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

Birmingham City University

May 2024

Supervisors
Dr Carlo Harvey, Dr Maite Frutos-Pascual

Abstract

Immersive technology permeates an increasing number of industry domains, causing a wide range of applications to benefit from visualising, experiencing, and manipulating digital reconstructions of the physical world through the lenses of a Head-Mounted Display (HMD). Augmented Reality (AR), as an emerging branch of immersive technology allows the projection of visual and audio stimuli onto the users' surroundings, enabling new interaction paradigms that find applications in a broad range of use cases, like training and simulation, for medical or military domains, as well as manufacturing, cultural heritage, accessibility, and more. The adoption of AR across these domains has incentivised the development of HMD technology and better and more optimal techniques for efficient and realistic holographic displays, leveraging sensing and capturing capabilities provided by the HMD. As a result, holograms 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 a body of literature on Extended Reality research domains. Physically based sound rendering is key to achieving realism in audio stimuli within immersive environments, as spatial features of the users' surroundings can be considered, simulating basic 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 activities and interactions in immersive experiences. There is a large family of techniques for simulating sound propagation effects within virtual environments, building from decades of research that provide a variety of techniques with varying degrees of realism and computational requirements. In the course of this thesis, the current state of sound rendering techniques is reviewed, discussing their application and feasibility for AR use cases, aiming at proposing a novel pipeline that can generate context-aware realistic audio for AR 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 complex scenes, and allowing audio stimuli to respond to spatial characteristics of the immersive environment. The experiments presented demonstrate applications of scene understanding techniques to game scenes and virtual reconstructions of real space to determine acoustic properties of scene geometry for automating realistic sound rendering, identifying the current state of automatic acoustic material recognition for virtual environments and proposing a novel evaluation framework to test objective and subjective accuracy against measurements from real environments. Proof-of-concept systems have been tested on state-of-the-art acoustic renderers to demonstrate their efficiency in offline procedures. Subjective testing conducted using a prototype deployment of the proposed pipeline suggests that audio stimuli generated using the proposed pipeline have a significant effect on task performance within AR. 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 AR platforms, where the acoustic space is constantly updating based on the surrounding, real world.

Preface

The Acknowledgements section may be used to thank your supervisor, family, research funding bodies, or any other applicable individuals or institutions.

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 Mattia Colombo Date: 7th May, 2024

Mattia Colombo

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

AR	Augmented Reality.
BRDF	Bidirectional Reflectance Distribution Function.
BSP	Binary Space Partitioning.
BVH	Bounding Volume Hierarchy.
CNN	Convolutional 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.
HAS	Human Hearing System.
HMD	Head-Mounted Display.
HRTF	Head-Related Transfer Function.
ILD	Interaural Level Difference.
IR	Impulse Response.

ISM	Image-Source Model.
ITD	Interaural Time Difference.
JND	Just-Noticeable Difference.
MSE	Mean Squared Error.
RIR	Room Impulse Response.
VE	Virtual Environment.
VR	Virtual Reality.

Chapter 1

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.

1.1 Background on Human Hearing

1.1.1 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: 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 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 1.1.2 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. The interaction of different senses, particularly how sound influences visual perception in virtual reality scenarios, affects how soundfields 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). The feature mismatch between visual and auditory cues can affect the externalization 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.

1.1.2 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

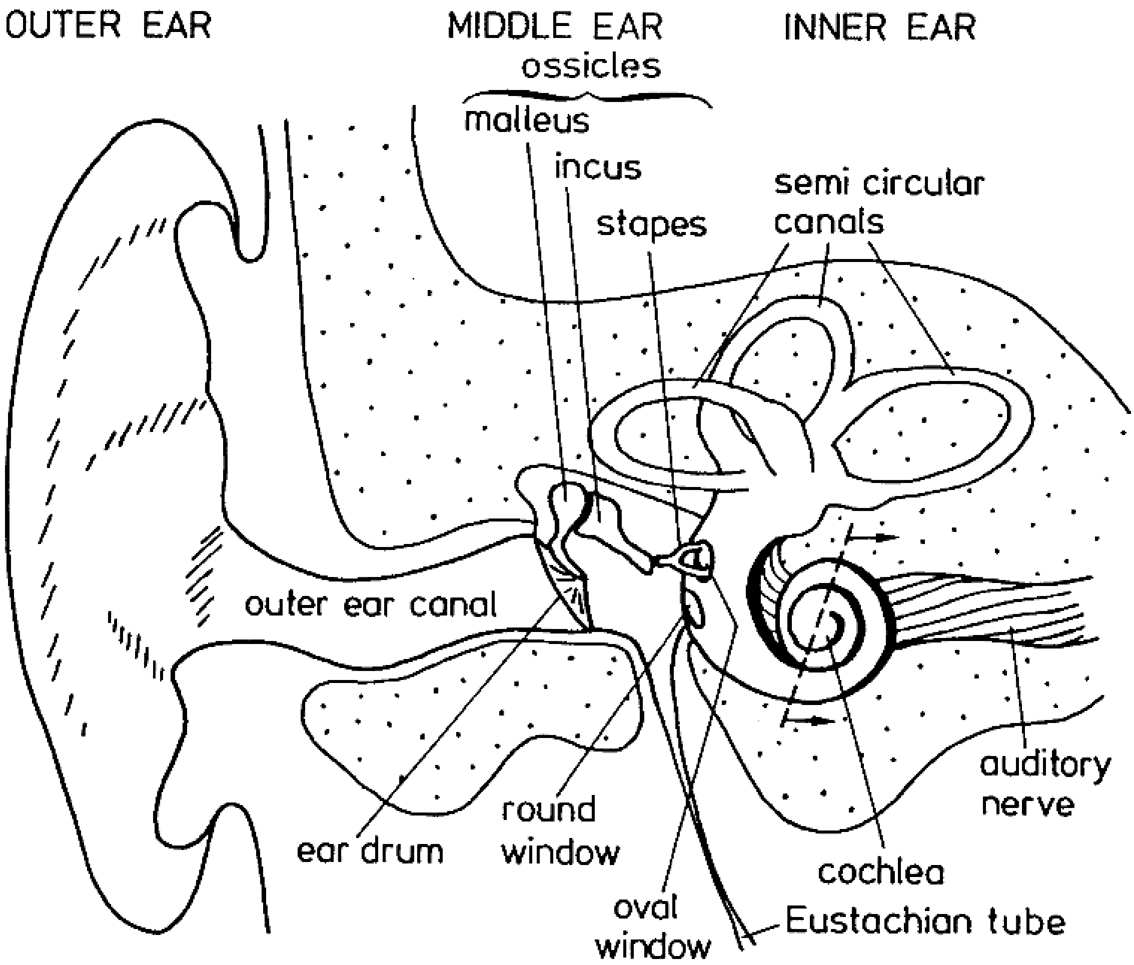
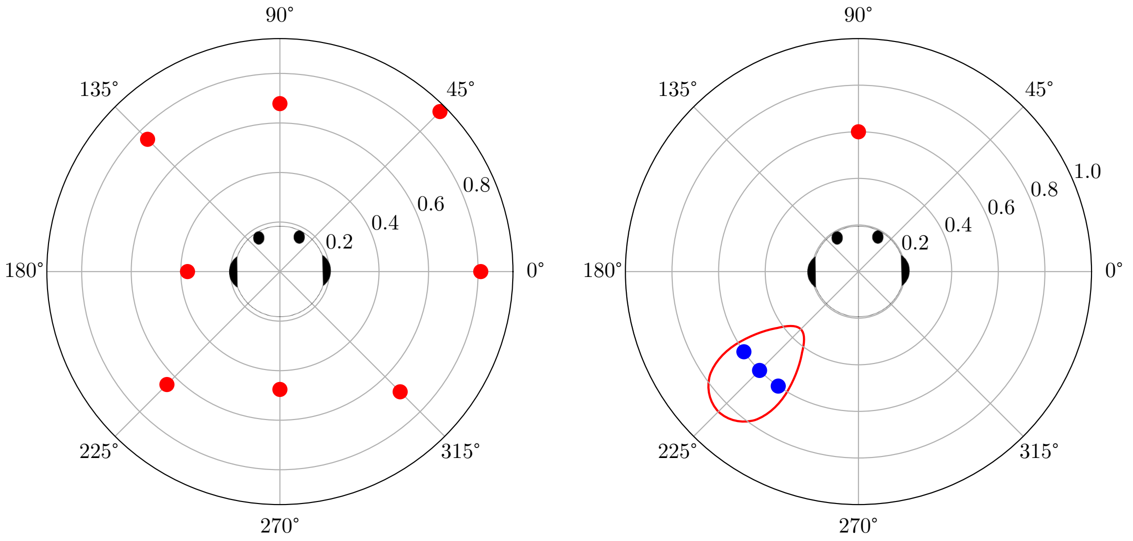


Figure 1.1: Overview of the human hearing system (Zwicker and Fastl, 2013)



(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 1.2: Visualisation of sound localisation and masking.

