CAD) to fully-featured virtual worlds engineered in modern game engines. Such a statement blurs the definition of a virtual environment, as one could represent a room by creating a photorealistic 3D model or by simply drawing a cuboid. Research on acoustic simulations conducted over the last decades has expanded

towards defining what is required from a virtual environment to produce a believable simulation. The representation of the environment geometry is a determinant of the resolution and perceptual quality of the acoustics simulation results, and, as a general rule, the higher the level of details expressed by the geometry, the more accurate the acoustic simulator is able to simulate how sound interacts with the environment. However, beyond certain levels of details, the increase in resolution does not have a significant perceptual response Pelzer and Vorländer (2010).

Combined with advances in real-space scanning technology and user-friendly 3D reconstruction software, it is now possible to create appropriate virtual environments for acoustic simulations without requiring expert computer graphics engineering knowledge.

2.3.5 Geometrical Acoustics Modelling Techniques

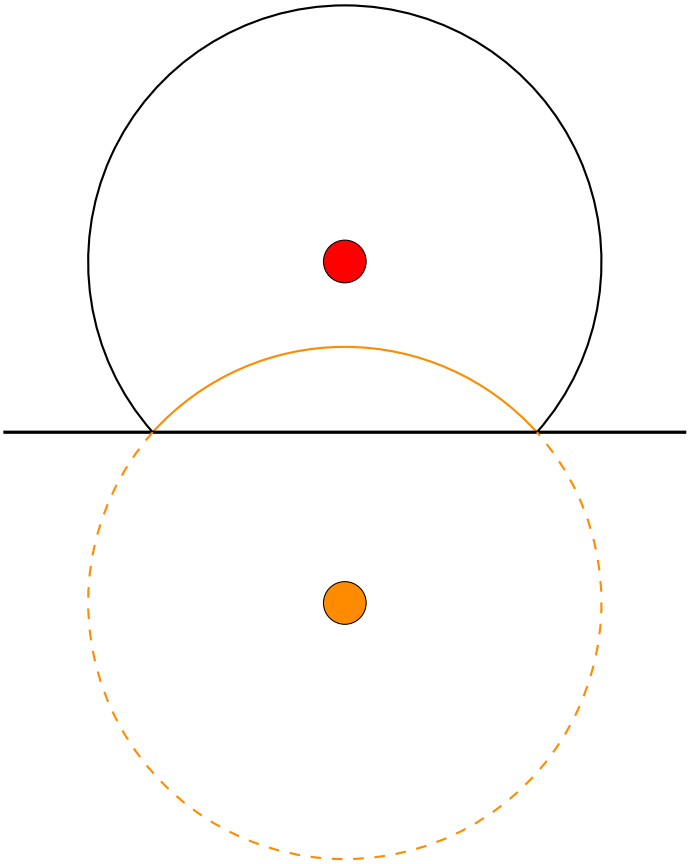
Geometrical acoustics is a family of acoustic modelling methods based on representing propagating sound waves with geometrical primitives, such as rays or cones. The foundations of these methods have sound propagating as straight lines, as opposed to moving particles, approximating the complex nature of acoustic energy transfer but neglecting phenomena related to the wave phenomena (Savioja and Svensson, 2015a). Figure 2.10 visualised the basic principle of an image-source, geometrical acoustics technique: sound propagation is calculated by approximating acoustic reflections bouncing off geometry.

Ray-based techniques, wave-ray hybrid techniques, and wave-based techniques have emerged as prominent methods to generate impulse responses in the field of acoustics, each contributing to understanding how sound propagates within a given space. Ray-based techniques, rooted in geometric acoustics, simulate sound by tracing rays that emanate from a source and bounce off various surfaces within a space. This method approximates reflections and reverberation, contributing to the overall energy expressed as an impulse response.

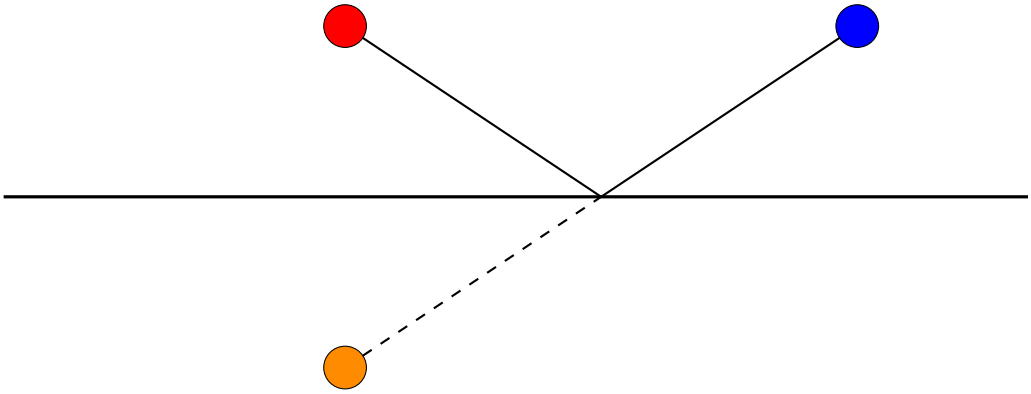
Geometrical acoustics are generally efficient as they are computationally less demanding than wave-based techniques, making them suitable for real-time applications or large-scale spaces often found in cultural heritage contexts. The geometrical nature of ray-based methods allows for easier integration with virtual reconstructions of space and dynamic geometry, and the method's inherent flexibility makes it easily adjustable to different acoustic scenarios (Vorländer, 2008).

Image-Source Model

The Image-Source Model (ISM) is a crucial technique in geometrical acoustics, particularly effective in modelling sound reflections within enclosed spaces. This method simplifies the calculation of reverberant sound fields by treating reflections as emitted from imaginary sources, see Figure 2.11. For each real sound source and reflective surface, a corresponding image source is created on the opposite side of the surface, at an equal but mirrored distance



(a) Visualisation of the basic principle of an image source: a wavefront propagating from an emitter collides with a reflective surface. Specular reflections are computed by creating a mirrored 'image source' reflected around the colliding surface.



(b) With the definition of a receiver (blue circle), an image source can determine propagation paths from an emitter (red circle).

Figure 2.10: A specular reflection calculated using the Image Source Model, demonstrated with a single reflective surface.

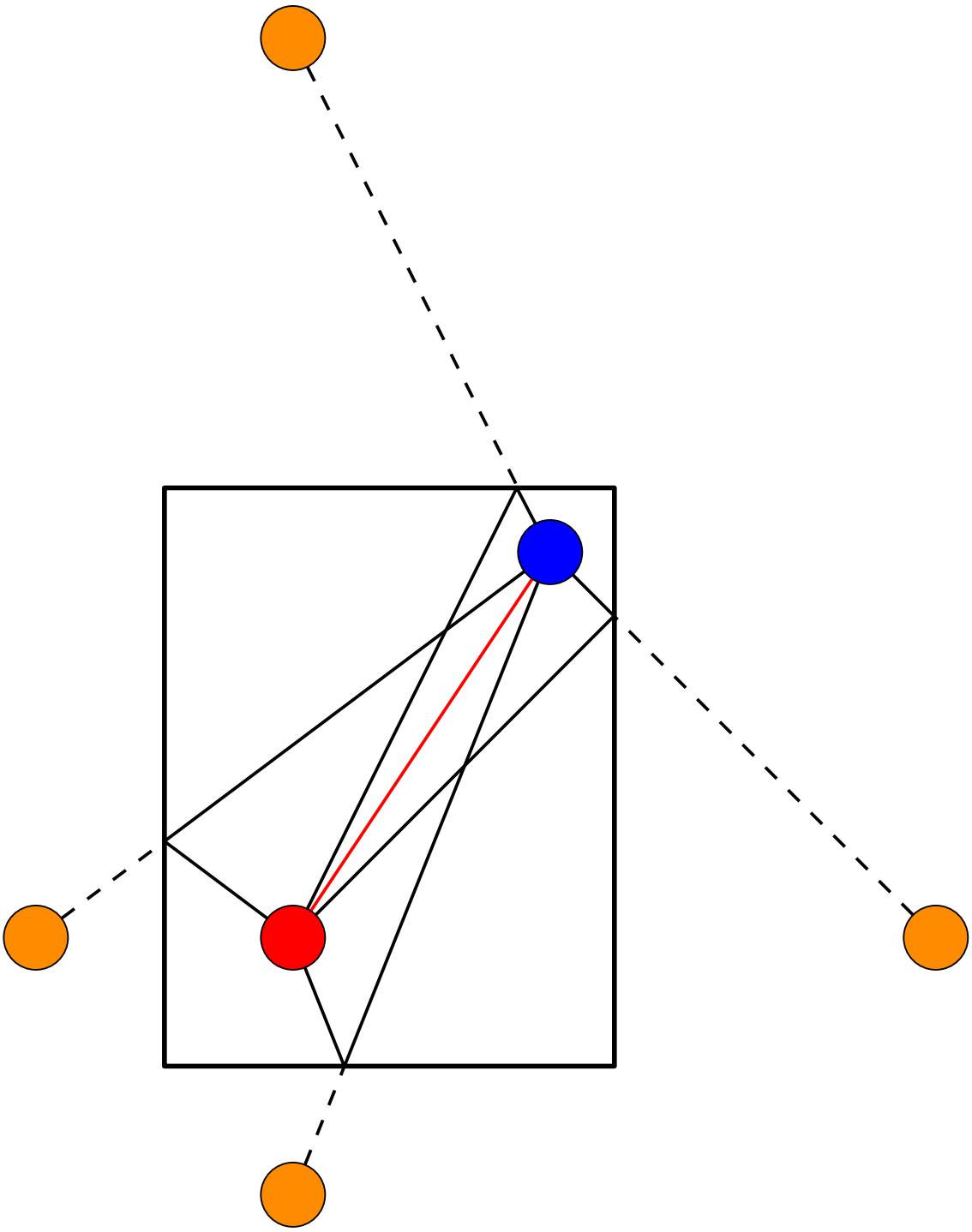


Figure 2.11: Visualisation of specular reflections from an emitter (red circle) to a receiver (blue circle) computed in a shoebox room. Image sources (orange circles) determine reflection points on the room geometry. The direct path (red segment) plus specular reflections computed by the Image Source Model can be used to approximate the basic acoustic features of a space.

from the point of reflection. This setup mimics the path that sound would travel if it directly reached the listener after reflecting off the surface. In environments with multiple reflective surfaces, image sources for higher-order reflections are generated recursively. Each new image source becomes a parent source for further reflections, potentially creating a complex network of sources depending on the geometry of the room and the number of reflections considered. The sound path from the real source to the listener is calculated directly. Paths from image sources are treated as if the image sources were real, with the distance and attenuation calculated based on the geometry and acoustic properties of the environment. The total sound field at the listener's position is the superposition of sound from the direct path, and all reflected paths. Each component is adjusted for delay (based on distance) and attenuation (due to both distance and material absorption properties).

Ray Tracing Techniques

Ray tracing, expanding from ISM techniques, is commonly used to model how sound propagates in an environment by simulating the path of sound rays. As a core method for implementing GA principles, ray tracing simplifies the sound field into discrete rays that carry sound energy, which are traced as they interact with various surfaces in a modeled space (Savioja and Svensson, 2015a). Ray tracing can be considered a Monte Carlo method due to the approximation of acoustic reflections by means of random sampling. At a high level, the technique has rays propagating from uniformly distributed origin points across a sphere, see Figure 2.12.

Ray tracing begins with the emission of rays from a sound source; they represent paths along which sound energy travels through the environment. As rays encounter surfaces, they can be absorbed, reflected, or transmitted based on the properties of the materials they encounter. Rays can typically form specular or diffuse reflections. A specular reflection occurs when rays bounce off smooth surfaces at angles equal to their incidence angles, akin to how light reflects off a mirror. Ray tracing's accuracy in modelling these reflections can vary based on the complexity of the environment, the number of rays emitted from the source and the number of bounces off geometry the model is able to compute (Thompson, 2005). Diffuse Reflections occur when sound rays strike rough surfaces and scatter in many directions; they require more complex algorithms to accurately predict the distribution of reflected sound energy. Specular reflections make the assumption that rays reflect off surfaces in a single, predictable direction. This model is simpler and computationally less demanding but can miss complex interactions in environments with varied surface textures, see Figure 2.12. Diffuse reflections scatter randomly upon striking a surface, often require probabilistic methods to compute, and can be computationally intensive but provide a more accurate depiction of realistic wave behaviors. Some advanced ray-tracing implementations use a combination of specular and diffuse models to capture a broader range of acoustic phenomena and improve realism. The computational load of ray tracing is generally lower than ISM implementations, especially at higher orders of reflection.

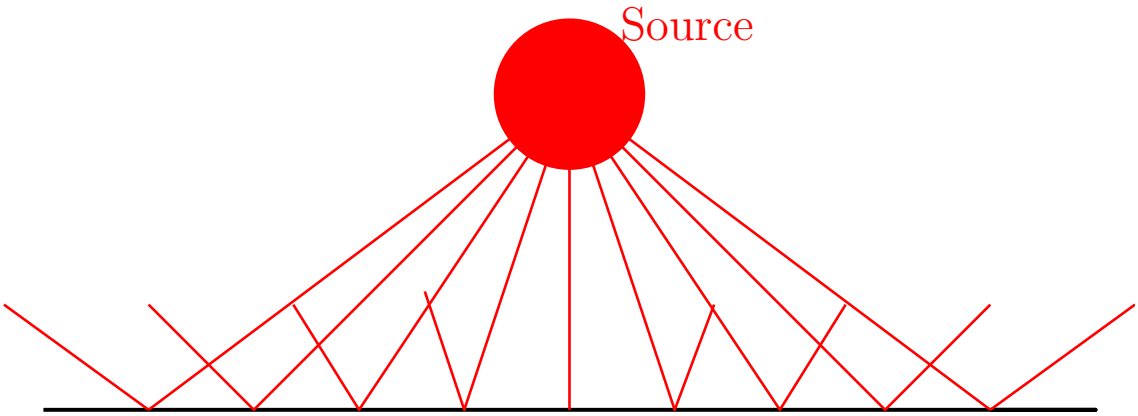


Figure 2.12: Visualisation of a ray tracing source: In a 3D environment, a given number of rays is often emitted from uniformly distributed points across the azimuth and elevation of the emitter surface. Rays emitted from the source and colliding with geometry are reflected around the surface normal at the collision point, generating a new ray.

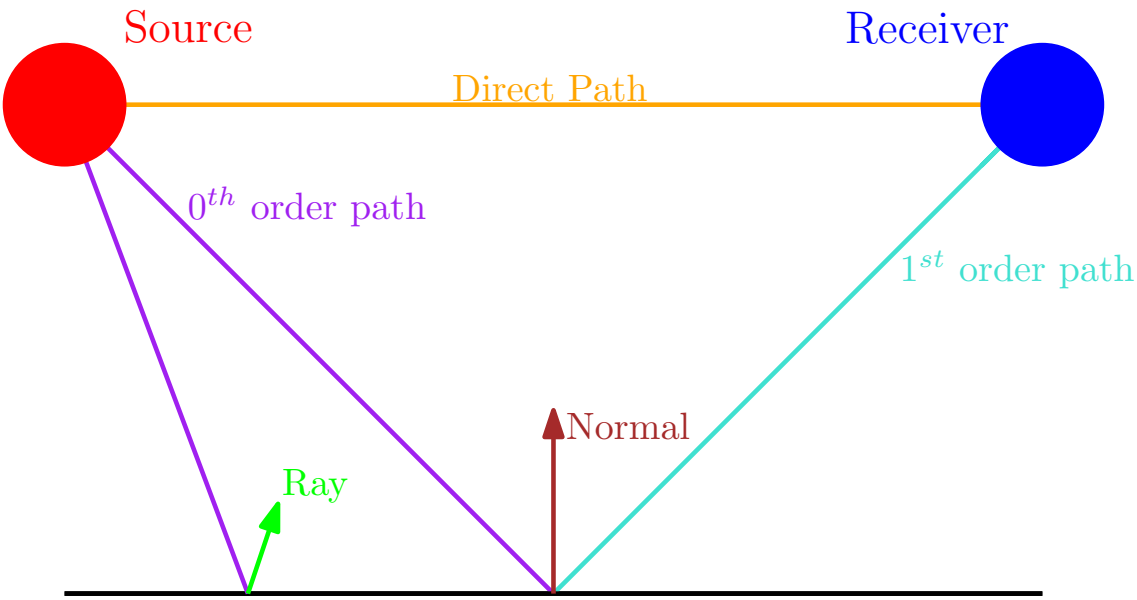


Figure 2.13: Visualisation of a basic Ray Tracing Model for computing specular reflections. Propagation paths from an emitter (red circle) to a receiver (blue circle) are calculated by checking whether reflecting rays intersect the receiver.

This efficiency makes ray tracing advantageous for scalability as it can more easily scale to larger and more complex environments (Schissler, Mehra and Manocha, 2014). Fast implementations also enable real-time applications and make it feasible for integration into game engines. Fast ray tracing implementations also allow for higher reflection order,

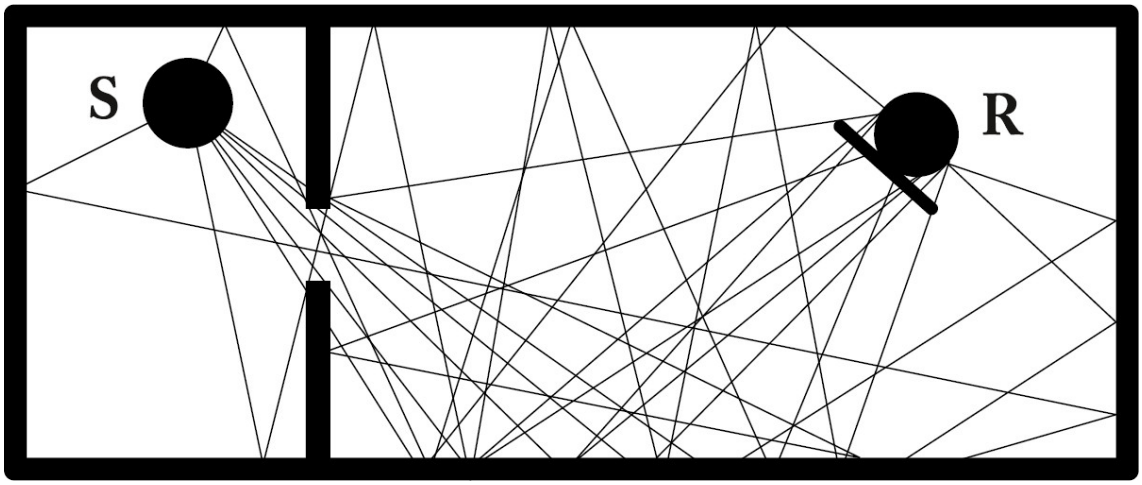


Figure 2.14: A Ray Tracing model computing specular reflections in a shoebox room. The example shows that the model lacks the ability to simulate diffraction effects caused by sound bending through a portal (Funkhouser, Tsingos and Jot, 2003).

achieving realism in simulating acoustic phenomena, whereas the ISM's computational cost increases exponentially with each added reflection order. While ray tracing is versatile and efficient, it still carries GA limitations, such as the high-frequency bias. Ray tracing tends to be more accurate at higher frequencies, where the wavelength is much smaller than the objects and spaces involved, and a ray can better approximate the path of a propagating wave. At lower frequencies, its accuracy diminishes as wave effects like diffraction become more significant as rays are unable to bend around obstacles, see Figure 2.14.

Source Paths

In most cases, all GA techniques discussed in previous Sections approximate acoustic reflections by constructing reflection paths originating from the emitter position and arriving at the receiver. Figure 2.15 shows an example of a set of paths constructed to represent reflections from a source emitter to a receiver for a given simulated environment and source-listener position pair. Each path is subdivided in segments, e.g. source description, room response, or receiver description. Each segment is comprised of nodes that describes how propagating energy is affected by factors relating to the segment, like the emissivity profile of the source or the bounces off the environment geometry to describe energy aborption. In the same Figure, these nodes are represented as functions, H_D, H_E, H_{Air}, H_{Head}. Respectively, these functions express how propagating energy is affected by the source description, the reflections occuring at the current state of the environment and source-listener position pair, the amount of air absorption given current conditions of humidity or temperature, and the current state of the listener expressed as HRTFs.

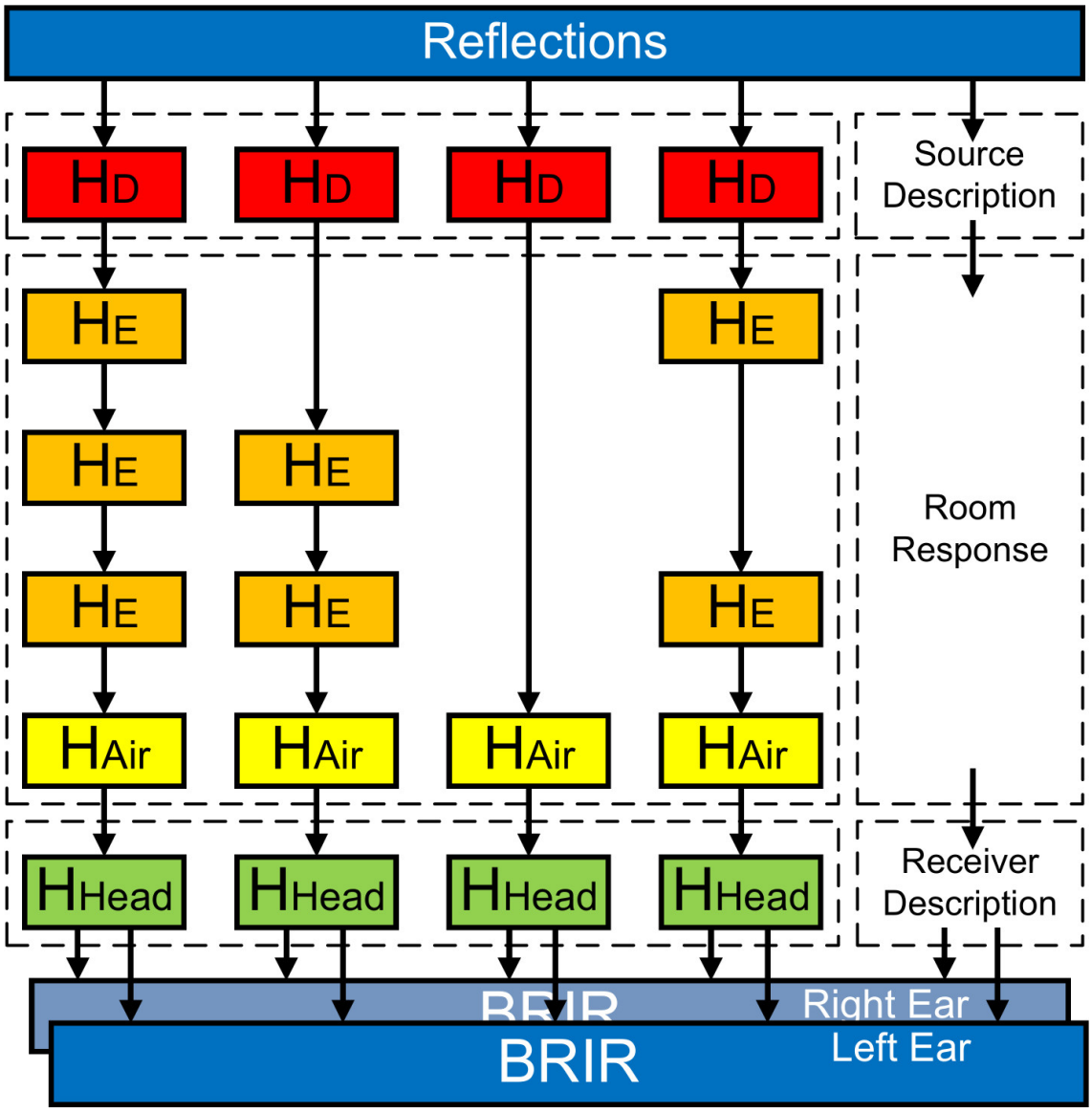


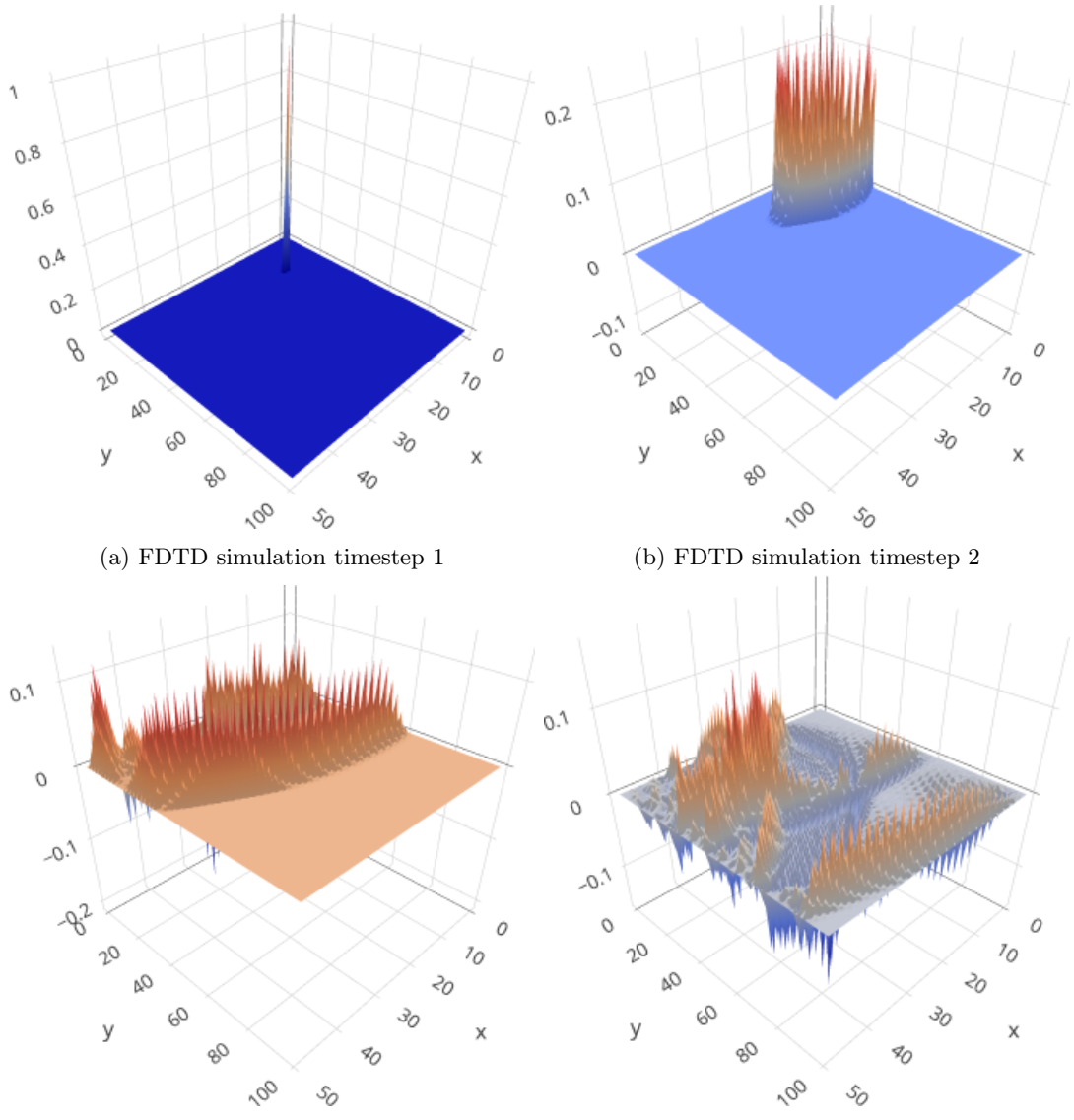
Figure 2.15: Diagram showing the anatomy of a reflection. Geometrical Acoustic techniques consider source directivity profiles and receiver descriptions (human listeners). The diagram shows how a reflection is generated from the source and is affected by source characteristics (directivity profiles). Bounces off the geometry and environment characteristics influence the reflection path through materials. Finally, Head-Related Transfer Functions can be used to produce binaural Impulse Responses (Schröder, 2011).

Wave-based Modelling Techniques

Wave-based techniques stand out for their precision, solving the wave equation to simulate how sound waves propagate through space, accurately modelling diffraction, scattering, and other complex wave phenomena. While highly accurate, wave-based techniques often require substantial computational resources, making them less suited for real-time or large-scale applications (Raghuvanshi and Snyder, 2014).

Wave-based methods account for the full complexity of sound waves, including diffraction, interference, and wavefront curvature—phenomena typically ignored by geometrical acoustics. These methods are based on solving the wave equation, which describes how sound pressure levels vary in space and time. Amongst popular techniques for sound propagation are Finite-Difference Time-Domain (FDTD) methods. These methods discretise the wave equation in both time and spatial domains, using a grid to simulate how waves propagate through a medium (Hamilton and Bilbao, 2017). FDTD methods are widely used in engineering and physics to model sound propagation in complex environments and to study the effects of diffraction and absorption or to model natural phenomena outside the sound domain (Teixeira et al., 2023). In FDTD, the simulation domain is the space truncated by the simulation region and discretised by the mesh. When an FDTD simulation runs, the acoustic energy fields are calculated from wave equations in every mesh cell and the solutions are repeatedly time-stepped, see Figure 2.16. Spatial discretisation allows for the representation of complex geometries and structures, while temporal discretisation captures the evolution of energy fields over time. Subfigures 2.16a-2.16d show propagating energy emission being simulated through timesteps.

These techniques can model wave phenomena with high accuracy and can account for varying medium properties but at the cost of high computational resource requirements. In large environments, they can express complex behaviours associated with wave propagations, like sound bending around obstacles or portals, enabling the generation of realistic stimuli and a coherent soundscape, see Figure 2.17. The Figure shows timestep simulations across timesteps of various environments (one environment for each row in the Figure). However, the computational costs associated with running simulations and computing the wave equation can limit their application for certain tasks. Alongside FDTD techniques, Boundary Element Methods solve the Helmholtz equation, a form of the wave equation applicable to steady-state problems, for the boundaries of a domain, reducing the dimensionality of the problem from the volume to the surface. They subdivide a large problem (such as a room or an outdoor environment) into smaller, simpler parts called finite elements. The sound field in each element is approximated by basis functions. Commonly used in architectural acoustics and the automotive industry, the technique helps in designing quieter and more acoustically pleasant spaces, designing spaces with specific acoustic properties, such as concert halls and lecture theatres, assessing the impact of noise on communities and designing sound barriers to mitigate unwanted noise (Gumerov and Duraiswami, 2021). Boundary methods are particularly effective for exterior problems,



(a) FDTD simulation timestep 1

(b) FDTD simulation timestep 2

(c) FDTD simulation timestep 3

(d) FDTD simulation timestep 4

Figure 2.16: Visualisations of Finite-Difference Time-Domain simulation timesteps. The figures represent several timesteps of a running simulation, showing how sound pressure emitted from a source changes across timesteps.

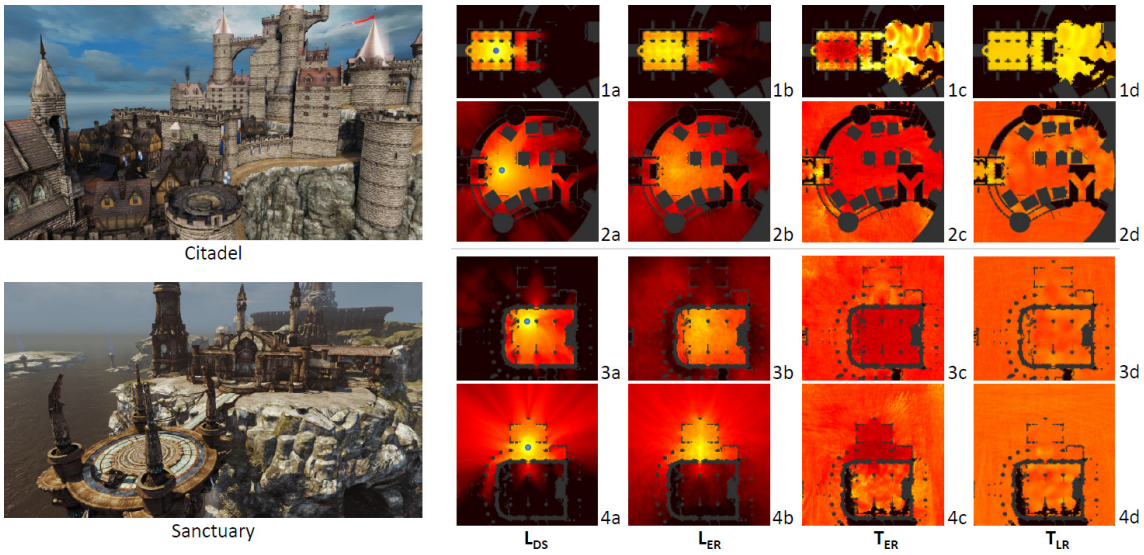


Figure 2.17: Demonstration of a state-of-the-art wave-based acoustic renderer: pressure fields are pre-computed within a complex virtual environment and encoded to allow game engines to decode and auralise sound sources in real-time (Raghuvanshi and Snyder, 2014).

such as noise propagation in an open environment, or noise profiling. They reduces the problem size significantly which can decrease computational demands. The method is less effective for high-frequency sounds where the wavelength is small relative to the dimensions of the modelling domain, as the surface elements need to be sufficiently small. They are highly accurate and versatile, capable of modelling complex material properties and geometries but computationally intensive, especially for three-dimensional problems and high frequencies (Kirkup, 2019).

Hybrid Modelling Techniques

On the other hand, wave-ray hybrid techniques present a more complex picture, combining aspects of ray-based and wave-based methods. Rays are utilised to model the high-frequency components of the sound, while wave equations handle the low-frequency behaviour, attempting to capture the best attributes of both methods. However, the hybrid nature often means more computational resources are needed, and it might not always be the most suitable choice for fast applications where platforms offer limited computational resources (Hulusic et al., 2012). An example experimental method from Southern et al. (2013) shows how a physical FDTD model is used for low-mid frequencies, while high frequencies are handled using beam-tracing or acoustic radiance transfer methods. This approach ensures a balance between computational load and physical accuracy.

2.4 Virtual Environments

Room acoustics simulations may involve the concept of virtual environments to represent sound-emitting objects, receivers, and space where these exist. Computer games technology has shaped the definition of virtual environments over decades of development and progress.

2.4.1 Representation of Virtual Environments

Graphics rendering pipelines display objects of a complex scene to viewers, determining the appearance of materials and geometry of the environment. In VEs, meshes are composed of triangles enabling game engines to organise geometry based on the semantics of scene objects. For E.g. a mug can be represented by triangles grouped in a mesh. They are essentially a network of triangles that connect, having adjacent vertices, to form objects. They are responsible for transforming the scene geometry and applying further processing, such as rasterisation, which generates fragments from geometry combined to create frames. A series of frames generated at interactive rates compose a frame buffer that allows users to experience scenes in real-time. Graphics pipelines describe geometry as vertices and triangles, applying shading techniques to control the appearance of surfaces depending on their lighting conditions and the viewer's spatial position. Here, texture images can also define the appearance of objects' geometry by painting their surfaces and controlling transparency (McAllister, Lastra and Heidrich, 2002; Marschner and Shirley, 2015).

Textures can determine the appearance of material composing objects in a scene adopting two-dimensional images. Texture mapping uses colour and transparency information contained in these images to paint triangles forming the geometry. Texture coordinates provide graphics pipelines with enough information to paint meshes.

2.4.2 Handling of Complex Scene Geometry

Implementing multimodal interactions in VES often requires handling and performing operations on the scene geometry, including searching interactions between entities and the environment. In computer games, physics systems are often fundamental components enabling game mechanics and interactions, which often involve computing intersections between scene entities and the environment. With the growing density of environment geometry and complexity of the scene elements, the computational requirements associated with evaluating these geometry searches have grown, demanding optimal solutions across the space and time domains.

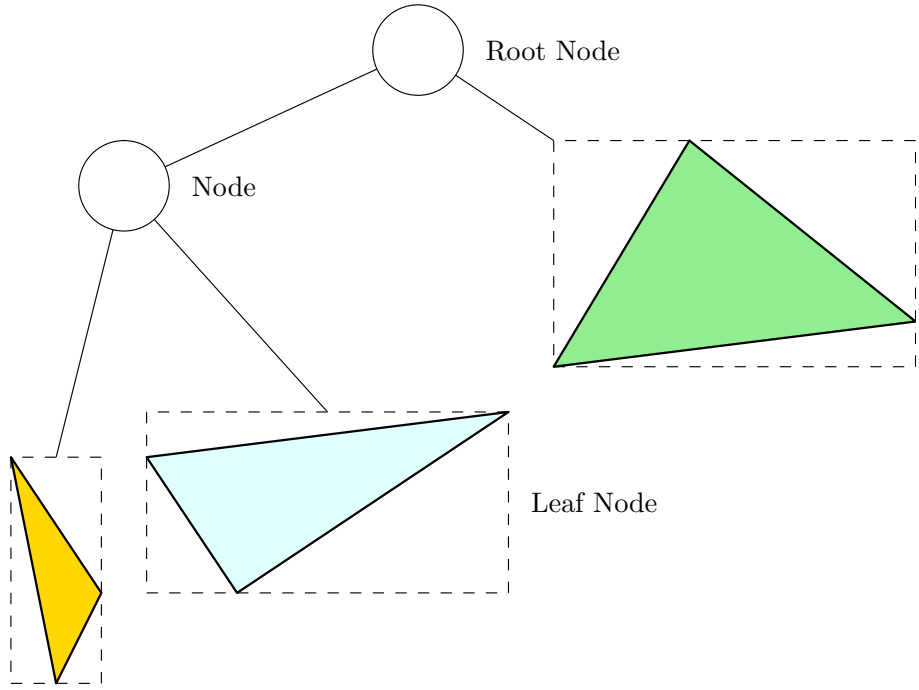
The goal of geometry handling systems is to allow searching intersections between volumes or primitives, such as rays or frustums, and the scene geometry and the engineering design of such systems are closely related to data structures and algorithm design. Data structures space and time complexity

Binary Space Partitioning

One of the first approaches in handling and indexing scene geometry in VEs is Binary Space Partitioning (BSP), which, motivated by performance aspects and limited computational resources available during the early developments of rendering pipelines in the field of computer graphics (Fuchs, Kedem and Naylor, 1980). BSPs allow graphics pipelines to organise the order of scene elements before drawing them or to determine the visibility of

(a) Triangulated Mesh

(b) Bounding Volumes



(c) Tree

Figure 2.18: A bounding volume hierarchy constructed on a given input scene represented as a triangulated mesh (a). Bounding volumes encapsulate mesh primitives, (b), which then represent nodes of the tree, (c).

surfaces.

The goal of BSPs is to index and search scene elements or geometry primitives, part of a given input scene. The technique works by subdividing the Euclidean space in which scene elements exist. The space is divided by partitioning planes, separating scene elements based on which side of the plane they exist. The process repeats recursively, subdividing space with further partitioning planes. Several criteria can determine the number of further subdivisions, such as the minimum size of regions generated by space subdivisions or the indexing complexity and granularity required for indexing and searching operations.

With subdivided regions obtained with partitioning planes, a binary tree, similar to the diagram shown in Figure 2.18, where a root node refers to the entire scene and branches into the first space subdivision, which recursively branches into further subdivisions, until a "leaf" region. A leaf region can hold a scene element, a geometry primitive, or a subset of primitives from the set of primitives representing the input scene.

Bounding Volume Hierarchies

A Bounding Volume Hierarchy (BVH) is a method closely related to BSP for handling scene geometry that optimises intersections between rays and the scene by adopting a binary tree to subdivide primitives that compose the scene geometry. A BVH can represent a scene by constructing a binary tree partitioning geometry primitives into a hierarchy of disjoint sets. In physically-based rendering applications, mesh triangles are often the primitives indexed by the constructed tree; see Figure 2.18 (Pharr, Jakob and Humphreys, 2023).

In a tree, bounding volumes are generated to fit primitives from a given triangulated mesh (Figure 2.18a) and aggregated based on proximity (Figure 2.18b). Bounding volumes encapsulating multiple primitives generate branches, and recursively, branches are encapsulated in volume until a root volume fits the entire input scene. A constructed tree (Figure 2.18c) can be queried and traversed by navigating branches from the root node to primitives within leaf nodes.

Thanks to branch subdivisions, ray-volume intersection tests performed on nodes allow filtering out entire segments of the scene, reducing the set of primitives that potentially intersect the ray to a subset of the input scene triangle set and improving the space complexity of the operation, much like in BSP techniques. Recent research trends are exploring tree rotations and balancing of branches in real-time, optimising search operations even further, and allowing the tree to reflect dynamic changes to the scene geometry (Kopta et al., 2012).

2.4.3 Materials

In computer graphics and multi-modal rendering, assigning physical properties to surfaces of the scene geometry has been addressed with the definition of materials. The definition of materials is often intrinsic to the rendering technique and to the engineering design of

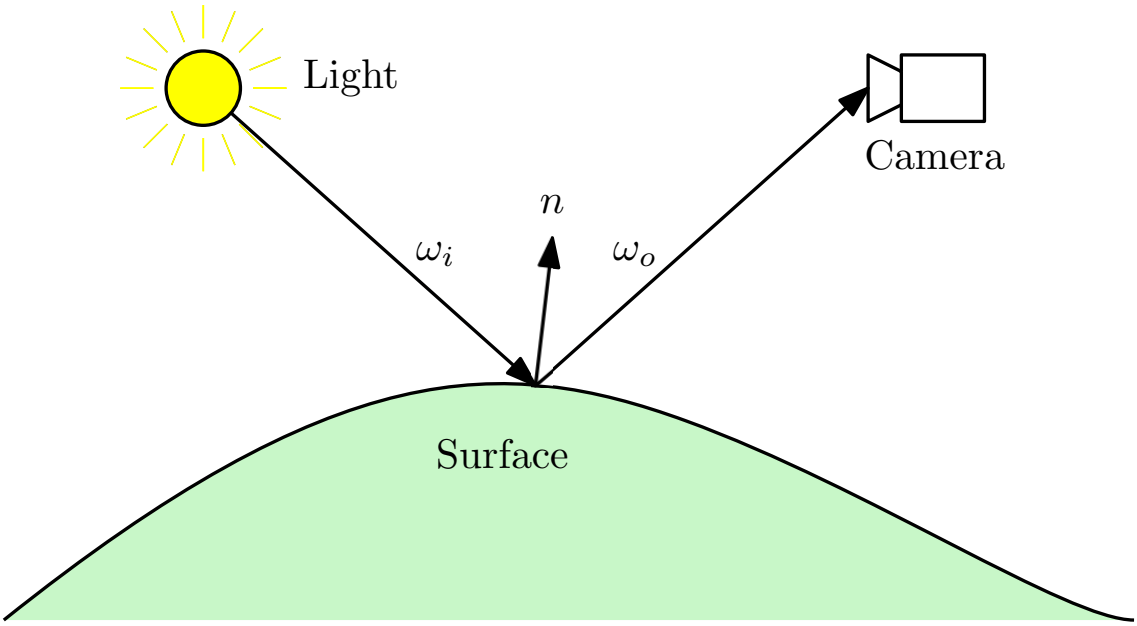


Figure 2.19: A simplistic material system visualised as a ray of light emitted by a source and colliding with a surface. The ray reflects around a surface normal and is detected by a virtual camera.

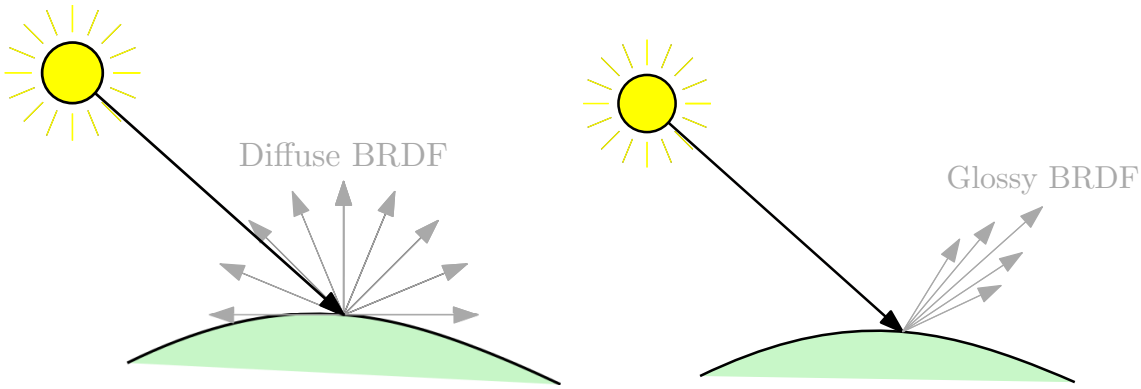
the rendering apparatus. In physically-based graphics rendering techniques, Pharr, Jakob and Humphreys (2023) define materials as a

> "description of its appearance properties at each point on the surface".

Materials in Rendering Pipelines

Rendering pipelines often model how surfaces in complex scenes reflect and respond to propagating energy by employing Bidirectional Reflectance Distribution Functions (BRDFs). In the light domain, rendering techniques often model reflected energy as a function $f_r(p, \omega_i, \omega_o)$ of a p BRDF, an incoming direction ω_i and an outgoing direction ω_o. A simplified diagram in Figure 2.19 shows how energy transmits from a light source and is sampled by a virtual camera, with a surface reflecting the light ray around the surface normal at the collision point. Though, materials in the real world have unique physical properties affecting reflectance, absorptions, or diffusion of incidental light, deviating from the idealised model shown in Figure 2.19.

A material using a reflectance function can model realistic behaviour, allowing surfaces to express varying physical attributes like roughness or metallic characteristics. Figure 2.20 shows example functions simulating a rough and a glossy material, Figure 2.20a and 2.20b respectively. These examples show BRDFs modelling the scattering of energy caused by the rough surface and the glossy reflections caused by a mirror-like surface.



(a) Rough Material (b) Glossy Material

Figure 2.20: Example BRDFs applied to a surface, defining a rough material (a) and glossy material (b). These functions emulate how properties of real surfaces respond to colliding light: glossy materials will reflect energy specularly, whereas uneven rough surfaces will cause diffuse scattering.

Materials in Sound Rendering

Defining materials translates to the acoustics domain, applying closely related principles defined in the visual domain. GA methods often share the same approach shown in Figure 2.19 by considering sound in VEs as propagating rays (or other geometry primitives) colliding with surfaces that reflect energy based on attributes assigned to the surface.

In real soundfields, acousticians and architects often plan the presence of certain materials to control aspects of sound propagation within a given environment. Studies show that strategic placements of surfaces with high acoustic absorption characteristics can have a positive subjective influence on perception in environments, improving the clarity of acoustic information transmitted within the space (Arvidsson et al., 2021). Absorption panels, diffusers, or bass traps are some example materials and surfaces that acousticians use to control how acoustic energy reflects around the environment, controlling parameters like T_{30} or T_{60} reverberation metrics or C_{50} and D_{50} clarity and definition metrics, respectively.

Modern game engines and acoustic simulation software aim to replicate the behaviour of these surfaces by encoding acoustic characteristics to scene geometry representing an environment. In GA simulation methods, material characteristics like absorption or scattering coefficients can influence of geometry primitive simulating propagating sound and interact with the environment, similarly to BRDFs (Rindel, 2000). Finally, acoustic material can encode frequency-dependent acoustic information, often expressed around Equivalent Rectangular Bandwidth (ERB) frequency region, to consider aspects of the HAS. Chapters 4 and 5 will discuss the use of materials in the context of the overarching aim of this thesis.

2.5 Deep Learning Background

Subsequent chapters of this thesis will leverage deep learning techniques to solve a subset of problems associated with developing a system targeting the overarching aim. Specifically, acoustic materials are central to applying the system to realistic, complex scenes, as they contribute towards the perceived quality and realism evoked by the auditory display. The problem arises from the complexity of mapping the appearance of surfaces within the complex scene to acoustic materials. Many factors in complex virtual environments influence the appearance of surfaces, making it hard to distinguish surfaces and map them against acoustic materials automatically.

Deep learning is a subset of machine learning comprising techniques and pipelines to address such mapping problems by learning from examples and providing a generalised model for unseen cases. The potential of deep learning lies in the feature extraction process, allowing models to learn from examples influenced by many factors. The term deep learning is associated with the feature extraction process, delegated to layered feature extraction components composing the model (Dolhasz, 2021).

2.5.1 Image Processing

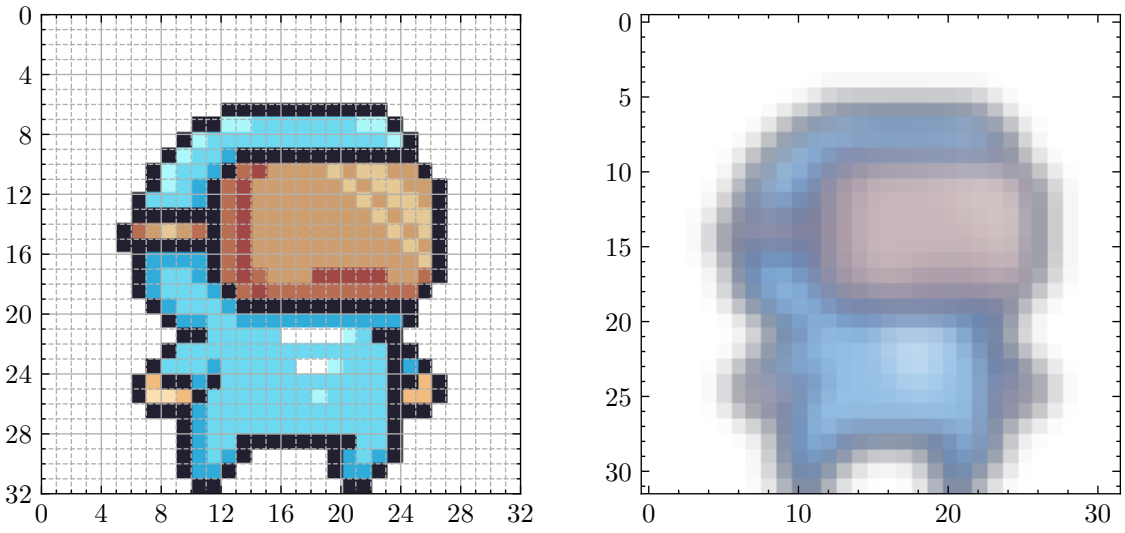


Figure 2.21: Visualisation of an image processing technique applied to an example image (left axes). On both axis pairs, abscissa and ordinate indicate pixel coordinates. A Gaussian blur filter is applied to the image by means of convolution applied to pixel intensity values representing the image. The right axes show the result.

Image processing is the act of performing operations on an image in order to enhance it or extract information. It involves manipulating pixel data to improve the image or to analyze it. This field intersects computer science, mathematics, and a broad range of engineering disciplines. Similarly to digital audio signals, as overviewed in Section 2.2.3, images are often represented in digital systems by encoding intensity values of colours registered by an analogue-to-digital converter (Marschner and Shirley, 2015). Figure 2.21

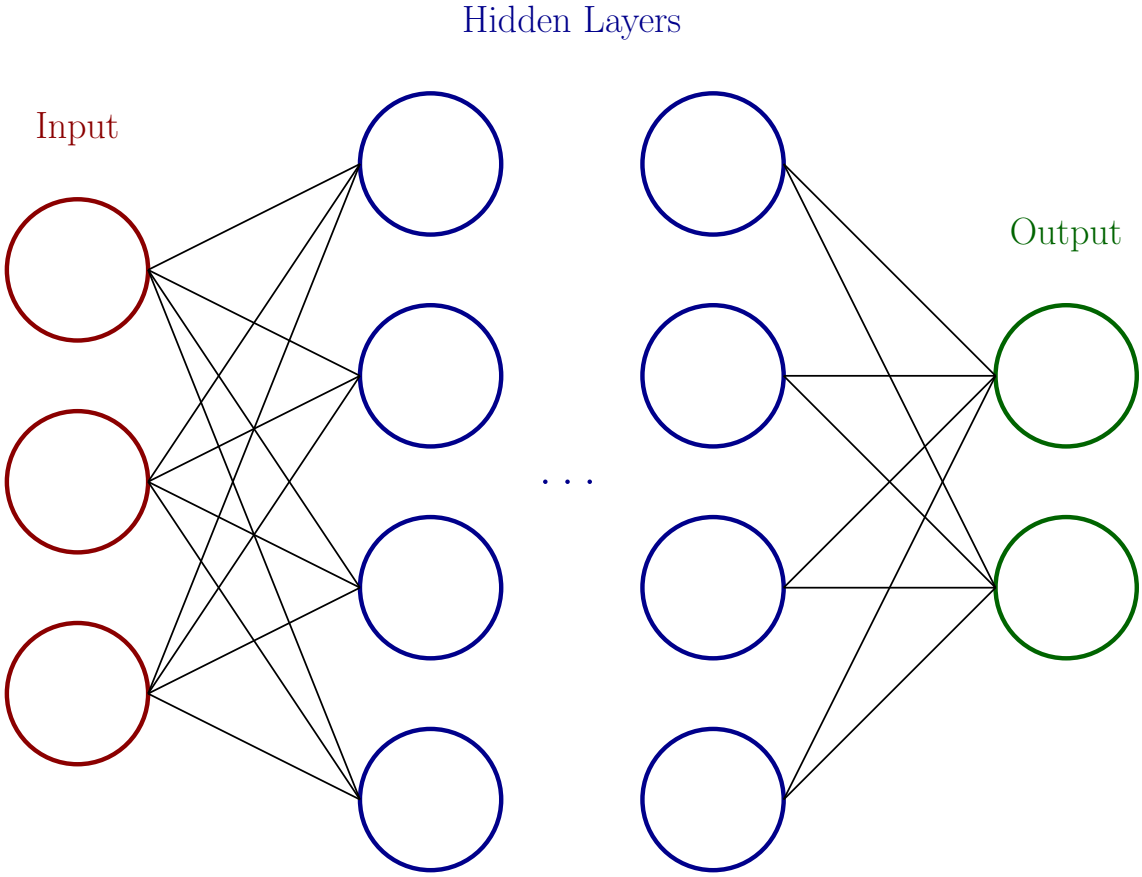


Figure 2.22: A visualisation of a basic neural network with fully connected hidden layers.

shows an example image processing operation where pixel values expressing pixel intensity values are represented as $\mathcal{RGB}$ triplets and expressed as a three-dimensional matrix that spans across the width and height of the image in pixels. Through convolution, a Gaussian filter is applied to the pixel values, rendering a processed image.

Image processing is fundamental to deep learning as operations like convolution are often used to apply filters for several applications (Goodfellow, Bengio and Courville, 2016). Convolutional neural networks build upon the process of decomposing a given image with multiple filters, from coarse to fine, matching patterns and shapes, expressed through kernels, which can be learned by neural networks as the next Sections will demonstrate.

2.5.2 General Machine Learning Tasks and Applications

Deep learning techniques discussed in this thesis can be broadly categorised based on the learning approach into supervised learning and unsupervised learning. Supervised learning is an example where the algorithm learns from a labelled dataset, understanding the relationship between the input features and the target output. The goal is to predict the output for new, unseen data based on this learned relationship. Unsupervised learning algorithms, on the other hand, analyse unlabeled data to find patterns or inherent structures.

The goal of a model is to provide inference on unseen data based on training on a set of

representative examples, emulating basic human abilities that are hard to programmatically engineer in computers. Typically, deep learning models consist of an input layer ingesting data such as images or audio signals, deeper layers extracting features from input data, and output layers transforming the extracted features to perform a task. A loss function is often employed to measure the error and accuracy of the output data and improve the fitness of the model through backpropagation: the tuning of weights and biases of neurons in hidden layers based on computer errors, see Figure 2.22. With layers fit on a dataset, the model can infer output from unseen data, generalising on the task at hand (Szeliski, 2022). General deep learning tasks related to this thesis include classification, regression, or synthesis.

Classification is a type of supervised learning where the goal is to predict the category or class of an input. The input data is fed into the algorithm, which then outputs a label from a predefined set. For example, a classification model might be used to predict the presence of objects in an input image. Classification can be binary (two classes) or multiclass (more than two) (Goodfellow, Bengio and Courville, 2016). Classification models are often evaluated by using Mean Squared Error (MSE) or error metrics that determine the distance between the model prediction and a data point sampled from an evaluation set.

Regression, another type of supervised learning, involves predicting a continuous quantity instead of a categorical label. The aim is to find the relationship or mapping between input variables and a continuous output variable. An example application is to predict parameters of audio engines like pitch or amplitude based on input physics factors like mass (Colombo et al., 2021). Regression models are evaluated using different metrics than classification models, such as Mean Squared Error (MSE). The choice of metric often depends on the specific requirements of the task. Deep learning techniques are becoming increasingly popular in virtual environment pipelines due to the flexibility and potential to adapt to various tasks.

Synthesis refers to techniques for the generation of novel data based on a set of input parameters. The domain of digital audio has benefitted from the evolution of machine learning techniques for synthesising audio data like sounds or synthesing acoustic environments based on an input representation of the soundscape (Liu and Manocha, 2020; Ahn et al., 2023). Synthesis tasks can often be evaluated using GAN-like architecture, as demonstrated by Singh et al. (2021), where a model that synthesises acoustic responses trains against a network, the discriminator, that distinguishes where the synthesized output is real or synthesized. Here, the synthesis network trains until the discriminator fails to detect whether the output is synthesized or part of the training set.