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 is a method 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.

1.1.3 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.

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 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).

1.1.4 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).

1.1.5 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.

1.1.6 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.

1.2 Sound Propagation Background

1.2.1 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:

$$c = (331.4 + 0.6\Theta)\ \frac{m}{s} \tag{1.1}$$

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.

1.2.2 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 60 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 80 ms or 50 ms) to late reflections (Reichardt, Alim and Schmidt, 1975). Equation 1.2 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.

$$C_{80} = 10\log\frac{\displaystyle\int_{0}^{80ms} \mathrm{p}^2(\mathrm{t})\mathrm{d}t}{\displaystyle\int_{80ms}^{\infty} \mathrm{p}^2(\mathrm{t})\mathrm{d}t} \tag{1.2}$$

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 1.3.

$$D_{50} = 10\log\frac{\displaystyle\int_{0}^{50ms} \mathrm{p}^2(\mathrm{t})\mathrm{d}t}{\displaystyle\int_{0}^{\infty} \mathrm{p}^2(\mathrm{t})\mathrm{d}t} \tag{1.3}$$

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).

1.2.3 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

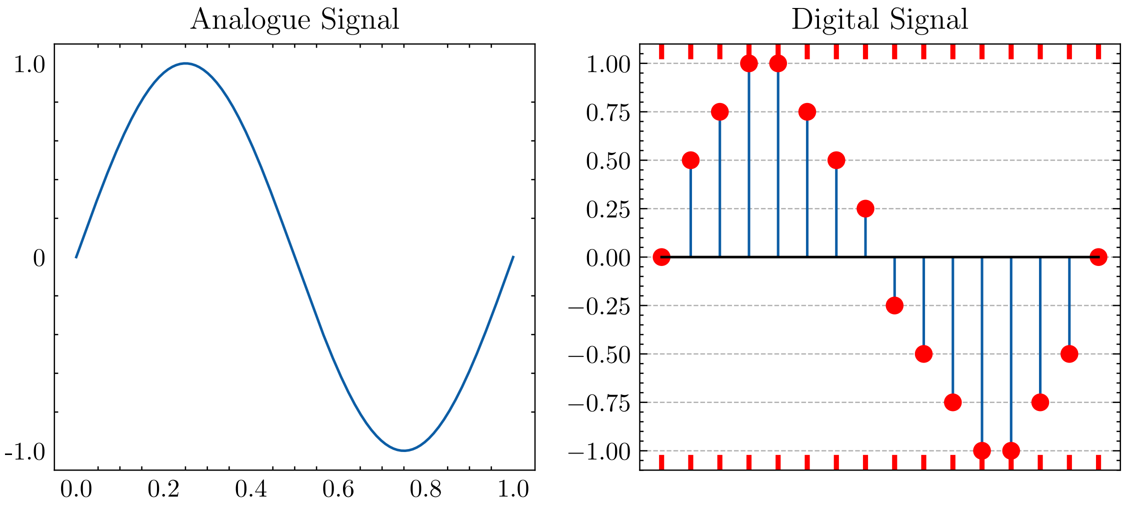


Figure 1.3: Analogue and digital signal: the left axes show a continuous signal and the right axes show a discrete sampled representation.

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 1.3 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 book 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 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 1.3. 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 1.3-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 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 1.4, for instance, a 50 Hz signal is sampled at 90 Hz, below the 100 Hz Nyquist sampling frequency, causing aliasing, an incorrectly reconstructed signal that will be able to oscillate at a maximum frequency of 45 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).

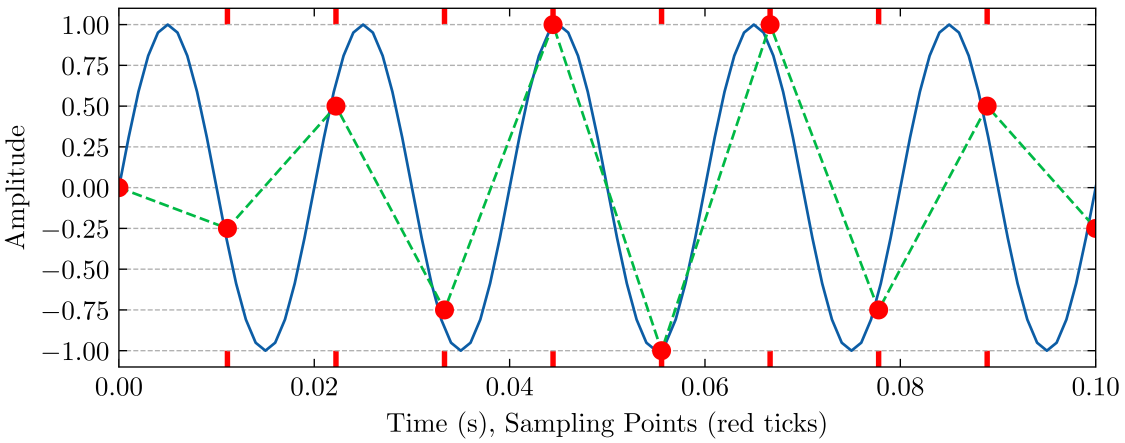


Figure 1.4: The blue signal is being sampled at a sampling frequency lower than the Nyquist frequency, causing aliasing, an incorrect reconstruction of the digital signal, as opposed to Figure 1.3 that shows a correctly reconstructed signal. As a result, the dotted green signal is created instead, having a frequency between 0 Hz and the sampling frequency.

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 Figure1.3-b and generates N complex numbers, representing power across N sinusoids. The DTFT, as defined by classic 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:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn} \tag{1.4}$$

1.2.4 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 Responses (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. Figure 1.9 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

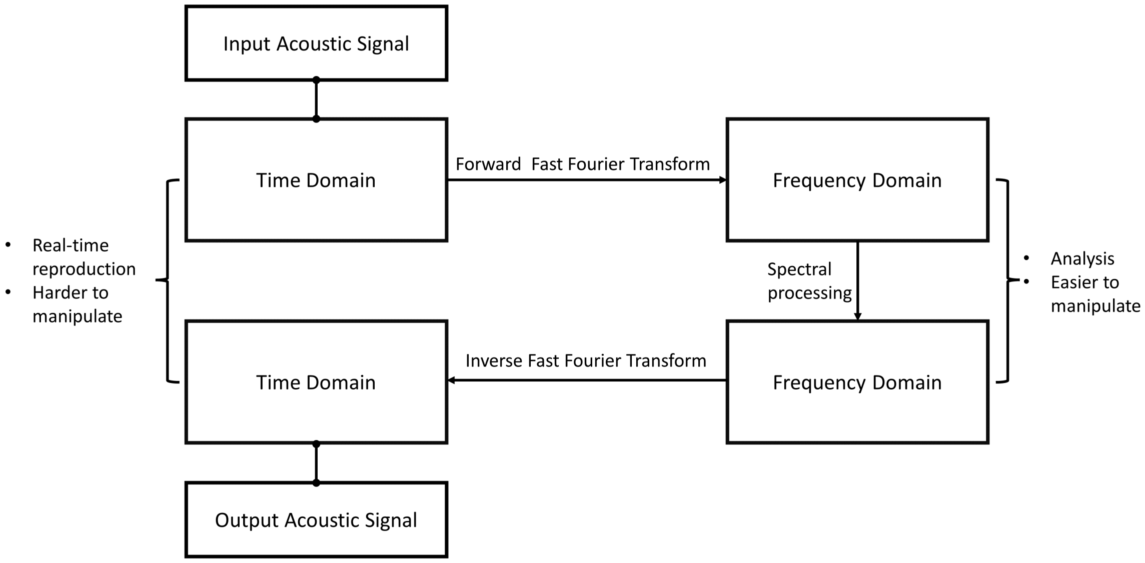


Figure 1.5: A basic chain for signal processing aimed at analysing or manipulating signal in auralisation, visualisation, or interactive applications. Analysis and processing of digital signals in the time domain is generally a hard task due to the complex nature of the function representing audio. The frequency-domain representation eases analysis and manipulation problems even with the added computational load of transforming between the two domains.

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.

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 IRs are defined as Binaural Room Impulse Responses: they extend RIRs 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 1.6 is an example of a monaural RIR, shown both in the time and frequency domain.