2.5.3 Deep learning Tasks Within Immersive Applications

Detecting the presence of certain objects in an image represents a milestone in the development of CNNs and computer vision techniques as it emulates a basic task of the human

visual system. Due to the nature of image representations in computers, as described in Section 2.5.1, recognising entities depicted by images is a central problem in computer vision (Szeliski, 2022). Classic computer vision algorithms have approached the problem by providing algorithms to recognise patterns programmatically by filtering the image or scanning for certain features. Thanks to advances in CNNs, object detection was addressed by extracting features using deep layers and learning from annotated examples expressing a set of classes captured in various contexts.

Object Detection Pioneering large-scale labelled datasets, such as the work by Deng et al. (2009) on ImageNet, enabled object detection networks to improve their efficiency and abilities of recognising classes. Of pioneering importance is Redmon et al. (2016)'s You Only Look Once (YOLO) network that introduced a state-of-the-art solution able to recognise thousands of classes with high accuracy and precision.

Image Segmentation Similarly to object detection, the task of image segmentation involves dividing a three-dimensional scene into its constituent entities, with the aim of identifying and categorising different segments based on features extracted. This task is crucial in robotics, autonomous driving, and MR, where understanding the structure and layout of the environment is essential for navigation and interaction. Techniques like point cloud segmentation and voxel-based approaches are commonly employed, leveraging deep learning models to process and classify 3D data (Minaee et al., 2022; Feng et al., 2020; Kalogerakis et al., 2017).

Pose Estimation Pose estimation refers to the task of determining the position and orientation of objects or individuals within a scene. In human pose estimation, this typically involves recognition of gestures from HMD cameras to enable human-computer interaction (Andriluka et al., 2014; Spittle et al., 2022).

Scene Reconstruction Scene reconstruction tasks can create a complete 3D model of a scene from a series of images or video frames or generally sparse input information. This can involve reconstructing the geometry of the environment, textures, and lighting conditions (Patow and Pueyo, 2003).

Sound Source Separation Sound source separation tasks involve isolating individual audio sources from a mixture of sounds. This is common in audio engineering to improve the clarity of speech in noisy environments, for instance. Techniques often involve signal processing methods and machine learning models designed to distinguish between different sound characteristics (Virtanen, 2006).

Audio Scene Understanding Audio scene understanding is the process of interpreting audio signals, detecting the presence of auditory elements like footsteps or speech, and

discerning the context or setting. It is akin to scene recognition in computer vision but applied to auditory inputs (Abeßer, 2020).

Sound Propagation Modelling Sound propagation modelling involves using deep learning models to simulate how sound waves travel and interact with the environment, including reflection, absorption, and diffraction around obstacles. This is crucial in acoustics engineering, game development, and architectural design to create realistic sound environments and to analyse the impact of sound in physical spaces (Liu and Manocha, 2022).

Measuring Perceptual Similarity Perceptual similarity tasks involve determining the perceptual distance between two stimuli and emulating human perception rather than computing pixel-level or waveform similarities. This requires understanding the features that humans consider important in judging similarity, a task with applications in image retrieval, content recommendation, and quality assessment (Dolhasz, Harvey and Williams, 2020).

Autonomous Behaviour Modelling Agency in games and reinforcement learning are deep learning approaches to simulate the capacity of players (or entities within a VE) to make choices and perform actions that affect the environment. Example uses of reinforcement learning include training agents to perform navigation in complex 3D environments using audiovisual stimuli, or training a robot to perform actions in a virtual environment that can be mirrored in the physical world (Yannakakis and Togelius, 2018; Matulis and Harvey, 2021).

2.6 Conclusions

The current state of interactive sound rendering allows for fast acoustic simulations, even on platforms with limited computational budgets, approximating the soundfield of any given environment, where a listener can experience realistic auditory interactions with virtual sound sources (Lakka et al., 2018; Hulusic et al., 2012). Sound rendering can be considered a fundamental component of computer games technology, responsible for reproducing everyday sound emitted by objects or agents in a virtual scene and perceived by a listener. This poses the challenging task of reflecting basic acoustic principles to render such auditory interactions realistic. In the real world, sound propagates from a sound source to a listener and interacts with objects in the environment and with the environment itself arriving at the listener's ears (Kuttruff, 2016). Sound cues alone are sufficient to enable users in VEs to pinpoint locations of sound-emitting entities in a scene by using auditory sound localisation, a natural ability associated with the human auditory system (Lokki and Grohn, 2005; Rubio-Tamayo, Gertrudix Barrio and García García, 2017).

As the acoustic principles that govern how sound propagates in space are difficult to reproduce in digital systems, many methods exist, providing variable orders of approximations, depending on the application. Such approaches emulate the wavefield of an environment, simulating how sound interacts with boundaries and scene objects. A subset of these can reproduce phenomena of sound, such as diffraction, reflection, and refraction, which are determinants of realism as they emulate how waves bend around obstacles. Such phenomena make the simulated wavefield dependent on the accuracy of scene geometry and materials represented in a VE.

There is a large tree of techniques and methods to simulate sound propagation, reflecting acoustic properties to any given sound source in a VE, adapting to perceptual requirements and computational budgets available Doukakis et al. (2019). As a general rule, the more computational budget available, the more complex techniques can be employed, allowing realistic sound rendering. Finite-difference Time Domain (FDTD) approaches shown by Hamilton and Bilbao (2017), or wave-based by Raghuvanshi and Snyder (2014) methods, on this end of the spectrum, obtain high degrees of accuracy and realism, but often require precomputation stages or GPU implementations to produce acoustic simulations at interactive rates. Wave-based methods provide a detailed and rigorous approach to acoustic modelling, ideal for scenarios where high fidelity and accuracy are necessary, particularly when dealing with low frequencies and complex interactions. While computationally more intensive than geometrical acoustics methods, the depth and realism they offer make them indispensable in many advanced acoustic studies and applications.

On the other end of the spectrum, there are fast geometrical acoustics methods, widely adopted in real-time applications due to their low computational requirements and highly parallelisable implementations (Cowan and Kapralos, 2010), which reduce simulated sound waves to rays or beams, that are much simpler to compute. Ray tracing is a robust and versatile technique in geometrical acoustics, widely used for modelling room acoustics due to its computational efficiency and adaptability to different reflection models. While it offers broad applicability and scalability, its accuracy in certain scenarios, particularly in complex, diffuse environments and at lower frequencies, may be limited. Finally, hybrid methods also exist to combine the strengths of the main families.

Acousticians and engineers have always employed classic sound rendering to solve practical problems as it requires the work of experts to adjust parameters and define the acoustic characteristics of a virtual scene. A constant here is the requirement of an accurate description of the environment, detailing the geometry of architectural components and objects contained within with acoustic information such as acoustic energy absorption, reflection, or scattering — this is essential to model the behaviour of sound waves interacting with the environment.

Only recently, with the increase of processing power available in computers, sound rendering has gained popularity in computer games and immersive technology for entertainment

and serious applications (Zhang et al., 2018). MR technology can particularly benefit from this as the increase in processing allows sound rendering on mobile devices, enabling listeners to experience virtual sound sources propagating in the reconstruction of real geometry, which is the main avenue that the planned thesis work aims to explore. The following Chapters will delve into elements of sound rendering and material recognition techniques to investigate the adoption of immersive acoustics for wearable technology.

Chapter 3

Advances in Visual-Acoustic Mapping Methods and Sound Rendering Pipelines

In light of the grounding provided around the domains of wave theory, digital signal processing, acoustics and soundfield simulations, as well as domains of computer graphics and vision, virtual environments, and immersive displays, this Chapter reviews the current state of research intersecting the primary aim of the thesis. Throughout this work, Sections of this thesis may re-iterate objectives to provide the reader with context relative to the Chapter or Section at hand.

As the overarching goal of this work is to explore the potential of computer vision within acoustic rendering pipelines for realistic sound transmissions between virtual sound-emitting objects perceived by a user in a virtual environment, experimental methods and novel systems are reviewed. The engineering process supporting the overarching aim addresses a set of problems arising from various facets of the sound rendering pipeline. With the major problems being addressed over the development of individual components of the system, this Chapter discusses how recent work has developed similar pipelines for tasks around visual-acoustic matching problems, auralisations, or sound rendering pipeline designs.

Advances in neural computing and computer vision are providing researchers with increasingly powerful and generalisable tools to address sound rendering tasks, generating an overwhelming volume of new research and novel systems. Although every effort is being made towards reviewing cutting-edge and state-of-the-art work on these domains, claims or information provided on techniques might become out of date or inaccurate at the time of reading this work. The following Sections gather pioneering work, state-of-the-art, and experimental methods to address problems or tasks associated with each component of the thesis aim. Methods and experiments are reviewed, considering limitations and expansion points to inform design choices in engineering systems proposed throughout the following

Chapters. Discussions around existing work aim at both orienting the reader towards the goal of each domain associated with the thesis component and defining the value of the contributions stemming from this work.

3.1 Introduction

The general trends of computing are heading towards ubiquitous wearable computing to facilitate specific or everyday tasks or provide entertainment with increasingly realistic and interactive multimodal interactions (Al-Ghaili et al., 2022; Slater et al., 2009; Park and Kim, 2022; Rubio-Tamayo, Gertrudix Barrio and García García, 2017). An increasing number of domains is able to involve wearable technology from the realm of XR in tasks. Wider industry domains are exploiting the potential generated by recent advances in graphics, game engines, acoustic rendering and neural computing. In industry applications, VEs are often experienced via immersive and interactive wearable computers, such as XR platforms, incentivising manufacturers of HMDs to accelerate the development of wearable computing platforms, equipping them with better sensing technology and more accurate interaction apparatuses.

Realistic interactions between the human sensory system and VEs allow users to perform tasks in a virtual world by transfering skills learned in the real world (Jot et al., 2021; Woodward, 2021). This can provide an effective and efficient tool for training purposes, as users can experience a variety of scenarios and learn to perform novel tasks whilst reducing the costs of setting up training scenarios in the physical world. Auditory interactions, as determinants of realism, can affect the perception of virtual world and ultimately affect task and learning performance. Hence, the focus of audio systems in wearable computing should aim at allowing users to use auditory interactions to perceive the virtual worls surrounding them (Yang, Barde and Billinghurst, 2022).

3.1.1 Current Trends of Interactions within Immersive Platforms

The domain of immersive acoustic has recently attracted more popularity thanks to the drive of XR platforms towards better and more efficient multimodal interactions. Over the last decades, the research fields of interaction, rendering, and visualisation in VEs have generated a significant body of literature (Kim et al., 2018).

Interaction techniques have seen an increasing number of experiments around immersive tasks performed by users within XR technology. The range of applications employing HMDs across domains of research and industry requires virtual interactions to simulate human-to-human interactions in their realism and completeness. The spectrum of interaction techniques reviewed by Spittle et al. (2022) shows that auditory interactions, such as speech input systems or language processing techniques, are widely adopted to alter or manipulate virtual scene elements.

Auditory interactions have generated various subdomains in XR domains with the aim of overcoming obstacles caused by factors influencing the relationship between aural cues and acoustic characteristics of the environment (Park and Kim, 2022). More specifically, Yang, Barde and Billinghurst (2022) have reviewed a body of literature around experiments towards auditory displays in augmented reality, forging the term Audio Augmented Reality (AAR). The authors review research focusing on auditory interactions in MR outlining crucial research problems affecting task performance, realism, presence or subjective factors associated with human perception. Some example research problems relating to the overall immersive experience revolve around acoustic factors of sound transmissions between physical or virtual entities and the user as a listener. For instance, Mansour et al. (2021) show that speech intelligibility in immersive environments perceived through ambisonics displays is a problem affecting XR in noisy soundscapes and hindering accessibility for users with hearing impairments. Such problems incentivise the field of AAR to develop audio pipelines considering and compensating for acoustic factors of the environment.

3.2 Review of Sound Rendering Pipelines for Immersive Environments

Naef, Staadt and Gross (2002) present a pioneering architecture for spatialised audio rendering for virtual environments experienced through immersive headsets. They define 3D sound localisation, room simulation, live audio input and efficiency as the main requirements the architecture should feature. The architecture draws from low-level rendering pipelines, such as graphic sub-systems, to integrate sound rendering procedures into existing scene-handling systems adopted by these pipelines.

3.2.1 Advances in Sound Rendering

Methods have been proposed to map visual representations of environments to their acoustic features. Sound rendering in virtual environments can leverage such mapping for producing audio stimuli conveying spatial information to the user. Recent work solves tasks within sound rendering for virtual environments, such as propagating audio within virtual environments.

Spatial sound has a significant effect on the sense of presence and immersion for a user in a VE (Poeschl, Wall and Doering, 2013). Factors of accurate and plausible acoustic rendering include geometry, material definitions and a room impulse response which describes the attenuation of sound from a sound source to a listener, and there exist approaches that tackle varying aspects of these factors. A common denominator in the sound rendering methods mentioned is the problem of mapping the visual representation of environments to corresponding acoustic materials, which intersects image processing and computer vision domains aimed at modelling how human vision recognise materials.

Acoustic rendering can reproduce spatial hearing abilities (Lokki and Grohn, 2005), supporting architectural acoustics, cultural heritage (Berardi, Iannace and Ianniello, 2016; Vorländer et al., 2015), and computer games (Raghuvanshi and Snyder, 2014; Mehra et al., 2015) to build compelling, realistic acoustic simulations. Recent advances in wavefield synthesis have made it easier and computationally feasible to apply to VEs (Raghuvanshi and Snyder, 2014). They draw on geometrical acoustics, wave-based or hybrid sound propagation algorithms, simulating sound propagation by tracing rays or beams (Hulusic et al., 2012); solving the wave equations at discretised junctures of the representation of the environment or by a combination of the former techniques. These techniques enable virtual complex scene designers to apply realistic, spatialised audio to immersive applications and are becoming part of standardised workflows in game engines.

In acoustics, it is common to capture an environment adopting measurement techniques such as the sine sweep, usually consisting of reproducing a logarithm sine chirp or a short burst, e.g., a gunshot, emulating a Dirac-delta function to excite frequencies in the audible spectrum and recording how the environment influenced the propagated sound at the listener position (Reilly and McGrath, 1995). Such measurements can determine a Room Impulse Response (RIR), a series of reflection paths over time, recreating the acoustic space for a given source-listener position pair. Wave-based acoustic simulations achieve the highest degrees of realism in generating acoustic fields as they compute sound propagation via simulations of high-dimensional pressure fields (Raghuvanshi and Snyder, 2014) or solving the wave equation with Finite-Difference Time-Domain schemes (Hamilton and Bilbao, 2017). Their inherently complex nature requires solving the wave equation to produce acoustic simulations for a given scene, and despite recent GPU-based solvers optimising complexity by orders of magnitude (Mehra et al., 2012). Their computational requirements are often impractical for real-time applications due to the nature of the wave equation, resulting in numerical complexity increases with frequency. On the other end of the spectrum, Geometrical Acoustics (GA) provide methods for fast approximations of acoustic space; they have gained popularity among extended reality platforms due to their highly parallelisable implementations (Savioja and Svensson, 2015a).

Schissler and Manocha (2016) introduced an acoustic rendering system based on ray-tracing, adapting to large complex scenes. Among their contributions is overcoming the problem of handling many sound sources in large-scale environments by clustering them based on the distance from the listener. Based on an octree representation of space, with respect to the listener position, their clustering aggregates increasing numbers of sources as their distance from the listener increases. Their approach highlights the need for dissecting the acoustic space for efficient selective rendering, resulting in rendering of fine perceptual details within the listener's close proximity and coarse approximations otherwise.

Schissler, Mückl and Calamia (2021) recently presented a novel method for computing acoustic diffraction in real-time, which can adapt to GA frameworks. They target complex scenes typical of virtual and augmented reality applications. Their approach can overcome

the shortcoming of GA techniques in approximating soundscapes, thanks to the ability to incorporate simulated propagation effects into their proposed sound rendering pipeline. The main contribution of their work is a mesh processing system that optimises diffraction simulations for environments expressed through dense geometry. Results gathered from their evaluation show that the technique is comparable to FDTD methods, obtaining a high degree of realism from generated propagation data. Due to the novelty and recency, there is a lack of psychoacoustic characterisation performed on their method, making it hard to evaluate whether the technique is feasible for wearable computing.

3.2.2 Differentiable Methods for Sound Rendering

Manocha et al. (2020) present a model for simulating sound fields using neural networks without pre-computing the wave field of an acoustic environment, predicting unseen objects with arbitrary shapes in a VE for sound propagation at interactive rates. They train a geometrical neural network on annotated meshes to infer acoustic data associated with the represented object. Chen et al. (2022a) introduce a novel task dependant on this mapping, *visual/acoustic matching*, which produces acoustic stimuli responding to a target space depicted in an image, given an input audio excerpt and an image of the environment in which excerpt propagated. The rapid development in DNN for multi-modal applications has opened new avenues in the field of sound propagation modelling, one of which tapped into visual-acoustic mapping, the process of determining relationships between visual and auditory features in audio-visual or immersive media.

One innovative experimental method, Singh et al. (2021)'s work into Image2Reverb, ventured into using DNNS to define mappings between images and reverberation, expressed as an IR. By observing photographs of real or virtual environments, our visual system is generally able to infer acoustic characteristics of the space; from a photograph of a cathedral, for instance, we can imagine its reverberant aural footprint. The authors leverage GANs to explore automated mappings between deep visual features extracted from a given input image, representing an environment, and an output spectrogram of an IR. Since many reverb metrics like the T_{60} are linearly correlated to the energy decay in RIRs, their network encodes reverb by representing a spectrogram with variable energy decay. The authors train and evaluate the network by comparing results to ground truths pairs of photographs and measured responses, achieving around $0.87s$ mean error in T_{60} estimations.

Improvements and new approaches to solving the task are being explored at increasing rates, such as Somayazulu, Chen and Grauman (2023)'s network presenting a self-supervised visual-acoustic matching system. With an input audio excerpt and a target image representing an environment, their system re-synthesises the audio excerpt to reflect the acoustic features of the target environment. A key novel aspect of their method is the handling of reverberant audio by leveraging a state-of-the-art network for audio dereverberation (which is a well-established task in the field of acoustic signal processing). The de-reverberated audio is passed to a GAN, which optimises the output audio until it

acoustically matches the extracted visual features.

Liang et al. (2023) present a method that improves on the adoption of Neural Radiance Fields for sound propagation. The authors present a method that allows a neural field to be learned on a real soundscape by providing emitter and receiver position input, a ground truth RIR, and acoustic context representations. The neural field fit on the input environment allows the generation of novel RIRs based on given emitter-receiver position pairs. The neural field learns from multimodal representations of the environment, expressed as visual information by RGB + depth images and acoustic data by emitter-receiver position information.

Tang et al. (2020) present a novel scene-aware sound rendering system aimed at rendering audio considering acoustic characteristics of a given room, providing real-time audio effects applied to novel signal matching the soundscape of the input room. The system uses a neural network to infer reverberation time and estimate resonance interferences caused by the room architecture using a recorded signal and 3D representation of the environment where the recording is generated. Their method uses the inferred acoustic properties as input to an acoustic simulator that generates and optimises acoustic materials by measuring simulation errors against the estimated room features. Once the acoustic simulator optimises materials, it convolves IRs with novel audio signals to emulate a sound source propagating in the input environment. As part of the testing procedures, the authors provide a benchmark for the material optimisation pipeline, outlining the error in estimated materials across rooms of increasing reverberation times; it increases with the size of the input room. Their system provides realistic acoustic stimuli as subjective tests show that the simulation error is not perceptually significant.

Chen, Su and Shlizerman (2023) use Audio-Visual receivers to sample reference features, generating joint audio-visual representations of input scenes to synthesise novel binaural audio. Their system takes visual information and uses a Joint Audio-Visual Representation to extract audio-visual features from space, which feed into an Integrated Rendering Head. The rendering head uses a ground-truth binaural waveform to optimise output binaural audio generated given a listener position. Their rendering pipeline improves state-of-the-art methods, such as few-shots learning-based techniques for sound rendering, by evaluating simulations on standard acoustic scenes and indoor space reconstruction datasets.

Ratnarajah et al. (2022) present a pioneering approach to neural networks for sound rendering as an alternative to physics-based IR computation methods like geometrical acoustics or wave-based methods. The core task of their method is to match auditory stimuli with visuals of a VE for applications around audio-visual navigation, auralisations, speech enhancement, dereverberation and more.

Their method takes a triangulated mesh representing the environment, which is fed into a series of graph NN to process vertex and edge information to create a graph encoding of the input scene, simplifying topology information. A modified Generative Adversarial

Network (GAN) uses the constructed graph representation to generate an IR by computing a decay curve and optimising acoustic characteristics encoded in the graph representation using a generator and discriminator. Their method is tested on indoor scenes, evaluating T_{60} reverberation, direct-to-reverberant ratio and early decay times. The evaluation shows less than 10% error across all metrics, placing the method amongst one of the pioneering NN-based approaches.

Yang et al. (2020) present a method for synthesising RIRs, reproducing perceptually-convincing acoustics of real environments based on a small number of ultrasonic measurements. The method consists of using a loudspeaker and a microphone to record an ultrasonic IRs that can be transformed into an octave-IRs by approximating reflection decay curves. Octave-IRs constructed from Gaussian noise and modelled using estimated decay curves are combined into a final monoaural response. The authors test the approach in two indoor spaces (a lounge and a classroom) demonstrating that the technique can generate perceptually plausible auditory stimuli and showing potential application for MR platforms. However, the apparatus adopted would require the wearable MR device to be equipped with a recording setup to sample ultrasonic responses.

Li, Langlois and Zheng (2018) identify a novel method for acoustic simulations using convolutional neural networks to perform acoustic analysis on videos, veering away from more formal 3D scene definition. This approach synthesises RIRs for environments' representations from audio-visual scenes. Their system extracts high-level acoustic properties such as reverberation time T_{60} and frequency-dependent amplitude level equaliser.

3.2.3 Discussion

The spectrum of sound rendering techniques continues to refine existing methods, increasing their efficiency and applying them to newer platforms and use cases, as well as present novel methods leveraging recent advances in deep learning. Table 3.1 summarises the crucial techniques, reporting advantages and limitations in relation to their employment in real-time auralisation applications.

A common shortcoming of experimental and novel methods for interactive sound rendering lies in the limited testing and benchmarking conducted on the techniques. Such a shortcoming does not necessarily invalidate the value of the contributions or their potential for real-time application in XR domains though provides research directions for future work. Although deep learning approaches are capable of generating auralisations at interactive rates, there are still challenges along the avenues of applying them to wearable computers like HMD due to the computational requirements. Given the current state of research towards deep learning-based techniques, their deployment to MR platforms would require specialised hardware, such as TPUs. Cloud computing alternatives could also facilitate the deployment of deep learning models, requesting inference from HMDs.

GA-derived methods, such as Schissler and Manocha (2016); Savioja and Svensson (2015b);

Schröder (2011)'s, share some inherent limitation of the wider geometrical acoustics family; though there is a deeper understanding of their perceptual impact and are generally easier to scale for platforms with limited computational resources. Considering the thesis objectives, the role of this work is to provide an overview of existing, relevant techniques for immersive acoustics and determine their feasibility for wearable computing platforms. Based on their feasibility, following Chapters will discuss their integration to a prototype rendering pipeline. Table 3.1 reports relevant methods, summarising their architecture and requirements.

Table 3.1: A summary of experimental methods for sound propagation reviewed. These methods vary depending on the task perfomed within the VE, its underline architecture. These are compared based on their inputs and efficiency.

Method	Task	Architecture	Inputs	Requirements	Notes
Schröder (2011)	BRIR est.	GA	S-R, 3D mesh	Physics computations	Fast, optimisable. Limited wave propagation effects.
Mehra et al. (2015)	BRIR est.	wave-based	S-R, 3D mesh	precomputation	Realistic, high-accuracy. Limited applications to dynamic environments.
Schissler and Manocha (2016)	BRIR est.	path tracing	S-R, 3D mesh	Physics computations	Fast, adapts to large-scale scenes. Psychoacoustics-driven optimisations. Limited wave effects.
Tang et al. (2020)	BRIR est.	NN	Room, samples[1]	NN, GA	Fast. Requires both NN and GA computational resources.
Singh et al. (2021)	Reverb est.	GAN	RGB image	DNN forward pass	Fast. No consideration for source-emitter receiver.
Ratnarajah et al. (2022)	IR est.	Graph NN	S-R, 3D Mesh	DNN forward pass	10.000 IR per second[2]. Limited control over acoustic materials.
Liang et al. (2023)	IR query	NAF	S-R, RGBD data	NAF query	Fast and highly generalisable. Not tested on real-time dynamic scenes.
Chen, Su and Shlizerman (2023)	BRIR est.	NN	A/V samples	NN forward pass	Fast[3]. Adapts to real-time dynamic scenes.

[1]Simplified room geometry, audio recordings.
[2]on NVIDIA GeForce RTX2080 Hardware.
[3]around 30.34ms to render BRIRs on NVIDIA GeForce RTX2080Ti.

3.3 Material Recognition for Rendering Tasks

Material recognition for rendering pipelines is a generally narrow research domain with a niche application, and there is a limited body of literature and development toward solutions. This niche technique derives from the thriving and popular superset of literature on the recognition and understanding of material information, benefitting from various advances in deep learning methods for understanding, classification, or detection tasks. Reviewing the superset of techniques is outside the scope of this work; hence, this Section focuses on applications to rendering tasks, discussing the relevance of novel techniques and their shortcomings in relation to the relevant components of this work.

Considering works related to the design of a system for extracting material information from VEs, techniques are categorised into supervised and unsupervised algorithms.

3.3.1 Supervised Material Recognition Techniques

Schissler, Loftin and Manocha (2017) present a two-stage system for sound rendering based on scene understanding performance on scans of physical space, requiring reconstruction of physical space and acoustic measurements as input and, leveraging recent advances in semantic segmentation for audio-visual rendering tasks. The first stage of the system uses multiple camera viewpoints to reconstruct a dense 3D triangle mesh representing the environment and generate input to a CNN to classify acoustic materials from camera renders. A Least Square Solver algorithm uses real measurements to optimise the inferred materials by calculating the distance from estimated IRs to the ground-truth IRs.

Semantic segmentation tasks aim to assign a semantic class label to every pixel in the input image. Examples of applications in scene understanding include PixelNet (Bansal et al., 2016), which performs semantic segmentation and edge detection; EdgeNet (Dourado et al., 2019), which combines depth information with semantic scene completion, using RGB-D input data. For synthetic data generation, UnrealCV provides a pipeline that generates images from VEs providing semantic segmentations (Qiu and Yuille, 2016), allowing for easy generation of training data.

Large-scale datasets, including semantic and 3D information, have been released, e.g. the Matterport3D dataset (Chang et al., 2017), which provides panoramic images generated across real environments. Various domain-specific applications of these methods have been proposed, e.g. in mixed and augmented reality (Chen et al., 2018), where semantic information about surfaces can guide contextual interactions between virtual elements and real-world structures; or surveillance (Mao et al., 2018), where the semantics of objects in the scene determine its subsequent processing.

However, few examples of applying computer vision to realistic audio rendering exist. Kim et al. (2019), as an alternative approach, adopts 360° photographs and depth estimates to generate 3D geometry and semantic information, which is then used for physically-based

audio rendering and can also adapt to VEs. In this context, even approximate semantic information could allow for gains in efficiency and a decrease in the costs of applying physically-based audio rendering to VEs.

Recent developments in deep learning techniques have contributed to a dramatic increase in accuracy in tasks such as image classification. Specifically, convolutional neural networks have been broadly adopted to learning functions mapping between image data and various semantic descriptors, such as local object classes (Long, Shelhamer and Darrell, 2015), or subjective quality (Bosse et al., 2017). For example, Lagunas et al. (2019) present a method to learn similarities between materials based on their appearance and distinguish them in a feature space, informed by human perception. They describe the mappings between subjective perception and physical material parameters. This is a challenging task due to the impact of low-level properties, such as illumination and reflectance on the appearance of materials. The authors address this problem using deep features learned by a neural network trained on a bespoke dataset, annotated with around about one hundred classes of materials, captured under different conditions, including surface shape, illuminance and reflectance, expressed by environment maps and bidirectional reflectance distribution functions. In a subjective study, they encode materials in a perceptually informed feature space, outlining perceptual distance information relating to material pairs.

3.3.2 Unsupervised and Semi-Supervised Alternatives

Schwartz and Nishino (2019) address the problem of material recognition from local visual information of materials to better model human interaction. They aim to reduce manual supervision in the process of encoding material characteristics, explaining visual attributes such as shiny or metallic and material properties that may not be visually or locally discoverable such as softness. They present a novel method for material recognition consisting of perceptually informed distances between materials and attribute spaces based on the distances.

Semi-supervised and unsupervised approaches have also been adopted in tackling such problems. For example, Gaur and Manjunath (2020) propose a novel deep learning architecture to cluster materials from a given dataset, improving state-of-the-art superpixel algorithms by combining segmentation of images into perceptually meaningful pixel clusters with a novel unsupervised clustering method based on superpixel embeddings. A novel loss function uses a variable margin that compensates for the limitations of classic superpixel algorithms in segmenting texture patterns, allowing the convolutional neural network to cluster superpixel labels based on their embeddings requiring no manual supervision or annotations.

Xia and Kulis (2017) introduce a novel deep learning model for unsupervised image segmentation tasks. Their network is composed of an encoder and a decoder connected together to reconstruct an input image, producing a segmentation map, and distinguishing different

materials depicted by the input image.

Kiechle et al. (2018) present a novel method for segmenting textural patterns in input image data reducing the requirements for large-scale datasets representing exemplary features that the model trains to predict. Instead, they propose a framework that learns convolutional features from a small set of images or image patches. Their method shows competitive performance metrics against standard texture segmentation benchmarks, revealing the potential of this experimental method for material tagging.

3.3.3 Discussion

Considering the broad field of computer vision and focusing on techniques that have a direct application to the retrieval of acoustic characteristics from VEs and assigning properties to environment geometry, the area is generally underdeveloped and has potential for improvements. Table 3.2 summarises relevant experimental methods discussed.

A crucial finding within this area derives from Schissler, Loftin and Manocha (2017), introducing some of the first approaches of scene understanding systems for sound rendering, projecting these into use cases for multi-modal MR and identifying limitations that future work should address to be around improved material recognition and inference on outdoor scenes. In general, the problem of recognising materials both in physical and virtual environments remains an open research question within these domains due to the challenging task of associating semantics to the visual appearance of surfaces in complex scenes, which depends on factors associated with the physical properties of surfaces or lighting conditions.

Supervised methods, especially considering Kim et al. (2019)'s work, can classify materials from their visual representation and provide input acoustic rendering pipelines. However, one drawback is the specificity of these methods to the acoustic materials expressed by data used to train the model.

Thanks to their abilities to handle large amounts of unlabelled data or a very small set of representative images, unsupervised methods have a lot of potential in addressing acoustic material tagging in VEs. This approach would require an additional step toward mapping the latent representation of clusters defined by the segmentation network to acoustic characteristics, matching the visual features learned by the feature extractor to their acoustic absorption or reflection characteristics. However, this shortcoming can become an advantage when artistic control is wanted, as existing acoustic materials within a complex scene could be re-mapped and controlled by their visual features.

Overall, CNNs are becoming optimised and fast enough that can be embedded in real-time systems, though generalising on a diverse set of surfaces and use cases is still an open research domain.

Table 3.2: A summary of key techniques for material recognition that can apply to VEs. These are largely based on convolutional neural network layers extracting features from input image data, which virtual cameras can often provide as renders. These methods predict semantic features of materials represented in the input image data, allowing mappings visual features to acoustic characteristics of enviroment geometry.

Method	Task	Architecture	Requirements	Outputs	Type
Schissler, Loftin and Manocha (2017)	Classification	GoogleLeNet	Camera renders	Semantic Materials	Supervised
Dourado et al. (2019)	Scene completion	EdgeNet	RGB-D	Semantic Materials	Supervised
Kim et al. (2019)	Scene segmentation	SegNet	360° stereo photographs	Materials, env. geometry	Supervised
Schwartz and Nishino (2019)	Classification	MAC-CNN	RGB	Material attributes	Supervised
Gaur and Manjunath (2020)	Segmentation	UNet-like	RGB	Semantic materials	Unsupervised
Xia and Kulis (2017)	Segmentation	W-Net	RGB	Segmentation Map	Unsupervised
Kiechle et al. (2018)	Segmentation	Conv filters	RGB	Semantics	Unsupervised

3.4 Human Factors and Perceptual Rendering

Bonneel et al. (2010)'s study investigates the influence of audio-visual stimuli, as well as the interaction of graphics and audio, on material perception. They designed an experiment testing whether graphics and audio have significant effects on the subjective perception of material qualities. The goal is to determine the minimum level of detail expressed by visual stimuli needed to evoke realism in observers, establishing a set of guidelines that can improve the performance of rendering techniques by culling and simplifying geometry maintaining significant perceptual responses.

Dolhasz, Harvey and Williams (2020)'s work around areas of perceptually-informed rendering expands the goal of investigating LoD thresholds in image compositions. The authors expand towards encoding perceptual responses into a latent space than automate the generation of perceptually-valid stimuli. They sample a large dataset of perceptual responses by prescribing a test to participants who were tasked with discriminating images with transformations from a given set. The goal of the authors is to fit a model on subjective responses that can then automate the suprathreshold detection process, which can be used within GAN-like models around automatic generation or transformations of content. This work can have a significant impact on multimodal rendering domains, as encoding perceptual responses can feed into sound rendering pipelines, enabling GAN to leverage discriminators learned on human perception.

Very recently around this area, Manocha et al. (2021) presented a perceptual similarity metric by encoding perceptual distances between audio signals. The authors use CNNs to train a model encoding subjective responses associated with pairs of audio signals, outputting a perceptual distance metric. The model is a great contribution to the field of audio quality evaluation as it can express JNDs between unheard pairs of audio signals, measuring perceptual distances or detecting transformations or perturbation audio data. In the domain of sound rendering, this model could optimise the laborious process of testing and sampling human perception to measure subjective factors of simulated auditory displays.

3.4.1 Perception of Audio Quality

Rummukainen et al. (2018) pioneered the field of audio quality evaluation in immersive technology by porting MUSHRA-like testing to VR platforms, evaluating the impact of audio engines in interactive multi-modal VEs. The MUlti-Stimulus ranking test with Hidden Reference and Anchor (MUSHRA), described in the International Standard BS1534 (Liebetrau et al., 2014), is a standard approach for evaluating the perceived audio quality of a system, often employed to evaluate coding, compression or processing tasks in the audio domain (Series, 2014). Thanks to Jillings et al. (2015)'s web implementations, MUSHRA methods have been providing an essential tool for A/B comparisons of audio effects or algorithms and can be used to evaluate the quality of acoustic phenomena simulated with

rendering techniques, such as reverberation, echo, diffraction or other soundscape characteristics that can be encoded in RIRs.

Rummukainen et al. (2018)'s framework gives MUSHRA additional dimensions by implementing the method as VR scenarios, enabling the evaluation of renderer or spatialiser systems such as HRTF spatialisers. The VR nature of the framework can provide a wide breadth of metrics associated with the interaction between the listener, sound-emitting entities, the environment and tasks or procedures. The user study the authors conducted demonstrates how these metrics can provide further insights into perceptive aspects, for instance, showing how participants dwelled around testing areas during the execution of the procedure. With modern HMDs providing more and better interaction and sensing technology, researchers have access to eye, head, or hand-tracking data, as well as more information regarding scene elements of the VE. Such data is generally unexplored, and investigation should explore how acoustic renderers affect subjective responses to audio stimuli.

3.4.2 Psychoacoustic Characterisation of Sound Propagation Methods

In light of the discussion on novel sound rendering pipelines in earlier sections (see Section 3.2.3), there is a rising need for profiling the psychoacoustic factors of simulated auditory stimuli. González-Toledo et al. (2023) present a toolbox providing an acoustic binaural rendering system, exposing parameters that researchers can measure for subjective evaluations. Their proposed framework provides control over listener pose information, sound source spatial information and binaural rendering parameters. In addition, they enable annotations on audio stimuli, allowing participants to save contextual information related to tasks administered. A point of expansion for this method may be the limited reverberation models available: the toolbox would benefit from interfacing with arbitrary sound propagation systems, allowing researchers to test recent experimental advances in sound rendering.

With the human listener as the central and final link in the chain of a sound rendering system, it is essential to consider how the audio display presented to the listener is affected by aspects related to human perception and psychoacoustic abilities performed by the HAS. Due to the applications of sound rendering in VEs within serious and entertainment domains, researchers often base subjective evaluations of sound rendering techniques on task performance, studying how sound rendering techniques affect interactions, navigation, localisation or other activities influenced by the hearing sense.

Mehra et al. (2015) presented a novel, wave-based sound rendering technique aimed at VR applications, advancing the domain of particle simulations for interactive sound rendering. One of the key contributions of their approach is providing a system offering realistic sound propagation between moving sound sources and listeners and can adapt to large, complex scenes. Their system offers spatial audio reproduction based on head tracking features of

HMDs and position information of the listener in the VE. The inherent limitation of their approach is the required pre-computation stage for evaluating acoustic energy transfers between geometry and objects in order to solve particle equations and generate the wave propagation field that can then be solved at runtime using general-purpose GPUs.

Here, the need for an evaluation of psychoacoustic factors arises for considering whether the perceived quality, subjective and psychoacoustic benefits outweigh the limitation of the pre-computation phase. They gathered 30 participants for their between-subjects experiment, 13 of whom had prior experience with VR technology, and the procedure they were asked to follow was the localisation of a sound-emitting object. The authors delegated a group for the navigation procedure using their renderer and a group using a geometrical acoustics renderer. They show that their wave-based sound renderer allowed a 27% increase in localisation abilities in participants. Some of the limitations of this evaluation lie in the employment of an outdated image source-based renderer with edge diffraction as a comparison, altering the fairness of the study and the singular procedure used, as opposed to a range of different psychoacoustic-based activities that could be tested.

Hacihabiboglu et al. (2017) discuss how perceptual aspects should influence the design of sound rendering pipelines. By reviewing a body of work around auralisation systems and sound propagation for interactive applications, they draw an effective pipeline design. Their design has the auralisation system revolving around a simplified model of the environment geometry and considering material properties, reverberation characteristics, and source directivity patterns as well as spatial information on the listener provided as input to the pipeline. Due to the complexity and computational requirements associated with the propagation algorithms and the rendering aspect of the system, the authors recommend perceptual culling to reduce the load and optimise the process for complex scenes with concurrent sound sources.

Arce, Fuchs and McMullen (2017) conducted the first investigation of psychoacoustic factors in MR, where researchers tested how well holographic audio could be used to attract users' attention towards a given location. Effectively, the study represents a pioneering methodology towards the effectiveness of spatialised audio for psychoacoustic tasks, demonstrating the significance of sound rendering pipelines within the realm of interactions in MR platforms.

3.4.3 Findings and Limitations

Recent work in multimodal rendering has highlighted the rising need for sampling human perception to measure and evaluate subjective factors in multimodal displays. A common denominator in rendering problems is the lack of data on perceptual responses obtained by simulated displays, allowing for dynamic changes in LoD and optimisation of computational resources. This problem becomes central in rendering pipelines that consider dynamic scenes with geometry being manipulated online or in the case of MR platforms that work

Table 3.3: A summary of current fields of research intersecting the thesis work. The Table summarises current state-of-the-art, gaps, and potential avenues across these areas.

Area of Research	State of the Art	Current Limitations	Potential Avenues
Sound rendering techniques	Fast techniques that can produce real-time acoustic simulations	Lack of support for immersive platforms	Testing on XR HMDs
Material recognition	Fast and generalisable techniques that can predict material semantics from any representation of geometry	Lack of benchmarking on material recognition on virtual reconstructions of real space	Implementation of material recognition that embeds into spatial understanding in XR platforms.
Computer vision in sound rendering	Material recognition has been proven effective in sound rendering pipelines	Few example of integrating material recognition with sound rendering and only with a subset of techniques	Propose a generalisable integration of material recognition into sound rendering, working with diverse environments
Psychoacoustic factors of sound rendering	Existing characterisation of psychoacoustics within sound rendering pipelines	Little testing conducted on XR platforms	Studying the application of sound rendering pipelines in immersive application by evaluating psychoacoustic effects

with reconstruction of real space surrounding the viewer. The dynamic nature of the platform often presents varying accuracy and precision in recognising and tracking real space and it is crucial to define the error margin in simulated displays before the user notices incoherence in stimuli. Table 3.3 summarises the current state of research in key areas intersecting the domain of the thesis work, showing gaps and limitations, as well as highlighting potential avenues.

3.5 Conclusions

Overall, the fields of sound rendering and sound propagation in immersive environments have advanced by significant strides into cutting-edge differentiable methods for simulating auditory stimuli. From approximating acoustic characteristics of a soundscape from a single photograph to generating thousands of IRs from a 3D representation of a given environment, there is now a plethora of methods that can adapt to varying needs of realism and accuracy. The rising development of computer vision methods is generating significant momentum towards the creation of novel sound propagation techniques. These advancements hold the potential to revolutionise sound rendering in MR platforms by enabling more accurate and realistic simulations of how sound interacts with dynamic and complex environments. However, before fully integrating these experimental sound propagation methods into MR systems, it is crucial to thoroughly profile the psychoacoustic factors associated with standard sound propagation methods. Understanding the perceptual impacts of these established techniques is essential to ensure that new methods do not inadvertently degrade the user experience or introduce unforeseen auditory artefacts. Despite the growing interest in novel techniques, it is important to acknowledge the value of traditional geometrical acoustics methods. Although they may be considered outdated in the face of newer technologies, GA methods come with well-defined requirements and a solid understanding of their perceptual responses. This foundational knowledge provides a reliable benchmark against which new methods can be evaluated. The following Chapters will start the journey towards developing and testing the core components of a novel acoustic rendering pipeline, in light of the related work and current state of research highlighted by this Chapter.

Chapter 4

Methods for Acoustic Characteristics Retrieval from Complex Virtual Environments

This work was published in:

- Colombo, M., Dolhasz, A. and Harvey, C., 2020, August. A computer vision-inspired automatic acoustic material tagging system for virtual environments. In 2020 IEEE Conference on Games (CoG) (pp. 736–739). IEEE.
- Colombo, M., Dolhasz, A. and Harvey, C., 2021, May. A texture superpixel approach to semantic material classification for acoustic geometry tagging. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–7).

The following Chapter introduces two systems to retrieve acoustic characteristics from space surrounding users in MR, providing fundamental building blocks for subsequent acoustic rendering techniques to simulate sound transmissions between an entity in MR space and the user. The acoustic characteristics extracted from the physical and virtual scene experienced by the user allow context-aware rendering, enabling the acoustic renderers discussed in Chapter 5 to approximate sound waves with geometrical primitives and compute propagation paths from sources to a listener, i.e., the user. Propagation paths, calculated by intersecting rays (or other primitives) with the environment, can respond to the physical properties of surfaces or scene entities, attributing characteristics to portions of the scene geometry. This allows propagation paths to model how acoustic energy propagating from an emitter in MR space is affected by diverse surfaces in space, enabling context-aware auditory interactions.

The following Chapter is structured around two main sections presenting novel workflows for acoustic information retrieval:

1. a **camera-based** system to understand complex scenes and project acoustic information based on visual renders,

2. and a **texture-based** extension that improves the above by abstracting away from camera render and analyses texture data from scene geometry.

The two methods predict acoustic characteristics of space from their visual representations to inform sound rendering and produce believable acoustic stimuli in interactive applications. The methods are tested on various complex scenes, ranging from authored virtual scenes to reconstructions of physical space with LiDAR scanners, emulating input environments that are typically available to MR HMDs. The methods presented demonstrate applications of scene understanding techniques to virtual environments and digital reconstructions of real space to determine acoustic properties of scene geometry for automating realistic sound rendering, and they are evaluated on state-of-the-art acoustic rendering systems, measuring objective and subjective metrics relating to simulated soundfields.

4.1 Introduction

Modern sound rendering, as discussed across Chapters 3 and 2, often requires the positions of sound sources and listeners, as well as the scene geometry with acoustic characteristics expressing the behaviour of each material. The accuracy of acoustic simulations depends on material information assigned to the scene geometry. The scene geometry, tagged with frequency-dependent absorption and scattering information, determines how sound behaves in space and affects the resulting wavefield. Attributing material characteristics to scene geometry segments and entities is commonly performed manually and often at a high cost and resources due to the human-in-the-loop.

In game development, for instance, tagging materials with appropriate acoustic data often requires the work of experts, raising costs and resource needs for large scenes. This Chapter illustrates progress towards creating an automatic process for the generation of material tagging data that can provide input to sound rendering pipelines. Specifically, here are proposed proof-of-concept systems for vision-based material information retrieval, allowing for tagging an object's acoustic properties based on its visual features. These systems can tag meshes in VEs representing boundaries in sound propagation paths with a noticeable perceptual impact, facilitating acoustic renderers on complex scenes. The goal is to remove the human-in-the-loop to allow virtual reconstructions of real space to provide input to acoustic rendering systems automatically.

At the core of these methods lies the problem of scene segmentation. In computer games, often, a set of meshes composes a scene, where each mesh represents an object in the scene. Acoustic materials are often assigned to all triangles composing a given mesh,

allowing audio engineers to group scene geometry when assigning acoustic data. Hence, the resulting acoustic model's accuracy depends on the separation of the geometry, where ideal conditions would have each triangle mapped to its specific acoustic data. A naive approach would have the entire geometry mapping to a single acoustic material. Besides, the representation of materials in real and virtual environments adds further dimensions to the material tagging problem due to complex links between the visual representation of materials of an object and its perceptual effects on the soundscape of the environments in which it exists.

The systems developed throughout this Chapter are trained and tested on exemplary use cases and environments with a limited spectrum of acoustic materials. The camera-based models trained as part of this Chapter are not generalisable for unseen new environments. However, by scaling and applying the methodologies on a diverse and complete set of scene, it is possible to provide a generalisable model, which is a task whose scope is beyond this thesis. The texture-based model, conversely, trains on a generalised set of surface representations, which is generalisable to the standards defined by Bell et al. (2013), making the model applicable to unseen environments. This Chapter addresses acoustic material tagging by providing the following contributions:

- a novel methodology for acoustic material tagging using a camera-based system and computer vision techniques to infer from the scene;
- an alternative system that abstracts from the use of cameras and operates on reconstructed virtual geometry;
- objective and subjective methodologies to evaluate the efficacy of the systems comparing against objective and subjective metrics;

4.2 Camera-based Acoustic Material Tagging

4.2.1 System Overview

The camera-based acoustic material tagging pipeline is the first approach, within the development of this work, to attribute physical properties to portions of scene geometry representing a virtual environment. At a high-level overview, the pipeline adopts a Convolutional Neural Network (CNN) to understand the scene features of the environment, expressed as pixel-wise semantic information predicted from camera renders obtained by a perspective camera within the virtual environment. The network generates a prediction map where pixel-wise semantic information maps to acoustic absorption information, e.g., a pixel belonging to a wooden table in the virtual scene may be attributed with "wood" semantics, which maps to absorption data related to the semantic material. By using camera transformation matrices, semantic information is projected onto the scene from camera space. The system has two phases: *training* and *inference*.

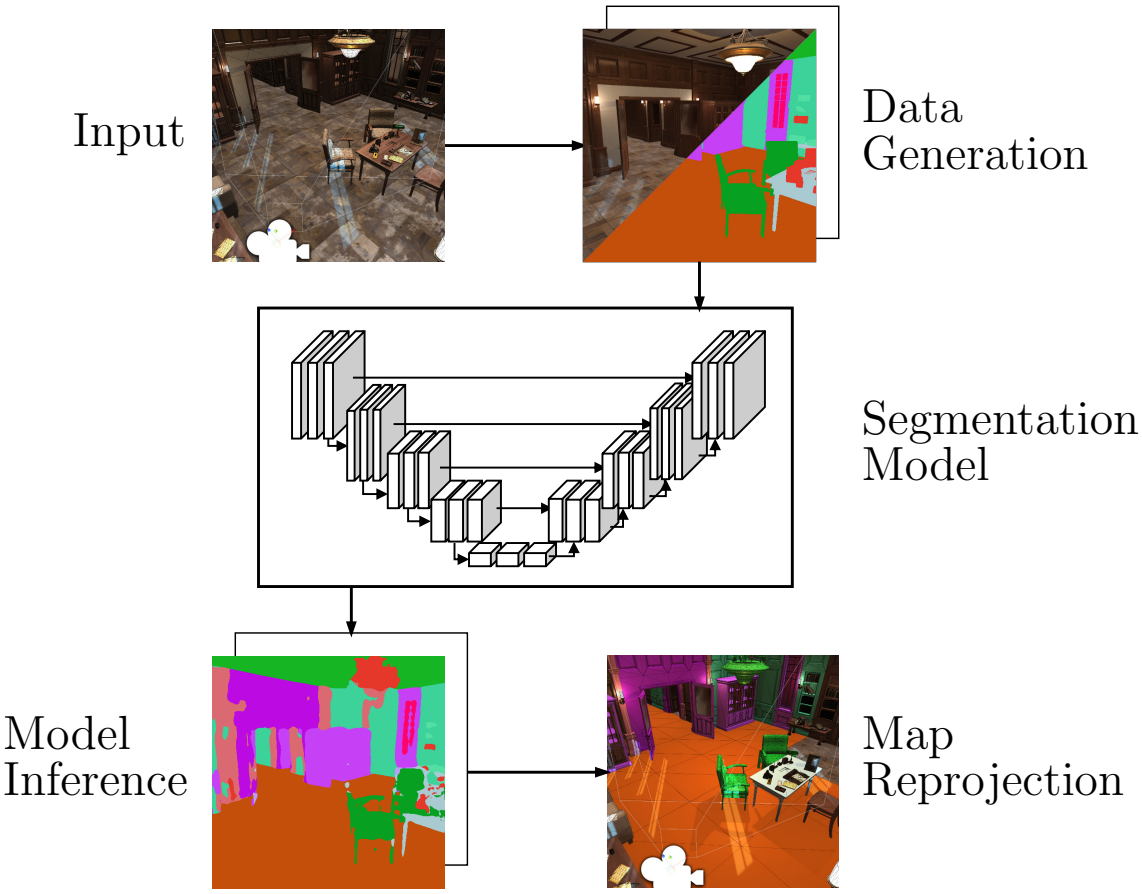


Figure 4.1: Overview of the proposed system: given a set of views captured in a VE, a convolutional neural network trained on samples from the test scenes performs semantic segmentation. The predicted semantic maps are then reprojected onto objects in the virtual scene, associating predicted semantic classes with acoustic profiles that are attributed to the scene geometry. Tagged scene geometry provides input to acoustic renderers or physically-based audio engines for sound propagation or synthesis tasks.

Training Phase The training phase used an environment with scene geometry tagged with ground truth acoustic materials, reproducing the workflow of audio engineers authoring virtual complex scenes by assigning acoustic properties to scene elements. The system is trained by adopting a model that is pre-trained on ImageNet (Deng et al., 2009) on these scenes by generating a set of camera renders with associated segmentation mapping to the correct pixel-wise category. The model, once pretrained, is fine-tuned on selected scenes, and hence, not generalisable.

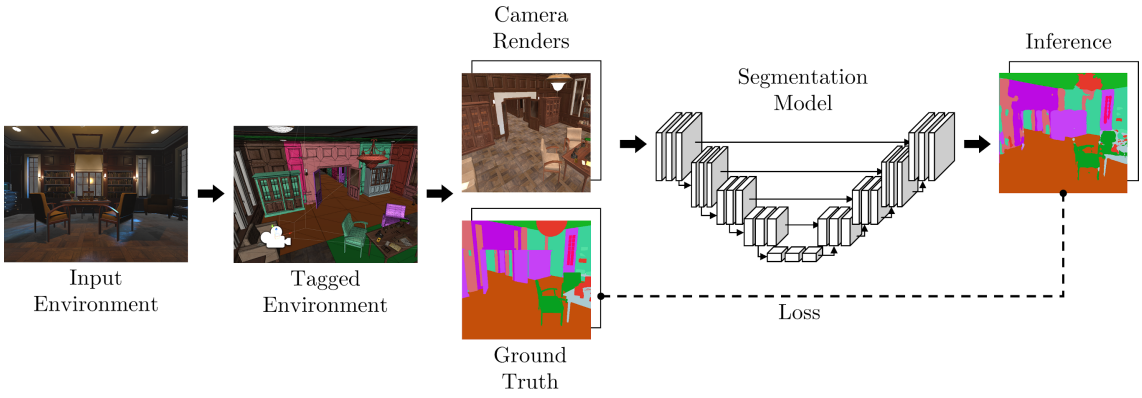


Figure 4.2: Training phase of the system pipeline: manual acoustic material tagging is performed on an input environment, generating pairs of camera renders and segmentation maps via ray casting, which are then used to train and evaluate the convolutional neural network.

Inference Phase Once the network is fit on the ground truth set, it is deployed to a set of test scenes with no material tags, obtaining camera renders across uniformly spaced camera probes scattered around the walkable space of each scene. The scene is then tagged by predicting segmentation maps and reprojecting acoustic material to virtual geometry.

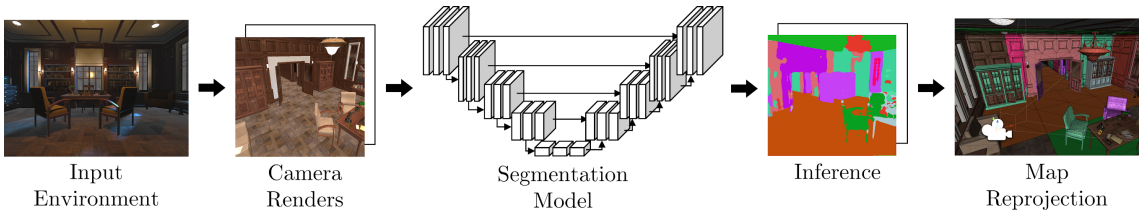


Figure 4.3: Inference phase of the system pipeline: camera renders are generated from an input environment, providing input to the convolutional neural network. With camera transformations, segmentation maps generated by the network are used to attribute acoustic properties via semantic class mapping to scene geometry.

4.2.2 Method

Based on advances in scene understanding and the current state of sound renderers, an architecture that enables the process of semantic mesh labelling in complex scenes and associating every category with a frequency-dependent acoustic absorption function is prototyped and tested. Methods based on perceptual metrics should consider only meshes that are relevant to the acoustic environment. Scene understanding methods and inference

should be optimised depending on the scene geometry.

Input

This method demonstrates the usage of the material tagging system in two scenes: an open, urban environment, *City*, and an indoor wooden room, *Office*. City has $6.3M$ triangles and $8.6M$ vertices. Office has $3.3M$ triangles and $3.8M$ vertices. A set of classes using tables of measured acoustic absorption of construction materials is defined, grouping materials in categories specifying a vector α of absorption coefficient values across an approximated Equivalent Rectangular Bandwidth (ERB) frequency scale ranging from 1.25×10^2 Hz to 4 kHz. Two levels are defined for every major material category in the material database, representing the low and high bounds of mass density ρ in that category. Mass density is a physical property allowing for the acoustic properties of two objects made of the same material to be perceptually distinguishable (Giordano and McAdams, 2006). There are 23 material classes constituted by the two density levels for each of the 11 material categories and an additional class representing "air"; see Table 4.2.

Data Generation

The data generation pipeline of the material tagging system, implemented in a standard game engine, Unity, uses probe points uniformly scattered across a complex scene to generate labelled data for training and inference purposes. Segmentation masks associated with each view are generated by ray-casting through each point of C_n, the *near* camera clipping plane, to ∞; see Figure 4.5a.

The areas where rays intersect with C_f, the *far* camera clipping plane, are labelled as air; objects that are hit by a ray determine the pixel value of the mask, which points to the corresponding material. The dataset consists of 3500 labelled images with 512×512 pixel resolution, split into 3000 training images and 500 validation images.

In the City and Office complex scenes, rendered views are generated in different regions of the environments. The different regions delimit spaces for the collection of training and validation data. For each delimited region, a set of 100 points for each scene are scattered to cover the walkable space. The camera position is interpolated across points in these sets and rotated between 0 and 2π along the azimuth and between 0 and π along the elevation. The camera rotates 6 steps along the azimuth, and 6 steps along the elevation distance.

Table 4.1: CNNs used to produce segmentation maps with the camera-based material tagging system, detailing architecture type and performance metrics.

Type	**Backbone**	**Capacity**	$\mathbf{F_1}$	**IOU**	**Epochs**	**LR**	**Loss**
Unet	ResNet-34	$24.5M$	0.51	0.58	70	1×10^{-4}	$2 * 10^{-3}$
Unet	ResNet-50	$32.6M$	0.54	0.47	70	1×10^{-4}	$7 * 10^{-4}$

Semantic Segmentation Model

A convolutional neural network is used to discriminate materials of objects represented in the camera-rendered views. This task is performed with pixel-level semantic segmentation using a ResNet-34-based UNet (Ronneberger, Fischer and Brox, 2015). The ResNet backbone offers a topology that is easy to train and has excellent generalisation performance. It also provides a compromise between accuracy and the number of parameters as demonstrated by He et al. (2016). The model, pre-trained on the ImageNet dataset (Deng et al., 2009), is fine-tuned for 70 epochs minimising focal loss (Lin et al., 2017). Table 4.1 shows the information on the networks trained, including the total number of parameters, F_1-score, Intersection Over Union (IOU), and the number of epochs to obtain fit models, see Figure 4.4.

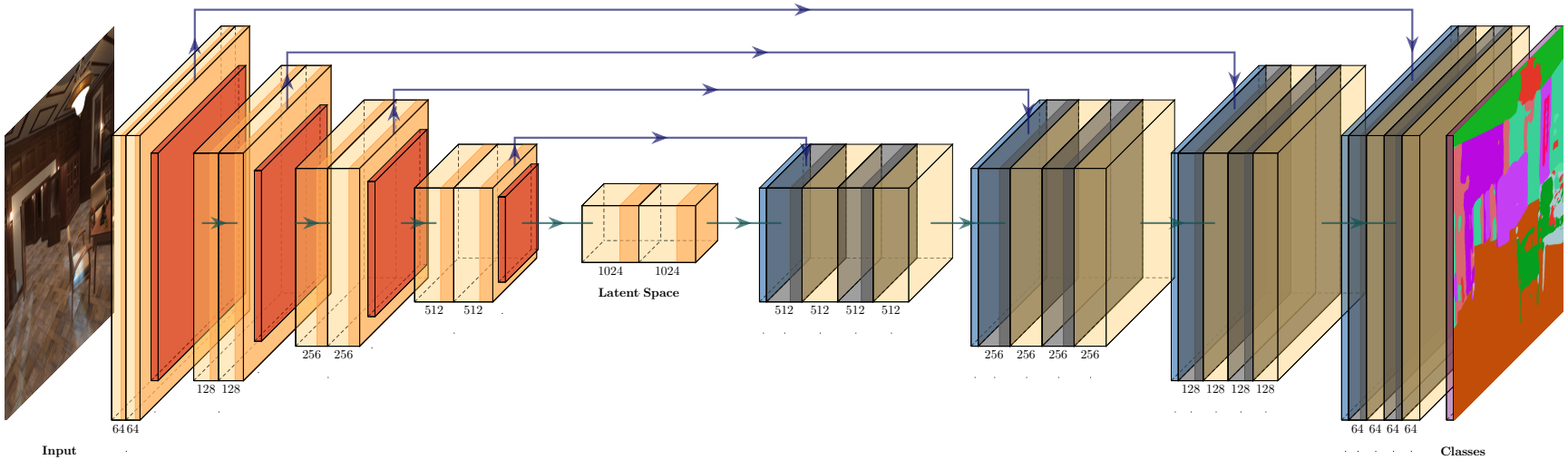


Figure 4.4: Convolutional neural network used for the camera-based acoustic material tagging system: given an input camera render of the virtual environment, the model performs semantic segmentation, predicting a map of semantic classes associated with each pixel in the input image. The model is trained on a set of renders from various virtual environments. The output semantic maps are then reprojected onto objects in the environment. By associating the semantic classes with acoustic characteristics and scene geometry, this information can be used in rendering operations.

Model Inference

The trained model outputs an $m \times n \times k$ matrix M, where m and n are the input image resolution and k is the number of classes. For each pixel, the k channels encode a probability distribution across the classes. Per-pixel classes are determined with the member having the highest presence probability, reducing M to an $m \times n$ matrix where entries encode the semantic class; see Table 4.2. In addition, counts of unique entries in M determine the number of pixels describing the associated material. Scaling based on the distance between a target object and C_n allows material exclusion below a threshold.

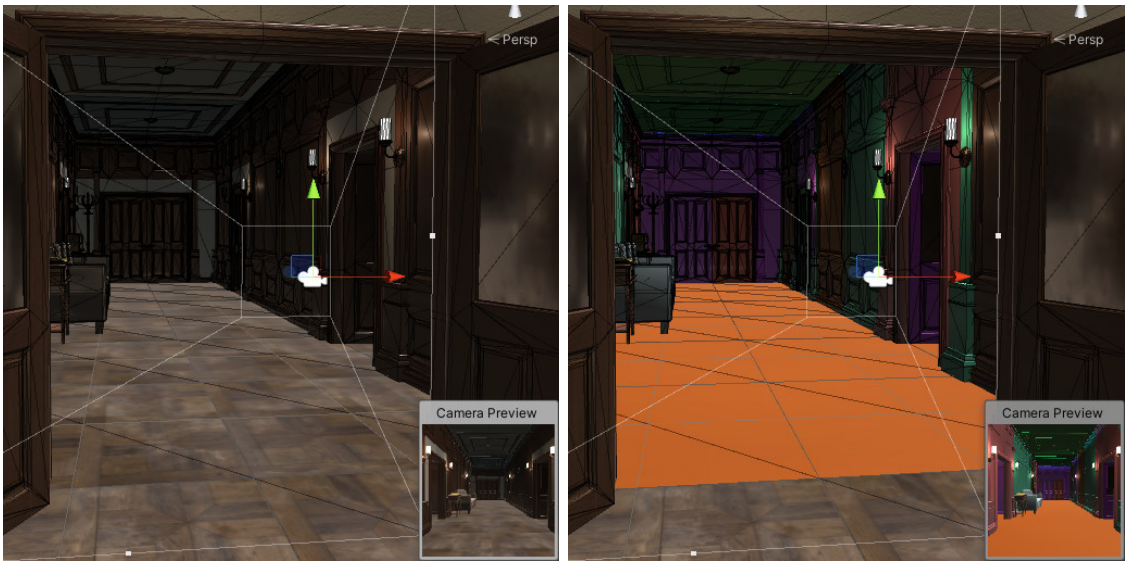
Map Reprojection

Using the segmented images, meshes are labelled by raycasting through C_n divided in strides, targeting entities within the camera frustum. The process depends on the distance between a target mesh in the game engine and the camera, the stride size is determined by the lowest structural dimension of each mesh, scaled according to its distance to C_n. This allows consideration of filtering objects by wavelength, $\lambda = 0.7m$, from the reprojection process. As a result, the material tagging system can dismiss objects that are too small to have a significant impact on the human perception of the soundscape. Through this Level of Detail (LoD) graduation, scene geometry is simplified, excluding structures having smaller perceptual impacts on the resulting acoustic model and optimising input for sound rendering systems applied to complex scenes. The map reprojection process attributes acoustic materials to geometry based on the predicted segmentation map, obtained from the trained model. The result is a scene with acoustic materials automatically assigned to entities.

Among the factors affecting the performance and accuracy of acoustical simulation methods is the polygon count of the acoustic VE, dependent on the complexity of a scene and the presence of detail and small objects. In acoustic environments, smaller structures on surfaces tend to induce scattering of incidental high-frequency waves reflected, and they are neglected by lower frequencies whose wavelength is greater than the structure dimensions. As a consequence, the amplitude of lower frequencies is more likely to be affected by first-order room modes, given by walls and large boundaries, affecting the frequency response of the sound field, resulting in a more noticeable perceptual effect. As opposed to frequencies higher than the Schroeder Frequency, which tend to scatter chaotically (Kuttruff, 2016; Blauert, 1997).

Small structures on surfaces can be excluded from the geometry tagging process due to their insufficient perceptual contributions to the soundfield. Filtering of small geometry can be seen in Table 4.3 (Office, Tagged) where smaller objects than a λ of $0.7m$ do not receive a material tag.

Upon tagging meshes based on pixel-level data extracted from segmentation maps, the confidence value provided by the network when attributing classes to pixels is stored at a



(a) Input Camera Render (b) Re-projected Acoustic Materials

Figure 4.5: Map Reprojection performed on an input camera render: based on a predicted segmentation map, acoustic materials determined by pixel-wise semantic information are re-projected onto scene objects captured by the camera.

mesh level. This is used whenever a subsequent map reprojection overlaps with a previous one, and the same mesh is being reprojected multiple times: in this case, the class with the highest confidence value is used.

Finally, in cases where geometry cannot be rendered by the camera due to obstructions or invalid camera placements, the geometry is tagged with the "Other" low category, see Table 4.2.

4.2.3 Acoustic Materials

Table 4.2 shows acoustic material definitions for the camera-based system. In light of the introduction to materials in Section 2.4.3, acoustic materials in the experiments illustrated in this thesis are defined as a set of absorption coefficients mapped to the scene geometry.

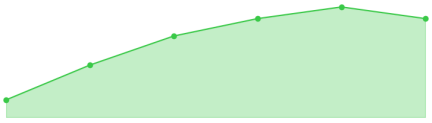
Each entry in the table has two variants for a given semantic material: a low and high absorption variant. This categorisation allows the reduction of dimensionality of standard acoustic material databases[1] whilst keeping the differentiation between instances of the same semantic material class.

Evaluation

An acoustic renderer is used to test the validity of this method by producing auralisations of Office. City is excluded because of its computationally expensive scene complexity. The evaluation uses a state-of-the-art acoustic renderer proposed by Raghuvanshi and Snyder (2014), Project Acoustics. It integrates into Unity, to generate a model of the acoustic

[1] https://odeon.dk/downloads/materials/

Table 4.2: Common acoustic absorption coefficients with ranges (low-high) of α absorption characteristics across the frequency bands for those material types. The last row shows the centre of each frequency band (in Hz), corresponding to each point of the absorption curve. It should be noted that these are regressions and averages of generally adopted materials and existing measurement tables; realistic or surveying acoustic simulations should adopt absorption measurements of real materials.

Material	**Low** α	**High** α
Glass and glazing		
Masonry walls		
Stud-work and lightweight walls		
Wood and wood panelling		
Floors		
Panels and doors		
Other		
Wall treatments and construction		
Ceilings		
Mineral wool and foams		
Audience and seating		
Frequency band (Hz)	125 250 500 1000 2000 4000	125 250 500 1000 2000 4000

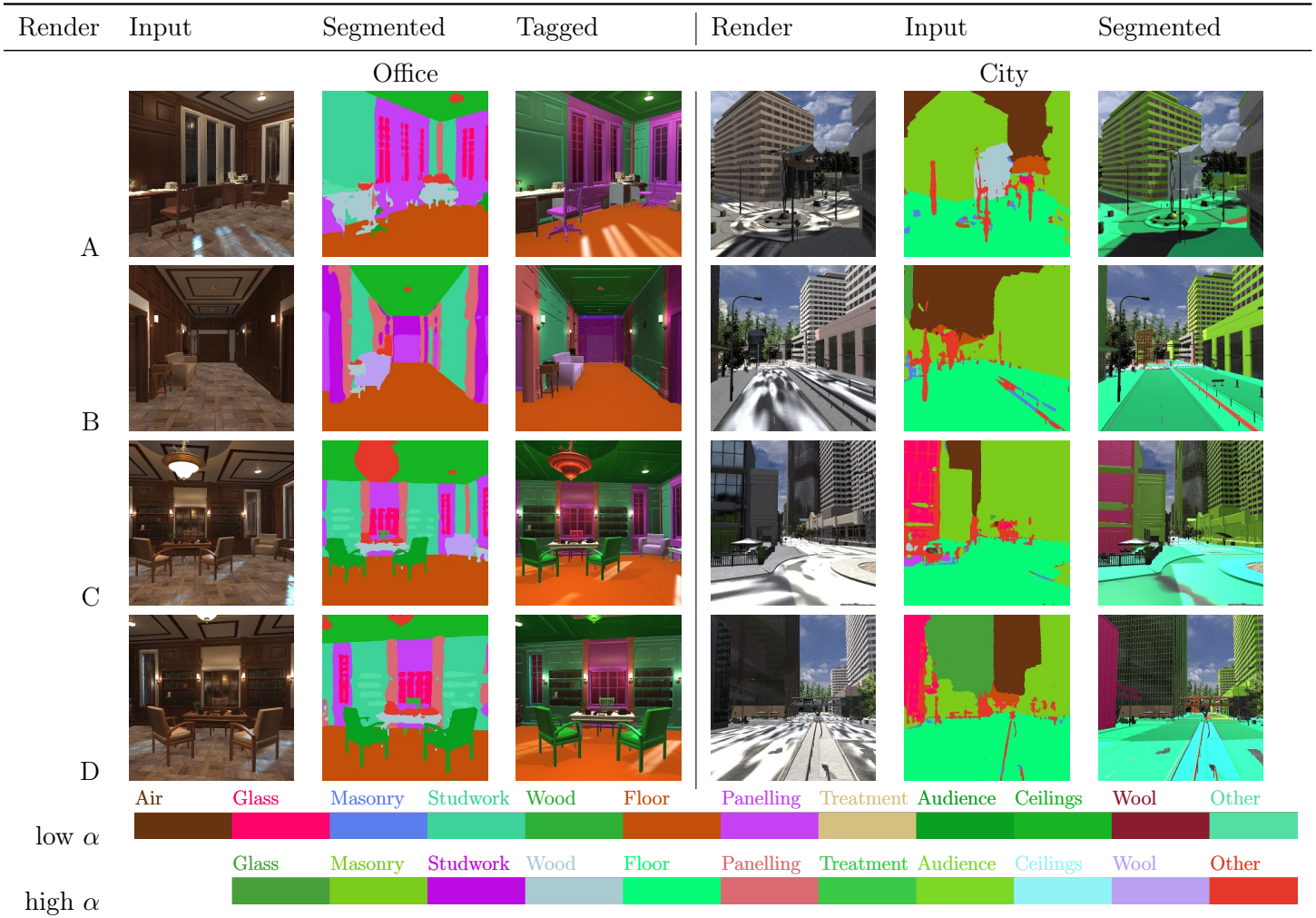
environment using the parametric wavefield coding technique to compute effects between sources and receivers through wave-based acoustic rendering. The renderer determines per-mesh absorption information based on the texture meta-data as per default behaviour. A sound source and listener are placed at human height in the scene; the listener captures a $30s$ chirp signal sweeping logarithmically from 2.0×10^1 Hz to 2.0×10^1 kHz emitted by the sound source to measure an IR. Maintaining the same settings and positions of source and listener, the procedure of supplying meshes and absorption information inferred by the system is repeated, generating the Tagged data. The evaluation compares the two IRs generated by the former (Default) and the latter acoustic model (Tagged) through comparisons of frequency responses.

4.2.4 Results

Figure 4.11 and Table 4.3 show data from computed RIRs across the scenes tested. The model inference takes an average of $400ms$ and the re-projection process takes an average of $96ms$. These figures are quoted per camera probe that is used to generate acoustic labels for surfaces in the scene. Images to be inferred are of a fixed size from the scene frame buffer, 512×512 pixels. The time taken for inference is largely invariant to typical scene complexities such as shape, polygon count, materials etc. The Office scene requires 12 probes to completely tag the environment, requiring ∼6s to complete the tagging process. The City scene shown has extra complexity and requires the use of solutions to the Art Gallery problem to deduce the minimum number of probes to cover the space and tag all objects.

As shown in Table 4.2, acoustic properties can be associated with geometry in the scene, and can be tagged from camera probes. These materials are used in an acoustic rendering process, either directly in audio engines or external offline acoustic renderers. This can result in more realistic aural spatialisation, using IRs to encode early and late reflections.

Table 4.3: Results from tests conducted on the City and Office complex scenes by applied the camera-based acoustic material tagging system. Each row of the Table is split into the two scenes, the left hand side being Office, and the right being City. Each row shows a different camera placement. The bottom two rows show legends for the colour-coded acoustic materials in the Segmented and Tagged columns.



4.2.5 Discussion

Currently, this camera-based approach works by providing inference for camera views within the scene. These camera views are manually placed and would need to be placed in many positions in order to tag materials accurately for the entire scene. The process still requires a human-in-the-loop and needs to be addressed to ensure the goal of having the system as an end-to-end autonomous vision-based material tagging system.

To extrapolate materials tagged to the entire scene, solutions to the Art Gallery problem would optimise the number of predictions required, automating the process of generating data for training the model Devadoss and O'Rourke (2011); Bajuelos et al. (2008). Considering the polygons encapsulating the walkable space W of a scene, minimum vertex guard algorithms suggest that $\lfloor n/3 \rfloor$, where n indicates the total vertices of W, is the least number of positions from where the entire scene can be seen. Based on the depth of the camera, additional intermediate positions $\mathbf{p}$ might be needed to accurately represent objects, this also depends on the number of pixels per object allowing the the neural network to infer materials from the set of camera views that facilitate the whole scene to be visible. For each camera probe position, rotation steps are needed to ensure that all points of W are inside the camera frustum. For an omnidirectional camera probe, these rotation steps $\mathbf{r}_\theta$ for azimuthal steps and $\mathbf{r}_\phi$ for elevation steps should cover the space in 2π azimuth and π elevation. The resulting complexity of the material tagging process for the scene would then be determined by $O(\lfloor n/3 \rfloor + \mathbf{p} + \mathbf{r}_\theta + \mathbf{r}_\phi)$.

Acoustic modelling and audio rendering methods can benefit from research and development of computer vision methods. The current status of this method does not fully eliminate the human-in-the-loop; however, it can generalise and operate on large sets of complex scenes. As a result, artistic and creative workflows for level design can benefit from an automated material tagging system that is agnostic of scene complexity and allows for easy integration of wave-based acoustic renderers. The next steps should include the development of a generalised system to perform material tagging in complex scenes, considering optimisation methods to allow the inference of entire scenes automatically with the minimal set of camera probes to consistently tag every acoustically congruent object that is contributory to the VE, considering its size.

One crucial advantage of a camera-based acoustic material recognition system is the potential to tailor the recognition of materials based on their appearance to a defined ecosystem of material appearances expressed by a set of virtual environments. Although this aspect contrasts the goal of CNN of providing generalisable models, it addresses the problem of the large variance found in visual features of materials by constraining the network within a range of material appearances that interest VE designers. Resembling workflows common in Neural Acoustic Fields (NAFs) (Mildenhall et al., 2020), game designers or VE artists would need to provide exemplary scenes tagged with a customisable set of acoustic and semantic materials to the camera-based system, enabling acoustic material tagging of

unseen or novel scenes sharing the same nature. With the rising availability of computational resources allocated to rendering VEs, performing a forward pass with a ResNet50 CNN (He et al., 2016) is becoming feasible at runtime, unlocking the potential for online sound propagation on dynamic scenes.

4.3 Texture-based Acoustic Material Tagging

In this Section, a novel architecture for tagging acoustic material in virtual environments, abstracting away from camera-based systems, is developed and tested. By working on virtual reconstructions of complex scenes, this approach is agnostic of the technology and architecture of the target platform. Despite the significant progress made in sound propagation over the last decades, there are still many limitations in simulations for indoor and outdoor environments due to the complexity of the factors that describe a wavefield (Liu and Manocha, 2020). The texture-based system considers the discussion around the camera approach and expands towards:

- a more efficient application of acoustic rendering to virtual environments;
- a novel architecture for recognising materials from textured meshes in complex scenes, reducing the need for manual tagging of acoustic materials and eliminating the needs for camera-based workflows;
- an objective evaluation of the architecture conducted on a virtual reconstruction of a real conference room.

4.3.1 Method

The texture-based system is presented as a method for processing scene geometry, generating input for sound rendering pipelines by predicting materials of objects composing complex scenes. A system overview shown in Figures 4.6 and 4.8 illustrates how visual representations of the environment map to acoustic data used by sound renderers to model sound propagation in a scene. Similarly to the camera-based approach, there is a *training* and *inference* phase.

In summary, the unwrapped texture of each mesh in the scene geometry provides representations of their materials, which provide input for a convolutional neural network. Based on features extracted from textures, the network recognises different material labels in textures and maps them to acoustic data from a database, expressed as acoustic materials.

Material Recognition

According to Schwartz and Nishino (2019), small image patches contain enough information to distinguish materials and hence, this method decomposes input image textures into small image patches.

Training

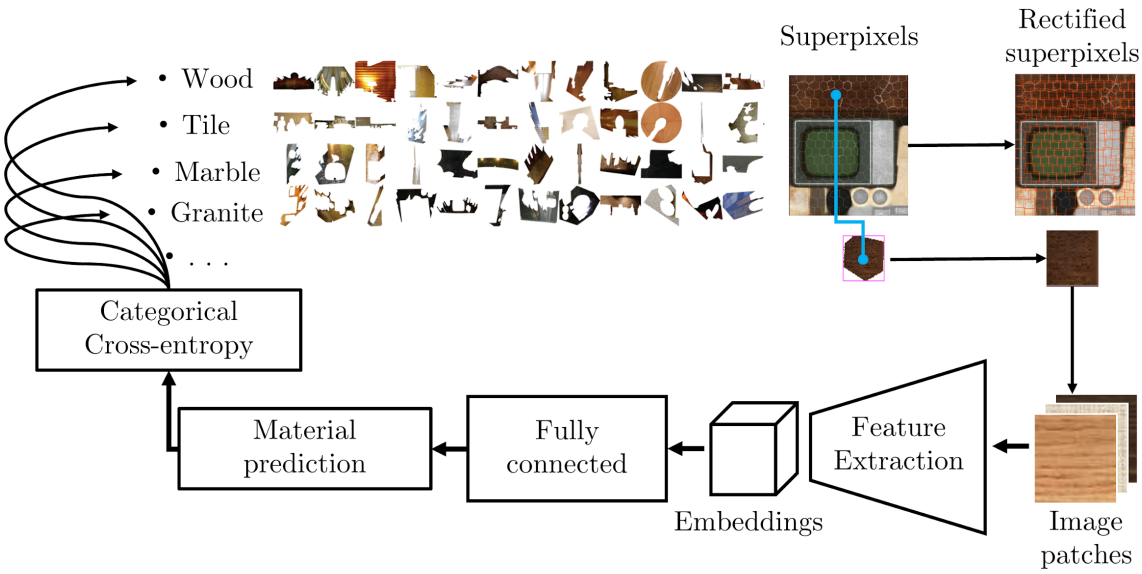


Figure 4.6: Overview of the texture-based system set up for the *training* phase: a network extracts features from superpixel data, obtained from image textures from scene objects in virtual environments. Extracted features are used to classify semantic materials. The system trains on a large dataset of annotated image patches gathered from in-the-wild photographs of real surfaces.

The training pipeline determines the visual material space by applying transfer learning to the OpenSurfaces dataset (Bell et al., 2013), which comprises 36 classes of surface photographs. To do so, a SLIC superpixel segmentation algorithm (Achanta et al., 2012) segments input surface photographs into a set of superpixel labels, determining regions correlated with boundaries of objects. These resulting superpixel labels are then used to generate rectified image patches encapsulating their contours through edge detection, adopting Ding and Goshtasby (2001)'s method. Rectified image patches provide input to a ResNet50 (He et al., 2016), used as a feature extractor for a classification network using a standard fully connected layer to predict class labels based on embeddings of 32×32 pixel resolution input patches. The network trains the network on $1.367\,781\,9 \times 10^7$ input patches, composing a train set of about 9.1M images and an evaluation set of about 4.5M, adopting the standard Adam optimiser (Kingma and Ba, 2014) to update the weights initialised from the ImageNet dataset (Deng et al., 2009). The model usually converges in 45 epochs with a training and validation accuracy of about 94% and 83% respectively. Figure 4.7 shows the CNN used for both the inference and training phases to generate semantic labels based on input image patches.

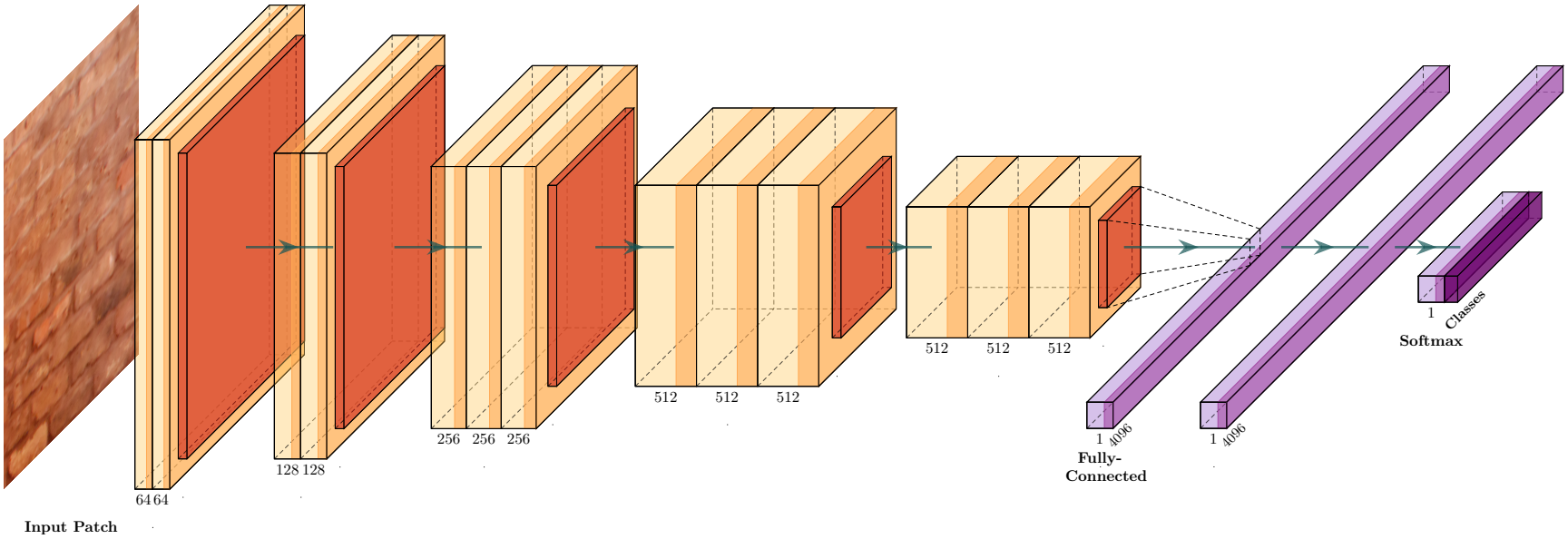


Figure 4.7: Unlike the image-to-image model used for the camera-based material tagging system, the texture-based approach outputs semantic classes based on an input image patch. The system extract image patches from textures associated with objects in the virtual environment. Output classes are then used to create acoustic materials via semantic mappings.

Inference

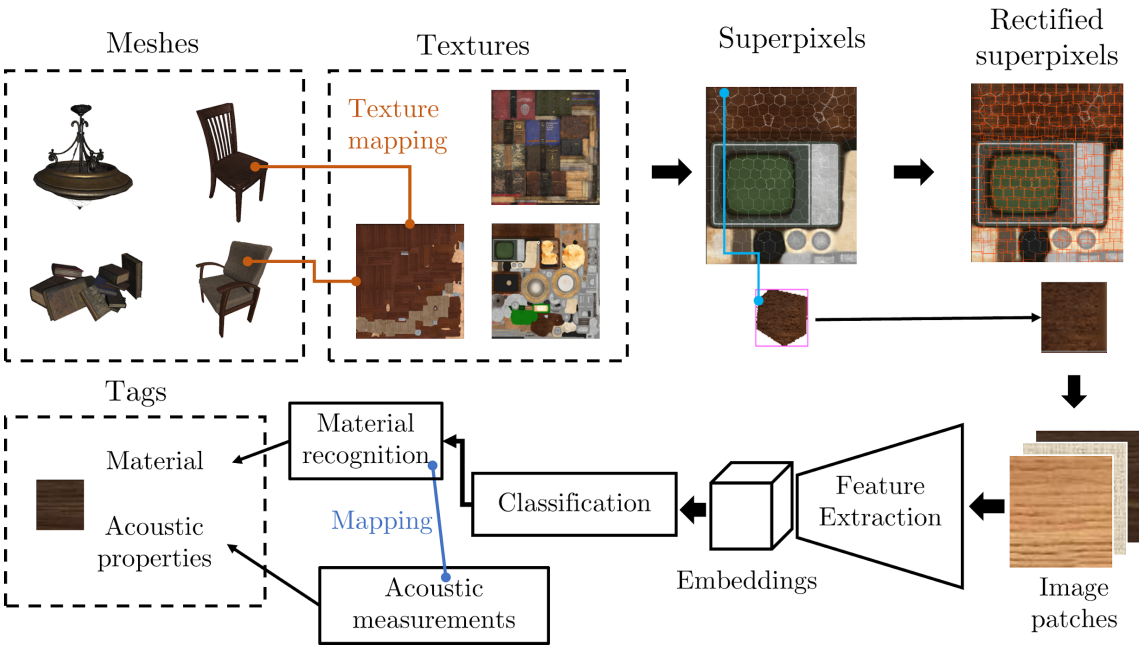
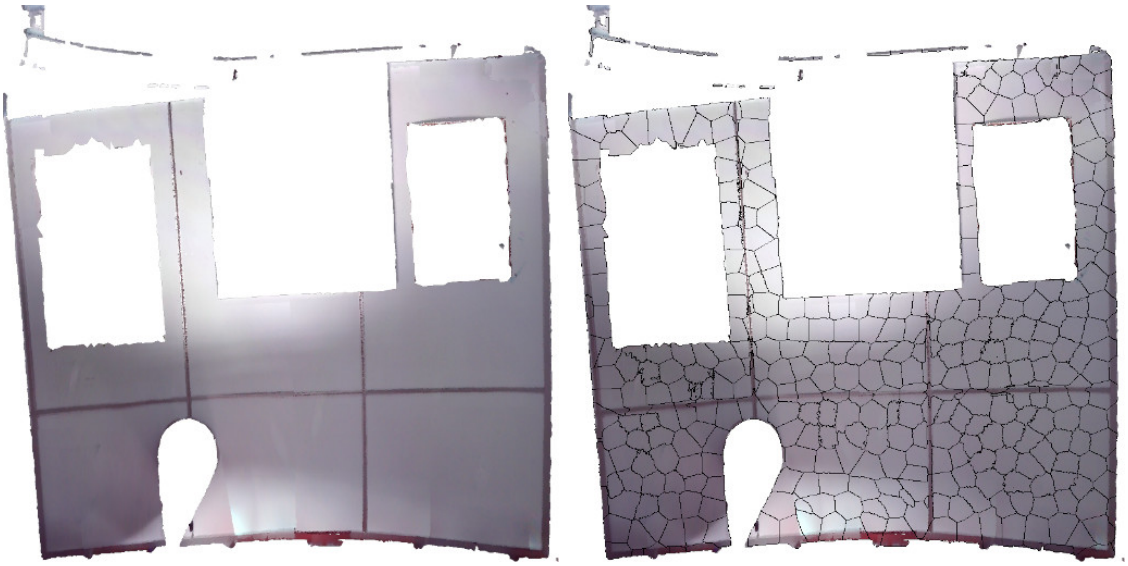


Figure 4.8: Overview of the texture-based system set up for the *inference* phase: textures mapped to meshes representing scene objects are decomposed into image patches using SLIC superpixel segmentation. A trained classifier extracts features from image patches and generates acoustic materials, which are then mapped against geometry representing the given scene object.

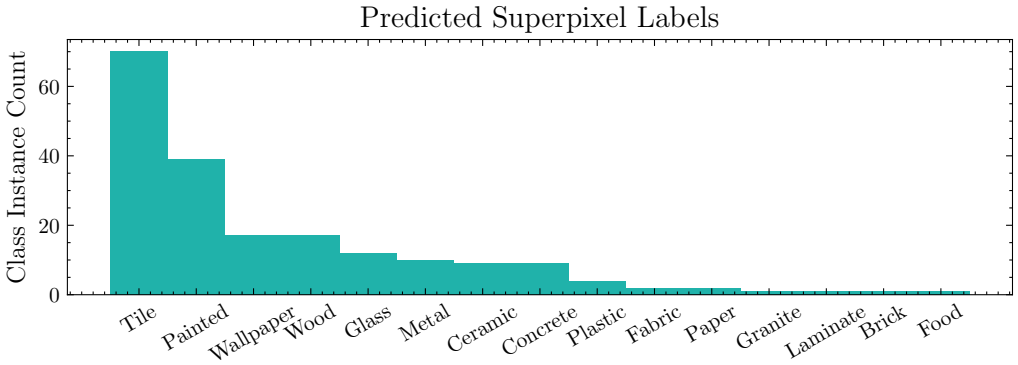
During the *inference* phase, the set of textured meshes in a given scene is used to unwrap textures as images, providing input to the trained network to recognise acoustic materials; see Figure 4.8. The trained ResNet50 extracts features from input image textures in complex scenes, whose embeddings enable the classifier to predict class labels associated with each input superpixels. The most frequent prediction maps to an acoustic measurement database, defining the output acoustic material. On average, the classifier takes $11.2s$ to determine the acoustic material for a given texture, see Figure 4.9, divided in $3.8s$ for generating rectified patches and $7.3s$ to extract features and compute the mapping. The transparency information contained in the mesh texture is preserved, allowing the network to dismiss image patches that do not contain visual features as demonstrated in Figure 4.9b.

Acoustic Material Mapping

Material labels inferred from textures are associated with acoustic measurements of absorption coefficients. For every label, a one-to-many mapping group measurements of the given material. Following the methodology developed by Kim et al. (2020), the system uses median frequency-dependent values to determine acoustic absorption, defining acoustic materials. A single acoustic material maps to each given mesh, associating a vector of acoustic absorption coefficients to its triangles, determining the overall acoustic mapping accuracy to depend upon the mesh separation of the scene geometry. In the example texture shown in Figure 4.9, "Tile" defines the acoustic material, as per predictions shown in Figure 4.9c.



(a) Input Texture (b) SLIC Superpixels



(c) Predicted Class labels from input image patches computed from the texture shown on the top left (Figure 4.9a)

Figure 4.9: SLIC Superpixel computation on real texture from a virtual reconstruction of a conference room. Given an input object from a virtual environment, its texture is extracted, top left, and fed to a SLIC superpixel algorithm, top right. Each superpixel region is then rectified and fed to a material classifier that predicts a list of materials associated with each superpixel, see bottom Figure. The class that recurs the most, "Tile" in this example, is used to attribute a material to the texture provided.

4.3.2 Evaluation

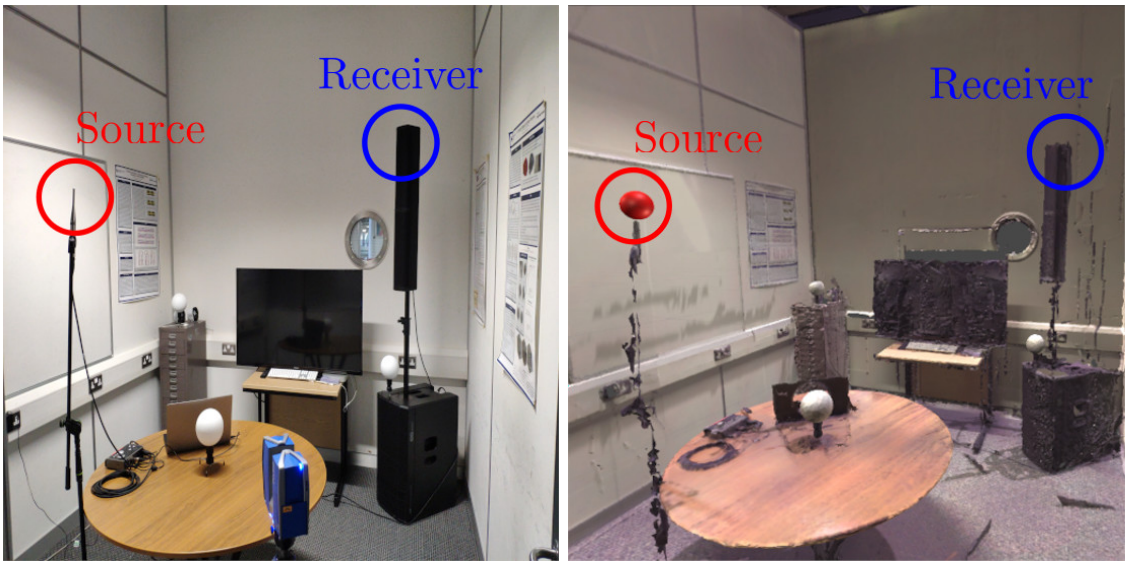
The texture-based acoustic material tagging system is evaluated by deploying it to a real-world scene, using a state-of-the-art GA acoustic renderer. The system is compared using acoustic models by generating these from both recordings of a physical soundfield, and from a virtual enrironment by producing acoustic simulations. The evaluation has acoustic simulations being supplied different sets of input materials, one of which is generated by the material tagging system. The other sets are a generic set (i.e. one material for all surfaces), and a manually-tagged set (i.e. produced by an audio engineering manually attributing materials).The experimental evaluation uses a real environment as a benchmark for testing how predicted models express acoustic parameters of measurable space. The goal is to measure which material set allows acoustic simulations to achieve the most similarity to the real, measured soundfield.

The evaluation compares an artistic material tag, *tagged*, which assigns acoustic geometry in the scene by applying the methodology used by game designers when tagging scenes; a *generic* tag, where all meshes map to a single acoustic material that expresses 0.1α acoustic absorption coefficient across all frequency bands; and finally, *predicted*, having meshes mapping to acoustic materials inferred by the proposed system. For each acoustic material, the sound renderer assigns its acoustic absorption coefficients to all triangles of the corresponding mesh.

Scene Geometry Reconstruction

Wavefields are simulated, via the wave-based renderer, from a real conference room that is $3.53m$ deep, $2.84m$ wide and $3.51m$ tall. Theoretically, the dimensions determine a Schroeder frequency of $261Hz$. The room geometry is composed of $2.9M$ of triangles and $6.6M$ vertices and is handled by the Unity game engine. The geometry of a conference room is reconstructed to deploy the proposed system in a real environment and conduct the subsequent subjective experimental evaluation. The reconstruction process adopts a LiDAR scanner, the FARO Focus^{3D} X300[2], to capture several point clouds scans of the room across 8 positions: 4 position points for each corner of the room at $1m$ height and 4 additional positions at $0.2m$ height to capture furniture and materials from different angles and enhance the spatial resolution. The reconstructed mesh obtained by registering and triangulating the point clouds using the FARO SCENE Software [3] is then manually segmented using Blender[4], to ensure that a separate textured mesh represents every scene object.

[2] https://www.faro.com/en/Products/Hardware/Focus-Laser-Scanners
[3] https://www.faro.com/en/Products/Software/SCENE-Software
[4] https://www.blender.org/



(a) Real space (b) Reconstructed environment

Figure 4.10: The evaluation design uses a real space to deploy the texture-based system. The real space, on the left, is captured with a LiDAR scan, obtaining a triangulated mesh. The virtual reconstruction of the space, on the right, is used to simulate a soundfield using acoustic materials inferred by the system. The simulated soundfield is then compared to the real counterpart.

Soundfield Measurements

The experimental evaluation compares simulated wavefields generated from the reconstructed environment to a sample measurement of the real counterpart's wavefield. The evaluation compares wavefields using RIR to describe acoustic properties dependent on geometry and materials surrounding a sound transmission between a source and a listener (Stan, Embrechts and Archambeau, 2002). The soundfield measurement process captures the environments' acoustic characteristics emitting and recording an exponentially swept sine, ranging from $20Hz$ to $20kHz$, emitted from a speaker and received by a microphone. The speaker-microphone position pair is reproduced in the virtual reconstruction of the space using a virtual sound source and receiver, see Figure 4.10. Applying inverse filtering, the RIRs is recovered from the captured signal (Holters, Corbach and Zölzer, 2009). Due to the objective nature of the evaluation, characteristics of human listeners are not considered by the evaluation, hence omnidirectional sources and receivers are used.

In the conference room, a public address system, the dB Technologies ES 1002[5], emitted the exponentially swept sine converted from a laptop using an Audient ID14[6] DAC and ADC, which captured the signal back through an omnidirectional, flat-frequency response soundfield measurement microphone, the Earthwork M30[7]. Steam Audio (Audio, 2020) is used as a standard acoustic renderer to simulate wavefields from the reconstructed environment. This allows the synthesis of wavefields based on acoustic geometry, considering

[5] https://www.dbtechnologies.com/en/products/es/es-1002/
[6] https://audient.com/products/audio-interfaces/id14/overview/
[7] https://earthworksaudio.com/measurement-microphones/m30/

absorption coefficients expressed across three frequency bands: low, medium and high. All simulations share the same resolution of 65536 and 16384 direct and secondary rays, respectively, with 256 bounces off solid geometry.

Maintaining the same settings for the generation of acoustic simulations, three sets of input acoustic materials are generated: the *generic* with a single acoustic material for the entire scene geometry; the *tagged*, with acoustic materials assigned through manual material tagging; and *ours*, using the proposed automatic tagging.

Perceptual test

The evaluation design also considers subjective factors by conducting a perceptual comparison between simulated wavefields at the same positions of source and listener, see Figure 4.10, by using an automated perceptual metric learned on subjects' responses (Manocha et al., 2020). The metric consists of a 14-layer deep neural network with filters trained on features extracted from paired input audio samples; it expresses a distance $D(x_{ref}, x_{per})$ between two input signals, where x_{ref} is a reference signal, and x_{per} is a perturbated signal. The function D considers factors including reverb and the ratio between direct and reverberated signal. The test evaluates whether the metric expresses a closer perceptual distance between the measured ground truth and the synthesised wavefields, by convolving the RIRs to samples from the evaluation subset of a database for acoustic scene classification, which comprises 15 different acoustic environments, generating a total of 18 minutes of audio over 1620 N samples (Mesaros, Heittola and Virtanen, 2016). The learned metric determines the distance between the ground truth convolutions and each of the relative simulated RIRs: for each audio sample k in N, the evaluation determines perceptual distances $D(x_{ref,k}, x_{per_i,k} \; \forall \; i \in \{generic, tagged, ours\})$, where ref and per_i are the measured and simulated RIRs.

4.3.3 Evaluation Results

The system takes around 7 minutes to compute acoustic materials for the 38 meshes composing the reconstructed environment. Figures 4.11, and 4.12 show the magnitude of early reflections for each acoustic simulation. Note that the ground truth shows distortion errors caused by the constant noise floor of $-55dB$ at the time of measurement. The *predicted* acoustic simulation marks the fastest decay in reverberation energy with a T_{60} of 4.4×10^{-1} s, computed parameterising the RIRs according to de Lima et al. (2009). The ground truth is 4.3×10^{-1} s, and generic and tagged are respectively $0.054s$ and 0.073.

The perceptual metric between the measured ground truth and the synthesised wavefields yields three variables describing 1620 pairwise distances between the measured and synthesised RIRs, see Table 4.4. Given the test conditions, the correlation test uses Pearson r scores to measure the relationship between RIRs with different acoustic geometry.

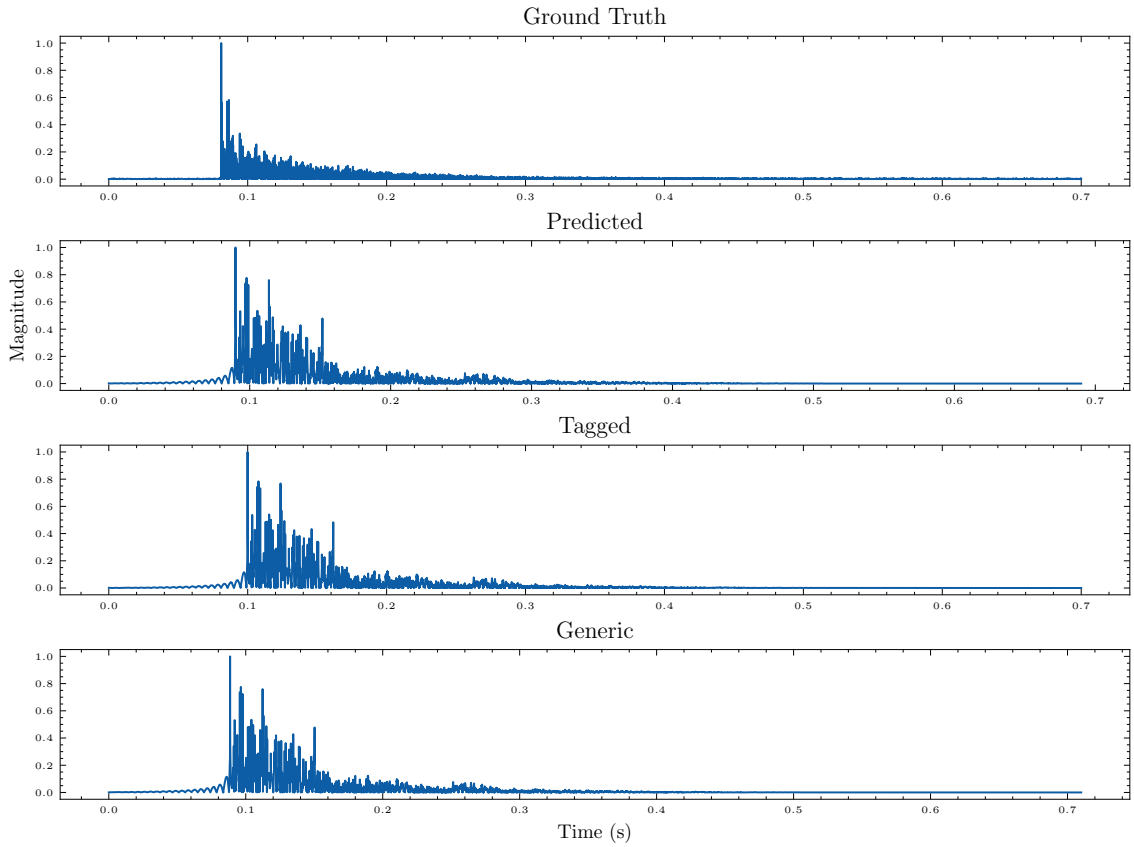


Figure 4.11: Measured impulse responses from measurements shown as normalised direct energy of reflections as a function of time $h(t)$. Energy functions of measured impulse responses, shown in the time domain where the horizontal axis indicates time, and the vertical axis indicates magnitude of reflection energy from 0 to 1.

Table 4.4: A table showing Pearson r scores for correlation measuring similarity across audio stimuli from simulated soundfields using different acoustic material sets. Generic acoustic materials are defined as 0.1 acoustic absorption across all frequency bands. Tagged materials are manually assigned, and Predicted materials are inferred from the texture-based material recognition system

		Generic	Tagged	Predicted
Generic	r	1	0.85	0.854
	Sig. (2-tailed)		0	0
Tagged	r	0.85	1	0.96
	Sig. (2-tailed)	0		0
Predicted	r	0.854	**0.96**	1
	Sig. (2-tailed)	0	**0**	

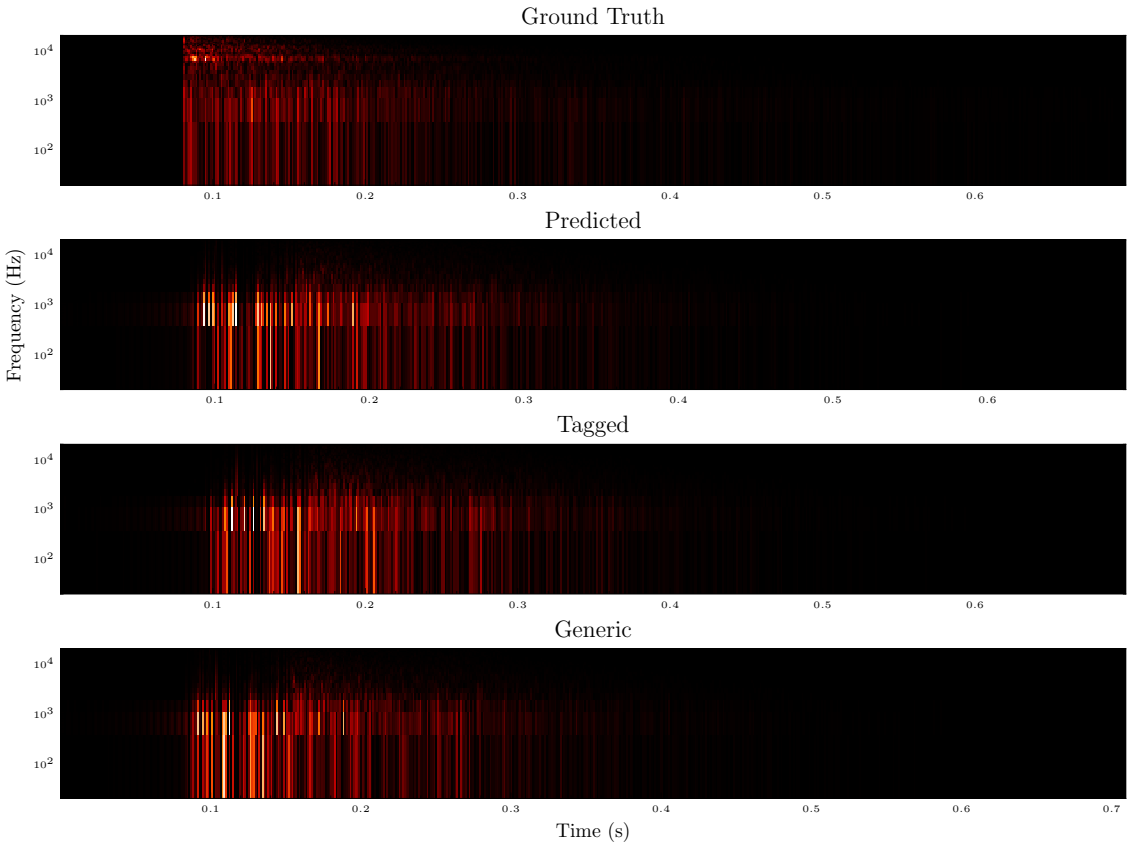


Figure 4.12: Spectrograms, computed with triangular windows, showing decay of early reflections across a logarithmic frequency range from $10^2 Hz$ to $10kHz$. Spectrograms comparing the frequency content of measured and simulated impulse responses. The impulse response generated with the proposed system maintains more energy over time.

4.3.4 Discussion

Experimental results show correlations between wave-fields generated from geometry that has been automatically tagged by the improved material tagging system; geometry tagged with a single generic material, and geometry tagged with human supervision. A preliminary perceptual test shows that tagging the acoustic geometry using a material classifier mapped to acoustic absorption data can produce acoustic simulations that correlate to their manually tagged counterparts.

As shown in the example of superpixel prediction in Figure 4.9b, the classifier can recognise many patches from the input image texture, associating them with labels that visually relate to the material represented, excluding outliers by selecting the most frequent prediction. In the evaluation, the mesh reconstruction process can introduce noise to the resulting texture due to the geometrical approximations caused by the point cloud's triangulation Turner, Cheng and Zakhor (2014). These approximations cause artefacts, resulting in incorrectly predicted material labels. While the superpixel segmentation process excludes these outliers, it prevents the feature extractor from capturing larger structures in the input texture.

The generic and tagged simulations, having a single acoustic material and approximated absorption data, are likely to subtract less energy from rays, maintaining more energy over late reflections. Hence, the limited number of rays and more specific acoustic absorption assigned to scene geometry determine shorter reverberation levels in the simulation with predicted materials, see Figure 4.12. According to the spectral analysis, this also causes the simulation to preserve more harmonic details above $10^3 Hz$. Overall the experimental evaluation shows that acoustic materials mapping through the texture-based system yields results that correlate to artistic material tags. Consequently, the improved system can tag acoustic materials in unseen complex scenes, reducing the costs of applications of wave-based methods in VEs. However, there are limitations in the test design that can affect the generalisability of the system, especially considering the single measurement point and the limited set of materials in the real environment. These limitations raise the need for an extended evaluation including different scene scales, i.e. larger rooms or outdoor environments having more sustained late reflections in their RIRs. Besides, task-based subjective experiments in VEs with subjects would further define the perceptual impact of acoustic material tagging on acoustic simulations.

4.4 Considerations on Acoustic Material Tagging Systems

In light of the two acoustic material tagging systems illustrated, this Section will draw conclusions from the data gathered and interpreted from tests run on both the camera-based and the texture-based prototype.

4.4.1 Conclusions Drawn From the Two Systems

Acoustic modelling and audio rendering methods can benefit from research and development of computer vision methods. The current status of this work does not eliminate the human-in-the-loop; however, it can generalise and operate on large sets of complex scenes. As a result, artistic and creative workflows for level design can benefit from an automated material tagging system that is agnostic of scene complexity and allows for easy integration of wave-based acoustic renderers.

The proposed systems addressed the problem of mapping acoustic absorption data to scene geometry in virtual environments for acoustic simulations. Methods and frameworks for material recognition have become efficient enough in recognising materials in the wild with varying factors of illumination, shape and surface characteristics. Despite their limited resolution in computer games applications, sound propagation systems benefit from acoustic material tagging. The proposed solutions aim to integrate material recognition into wave-based methods to determine materials' acoustic properties.

4.4.2 The Problem with Manual Material Tagging

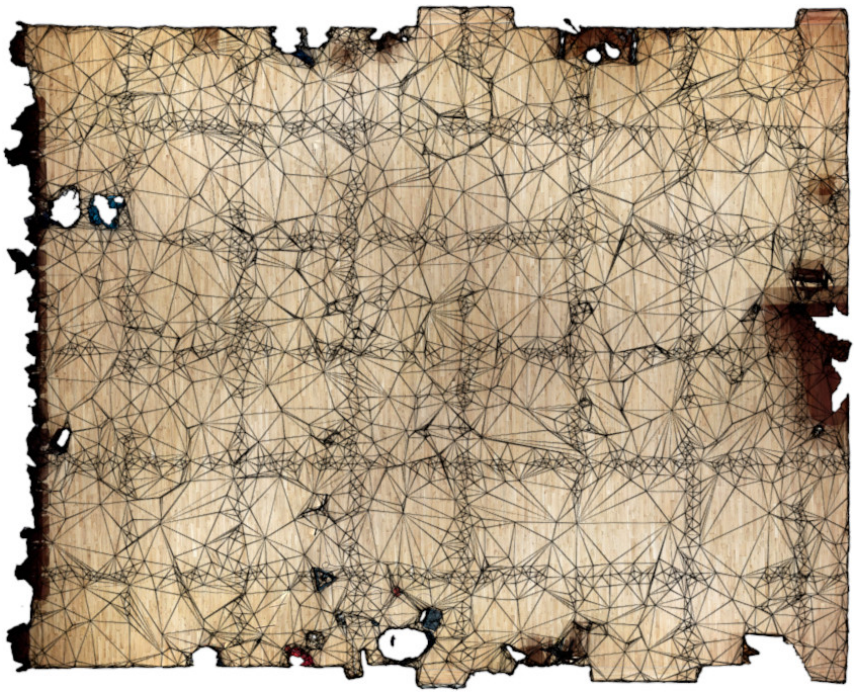
The closest approximation to ground truth for the design of evaluations is derived from scenes where an acoustic engineer manually tags materials using a pre-compiled database. However, this method introduces inherent biases due to the choice of materials being restricted to those available in the database. As such the manual tagging process has the engineer selecting from the closest match rather than an exact representation. The manually crafted ground truth contains biases caused by the limitations of the database used. This issue underscores the need for a more dynamic and comprehensive database that can adapt to include new materials and more nuanced categorisations as the field evolves.

However, the generation of acoustic materials has significant challenges in costs and resources as they often require the use of impedance tubes (Jones and Stiede, 1997). This can be unfeasible in diverse outdoor environments with complex sets of diverse materials, or where samples of materials are not easily available for impedance testing.

4.4.3 Advantages and Limitations of the Two Approaches

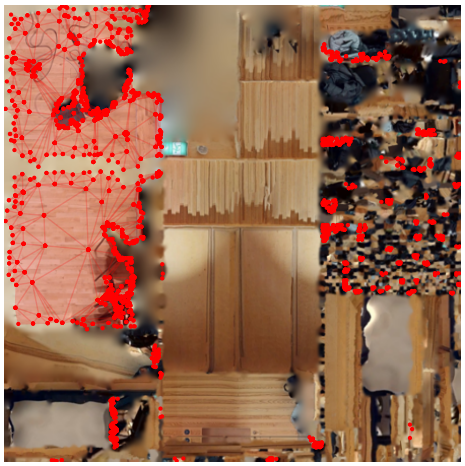
The camera approach can provide useful implementations in the context of MR platforms, as the camera render requirements can be satisfied by built-in HMD cameras. Scene reconstruction systems such as the Matterport cameras (Chang et al., 2017) share the requirements of capturing camera renders covering the entire real space.

The texture-based system can be highly efficient due to the mapping between texture data and scene geometry shown in Figure 4.13.



(a) Scene Geometry: Floor

(b) Texture Mapping 1



(c) Texture Mapping 2

Figure 4.13: UV mapping of one scene geometry segment to many texture segments.

Mesh Segmentation

Current space reconstruction algorithms that generate a triangulated mesh representing the physical space scanned have limited knowledge of the semantics of the reconstruction. Until the definition of a scene entity is given and a solution to distinguish physical scene entities in the reconstructed space, it is impossible to separate the triangulated mesh into semantically meaningful segments.

Such limitation is beyond the scope of this thesis work. It is a problem that can be addressed with complex mappings between defining a scene entity belonging to a physical environment and its reconstruction as a triangulated mesh. A family of techniques exists to address the problem of segmenting 3D scenes expressed as triangulated meshes or point clouds. However, the complex nature of three-dimensional space reconstructions makes the problem an open challenge. The work by Chen, Golovinskiy and Funkhouser (2009) is a significant stepping stone towards the problem of defining a physical entity represented in a complex mesh or point cloud, as the authors gather annotations from humans to investigate the consistency in segmenting entities represented in reconstructed data.

Such experiments have shown a consensus in the segmentation process and provided a benchmark as an evaluation tool and starting point for automated methods. There are techniques available that address the problem, such as Rusu and Cousins (2011)'s work providing a set of algorithms in Point Cloud Library, as well as new methods leveraging geometrical CNNs to segment 3D meshes and point clouds, such as Feng et al. (2020)'s work. Despite these techniques being available, there are still large margins of errors when deployed on in-the-wild reconstructions that require fine-tuning and human authoring.

Hence, the mesh segmentation stage of the material recognition pipeline is performed manually for future experiments discussed in this thesis, separating the reconstructed data into semantically meaningful segments with a minimum structural size of $70cm$, following Pelzer and Vorländer (2010)'s perceptual threshold for GA pipelines. Using Blender as a 3D data editor[8], the reconstruction obtained from the Matterport camera is separated into individual meshes, see Figure 4.14.