1.2.5 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

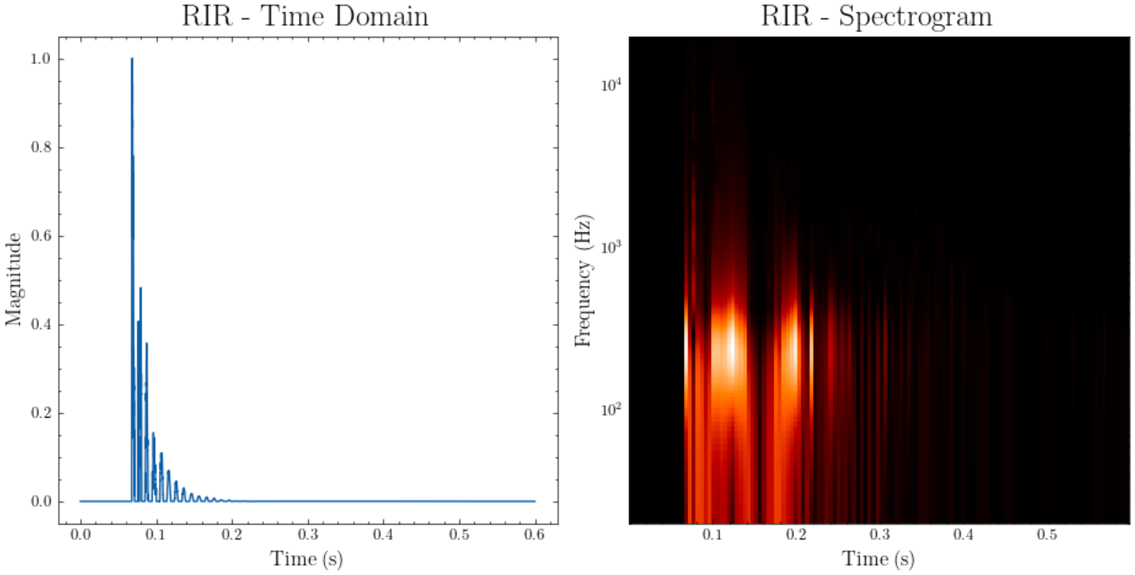


Figure 1.6: Both time and frequency-domain, left and right respectively, representations of a Room Impulse Response (RIR). The time domain representation shows the magnitude of sound paths from a sound source to a receiver over time. The frequency domain representation shows how the energy of sound paths is distributed across the frequency spectrum over time, which is visualised with an infrared colour map.

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 1.3.2 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). 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 1.6 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).

1.3 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 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).

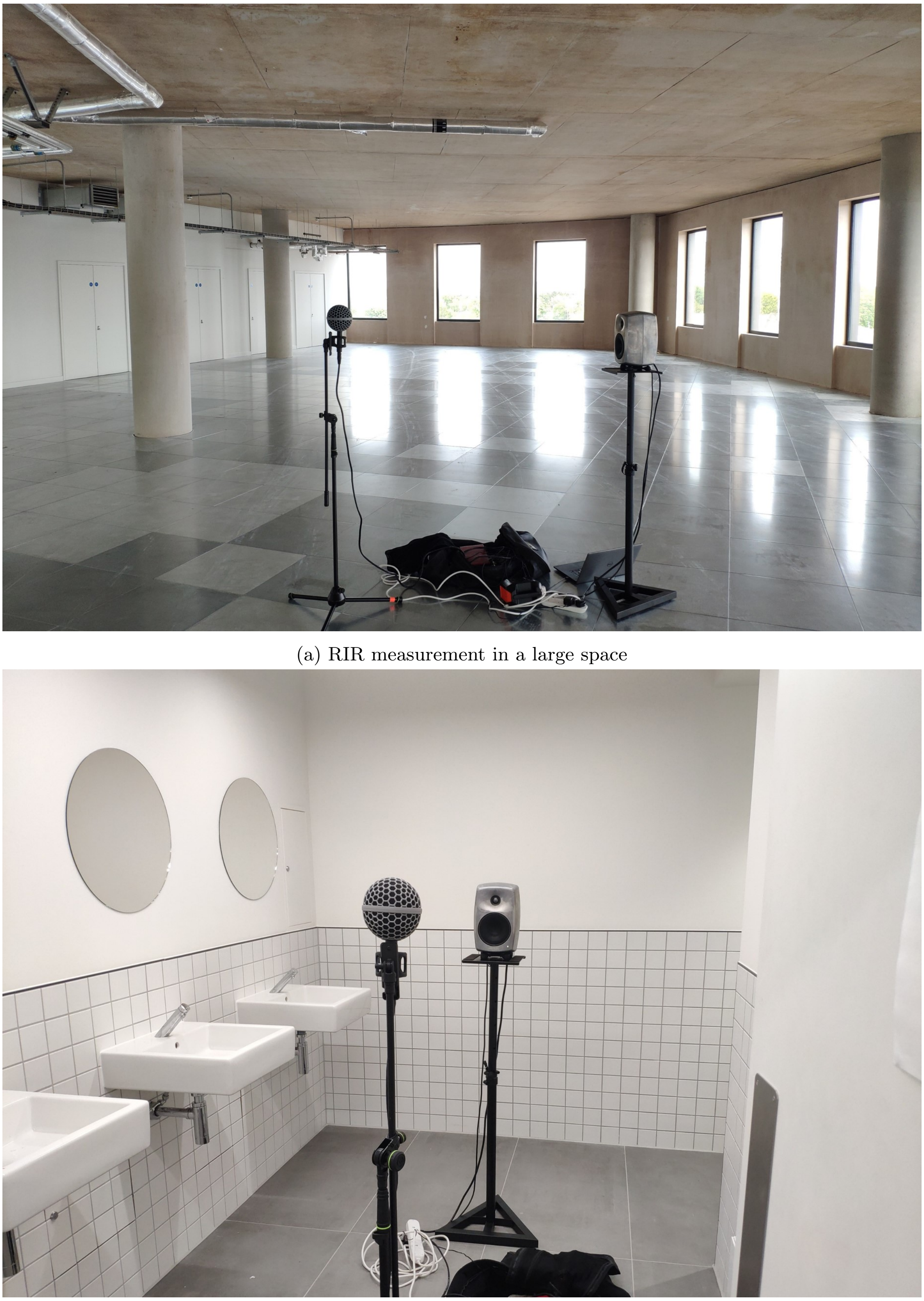
1.3.1 3D Sound Reproduction Techniques

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.

1.3.2 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



(a) RIR measurement in a large space

(b) RIR measurement in a small space

Figure 1.7: 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.

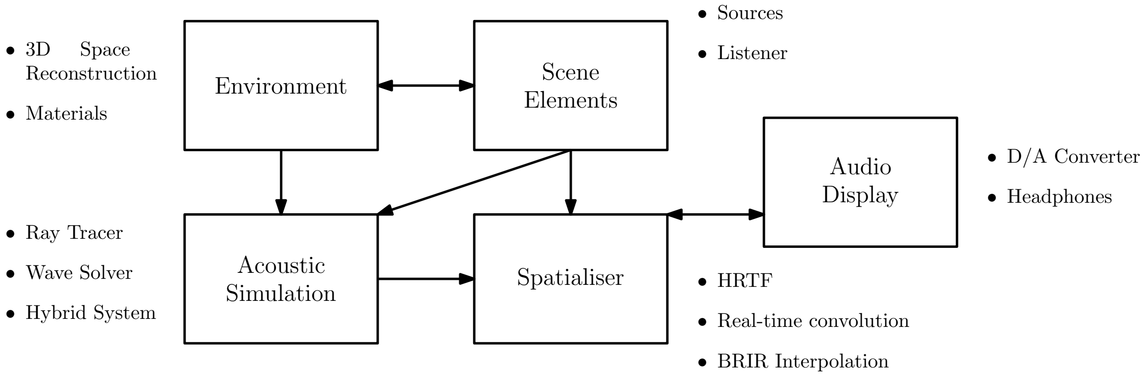


Figure 1.8: An example chain of 3D reproduction based on acoustic simulators: an environment is fed into acoustic simulators to produce BRIRs. A spatialiser system consumes these to generate an audio signal considering the listener and scene elements.

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:

$$y[n] = x[n] * h[n]. \tag{1.5}$$

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:

$$(x * h)[n] = \sum_{m=0}^{M} x[n-m]h[m]. \tag{1.6}$$

As shown in Figure 1.5, 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:

$$(x * h)[n] = IDTFT_N(DTFT_N(x[n]) \cdot DTFT_N(h[n])), \tag{1.7}$$

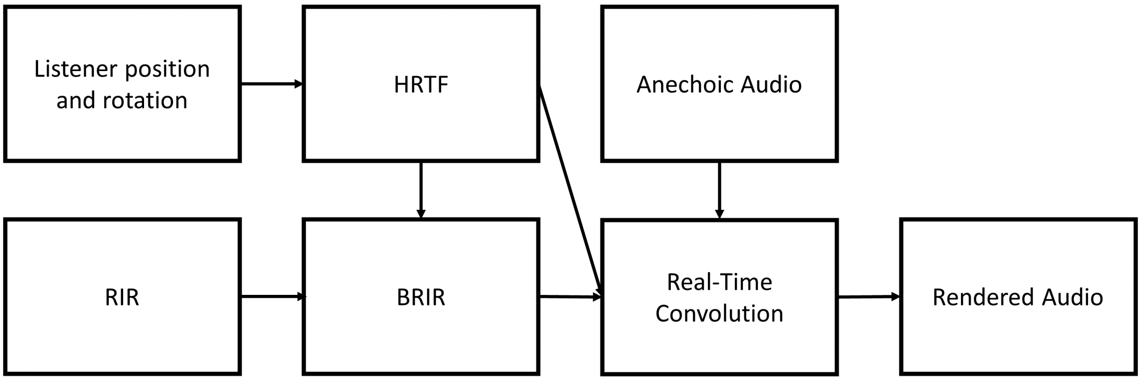


Figure 1.9: Common spatial audio pipeline: the listener's position and rotation in the scene is used to sample an interpolated HRTF from the currently loaded bank, combined with the generated IR, the real-time convolution algorithm applies the BRIR to an anechoic audio signal. The result is an audio signal arriving at the listener, taking into account their position and rotation with respect to the sound-emitting object in the virtual environment.