4.4.4 Contributions and Future Research Directions

The prototypes developed in this Chapter have demonstrated their capability to effectively tag entire unseen virtual environments with appropriate acoustic materials. This is due to the training phases where the systems learned to recognise and categorise various materials based on visual features. By employing machine learning techniques and a structured approach to feature extraction and material classification, these systems can autonomously apply accurate acoustic tags across diverse virtual environments, enabling more realistic and dynamically responsive acoustic simulations.

[8] https://www.blender.org/

Figure 4.14: Mesh segmentation process performed manually on a reconstructed scene. The figure shows the virtual reconstruction of the small test room after the mesh segmentation process. Scene elements are segmented using Blender8, allowing submeshes to represent each scene element. This allows the material tagging system to target scene elements.

Despite these promising results gathered from the camera and texture-based systems, several challenges persist, primarily due to the lack of comprehensive datasets and standardized definitions for acoustic materials. Limitations remain around the practicality and generalisability of the systems, requiring wider research areas to address problems around unified data on acoustic materials and the problem of segmenting unseen virtual reconstructions of real space.

The next steps of this work aim at extending the system to broader scenes with larger sets of materials as well as improving the material recognition by performing multi-scale analysis of superpixels and adopting clustering paradigms to overcome the limitations of finite material space definitions. Finally, the two systems aim at employing these methods to eliminate the human-in-the-loop for the process of labelling materials for acoustic renderers or assist in artistic and creative processes for level design. Both systems have potential in the context of MR platforms, where HMDs provide camera renders and virtual reconstructions of real space at interactive rates. Future work should expand towards extending the systems to infer acoustic materials in real environments, leveraging access to real soundfields to optimise predicted acoustic materials.

Chapter 5

Geometrical Acoustics Rendering Pipelines for Augmented Acoustics

This work was published in:

Colombo, M., Dolhasz, A., Hockman, J. and Harvey, C., 2022, August. Acoustic rendering based on geometry reduction and acoustic material classification. In 2022 IEEE Conference on Games (CoG) (pp. 409–416). IEEE.

This Chapter shows the prototyping of a bespoke acoustic rendering pipeline for dynamic virtual environments that integrates material recognition that can extend to immersive technology applications. The design process started with a simplistic form of acoustic rendering, using a standard method to approximate reverb, and expanded towards a bespoke ray tracing-based system that handles dynamic geometry.

5.1 Image Source-Based Rendering

The initial prototype acoustic renderer developed for this research utilises the Image-Source Model (ISM), a method for simulating reverberation in cuboid environments. This Chapter delves into the principles, implementation, and evaluation of several acoustic renderers, providing insights into their advantages and limitations. The ISM is a simple approach to rendering that approximates reverberation by simulating the reflections of sound waves off the surfaces of a cuboid environment, representing a useful choice for prototyping. By treating each reflection as an image source, the model can efficiently compute the paths that sound waves take from the source to the listener, accounting for multiple reflections. This method is particularly effective in environments with rectangular geometries, where the predictability of reflections simplifies the calculations involved. As discussed in Chapter 2,

the ISM technique is well-established in the sound rendering research domain, offering a balance between computational efficiency and auditory realism (Savioja et al., 1999; Allen and Berkley, 1979). The technique's ability to quickly simulate reverberation makes it an attractive choice for real-time applications, such as MR, where responsiveness is crucial. By leveraging the predictable nature of reflections in cuboid environments, ISM can generate realistic reverberation effects with relatively low computational overhead.

Figure 5.1 shows how the technique has been integrated into a sound rendering pipeline to compute reflection paths in a given virtual scene reconstruction. As a high-level summary, the *geometry reduction* component of this pipeline decomposes the virtual scene into a simpler cuboid volume. Vision-based material recognition techniques are used to understand the acoustic characteristics of surfaces in the input scene; these characteristics are then mapped onto the appropriate surfaces of the acoustic volume representing the given virtual scene. Finally, the ISM is supplied with spatial information on emitters and listeners to produce RIRs from the acoustic volume.

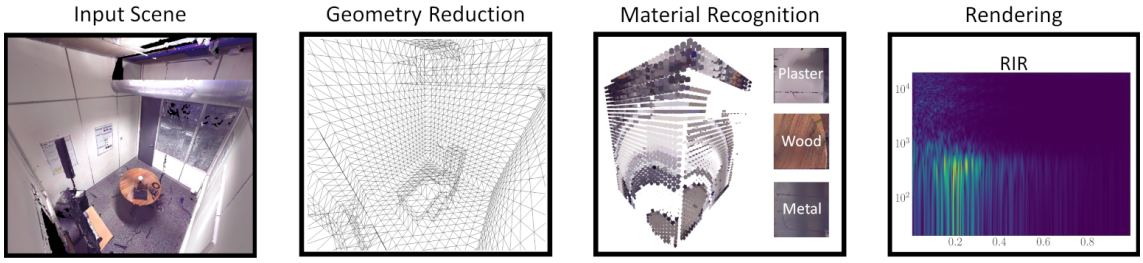


Figure 5.1: The Image Source Model (ISM) pipeline showing the four stages of the rendering process (left to right): an input scene is used as input; a geometry reduction process simplifies the complexity of the virtual environment to a cuboid volume; based on portions of the geometry dictated by the cuboid volume, materials of the input environment are determined; materials and simplified geometry provide input to the ISM to produce reverberation responses.

5.1.1 Geometry Reduction

The *geometry reduction*, Figure 5.1, generates a binary field, sampling from the input geometry at a given number of points. The field shapes the acoustic volume using a marching cubes algorithm, where each cell captures the appearance of surfaces from textures associated with meshes of the input geometry. A cuboid volume encapsulating the reconstructed surface determines the dimensions of the acoustic space simulated with the ISM. Using the vision-based techniques discussed in Chapter 4 to analyse surface appearance, acoustic materials are inferred per image patch, effectively mapping acoustic materials to the surface of the acoustic volume.

Input geometry is decomposed into a binary field expressed as $\mathbf{B} = (b_{x,y,z}) \in \{0,1\}^P$, where P is the number of points sampled across the three dimensions. The definition of $\mathbf{B}$ is relevant to the quality of the resulting acoustic simulation and determines the accuracy of the acoustic material recognition, discussed in the following section. Given the set of input meshes from the complex scenes, an Axis-Aligned Bounding Box (AABB) is defined

to encapsulate all vertices of a given scene object, normalising their coordinates to the $[-1, +1]$ range. Values in $\mathbf{B}$ equate to 1 whenever a cell is within the coordinates of any AABB defined and to 0 otherwise. The term AABB throughout this Chapter, will refer to the collection of all bounding boxes associated with scene objects throughout the rest of this Section. A multi-threaded implementation of the marching cubes algorithm (Bourke, 1994; Lengyel, 2019) shapes the binary field into a volume, representing the acoustic environment for reflection path computation. In this process, the use of the AABB to probe the input scene objects introduces error in cases with concave geometry, which negatively correlates to the resolution of $\mathbf{B}$ as collision checks with AABBs increase with the number of cells. A cuboid encapsulating the reconstructed surface from the marching cubes determines the ISM volume for the reflection computation stage.

Isosurface Reconstruction Error

The marching cubes algorithm is a powerful tool and a crucial step in the pipeline for reconstructing acoustic volumes by sampling complex geometries at various resolution levels. This technique is useful in acoustic rendering, where accurate representation of three-dimensional surfaces is crucial for simulating sound propagation and interaction with the environment. One of the key strengths is its speed and efficiency, as the algorithm is highly suitable for parallel processing. This parallelisation capability allows for significant improvements in processing times, which is essential for real-time applications in acoustic rendering. By leveraging modern multi-core processors and GPU acceleration, as seen in similar acoustic rendering using parallelisable computer shaders on GPUs (Savioja, Manocha and Lin, 2010), the algorithm can handle large scenes efficiently.

Despite its efficiency and adaptability, the marching cubes algorithm has inherent limitations that can affect its performance and accuracy in certain scenarios. It may oversimplify complex geometrical features, for instance, leading to a loss of important details that are crucial for accurate acoustic modeling. This can result in a less precise representation of the surface, affecting the fidelity of sound propagation simulations. The discrete nature of the algorithm can introduce artefacts, especially at lower resolution levels. These artefacts can manifest as unwanted distortions, irregularities, or floating pieces of geometry in the reconstructed acoustic volume, which can impact the accuracy of the acoustic rendering. The algorithm may struggle with accurately capturing sharp edges and corners, which are often critical for realistic sound reflections and diffractions. This limitation can lead to a smoothing effect that diminishes the realism of the acoustic model.

5.1.2 Camera Projection

Image patches are generated whenever a marching cube intersects a surface by an entire side, i.e. four consecutive corners of each marching cube. An orthographic camera, positioned at the centre of the neighbouring cube, rasterises sub-surfaces delimited by the cube. The camera is positioned on the opposite side of the four intersected corners, facing

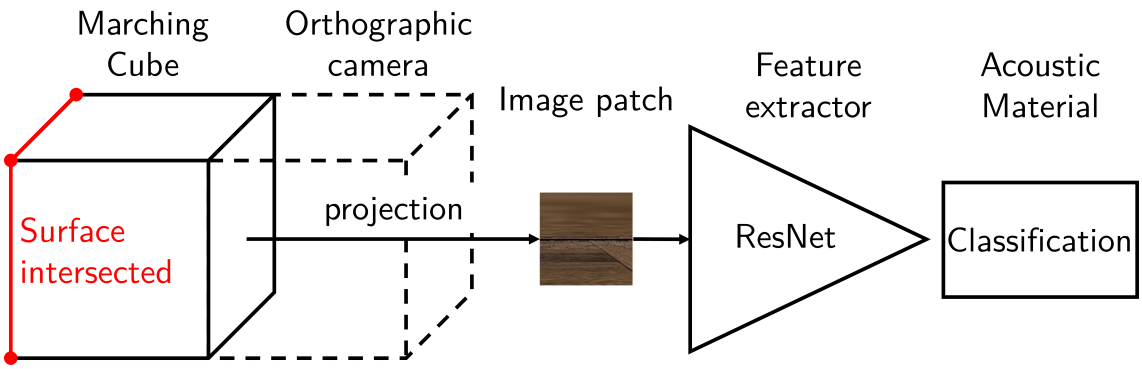


Figure 5.2: Image patch generation: surfaces intersected by marching cubes are projected onto image patches with a camera by rasterising vertices via orthographic projections of the delimited portion of the surface. The resulting image patch is then fed to a neural network to extract features for semantic classification. Through a one-to-many mapping, semantics attributed to a patch, via embeddings classification, map to acoustic parameters.

the surface that is intersected; see Figure 5.2. By using OpenGL rasterisation, surfaces are renderered orthographically from world space, defining the camera clipping space as the volume of the marching cube. Sampling image data from textures, the rasterisation stage produces an image patch representing the projection of the portion of the surface inscribed by a marching cube. Marching cubes intersecting in four consecutive corners ensure that the orthographic camera is perpendicular to the surface.

5.1.3 Acoustic Material Classifier

The acoustic renderer leverages material recognition techniques developed throughout Chapter 4, specifically adopting surface material classifiers within the context of predicting acoustic characteristics of handled scene geometry. This pipeline uses the geometry reduction component illustrated in earlier Sections to obtain a view of entities and surfaces on which material classifiers are applied. Image patches projected from marching cubes provide input to a neural network, a ResNet50 (He et al., 2016) backbone that operates as a feature extractor, whose output is forwarded to a densely connected layer. This last layer classifies predicted embeddings into acoustic material categories. The network was trained on the OpenSurfaces dataset (Bell et al., 2013), learning mappings between visual appearances of surfaces from 34 material classes and semantic labels, as described in Chapter 4. The network, pre-trained on ImageNet (Deng et al., 2009), learns on 32×32 pixel resolution image patches that are extracted from appearances sampled from OpenSurfaces, assembling a dataset of about 13M images, split into 9M and 4M train and evaluation sets, respectively. Via a one-to-many mapping, 34 categories from the OpenSurfaces dataset map to acoustic materials, and, given 11 acoustic materials, subdividing into two levels of mass density, the visual labels are mapped to 22 acoustic materials.

Classification Accuracy and Precision

The process of using visual representations of scene geometry to perform acoustic material inference can be prone to error due to the accuracy and precision of the classifier adopted. As discussed in Section 4.3, one crucial problem of material tagging is matching the ecosystem of materials represented by the classifier's training data and the target scene geometry to be predicted. State-of-the-art CNNs perform around 94% on in-the-wild datasets like OpenSurfaces (He et al., 2016; Bell et al., 2013). Hence, the acoustic rendering pipeline design should account for misclassified scene geometry segments.

The subdivision parameter offered by the geometry reduction components can have control over the impact of prediction errors caused by the classifier by offering coarse or fine subdivisions of the environment. With more subdivisions and a fine subdivision of the environment, the classifier can have more views of the scene geometry. Coarse subdivisions allow fewer predictions but increase the impact of misclassifications. Furthermore, misclassifications have varying degrees of impact on the resulting acoustic simulation: a misclassification of a segment of scene geometry that has low absorption as ground truth will have a higher impact on the simulation if tagged with a high-absorption material. Using materials defined in Table 4.2 as an example, the impact of a misclassification varies depending on the distance between the classified material and the ground truth on the spectrum of acoustic absorption.

5.1.4 Frequency-Dependent Reverb Approximation

As discussed in Section 2.3, the Image-Source Model (ISM) is a foundational technique used in acoustic rendering to simulate reverberation by modeling the reflections of sound waves within an environment. This model approximates how sound interacts with surfaces, creating virtual "image" sources that replicate the effect of reflections. While the ISM is computationally efficient and straightforward, it inherently suffers from limitations that affect its accuracy, particularly in approximating low-frequency reverberation. Techniques exist to overcome the limited representation of low-frequency information, such as wave-based acoustic simulations (Hamilton and Bilbao, 2017) or hybrid techniques, using wave-based but delegating high-frequency approximation to geometrical acoustics (Southern et al., 2013). However, these techniques have considerable computational resource footprints, making it difficult to integrate them into real-time applications.

Considering the pipeline at hand, the ISM offers fast and efficient reverb approximations at a trivial computational load, allowing the integration of computer vision techniques. The ISM estimates attenuation functions for a given listener position, with respect to a source in space, drawing from Habets (2006) implementation to estimate h reverb response functions based on the spatial $[x, y, z]$ coordinates of a sound source, $\mathbf{s}$, a listener, $\mathbf{l}$, at each

time step t as

$$h(\mathbf{l},\mathbf{s},t)=\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{m}\in\mathcal{M}}\beta_{-x}^{|m_x-q|}\beta_{+x}^{|m_x|}\beta_{-y}^{|m_y-j|}\beta_{+y}^{|m_y|}\beta_{-z}^{|m_z-k|}\beta_{+z}^{|m_z|}\frac{\delta(t-\tau)}{4\pi d}. \tag{5.1}$$

Here, $\mathcal{M}=\{(m_x,m_y,m_z):-N\leq m_x,m_y,m_z\leq+N\}$ determines the number of points per dimension based on the size of the input acoustic volume and the timestep, dependent on the desired sampling frequency, on the length of the output RIR, and the order of reflections N. $\mathcal{P}=\{(q,j,k):q,j,k\in 0,1\}$ determines possible combinations of image sources mirrored in the three dimensions from every boundary in the acoustic volume to consider higher-order reflections, which are computed by $\delta(t-\tau)$, where $\tau=\frac{\|\mathbf{R_p}+\mathbf{R_m}\|}{c}$ indicates the reflection time delay by dividing the measured distance between mirrored image positions $\mathbf{R_p}+\mathbf{R_m}$ and the listener by the speed of sound c, d is the distance term and is calculated as $\sqrt{(\mathbf{R_m}+\mathbf{R_p})^2}$. With combinations in $\mathcal{P}$, image positions are determined by

$$\mathbf{R_p}=[(1-2q)\mathbf{s_x}-\mathbf{l_x},(1-2j)\mathbf{s_y}-\mathbf{l_y},(1-2k)\mathbf{s_z}-\mathbf{l_z}]\quad\text{and}$$
$$\mathbf{R_m}=[2m_xl_x,2m_yl_y,2m_zl_z].$$

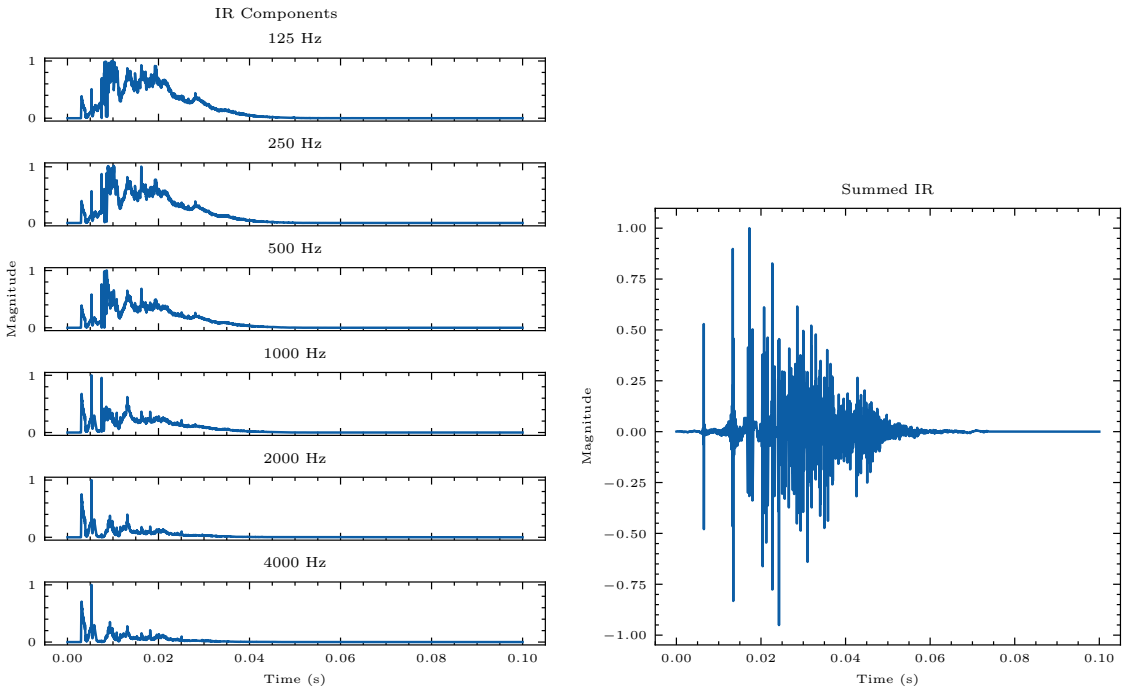
The ISM computes multiple h functions across frequency-dependent reflection coefficients, increasing the accuracy of simulated reflections from boundaries of the acoustic volume. In Equation 5.1, β reflection coefficients determine the energy attenuation of a computed reflection, specifying a single reflection coefficient mapping to each side of the acoustic volume; namely, $-x$ to $+z$ (left, right, top, bottom, front and back side). As these coefficients imply that materials apply a constant attenuation over the frequency spectrum, they can be redefined as $\beta_{-x,f},\beta_{+x,f},\ \ldots,\ \beta_{+z,f}\ \forall f\in\mathcal{F}:\{125,250,500,1000,2000,4000\}Hz$, where f indicates the frequency bin mapping to a reflection coefficient, adding a further dimension to Equation 5.1, which can be defined as:

$$h(\mathbf{l},\mathbf{s},f,t)=\sum_{\mathbf{p}\in\mathcal{P}}\sum_{\mathbf{m}\in\mathcal{M}}\sum_{f\in\mathcal{F}}\cdot\beta_{-x,f}^{|m_x-q|}\beta_{+x,f}^{|m_x|}\beta_{-y,f}^{|m_y-j|}\beta_{+y,f}^{|m_y|}\beta_{-z,f}^{|m_z-k|}\beta_{+z,f}^{|m_z|}\frac{\delta(t-\tau)}{4\pi d} \tag{5.2}$$

Frequency-dependent reflection coefficients enable the mapping between boundaries in the environment and acoustic materials. Equation 5.2 produces separate h attenuation functions for frequency bins in $\mathcal{F}$, associated with the common six-octave bands defined by the acoustic materials to cover the equivalent rectangular bandwidth-number scale (Kuttruff, 2016; Savioja and Svensson, 2015a).

5.1.5 Audio Rendering

The generated frequency-dependent h functions can be treated as finite impulse responses and can be used to produce a broadband RIR. Filters based around frequency-dependent absorption coefficients (six in this case) can generate a new set of h functions that contribute to a specific band of the equivalent rectangular bandwidth scale. Phase-invariant



(a) IR components

(b) Broadband RIR

Figure 5.3: Sum of IR components, creating a broadband response by filtering frequency-dependent energy logged by the ISM model.

low-pass filters, based on designs from Smith et al. (1997), are combined with their corresponding frequency-inverted counterparts, chained to produce band-pass filters. Filters are complementary to each other in the frequency domain ($20Hz$ to $20kHz$), summing to a flat magnitude response. Processed h functions are then summed into a resulting RIR, which are convolved to anechoic audio to propagate audio in the simulated acoustic environment.

The filter used in this process has an attenuation level exceeding 120dB between the passband and the stopband, removing unwanted frequency components from each frequency-dependent h function, resulting in a precise sum representing a broadband response, see Figure 5.3. The attenuation achieved by the filter is below the audible threshold of human hearing, having no perceptible impact on resulting auralisations.

5.1.6 Acoustic Volume Absorption

As discussed in Section 2.4.3, acoustic materials have a significant impact on acoustic simulations and auralisations. In this pipeline, a computer vision algorithm allows the prediction of absorption and reflections of portions of the acoustic volume. Historically, the Sabine and Eyring equations provided tools to predict the amount of absorption in a given environment, as they provide generalised metrics for acoustic design and analysis (Beranek, 2006). These formulas assume that the absorption is uniformly distributed and is most accurate in environments with moderate levels of absorption. The Eyring

formula, an extension of Sabine's work, provides a more precise calculation in highly absorptive environments by accounting for the logarithmic relationship between absorption and reverberation. However, an engineered computer vision system has higher precision and is able to predict segments of a given environment, rather than providing an estimate of the overall reverberation (Schissler, Loftin and Manocha, 2017). This ISM adapts to non-enclosed or partially enclosed space by determining acoustic materials associated with the six sides of the cuboid acoustic volume: when no image patches are assigned to a given side, its respective reflections are ignored, i.e. maximum attenuation. Otherwise, let $\mathbf{M}_{-x \ \ldots \ +z} = \{\alpha_{0,f}, \alpha_{1,f}, \ \ldots \ , \alpha_{n,f}\}$ be the vectors of acoustic materials corresponding to marching cubes, describing frequency-dependent α acoustic absorption. Considering marching cubes intersecting surfaces associated with a side of the volume at n points, acoustic materials generated by the *material recognition* stage substitute elements of vector $\mathbf{M}$, while the remaining elements default to air absorption (Kates and Brandewie, 2020).

Acoustic materials contribute to a final set of acoustic absorption coefficients

$$\alpha_{-x,f}, \alpha_{+x,f}, \ \ldots \ \alpha_{+z,f} \tag{5.3}$$

equivalent, as $\beta = \sqrt{1-\alpha}$ (Allen and Berkley, 1979), to the reflection coefficients considered in Equation 5.2. The contribution of each acoustic material is weighted based on the number of image patches found for each side and the total number of possible image patches P^3 that can be associated with each side, dependent on the resolution of the *geometry reduction* stage. Hence, the weighted average determining acoustic absorption for each side of the acoustic volume can be defined as:

$$\alpha_{-x,f}, \alpha_{+x,f}, \ \ldots, \ \alpha_{+z,f} = \frac{\sum_{i=1}^{P^3} w_i \ \mathbf{M}_{-x,i}, \mathbf{M}_{+x,i}, \ \ldots, \ \mathbf{M}_{+z,i}}{\sum_{i=1}^{P^3} w_i}, \tag{5.4}$$

where w indicates the weights vector defining acoustic material contribution:

$$w_i = \begin{cases} \frac{n}{P^3} & \text{if } i \leq n \\ \frac{(P^3 - n)}{P^3} & \text{otherwise} \end{cases} \tag{5.5}$$

Hence, the contribution of acoustic materials to the ISM volume depends on the surface intersected by marching cubes.

5.1.7 Preliminary Evaluation Using Learned Similarity

A preliminary objective evaluation was conducted on the initial renderer prototype by deploying the proposed pipeline on a set of scenes, providing coverage of a wide range of practical use cases: "Room", "Office", "Church" and "Village". These have increasing volume as reported in Table 5.1. "Room" is a real conference room captured using a LiDAR scanner, FARO Focus^{3D} X300. The remaining scenes are virtual environments

that are common in game development.

Using the initial ISM prototype renderer, RIRs are generated across the four scenes, applying all stages of our pipeline as offline procedures, comparing acoustic simulations with inferred acoustic materials to simulations with manually tagged materials. These acoustic simulations assume omnidirectional sources and receivers, as no human listeners are involved in the comparison. Hence, simulations disregard directivity patterns or head-related acoustic phenomena. Source-receiver position pairs are consistent across RIRs pairs. Objective metrics associated with room acoustic parameters are extracted to compare the output of each acoustic simulation.

5.1.8 Similarity Metric Test

Overview

To gather insights on the potential deployment of the pipeline on auralisation systems for interactive applications, a learned similarity metric was employed to collect data from a set of scenes. The learned similarity metric compares sets of audio transmissions simulated using the rendering pipeline differing by input acoustic materials. This allows to determine whether the pipeline can render the difference between materials, generating discussions on the feasibility of the implementation of material recognition into acoustic rendering for wearable computing.

Procedure

The similarity procedure uses the prototype renderer to construct a cuboid acoustic volume encapsulating the source-listener pair, producing image patches with captured surfaces and infering acoustic materials to generate acoustic materials associated with the volume. For each scene, *two* RIRs are generated to allow auralisations: *one* using automatically-generated acoustic materials and *one* using manually-tagged acoustic materials. RIRs are compared by utilising a pre-trained network for subjective comparison between auralisations produced via convolution of RIRs to audio from a database of anechoic recordings of sound events from the TUT Sound Event database (Adavanne, Politis and Virtanen, 2018). The deep audio perceptual similarity metric, CDPAM (Manocha et al., 2021), trained on a dataset of human judgements, expresses distances between two audio signals and a reference. The network can be adopted as a metric that measures the Just-Noticeable Difference (JND) as a distance $D(x_{per}, x_{ref})$ between a reference signal x_{ref} and a perturbed x_{per}. Reverberation or equalisation are among the perturbation factors affecting the measured distance. Perceptual distance scores greater than 1 indicate that a human would distinguish them as distinct. The objective of the evaluation is to test whether sound propagated using the rendering pipeline with inferred acoustic materials is perceptually indistinguishable from sound propagated with manually tagged materials. The rationale for setting this test objective aims at removing the human from the material tagging workflow: if the system is able to correctly predict materials without relying on manual authoring of

acoustic materials, the sound rendering pipeline can autonomously work on unseen complex scenes. For this test, RIRs are used to generate convolutions of a collection S of 2700 audio samples to determine the perceptual distances $D(x_{predicted,i},\ x_{tagged,i})\ \forall i \in S$ between convolutions using predicted and manually tagged materials.

5.1.9 Results

The classifier takes an average of $0.03s$ to infer the acoustic material from an image. The prototype rendering pipeline is tested by comparing acoustic energy decay across computed RIRs, as it expresses reflection paths computed on scene geometry for a given source-listener position pair. Metrics from impulse responses are extracted following de Lima et al. (2009)'s feature analysis definitions. By fitting energy decay curves, the T_{60} reverberation metric, the C_{50} clarity index, and the D_{50} definition index are determined. C_{50} and D_{50} indices are dependent upon the ratio between the power of early and late reflections. See Table 5.1 for estimated reverberation, clarity and definition scores across scenes. Figure 5.5 shows distributions of perceptual distances from RIRs generated using automatically tagged materials to RIRs generated with manually assigned materials.

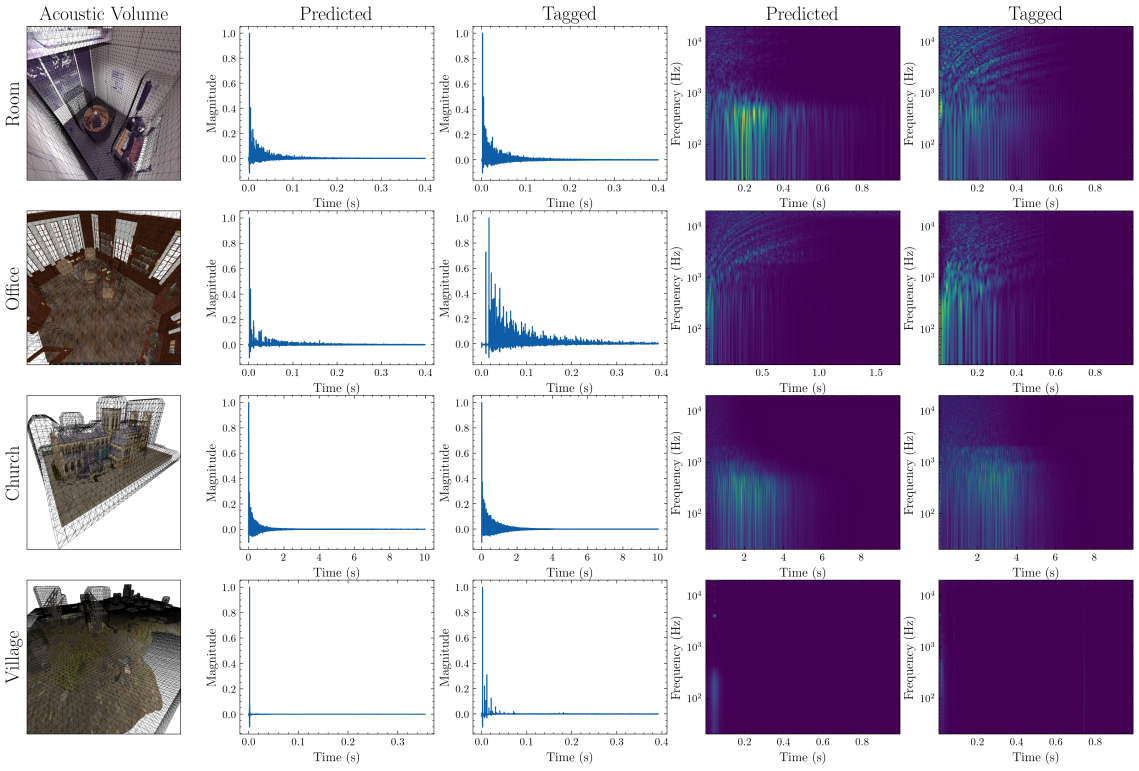
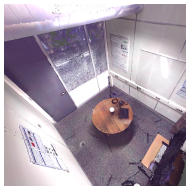
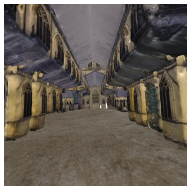
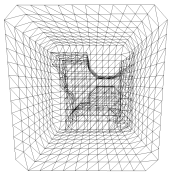
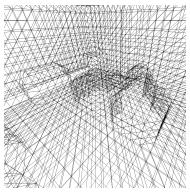
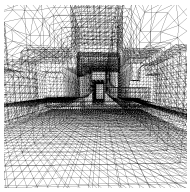
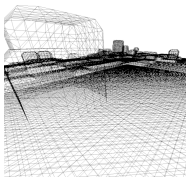


Figure 5.4: Comparison between room impulse responses, generated using the proposed framework.*Predicted* RIRs are produced using materials inferred by the automatic acoustic material classification, while the *tagged* counterparts have manually tagged acoustic materials. Rows show scenes in ascending order of volume; Columns from left to right show a render of the scene with an overlapped polygonised acoustic volume resulting from the marching cubes algorithm; a time-domain visualisation of the impulse response using *predicted* materials, followed by the counterpart with *tagged* materials; finally, the last two columns show spectrograms of the two. RIR pairs are generated, maintaining the same positions of source and listener.

Table 5.1: Features extracted from room impulse response pairs generated using the proposed system and metrics of corresponding input environments. Each pair has a *predicted* and *tagged* RIR, referring to acoustic materials being inferred with acoustic material classification or tagged manually. t refers to the time taken to compute reflections, and P indicates the number of sampling points per dimension.

	Room	Office	Church	Village
Scene				
Acoustic Volume				
Volume (m^3)	3.43×10	8.64×10^3	4.83×10^4	3.45×10^8
Triangles	15.4M	0.973M	0.49M	9.6M
P	16^3	32^3	64^3	128^3
Order	3	4	5	1
$t(s)$	2.824	3.53	2.791	1.25
predicted $T_{60}(s)$	0.997	1.705	5.982	0.103
tagged $T_{60}(s)$	0.331	0.734	9.6	0.124
T_{60} error	0.666	0.971	3.618	0.021
predicted C_{50}	1.278	-2.613	-12.341	-5.724
tagged C_{50}	0.709	-1.193	-13.706	-3.036
C_{50} error	0.569	1.42	1.365	2.688
predicted D_{50}	0.499	0.384	0.048	0.275
tagged D_{50}	0.509	0.619	0.035	0.36
D_{50} error	0.01	0.235	0.013	0.085

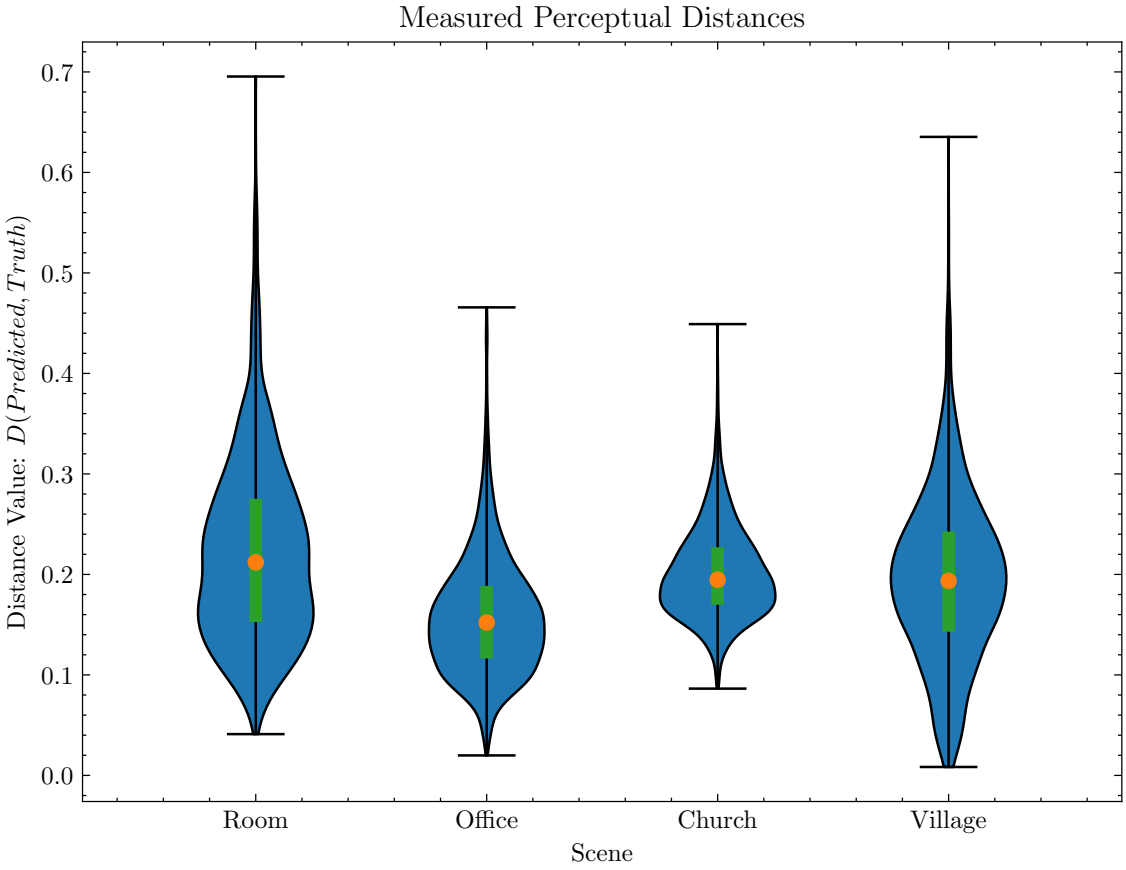


Figure 5.5: Distributions of perceptual distances between pairwise comparisons of audio recordings convolved to generated RIR pairs. A collection of everyday sounds is propagated for each pair, producing convolution pairs. By employing a learned metric, the perceptual distance between RIR with *predicted* and *tagged* materials is measured. The violin plot reports that all distances fall below one just noticeable difference, indicating that a human would be unlikely to distinguish between the two convolutions. Distributions and maxes are outlined in blue and black, respectively. Medians and both lower and upper quartiles are outlined in orange and green, respectively.

5.1.10 Discussion

Overview

Given that procedures of isosurface extraction and computation of frequency-dependent impulse responses run on CPU programs, the timings recorded across the four scenes with increasing spatial resolution suggest that the pipeline would be practical for real-time platforms. This could hold true even when considering the dynamic geometry typical of XR systems. Furthermore, GPU implementations of the ISM efficiently distribute sources across parallel workers, allowing for real-time RIR generation (Diaz-Guerra, Miguel and Beltran, 2021). The technique illustrates promise in the domain of streaming geometry, where virtual environments are constructed via spatial mapping services for visualisation on extended reality displays.

There is a necessity for understanding the acoustic information to be associated with this streamed mesh data. The nature of this approach facilitates mesh ingestion and updates to the binary field; subsequent subregions of the marching cube volume can be iteratively updated, and the resultant RIR to take stock of updated geometry can thus be generated. Despite overcoming the limitations of the ISM in propagating sound in non-enclosed environments, phenomena such as the occlusion of sound sources and arbitrary shapes of the environments are not considered. Occlusion and visibility of sound sources can be solved by combining the ISM with ray tracing, allowing for checking source visibility and introducing limited overhead thanks to their GPU implementations (Taylor et al., 2012).

Acoustic Material Approximation

The complexity of the pipeline depends upon the nature of the complex scenes and the number of points in which surfaces are sampled. Considering the overhead introduced by the acoustic material classifier, the complexity of the pipeline scales linearly with the number of surface intersections in the environment. Hence, the worst complexity occurs when each marching cube intersects scene geometry. In addition, despite the classifier's reasonable accuracy on test data, no ablation studies have been conducted to reduce the architecture to a minimum topology and further optimise complexity. Acoustic materials associated with each side of the calculated volume contribute to a mean, causing the ISM to approximate the computation of specular reflections, neglecting characteristics of surfaces, such as position or orientation. This approximation can be overcome by eliminating the process of averaging acoustic materials and reformulating the computation of image sources defined in Equation 5.2 to refer to individual acoustic materials mapped to the acoustic volume. However, the reformulation would affect the timestep of computed h functions, constraining it to the geometry reduction resolution, resulting in an arbitrary scale that should be interpolated to the time scale.

The benefits of removing mean acoustic material would include the ability to simulate arbitrary shapes of surfaces by having acoustic materials mapped to marching cubes. This

process would need to consider Nyquist sampling theory to determine the appropriate cube sizes to simulate accurate acoustics whilst maintaining specular plausibility in the frequencies simulated. (Pelzer and Vorländer, 2010)'s work, in addition, suggests that the resolution of the geometry reduction process can be set according to perceptual responses in the resulting simulation. Their experiments reveal that geometry with small structural details can be excluded from acoustic modelling, maintaining the perceived quality, and this acts as a motivating basis for this future iterative study.

Geometry Reduction

Determining space subdivision through the resolution factor P^3 of the *geometry reduction* stage has an effect on the volume reconstruction and generation of image patches, which directly map to acoustic materials. Larger resolutions require more marching cubes, causing the number of orthographic projections from surfaces to increase, resulting in a higher number of forward passes through the feature extractor, finally resulting in increased computational overhead. In order to maintain perceptual accuracy and produce plausible acoustic simulations whilst minimising the spatial resolution to optimise execution times, further work would require subjective evaluations to derive optimal spatial resolution across varying scene geometry.

Acoustic Volume Reconstruction Error

As discussed in Section 5.1.1, the use of the marching cubes algorithm for the *geometry reduction* provides an efficient tool for sampling and reconstructing complex virtual environments. However, it can be prone to error as the isosurface reconstruction process can fail to represent important features of the input geometry.

One promising approach to mitigate the issue of information loss due to isosurface reconstruction is the integration of neural importance sampling within the geometry reduction process. Neural importance sampling leverages machine learning techniques to prioritise and refine the sampling of critical geometrical features, improving the overall detail and accuracy of the reconstructed mesh (Müller et al., 2019). Neural importance sampling involves training neural networks to identify and focus on the most significant features of geometry during the sampling process. By learning which regions of the geometry contribute most to the acoustic properties, the algorithm can allocate more computational resources to accurately reconstruct these areas. This targeted approach helps preserve essential details that are often lost in traditional isosurface approximations.

Objective Evaluation

The most noticeable differences between renders with predicted and manually tagged materials are due to different decays of acoustic energy. By considering spectrograms of generated RIRs shown in Figure 5.4, the different acoustic materials influence the decay

of energy over time. As a result, there are errors relative to reverberation, definition and clarity; see Table 5.1.

5.1.11 Conclusions and Proposed Design Strategy

While the ISM is a valuable tool in acoustic rendering, its limitations in representing low-frequency information pose challenges for creating fully immersive auditory experiences. Mitigation strategies, such as incorporating frequency-dependent acoustic absorption data and generating frequency-dependent energy functions, offer effective solutions to these challenges. By enhancing the ISM with these techniques, it is possible to achieve a more accurate and realistic representation of reverberation, ultimately improving the quality and immersion of acoustic simulations.

The initial prototype represents a novel pipeline for acoustic rendering that is able to capture acoustic material characteristics of space around a listener using computer vision paradigms. These predicted acoustic material characteristics are used to generate input for sound propagation methods, producing plausible acoustic simulations. The proof-of-concept prototype is executed as offline procedures implemented as CPU programs, demonstrating the generation of RIRs that can be used in downstream convolution auralisations in real-time audio engines to propagate audio from virtual sound sources in simulated environments. The automated mapping between visual appearances and acoustic characteristics directly applies to extended reality platforms where the virtual environment is incrementally reconstructed as a listener explores their surroundings, enabling sound rendering to produce plausible acoustic simulations and removing human experts from the scene authoring process.

The preliminary design and evaluation cycle shows that the bottleneck of the rendering pipeline is at the core reverb approximation method, which should be improved by exploiting the space partitioning capabilities of the geometry reduction system. A ray tracer could be a more suitable design choice to interact with individual partitions of the generated acoustic volume.

5.2 Ray Tracing-Based Rendering Pipeline

Considering the limiting factors of the ISM-based renderer, such as the inability to render complex architectures of environments or the inability to model acoustic phenomena beyond basic reverberation, the pipeline design was improved by adopting a ray tracer system. The targets for the rendering pipeline should be oriented towards:

- approximating basic acoustic phenomena based on architectural features of the environment where auditory interactions take place;
- enabling dynamic partitioning and indexing of the space, considering material recognition systems;

- leveraging fast rendering and processing techniques that can be optimised for wearable computing platforms such as HMDs.

Overview

Ray tracing techniques have inherent limitations in modelling sound propagation within a space due to their nature of representing sound waves as rays. As discussed in Section 2.3.5, such limitations can cause physical inaccuracies in the simulated acoustic model, affecting the resulting audio interactions due to incorrectly simulated phenomena. An example is occlusion, where a sound source might be inaudible due to the obstruction and absence of paths between source and listener, though in the physical space, some low-frequency acoustic energy will still reach the listener due to their omnidirectional propagation patterns.

Although such limitations can cause inaccurate acoustic models, ray tracers still provide accurate reverberation and reflection approximations and are widely employed in modern acoustic surveying frameworks like ODEON[1] due to their ability to estimate key properties of soundfields and fast auralisations. Moreover, the constant rise in computing power offered by modern general-purpose GPUs offers fast implementations of geometrical acoustics that can be accelerated thanks to their architecture and the parallelisable nature of ray tracing (Savioja, Manocha and Lin, 2010).

5.2.1 Geometrical Search

As discussed in Section 2.4.2, the foundations of an acoustic ray tracer lay in the interactions between the primitive used to approximate a propagating sound wave and the environment. The modelling of acoustic phenomena in acoustic simulations produced by ray tracers is a process dependent on physics interaction between the geometrical primitives that compose the complex scene. Ray Tracing uses rays and line segments as geometrical primitives, requiring the environment handling system to facilitate operations between rays and the complex scene.

In acoustic simulation domains, as overviewed in Chapter 3, handling scene geometry is a central problem of the task, which has been approached from several angles ranging from classic computer graphics techniques to computer vision and deep learning solutions. A Bounding Volume Hierarchy (BVH) or Binary Space Partitioning (BSP) is often employed for handling scene geometry in acoustic simulations (Schröder, 2011). BVHs organise the scene geometry into a tree structure of nested volumes, typically using Axis-Aligned Bounding Boxs (AABBs). Each node in the hierarchy contains a bounding volume that encompasses the scene's triangles. This hierarchical approach allows for rapid traversal and collision detection, significantly reducing the computational load when determining sound interactions within complex scenes.

[1] https://odeon.dk/

BSP, on the other hand, divides the scene into a series of convex subsets using hyperplanes. The BSP tree is constructed by recursively splitting the scene with planes until each leaf node contains a manageable number of triangles. This method provides an efficient way to manage visibility and occlusion, as the BSP tree can quickly determine the spatial relationships between objects. Both BVH and BSP trees enhance the performance of acoustic rendering algorithms by optimising the geometric data processing, enabling more accurate and real-time simulations.

Recent advancements in deep learning have introduced novel techniques for handling scene geometry by creating latent representations of the environment (Liang et al., 2023). These techniques leverage neural networks to learn compact, high-dimensional representations of complex scenes from raw geometric data. By training on large datasets of scene geometries, deep learning models can encode essential features and spatial relationships into a latent space, capturing the critical aspects of the environment that influence sound propagation (Ratnarajah et al., 2022; Ahn et al., 2023). As illustrated in Chapter 3, deep learning-based approaches have not seen as much progress as tree structures due to their recent nature; BVHs have become fast and feasible for interactive applications (Kopta et al., 2012).

The ray tracing-based renderer design implements a Bounding Volume Hierarchy (BVH) as a geometry handling system, allowing the dynamic indexing and searching of a given complex scene. The BVH implementation constructs a binary tree of AABBs encapsulating triangles of the mesh representing the VE and any entity interacting with the acoustic environment, e.g. a character representing a listener. A dynamic BVH implementation uses the environment, represented as a set of triangulated meshes, to construct a set of bounding boxes that facilitate geometrical search operations such as ray-box intersections or point-in-volume tests.

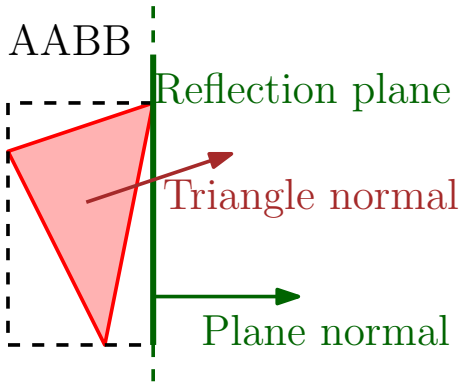


Figure 5.6: An Axis-Aligned Bounding Box (AABB), leaf node of a scene BVH, encapsulating a mesh triangle. The AABB facilitates geometry indexing and searching, optimising ray tracing operations such as ray-box intersections, and is used to provide reflection planes constructed based on the encapsulated triangle normal for reflecting rays.

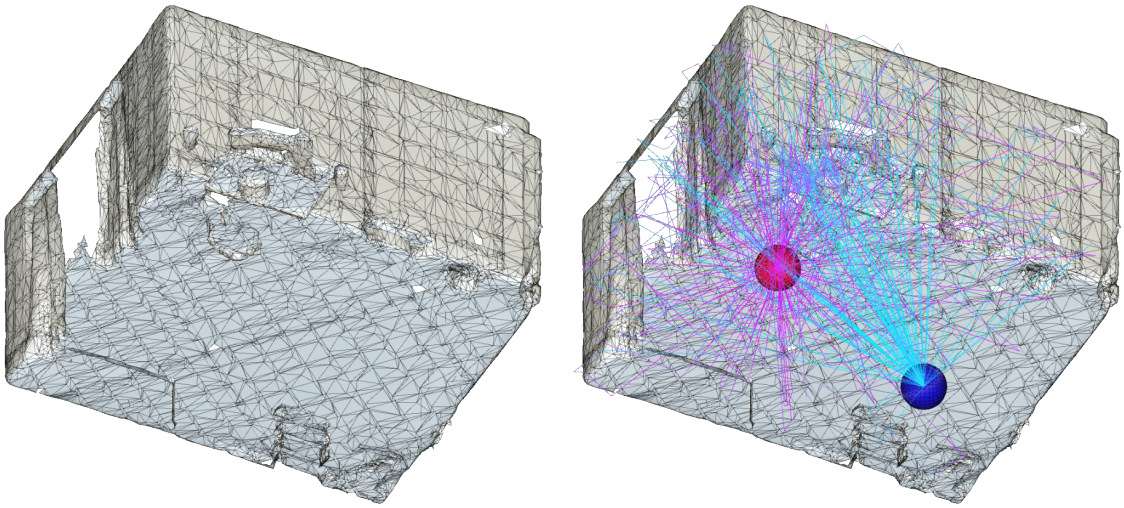


Figure 5.7: Visualisation of propagated rays from a sound source, red sphere, to a listener, blue sphere. The left image shows a visualisation of a triangulated mesh provided to the ray tracer; the right image shows rays originating from across uniformly distributed points on a sphere, reflecting off mesh triangles, and arriving at the listener sphere volume.

5.2.2 Sound Propagation

Rays propagate from a given spherical object, the sound source, visualised as a red sphere in Figure 5.7, originating from N points uniformly distributed across the azimuth and elevation of the sphere. This chapter deals with the approximation of basic acoustic phenomena achieved by rendering pipelines and excludes directivity patterns for sources and receivers; hence, sources and listeners are omnidirectional emitters and monopole detectors unless explicitly defined otherwise. Propagation paths between source and receiver are expressed as three-dimensional segments determined by intersections caused by rays, containing acoustic features such as acoustic energy carried. The current prototype does not consider directivity patterns, expressing sound sources as omnidirectional emitters and listeners as omnidirectional receivers. The problem of emitter and receiver directionality profiles is considered and discussed throughout Chapter 7, alongside the introduction of human factor to the overarching thesis problem.

The number of rays emitted from the source determines the initial energy e_0 of propagation paths, $e_0 = 1/N$, where the sum of all propagation paths amounts to 1.

Algorithm 1 shows an overview of the main procedure of the ray tracer renderer. Starting from uniform sampling across azimuth and elevation points of the spherical volume of S, rays, defined by their origin and direction, are emitted from the spatial position of S and oriented based on the sampled points across the sphere (Shirley and Morley, 2008). The number of rays required to obtain perceptually valid propagation paths is generally between about $60,000$ and $100,000$ (Le Bot and Bocquillet, 2000; Pompei, Sumbatyan and Todorov, 2009; Pelzer and Vorländer, 2010). The complexity of the simulation scales rapidly with the order: as more bounces off geometry are computed, more geometrical

search operations are required, increasing the number of BVH traverses and the overall computational complexity.

Reflection

Figure 2.12 illustrates the computation of propagation paths overview in Algorithm 1. For the sake of simplicity, only reflection paths are illustrated in the pseudocode listing, as shadow paths follow the same logic and registration procedure. The algorithm checks for intersection along rays propagating from the Source object bounds (the red sphere volume in Figure 5.7). Intersections found at surfaces create reflections around the computed surface normal by considering the leaf node AABBs encapsulating mesh triangles: rays are checked for collisions by traversing nodes of the scene BVH and testing ray-box collisions until a leaf node is reached. Leaf nodes are AABBs encapsulating mesh triangles; they are the primitives of the reconstructed acoustic volume. As shown in Figure 5.6, propagation paths reflect around normals of reflections planes fit on surfaces of the AABB. Upon intersection of a ray with a leaf AABBs, the dot product of the encapsulated triangle normal is pre-computed for all six faces of the box; a plane is then fit on the face that aligns the most with the triangle normal, approximating the acoustic geometry to a Manhattan space.

The process of computing reflections is central to sound propagation as it introduces materials and the effect of surface acoustic characteristics into the acoustic simulation. Propagating paths with initial energy e_0 have their energy detracted by intersections with AABBs. Material properties, here, are determinant of perceptual characteristics of the final simulation due to their absorption characteristics affecting energy modelling. Materials can express several characteristics beyond acoustic absorption, modelling surface properties and affecting phenomena of sound propagation (Kuttruff, 2016). The scope of the model at hand is limited to reverberation, which correlates to how energy is distributed throughout the generated impulse response, due to it being a primary determinant of realism and has a direct impact on psychoacoustic factors (Liu and Manocha, 2022; Rungta et al., 2016).

Absorption

Leaf nodes of the scene BVH carry frequency-dependent acoustic absorption coefficients that affect the energy of propagation paths colliding with scene geometry. As discussed in Section 4.2.2, coefficients express acoustic energy absorption over six frequency bands distributed over the human hearing frequency range. Propagation paths have e_0 initial energy across these six frequency bands, and upon collision with scene geometry, energy is absorbed based on the acoustic material assigned to the mesh triangle encapsulated in the colliding scene BVH leaf node. In pseudocode, $path.currentEnergy_f = path.currentEnergy * (1 - reflector.absorption_f)$ shows energy subtraction upon collision, with f denoting frequency bands $\{125, 250, 500, 1000, 2000, 4000\}Hz$ Table 4.2 shows

Algorithm 1 Main procedure of the ray tracer: propagation paths are computed given Source/Listener bounds S and L, a number or rays N to propagate, and a max order o. for each propagation point on the emitter bounds, the geometry is searched for intersections a number of times depending on found intersections and the max order. These paths are registered if they intersect with the listener bound L.

define $Ray(origin,\ direction)$
define $Vector3(x, y, z)$
define $Path(e_0)$ ▷ e_0: initial path energy
function *RandomOnUnitSphere* ▷ Random direction vector, uniformly sampled on unit sphere
    $u_\theta \backsim U(0, 1)$
    $u_\varphi \backsim U(0, 1)$
    $\theta \leftarrow 2\pi u_\theta$ ▷ Uniform random between 0 and 2π
    $\varphi \leftarrow \arccos(2u_\varphi - 1)$ ▷ Uniform random between 0 and π
    $x \leftarrow \sin(\varphi)\cos(\theta)$ ▷ Spherical to cartesian coordinate conversion
    $y \leftarrow \sin(\varphi)\sin(\theta)$
    $z \leftarrow \cos(\varphi)$
    return $Vector3(x, y, z)$
end function
procedure ComputePropagationPaths(S, L, N, o)
    $e_0 \leftarrow 1/N$
    for all $i \in [0..\sqrt{N}]$ **do** ▷ Uniform sphere sampling: azimuthal steps
        for all $j \in [0..\sqrt{N}]$ **do** ▷ Uniform sphere sampling: elevation steps
            $dir \leftarrow RandomOnUnitSphere()$
            $path \leftarrow new\ Path(e_0)$
            $path.addRay(new\ Ray(S.position, dir))$
            for all $order \in [0..o]$ **do**
                if *path.currentRay* Intersects *Listener.bounds* **then**
                    register specular path ▷ occlusion between path and L is computed
                    break
                end if
                if *intersection* between Geometry and *path.currentRay* **then**
                    $reflection \leftarrow$ reflect *path.currentRay* with *intersection*
                    $path.addRay(new\ Ray(intersection,\ reflection))$
                    $path.substractEnergy(intersection.absorption)$
                end if
            end for
        end for
    end for
end procedure

example acoustic materials that can be attributed to leaf nodes.

The prototype acoustic render considers constant air absorption due to the deployment of the prototype in an environment with constant temperature and humidity, as shown in Chapter 7. Air absorption is applied to the ray tracer based on standard measurements, See Table 5.2. Absorption is calculated over the six frequency bands at $2.0 \times 10^{1\circ}$ environment temperature and 60% relative air humidity (Nikolic, 2002; Harris, 1966). The temperature is considered to remain constant for the duration of each acoustic simulation.

Table 5.2: Acoustic absorption coefficients applied to propagation paths to simulate air absorption. Coefficients are computed for 20° air temperature.

Relative	Frequency (Hz)					
Humidity (%)	125	250	500	1000	2000	4000
40	0.000	0.000	0.001	0.001	0.004	0.024
50	0.000	0.000	0.000	0.001	0.003	0.02
60	0.000	0.000	0.001	0.001	0.003	0.017
70	0.000	0.000	0.001	0.001	0.003	0.014

Shadow Paths

Diffraction simulations, as demonstrated by diffraction modelling using the Uniform Theory of Diffraction (UTD) Tsingos et al. (2001), use reflecting geometry and edges as sources of new diffracted rays, complementing specular reflections. Similarly, the ray tracer considers the possible connections between intersections at intermediate nodes of propagation paths and the receiver volume, see Figure 2.12. Such shadow paths contribute to the modelling of the energy decay and exploit computations involved in geometry searches and occlusion tests.

This model can be extended to dynamic scene geometry with real-time BVH implementations that allow fast rebuilding and rotations of tree (Kopta et al., 2012).

5.2.3 Energy Decay Modelling

Due to frequency-dependent absorption properties of the scene geometry, specular and shadow rays generate propagation paths having varying remaining energy across the six frequency bands. Upon the execution of the procedure in Algorithm 1, energy data can be extracted at the detector bounds by measuring all registered paths, creating frequency-dependent impulse responses, see Figure 5.8. To mitigate the deterministic nature of a ray tracer, a Poisson distribution is used to introduce randomness in the generation of an RIR, simulating chaotic aspects of acoustic energy reflected in an environment.

As Schröder (2011)'s work demonstrates the efficacy of Monte Carlo methods combined with deterministic geometrical acoustics for physically-based acoustic rendering, the energy detected E_f is used to control the reflection magnitude using Poisson distributions across

the six frequency bands f. A time-varying Poisson distribution of Dirac-Delta pulses is generated, shown in Figure 5.10. The distribution models how, given a hypothetical environment with an emitter and detector, reflections would land at the detector over time: the transition from the direct signal and early reflections to late reflections is expressed as the density of detected reflections increased over time.

Following Schröder (2011)'s model, the probability of the occurrence of reflections $w_n(\Delta t)$ within a time interval Δt can be determined as:

$$w_n(\Delta t) = \frac{(\mu \Delta t)^n}{n!} e^{-\mu \Delta t} \quad n \in \mathbb{N}_0, \quad \mu > 0, \quad \Delta t \geq 0, \tag{5.6}$$

where μ is the mean reflection occurrence. The interval can be determined as a function of a random number z sampled from a uniform distribution:

$$\Delta t(z) = \frac{1}{\mu} ln(\frac{1}{z}) \quad 0 < z <= 1, \tag{5.7}$$

with μ and the starting time t_0 dependent on the size of the environment V, determined by the root node of the scene BVH:

$$\mu = \frac{4\pi c^3 t^2}{V}, \quad t_0 = \sqrt[3]{\frac{2V ln2}{4\pi c^3}}. \tag{5.8}$$

The use of Poisson distributions allows the generation of perceptually valid RIRs thanks to the layer of stochastic statistic modelling added to the process of constructing the response. This layer allows the Dirac-Delta series to be modelled with an approximated energy decay function, obtained via ray tracing. Without the distributions of Dirac-Delta pulses, more rays are needed to construct the final response.

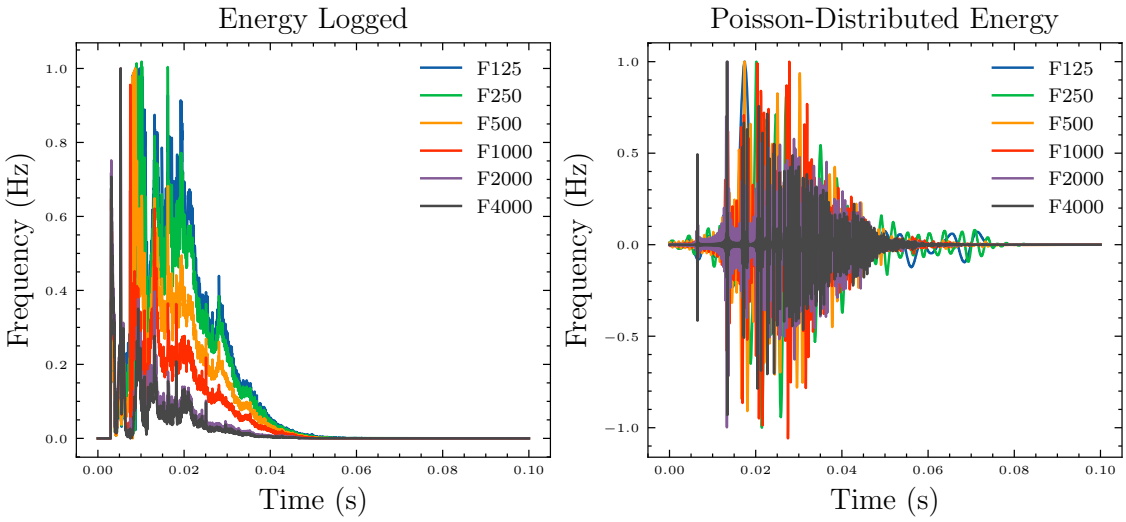


Figure 5.8: Visualisation of frequency-dependent RIRs, represented as energy logging resulting from generated propagation paths for a given source-receiver computation of Algorithm 1. Poisson distribution calculated based on the environment dimensions obtained from the generated BVH are then used to generate bipolar impulse responses.

5.2.4 Impulse Response Construction

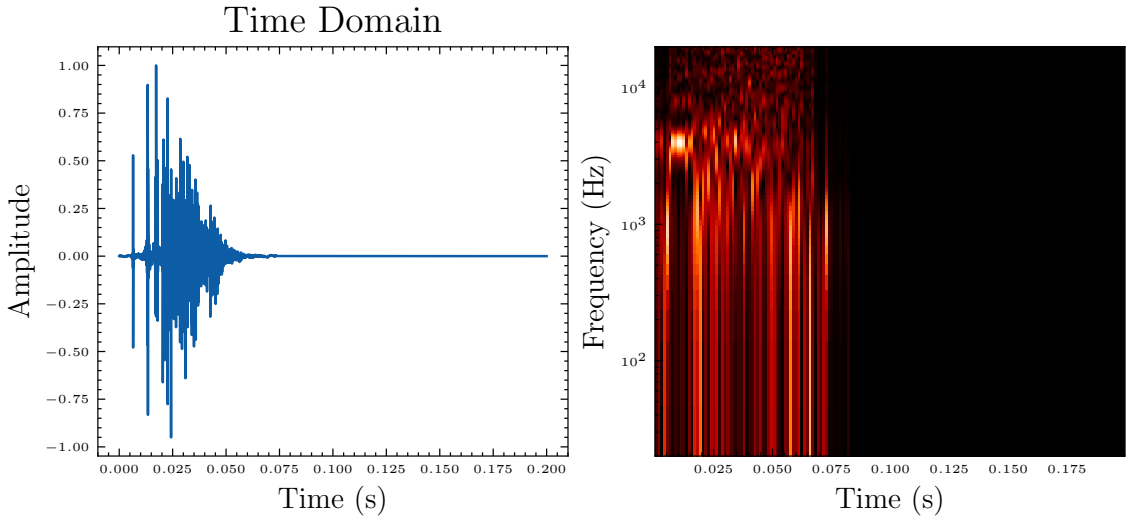


Figure 5.9: Time and frequency-domain representations left and right subfigures, respectively, of frequency-dependent RIRs combined into a broadband monoaural response.

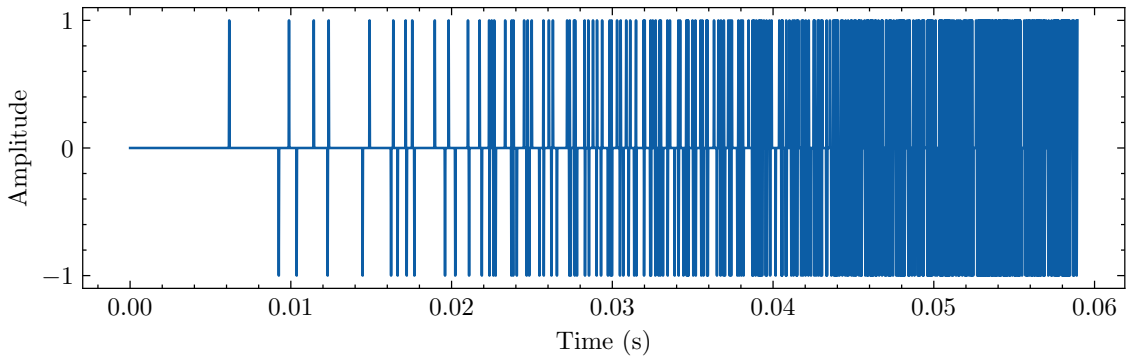


Figure 5.10: A Poisson Distributions of Dirac-Delta functions generated for an environment with an approximate volume of $60m^3$. The distribution represents, using unit pulses, the time-dependent likelihood of reflections registered at a hypothetical receiver from a hypothetical source within an environment of said volume. Considered a stochastic model, the distribution is used to generate a synthetic RIR from energy logged by a ray tracer.

Using Dirac-Delta pulses to represent reflections, six sequences are generated with Equation 5.6, obtaining distributions similar to Figure 5.10 for each frequency band f. The magnitude of these Poisson sequences $w_{n,f}$ follows the energy decay computed by the ray tracer:

$$h_f(t) = E_f(t)w_{n,f}(t), \tag{5.9}$$

determining frequency-dependent h functions that can construct a monoaural RIR by filtering each frequency component h_f and summing their aggregate to a broadband response, see Figure 5.9. A filter bank determines the contribution of each h_f towards its specific portion of the spectrum, dictated by the spectrum partitioning of the absorption coefficients used; it is composed of Finite Impulse Response (FIR) filters applied to the sequences. The filter bank is constructed using low-pass window-sinc filters for efficient computation

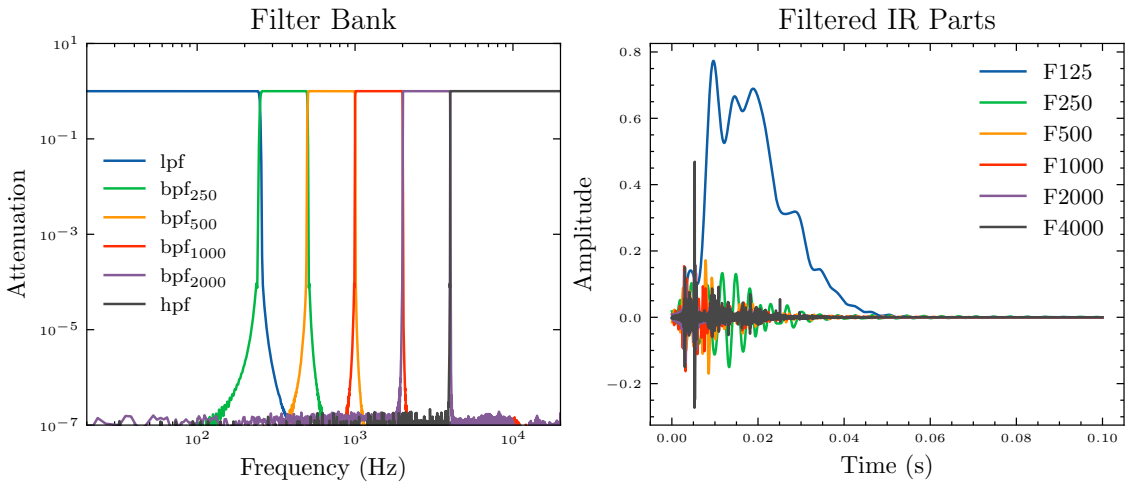


Figure 5.11: A filter bank partitioning the hearing range into octave bands associated with frequency-dependent energy logged by the geometrical acoustics renderer. The monoaural broadband IR visualised in Figure 5.8 is constructed by having each frequency band filtered and contributing to its specific region.

against attenuation factors. High and band-pass filters are obtained via spectral inversions of window-sinc low-pass filters. Using Smith et al. (1997)'s implementation, the h function is calculated for a low-pass lpf filter for a digital sequence of M points with sampling frequency F_s:

$$lpf(i) = \frac{\sin(2\pi f_c i)}{i\pi}, \tag{5.10}$$

where f_c indicates the cut-off frequency relative to F_s. Filters are computed over $M = 2^{14}$ points, obtaining over 120dB of stop-band attenuation, see Figure 5.11. A Blackman window wb is applied to the filter, reducing noise or artefacts caused by abrupt changes to the frequency response, given by:

$$wb(i) = 0.42 - 0.5\cos(\frac{2\pi i}{M}) + 0.08\cos(\frac{4\pi i}{M}). \tag{5.11}$$

5.3 Acoustic Simulation Evaluation

To consider the suitability of the geometrical acoustics-based acoustic rendering pipeline for application in the overarching immersive technology system, an evaluation should investigate the performance of the rendering pipeline in simulating the soundfield. The evaluation compares a real soundfield, measured using standard techniques for acoustic measurements, to a simulated soundfield generated from a virtual reconstruction of the physical, measured space. The objectives of this evaluation are:

- to investigate the suitability of geometrical acoustics for interactive, immersive applications that leverage virtual reconstructions of real space for auditory interactions;
- to measure the error of the simulated soundfield along standard acoustic metrics

against measurements of the physical soundfield;

- to discuss the adoption of geometrical acoustics for interactive platforms considering computational resources and material recognition integration.

5.3.1 Method

The experimental evaluation apparatus includes measurement equipment for capturing the acoustic space, space-capturing techniques to generate a virtual reconstruction of the room, and the geometrical acoustics-based rendering pipeline illustrated. Evaluating the effectiveness of acoustic rendering techniques involves comparing the simulated soundfield generated by the Ray Tracing-based sound rendering pipeline against the actual soundfield measured in the real environment. This comparison uses objective metrics to assess the accuracy of the simulation, including reverberation time, early decay time, or clarity.

The primary focus of this evaluation is to gain insights on an early ray tracer prototype designed for wearable computing devices. Wearable computing imposes significant constraints on computational resources, necessitating efficient algorithms. The prototype under investigation employs a simplified ray tracing approach with limited resolution to balance the need for real-time performance with the available processing power of wearable devices. The main goal of this evaluation is to gain insights into how closely a simplistic ray tracer can approximate the real-world soundfield, evaluating how the limited resolution impacts the accuracy of simulated reflections in a real-world environment. Effects relating to human hearing are dismissed by using omnidirectional emitters and receivers, delegating the investigation of human factors and listener-related effects to the subjective study illustrated in Chapter 7.

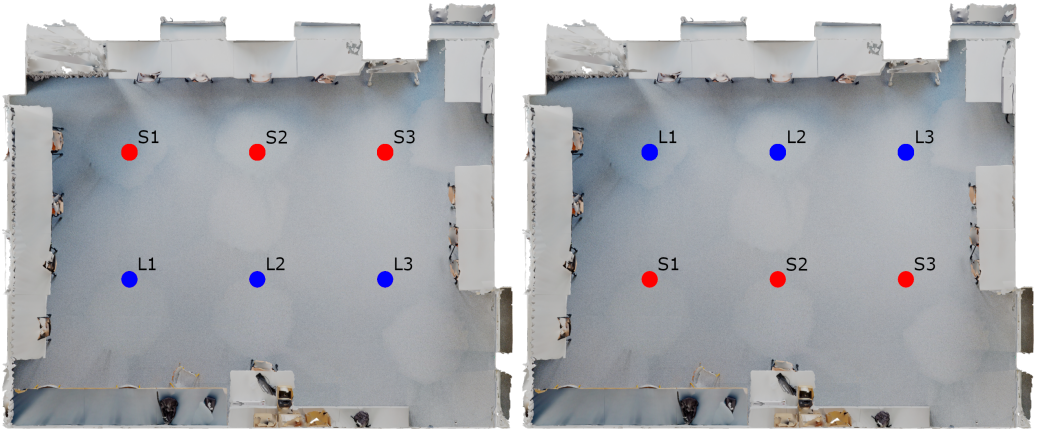
5.3.2 Apparatus and Test Procedure

The method compares a real soundfield and a simulated soundfield based on the same space: a lecture room within the City Centre campus of Birmingham City University. A real sound source and a microphone are used to capture the acoustic space at several probe points distributed uniformly across the space. The space is chosen as an exemplary environment, representative of indoor performance and offices spaces that are common amongst target use cases expressed in Chapter 1. Figure 5.12 shows top views of the space utilised for the experimental evaluation with overlaid recording probes as grids of source-receiver position pairs. Measurements are taken by permutating source positions and receiver positions **S1**, **S2**, **S3** and **L1**, **L2**, **L3**, respectively, in both grid formations illustrated in Figure 5.12a and 5.12b; e.g. measurements are taken between **S1** and **L1**, **S2** and **L1**, **S3** and **L1**, etc.

Source-listener position pairs are determined by uniformly distributing measurement positions to sample the entire space. According to the British Standard, position pairs should

Table 5.3: Distance (m) matrix source-listener position pairs illustrated in Figure 5.12.

	L1	**L2**	**L3**
S1	2.6	3.6	5.6
S2	3.6	2.6	3.6
S3	5.6	3.6	2.6



(a) Grid A (b) Grid B

Figure 5.12: Source-Listener location points in the recording environment. RIR are recorded between source positions **S1, S2, S3** and listener positions **L1, L2, L3**; these pairs are then inverted, obtaining uniformly spaces probe points across the soundfield.

be representative of the location of receivers and emitters in sound transmissions normally occurring within the space (3382-1, 2009). Furthermore, the standard recommends maintaining at least 1 m between the source and emitter. Table 5.3 shows the obtained measurement distances accounting for these factors and recommendations. Positions of source and emitter are permutated to represent proxemic theory distances in MR: near, medium, and far (Huang et al., 2022).

Real Soundfield Measurements

Following standard acoustic measurement procedures described by the British Standard (3382-1, 2009), sampling a space uniformly by recording measurement signals emitted by a speaker and recorded by a microphone. The test uses a dodecahedron speaker to fit the emitter directionality requirements established by the BS 3382-1 and to maximise the excitation of the soundfield, a Nor276 Dodecahedron Loudspeaker driven by a Nor282 Amplifier. A Focusrite Scarlett 18i8[2] microphone pre-amplifier and Analogue/Digital, Digital/Analogue Converter is used to emit a 30s long logarithmically swept sine signal x ranging from 20Hz to 20kHz. A Rode NT-SF1[3] flat-frequency response ($\pm 4dB$ between 20Hz to 20kHz) ambisonic microphone captures the swept sine emitted by the speaker via the Focusrite Scarlett 18i8.

[2] https://focusrite.com/products/scarlett-18i8
[3] https://rode.com/en/microphones/360-ambisonic/nt-sf1

The same procedure was repeated for all source-listener position pairs illustrated in Figure 5.12 across both grid formations: the speaker on the red dot and the microphone is placed on the blue dot, facing the speaker; the swept sine signal is recorded y, recovering the response h by convolving it to the time reversal sequence of the original sweep, $h(t) = y(t) * x(-t)$, as demonstrated by Farina (2007).

Simulated Soundfield Measurements

The lecture room was captured using a Matterport Pro3[4] space-scanning camera, generating a triangulated 3D mesh that the geometry handling system of proposed acoustic rendering pipeline uses to construct a scene BVH for geometrical operations of the ray tracer.

The acoustic rendering pipeline is deployed on the scene, defining two $0.5m$ radius spheres as source and receiver volumes and emitting $3 * 10^5$ rays, allowing 4 orders of reflections. RIRs are constructed in about $60s$ with an unoptimised and unparallelised implementation running on the Unity game engine scripting runtime. The same procedure described in the previous Section is repeated.

5.3.3 Results

Table 5.4: Standard acoustic metrics collected across source-receiver position pairs for Grid A and B, see Figure 5.12. The T_{30} reverberation time, C_{50} clarity index, and D_{50} definition index are dependent on the magnitude and distribution of acoustic energy expressed by the respective RIR. In this case, simulated and measured RIRs are compared.

Grid	Pos. Pair	T_{30} Sim.	T_{30} Real	C_{50} Sim.	C_{50} Real	D_{50} Sim.	D_{50} Real
A	**S1 L1**	0.12	0.43	4.33	3.20	0.27	0.32
	S1 L2	0.12	0.41	2.96	3.33	0.34	0.32
	S1 L3	0.12	0.41	2.74	2.45	0.35	0.36
	S2 L1	0.12	0.45	5.81	3.40	0.21	0.31
	S2 L2	0.12	0.43	2.02	3.16	0.39	0.33
	S2 L3	0.12	0.43	3.90	1.74	0.29	0.40
	S3 L1	0.11	0.42	1.46	2.45	0.42	0.36
	S3 L2	0.11	0.47	1.72	2.96	0.40	0.34
	S3 L3	0.12	0.45	1.92	2.71	0.39	0.35
B	**S1 L1**	0.12	0.44	3.74	2.43	0.30	0.37
	S1 L2	0.10	0.40	2.13	2.83	0.38	0.34
	S1 L3	0.11	0.41	2.80	2.45	0.34	0.36
	S2 L1	0.11	0.43	9.02	2.95	0.11	0.34
	S2 L2	0.11	0.49	7.96	3.93	0.14	0.29
	S2 L3	0.10	0.49	3.33	3.62	0.32	0.30
	S3 L1	0.11	0.45	6.24	3.57	0.19	0.31
	S3 L2	0.10	0.46	2.71	2.42	0.35	0.36
	S3 L3	0.10	0.39	2.76	2.88	0.35	0.34

[4] https://matterport.com/pro3

Table 5.5: Descriptive statistics computed on acoustic metrics gathered from source-listener position pairs.

Grid	Pair	T_{30} Err.	T_{30}	C_{50} Err.	C_{50}	D_{50} Err.	D_{50}
	S1 L1	0.31		-1.13		0.05	
	S1 L2	0.29	$\mu = 0.32$	0.37	$\mu = -0.16$	-0.02	$\mu = 0.00$
	S1 L3	0.29		-0.29		0.01	
	S2 L1	0.33		-2.41		0.10	
A	**S2 L2**	0.31	$\sigma = 0.02$	1.14	$\sigma = 1.34$	-0.06	$\sigma = 0.06$
	S2 L3	0.31		-2.16		0.11	
	S3 L1	0.31		0.99		-0.06	
	S3 L2	0.36	$MSE = 0.10$	1.24	$MSE = 1.82$	-0.06	$MSE = 0.00$
	S3 L3	0.33		0.79		-0.04	
	S1 L1	0.32		-1.31		0.07	
	S1 L2	0.30	$\mu = 0.33$	0.70	$\mu = -1.51$	-0.04	$\mu = 0.06$
	S1 L3	0.30		-0.35		0.02	
	S2 L1	0.32		-6.07		0.23	
B	**S2 L2**	0.38	$\sigma = 0.04$	-4.03	$\sigma = 2.29$	0.15	$\sigma = 0.09$
	S2 L3	0.39		0.29		-0.02	
	S3 L1	0.34		-2.67		0.12	
	S3 L2	0.36	$MSE = 0.11$	-0.29	$MSE = 6.97$	0.01	$MSE = 0.01$
	S3 L3	0.29		0.12		-0.01	

Table 5.6: G Sound strength values (dB) across the two source-listener position pair grids, see Figure 5.12. For each grid, two receiver positions are used to measure the sound strength by computing the ratio of instantaneous energy between a near and a far source. Hence, for **L1** the ratio of energy between **S1** and **S3** is measured (near and far, respectively), and for **L3**, is measured between **S3** and **S1**.

Receiver Position	Grid A		Grid B	
	Simulated	Real	Simulated	Real
L1	-1.156	0.505	1.607	-2.89
L3	-2.169	-2.203	2.022	-3.39

RIRs obtained from both real and simulated soundfields are reported in Figures 5.15, 5.16, 5.17, 5.18; these plots show Simulated and Measured RIRs in the first and second columns, respectively, reporting comparisons of decay curves obtained from smoothed analytic signals from the absolute value of a Hilbert transform of the respective RIR. The Measured RIRs have around $45dB$ of signal-to-noise ratio, and the Simulated ones are over $120dB$.

From these responses, metrics, shown in Table 5.4, describing the acoustic space are computed, including the T_{30} reverberation, C_{50} clarity and D_{50} definition indexes that relate to factors of sound transmissions occurring within the space, affecting tasks such as speech intelligibility or perception of spatial features of sound sources via auditory information. Changes in clarity and definition parameters are observed as a function of distance from the source across all listeners; see Table 5.3 for distance values.

The T_{30} reverberation metric is obtained from the computed analytic signal, and the D_{50} and C_{50} are computed from the responses, following the BS 3382-1 (2009) definitions,

$$D_{50} = \frac{\int_0^{0.05} p^2(t)\mathrm{d}t}{\int_0^{\infty} p^2(t)\mathrm{d}t} \quad \text{and} \quad C_{50} = 10\log_{10}\left(\frac{D_{50}}{1-D_{50}}\right), \tag{5.12}$$

where p is the magnitude of sound events expressed by the response. Additionally, the sound strength metric G between a *near* and *far* sound source is defined as:

$$G = 10log_{10}\frac{\int_0^{\infty} p^2_{near}(t)\mathrm{d}t}{\int_0^{\infty} p^2_{far}(t)\mathrm{d}t}; \tag{5.13}$$

G scores are reported in Table 5.6.

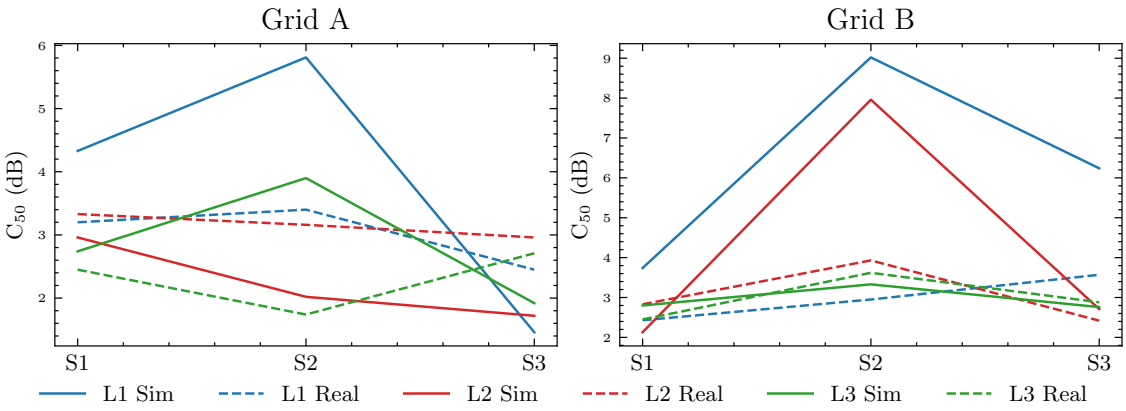


Figure 5.13: C_{50} index values for both the simulated and measured soundfield, observed over distance from the sound source.

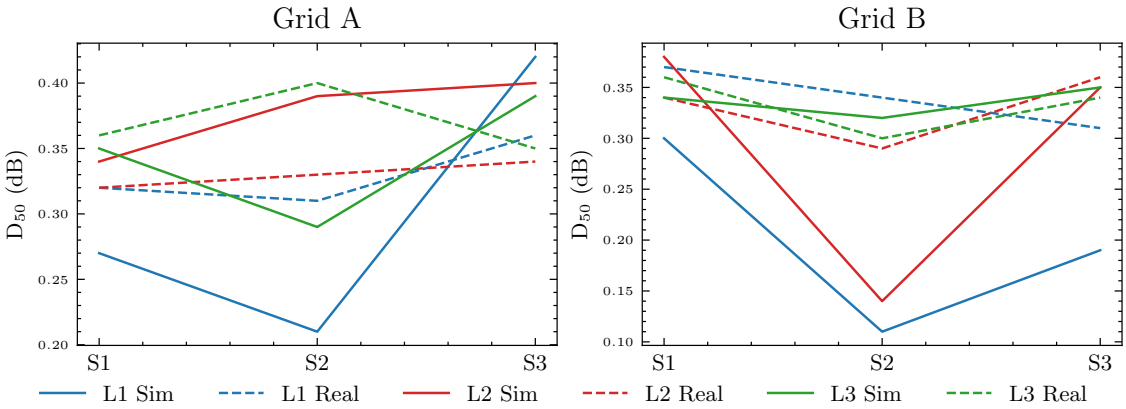


Figure 5.14: D_{50} index values for both the simulated and measured soundfield over distance from the sound source.

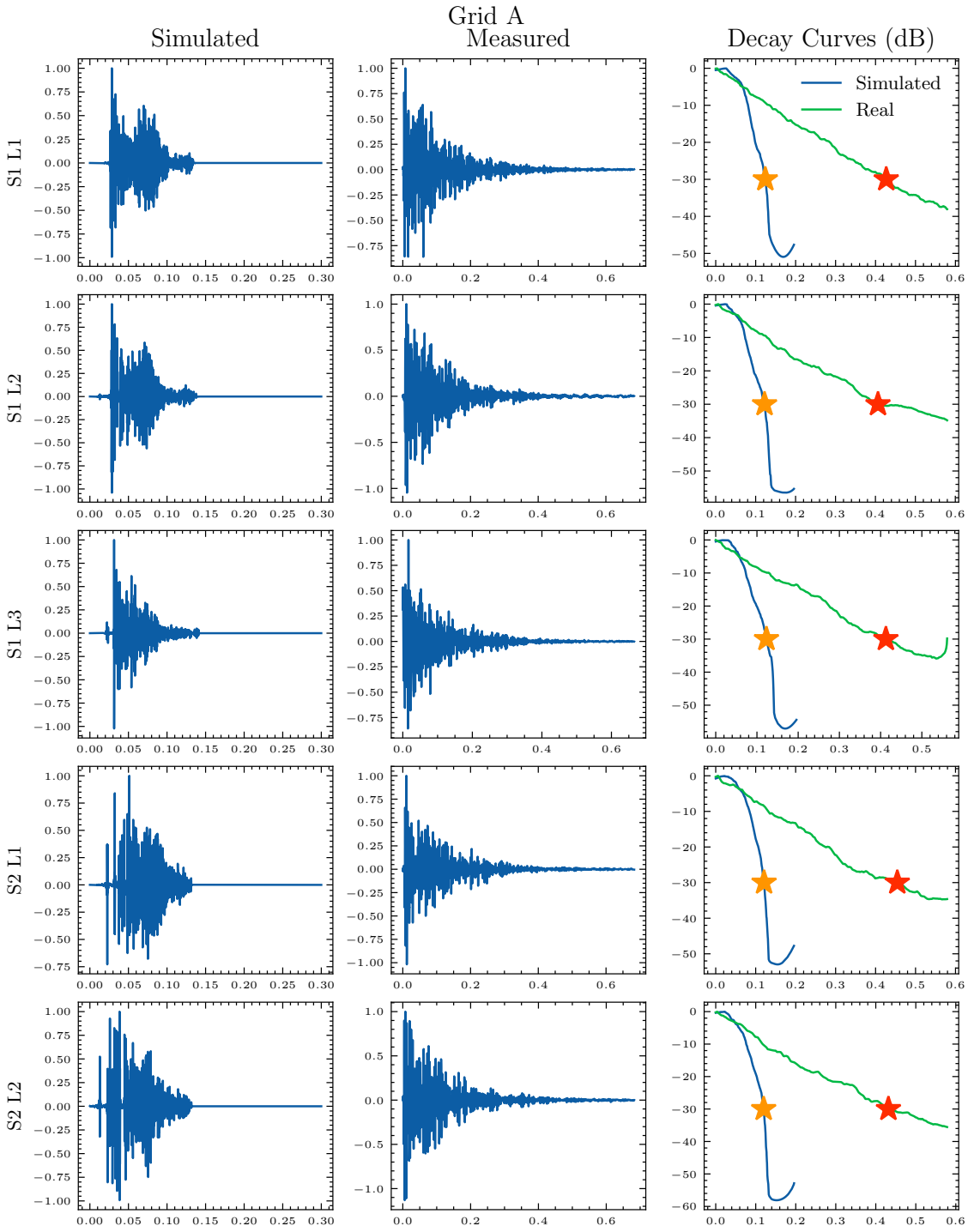


Figure 5.15: Time-domain representation of responses, as magnitude over time (s), obtained from simulations of the soundfield of a virtual reconstruction of a lecture room, first column, across permutations of source-listener position pairs illustrated in Figure 5.12a, rows. These are compared to responses obtained from acoustic measurements of the physical space, second column. The third column shows comparisons of the acoustic energy, as attenuation (dB) over time (s), decay across the two responses with overlaid light, yellow star and dark, red star indicating T_{30} values for simulated and measured responses, respectively.

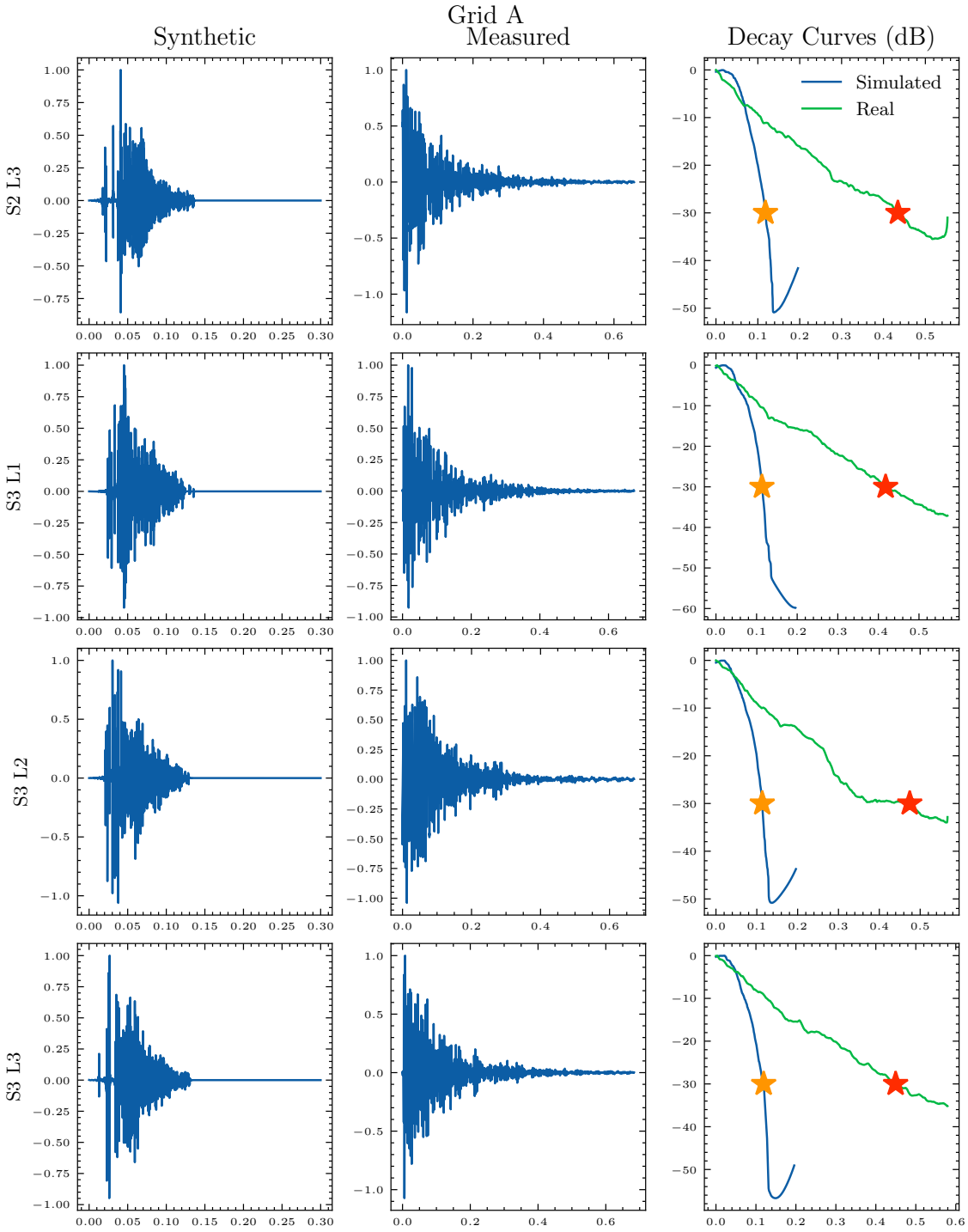


Figure 5.16: Continuation of rows from Figure 5.15, showing remaining source-listener position pairs.

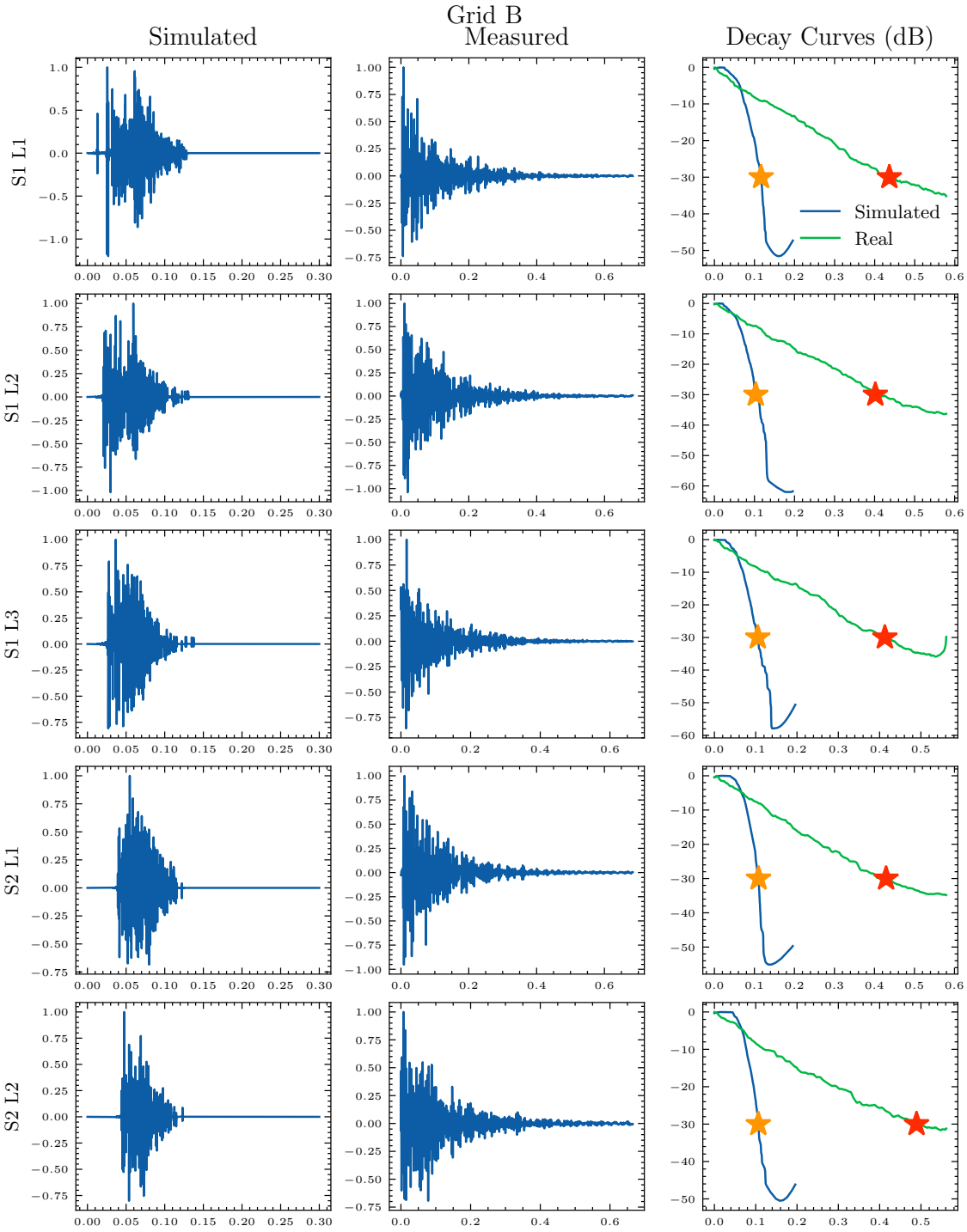


Figure 5.17: Continuation of rows from Figure 5.15, showing responses from position pairs illustrated in Figure 5.12b.

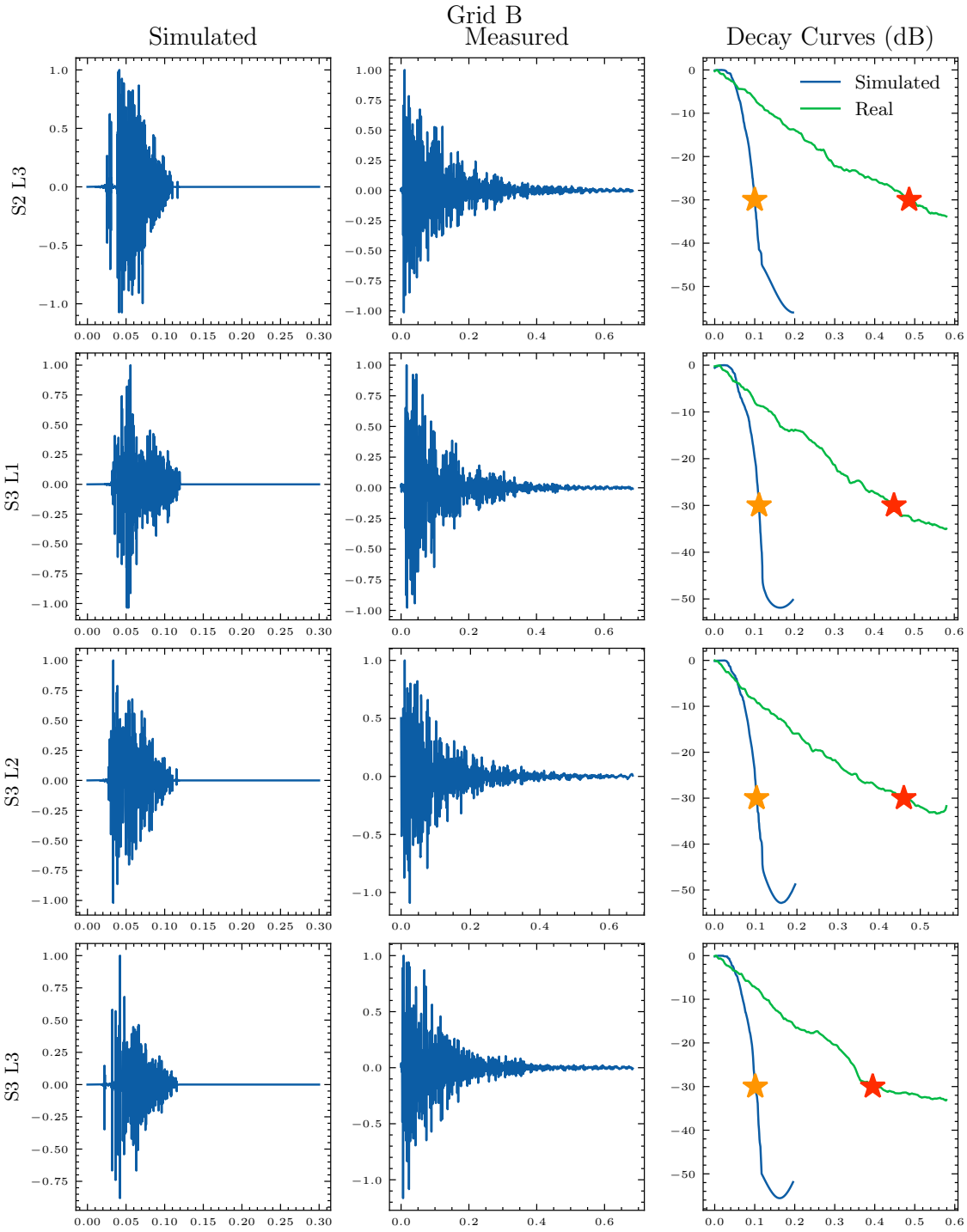


Figure 5.18: Continuation of rows from Figure 5.17, showing remaining source-listener position pairs.

5.3.4 Discussion

Reverberation is the parameter relating to the acoustic soundfield where the largest discrepancy between the simulated and real responses is observed, as shown in Tables 5.4 and 5.5. By observing decay lines fit over the responses shown in Figures 5.15 to 5.18, the acoustic rendering pipeline has a sharper energy decay, with limited capabilities in reproducing late energy transfers from emitter to receiver. The limited resolution of low-frequency energy in the simulated soundscape contributes to the reverberation discrepancy from the ground truth caused by incorrect modelling of directionality profiles of radiating waves. Geometrical acoustics pipelines can be integrated with modern acoustic radiance modelling techniques, as demonstrated by Siltanen, Lokki and Savioja (2010). Despite the measured reverberation time error in the simulated soundfield, the improved pipeline demonstrates an extensive improvement from the first prototype, as shown in Table 5.1. In addition, the limitation of the ISM model in adapting to architectural features of the scene geometry.

The measured T_{30} error is analysed, using reverberation as a determinant of perceptual aspects of sound transmissions in a soundfield, the proposed geometrical acoustics-based rendering pipeline aligns with state-of-the-art and modern realistic methods for acoustic rendering. As discussed in Section 3.2, modern rendering pipelines adopt learned Generative Adversarial Neural Networks to determine reverberation features of multi-modal representations of complex scenes. The proposed acoustic rendering pipeline estimates reverberation with similar accuracy compared to Singh et al. (2021)'s work. Such error is acceptable even when considering advanced realistic multi-modal pipelines such as Schissler, Mehra and Manocha (2014)'s work.

5.4 Conclusions

The evaluation and testing of the ISM prototype and ray tracer implementation have provided critical insights into the integration of material recognition systems and the development of streamlined acoustic rendering pipelines for interactive applications. The ISM prototype was instrumental in testing the integration of material recognition systems within the acoustic rendering pipeline. It, however, suffered from limitations around handling complex environment geometry architectures and only considering specular reflections. The ray tracer implementation provided valuable insights into how an acoustic rendering pipeline can be optimised for interactive applications, particularly in the context of wearable computing by integrating stochastic models to reduce the number of rays and computations needed. The simplified ray tracing approach, despite its limited resolution, showed the feasibility of achieving reasonable accuracy against ground truth soundfields. The work in this Chapter highlighted the potential of using efficient algorithms to balance performance and accuracy, ensuring that users receive timely auditory feedback without

significant computational delays. Chapter 6 will provide discussions around real-time implementation of the sound rendering pipelines illustrated, introducing human factors in audio reproduction chains. In conclusion, the integration of material recognition systems and the insights gained from the ray tracer implementation represent significant progress in the field of acoustic rendering. While challenges remain, the ability to approximate basic acoustic phenomena and optimise performance for interactive applications marks a step forward.

Chapter 6

Towards Scene-Aware Acoustic Rendering Pipelines for Augmented Audio Reality

This work was published in:

Colombo, M., 2021, October. Vision-based acoustic information retrieval for interactive sound rendering. In 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (pp. 487–490). IEEE.

The following Chapter details the implementation to address the primary objective of the thesis: proposing a scene-aware interactive rendering system for realistic sound transmissions between sound sources and a listener in MR space. The following Sections will detail how the acoustic material recognition and rendering prototypes integrate with the overarching pipeline, discussing target functionalities, design choices, limitations, and future expansions based on evaluation data obtained from evaluations provided in earlier Chapters and recent advances in the field. This Chapter breaks down into the following three areas:

- **concept**,
- **implementation**, and
- **vision**.

Concept defines the proposed pipeline's architecture, objectives, and design principles; *implementation* demonstrates how the pipeline can deploy to consumer HMDs as a proof-of-concept system; finally, *vision* analyses limiting factors and expansion points, looking at perception factors to consider for optimisation and ablation studies.

Discussions on design principles, implementation, and limitations feed into methodologies for future research direction, as demonstrated by Chapter 7, where the evaluation design has foundations on the outcomes of these Sections.

6.1 Concept

The proposed system aims to produce realistic sound transmissions between sound-emitting holograms in MR space and a listener experiencing the holograms via a head-mounted holographic display and perceived auditory stimuli via headphones, see Figure 1.1. In this Chapter, the term "listener" is used interchangeably with the term "user" to refer to the user who is displayed audio-visual stimuli from both the holographic HMD and the headphones. The implementation of this method for interactive sound rendering is demonstrated on a Microsoft Hololens 2 Augmented Reality Head-Mounted Headset, an embedded wearable computer with a holographic display featuring space-sensing and spatial mapping technology that enables Simultaneous Localisation and Mapping (SLAM) methods (Davison, 2003; Ungureanu et al., 2020), see Figure 1.1.

6.1.1 Dynamic Environment Reconstruction

The first link in the chain of generating realistic sound transmission between a listener and holograms in MR space is generating and handling a virtual environment to host scene elements such as virtual sound-emitting entities. In MR space, the scene geometry includes a reconstruction of the real space surrounding the user. The space surrounding the user is reconstructed using space-sensing technology featured in the Hololens 2 HMD (Ungureanu et al., 2020). Using the Microsoft Mixed Reality Toolkit [1], an environment mesh is continuously extracted as a triangulated mesh and updated every 3 to 5 seconds, reflecting dynamic changes in the environment (e.g. furniture being moved or a crowd entering the space).

The reconstructed environment mesh optimises as the user walks around the scene and can produce reconstruction errors or holes in the mesh. This technology produces reconstructions with less accuracy and precision than LiDAR scanners with dedicated hardware, such as the one used for conducting experiments in Chapter 4, but it is still capable of producing watertight meshes, suitable for rendering applications[2].

The triangulated mesh, represented as a list of triangles and indices, is used to construct a Bounding Volume Hierarchy (BVH), where triangle primitives are encapsulated in Axis-Aligned Bounding Boxs (AABBs), optimising geometrical search operations for acoustic modelling such as ray-triangle intersection tests.

[1] https://github.com/microsoft/MixedRealityToolkit-Unity
[2] https://learn.microsoft.com/en-us/windows/mixed-reality/design/spatial-mapping

6.1.2 Environment Completion for Rendering Tasks

MR HMDs require techniques such as Simultaneous Localisation and Mapping (SLAM) to accurately reconstruct, map, and understand the environment surrounding the user as he engages with the immersive scene. SLAM is essential for enabling MR systems to understand and interact with the physical world in real-time. However, this process poses significant computational challenges, which can create obstacles for MR applications, particularly when it comes to rendering realistic audio environments. The computational load required to continuously map and update a complex environment limits the pool of techniques to a subset that is feasible to portable computing (Saeedi et al., 2018; Palazzolo et al., 2019), leading to delays and inaccuracies in the representation of the physical space.

HMDs like the HoloLens 2 tackle this challenge by reconstructing the environment as the user navigates through the scene. As the user moves around, the device gradually builds a mesh representation of the environment, which is then used for simulating soundfields or other rendering tasks. However, until the user has fully explored the entire scene, the environment mesh may remain incomplete. This incompleteness can have significant implications for sound rendering, as the accuracy of simulated soundfields heavily depends on the fidelity of the environmental geometry. Incomplete meshes can lead to errors in the simulation of reverberation and other sound phenomena that rely on accurate geometry, such as occlusion and reflection.

6.1.3 Material Tagging from Partitioned Space

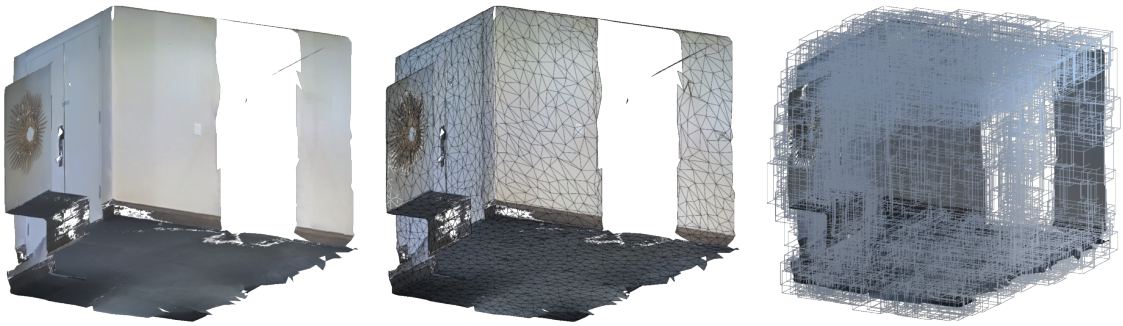


Figure 6.1: A visualisation of a BVH constructed on a portion of a scanned environment, left image, to demonstrate how AABBs encapsulate triangle primitives, visible in the centre image, forming a tree of boxes to partition the geometry.

One key advantage of the employment of a BVH for the handling of scene geometry is the space partitioning that can be used for the material tagging process. As discussed in Chapter 4, CNNs offer efficient approaches to mapping the appearance of surfaces to the acoustic characteristics of their materials, like absorption or scattering coefficients. The space partitioning provided by the BVHs here is used to capture the appearance of portions of the reconstructed environment in order to determine the material characteristics of its surfaces, see Figure 6.1.

The employment of a BVHs additionally allows the generation of ground truth data for acoustic material mapping processes: with access to the physical space, the acoustic material classification pipeline can test the accuracy of material tagging against labelled mesh segments. This would allow the pipeline to better target specific ecosystems of materials by creating sets of training data by labelling representative environments and generating image patches from the reconstructed space partitioned by the BVH.

6.1.4 Overview

As discussed in Chapter 3, the current state of research towards realistic auditory display in virtual environments leverages deep learning-based methods, using NNs to create latent representations of soundfields (Liu and Manocha, 2022; Ratnarajah et al., 2022; Chen, Su and Shlizerman, 2023). These methods often encode visual and geometrical features of the environment, as well as spatial features relating to the source and listener. These avenues are, however, experimental and require intense use of GPUs or hardware, making it hard to test on wearable computing platforms. Classic geometrical acoustics-based pipelines continue to receive attention from both research and professional domains, making ray tracing or hybrid techniques faster and more efficient by exploiting parallelisation paradigm in modern compute hardware (Cowan and Kapralos, 2010; Diaz-Guerra, Miguel and Beltran, 2021).

The geometry handling system provides the foundations for supporting real-time auralisations in MR, as it provides a solution to the problem of defining acoustic geometry: auralisations can be produced from simplistic cuboid volume to realistic reconstructions of space with orders of fidelity to the physical counterparts of millimetres. Computational costs associated with generating auralisations rise with the number of geometrical primitives used in the reconstructed space: in triangulated meshes, detailing smaller structures in a 3D model increases the number of triangles. Detailed scene geometry allows approximation of acoustic phenomena and generally contributes to high-resolution soundfield modelling, though the computational requirements make it unfeasible for interactive applications. As discussed in Chapter 3 and 5, studies outline a perceptual threshold and benchmark to define the minimum level of detail required from reconstructions, and recent research on acoustic simulations conducted over the last decades has expanded towards defining what is required from a virtual environment to produce a believable simulation. Combined with advances in real-space scanning technology and user-friendly 3D reconstruction software, it is now possible to create perceptually valid virtual environments for acoustic simulations without requiring expert computer graphics engineering knowledge.

6.1.5 Selecting Acoustic Rendering Techniques

A central problem inherent in the task of interactive soundfield approximations for generating dynamic auralisations is the task of adopting a method for propagating anechoic

audio in a dynamic scene. As discussed in Chapter 3 there are families of sound rendering techniques that produce auralisations with varying levels of realism evoked by audio propagated in the simulated soundscape and associated computational costs, as well as limitations due to their inherent architecture or nature.

Ray-based techniques offer efficiency as they are computationally less demanding than wave-based techniques, making them suitable for real-time applications or large-scale spaces often found in cultural heritage contexts. The geometrical nature of ray-based methods allows for easier integration with existing architectural models or historical reconstructions, and the method's inherent flexibility makes it easily adjustable to different acoustic scenarios. On the other hand, wave-ray hybrid techniques present a more complex picture, combining aspects of ray-based and wave-based methods. Rays are utilised to model the high-frequency components of the sound, while wave equations handle the low-frequency behaviour, attempting to capture the best attributes of both methods. However, the hybrid nature often means more computational resources are needed, and it might not always be the most suitable choice for cultural heritage applications where both high and low-frequency accuracy is not often the primary concern. Wave-based techniques stand out for their precision, solving the wave equation to simulate how sound waves propagate through space, accurately modelling diffraction, scattering, and other complex wave phenomena. While highly accurate, wave-based techniques often require substantial computational resources, making them less suited for real-time or large-scale applications. Moreover, this level of detail may exceed what is necessary for conveying the historical or cultural experience. Considering the landscape of immersive acoustics, ray-based techniques offer a compelling option for real-time applications and limited computational resources. Their computational efficiency, relative simplicity, and adaptability in handling various scenarios effectively capture the essential acoustic characteristics of historical spaces. Unlike wave-based or hybrid methods, ray-based techniques can prioritise the aspects most relevant to the experience and understanding of cultural heritage, aligning well with the objectives and constraints often found in this field. Therefore, while the high accuracy of wave-based methods or the comprehensive nature of hybrid methods may have specific applications in areas of cultural heritage investigations, it is the ray-based techniques (such as those employed by acoustic simulation software such as ODEON) that generally stand out as the most appropriate choice for the unique challenges and opportunities presented within the context of MR.

6.1.6 Real-Time Spatial Audio Rendering Techniques

The proposed pipeline approximates acoustic phenomena from a dynamic environment reconstructed as the user interacts and navigates the complex scene, considering physical and objective factors of the soundscape. The HAS, however, is the final and crucial link in the chain of the reproduction system, requiring the pipeline to consider psychoacoustic and human factors in generating auditory stimuli expressed as binaural audio. The integration

of factors of the HAS and the convolution of anechoic audio completes the 3D spatial audio chain.

Head-Related Impulse Responses (HRIRs) offer a compact solution to the integration of HAS features, much like room responses, as we can convolve a monaural signal to the frequency representation of a pair of HRIRs (one for each ear), the Head-Related Transfer Function (HRTF). Hence, the discussed real-time convolution algorithms (Kim et al., 2019) can be used, in combination with the h RIR signal, to generate binaural rendered audio channels y_{left} and y_{right} from a monaural signal x:

$$y_{left} = x * HRTF_{left} * h \quad \text{and} \quad y_{right} = x * HRTF_{right} * h. \tag{6.1}$$

HRTF channels left and right are evaluated in real-time, depending on the current position and rotation of the listener entity in the virtual environment, using a loaded bank or HRIRs that are often real measurements of subjects across a grid of points in the azimuth and elevation surrounding the subject. These measurements record the response of each ear canal from a source position in the surrounding sphere, allowing an interactive HRIR algorithm to interpolate between these measurements (Hoene, Patino Mejia and Cacerovschi, 2017).

Audio rendering tasks, such as performing the convolution operation indicated by Equation 6.1, generally require trivial computational resources, unlike sound propagation or acoustic modelling (Liu and Manocha, 2020).

6.2 Implementation of an AAR Prototype System

The following Section proposes a software implementation of the system, proposing a technical apparatus that can be employed for dynamic auralisation of sound-emitting holograms, considering approximated acoustic phenomena of MR space.