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 1.7 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 1.9, spatialiser systems or acoustic simulation systems provide filters in the forms of IRs that can be applied to audio chunks from the audio buffer.

1.3.3 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 1.2.3, 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 1.5. 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 1.9.

1.3.4 Common methods for Auralisations

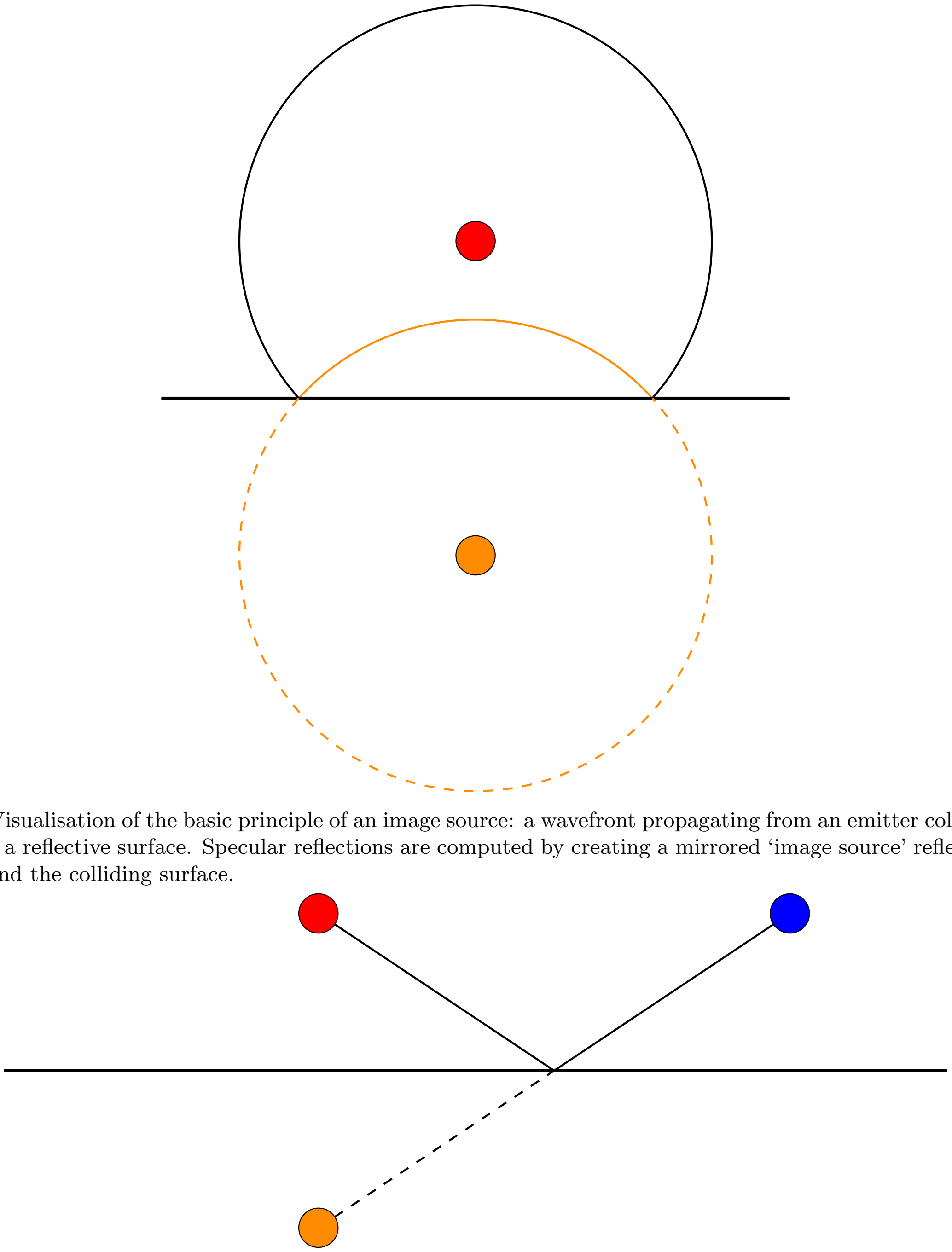
Methods for producing auralisation start from the creation of an environment, which is the first component of the system in Figure 1.9 hosting the virtual sound-emitting objects, e.g. an orchestra, and the virtual sound-receiving objects, e.g. the listener. In computers, environments are generally represented 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.

1.3.5 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, 2015).

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

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. 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 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, 2015). 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 1.12.

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

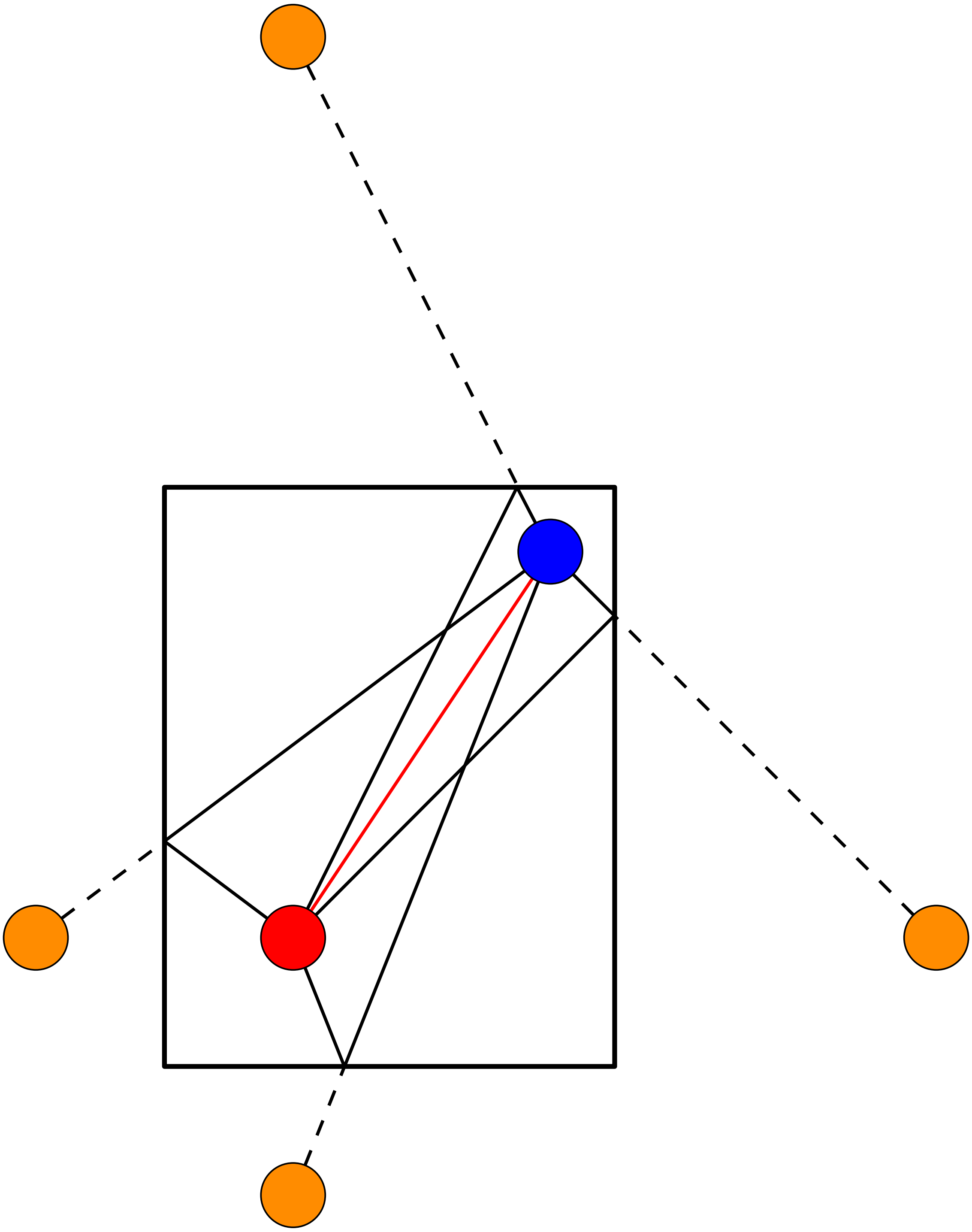


Figure 1.11: Visualisation of specular reflections from an emitter (red circle) to a receiver (blue circle) computed in a shoebox room. Image sources (orange circles) determine reflection points on the room geometry. The direct path (red segment) plus specular reflections computed by the Image Source Model can be used to approximate the basic acoustic features of a space.