6.2.1 Overview

The system, illustrated in Figure 6.2, has a game engine as its core component, responsible for querying the HMD for updates on the spare surrounding the user, indexing and handling reconstructed scene geometry, computing energy transfers between emitters and receivers, and controlling a DSP engine for performing audio rendering tasks.

6.2.2 Hardware Apparatus

This Section demonstrates the system implementation using a Hololens 2[3] HMD, composed of a built-in ARM32/64 computer equipped powered by a Qualcomm Snapdragon 850 CPU with 4GB LPDDR4 DRAM.

As shown in Figure 6.4, the HMD has a visor containing holographic visual displays and

[3] https://learn.microsoft.com/en-us/style-guide/developer-content/reference-documentation

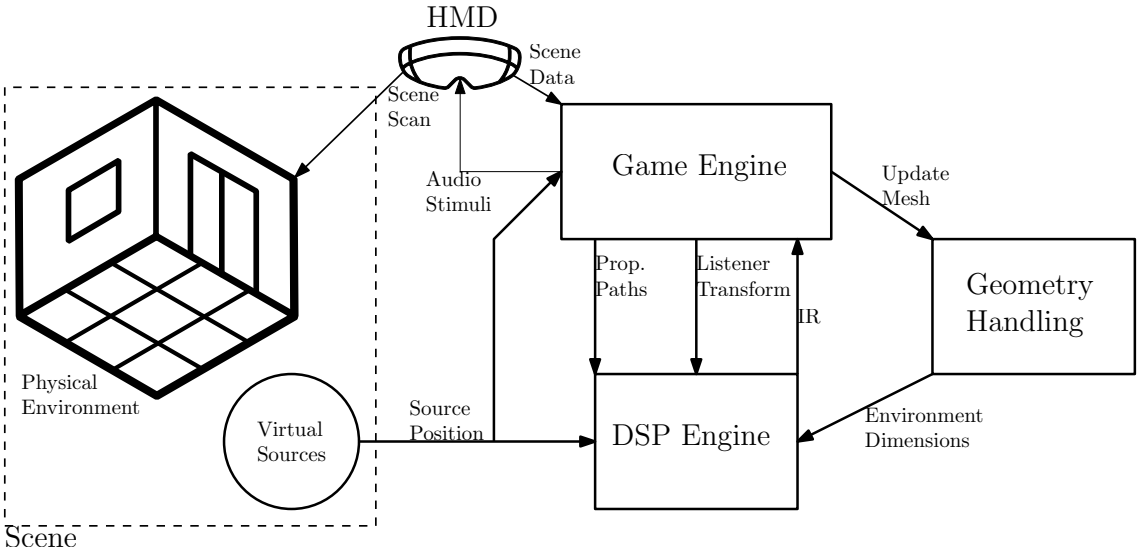


Figure 6.2: An overview of the technical implementation of the proposed pipeline for spatial audio rendering applied to an Augmented Reality interactive application.

featuring head-tracking sensors with four light cameras, two eye-tracking infrared cameras, 1 Megapixel depth sensor. The visual displays have 1.08mm focal length and 96.1° field of view. Additionally, an accelerometer, a gyroscope, and a magnetometer allow tracking of the user's position and orientation. The device supports spatial audio reproduction with built-in speakers and a Bluetooth-connected external audio DAC. The array of sensors enables several human understanding features, such as hand and eye tracking and six-degrees-of-freedom position tracking.

The Hololens 2 represents a platform offering average to low computational resources when compared against the current market of XR HMDs. The Meta Quest 2 HMD[4], for instance, offers a faster Qualcomm Snapdragon XR2 CPU with 6GB of LPDDR DRAM. The overall rapid increase in compute power and memory amongst XR technology sets up the proposed rendering pipeline for success.

6.2.3 Game Engine and Scene Management System

The game engine has a central role in managing operations and the lifecycle of the audio processing pipeline, as well as tracking the listener's position and orientation and managing sound sources in the complex scene. The game engine allocates computational resources to sound rendering-related operations and polls updates on the spatial reconstruction system of the HMD. Hence, the design for an acoustic rendering pipeline for applications should centre around the core principles of game engine architectures.

Upon receiving updates from the spatial reconstruction system, whose frequency depends on the parameters of the MRTK spatial understanding setup, the game engine is responsible for updating the BVH to take into account changes to the environment geometry. Here,

[4] https://www.meta.com/gb/quest/products/quest-2/

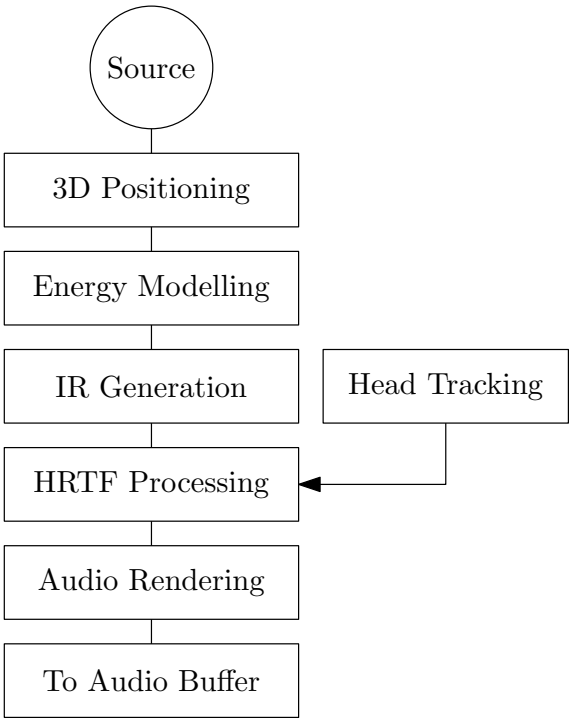


Figure 6.3: Rendering process of a single sound source: an input source description, positioned in AR space, is used to model its energy transfer to the listener, represented via a response function. Via head-related transfer functions based on the listener's head rotation provided by the HMD, audio from the audio is processed and sent to the output audio buffer.

the modularity of modern geometry handling systems, such as Kopta et al. (2012)'s or Lauterbach et al. (2009)'s fast and efficient implementations; designed with modularity and parallelisation in mind, allowing real-time updates and reconstruction of parts of the scene tree, making it a suitable choice for sound rendering geometry handling in MR.

The game engine, furthermore, is responsible for the handling and positioning of sound sources, starting the source-to-listener signal flow expressed in Figure 6.3. This includes managing anechoic audio associated with the sound source or the sound synthesis system that generates audio on the fly (Hamid and Kapralos, 2009).

6.2.4 Engine for Spatial Audio Rendering

Drawing from well-established and modern spatialiser system designs from Naef, Staadt and Gross (2002) and Lakka et al. (2021), respectively, a DSP engine receives control from the game engine, transforming the energy propagation modelling into binaural IRs that can propagate anechoic audio. Duties of the DSP engine include manipulating signals in the time and frequency domain, requiring a fast implementation of the DTFT, and a suite of tools and routines for handling and processing IRs and transfer functions.

As shown in Figure 6.3, input audio from sources undergo the time and frequency-domain manipulations based on aspects of the environment and the listener implemented by the DSP engine. Table 6.1 illustrates the interface to the DSP Engine, showing procedures

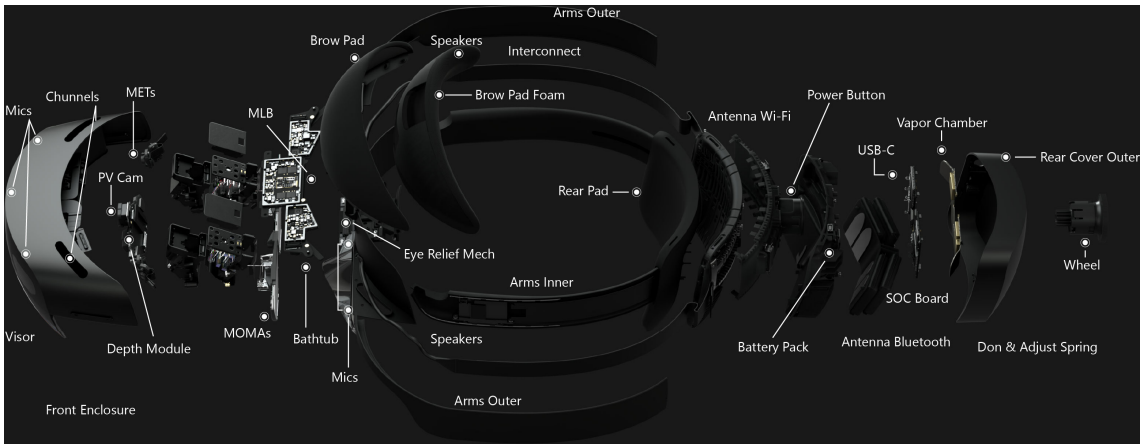


Figure 6.4: Exploded view of the Microsoft Hololens 2 Augmented Reality Head-Mounted-Display.

Table 6.1: A list of operations and routines provided by the proposed DSP engine to process audio in real-time from sources depending on acoustic energy transfers and listener's orientation information.

Procedure	Inputs	Outputs	Phase
Initialise Engine	Samplerate, filter kernel size, frequency bands definition	None	Setup
Setup Block Processing	Chunk size, kernel size, channels	None	Setup
Load HRTF Banks	HRTF database	None	Setup
Set Energy Transfer	Target frequency band, data	None	Runtime
Render IR	None	None	Runtime
Process Buffer Chunk	Chunk	Chunk	Runtime
Update HRTF	Head rotation	None	Runtime

and setup operations to support real-time audio manipulation of sources. The setup phase determines the sample rate for all audio manipulations, the number of points for evaluating kernel functions for filters, and the definition of frequency bands for frequency-dependent energy transfer modelling and processing. In addition, other setup procedures include loading HRTF banks from a participant record database and configuring block processing routines by specifying the current engine's audio buffer properties, such as the size of audio chunks and number of channels (Hoene, Patino Mejia and Cacerovschi, 2017).

Runtime procedures, executed as the game engine runs the scene and a sound source is emitting with the listener being within the propagation radius, include setting frequency-dependent energy transfer between source and receiver, rendering a monoaural IR, and processing audio emitted based on the listener's head rotation; these operations complete the chain in Figure 6.3.

6.2.5 DSP Engine Procedures

A Digital Signal Processing (DSP) engine is crucial for generating spatialised audio, enabling realistic and immersive auditory experiences in immersive applications. The DSP engine processes audio signals to simulate how sound propagates in a 3D environment, taking into account the positions of sound sources and listeners, as well as the acoustic characteristics of the environment. The initialisation phase is critical for setting up the DSP engine with the necessary parameters and resources to perform spatial audio processing effectively. The following procedures are executed during this phase.

Intialisation The DSP engine requires several key parameters to be configured, like sample rate, to determine the number of audio samples processed per second; or the buffer size, to define the number of audio samples processed in each block. Smaller block sizes can reduce latency but may increase CPU load, while larger block sizes can reduce CPU load at the cost of increased latency.

Loading HRTF Banks HRTF banks are crucial for spatial audio rendering as they model how sound waves interact with the human head, ears, and torso. During initialisation, the DSP engine loads the necessary HRTF banks from external files or databases. These banks provide the filters required to spatialise audio based on the listener's position relative to the sound source.

Processing Audio Buffer At runtime, the DSP engine processes audio chunks in real-time. This involves:

- Applying HRTFs: The engine applies the selected HRTF to the audio signals, simulating how the sound would be perceived from the listener's position;
- Convolving IRs: The audio signals are convolved with the pre-rendered IRs to simulate the acoustic characteristics of the environment. This step adds realism by

incorporating reflections and reverberation effects.

- The processed audio chunks are sent to the output device (e.g., headphones or speakers) via the game engine, providing the listener with a spatialised audio experience.

Table 6.1 summarises routines that compose the processing and audio signal processing of the proposed audio pipeline.

6.2.6 Interactive Acoustic Rendering

At runtime, the game engine manages operations of simulating sound propagation from a source to a listener, generating an Room Impulse Response (RIR) as a result of the operation. The implementation for this proposed pipeline leverages the work towards a ray tracing-based propagation model discussed in Section 5.2, underpinned by Schröder (2011)'s work in physically-based acoustic rendering. In summary, the propagation model approximates the energy transfer from an emitter to a receiver volume, using frequency-dependent acoustic to mitigate the low-frequency limitations of geometrical acoustics modelling techniques. Given a source-receiver pair, the game engine can generate propagation a set of propagation paths for each defined frequency band, which are defined based on frequency-dependent material properties assigned to the scene geometry, which the DSP engine can then use to generate a monoraural response for real-time auralisations.

The IR generation component of the acoustic rendering pipeline is designed with a modular architecture, allowing for significant flexibility and adaptability. This modularity ensures that the system can evolve to incorporate advanced techniques as they become available, enhancing the accuracy and realism of spatial audio rendering. The modular design of the IR generation component means that it can accommodate various underlying techniques without requiring rebuilding of the entire pipeline. This flexibility is crucial for future-proofing the system and ensuring that it can integrate cutting-edge technologies as they emerge. The modular framework supports the integration of different propagation techniques, ranging from traditional geometrical acoustics approaches to deep learning methods.

Methods like Singh et al. (2021)'s Image2Reverb can approximate reverb based on visual representations of the environment, leveraging neural networks to predict acoustic properties from images but do not consider source-listener position information within the scene and may introduce inaccuracies to the generated auralisations. Ray tracing, despite the low-frequency limitations, can still approximate basic occlusion phenomena (Savioja and Svensson, 2015a). Though, on the other hand, upcoming NAF-based approaches will be able to consider source-listener effects and generate responses at interactive rates (Luo et al., 2022). These models can potentially adapt to dynamic changes in the environment and provide real-time soundfield simulations, ensuring that the spatial audio remains accurate as the scene evolves. Additionally, neural networks can potentially overcome the limitations of ray tracing at low frequencies by learning to model diffraction and scattering

effects directly from data.

6.2.7 Audio Rendering

Approximated acoustic phenomena expressed in IRs are often applied to anechoic, unpropagated audio by application of the convolution operation, as discussed in Section 2.3.2. The resulting auralisation targets a sound reproduction system by applying responses to model characteristics of a human listener in the sound field.

Audio rendering describes operations related to such tasks, requiring DSP algorithms for real-time auralisation of audio signals from sources in a complex scene. The apparatus can display audio stimuli as a stereo signal via headphones connected to the master bus output of the game engine. The audio rendering implementation for the system needs to consider modelled source-receiver acoustic characteristics and apply spatialisation effects expressed as HRTF, inserting a spatialised system in the sound reproduction chain (Liu and Manocha, 2022).

Real-time acoustic effects apply to audio buffers from scene sources via real-time DSP convolution algorithms such as Overlap-Save or Overlap-Add. The former optimises the number of convolution operations. Though, this is unfavourable when the kernel function is updated frequently, leading to audio artefacts caused by outdated convolutions applied to the audio buffer. Hence, the implementation adopts the Overlap-Add algorithm, allowing for fast convolution of kernel functions, which can be updated across update calls within the engine audio thread to audio buffer chunks.

Figure 6.5 shows the algorithm implementation considering HRTFs combined with RIR applied to a given audio signal block from the audio buffer associated with a given scene source, showing the system as a single channel for simplicity. For each update call in the engine audio thread, the algorithm convolves a block of audio samples of length B containing interleaved stereo signal to a kernel function of length K. Zero padding, applied to both the kernel function and blocks, allows a DFT toolbox to multiply the spectrum of the two signals ($\bigotimes$ operation), generating a $B + K - 1$ convolution result. B-long blocks (the turquoise blocks) from past convolution results sum into a new block, discarding already summed blocks (the grey blocks). Finally, the new block replaces the given input block from the audio buffer.

6.2.8 Acoustic Geometry Handling

The geometry handling system within the acoustic rendering pipeline plays a crucial role in accurately representing the physical environment to facilitate realistic sound propagation. Environment meshes obtained from the spatial understanding system of the HoloLens Mixed Reality Toolkit, the system can effectively process geometry by constructing a BVH and integrating material classification for enhanced acoustic rendering.

The MRTK provides detailed environment meshes that serve as the primary input for

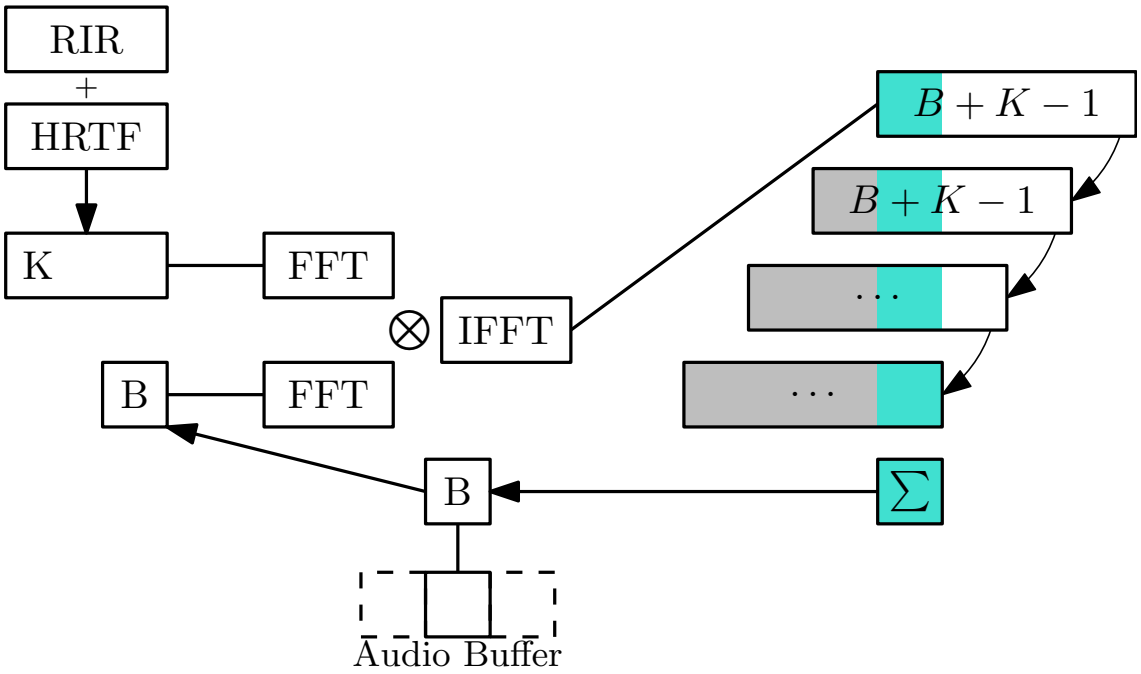


Figure 6.5: Overlap-Add algorithm used for real-time convolution of a kernel function expressing Room Impulse Responses and Head-Related Transfer Functions applied to chunks of audio data from an audio buffer.

the geometry handling system. The geometry handling system constructs a BVH. This hierarchical structure encapsulates the triangles from the Environment Mesh into a series of nested nodes, enabling rapid traversal and intersection testing. The BVH is constructed by grouping nodes based on their spatial proximity, creating a tree where each node represents a subset of triangles of the input mesh. According to Pelzer and Vorländer (2010)'s work, the minimum size for leaf nodes is set to 7 cm to prevent the construction of nodes that are too small to have a significant impact on the auralisation process.

In addition to encapsulating geometric data, nodes within the BVH are used to generate image patches that express the visual features of the environment, which are essential for the material classification process. This allows the prediction of acoustic characteristics of different surfaces within the environment. For each node, the system captures visual features by generating image patches that represent the appearance of the surfaces contained within the node. These patches provide a visual context that is crucial for accurate material classification. Following the methodologies presented throughout Chapter 4, the generated image patches are fed into a material classifier, which uses machine learning algorithms to predict the acoustic properties of the surfaces. The classifier analyses the visual features to determine attributes such as absorption coefficients. Once the material classifier has predicted the acoustic characteristics of the surfaces, these attributes are assigned to the corresponding triangles within the BVH leaf nodes. This process ensures that each surface in the input mesh is accurately represented in terms of its acoustic properties, allowing for realistic propagation modeling. The predicted acoustic characteristics are attributed to the triangles within each leaf node, enabling the DSP engine to account for variations

in surface materials during the sound rendering process.

6.2.9 Environment Geometry completion

As discussed in Section 6.1.2, the environment mesh reconstructed by the HMD may mismatch the physical space, causing errors within the simulated soundfield and the resulting stimuli.

To mitigate these issues, algorithms for mesh completion could be explored as part of outlining the vision of a sound rendering pipeline, providing a solution to the geometry mismatch. These algorithms would infer and reconstruct the unseen parts of the scene, allowing the system to create a more complete and accurate environment mesh even before the user has fully explored the space. By predicting the likely structure and materials of the missing areas, these mesh completion techniques could reduce errors in soundfield simulation, leading to more accurate and realistic auditory experiences. The field of computer vision offers techniques like Cao and De Charette (2022)'s work, allowing the prediction of 3D scene geometry from a single monocular RGB image. Such a technique expands towards the task of semantic scene completion, addressing the issue of reconstructing 3D environments from scarce or minimal information about the world.

Future work could explore the use of semantic scene completion to improve the state of sound rendering within AAR by predicting unexplored areas of the physical space around the user. This would mitigate issues around the correctness of the input model for sound propagation and reduce errors in reverberation estimation and simulation of other phenomena like occlusion or diffraction.

6.3 Vision for Dynamic Auralisation Systems

This section outlines the design of future prototypes for dynamic auralisations expanding from the proposed systems toward real-time interpolation of RIRs computed across a given environment. Beyond a baseline prototype for modelling acoustics of MR environments are potential avenues expanding towards areas of AAR. These can branch into domains of psychoacoustics and interactions between MR sound sources and listeners, adding to the overall aural experience beyond colouring "dry" propagating audio with acoustic effects, thus enabling the further potential for rich auditory displays in applications. Prior to expanding from the primary objective of rendering virtual audio-emitting entities that perceptually belong to the user's physical surroundings, there must be a standardised rendering pipeline that satisfies the minimum psychoacoustic requirements, as there is still a need for research defining thresholds and just noticeable differences of perceived aspects of soundfields, like reverberation or definition (Yang, Barde and Billinghurst, 2022).

Chapter 7 demonstrates a prototype deployment of the implementation discussed in the previous Sections, characterising the psychoacoustic factors associated with approximating

the soundfield surrounding the user. The implementation focuses on a single sound source, omitting optimisations that target multiple sources, real-time aspects, dynamic environments scenarios, and factors targeting realistic acoustic rendering. With the human listener being the target for the generated audio-visual stimuli, rendering procedures affecting the perceived quality require probing of human perception to measure performance aspects and realism of the system.

Despite the need to outline perceptual responses and characterise psychoacoustic factors of rendering pipelines, a vision for research directions and expansions toward context-aware and realistic auditory interactions in MR should be developed.

6.3.1 Combining Reinforcement Learning With Material Tagging

Tagging materials in MR space has been shown to significantly influence simulated soundfields and resulting auralisations. By accurately identifying and assigning acoustic properties to surfaces within an MR environment, it is possible to create highly realistic and immersive auditory experiences. This process involves matching virtual acoustic materials to their physical counterparts and validating the accuracy through impulse response measurements. The precision of material tagging directly impacts the quality of sound propagation and the user's overall auditory experience. So far, this work has primarily focused on matching virtual acoustic materials to their physical ground truth, especially considering the work in Chapters 4 and 5, involving testing to ensure that the virtual materials accurately replicate the acoustic properties of the real-world surfaces they represent. The error margins are assessed based on impulse responses measured from the physical environment, aiming to minimise discrepancies and enhance realism.

Reinforcement Learning (RL) offers tools that can target wider problems around XR and fields of acoustics by leveraging material tagging. Rather than achieving the closest match between virtual and real acoustic materials, RL agents can be trained to assign materials to geometry within a given environment to improve the parameters of its soundfield. RL agents combined with simulated soundfields are being explored as part of new research avenues around audio-visual navigation or locomotion for robotics applications, where models can be trained to understand acoustic characteristics or space by analysing visual and geometrical features of virtual environments (Chen, Al-Halah and Grauman, 2021; Gan et al., 2020). The training process would involve simulating various configurations of acoustic materials and evaluating their impact on the impulse response. The RL agent uses this feedback to iteratively adjust the material assignments, gradually converging on an optimal solution that meets the target criteria. The application of this has several practical benefits and potential use cases around architectural design. It can be employed to achieve specific acoustic properties that are crucial for spaces such as concert halls, recording studios, and lecture theaters. An RL agent can predict and select materials that optimise sound quality, ensuring that the designed spaces meet the desired acoustic standards such as ISO standards for assessing acoustics of space for musical performance. This is an open research

avenue that has been explored using classic sound rendering methods by Funkhouser et al. (2004) and Heimes, Imran and Vorländer (2019).

6.3.2 Hear-Through Displays and Context-Aware Dereverberation

Among the requirements and design principles of audio pipelines are safety and accessibility; the platform should interact with the user's safety and improve their perception of their surroundings.

An important task here is improving the perception of physical sound sources through dereverberation techniques, leveraging scene understanding techniques and human-computer interfaces to resolve the cocktail party problem in adverse acoustic environments with concurrent propagating sources. A real-time system to improve perception of physical sound sources in MR space could target hard-of-hearing users to increase intelligibility and reduce information loss in sound transmissions, targeting a range of use cases around accessibility, training, or education.

Speech dereverberation is still an open research question in signal processing domains where the primary task is to perform signal processing to reduce the impact of the environment on a given sound transmission. Modern approaches to dereverberation extract features from a given signal, as demonstrated by Santos and Falk (2018), understanding characteristics of the soundfield in which the signal propagates to counteract the colouring effects responsible for reverberation and other acoustic phenomena affecting the unpropagated sound, which hinder definition, clarity, and intelligibility and can render auditory information inaccessible to some users.

Speech dereverberation techniques directly benefit from advances in sound rendering algorithms targeted at MR as they allow reconstructions of characteristics of soundfield for arbitrary placements of sources and receivers, as well as architectures and materials of the environment. This information provides dereverberation algorithms with knowledge about a sound transmission and an RIR that can counteract the colouring affecting the signal, enhancing clarity and definition by recovering the unpropagated signal.

Chen et al. (2023b) demonstrate an example implementation using a multimodal DNN with RGB images, depth maps, an input audio signal, and a processed de-reverberated signal as output. Their system can remove reverb and recover characteristics of the soundfield across a set of large-scale 3D reconstructions of physical environments. MR platforms provide sensing technologies to satisfy the input of such example systems. A vision for audio pipelines should consider these processing techniques as MR scenes become increasingly complex and able to interact with physical entities surrounding the listener.

A crucial use case for this avenue would be creating accessible applications for improving speech intelligibility in noisy environments. For instance, receiving information from a tour guide in a museum, speaking may be a challenging task for some due to the highly

reverberant environment affecting the clarity and definition of the sound transmission between the tour guide and the recipient. Here, a dereverberation model can leverage hear-through displays by:

- registering the tour guide as a sound-emitting entity within the MR space and capturing their audio;
- simulating the acoustic soundfield based on the reconstructed physical environments obtained by the spatial understanding system;
- applying dereverberation to the captured audio from the tour guide, based on the simulated soundfield;
- presenting the processed audio with improved clarity and definition to the user to complete the hear-through display system.

6.3.3 Extension to multiple sources and source clustering

There is potential for sound rendering for large multi-source scenes, leveraging the work presented by Schissler and Manocha (2016); the authors introduce a novel sound rendering method that can handle from a few hundred to thousands of sound-emitting objects using ray tracing-based renderers. A factor to consider in the proposal of novel audio pipelines is how rendering processes are limited to the number of sound sources they can manage in a dynamic complex scene. Their method achieves this with a perceptual metric-driven source clustering paradigm based on the premise that sound sources from distant or occluded emitters can be perceptually difficult to distinguish. Thus, the audio engine can approximate multiple emitters as one, provided there is a line of sight among clustered emitters, i.e., no occlusion across emitter pairs.

As virtual complex scenes become increasingly rich in entities and more dynamic, source clustering is a psychoacoustic-related factor important to perceptual rendering. Audio pipelines can increase their efficiency by applying selective rendering in busy scenes, often essential for realism and optimal performance of generation of audio-visual stimuli for training applications and simulated realistic environments (Woodward, 2021). Hence, source clustering should be considered when designing novel pipelines by probing human perception to outline perceptual thresholds and evaluate how well listeners are able to distinguish individual sound sources in complex environments with concurrent source clusters.

6.3.4 Extension to Dynamic Impulse Response Interpolation

Recent advances in CNNs enable image-to-image networks to infer characteristics of a given soundfield. Singh et al. (2021) present pioneering work by proposing a novel network that infers an RIR from a single RGB image describing an environment where a hypothetical sound transmission occurs. They use image-to-image CNNs due to their ability to extract features from visual representations and map those to a latent space that can project onto

an image expressing the frequency domain of an IR. The single input image expands into an RGBD image using a state-of-the-art depth prediction network (Bauer et al., 2021), which provides insights into the architectural features of the space described by the input image. Their method can predict reverberation features of a given space and has the potential for sound rendering and dereverberation applications in MR platforms. Furthermore, such a system has promising performance and efficiency metrics thanks to the task of understanding and prediction sound transmissions within a soundfield compresses into a forward pass of a neural network, which is becoming more trivial by the year, as both tensor processing hardware provided in consumer platforms optimises for such operations and cloud computing becomes more ubiquitous.

One limitation, however, is the omission of source-listener position pair information considered in the reverberation estimation process, resulting in an IR agnostic of some crucial factors of sound transmissions. There is the potential for expanding this system to provide a more accurate output by considering the position of the hypothetical source and listener in MR space. The method proposed by Singh et al. (2021) could leverage depth estimation and scene understanding paradigms provided by the MR platform to provide information about the listener's position and orientation, the source position. The expanded network would combine visual and spatial features to understand acoustic phenomena affecting transmissions beyond reverberations, such as occlusion or diffraction. The network's training process would leverage artificial soundfields based on real environments by creating datasets expressing a range of sound transmissions affected by such acoustic phenomena. The work by Chen et al. (2022b) is becoming a pillar of modern sound rendering, thanks to the ability to simulate such acoustic phenomena in realistic environments that enable multi-modal perception. A vision for sound rendering MR should leverage the potential of such DNN-based approaches to infer characteristics of soundfields with high efficiency, sidestepping the computational load of using geometrical acoustics or other classic propagation techniques.

6.3.5 Extension to Radiance Fields for Inferring Local Acoustic Features of Reconstructed Space

Parametric soundfield coding for interactive applications, as recently advanced by Raghuvanshi and Snyder (2018), provides an efficient solution to real-time auralisations in complex scenes and game engines, allowing simulations of advanced acoustic phenomena by leveraging wave-based sound propagation techniques. Even more recently, building upon the domain of neural radiance fields, the pioneering work from Luo et al. (2022) proposes novel neural acoustic fields able to represent the soundfield of arbitrary complex scenes. Their NAF takes parameters of source-receiver positions pairs, as well as listener's orientation information, to produce a function that represents a RIR as a STFTs that can be represented as a time-domain response. The NAF can train by comparing generated spectrograms, using an MSE loss function against ground-truth counterparts from synthetic or

real datasets (Chen et al., 2022b).

NAFs learn using joint acoustic/visual representations to construct a local geometric grid that can infer acoustic energy transfer from emitter to receiver by querying the radiance field. The technique performs feature extraction on a triangulated environmental mesh, modelling reflection, occlusion, and other soundfield phenomena represented by a volumetric scene function. Similar to wavefield coding techniques, NAFs suffer from requiring training every time the state of the environment has changes that may have perceptual responses in the resulting auralisations, such as new holograms introduced to the complex scene or moving objects in the physical space, causing the feature grid to be outdated. The technique should be able to update segments of the feature grid by computing inference on the updated spatial reconstruction mesh provided the HMD.

Here is a potential alternative to leverage the feature extraction process, discussed in Section 6.3.4, using an image-to-image network to infer characteristics of the soundfield from visual descriptions of the environment. MR platforms offer arrays of cameras that can describe the environment surrounding the user. With image-to-image networks inferring acoustic properties of space photographed as the user looks around, it would be possible to project such features onto a grid that can be queried to produce acoustic responses. The benefit of such expansion would lie in the implementation of a system that continuously updates feature points, providing an online grid that can reflect dynamic changes to both the virtual and physical environment. Even by employing radiance fields, the interface between the game engine and the DSP engine would be maintained valid as the game engine would query the radiance field based on the current source and listener position.

6.3.6 Conclusions

Design principles and potential research avenues are identified to provide a vision that leverages the potential provided by MR platforms around sensing technologies shipped with consumer hardware and interfaces with the user and the physical world. This vision reflects on a base prototype expressed as a system composed of individually tested components to achieve several goals associated with standard audio pipeline design standards. This Chapter has shown how several research domains could benefit from expansions stemming from the prototypes and addressing tasks associated with use cases such as accessibility or enhanced perception of surroundings around users wearing MR HMDs.

Given the early stage of the prototype, the next step towards an end-to-end deployable solution should consider testing and the analysis of human factors, evaluating how the components discussed and developed so far work together to produce auditory stimuli with improved realism and presence compared to standard audio implementations in game engines. This should generate insights into the feasibility of the overall pipeline and provide data on the perceptual aspects. Chapter 7, aiming at obtaining early data on the pipeline, will illustrate and discuss preliminary subjective testing with an offline prototype.

Chapter 7

Psychoacoustic Characterisation of Rendering Pipelines for Augmented Acoustics

Chapter 6 demonstrates the engineering process for the implementation of the envisioned proposed pipeline, prototyped on a consumer MR HMD. This enables the generation of context-aware auditory stimuli that interact with the physical and virtual scene around the user. The deployment of the pipeline comes at non-trivial computational costs associated with requirements to process scene understanding techniques discussed in Chapter 4 for the retrieval of scene acoustic characteristics. Rendering techniques add further requirements to generate acoustic phenomena relative to a given source-listener sound transmission. The costs of the deployment raise the important question: does the computational load justify an increased objective or subjective measure attributed to auditory interactions? The data extracted by an early prototype will provide insights needed to formulate an answer for this question.

7.1 Introduction

Chapter 5 discusses increased objective measures evaluated on simulated soundfields and propagated audio. However, it is only by probing human perception that one can address the question concerning subjective measures and justify the computational costs of the engineering process associated with the prototype. With the human listener being the central link of the chain of the proposed system, factors of auditory perception can be investigated to study how acoustic rendering pipelines influence the psychoacoustic response. Perceptive aspects of sound transmissions in immersive applications can affect interactions and activities conducted by the user, especially if such interactions have relationships with spatial characteristics of the environment. The following Sections will investigate psychoacoustic factors of sound transmissions in virtual environments. The proposed acoustic rendering

pipeline is tested on an MR device and evaluated by investigating its impact on tasks requiring the application of psychoacoustic abilities. This Chapter tests how well subjects are able to match spatial information provided by an individualised auditory display with holograms projected onto their surroundings. The findings of this chapter describe how an acoustic spatialisation workflow applied to a common teaching room has a significant impact on ability in common psychoacoustic tasks as opposed to a no acoustic spatialisation condition.

This Chapter contributes towards:

1. the presentation of a novel methodology for studying psychoacoustic factors of sound transmissions between sound sources and a human listener within a simulated soundfield displayed in augmented reality;
2. the development of a novel bespoke testing framework for conducting a user evaluation in augmented reality using a custom audio engine and a prototype acoustic rendering pipeline;
3. outlining recommendations for future designs of audio rendering pipelines for immersive technology drawing from a dataset of perceptual responses captured from a user study on two tasks dependent on psychoacoustic abilities.

7.2 Methodology

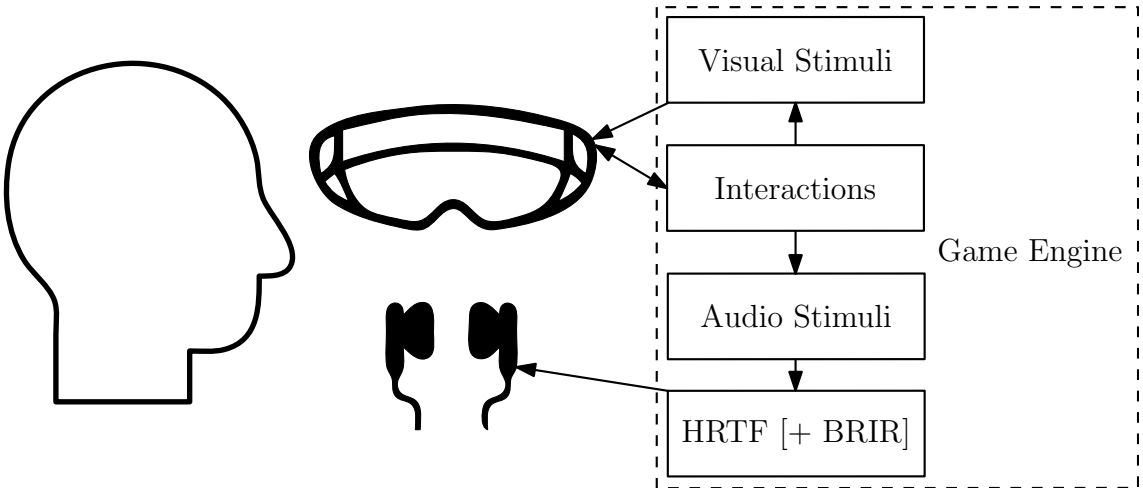


Figure 7.1: The experimental apparatus used to implement the testing framework and execute the two tasks, *localisation* and *clustering*, administered to recruited participants.

7.2.1 Overview

This evaluation aims at investigating how audio interactions in MR affected by simulated acoustic phenomena influence task perfomance in complex scenes. The methodology evaluates tasks that are dependent on psychoacoustic abilities, such as pinpointing the location

and resolving the direction of arrival of sound sources or determining the number of concurrent sound sources at varying angular distances between sources. To investigate these factors, a within-subjects evaluation is designed to probe the human auditory perception applied to practical tasks in a virtual environment. Localisation of sound-emitting objects is set as the primary psychoacoustic-related task affecting task performance in immersive environments, followed by testing the ability to detect clustered sound sources, as source clustering is a key factor in sound propagation system design for complex scenes that are typical of MR applications (Schissler and Manocha, 2016).

7.2.2 Task Definitions

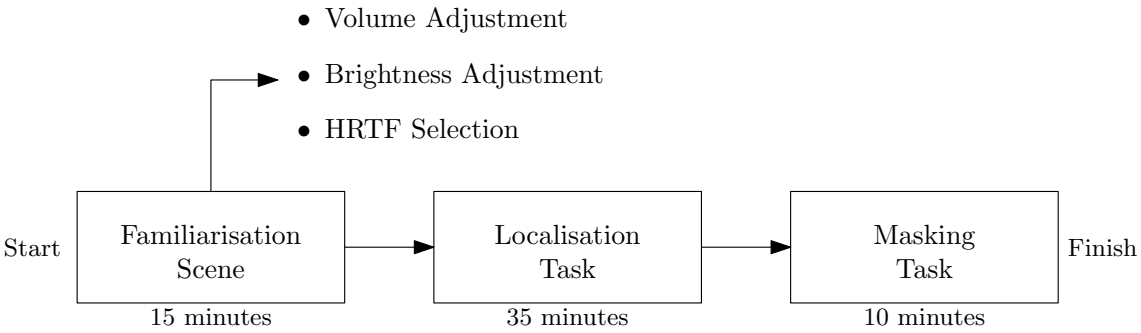


Figure 7.2: Experiment procedure administered to participants: the task starts with a familiarisation scene where the user is introduced to the UI system and is asked to calibrate the audio reproduction system. The user then performs the localisation and masking tasks after the familiarisation scene.

The subjective study is administered to participants as an MR application, see Figure 7.2 and focuses on two tasks that are:

1. the **localisation** task: where participants are asked to listen to a set of audio stimuli and indicate which holographic object is responsible for emitting each stimulus, choosing from arrays of holographic objects;

2. the **masking** task: where participants are displayed concurrent audio stimuli, and they are asked to determine whether these propagate from a singular or two invisible sound-emitting objects.

The rationale for defining the tasks centre around key aspects of sound rendering, namely the efficacy of propagation effects, measured by analysing user performance in tasks requiring the perception of spatial clues encoded in audio transmissions presented to the users within the complex scene. The other aspect centres around the problem of performance the limited computational resources in HMDs. The localisation task aims at evaluating whether there is a significant increase in perceived spatial information using a sound rendering pipeline, compared to unpropagated stimuli. This task is often evaluated to analyse the effectiveness of sound rendering methods (Rungta et al., 2016). The masking task measures how concurrent sound sources are perceived by the user, studying the efficacy of sound rendering pipelines in resolving masking effects occurring in sound sources at

varying distances from each other. Testing how users perceive concurrent sound sources can provide insights and guidelines for handling clustered sources in large-scale scenes, aiming at reducing the load of propagation methods simulating multiple sources (Schissler and Manocha, 2016). Section 7.2.10 details the procedure repeated for each participant to conduct the two tasks.

7.2.3 Participants

31 participants, of whom 16 females and 15 males, were recruited from staff and students of Birmingham City University and were invited to partake in the two tasks, which are defined as follows. Participants were asked to pinpoint the location of a set of sounds, resolving the direction of arrival of sound sources and determining the number of concurrent sound sources they could identify at varying angular distances. A within-subjects design was conducted with 31 participants (16 females, 15 males) to probe the human auditory perception of practical tasks in a virtual environment.

7.2.4 Apparatus and Evaluation Platform

The two tasks were implemented in Unity 2021.3.14f1 and deployed and executed in a Microsoft Hololens 2 HMD[1,2]. Figure 7.1 shows the experimental apparatus composing the evaluation framework: the user is presented with visual and auditory stimuli driven by the Unity game engine, responsible for managing the scene representing the evaluation tasks, rendering holographic interactable objects. Audio reproduction was done through the embedded audio interface in the Microsoft Hololens 2 HMD and using a pair of Bose Soundsport earbuds with an integrated D/A converter.

Spatial information relating to the scene and the user, as well as interactions detected by the gesture recognition features of the HMD, determine the reproduction of audio stimuli. These are processed with customised spatialisation tools to consider human factors in the audio reproduction chain. The system is tested in a real space, a teaching room of $10.7m$ by $8.6m$ dimensions, where participants stand in the middle of the room wearing the HMD and a headset to complete the two tasks with their freedom being restricted to 3 DoF.

7.2.5 User Interface System

The testing apparatus incorporates a diegetic user interface integrated within the immersive environment. Diegetic UI elements exist within the narrative world of the MR experience, creating a sense of immersion for the user. This approach has been proven to increase user engagement and realism by making the UI elements a part of the user's surroundings, rather than overlaying non-diegetic elements that can disrupt the task (Dickinson et al., 2021). The UI for the subjective testing was designed using the Microsoft Mixed Reality Toolkit (MRTK)[1]. MRTK provides a comprehensive collection of components and features

[1] https://learn.microsoft.com/en-us/windows/mixed-reality/develop/unity/unity-development-overview

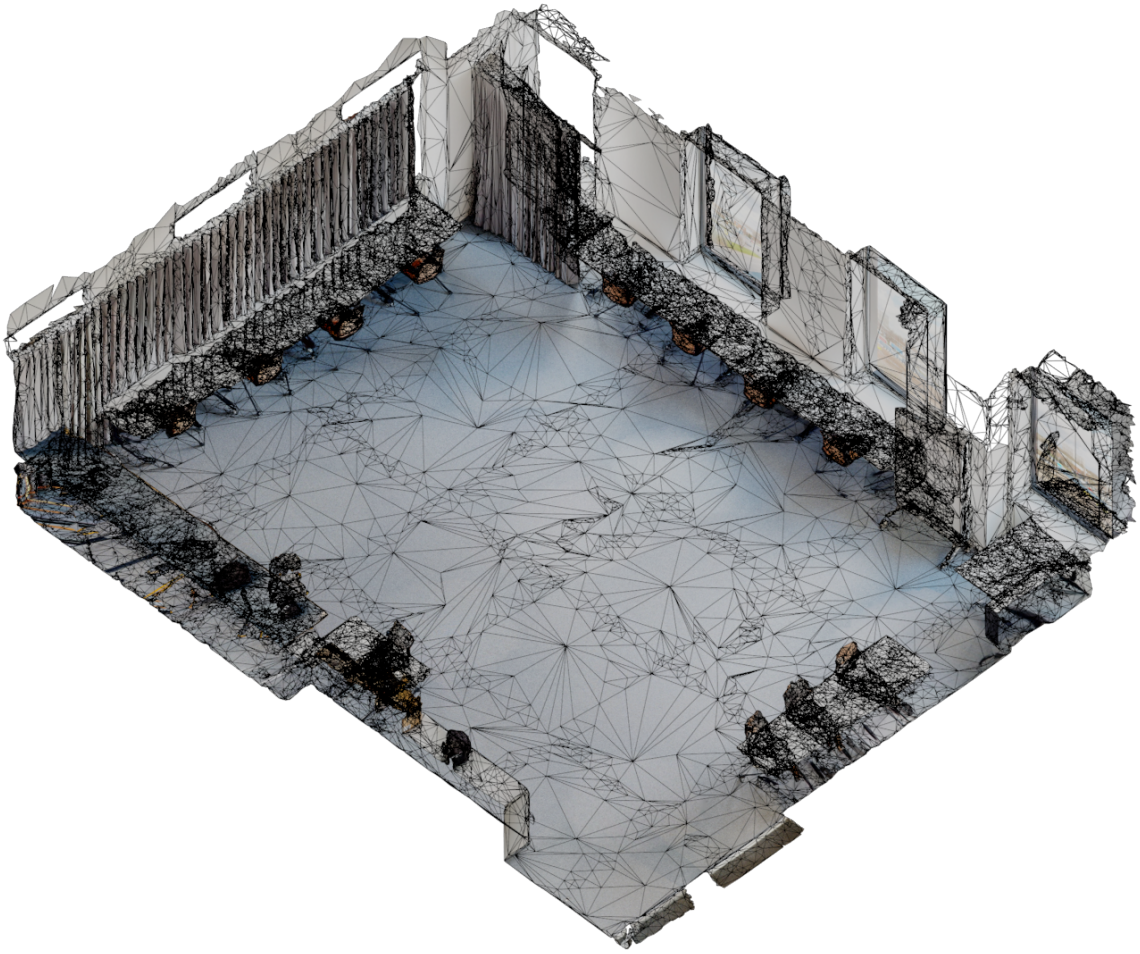


Figure 7.3: Visualisation of the virtual reconstruction of the space used for the user study. The space, scanned using a Matterport Pro3 camera, is manually segmented into semantically related portions of geometry aligning with acoustic material definitions (e.g. curtains, floor, walls) and imported into the game engine for scene alignment and spatial mapping. The scanned geometry is handled using a Bounding Volume Hierarchy to allow indexing and searching of the triangle primitive for acoustic material tagging and rendering operations. Acoustic rendering is performed between sources and the listener, generating impulse response pairs, based on propagation paths computed by a ray tracer bouncing rays off the scene geometry with tagged acoustic materials.

that simplify the development of MR applications. Leveraging MRTK allowed for the creation of intuitive and responsive UI elements, ensuring a consistent and high-quality user experience. The toolkit's support for spatial interactions and holographic elements was crucial in developing an interface that could dynamically adapt to the user's movements and interactions within the MR environment. The toolkit is used to create panels and controls for participants to adjust audio settings, such as volume or current HRTF, see Figure 7.8.

Sound sources in the localisation task (later illustrated by Section 7.2.10) are visually displayed to the user through virtual cubes arranged in three circular arrays corresponding to different distances: near, medium, and far. Each array represented a specific proxemic zone, which is crucial for understanding how users perceive and interact with sound sources in MR. Proxemics zones relate to physical distance as a dimension of interaction design, where the spatial relationship between users and devices influences interaction possibilities (Huang et al., 2022). In the near array, the inner circle of holographic cubes was placed within the personal space of the user, which is defined to be approximately 5×10^{-1} m to 1.5 m. This distance represents close interactions where sounds are expected to be more intimate and directly related to the user. The medium distance array has cubes positioned within the social space, around 1.5 m to 4 m meters from the user. This zone corresponds to social interactions where sounds need to be clearly audible but not as intimate as those in the personal space. The outer circle is placed within the public space, at 4 m distance or farther. Sounds originating from this distance are meant to represent broader environmental audio, providing context without being the primary focus (Büschel, Lehmann and Dachselt, 2021).

7.2.6 Geometry Reconstruction and Handling

The room, the physical test environment, is scanned using a Matterport Pro3[2] camera, from 9 camera placements covering the walkable space of the room uniformly, with the Near array being $0.9m$ from the listener, the Medium $2.24m$, and the Far $4.02m$; refer to Figure 7.3 for a visualisation of the mesh showing wireframe geometry and textures of the reconstruction. The textured triangulated mesh is partitioned into semantically related segments, separating architectural components of the room, as well as furniture and objects, into individual meshes to facilitate indexing and attributing of physical characteristics to portions of the scene geometry. Portions of the segmented geometry are manually tagged by attributing frequency-dependent absorption coefficients and mapping the appearance of the reconstructed textures with acoustic materials from measured absorption data[3], see Figure 7.5. In acoustic rendering operations, material properties assigned to scene geometry are required to approximate models of the environment soundscape. In the testing framework, the reconstructed geometry is used by the game engine to align

[2] https://matterport.com/pro3
[3] https://odeon.dk/downloads/materials/

holographic objects with the physical space. We use a dynamic BVH implementation, constructing the tree using mesh triangles with attributed acoustic absorption characteristics to organise, index, and search the scanned geometry within the game engine for sound propagation (Kopta et al., 2012).

7.2.7 Acoustic Rendering Apparatus

Figure 7.4: A photograph of a teaching space used for deploying and testing an implementation of the proposed system.

The experimental evaluation deploys the rendering pipelines presented in Chapter 6 as an offline process. The soundscape is simulated by generating IR for source-listener pairs in the MR environment by implementing the renderer proposed in Section 5.2, maintaining the validity of the offline renders by adopting static sound sources and environment and limiting the movement freedom of the listener. As shown in Figure 7.1, auditory stimuli use RIRs convolved to anechoic audio samples to reflect approximated acoustic phenomena in audio signals displayed to participants. The experimental evaluation uses ray tracing as an offline rendering to perform sound propagation between scene elements in MR. This is done by restricting sound transmissions to static sound sources and a static listener (the participant). Hence, the participant's position is fixed within the grey holographic barriers, as shown in Figure 7.6. We implement a minimal geometrical acoustics renderer, based on Savioja and Svensson (2015b)'s work, to approximate reverberation by modelling acoustic energy decay with propagation paths between an emitter and detector. Sound sources are considered omnidirectional emitters, from which $3 * 10^3$ rays originate at uniformly distributed points on a $1m$ radius sphere located at the sound source position in the scene, searching for geometrical intersections within the BVH constructed on the scanned space, see Figure 7.3. Intersections are computed between propagating rays and leaf nodes of the BVH, using the normal and acoustic absorption data attributed to the

encapsulated triangle, allowing propagation paths a maximum of four reflection orders. α_f acoustic absorption data provided by tagged materials and attributed to mesh triangles is frequency-dependent, $f \in \{125, 250, 500, 1000, 2000, 4000\}Hz$, following state-of-the-art sound propagation methods (Schissler, Loftin and Manocha, 2017).

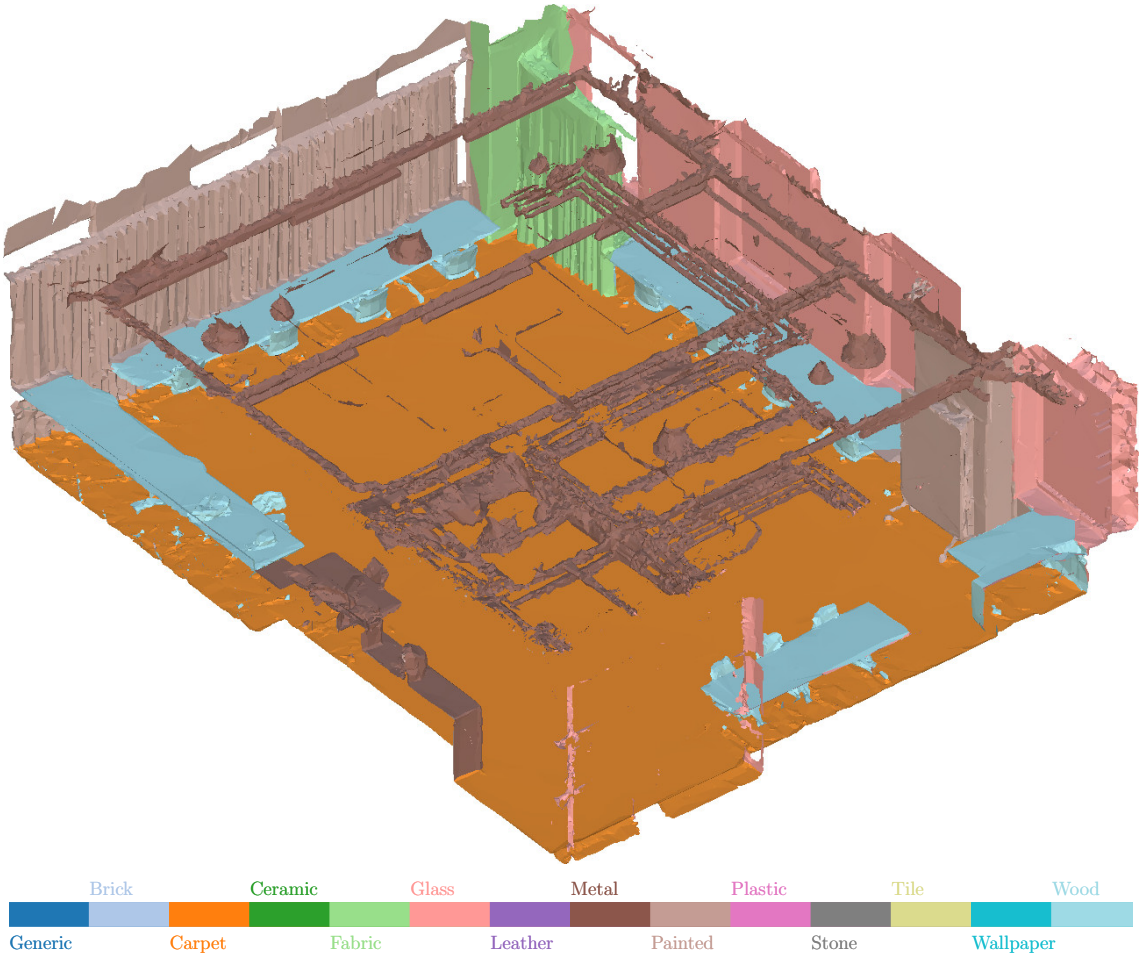


Figure 7.5: An isometric view of the room used for the experiment, visualising acoustic materials. Materials are mapped to the scanned environment geometry (see Figure 7.3) by segmenting the mesh into semantically meaningful submeshes that represent scene entities.

Propagation paths have $e_0 = 1$ initial energy values across six frequency bands relating to frequency-dependent material data, attenuated upon collisions with scene geometry and logged in a series of IRs that model frequency-dependent reverberation. To overcome the deterministic nature of the ray tracer, we adopt Schröder (2011)'s technique, consisting of generating Poisson-distributed sequences of Dirac-Delta pulses to simulate the likelihood of detecting acoustic reflections in a hypothetical environment. These sequences are convolved with the modelled energy from the ray tracer, filtered based on their respective frequency band contributing towards Equivalent Rectangular Bandwidth (ERB) portions of the spectrum centred around each f and having all frequency-dependent IRs contributing equally towards the spectrum, which are finally summed together into a RIR.

Listener and sound source positions were pre-determined by the *localisation* and *clustering*

task, see Figure 7.6; RIRs are pre-computed for all sound-emitting objects placed in the virtual reconstruction of the space.

7.2.8 Audio Rendering Apparatus

The pre-computed RIRs associated with the sound-emitting objects placed in the test scenes were combined with sets of Head-Related Transfer Functions (HRTFs). These were drawn from the public CIPIC database in real-time by implementing a fast convolution engine. This engine convolves RIRs and HRTFs with audio signals, creating two types of audio stimulus sets S:

$$\forall y \in S_A \; y_l(t) = x(t) * h(\theta, \phi, t) \tag{7.1}$$

$$y_r(t) = x(t) * h(\theta, \phi, t) \tag{7.2}$$

$$\forall y \in S_P \; y_l(t) = x(t) * RIR(t) * h(\theta, \phi, t) \tag{7.3}$$

$$y_r(t) = x(t) * RIR(t) * h(\theta, \phi, t) \tag{7.4}$$

where t, θ and ϕ indicate, respectively, time, the listener's azimuthal position and elevation. HRTFs are interpolated in real-time using AES SOFA containers Hoene, Patino Mejia and Cacerovschi (2017) and computed based on the listener's head orientation determined by the HMD tracking system. The final links of the audio reproduction chain consist of the embedded audio interface in the Microsoft Hololens 2 HMD that provides a display to the user via a pair of Bose Soundsport earbuds with an integrated D/A converter.

7.2.9 Audio Stimuli

The testing framework presents two sets of audio stimuli: a *set* of anechoic audio excerpts S_A, and a *set* containing corresponding audio excerpts S_P generated using a prototype deployment of the geometrical acoustics pipeline. The latter set of auralisations is produced based on the listener's position during the testing procedure and the position of the sound-emitting objects in the testing scene.