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 1.13. 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. 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, 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 1.14.

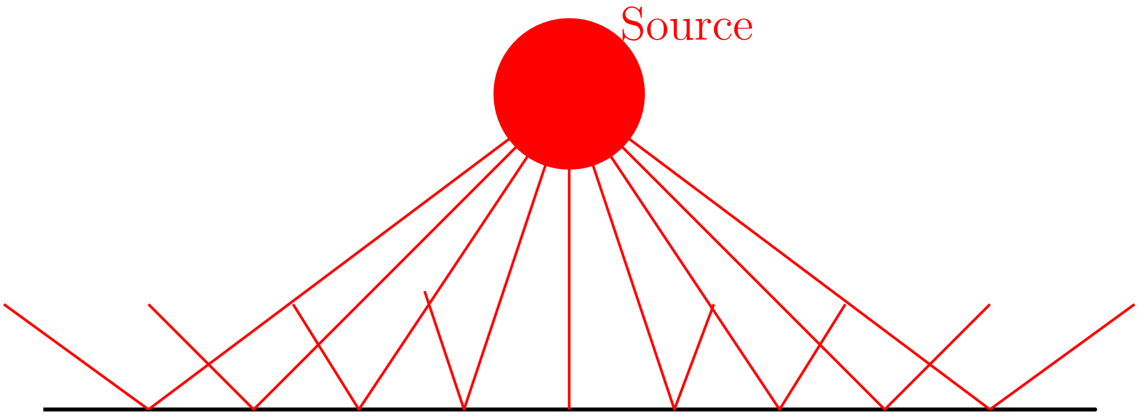


Figure 1.12: Visualisation of a ray tracing source: In a 3D environment, a given number of rays is often emitted from uniformly distributed points across the azimuth and elevation of the emitter surface. Rays emitted from the source and colliding with geometry are reflected around the surface normal at the collision point, generating a new ray.

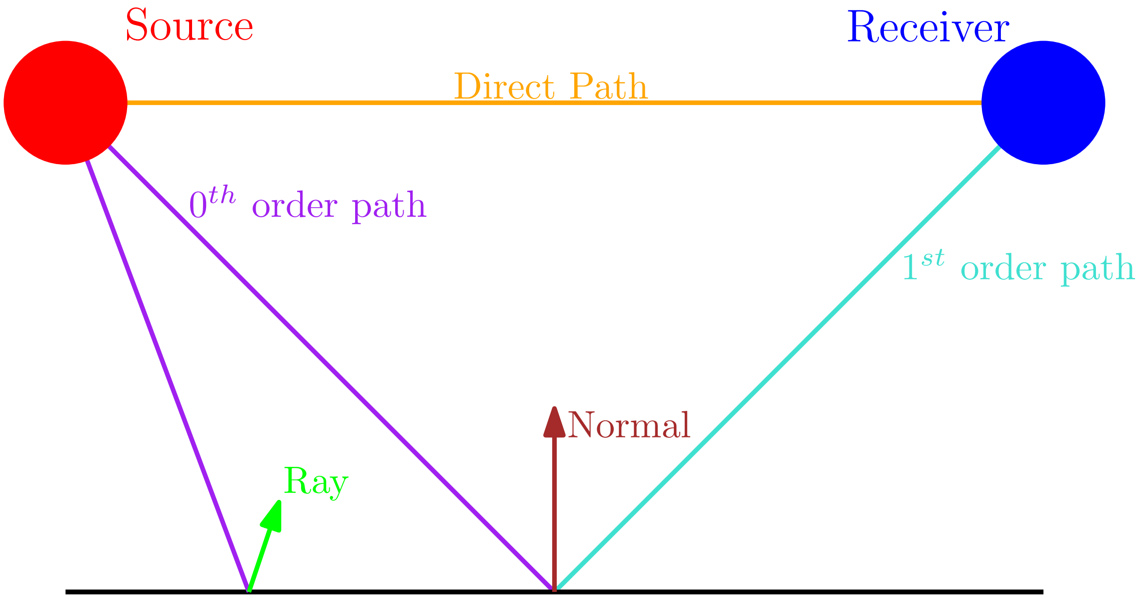


Figure 1.13: Visualisation of a basic Ray Tracing Model for computing specular reflections. Propagation paths from an emitter (red circle) to a receiver (blue circle) are calculated by checking whether reflecting rays intersect the receiver.

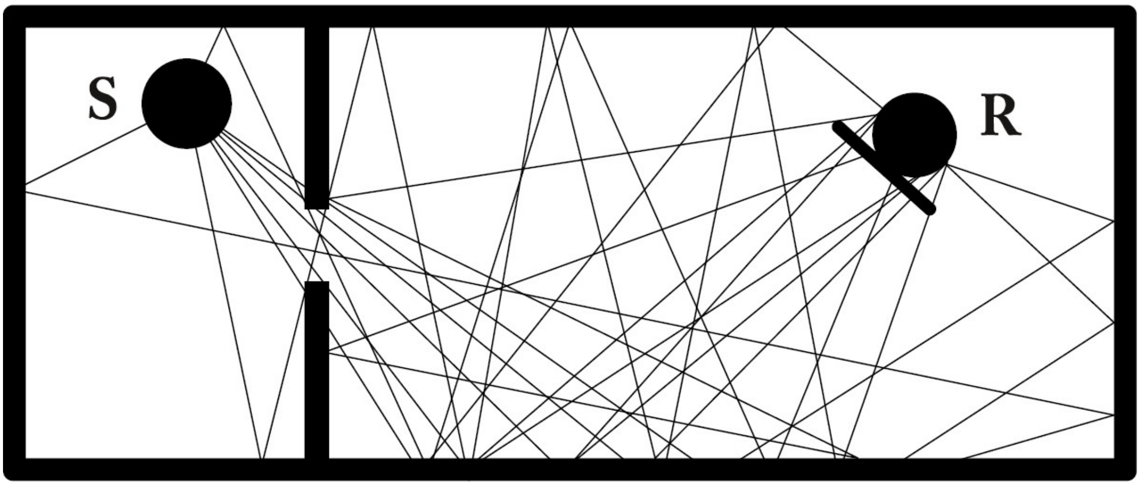


Figure 1.14: A Ray Tracing model computing specular reflections in a shoebox room. The example shows that the model lacks the ability to simulate diffraction effects caused by sound bending through a portal (Funkhouser, Tsingos and Jot, 2003).

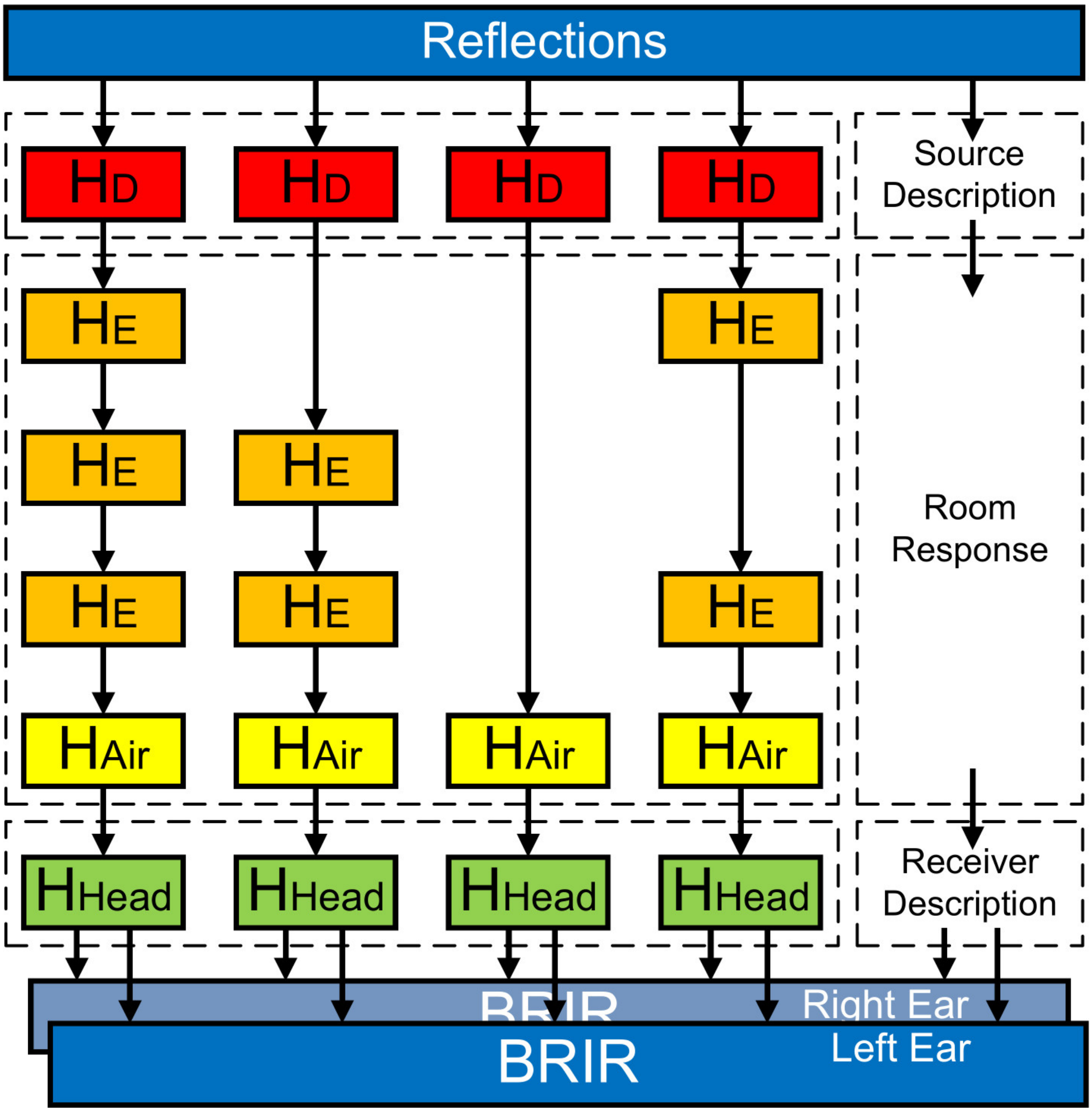
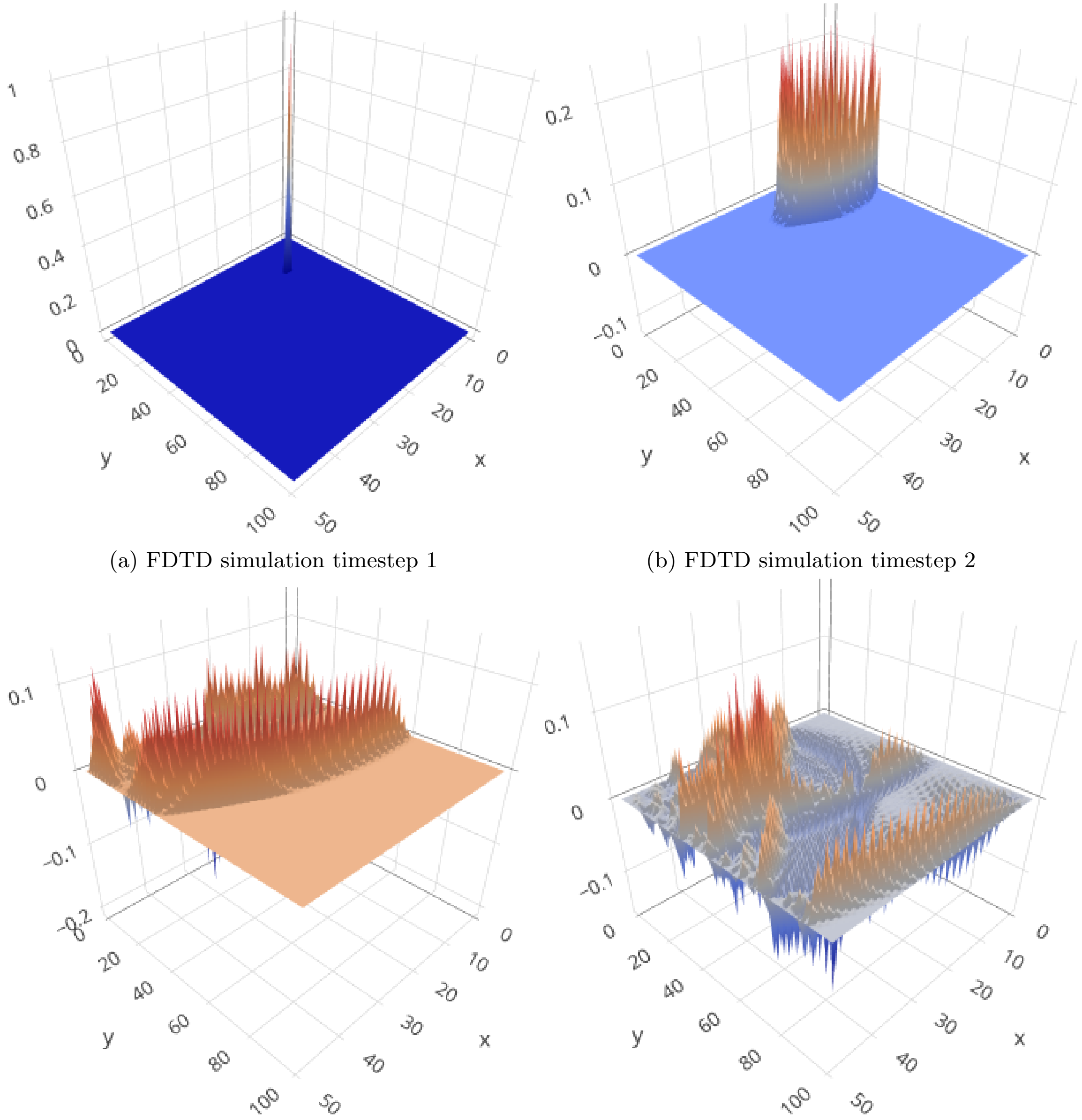


Figure 1.15: Diagram showing the anatomy of a reflection. Geometrical Acoustic techniques consider source directivity profiles and receiver descriptions (human listeners). The diagram shows how a reflection is generated from the source and is affected by source characteristics (directivity profiles). 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 1.16. Spatial discretisation allows for the representation of complex geometries and structures, while temporal discretisation captures the evolution of energy fields over time. 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 1.17. 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, 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



(a) FDTD simulation timestep 1

(b) FDTD simulation timestep 2

(c) FDTD simulation timestep 3

(d) FDTD simulation timestep 4

Figure 1.16: Visualisations of Finite-Difference Time-Domain simulation timesteps. Finite-Difference Time-Domain (FDTD) is a computational technique used to model sound propagation in virtual environments by solving the wave equation at discrete space and time intervals. The figures represent several timesteps of a running simulation, showing how sound pressure emitted from a source changes across timesteps.

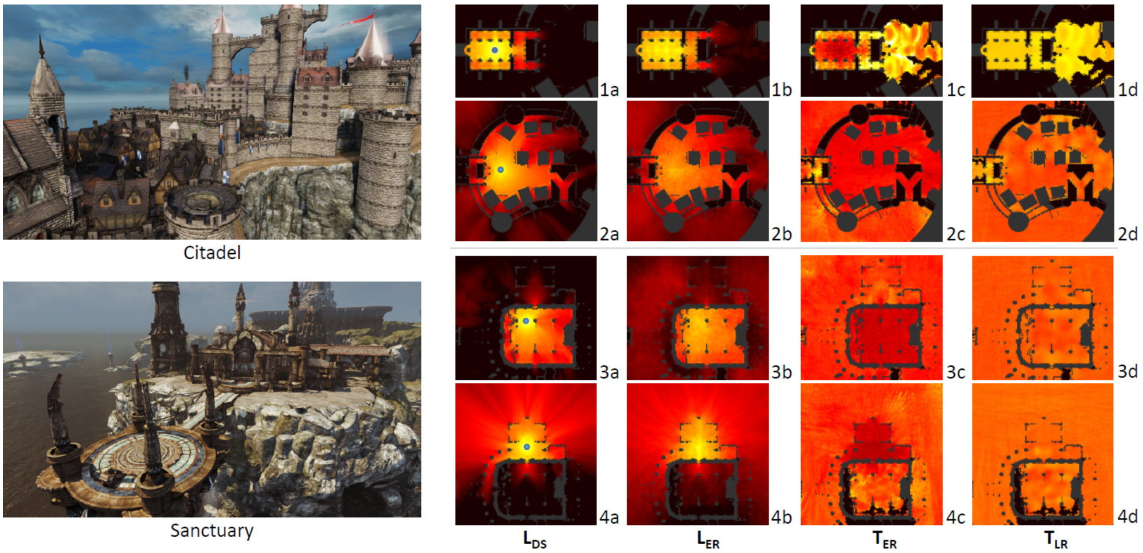


Figure 1.17: Demonstration of a state-of-the-art wave-based acoustic renderer: pressure fields are pre-computed within a complex virtual environment and encoded to allow game engines to decode and auralise sound sources in real-time (Raghuvanshi and Snyder, 2014).