In localisation test methods, it is common to use test tones or broadband noise as stimuli to evaluate the distance between ground truth source positions and target positions expressed by participants (Bertet et al., 2013; Kashino and Nishida, 1998). Psychoacoustic factors around the perception of loudness may affect the application of localisation abilities (Blauert, 1997). Hence, stimuli are selected to cover a wide range of frequencies around the audible range. For the *localisation* task, both sets are constructed with three anechoic audio samples: a 4.98s drums loop, a 4.87s singing clip, and a 7.14s violin clip. All sound clips are indefinitely looped until the subtask is completed or the participant pauses them with interactable panes presented every time audio is reproduced. The *clustering* task has eight pairs of 25s multi-instrument anechoic recordings[4]. For each pair, the two sound sources emit a different instrument from the same musical excerpt. Emitting the same

[4] https://odeon.dk/downloads/odeon-zip-archives/

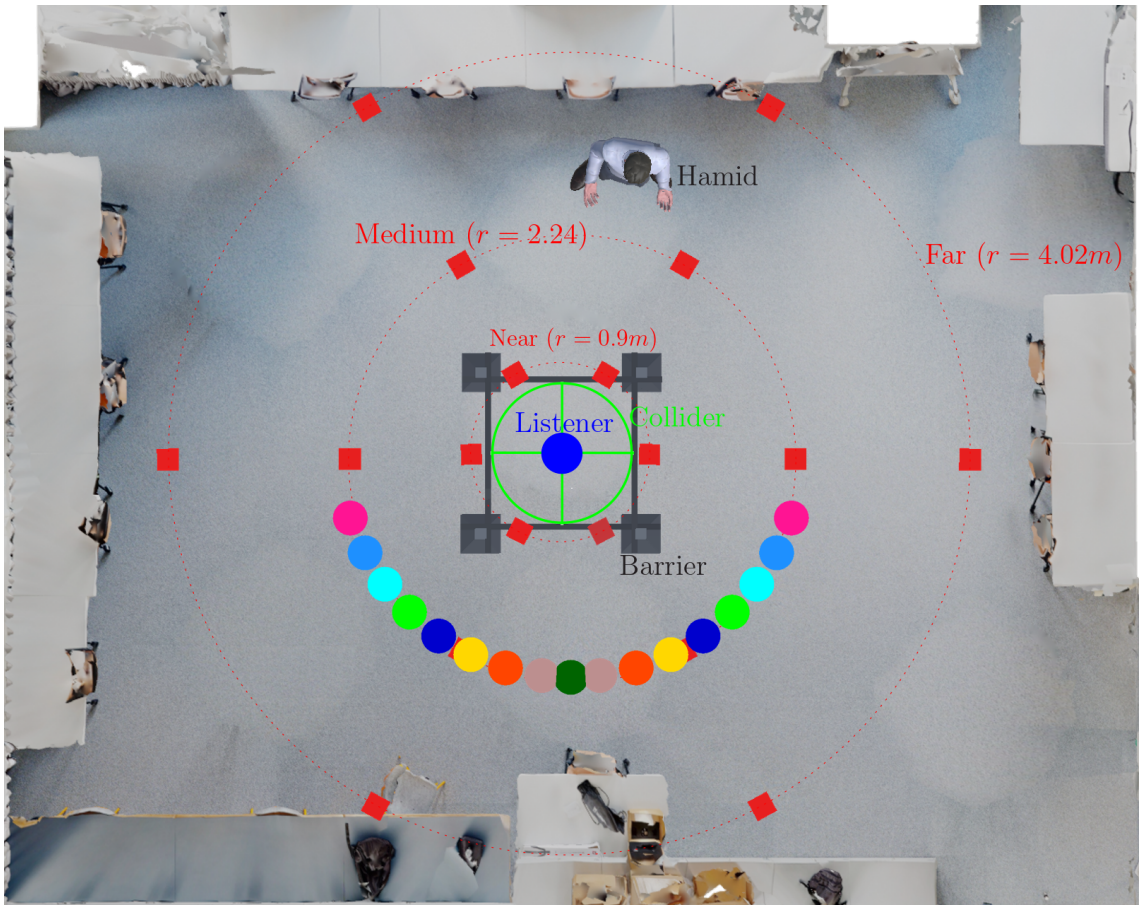


Figure 7.6: Top view of the testing environment overlaid on a top view photograph of the room used for running the user study. The overlay shows the listener at the centre of the room, the blue dot, surrounded by an invisible collider, visualised as a green wireframe, to detect and warn the participant moving outside the bounds. Crowd barriers, the grey objects surrounding the listener, restrict the participant's movement. The red squares visualise the holographic sound sources for the sound localisation task procedure. The colour circle pairs visualise source position localations for the masking test. The circular *Near*, *medium*, and *far* arrays are visualised with dotted circles, indicating the distances from the user, respectively, 0.9, 2.24, and $4.02m$. All holographic cubes have a $60°$ angle of distance between each other.

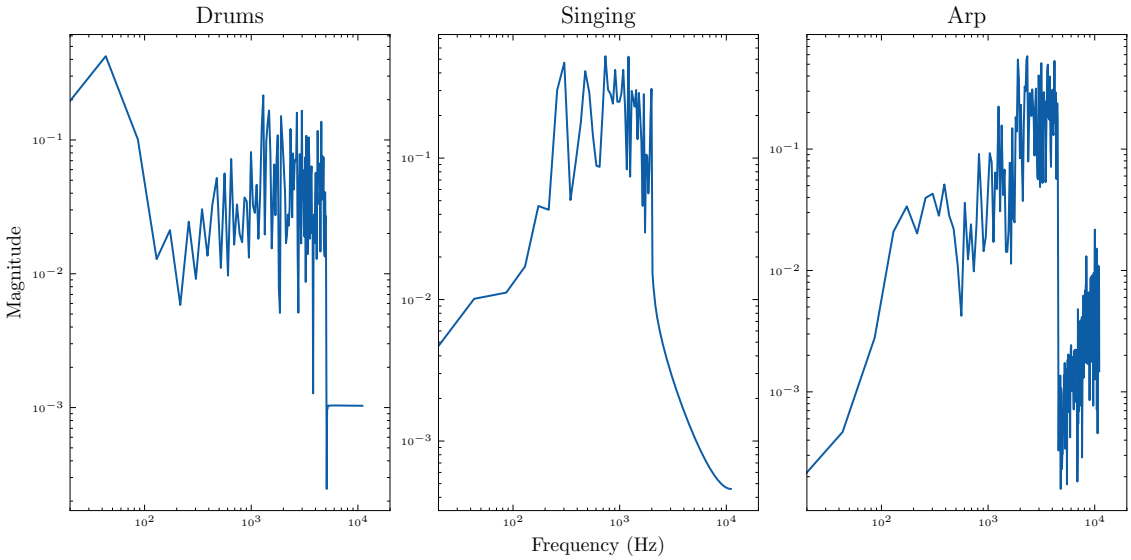


Figure 7.7: The evaluation design uses three auditory stimuli, reproduced as spatialised sound sources placed at positions defined by the near, medium, and far array. The figure shows a spectral analysis conducted over the three stimuli showing how the spectrum of "drums", "singing", and "arp" centre around low, medium, and high frequencies.

audio from the two sources is avoided to prevent potential reproduction artefacts of phase cancellations that could interfere with the task. The *clustering* task has eight anechoic pairs, and eight counterparts propagated with appropriate IRs for a total of 18 subtasks.

7.2.10 Procedure

Each study lasted between 60 and 90 minutes, the study protocol had the approval of the Institutional Review Board. The study was broken down into the following steps. Informed consent was attained from participants prior to each study session. After welcoming participants, the purpose of the study and test protocol was explained in detail. The researcher then measured visual acuity and colour blindness via a Snellen chart and Ishihara test. Participants were then asked to adjust and wear the HMD, so it was comfortable and stood at the initial starting position.

Familiarisation Scene

The testing framework introduces the participant to the tests via a familiarisation scene composed of panels to describe and allow access to the two tasks, barriers to limit the participant's freedom of movement, a panel to explain and control the spatialisation system, and a sound-emitting hologram. The sound-emitting hologram, visualised in Figure 7.6 as the character "Hamid", emits a $102s$ long anechoic recording of a narration summarising knowledge on spatialised audio and the nature of the two tasks; the recording is looped indefinitely until the participant pauses it with panel controls or enters any of the tasks. Users are asked to use this scene to adjust and calibrate the audio-visual setup, ensuring they are comfortable with the tasks and the controls of the User Interface.

Figure 7.8: A frame from the render buffer displayed to the user through the Microsoft Hololens 2 Head-Mounted Display: the frame shows the holographic content displayed to the participant when entering the testing environment, consisting of, from left to right, a set of interactive panels for Head-Related Transfer Function explanations and controls; two sets of panels for learning and accessing the two tasks, localisation and masking; a virtual hologram of Hamid, a spatialised sound-emitting character briefing and summarising key information to the participant. At the bottom is a portion of the crowd barrier object that limits the participant's freedom of movement.

A panel briefs subjects on the basic features of the audio spatialiser system, allowing the selection of different HRTF containers, see Figure 7.8 for a first-person view render of the panel. There are 10 available HRTF selections drawn from the CIPIC database, as described in Section 7.2.8. These are sampled uniformly to represent the anthropometric spectrum of the database. There are no time constraints to the familiarisation scene, and subjects are asked to test each HRTF selection with the sound-emitting hologram until they find the selection that allows the best perceptive visual-acoustic spatial matching. After the familiarisation task was complete, participants were then asked to proceed to the study tasks, with half of them completing the localisation task first and the other half starting with the source clustering tasks.

Localisation Task

The goal of the *localisation* task is to evaluate the success rate of participants identifying the hologram responsible for emitting audio stimuli that are displayed via a bespoke sound system.

The procedure starts with the participant standing on the marker indicated on the floor, within the holographic barriers in the virtual testing environment; a holographic panel presents a quick summary of the tasks and asks whether the participant is ready to proceed. Upon interaction with the proceeding button on the panel, the first circular array of sound sources, the *near* array, is projected via the HMD around the user as static red cubes elevated at $1.4m$, a standard altitude adopted in source-receiver measurements in

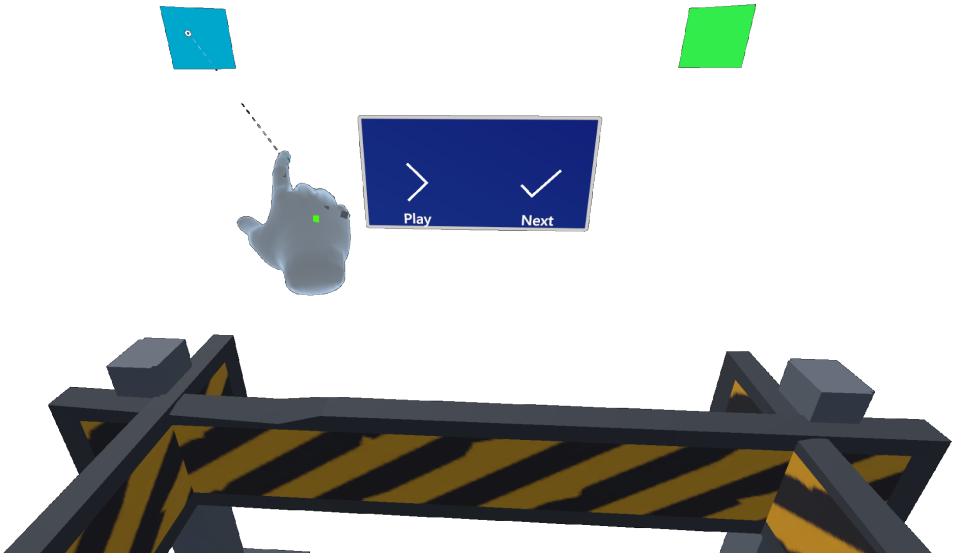


Figure 7.9: First-person view of the localisation task: the participant, surrounded by circular arrays of sound-emitting cuboids, perceives spatialised audio and is asked to determine which of the holographic cubes is responsible for emitting sounds. The participant selects the cube that best matches the spatial position and direction conveyed by the auditory information by indicating with their hands, tracked by the HMD.

soundfield. With the array of sound sources around the user, a spatialised audio stimulus is displayed through the sound reproduction setup; the participant is tasked with indicating the holographic cube they deem responsible for emitting the stimulus. The cube tint transitions from red to green, indicating the user's selection. The stimulus is reproduced in a loop, without time limits, until the user proceeds to the successive stimulus using a panel always shown in front of the user. The stimuli are drawn randomly from both sets $S_A \cup S_P$.

The task has three phases: the near array, the medium array, and the far array. These arrays are distributed across the walkable radius of the testing room, see Figure 7.6. All phases have six sound-emitting objects, each emitting six stimuli, the three anechoic audio clips and three corresponding propagated clips. Hence, each participant is presented with 108 stimuli. For all three phases, participants were to complete all subtasks. For each subtask, an auditory stimulus was presented, drawing from a shuffled set of anechoic and propagated audio clips associated with each phase.

Clustering Task

The goal of the *clustering* task is to investigate how well participants can resolve clustered invisible sound sources across the two sets of audio stimuli. The procedure starts with the participant standing on the marker indicated on the floor within the holographic barriers; a holographic panel summarises the task to the participants, and upon interaction with the

Figure 7.10: First-person view of the clustering task: the participant is presented with spatialised audio that may be emitted by a *single* or *multiple* invisible sound-emitting object(s) and is asked to classify each stimulus using buttons on a holographic panel. The render shows coloured spheres (the perspective camera might affect the appearance of the sphere objects) to visualise the position and distance of sound source pairs; they are invisible to the participants.

proceeding button on the panel, the first subtask is administered. For each subtask, the participant is displayed an auditory stimulus and asked to determine whether it is emitted by a *single* or *multiple* invisible sound-emitting object(s). The participant selects one or the other using a panel showing two interactable buttons, "Single" and "Multiple".

Audio stimuli are emitted from a single or pairs of invisible sound sources, represented as colour-coded circles in Figure 7.6. Pairs are positioned along a semicircular array with a radius of $2.24m$, with the listener at the centre of the semicircle. Starting from a singular source in front of the listener at the midpoint of the semicircle, there are eight source pairs with increasing angular distance, about 15 to 20° between pairs. Subtasks are random samplings of the eight sources plus the single source, repeated twice, one for anechoic audio and one for propagated audio. Hence, each participant completes 18 selections.

7.2.11 Evaluation

Data collected from procedures is analysed with the goal of evaluating the efficacy of the proposed pipeline and assessing its impact on immersive applications. One of the primary techniques used to compare populations and evaluate significant differences between anechoic stimuli and stimuli generated by the sound rendering pipeline is the Analysis of Variance (ANOVA) or Mann-Whitney U non-parametric test, both particularly useful for comparing the means of multiple groups to determine if there are statistically significant differences among them. In cases where the data do not meet the assumptions required

for ANOVA, such as normality or homogeneity of variances, non-parametric tests like the Kruskal-Wallis test are utilised. These tests provide a robust alternative by comparing the medians of the groups, offering insights into whether the sound rendering pipeline produces significantly different auditory experiences compared to anechoic conditions.

Another aspect of the evaluation involves measuring localisation error, which assesses the accuracy with which participants can identify the spatial origin of sound sources. This metric is quantified by calculating the angular deviation between the actual and perceived locations of the sound sources. In addition to localisation error, clustering error and accuracy are measured to evaluate how well the participants can group similar sounds together based on their perceived spatial and spectral characteristics. Clustering error is assessed by analysing the consistency of participants' responses in grouping sounds that originate from the same or similar locations. Accuracy is determined by comparing these groupings to the actual configurations of the sound sources. Statistical methods such as cluster analysis and multidimensional scaling are employed to visualize the data and identify patterns. Furthermore, measures like the receiver operating characteristic curve (ROC) can be used to quantify the accuracy between the participant-derived clusters and the true clusters.

7.3 Results

The evaluation framework is deployed to the HMD, and responses are gathered from play sessions with the recruited participants, investigating whether the geometrical acoustics-based rendering pipeline has a significant effect on psychoacoustic-related abilities performed in MR. The overall null hypothesis set for this evaluation is that task performance is not affected by applying approximated acoustic phenomena to auditory stimuli. A total of 3348 responses were drawn from the *localisation* task and 522 responses from the *clustering* task, as two participants withdrew from the latter task.

7.3.1 Localisation Task

The observed metric for the *localisation* task is the angular distance between the position of the sound source and the sound source that the participant determines as responsible for reproducing the stimulus and the position of the true sound source, referred to as localisation "Error", see Figure 7.12. To determine whether this task is affected by acoustic phenomena applied to auditory stimuli, we set a null hypothesis H_0 to determine no significant difference between localisation error samples associated with anechoic audio and localisation error samples associated with propagated audio.

The two Error populations are tested for normality to evaluate whether the assumptions for conducting variance tests by setting a hypothesis H_n determine that the two populations come from a normal distribution (Pearson, D'Agostino and Bowman, 1977). Normality tests fail to prove H_n for both distributions with $7.9 * 10^2$, $7.3 * 10^2$ statistical scores and $\rho = 6.2 * 10^{-17}$, $\rho = 3.4 * 10^{-158}$ significance values for Error samples associated with

the *anechoic* and *pipeline* stimulus sets, respectively. Given the departure from normality in the recorded responses, the Mann-Whitney U non-parametric test is used to compare the distributions, rejecting H_0 with $1.34 * 10^6$ U score and $\rho = 0.042$ significance value. The significance value threshold is set at 0.05 for all hypotheses.

Table 7.1: The table shows mean, mean standard error, standard deviation, variance score, kurtosis, and skewness factors computed on **distances** between true values and responses indicated by subjects performing the localisation task across the two sets of audio stimuli, anechoic S_A and pipeline-generated S_P. These distances measure the angle between the position of the object responsible for propagating sounds and the object the participants deemed responsible for propagating sounds. The Standard Error for Kurtosis and Skewness scores are, respectively, 0.2 and 0.1 across all distances for both S_A and S_P.

Distance	Stimulus Set	μ	S.E.μ	σ	Variance	Kurtosis	Skewness
Near	S_A	30.93	1.62	38.03	1446.32	4.6	2.32
	S_P	**24.76**	1.37	31.55	995.56	7.24	2.76
Medium	S_A	24.85	1.77	42.11	1773.43	4.08	2.17
	S_P	**21.36**	1.53	37.22	1385.09	5.58	2.4
Far	S_A	**22.84**	1.89	44.67	1995.84	4.38	2.28
	S_P	25.16	1.92	45.28	2050.01	3.32	2.04

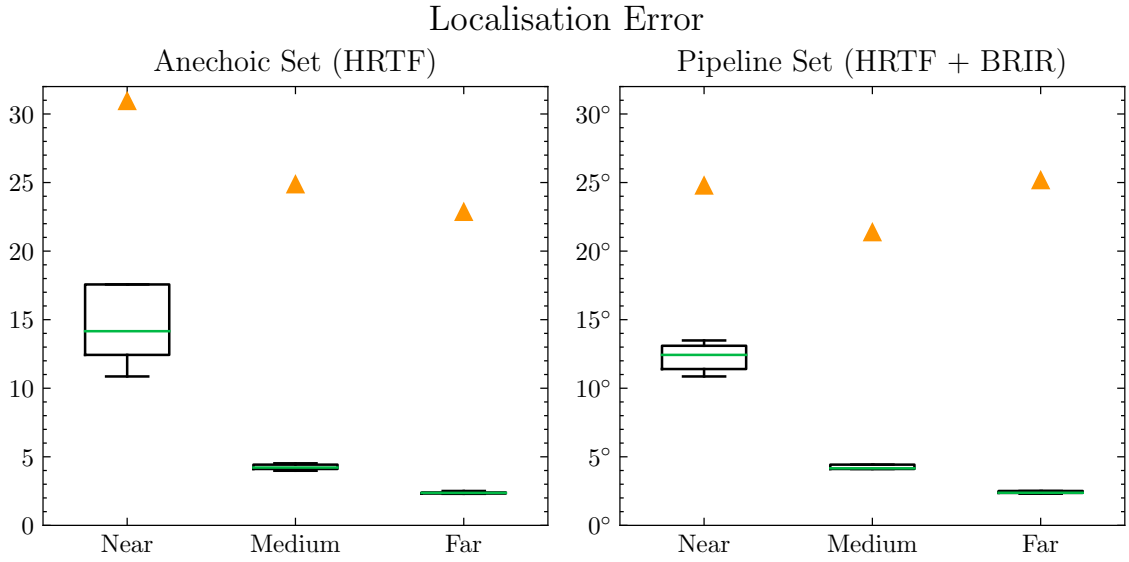


Figure 7.11: Boxplots comparing localisation errors across near, medium, and far sound source distances, showing Interquartile range (the box), the median value (the green line), and the mean value for each group, across the S_A and S_P stimulus sets.

7.3.2 Clustering Task

The observed metric for the *clustering* task is the rate of correct answers provided by subjects for each subtask, where they are asked to identify whether a single or multiple invisible sources emit the sound. The responses are analysed as a binary classification task, measuring accuracy and precision across the two stimulus sets. Table 7.2 shows the accuracy of sound source cluster classification for two stimulus sets and their union across

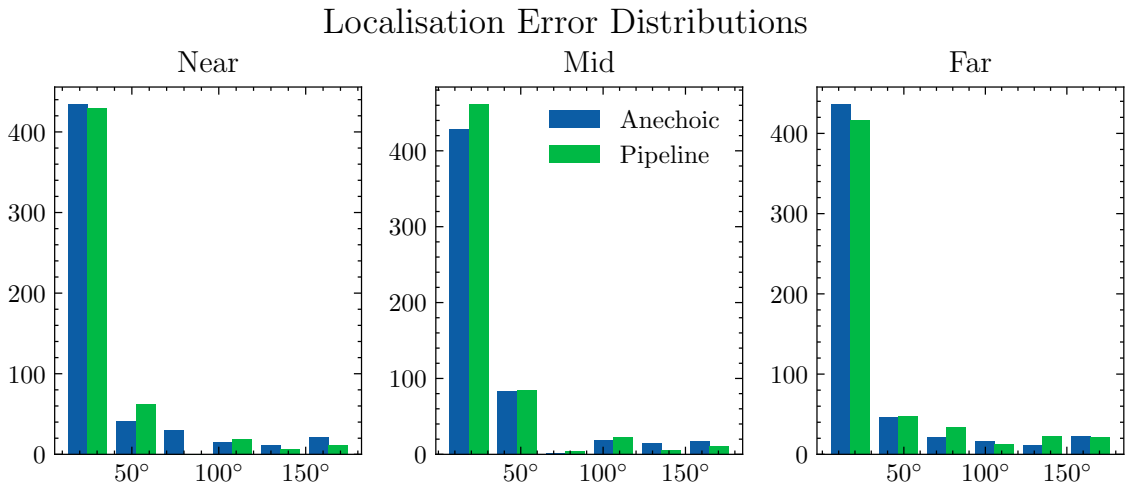


Figure 7.12: Distributions of angular errors sampled from the localisation task, shown across the distance between listener and array of sound sources. The histograms compare samples from the set of anechoic audio stimuli and samples from the set of stimuli generated with the deployed geometrical acoustics-based rendering pipeline. Both sample sets have stimuli conveyed to participants using individualised HRTFs and maintain the same spatialisation apparatus.

all angular distances between sources in each pair.

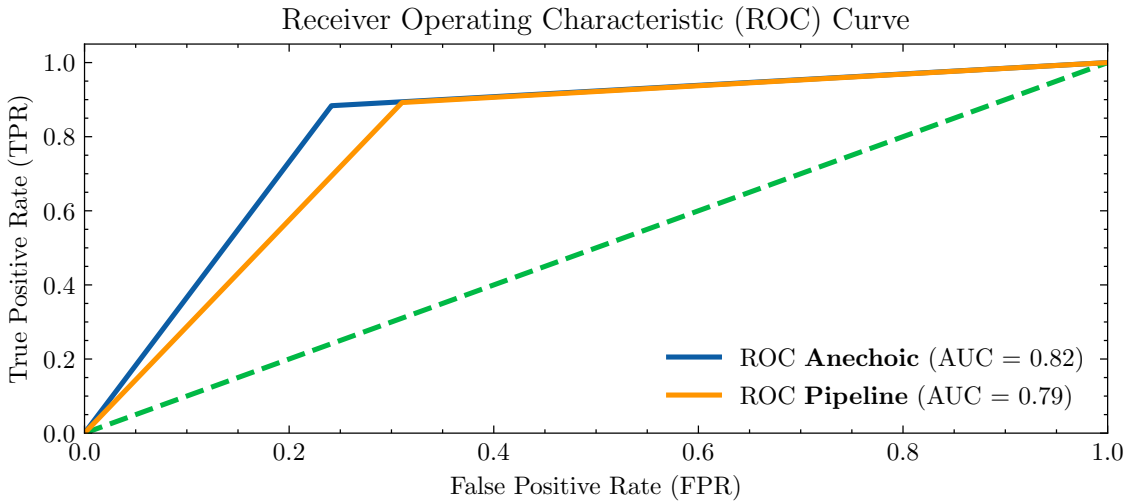


Figure 7.13: Receiver Operating Characteristic (ROC) curves showing the success rate and false positive of responses provided by participants indicating whether perceived audio stimuli are emitted by a singular or multiple invisible sound-emitting objects placed in the testing scene.

7.4 Discussion

The overarching goal of this research is to investigate whether approximated acoustic phenomena applied to auditory interactions in AR affect task performance for activities related to psychoacoustic abilities performed naturally by the Human Hearing System. The comparison of distributions with a non-parametric test shows a significant difference between responses associated with the anechoic set S_A and the pipeline set S_P, as presented in

Table 7.2: F_1 score evaluated over responses gathered from the clustering test, expressing participants' accuracy in discriminating single and multiple invisible sound-emitting objects at increasing angular distances. Scores are computed over the set of anechoic audio stimuli S_A; audio stimuli generated from the pipeline S_P and their aggregate.

Angle	F_1 Anechoic	F_1 Pipeline	F_1 Aggregate
143°	1.0	1.0	1.0
125°	1.0	1.0	1.0
106°	1.0	1.0	1.0
88°	0.97	1.0	0.98
69°	1.0	1.0	1.0
51°	1.0	0.72	0.86
33°	0.59	0.9	0.74
14°	0.52	0.52	0.52
0°	0.76	0.69	0.72

Section 7.3.1. The approximated acoustic reverberation convolved to unpropagated audio computed by a ray tracer is sufficient to affect the perception of spatial resolution of sound-emitting objects in AR. However, comparing means of angular error distributions across distances from the listener shows that as the distance from the sound source increases, this effect hinders the spatial resolution abilities of subjects.

The localisation accuracy drops as the distance increases, as shown in Table 7.1, with the angular error favouring the anechoic stimulus set for far sound sources. The significance of the results gathered from the localisation task aligns with Rungta et al. (2016)'s work, demonstrating that physically accurate acoustic simulations have a perceptual impact on tasks in virtual environments. For the space in question, approximations of acoustic phenomena, computed with only $3 * 10^3$ rays and fourth-order reflections, affect task performance in AR for activities related to psychoacoustic abilities, hence determining that stimulus sets produced using the acoustic rendering pipeline affect matching of visual-acoustic information.

The analysis conducted on responses gathered from the clustering task indicates that acoustic responses allow subjects to better resolve clustered sound sources. Figure 7.13 shows better classification performance in subjects utilising the *pipeline* stimulus set. However, this is only true for clustered sources that are farther than 15° apart. The practical implications of this study lie within the ability to design future auditory interactions in AR. When designing tasks in AR involving the application of psychoacoustic abilities, such as pinpointing the position of sound-emitting objects or resolving the nature of concurrent sound sources, acoustic modelling based on a virtual reconstruction of the space can suffice to affect task performance. From this practical implication stems the main theoretical contribution that transfers the established psychoacoustic characterisation of sound propagation effects to AR space, which has unlocked new potential in the realm of audio interactions. However, the theoretical contribution will require a set of complex scenes with diverse architectural and acoustic characteristics, evaluating the generalisability of

the findings discussed in this work.

Despite the limited generalisability, there are important new research directions that branch from these results: mainly the need for an end-to-end pipeline for acoustic rendering in AR based on spatial understanding technology supplied by HMDs and the relationship between auditory interactions in AR and acoustic phenomena modelled from spatial understanding. The latter could have a potential impact on improving task performance in reverberant spaces since the data gathered shows that distance and reverberant sources can impact localisation, future research would explore whether physical sound sources can be affected by cross-talk cancellation techniques to remove wet components from signals, addressing such impact wherever negative. Such techniques could be based on real-time modelling techniques, such as neural networks for IR generation or acoustic novel view synthesis recently pioneered by Ratnarajah et al. (2022) and Chen et al. (2023a), to improve clarity and definition of auditory interactions with inverse rendering approaches.

7.4.1 Significance

The results demonstrate that the proposed audio rendering pipeline enhances localisation resolution and allows better resolution of masking effects in audio stimuli within MR environments. Due to the ability of audio stimuli of conveying scene information, this can have an effect on the wider domains that leverage spatial audio in immersive environments (Yang, Barde and Billinghurst, 2022). The ability to accurately localise sound sources within an MR environment is crucial for many applications. Improved spatial awareness can enhance user interactions in training simulations for medical and military applications, where precise sound localisation can be critical for success (Harris et al., 2023). For instance, in medical training, the ability to pinpoint the origin of alarms or equipment sounds can improve response times and decision-making. Enhanced audio realism contributes to a more immersive and natural user experience. In the entertainment and gaming industries, where user engagement is paramount, realistic audio cues that correspond accurately to visual elements can significantly boost the overall experience (Slater et al., 2009).

7.5 Conclusions

This research has explored the psychoacoustic characterisation of geometrical acoustics-based rendering pipelines in AR, providing a novel lens through which the auditory interactions in AR environments can be understood and enhanced. The designed study revealed the criticality of integrating fundamental acoustic principles to augment the realism and immersive quality of auditory experiences in AR environments — illuminating the necessity in the proposed pipeline to deliver more correct human responses to acoustic tasks within AR contexts.

The research involved human participants to enable us to decipher the efficacy of sound

rendering techniques in AR, particularly focusing on their capacity to emulate psychoacoustic abilities. Through the experimental design, the study illuminated the nuances of sound transmission in AR contexts, providing a large dataset enabling comparative analysis of performance in psychoacoustic tasks, localisation and clustering.

The contributions of this work were threefold: firstly, it introduced a pioneering methodology for studying psychoacoustic factors within simulated sound fields in AR; secondly, it developed a bespoke testing framework, employing a custom audio engine and a prototype acoustic rendering pipeline for an AR context, facilitating AR evaluation; and thirdly, it provided insightful recommendations for the future design of audio rendering pipelines in immersive technology, derived from a comprehensive dataset of perceptual responses.

The findings underscored the significant impact of acoustic approximations and spatial understanding in enhancing auditory interactions in AR, revealing that the convolution of approximated acoustic phenomena to audio stimuli can notably influence psychoacoustic abilities and potentially modulate task performance in AR applications. This research not only paves the way for further exploration into sound rendering in AR but also establishes a foundational framework for developing more immersive and acoustically precise or beneficial AR applications across various domains.

The experimental evaluation explores the first steps towards the question concerning subjective factors affected by the deployment of a context-aware acoustic rendering pipeline, showing significant differences in stimuli considering approximated acoustic phenomena against unpropagated audio. Future work may seek to explore the application and impact of these findings across diverse AR platforms and user demographics, thereby enriching our understanding and development of acoustically immersive and perceptually coherent AR environments. Important next steps towards defining the relationship between vision-based acoustic pipelines and auditory interactions in AR should aim at defining the requirements from both the acoustic rendering and the acoustic characteristic retrieval, performing ablation studies on both components to evaluate Just-Noticeable differences across acoustic rendering resolution levels or to evaluate the required granularity of acoustic characteristics to be attributed to the scene geometry.

Chapter 8

Conclusion

The primary goal of this thesis has been to explore a novel system for acoustic rendering bespoke to MR platforms, enabling context-aware rendering to consider characteristics of the environment surrounding the player. The objectives defined by this thesis developed and evaluated individual systems composing the pipeline proposed as a solution modularised into standalone contributions. Overall, the development of components evaluations conducted on test cases showed that the domain of acoustic rendering for MR would benefit from applications of acoustic rendering pipelines, and the discussions on the results gathered open several potential avenues for future research heading towards more interactive and realistic auditory interactions.

8.1 Summary of Contributions

The following Sections will reflect on the objectives posed by the introductory chapter of this work (Section 1.4), highlighting significant results by connecting findings gathered from studies conducted throughout the development of the pipeline. For contextualisation, clarity, and accessibility, Table 8.1 summarises the thesis aims and how Chapters contribute to each objective.

The contributions of this thesis advance the state of the art by integrating computer vision techniques into sound rendering systems, specifically by inferring acoustic data associated with materials and boundaries within complex environments. This novel approach allows for dynamic adaptation of sound rendering systems to unseen environments, which has implications across several key applications of Audio Augmented Reality (AAR). For instance, in the context of architectural acoustics, the ability to automatically reconstruct auditory spaces from real environments allows to more efficiently create and test acoustic models within MR applications or game engines. This not only streamlines the design process but also enables more precise simulations of how sound will behave in newly designed spaces, ultimately leading to better-informed architectural decisions. As an example, designers could use this technology to simulate the acoustics of concert halls or public spaces

Table 8.1: A summary of the objectives, highlighting connections between the aims of the work and the Chapters of this thesis. Key Performance Indicators (KPIs), expressed in the introductory Chapter are reported.

N	Objective	Contributions	KPIs
1	To review the current state of audio rendering in virtual environments, assessing their limitations, realism and computational requirements to investigate their applications in real-time acoustic simulations for complex scenes.	Chapter 2, 3	Reviewed key papers and identified limitations and expansion points, informing the development of the proposed rendering pipeline.
2	To explore how, in modern approaches to acoustic simulations, visual representations of materials relate to simulated soundfields.	Chapter 3, 4	Developed and tested two material recognition systems. Tested on a real scene and published results.
3	To design and test systems to automatically attribute acoustic materials to scene geometry in virtual environments, recognising and distinguishing between materials in the acoustic and visual domains.	Chapter 4, 5	Integrated the material recognition systems on standard acoustic rendering pipelines. Tested on various scenes and published results.
4	To study and evaluate the application of sound rendering methods with acoustic material recognition.	Chapter 5	Developed bespoke acoustic rendering pipeline to integrate material recognition. Tested objective and subjective metrics and published results.
5	To design and propose a novel pipeline for acoustic rendering applied to augmented reality platforms.	Chapter 6	Defined a vision for acoustic rendering in AR based on data obtained from tests.
6	To investigate psychoacoustic and human factors in auditory displays created by the proposed pipeline by testing an augmented reality prototype.	Chapter 7	Conducted subjective testing on sound rendering in AR and tested the proposed rendering prototype on a real application.

during the planning stages, ensuring optimal sound quality before construction begins.

In the realm of digital tourism, the proposed methods enhance the realism and interactivity of virtual reconstructions of cultural sites. By improving sound rendering pipelines, users can experience more immersive virtual tours where the acoustic environment closely mirrors the original site by accurately simulating the reverberation, and other acoustic phenomena, in characteristics soundscapes, making digital tourism more engaging and educational. The recreation of the "Roaring Drago" by Yokota et al. (2008), a characteristic historical soundfield, using FDTD techniques, is an example of sound rendering applied to cultural heritage aroud this field of research.

For the gaming industry, the integration of realistic audio based on environmental cues elevates the immersive quality of games. The spatialisation of sound, tailored to reflect the unique acoustic characteristics of different virtual environments, enhances players' ability to navigate, locate, and interact within these worlds. Such improvements are not merely aesthetic; they can improve gameplay by providing auditory cues that guide player actions, such as identifying the direction of approaching threats or locating hidden objects, thereby enhancing both the engagement and the competitive aspects of gaming. In industrial training and simulation scenarios, the proposed sound rendering advancements enable trainees to detect and respond to auditory cues that represent critical, but visually hidden, information. For example, in a simulated industrial environment, accurate audio cues can alert trainees to malfunctioning machinery or other hazards that are not immediately visible, thereby improving situational awareness and response times. This capability is crucial for preparing workers for real-world conditions, where auditory information often plays a pivotal role in safety and efficiency. Furthermore, the contributions of this work extend to the emerging Metaverse, where realistic auditory interactions are fundamental to creating immersive and interactive experiences. In this context, the methods developed in this thesis facilitate more realistic virtual collaborations and interactions by ensuring that sound accurately reflects the virtual environment, enhancing user engagement and the overall sense of presence within the Metaverse.

Overall, the significance and contribution to knowledge of this thesis work lies in the proposal of a novel rendering pipeline, expressed by Chapter 6, which illustrates a design of the end-to-end system, targeting the primary objective, illustrating its concept and design principles, demonstrating implementations and apparatus design, and outlining the vision and potential expansions.

Chapter 7 evaluated a prototype deployed to an MR platform, evaluating psychoacoustic factors in auditory displays, demonstrating the validity of the overarching aim by proving that a pipeline for realistic auditory displays has effects on interactions in virtual environments.

8.1.1 Integration of Computer Vision into Sound Rendering

As a high-level discussion around the interpreted results gathered from evaluating the system and its components, there are conclusions drawn that determine contributions towards the field of Audio Augmented Reality (AAR) and domains of computer graphics and vision in the context of simulating soundfields and generating realistic auditory stimuli. The following Sections will elaborate on the significance of the contributions, discussing their relevance to the broader research domains, target use cases, their limitations, and how future work can overcome them.

8.1.2 System Overview

A prototype system for realistic auditory displays in MR, considering basic features of the user's surrounding environment, has shown potential in evoking perceptual responses within applications and tasks in immersive experiences. Extrapolating from research trends in acoustic rendering domains, as shown in Chapter 3, deep learning-based real-time sound rendering is becoming increasingly accessible and feasible to deploy in XR platforms. A summary of related work concerning areas of sound rendering and computer vision was discussed in light of the overarching aim of the thesis, targeting Objectives 1 to 6 in Table 8.1. The modularity of the system proposed to address the thesis objective provides methodologies and evaluations that can apply to future extensions, facilitating the use of experimental sound rendering for wearable computing.

8.1.3 Acoustic Characteristics Retrieval Methods

The solutions proposed to target Objectives 2 and 3 addressed the problem of extracting soundscape features in context-aware acoustic rendering target platforms equipped with space-sensing technology. The two proposed novel systems addressed the problem of retrieving acoustic characteristics from environments from their visual representations by adopting computer vision techniques. This thesis work and recent methods show that such a task is still an open research question due to the complexity in the mappings between visual representations of space and auditory stimuli generated from simulated soundfields based on properties assigned from inferring their visual appearance.

Despite deep learning methods exist to create mappings between semantics and visual characteristics of materials and surfaces (Gaur and Manjunath, 2020; Chen et al., 2022a), there are still open challenges around the areas of recognising and attributing properties to virtual geometry via its visual features. Mapping acoustic properties to representations of space has been defined in this work as the "material tagging" process, expressing the task engineers face when constructing immersive, complex scenes in game engines, where portions of scene geometry are assigned acoustic material properties. As acoustic simulations become more feasible for increasing use cases and applications (Liu and Manocha, 2020), material tagging will incur higher manual authoring, which the application of computer vision techniques can automate, enabling and rendering procedural workflows (Schissler,

Loftin and Manocha, 2017).

There are benefits to the two proposed systems that aim at automating material tagging. With the camera-based acoustic material tagging system demonstrated in Section 4.2, a novel system that uses camera renders to infer portions of scene geometry they depict, provides a solution that can adapt to cameras equipped by HMDs. It allows the system to scan and map acoustic characteristics to the environments as the user walks around their MR scene.

The texture-based alternative system demonstrated in Section 4.3 expanded towards a more efficient workflow that also has applications to HMDs by using spatial reconstruction systems that provide and update scene geometry as the user navigates the scene.

8.1.4 Acoustic Rendering

Geometrical acoustics-based rendering pipelines have the advantage of controlling the number of rays and other aspects associated with the simulated acoustic phenomena. Thus, in order to target Objectives 4 and 5, perceptual evaluations aimed at measuring psychoacoustic factors associated with auditory stimuli produced from simulated soundfields benefit from control over the resolution of the sound propagation technique. The degree of control is crucial for evaluating perceptual responses and defining the minimum requirements for psychoacoustic factors of the generated stimuli.

Chapter 5 provided important insights into the development of sound rendering pipelines targeted at wearable computing platforms, adopting geometrical acoustics techniques. The testing performed on renderer prototypes generated data on the deployment of the renderers on virtual and real real-world scenes, providing crucial discussion points around realism and performance aspects of seleced rendering techniques.

The current state of the field of sound rendering has emerging techniques that range from classic approaches to sound propagation such as geometrical acoustics (Savioja and Svensson, 2015a) or deep learning methods like Singh et al. (2021)'s Image2Reverb. However, there is a lack of literature around developing and testing techniques for HMDs or the wider field of wearable computing platforms. The work towards Objective 5 is a step towards outlining guidelines for the development of sound rendering pipelines for wearable computing platforms, considering limited resolution sound propagation and geometry reduction paradigms.

8.1.5 Objective and Subjective Evaluations

Despite the modularity of the proposed system and the standalone nature of individual components developed, evaluations and preliminary tests to evaluate their performance, efficacy, and human factors require consideration of the entire pipeline, including the final link of the chain, the listener. This final link is essential for determining perceptual responses and conducting ablation studies to evaluate the effect and the priority of individual

components, targeting Objectives 5 and 6.

Of significant impact here are the methodologies in Chapters 4 and 5, where cutting-edge perceptual encoding techniques are leveraged to optimise preliminary testing of perceptual factors of systems for retrieving acoustic features from virtual environments. For instance, employing a deep perceptual audio similarity metric allows fast preliminary testing of the sound rendering pipeline, eliminating the need to recruit human subjects.

8.2 Limitations

Despite the contributions achieved through this thesis work in the integration of sound rendering with computer vision, it is important to acknowledge the inherent limitations that need to be considered when interpreting findings. This section elaborates on specific limitations relative to the proposed system, and the experiments conducted range from the scalability of our algorithms in real-world applications to the generalisability of our results across different environments and scenarios. By openly discussing these aspects, we aim to provide a comprehensive understanding of the context within which our findings should be interpreted, thereby enriching the discourse and encouraging subsequent innovation in the field.

8.2.1 Acoustic Material Tagging

A limitation of the acoustic material tagging system arises from the methodology used for acoustic validation. Specifically, the testing protocols did not include comparisons against measurements obtained from impedance tubes, which are standard in evaluating the acoustic properties of materials. Impedance tubes offer precise data on how materials absorb and reflect sound, providing a critical benchmark for the accuracy of any sound rendering system. The absence of this comparison means our system's effectiveness in accurately rendering material-based acoustic responses might not fully align with real-world acoustic properties, potentially impacting the system's reliability in applications requiring high fidelity in sound simulation.

The two presented acoustic material tagging methods primarily utilised a single sound rendering method to evaluate the system's performance in recognising different materials. This limits the understanding of how the system might perform across the diverse spectrum of sound rendering techniques. Sound rendering methods, each with unique strengths and handling of acoustic phenomena, could significantly influence the accuracy and efficiency of material recognition required. Expanding the testing to include a broader range of sound rendering methods would provide a more comprehensive assessment of the system's versatility and applicability in various simulation contexts.

The choice of deep learning architecture plays a pivotal role in the system's ability to recognise materials based on their acoustic signatures. Exploring alternative deep learning

models, like self-supervised or unsupervised approaches (Gaur and Manjunath, 2020), could provide more suited systems for capturing the subtle nuances of surface characteristics, enhancing the overall accuracy and adaptability of the tagging process. The training dataset used for the two systems plays a critical role in the development of any machine learning system. The texture-based acoustic material tagging system was trained on the OpenSurfaces (Bell et al., 2013) dataset, which, while extensive, does not encompass the full variety of materials encountered in real-world or high-fidelity virtual environments. This limitation could hinder the system's ability to generalise across unseen materials or unique surface characteristics not represented in the dataset. Enriching the training dataset with a broader and more diverse collection of surface types and material properties would enhance the robustness and accuracy, enabling it to better adapt to a wider range of virtual simulation scenarios.

8.2.2 Rendering Pipelines for Augmented Acoustics

The architecture of the sound rendering system primarily relies on principles of geometrical acoustics, which, while providing a solid foundation for simulating sound propagation, introduces certain limitations. Geometrical acoustics often simplifies wave behaviors, neglecting complex phenomena such as diffraction and the nuanced interaction of sound with irregular surfaces. This simplification can lead to less accurate rendering in environments where these effects are significant, potentially affecting the immersive quality and realism in augmented reality applications, despite comparisons between real and virtual soundfields revealing acceptable error margins.

The efficacy of the sound rendering system in dynamic geometry scenarios requires further investigations around real-time uses by deploying an online sound renderer to a prototype. Furthermore, dynamic geometry, involving moving objects and changing environments, presents unique challenges for sound rendering, necessitating continuous data updates for accurate spatial audio. Testing an online sound renderer in MR would address the responsiveness and adaptability in live dynamic environments, highlighting a crucial area for future development.

The testing conducted has predominantly focused on static sound sources, thereby not fully exploring the complexities introduced by non-static or moving sources due to moving sound sources requiring the deployment of a tested online sound propagation technique on a wearable computing platform These scenarios are critical for augmented reality, where sound sources often change position relative to the user. Additional testing is required to evaluate the system's ability to accurately render sound phenomena such as occlusion and reflection based on the dynamic positioning of sources and the geometry of the surrounding environment. Addressing this limitation is essential for achieving higher complexity in auditory interactions.

To date, the sound rendering system has not extensively explored the potential of deep

learning-based approaches, which could offer significant advancements in the accuracy and realism of audio simulations. Deep learning algorithms have the potential to model complex acoustic phenomena more effectively than traditional methods, learning from vast datasets to predict sound behavior in varied environments. Further development and testing of these approaches are required to ascertain their feasibility and effectiveness in augmenting or replacing current sound rendering techniques, potentially leading to breakthroughs in how sound is simulated and experienced in augmented reality platforms.

8.2.3 Psychoacoustic Evaluation

A primary limitation of the investigation into the psychoacoustic effects of the sound rendering system is its restriction to data collected from a single environmental space. This constraint significantly impacts the generalisability of our findings, as acoustic characteristics and listener perceptions can vary widely across different spaces due to variations in size, geometry, and materials.

Future studies should aim to address generalisability and include a more diverse array of spaces, encompassing a range of architectural designs and acoustic properties. The use of a single sound rendering architecture limits the understanding of how different rendering approaches affect multimodal perception. Sound rendering systems vary in inputs/outputs and perceptual factors, each potentially offering distinct advantages or drawbacks in simulating accurate and believable soundscapes. Expanding future research to include multiple sound rendering architectures would provide a more comprehensive understanding of their psychoacoustic implications, allowing for a comparative analysis that could develop more effective and immersive audio solutions. A single and controlled environment was used due to the need for accessing the physical space to obtain ground truth IRs.

The current investigation has explored a minimal set of psychoacoustic abilities, leaving a spectrum of human auditory processing capabilities unexamined. Psychoacoustic phenomena such as sound localisation accuracy, auditory scene analysis, and the perception of sound quality under different conditions are critical for evaluating the effectiveness of sound rendering systems. Future investigations should aim to assess a broader range of psychoacoustic abilities to gain a deeper insight into the system's impact on the auditory experience and to identify areas where enhancements are needed to achieve a more realistic and engaging auditory simulation.

The study predominantly utilised static sound sources, neglecting the potential psychoacoustic effects introduced by dynamic and moving sound sources interacting with environmental geometry, such as occlusion. Sound sources' movement and interaction with the listener's environment can significantly affect spatial audio perception, offering a more complex and realistic auditory experience. Investigating the psychoacoustic effects with dynamic sound sources that exploit environmental geometry would provide insights into

more realistic audio rendering and enhance the system's applicability to real-world scenarios with non-static sound sources.

8.3 Future Work

While this research has taken steps towards integrating sound rendering with computer vision to enhance immersive experiences, it has also uncovered numerous pathways for future investigations. The evolving landscape of technology and the increasing need for multimodal applications in XR domains present both challenges and opportunities for advancing the state of rendering domains. Future work should aim to address the limitations encountered in the current study, considering scalability, real-time processing demands, and the incorporation of more complex environmental interactions. Specific areas where future research could not only extend the findings of this study but also open new avenues for innovation in sound rendering and computer vision are outlined here, aiming at the next generation of immersive technologies.

8.3.1 Visual-Acoustic Mappings

Future expansion points to the acoustic material tagging systems developed as part of the studies illustrated in Chapter 4 include experimenting with the use of unsupervised methods for material recognition tasks. Unsupervised approaches would constitute an alternative method to the two systems provided and allow improved workflows around training a model to recognise an ecosystem of acoustic materials given a set of virtual scenes. Unsupervised networks can create latent representations of acoustic materials as clusters of visual features extracted from the given set of scenes. The contribution of this expansion point would lie in the mappings between classes learned by extracting visual features via the unsupervised network and acoustic characteristics attributed to each class. Artistic control would be possible as a byproduct of this approach, allowing the network to specialise around selected environments and provide the ability to remap learned visual material classes to personalised acoustic characteristics. This would feed into procedural content creation, providing artists and worldbuilders with control over procedurally generated soundfields, given a set of representative environments. An extension building from Singh et al. (2021)'s would be an improved network considering listener and speaker position information in the mapping between RGB-D input and an approximated RIR. Instead, the fast image-to-image network would learn to manipulate existing IRs by using depth maps. The system would approximate how features of the environment expressed by the depth affect the response based on the listener and source position information encoded in the depth map data. The network would approximate propagation effects by leveraging systems like Chen et al. (2022b)'s, learning how an IR is affected by the positioning of source and receiver and scene geometry. In the light domain, this is demonstrated by Wang et al. (2020) and Gardner et al. (2019)'s methods, which estimate and manipulate light propagation effects depicted in RGB input image data, allowing rendering scenes with little knowledge of the

virtual environment. This system would optimise the need for online acoustic rendering in MR, as a generic RIR provided as input to the system would suffice for auralisations of multiple source-receiver position pairs.

8.3.2 Acoustic Material Tagging

By addressing the limiting factors of the acoustic material tagging systems discussed, ablation studies of deep learning-based systems would provide insights into the minimum requirements for a network to classify surfaces in complex scenes successfully. Future work can expand this work and feed into cross-domain applications by intersecting elements of the proposed pipeline with reinforcement learning and robotics. This thesis work could be applied to autonomous audio-visual navigation in 3D environments, where agents are trained to learn to perceive and respond to audio-visual entities with the goal of navigating to a specific target. This demonstrates that a robust approach to retrieving acoustic information from virtual environments can benefit audiovisual-based navigation applications. Generalisable and robust acoustic material systems have a direct impact on training models in differentiating and resolving features of soundscapes thanks to the complex mapping between visual features of surfaces in virtual environments and acoustic features in audio from simulated soundscapes.

Reinforcement learning models could also open avenues towards novel systems for acoustic material tagging. With the increase in multimodal autonomous navigation systems, like the work from Chen et al. (2022b), RL agents can interact with the environment based upon auditory stimuli generated via acoustic rendering. Future work could explore the use of RL agents to pain surfaces with acoustic material, optimising aspects of the soundfield. This work would contribute to architectural applications, where control over aspects of the soundfield is required as part of the building planning process, as discussed by Vorländer et al. (2015).

8.3.3 Dynamic Geometry Handling for Augmented Audio Reality

As discussed in Chapter 6, the environment mesh reconstructed by the HMD may not always accurately match the physical space, leading to errors in the simulated soundfield and the resulting auditory stimuli. To address these challenges, future work could focus on the development and integration of mesh completion algorithms within the sound rendering pipeline. These algorithms would be designed to infer and reconstruct the unseen portions of the scene, enabling the system to generate a more complete and accurate environment mesh even before the user has fully explored the space. By predicting the likely structure and materials of the missing areas, such techniques could significantly reduce errors in soundfield simulation, leading to more precise and realistic auditory interactions. In particular, the application of advanced computer vision techniques, such as those demonstrated by Cao and De Charette (2022), which predict 3D scene geometry from a single monocular RGB image, could address the issue of reconstructing 3D environments from

limited information, improving the accuracy of the input model for sound propagation. Future research could explore how these techniques could be leveraged to enhance sound rendering in augmented reality, particularly in predicting unexplored areas of the physical space around the user.

8.3.4 Perceptual Response Analysis

Chapter 5 conducts extensive testing on the acoustic rendering pipeline employed by the MR prototype and analyses objective and subjective factors of simulated soundfields. However, some unexplored avenues could expand the testing conducted, providing generalised guidelines for future sound rendering methods. A crucial expansion point here is the definition of a minimum input to a rendering pipeline to obtain a perceptually valid response — an investigation that could be performed both on GA and deep learning-based acoustic rendering systems.

Future work should analyse the minimum simulation resolution required for GA pipelines, such as the number of rays or the order, as well as study the disparity between the reconstructed environment and the real space with respect to perceptual factors. The study would provide insights into the margins of error before the listener starts noticing discrepancies in perceived realism evoked by the simulated auditory display.

8.3.5 Further Psychoacoustic Analysis

The significance of the findings obtained from conducting the psychoacoustic study illustrated by Chapter 7 has contributed towards creating guidelines for designing realistic interactions in MR. When designing tasks in MR involving the application of psychoacoustic abilities, such as pinpointing the position of sound-emitting objects or resolving the nature of concurrent sound sources, acoustic modelling based on a virtual reconstruction of the space can suffice to affect task performance. From this practical implication stems the main theoretical contribution that transfers the established psychoacoustic characterisation of sound propagation effects to MR space, which has unlocked new potential in the realm of audio interactions. However, the theoretical contribution will require a set of complex scenes with diverse architectural and acoustic characteristics, evaluating the generalisability of the findings discussed in this work. Future work could benefit from the application of the testing methodology to a generalisable set of soundfields, including environments from high absorption and controlled acoustics, such as audio production rooms, to reverberant spaces like churches.

8.4 Conclusions

8.4.1 Significance of Contributions

One of the primary benefits of improved sound rendering in MR through the proposed rendering pipeline is the enhancement of task performance. Accurate and realistic auditory

information enables users to better interact with the immersive scene, leveraging sound cues to gain insights and make informed decisions (Rubio-Tamayo, Gertrudix Barrio and García García, 2017; Zimmons and Panter, 2003). The integration of context-aware and spatially accurate audio allows users to perceive their environment more naturally, similar to how they would in the real world. This auditory augmentation supports multitasking, increases situational awareness, and enhances the ability to locate and identify objects within the MR environment.

Impact on Modern MR Use Cases

Examples of the impact of the thesis work on modern MR applications include training and serious games where users can perform tasks. Generally, users can benefit from auditory information to guide actions and obtain feedback on task performance. Realistic spatialised audio generated via the proposed pipeline allows, additionally, the ability to discover visually hidden entities through means of psychoacoustics; e.g. via sound localisation. This enhances task performance to respond quickly and accurately to various situations, mirroring real-world conditions where sound-emitting entities play a critical role in situational awareness. This is essential in industrial applications (Machała, Chamier-Gliszczyński and Królikowski, 2022) or training (Soltani and Morice, 2020).

Some examples of this are in remote working environments, for instance, where team members are geographically dispersed, enhanced audio rendering can facilitate more natural and effective communication. Spatial audio can help users identify the direction and distance of users voices, making virtual meetings and collaborative work sessions more intuitive and engaging. Or in colocated working scenarios, where individuals share a physical space augmented with digital information, accurate sound rendering can support better coordination and interaction with shared virtual objects and information (Park and Kim, 2022; Al-Ghaili et al., 2022).

In military and emergency response training, sound rendering can simulate realistic battlefield or disaster environments, where trainees must rely on auditory information to detect threats, locate teammates, or navigate through complex terrains. The ability to hear footsteps, weapon discharges, or environmental sounds can significantly enhance the realism and effectiveness of the training, complementing the lack or absence of visual information (Harris et al., 2023; Ahir et al., 2020).

In cultural heritage applications, XR provides a platform to reconstruct historical soundscapes that are lost or decayed through time and damage. As discussed by Chapter 6, MR HMD provide technology to reconstruct physical space and project reconstructed and virtual geomtry to the user. Here, sound rendering allows historical changes applied to the virtual geometry to reconstruct lost soundscapes and provide users with experiences that reflect the orginal conditions of the historical site (Katz, Murphy and Farina, 2020; Schofield et al., 2018).

8.4.2 Conclusive Remarks

This thesis work builds on decades of advances toward systems for realistic auditory interactions for immersive media, adopting classic techniques for simulating soundfields and leveraging recent works within the domains of graphics and vision to optimise and automate several tasks associated with rendering believable auditory displays.

The primary goal of researching and proposing a novel system for MR platforms was achieved, grounding its design choices on data collected from a series of experimental tests. Overall, the findings not only advance the understanding of sound rendering integrated with computer vision but also lay the groundwork for significant advances toward realistic auditory interactions in immersive virtual environments. This has implications not only for the entertainment industry, where enhanced immersive experiences can greatly elevate user engagement, but also for educational and training simulations, where realism can significantly improve learning outcomes and user task performance.

A conclusive remark that can serve the reader food for thought is the unexplored potential that can be leveraged by combining machine vision with acoustic rendering applied to immersive platforms. Finally, the potential for these technologies to assist visually impaired individuals by providing more intuitive and context-aware navigational aids represents a meaningful step towards technology that provides context-aware sonic interactions and realistic soundscapes.

Bibliography

3382-1, B., 2009. Iso 3382-1: Acoustics - measurement of room acoustic parameters. Available from: https://www.iso.org/standard/40979.html.

Abeßer, J., 2020. A review of deep learning based methods for acoustic scene classification. *Applied sciences*, 10(6), p.2020.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. *Ieee transactions on pattern analysis and machine intelligence*, 34(11), pp.2274–2282. Available from: https://doi.org/10.1109/TPAMI.2012.120.