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.

1.3.6 Summary

Ray-based techniques offer a compelling option for interactive applications. Their computational efficiency, relative simplicity, and adaptability in handling various scenarios effectively capture the essential acoustic characteristics of sounscapes. Unlike wave-based or hybrid methods, ray-based techniques can easily adapt to dynamic environments, aligning with the objectives of this work and constraints often found in AR platforms. Therefore, while the high accuracy of wave-based methods or the comprehensive nature of hybrid methods may have specific applications in areas of high-fidelity and complex propagation effects, 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 fast and efficient acoustic simulation in immersive applications.

1.4 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.

1.4.1 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.

1.4.2 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 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.

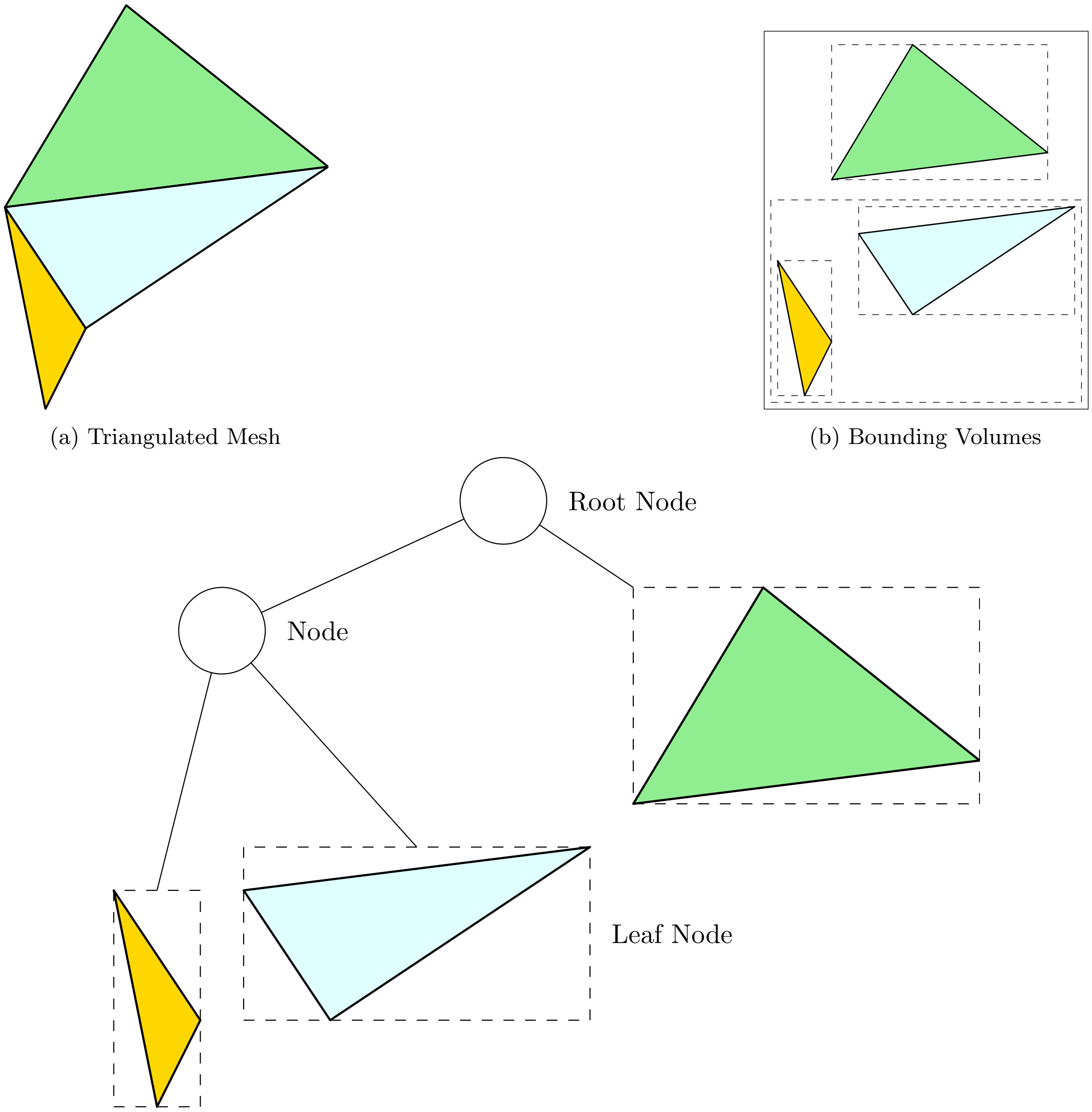
With subdivided regions obtained with partitioning planes, a binary tree, similar to the diagram shown in Figure 1.18, where a root node refers to the entire scene and branches into the first space subdivision, which recursively branches into further subdivisions, until a "leaf" region. A leaf region 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 1.18 (Pharr, Jakob and Humphreys, 2023).

In a tree, bounding volumes are generated to fit primitives from a given triangulated mesh (Figure 1.18a) and aggregated based on proximity (Figure 1.18b). 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 1.18c) 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



(a) Triangulated Mesh

(b) Bounding Volumes

(c) Tree

Figure 1.18: 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).

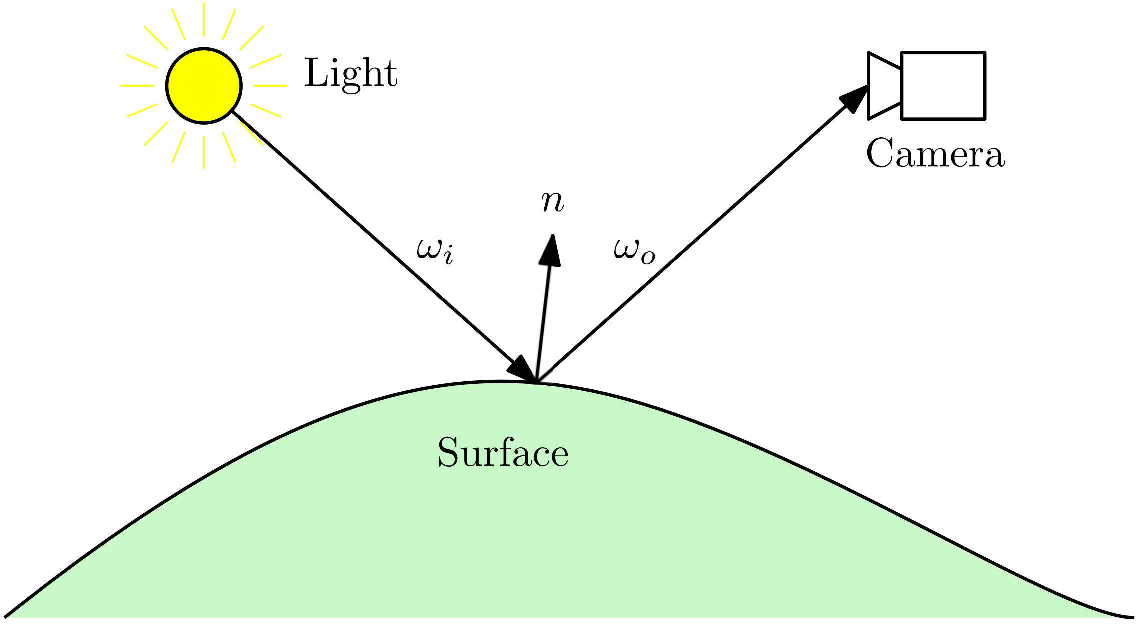


Figure 1.19: A simplistic material system visualised as a ray of light emitted by a source and colliding with a surface. The ray reflects around a surface normal and is detected by a virtual camera.

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).

1.4.3 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 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 1.19 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 1.19.

A material using a reflectance function can model realistic behaviour, allowing surfaces to express varying physical attributes like roughness or metallic characteristics. Figure 1.20 shows example functions simulating a rough and a glossy material, Figure 1.20a and 1.20b 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.