Adavanne, S., Politis, A. and Virtanen, T., 2018. TUT Sound Events 2018 - Circular array, Anechoic and Synthetic Impulse Response Dataset. Available from: https://doi.org/10.5281/zenodo.1237752.

Ahir, K., Govani, K., Gajera, R. and Shah, M., 2020. Application on virtual reality for enhanced education learning, military training and sports. *Augmented human research*, 5, pp.1–9.

Ahn, B., Yang, K., Hamilton, B., Sheaffer, J., Ranjan, A., Sarabia, M., Tuzel, O. and Chang, J.H.R., 2023. Novel-view acoustic synthesis from 3d reconstructed rooms. *arxiv preprint arxiv:2310.15130*.

Al-Ghaili, A.M., Kasim, H., Al-Hada, N.M., Hassan, Z.B., Othman, M., Tharik, J.H., Kasmani, R.M. and Shayea, I., 2022. A review of metaverse's definitions, architecture, applications, challenges, issues, solutions, and future trends. *Ieee access*, 10, pp.125835–125866.

Allen, J.B. and Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. *The journal of the acoustical society of america*, 65(4), pp.943–950.

Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.3686–3693.

Arce, T., Fuchs, H. and McMullen, K., 2017. The effects of 3d audio on hologram localization in augmented reality environments. *Proceedings of the human factors and ergonomics society annual meeting*, 61(1), pp.2115–2119. https://doi.org/10.1177/1541931213602010, Available from: https://doi.org/10.1177/1541931213602010.

Arvidsson, E., Nilsson, E., Bard-Hagberg, D. and Karlsson, O.J., 2021. Subjective experience of speech depending on the acoustic treatment in an ordinary room. *International journal of environmental research and public health*, 18(23), p.12274.

Audio, S., 2020. Git repository steam audio. *Url: https://valvesoftware. github. io/steamaudio/downloads. html,[accessed 2019, february 27]*, 1(1), p.1.

Bajuelos, A.L., Canales, S., Hernández, G. and Martins, A.M., 2008. Optimizing the minimum vertex

guard set on simple polygons via a genetic algorithm. *Wseas transactions in information science and applications*, 5(11), pp.1584–1596.

Ballou, G., 2013. *Handbook for sound engineers*. Taylor & Francis.

Bansal, A., Chen, X., Russell, B., Gupta, A. and Ramanan, D., 2016. Pixelnet: Towards a general pixel-level architecture. *arxiv preprint arxiv:1609.06694*.

Bauer, Z., Li, Z., Orts-Escolano, S., Cazorla, M., Pollefeys, M. and Oswald, M.R., 2021. Nvs-monodepth: Improving monocular depth prediction with novel view synthesis. *2021 international conference on 3d vision (3dv)*. IEEE, pp.848–858.

Bell, S., Upchurch, P., Snavely, N. and Bala, K., 2013. Opensurfaces: A richly annotated catalog of surface appearance. *Acm transactions on graphics (tog)*, 32(4), pp.1–17.

Beranek, L.L., 2006. Analysis of sabine and eyring equations and their application to concert hall audience and chair absorption. *The journal of the acoustical society of america*, 120(3), pp.1399–1410.

Berardi, U., Iannace, G. and Ianniello, C., 2016. Acoustic intervention in a cultural heritage: The chapel of the royal palace in caserta, italy. *Buildings*, 6(1), p.1.

Bertet, S., Daniel, J., Parizet, E. and Warusfel, O., 2013. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta acustica united with acustica*, 99(4), pp.642–657.

Blauert, J., 1997. *Spatial hearing: the psychophysics of human sound localization*. MIT press.

Bollmer, G. and Suddarth, A., 2022. Embodied parallelism and immersion in virtual reality gaming. *Convergence*, 28(2), pp.579–594.

Bonneel, N., Suied, C., Viaud-Delmon, I. and Drettakis, G., 2010. Bimodal perception of audio-visual material properties for virtual environments. *Acm transactions on applied perception (tap)*, 7(1), pp.1–16.

Bork, F., Fuers, B., Schneider, A.K., Pinto, F., Graumann, C. and Navab, N., 2015. Auditory and visio-temporal distance coding for 3-dimensional perception in medical augmented reality. *2015 ieee ismar*. IEEE, pp.7–12.

Bosse, S., Maniry, D., Müller, K.R., Wiegand, T. and Samek, W., 2017. Deep neural networks for no-reference and full-reference image quality assessment. *Ieee transactions on image processing*, 27(1), pp.206–219.

Bourke, P., 1994. Polygonising a scalar field.

Büschel, W., Lehmann, A. and Dachselt, R., 2021. Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. *Proceedings of the 2021 chi conference on human factors in computing systems*. pp.1–15.

Cao, A.Q. and De Charette, R., 2022. Monoscene: Monocular 3d semantic scene completion. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.3991–4001.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y., 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arxiv preprint arxiv:1709.06158*.

Chen, C., Al-Halah, Z. and Grauman, K., 2021. Semantic audio-visual navigation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.15516–15525.

Chen, C., Gao, R., Calamia, P. and Grauman, K., 2022a. Visual acoustic matching. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.18858–18868.

Chen, C., Richard, A., Shapovalov, R., Ithapu, V.K., Neverova, N., Grauman, K. and Vedaldi, A., 2023a. Novel-view acoustic synthesis. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.6409–6419.

Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P. and Grauman, K., 2022b. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in neural information processing systems*, 35, pp.8896–8911.

Chen, C., Sun, W., Harwath, D. and Grauman, K., 2023b. Learning audio-visual dereverberation. *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp.1–5.

Chen, L., Tang, W., John, N., Wan, T.R. and Zhang, J.J., 2018. Context-aware mixed reality: A framework for ubiquitous interaction. *arxiv preprint arxiv:1803.05541*.

Chen, M., Su, K. and Shlizerman, E., 2023. Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. *Proceedings of the ieee/cvf international conference on computer vision*. pp.7853–7862.

Chen, X., Golovinskiy, A. and Funkhouser, T., 2009. A benchmark for 3d mesh segmentation. *Acm transactions on graphics (tog)*, 28(3), pp.1–12.

Colombo, M., Dolhasz, A., Hockman, J. and Harvey, C., 2021. Psychometric mapping of audio features to perceived physical characteristics of virtual objects. *2021 ieee conference on games (cog)*. IEEE, pp.1–4.

Cowan, B. and Kapralos, B., 2010. Gpu-based real-time acoustical occlusion modeling. *Virtual reality*, 14, pp.183–196.

Davison, 2003. Real-time simultaneous localisation and mapping with a single camera. *Proceedings ninth ieee international conference on computer vision*. IEEE, pp.1403–1410.

de Giorgio, A., Monetti, F.M., Maffei, A., Romero, M. and Wang, L., 2023. Adopting extended reality? a systematic review of manufacturing training and teaching applications. *Journal of manufacturing systems*, 71, pp.645–663. Available from: https://doi.org/https://doi.org/10.1016/j.jmsy.2023.10.016.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 ieee conference on computer vision and pattern recognition*. Ieee, pp.248–255.

Devadoss, S.L. and O'Rourke, J., 2011. *Discrete and computational geometry*. Princeton University Press.

Diaz-Guerra, D., Miguel, A. and Beltran, J.R., 2021. gpurir: A python library for room impulse response simulation with gpu acceleration. *Multimedia tools and applications*, 80(4), pp.5653–5671.

Dickinson, P., Cardwell, A., Parke, A., Gerling, K. and Murray, J., 2021. Diegetic tool management in a virtual reality training simulation. *2021 ieee virtual reality and 3d user interfaces (vr)*. IEEE, pp.131–139.

Ding, L. and Goshtasby, A., 2001. On the canny edge detector. *Pattern recognition*, 34(3), pp.721–725.

Dolhasz, A., 2021. *Perceptually-based modelling for image composite harmonisation*. Ph.D. thesis. Birmingham City University.

Dolhasz, A., Harvey, C. and Williams, I., 2020. Learning to observe: Approximating human perceptual thresholds for detection of suprathreshold image transformations. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*.

Doukakis, E., Debattista, K., Bashford-Rogers, T., Dhokia, A., Asadipour, A., Chalmers, A. and Harvey, C., 2019. Audio-visual-olfactory resource allocation for tri-modal virtual environments. *Ieee transactions on visualization and computer graphics*, 25(5), pp.1865–1875.

Dourado, A., Campos, T.E. de, Kim, H. and Hilton, A., 2019. Edgenet: Semantic scene completion from rgb-d images. *arxiv preprint arxiv:1908.02893*.

Eckhardt, E.A., 1923. The acoustics of rooms. reverberations. *Journal of the franklin institute*, 195(6), pp.799–814.

Farina, A., 2007. Advancements in impulse response measurements by sine sweeps. *Audio engineering society convention 122*. Audio Engineering Society.

Feng, M., Zhang, L., Lin, X., Gilani, S.Z. and Mian, A., 2020. Point attention network for semantic segmentation of 3d point clouds. *Pattern recognition*, 107, p.107446. Available from: https://doi.org/https://doi.org/10.1016/j.patcog.2020.107446.

Frank, M., Zotter, F. and Sontacchi, A., 2015. Producing 3d audio in ambisonics. *Audio engineering society conference: 57th international conference: The future of audio entertainment technology–cinema, television and the internet*. Audio Engineering Society.

Fuchs, H., Kedem, Z.M. and Naylor, B.F., 1980. On visible surface generation by a priori tree structures. *Proceedings of the 7th annual conference on computer graphics and interactive techniques*. New York, NY, USA: Association for Computing Machinery, SIGGRAPH '80, p.124–133. Available from: https://doi.org/10.1145/800250.807481.

Funkhouser, T., Tsingos, N., Carlbom, I., Elko, G., Sondhi, M., West, J.E., Pingali, G., Min, P. and Ngan, A., 2004. A beam tracing method for interactive architectural acoustics. *The journal of the acoustical society of america*, 115(2), pp.739–756.

Funkhouser, T., Tsingos, N. and Jot, J.M., 2003. Survey of Methods for Modeling Sound Propagation in Interactive Virtual Environment Systems. *Presence: Teleoperators and Virtual Environments*, pp.–. No note. Available from: https://inria.hal.science/inria-00606737.

Gan, C., Zhang, Y., Wu, J., Gong, B. and Tenenbaum, J.B., 2020. Look, listen, and act: Towards audio-visual embodied navigation. *2020 ieee international conference on robotics and automation (icra)*. IEEE, pp.9701–9707.

Gardner, M.A., Hold-Geoffroy, Y., Sunkavalli, K., Gagné, C. and Lalonde, J.F., 2019. Deep parametric indoor lighting estimation. *Proceedings of the ieee/cvf international conference on computer vision*. pp.7175–7183.

Gaur, U. and Manjunath, B.S., 2020. Superpixel embedding network. *Ieee transactions on image processing*, 29, pp.3199–3212. Available from: https://doi.org/10.1109/TIP.2019.2957937.

Giordano, B.L. and McAdams, S., 2006. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The journal of the acoustical society of america*, 119(2), pp.1171–1181.

González-Toledo, D., Molina-Tanco, L., Cuevas-Rodriguez, M., Majdak, P. and Reyes-Lecuona, A., 2023. The binaural rendering toolbox. a virtual laboratory for reproducible research in psychoacoustics. *Forum acusticum*.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning*. MIT press.

Gumerov, N.A. and Duraiswami, R., 2021. Fast multipole accelerated boundary element methods for room acoustics. *The journal of the acoustical society of america*, 150(3), pp.1707–1720.

Habets, E.A., 2006. Room impulse response generator. *Technische universiteit eindhoven, tech. rep*, 2(2.4), p.1.

Hacihabiboglu, H., De Sena, E., Cvetkovic, Z., Johnston, J. and Smith III, J.O., 2017. Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. *Ieee signal processing magazine*, 34(3), pp.36–54.

Hamid, F. and Kapralos, B., 2009. A review of spatial sound for virtual environments and games with graphics processing units. *The open virtual reality journal*, 1(1).

Hamilton, B. and Bilbao, S., 2017. Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time. *Ieee/acm transactions on audio, speech, and language processing*, 25(11), pp.2112–2124.

Harris, C.M., 1966. Absorption of sound in air versus humidity and temperature. *The journal of the acoustical society of america*, 40(1), pp.148–159.

Harris, D.J., Arthur, T., Burgh, T. de, Duxbury, M., Lockett-Kirk, R., McBarnett, W. and Vine, S.J., 2023. Assessing expertise using eye tracking in a virtual reality flight simulation. *The international journal of aerospace psychology*, 33(3), pp.153–173. https://doi.org/10.1080/24721840.2023.2195428, Available from: https://doi.org/10.1080/24721840.2023.2195428.

Harris, D.J., Buckingham, G., Wilson, M.R., Brookes, J., Mushtaq, F., Mon-Williams, M. and Vine, S.J., 2020. The effect of a virtual reality environment on gaze behaviour and motor skill learning. *Psychology of sport and exercise*, 50, p.101721.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.770–778.

Heimes, A., Imran, M. and Vorländer, M., 2019. *A real-time virtual reality building acoustic auralization framework for psychoacoustic experiments with contextual and interactive features*. Universitätsbibliothek der RWTH Aachen.

Hoene, C., Patino Mejia, I.C. and Cacerovschi, A., 2017. Mysofa—design your personal hrtf. *Audio engineering society convention 142*. Audio Engineering Society.

Holters, M., Corbach, T. and Zölzer, U., 2009. Impulse response measurement techniques and their applicability in the real world. *Proceedings of the 12th international conference on digital audio effects (dafx-09)*. Italy: DAFX, pp.1–5.

Howard, D. and Angus, J., 2013. *Acoustics and psychoacoustics*. Routledge.

Huang, A., Knierim, P., Chiossi, F., Chuang, L.L. and Welsch, R., 2022. Proxemics for human-agent interaction in augmented reality. *Proceedings of the 2022 chi conference on human factors in computing systems*. pp.1–13.

Hulusic, V., Harvey, C., Debattista, K., Tsingos, N., Walker, S., Howard, D. and Chalmers, A., 2012. Acoustic rendering and auditory–visual cross-modal perception and interaction. *Computer graphics forum*. Wiley Online Library, vol. 31, pp.102–131.

Jillings, N., Moffat, D., De Man, B. and Reiss, J.D., 2015. Web Audio Evaluation Tool: A browser-based listening test environment. *12th sound and music computing conference*.

Jones, M.G. and Stiede, P.E., 1997. Comparison of methods for determining specific acoustic impedance. *The journal of the acoustical society of america*, 101(5), pp.2694–2704.

Jordan, V.L., 1970. Acoustical criteria for auditoriums and their relation to model techniques. *The journal of the acoustical society of america*, 47(2A), pp.408–412.

Jot, J.M., Audfray, R., Hertensteiner, M. and Schmidt, B., 2021. Rendering spatial sound for interoperable experiences in the audio metaverse. *2021 immersive and 3d audio: from architecture to automotive (i3da)*. IEEE, pp.1–15.

Kalogerakis, E., Averkiou, M., Maji, S. and Chaudhuri, S., 2017. 3d shape segmentation with projective convolutional networks. *proceedings of the ieee conference on computer vision and pattern recognition*. pp.3779–3788.

Kashino, M. and Nishida, S., 1998. Adaptation in the processing of interaural time differences revealed by the auditory localization aftereffect. *The journal of the acoustical society of america*, 103(6), pp.3597–3604.

Kates, J.M. and Brandewie, E.J., 2020. Adding air absorption to simulated room acoustic models. *The journal of the acoustical society of america*, 148(5), pp.EL408–EL413.

Katz, B.F., Murphy, D. and Farina, A., 2020. The past has ears (phe): Xr explorations of acoustic spaces as cultural heritage. *International conference on augmented reality, virtual reality and computer graphics*. Springer, pp.91–98.

Ke, F. and Moon, J., 2018. Virtual collaborative gaming as social skills training for high-functioning autistic children. *British journal of educational technology*, 49(4), pp.728–741.

Kiechle, M., Storath, M., Weinmann, A. and Kleinsteuber, M., 2018. Model-based learning of local image features for unsupervised texture segmentation. *Ieee transactions on image processing*, 27(4), pp.1994–2007.

Kim, H., Hernaggi, L., Jackson, P.J. and Hilton, A., 2019. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. *2019 ieee conference on virtual reality and 3d user interfaces (vr)*. IEEE, pp.120–126.

Kim, H., Remaggi, L., Fowler, S., Jackson, P. and Hilton, A., 2020. Acoustic room modelling using 360 stereo cameras. *Ieee transactions on multimedia*, 1, p.1.

Kim, K., Billinghurst, M., Bruder, G., Duh, H.B.L. and Welch, G.F., 2018. Revisiting trends in augmented reality research: A review of the 2nd decade of ismar (2008–2017). *Ieee transactions on visualization and computer graphics*, 24(11), pp.2947–2962.

Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arxiv preprint arxiv:1412.6980*.

Kirkup, S., 2019. The boundary element method in acoustics: A survey. *Applied sciences*, 9(8), p.1642.

Kopta, D., Ize, T., Spjut, J., Brunvand, E., Davis, A. and Kensler, A., 2012. Fast, effective bvh updates for animated scenes. *Proceedings of the acm siggraph symposium on interactive 3d graphics and games*. pp.197–204.

Kuttruff, H., 2016. *Room acoustics*. Crc Press.

Lagunas, M., Malpica, S., Serrano, A., Garces, E., Gutierrez, D. and Masia, B., 2019. A similarity measure for material appearance. *arxiv preprint arxiv:1905.01562*.

Lakka, E., Malamos, A.G., Pavlakis, K.G. and Ware, J.A., 2018. Spatial sound rendering–a survey. *Ijimai*, 5(3), pp.33–45.

Lakka, E., Papadaki, M., Brutzman, D., Puk, R. and Malamos, A.G., 2021. X3d audio graph for the consistent declarative representation of the w3c audio api. *The 26th international conference on 3d web technology*. pp.1–5.

Lauterbach, C., Garland, M., Sengupta, S., Luebke, D. and Manocha, D., 2009. Fast bvh construction on gpus. *Computer graphics forum*. Wiley Online Library, vol. 28, pp.375–384.

Le Bot, A. and Bocquillet, A., 2000. Comparison of an integral equation on energy and the ray-tracing technique in room acoustics. *The journal of the acoustical society of america*, 108(4), pp.1732–1740.

Lee, H., 2011. The relationship between interchannel time and level differences in vertical sound localization and masking. *Audio engineering society convention 131*. Audio Engineering Society.

Lengyel, E., 2019. *Foundations of game engine development: Volume 2: Rendering*. Terathon Software.

Li, D., Langlois, T.R. and Zheng, C., 2018. Scene-aware audio for 360 videos. *Acm transactions on graphics (tog)*, 37(4), pp.1–12.

Liang, S., Huang, C., Tian, Y., Kumar, A. and Xu, C., 2023. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arxiv preprint arxiv:2309.15977*.

Liebetrau, J., Nagel, F., Zacharov, N., Watanabe, K., Colomes, C., Crum, P., Sporer, T. and Mason, A., 2014. Revision of rec. itu-r bs. 1534. *Audio engineering society convention 137*. Audio Engineering Society.

Lima, A.A. de, M. Prego, T. de, Netto, S.L., Lee, B., Said, A., Schafer, R.W., Kalker, T. and Fozunbal, M., 2009. Feature analysis for quality assessment of reverberated speech. *2009 ieee international workshop on multimedia signal processing*. pp.1–5. Available from: https://doi.org/10.1109/MMSP.2009.5293326.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the ieee international conference on computer vision*. pp.2980–2988.

Liu, S. and Manocha, D., 2020. Sound synthesis, propagation, and rendering: A survey. *arxiv preprint arxiv:2011.05538*, 1(1), p.1.

Liu, S. and Manocha, D., 2022. Sound rendering. *Sound synthesis, propagation, and rendering*. Springer, pp.45–52.

Lokki, T. and Grohn, M., 2005. Navigation with auditory cues in a virtual environment. *Ieee multimedia*, 12(2), pp.80–86.

Lokki, T., Savioja, L., Väänänen, R., Huopaniemi, J. and Takala, T., 2002. Creating interactive virtual auditory environments. *Ieee computer graphics and applications*, 1(4), pp.49–57.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the ieee conference on computer vision and pattern recognition*. USA: IEEE, pp.3431–3440.

Luo, A., Du, Y., Tarr, M., Tenenbaum, J., Torralba, A. and Gan, C., 2022. Learning neural acoustic fields. *Advances in neural information processing systems*, 35, pp.3165–3177.

Machała, S., Chamier-Gliszczyński, N. and Królikowski, T., 2022. Application of ar/vr technology in industry 4.0. *Procedia computer science*, 207, pp.2990–2998.

Malpica, S., Serrano, A., Allue, M., Bedia, M.G. and Masia, B., 2020. Crossmodal perception in virtual reality. *Multimedia tools and applications*, 79, pp.3311–3331.

Manocha, P., Finkelstein, A., Jin, Z., Bryan, N.J., Zhang, R. and Mysore, G.J., 2020. A differentiable perceptual audio metric learned from just noticeable differences. *arxiv preprint arxiv:2001.04460*, 1(1), p.1.

Manocha, P., Jin, Z., Zhang, R. and Finkelstein, A., 2021. Cdpam: Contrastive learning for perceptual audio similarity. *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp.196–200.

Mansour, N., Marschall, M., May, T., Westermann, A. and Dau, T., 2021. Speech intelligibility in a realistic virtual sound environment. *The journal of the acoustical society of america*, 149(4), pp.2791–2801.

Mao, T., Zhang, W., He, H., Lin, Y., Kale, V., Stein, A. and Kostic, Z., 2018. Aic2018 report: Traffic surveillance research. *Proceedings of the ieee conference on computer vision and pattern recognition workshops*. pp.85–92.

Marschner, S. and Shirley, P., 2015. *Fundamentals of computer graphics*. CRC Press.

Matulis, M. and Harvey, C., 2021. A robot arm digital twin utilising reinforcement learning. *Computers & graphics*, 95, pp.106–114.

McAllister, D.K., Lastra, A. and Heidrich, W., 2002. Efficient rendering of spatial bi-directional reflectance distribution functions. *Proceedings of the acm siggraph/eurographics conference on graphics hardware*. pp.79–88.

Mehra, R., Brimijoin, O., Robinson, P. and Lunner, T., 2020. Potential of augmented reality platforms to improve individual hearing aids and to support more ecologically valid research. *Ear and hearing*, 41(Suppl 1), p.140S.

Mehra, R., Raghuvanshi, N., Savioja, L., Lin, M.C. and Manocha, D., 2012. An efficient gpu-based time domain solver for the acoustic wave equation. *Applied acoustics*, 73(2), pp.83–94.

Mehra, R., Rungta, A., Golas, A., Lin, M. and Manocha, D., 2015. Wave: Interactive wave-based sound propagation for virtual environments. *Ieee transactions on visualization and computer graphics*, 21(4), pp.434–442.

Mesaros, A., Heittola, T. and Virtanen, T., 2016. Tut database for acoustic scene classification and sound event detection. *2016 24th european signal processing conference (eusipco)*. IEEE, Europe: IEEE, vol. 1, pp.1128–1132.

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

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. and Terzopoulos, D., 2022. Image segmentation using deep learning: A survey. *Ieee transactions on pattern analysis and machine intelligence*, 44(7), pp.3523–3542. Available from: https://doi.org/10.1109/TPAMI.2021.3059968.

Müller, T., McWilliams, B., Rousselle, F., Gross, M. and Novák, J., 2019. Neural importance sampling. *Acm transactions on graphics (tog)*, 38(5), pp.1–19.

Naef, M., Staadt, O. and Gross, M., 2002. Spatialized audio rendering for immersive virtual environments. *Proceedings of the acm symposium on virtual reality software and technology*. pp.65–72.

Nikolic, I., 2002. Improvements of artificial reverberation by use of subband feedback delay networks. *Audio engineering society convention 112*. Audio Engineering Society.

Palazzolo, E., Behley, J., Lottes, P., Giguere, P. and Stachniss, C., 2019. Refusion: 3d reconstruction in

dynamic environments for rgb-d cameras exploiting residuals. *2019 ieee/rsj international conference on intelligent robots and systems (iros)*. IEEE, pp.7855–7862.

Park, S.M. and Kim, Y.G., 2022. A metaverse: Taxonomy, components, applications, and open challenges. *Ieee access*, 10, pp.4209–4251.

Patow, G. and Pueyo, X., 2003. A survey of inverse rendering problems. *Computer graphics forum*. Wiley Online Library, 4, pp.663–687.

Pearson, E.S., D'Agostino, R.B. and Bowman, K.O., 1977. Tests for departure from normality: Comparison of powers. *Biometrika*, 64(2), pp.231–246.

Pelzer, S. and Vorländer, M., 2010. Frequency-and time-dependent geometry for real-time auralizations. *Proceedings of 20th international congress on acoustics, ica*. pp.1–7.

Pharr, M., Jakob, W. and Humphreys, G., 2023. *Physically based rendering: From theory to implementation*. MIT Press.

Plinge, A., Schlecht, S.J., Thiergart, O., Robotham, T., Rummukainen, O. and Habets, E.A., 2018. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information. *Audio engineering society conference: 2018 aes international conference on audio for virtual and augmented reality*. Audio Engineering Society.

Poeschl, S., Wall, K. and Doering, N., 2013. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. *2013 ieee virtual reality (vr)*. USA: IEEE, 1, pp.129–130. Available from: https://doi.org/10.1109/VR.2013.6549396.

Pompei, A., Sumbatyan, M. and Todorov, N., 2009. Computer models in room acoustics: The ray tracing method and the auralization algorithms. *Acoustical physics*, 55, pp.821–831.

Qiu, W. and Yuille, A., 2016. Unrealcv: Connecting computer vision to unreal engine. *European conference on computer vision*. Springer, pp.909–916.

Raghuvanshi, N. and Snyder, J., 2014. Parametric wave field coding for precomputed sound propagation. *Acm transactions on graphics (tog)*, 33(4), pp.1–11.

Raghuvanshi, N. and Snyder, J., 2018. Parametric directional coding for precomputed sound propagation. *Acm transactions on graphics (tog)*, 37(4), pp.1–14.

Ratnarajah, A., Tang, Z., Aralikatti, R. and Manocha, D., 2022. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. *Proceedings of the 30th acm international conference on multimedia*. pp.924–933.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*.

Reichardt, W., Alim, O.A. and Schmidt, W., 1975. Definition and basis of making an objective evaluation to distinguish between useful and useless clarity defining musical performances. *Acta acustica united with acustica*, 32(3), pp.126–137.

Reilly, A. and McGrath, D., 1995. Convolution processing for realistic reverberation. *Audio engineering society convention 98*. Audio Engineering Society.

Rindel, J.H., 2000. The use of computer modeling in room acoustics. *Journal of vibroengineering*, 3(4), pp.219–224.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention*. Springer, pp.234–241.

Rubio-Tamayo, J.L., Gertrudix Barrio, M. and García García, F., 2017. Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal technologies and interaction*, 1(4), p.21.

Rummukainen, O., Robotham, T., Schlecht, S.J., Plinge, A., Herre, J. and Habels, E.A., 2018. Audio quality evaluation in virtual reality: multiple stimulus ranking with behavior tracking. *Audio engineering society conference: 2018 aes international conference on audio for virtual and augmented reality*. Audio Engineering Society.

Rungta, A., Rust, S., Morales, N., Klatzky, R., Lin, M. and Manocha, D., 2016. Psychoacoustic characterization of propagation effects in virtual environments. *Acm transactions on applied perception (tap)*, 13(4), pp.1–18.

Rusu, R.B. and Cousins, S., 2011. 3d is here: Point cloud library (pcl). *2011 ieee international conference on robotics and automation*. pp.1–4. Available from: https://doi.org/10.1109/ICRA.2011.5980567.

Saeedi, S., Bodin, B., Wagstaff, H., Nisbet, A., Nardi, L., Mawer, J., Melot, N., Palomar, O., Vespa, E., Spink, T. et al., 2018. Navigating the landscape for real-time localization and mapping for robotics and virtual and augmented reality. *Proceedings of the ieee*, 106(11), pp.2020–2039.

Santos, J.F. and Falk, T.H., 2018. Speech dereverberation with context-aware recurrent neural networks. *Ieee/acm transactions on audio, speech, and language processing*, 26(7), pp.1236–1246.

Savioja, L., Huopaniemi, J., Lokki, T. and Väänänen, R., 1999. Creating interactive virtual acoustic environments. *Journal of the audio engineering society*, 47(9), pp.675–705.

Savioja, L., Manocha, D. and Lin, M., 2010. Use of gpus in room acoustic modeling and auralization. *Proc. int. symposium on room acoustics*. p.3.

Savioja, L. and Svensson, U.P., 2015a. Overview of geometrical room acoustic modeling techniques. *The journal of the acoustical society of america*, 138(2), pp.708–730.

Savioja, L. and Svensson, U.P., 2015b. Overview of geometrical room acoustic modeling techniques. *The journal of the acoustical society of america*, 138(2), 08, pp.708–730. https://pubs.aip.org/asa/jasa/article-pdf/138/2/708/13242401/708_1_online.pdf, Available from: https://doi.org/10.1121/1.4926438.

Schissler, C., Loftin, C. and Manocha, D., 2017. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *Ieee transactions on visualization and computer graphics*, 24(3), pp.1246–1259.

Schissler, C. and Manocha, D., 2016. Interactive sound propagation and rendering for large multi-source scenes. *Acm transactions on graphics (tog)*, 36(4), p.1.

Schissler, C., Mehra, R. and Manocha, D., 2014. High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *Acm transactions on graphics (tog)*, 33(4), pp.1–12.

Schissler, C., Mückl, G. and Calamia, P., 2021. Fast diffraction pathfinding for dynamic sound propagation. *Acm transactions on graphics (tog)*, 40(4), pp.1–13.

Schofield, G., Beale, G., Beale, N., Fell, M., Hadley, D., Hook, J., Murphy, D., Richards, J. and Thresh, L., 2018. Viking vr: designing a virtual reality experience for a museum. *Proceedings of the 2018 designing interactive systems conference*. pp.805–815.

Schröder, D., 2011. *Physically based real-time auralization of interactive virtual environments*, vol. 11. Logos Verlag Berlin GmbH.

Schwartz, G. and Nishino, K., 2019. Recognizing material properties from images. *Ieee transactions on pattern analysis and machine intelligence*, 1(1), p.1.

Schäfer, P., Palenda, P., Aspöck, L. and Vorlaender, M., 2024. Plug-and-play tutorials for the auralization of complex scenarios using an open-source simulation framework.

Selmanović, E., Rizvic, S., Harvey, C., Boskovic, D., Hulusic, V., Chahin, M. and Sljivo, S., 2020a. Improving accessibility to intangible cultural heritage preservation using virtual reality. *J. comput. cult. herit.*, 13(2), May. Available from: https://doi.org/10.1145/3377143.

Selmanović, E., Rizvic, S., Harvey, C., Boskovic, D., Hulusic, V., Chahin, M. and Sljivo, S., 2020b. Improving accessibility to intangible cultural heritage preservation using virtual reality. *J. comput. cult. herit.*, 13(2), may. Available from: https://doi.org/10.1145/3377143.

Series, B., 2014. Method for the subjective assessment of intermediate quality level of audio systems. *International telecommunication union radiocommunication assembly*.

Shenoi, B.A., 2005. *Introduction to digital signal processing and filter design*, vol. 169. John Wiley & Sons.

Shirley, P. and Morley, R.K., 2008. *Realistic ray tracing*. AK Peters, Ltd.

Shivappa, S., Morrell, M., Sen, D., Peters, N. and Salehin, S., 2016. Efficient, compelling, and immersive vr audio experience using scene based audio/higher order ambisonics. *2016 aes international conference on audio for virtual and augmented reality*.

Siltanen, S., Lokki, T. and Savioja, L., 2010. Room acoustics modeling with acoustic radiance transfer. *Proc. isra melbourne*.

Singh, N., Mentch, J., Ng, J., Beveridge, M. and Drori, I., 2021. Image2reverb: Cross-modal reverb impulse response synthesis. *Proceedings of the ieee/cvf international conference on computer vision (iccv)*. pp.286–295.

Slater, M., Khanna, P., Mortensen, J. and Yu, I., 2009. Visual realism enhances realistic response in an immersive virtual environment. *Ieee computer graphics and applications*, 29(3), pp.76–84.

Smith, S.W. et al., 1997. *The scientist and engineer's guide to digital signal processing*. California Technical Pub. San Diego.

Soltani, P. and Morice, A.H., 2020. Augmented reality tools for sports education and training. *Computers & education*, 155, p.103923.

Somayazulu, A., Chen, C. and Grauman, K., 2023. Self-supervised visual acoustic matching. *arxiv preprint arxiv:2307.15064*.

Southern, A., Siltanen, S., Murphy, D.T. and Savioja, L., 2013. Room impulse response synthesis and validation using a hybrid acoustic model. *Ieee transactions on audio, speech, and language processing*, 21(9), pp.1940–1952.

Spittle, B., Frutos-Pascual, M., Creed, C. and Williams, I., 2022. A review of interaction techniques for immersive environments. *Ieee transactions on visualization and computer graphics*.

Stan, G.B., Embrechts, J.J. and Archambeau, D., 2002. Comparison of different impulse response measurement techniques. *Journal of the audio engineering society*, 50(4), pp.249–262.

Sutherland, I.E., 1968. A head-mounted three dimensional display. *Proceedings of the december 9-11, 1968, fall joint computer conference, part i*. pp.757–764.

Szeliski, R., 2022. *Computer vision: algorithms and applications*. Springer Nature.

Tang, Z., Bryan, N.J., Li, D., Langlois, T.R. and Manocha, D., 2020. Scene-aware audio rendering via deep acoustic analysis. *Ieee transactions on visualization and computer graphics*, 26(5), pp.1991–2001.

Taylor, M., Chandak, A., Mo, Q., Lauterbach, C., Schissler, C. and Manocha, D., 2012. Guided multiview ray tracing for fast auralization. *Ieee transactions on visualization and computer graphics*, 18(11), pp.1797–1810.

Teixeira, F., Sarris, C., Zhang, Y., Na, D.Y., Berenger, J.P., Su, Y., Okoniewski, M., Chew, W., Backman, V. and Simpson, J., 2023. Finite-difference time-domain methods. *Nature reviews methods primers*, 3(1), p.75.

Thompson, P.R., 2005. A graphic representation of acoustics using ray tracing. *The journal of the acoustical society of america*, 82(S1), 08, pp.S45–S45. https://pubs.aip.org/asa/jasa/article-pdf/82/S1/S45/12119055/s45_2_online.pdf, Available from: https://doi.org/10.1121/1.2024820.

Tsingos, N., Funkhouser, T., Ngan, A. and Carlbom, I., 2001. Modeling acoustics in virtual environments using the uniform theory of diffraction. *Proceedings of the 28th annual conference on computer graphics and interactive techniques*. pp.545–552.

Turner, E., Cheng, P. and Zakhor, A., 2014. Fast, automated, scalable generation of textured 3d models of indoor environments. *Ieee journal of selected topics in signal processing*, 9(3), pp.409–421.

Ungureanu, D., Bogo, F., Galliani, S., Sama, P., Duan, X., Meekhof, C., Stühmer, J., Cashman, T.J., Tekin, B., Schönberger, J.L. et al., 2020. Hololens 2 research mode as a tool for computer vision research. *arxiv preprint arxiv:2008.11239*.

Virtanen, T., 2006. *Sound source separation in monaural music signals*. Tampere University of Technology.

Vorländer, M., 2008. *Simulation of sound in rooms*. Springer.

Vorländer, M., Schröder, D., Pelzer, S. and Wefers, F., 2015. Virtual reality for architectural acoustics. *Journal of building performance simulation*, 8(1), pp.15–25.

Wang, L.W., Siu, W.C., Liu, Z.S., Li, C.T. and Lun, D.P., 2020. Deep relighting networks for image light source manipulation. *Computer vision–eccv 2020 workshops: Glasgow, uk, august 23–28, 2020, proceedings, part iii 16*. Springer, pp.550–567.

Woodward, B.F., 2021. *Cost-effective and eco-friendly fire investigation training using photorealistic interactive room scale virtual reality*. Cham: Springer International Publishing, pp.147–156. Available from: https://doi.org/10.1007/978-3-030-72781-9_18.

Xia, X. and Kulis, B., 2017. W-net: A deep model for fully unsupervised image segmentation. *arxiv preprint arxiv:1711.08506*.

Yang, J., Barde, A. and Billinghurst, M., 2022. Audio augmented reality: a systematic review of technologies, applications, and future research directions. *journal of the audio engineering society*, 70(10), pp.788–809.

Yang, J. and Mattern, F., 2019. Audio augmented reality for human-object interactions. *Adjunct proceedings of the 2019 acm international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 acm international symposium on wearable computers*. pp.408–412.

Yang, J., Pfreundtner, F., Barde, A., Heutschi, K. and Sörös, G., 2020. Fast synthesis of perceptually adequate room impulse responses from ultrasonic measurements. *Proceedings of the 15th international audio mostly conference*. pp.53–60.

Yannakakis, G.N. and Togelius, J., 2018. *Artificial Intelligence and Games*. https://gameaibook.org. Springer.

Yokota, T., Sakamoto, S., Tachibana, H. and Ishii, K., 2008. Numerical analysis and auralization of the'roaring dragon' phenomenon by the fdtd method. *Journal of environmental engineering(transaction of aij)*, 73(629), pp.849–856.

Yuen, S.C.Y., Yaoyuneyong, G. and Johnson, E., 2011. Augmented reality: An overview and five directions for ar in education. *Journal of educational technology development and exchange (jetde)*, 4(1), p.11.

Zhang, Z., Raghuvanshi, N., Snyder, J. and Marschner, S., 2018. Ambient sound propagation. *Acm trans. graph.*, 37(6), December. Available from: https://doi.org/10.1145/3272127.3275100.

Zimmons, P. and Panter, A., 2003. The influence of rendering quality on presence and task performance in a virtual environment. *Ieee virtual reality, 2003. proceedings.* IEEE, pp.293–294.

Zotkin, D., Hwang, J., Duraiswaini, R. and Davis, L.S., 2003. Hrtf personalization using anthropometric measurements. *2003 ieee workshop on applications of signal processing to audio and acoustics (ieee cat. no. 03th8684)*. Ieee, pp.157–160.

Zotter, F. and Frank, M., 2019. *Ambisonics: A practical 3d audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature.

Zwicker, E. and Fastl, H., 2013. *Psychoacoustics: Facts and models*, vol. 22. Springer Science & Business Media.

Appendix A

Proof-of-Concept Material Tagging System

A proof of concept implementation is designed to work on real-world scenes, gathering insights into the performance of the proposed material tagging prototypes in attributing acoustic properties to segments of the scene geometry of a given environment expressed as a set of triangulated meshes. The design focuses on scenes scanned with space reconstruction techniques applied to real environments, emulating space reconstruction paradigms often adopted by MR HMDs.

The design prototype uses standard space-sensing cameras to produce triangulated meshes from a set of real environments. These virtual reconstructions are then processed, segmenting the scene of each environment by generating portions of scene geometry into a set of semantically meaningful sub-meshes expressing elements of the scene. Submeshes are tagged with acoustic material by assigning each scene element, referred to by its sub-mesh, a semantic acoustic material drawn from a generated list of materials defined by each of the two prototypes.

The two prototype systems are deployed on the virtual reconstructions of the real spaces, inferring acoustic materials across all scene elements.

A.1 Proof of Concept System

The system is deployed across various test scenes that are selected to represent diverse environments with varying geometric and material complexities to demonstrate the application of the proof of concept system on real-world reconstructed data. The system deploys the tested texture-based method illustrated in Section 4.3. As an overview, the system starts by extracting textures from the environment, as they express the visual features of surfaces within the scene, which can be used to infer material properties. Following the tested methodology, image textures are segmented using a superpixel segmentation technique.

The SLIC (Simple Linear Iterative Clustering) superpixel algorithm is employed (Achanta et al., 2012). This algorithm partitions the extracted textures into superpixels, providing clusters of pixels with similar characteristics. The SLIC algorithm allows the partitioning of image textures into perceptually meaningful segments separating contextual elements expressed in image data into clusters. Once the superpixels are formed, they are converted into rectified patches, enabling CNNs to treat clusters of pixels as input image data. This step involves transforming the clustered pixel data into $32x32$ image patches. The rectified patches are then fed into the pre-trained ResNet50 feature extractor, identifying the distinctive features of each patch, which are indicative of semantic materials found in the texture data. Upon extraction of features, a classifier layer appended to the feature extractor assigns semantic material labels to each patch. This classification is based on the textural and color features identified during feature extraction, allowing the system to discern between different material types such as wood, metal, fabric, etc. Each semantic material type is associated with specific acoustic absorption coefficients. These coefficients are critical for simulating how sound interacts with different materials, regardless of the nature of the acoustic rendering method adopted, impacting the acoustic properties of the environment. In the final step, the semantic materials, along with their corresponding acoustic properties, are mapped onto a set of meshes that represent the physical structure of the environment. This mapping uses UV mapping data to accurately align the material properties with the 3D geometry of the meshes, ensuring that the acoustic simulation reflects real-world conditions.

A.2 Real Spaces for Demonstration

The proof of concept is demonstrated in three real spaces: a small, a medium, and a large environment, see Table A.1. These environments have diverse ecosystems of surfaces and materials that characterise and determine their respective soundscape and acoustic fingerprints. Figures A.4, A.5, and A.6 show demonstrations of superpixel-based material tagging on textures extracted from the three spaces.

Scene	Dimensions (L x W x H)	Volume	Vertices	Faces
Mastering Suite	$5.4m \times 3.5m \times 2m$	$3.78 \times 10^1\,\mathrm{m}^3$	264939	281425
Recital Hall	$15m \times 12m \times 5.5m$	$9.90 \times 10^2\,\mathrm{m}^3$	606331	486811
St Mary's Guildhall	$23m \times 14m \times 11m$	$3.542 \times 10^3\,\mathrm{m}^3$	231116	439844

Table A.1: Summary of the scenes used for the testing procedure of the acoustic material tagging prototypes.

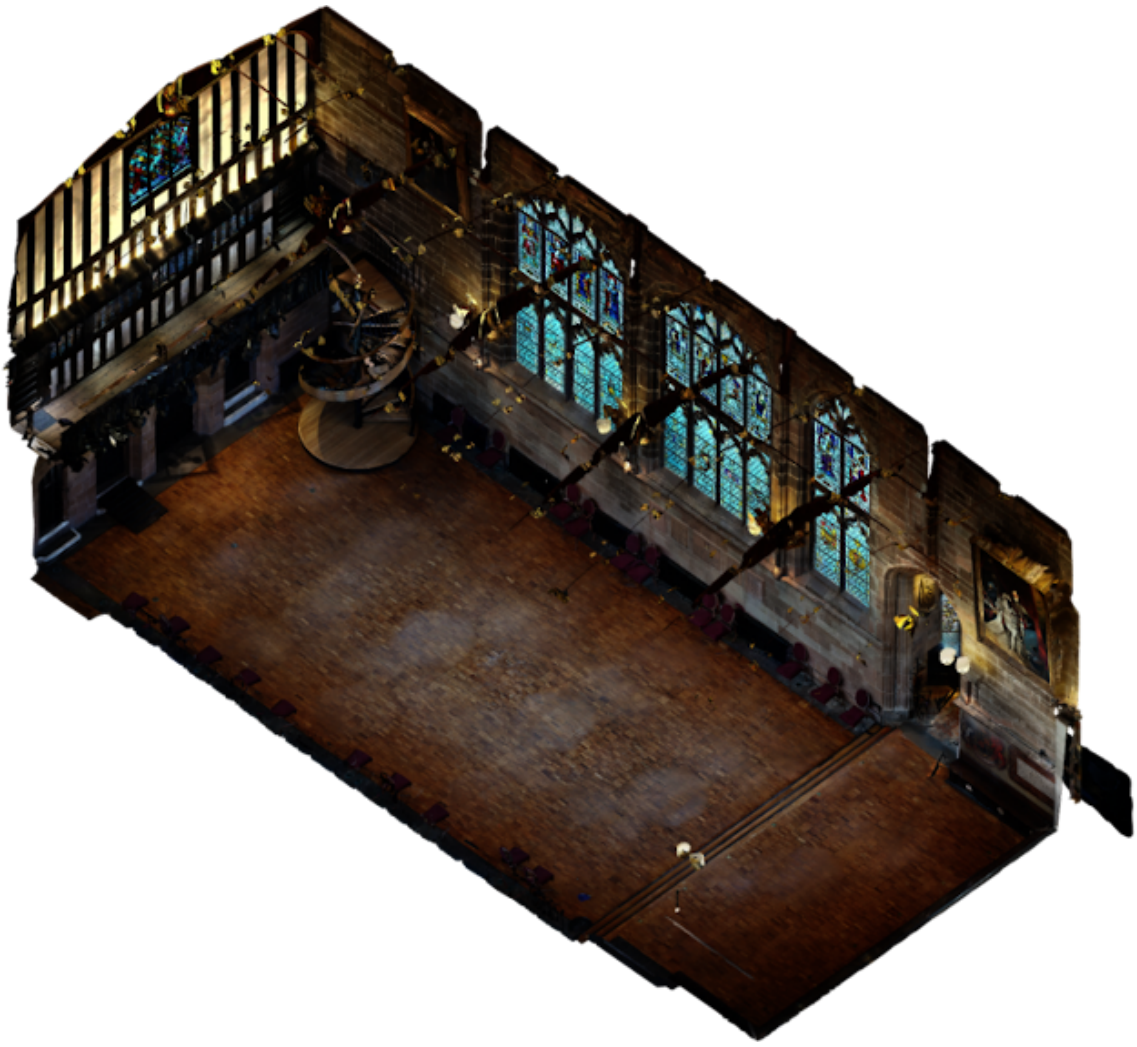
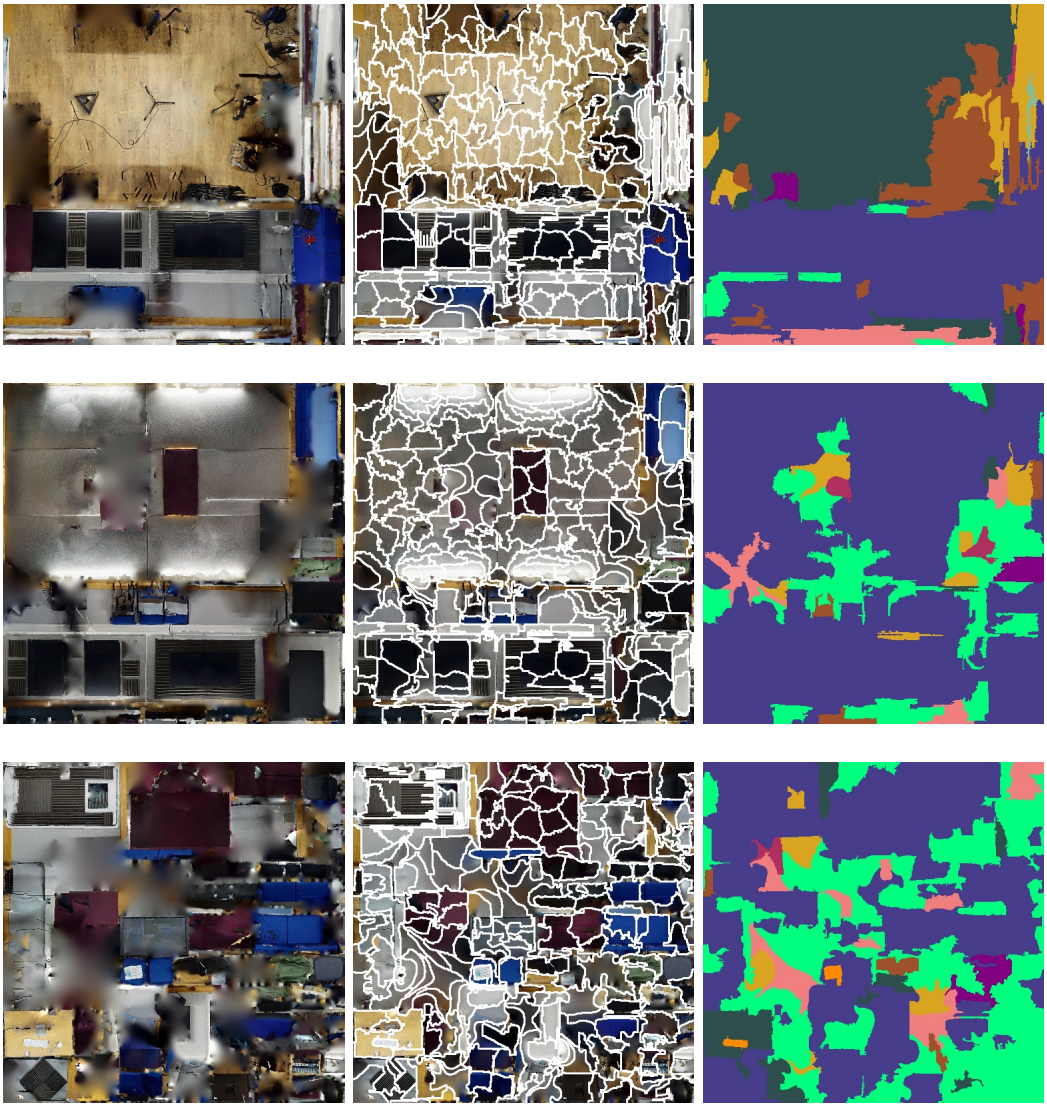


Figure A.1: St Mary's Guildhall: a Medieval-style church in Coventry, West Midlands, England. The large environment has a unique soundscape, characterised by a recorded T_{60} reverberation time of over a second.

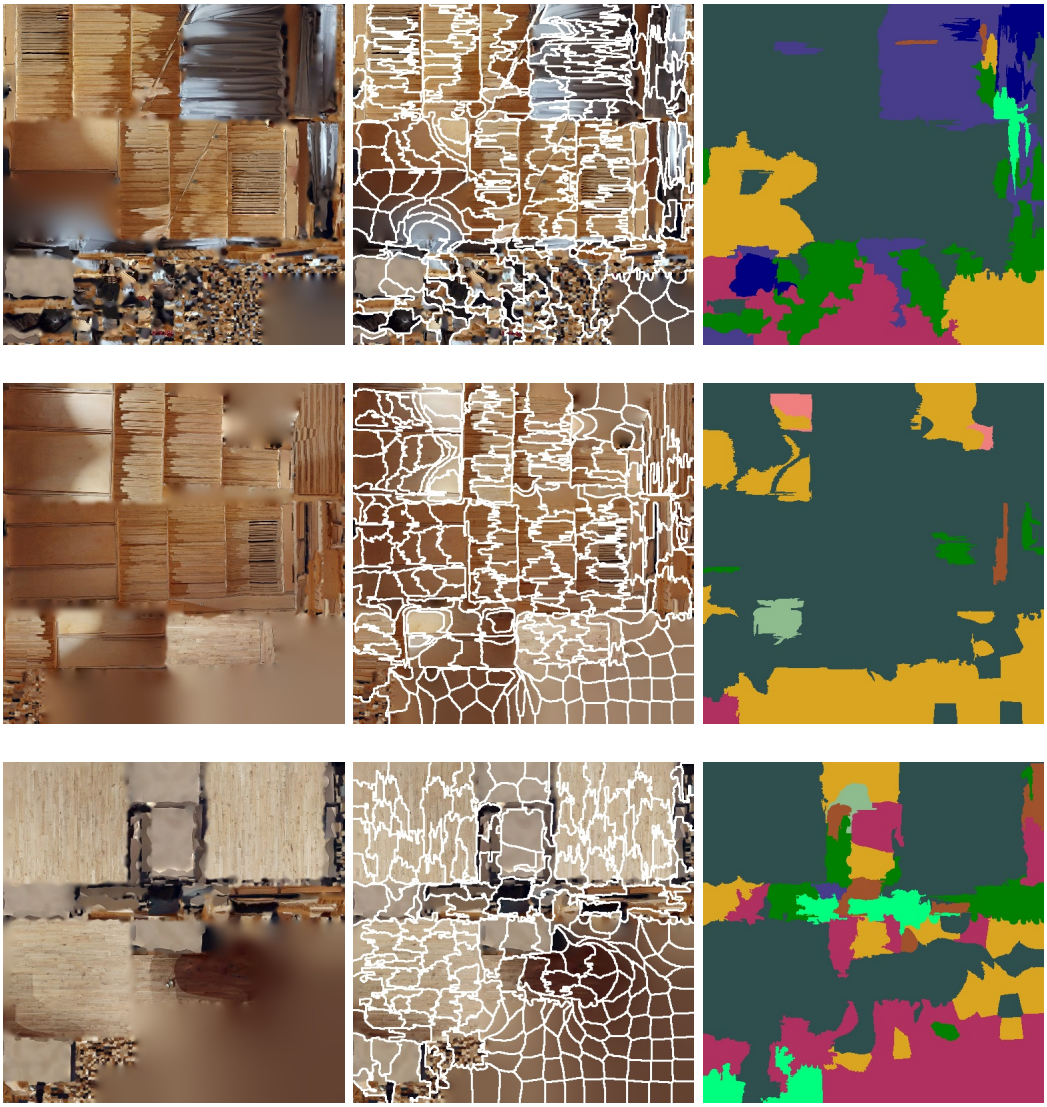
Figure A.2: Recital Hall: a wooden space serving as a stage for musical performances. The soundscape is characterised by a recorded T_{60} reverberation time within a second.

Figure A.3: Mastering Suite: a small audio production studio with acoustic treatments to minimise the effect of the soundscape on the sound reproduction system within.



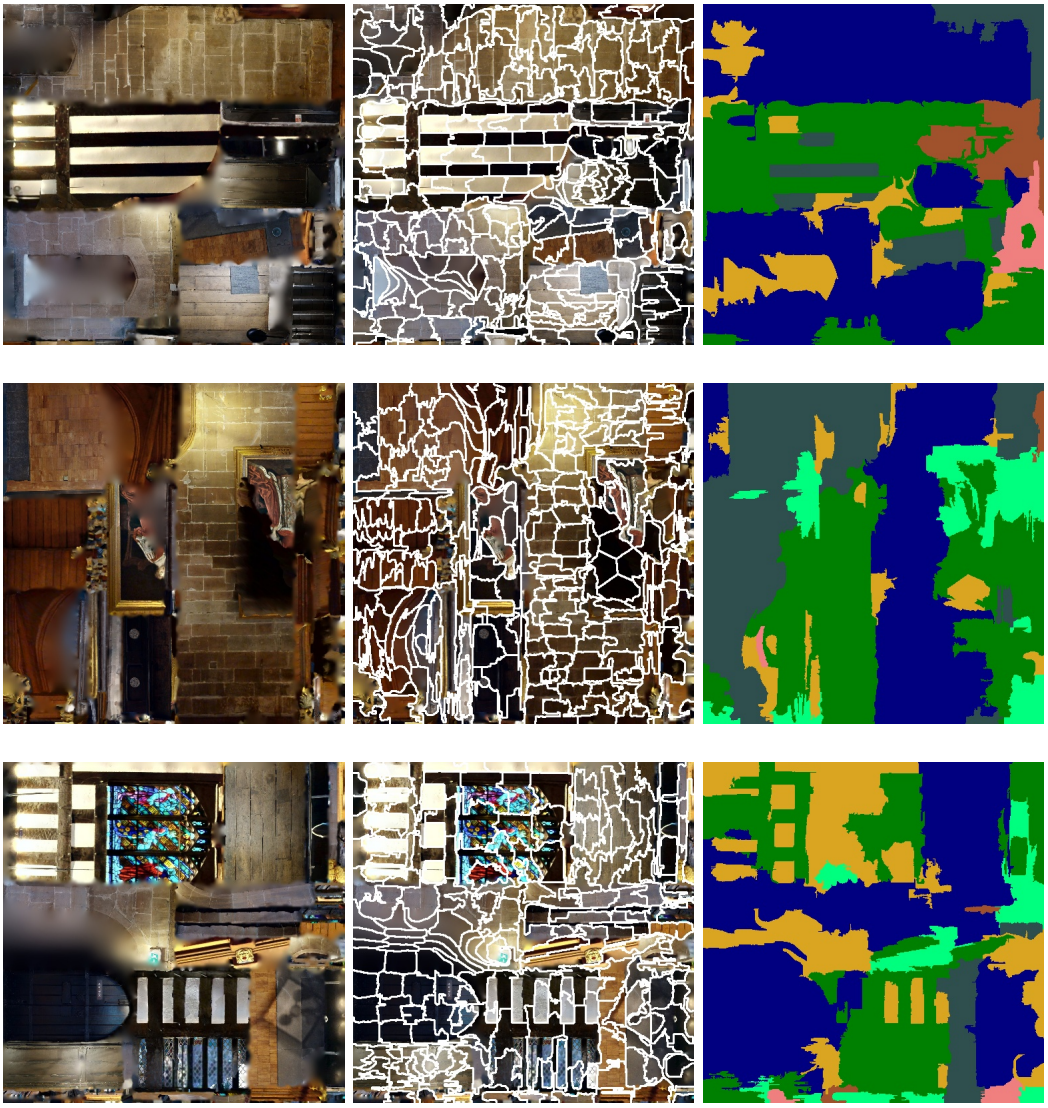
(a) Input Texture (b) Superpixel Segmentation (c) Material Tagging

Figure A.4: Material Tagging performed on textures from the Mastering Suite scene.



(a) Input Texture (b) Superpixel Segmentation (c) Material Tagging

Figure A.5: Material Tagging performed on textures from the Recital Hall scene.



(a) Input Texture (b) Superpixel Segmentation (c) Material Tagging

Figure A.6: Material Tagging performed on textures from the St Mary's Guild Hall scene.