Materials in Sound Rendering

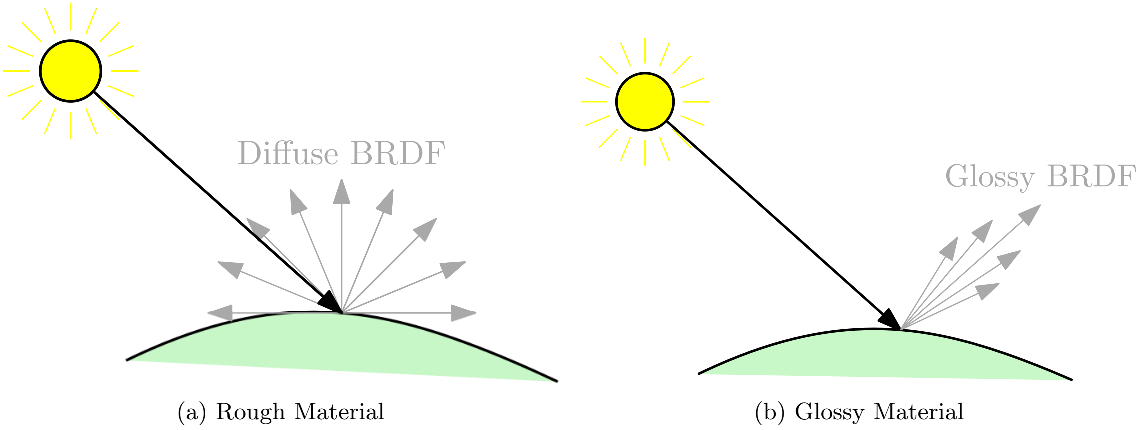
Defining materials translates to the acoustics domain, applying closely related principles defined in the visual domain. GA methods often share the same approach shown in Figure 1.19 by considering sound in VEs as propagating rays (or other geometry primitives) colliding with surfaces that reflect energy 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 **??** and **??** will discuss the use of materials in the context of the overarching aim of this thesis.

1.5 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



(a) Rough Material

(b) Glossy Material

Figure 1.20: 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.

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).

1.5.1 Image Processing

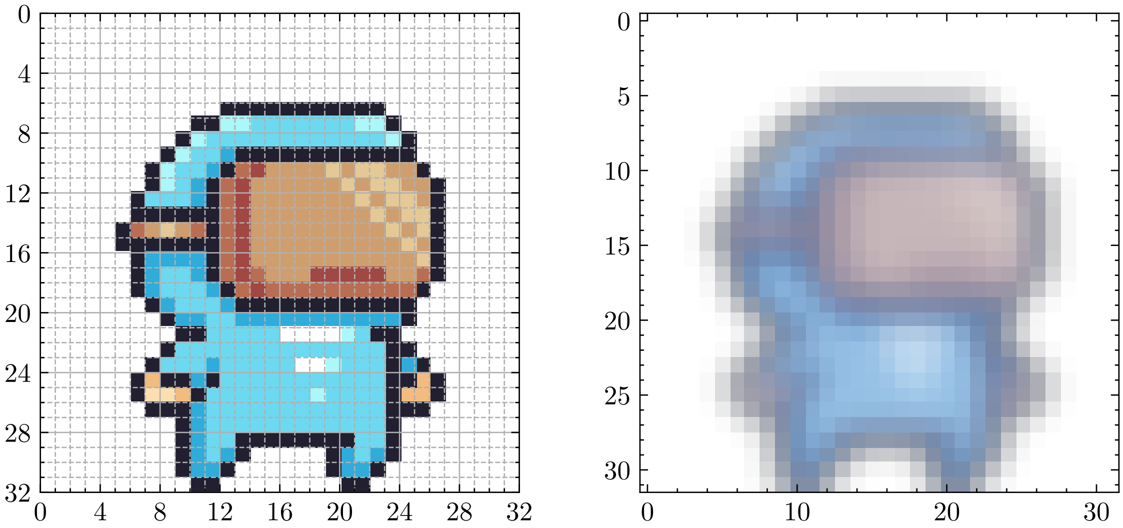


Figure 1.21: Visualisation of an image processing technique applied to an example image (left axes). On both axis pairs, abscissa and ordinate indicate pixel coordinates. A Gaussian blur filter is applied to the image by means of convolution applied to pixel intensity values representing the image. The right axes show the result.

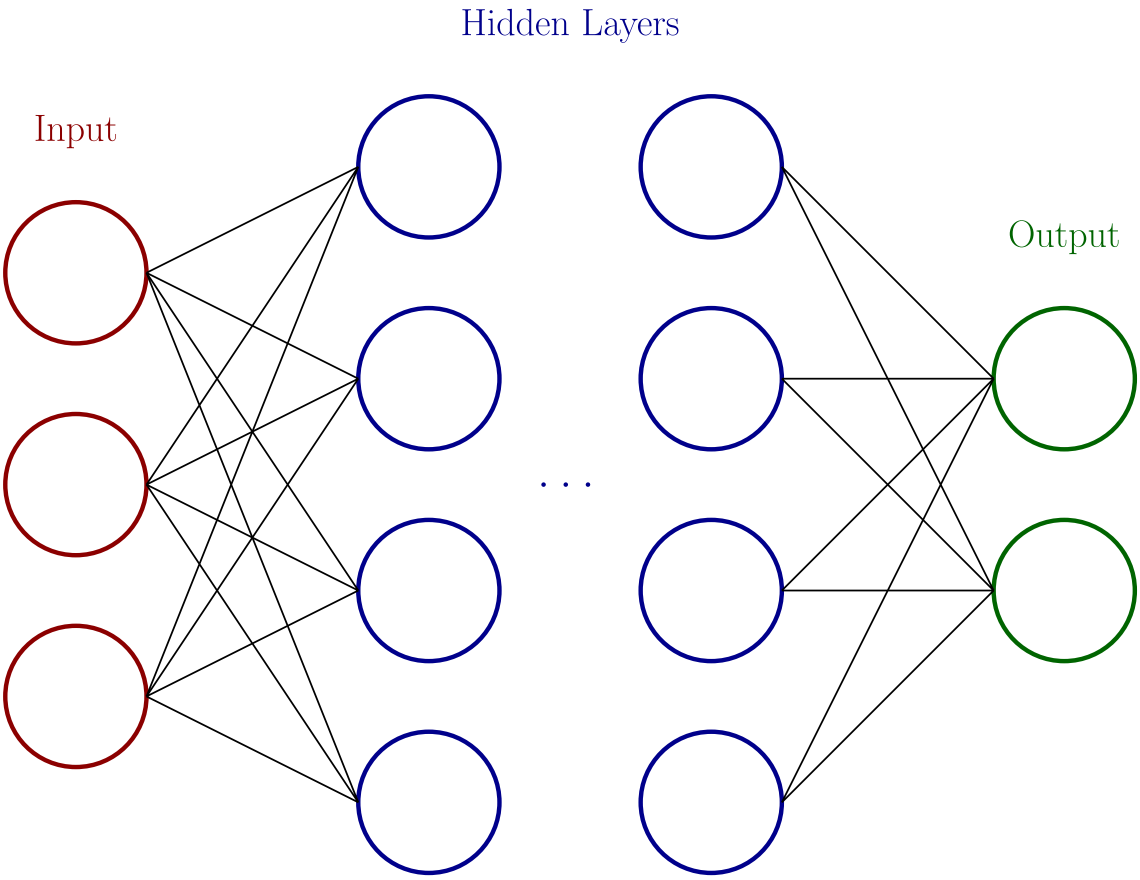


Figure 1.22: A visualisation of a basic neural network with fully connected hidden layers.

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 1.2.3, 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 1.21 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.

1.5.2 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 1.22. 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).

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.

1.5.3 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 **??**, 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 AR, 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).

1.6 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 Virtual Environments (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 pre-computation 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, it has gained popularity in computer games and immersive technology for entertainment and serious applications (Zhang et al., 2018). AR 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.

Hearing is a complex function that requires modelling of a human listener as well as considering environmental aspects.

Bibliography

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

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.

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.

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

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

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.

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.

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.

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.

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. 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.

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.

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.

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.

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.

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.

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.

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.

Lakka, E., Malamos, A.G., Pavlakis, K.G. and Ware, J.A., 2018. Spatial sound rendering–a survey. *Ijimai*, 5(3), pp.33–45.

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.

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.

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.

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.

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.

Patow, G. and Pueyo, X., 2003. A survey of inverse rendering problems. *Computer graphics forum*. Wiley Online Library, 4, pp.663–687.

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.

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.

Raghuvanshi, N. and Snyder, J., 2014. Parametric wave field coding for precomputed sound propagation. *Acm transactions on graphics (tog)*, 33(4), pp.1–11.

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.

Rindel, J.H., 2000. The use of computer modeling in room acoustics. *Journal of vibroengineering*, 3(4), pp.219–224.

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.

Savioja, L. and Svensson, U.P., 2015. Overview of geometrical room acoustic modeling techniques. *The journal of the acoustical society of america*, 138(2), pp.708–730.

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.

Schröder, D., 2011. *Physically based real-time auralization of interactive virtual environments*, vol. 11. Logos Verlag Berlin GmbH.

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.

Shenoi, B.A., 2005. *Introduction to digital signal processing and filter design*, vol. 169. John Wiley & Sons.

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*.

Szeliski, R., 2022. *Computer vision: algorithms and applications*. Springer Nature.

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.

Virtanen, T., 2006. *Sound source separation in monaural music signals*. Tampere University of Technology.

Vorländer, M., 2008. *Simulation of sound in rooms*. Springer.

Yannakakis, G.N. and Togelius, J., 2018. *Artificial Intelligence and Games*. https://gameaibook.org. Springer.

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.

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.