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

March 2024

Supervisors
Dr Carlo Harvey, Dr Maite Frutos-Pascual

Abstract

The planned thesis work 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. 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. 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 Augmented Reality 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  Date: 15th March, 2024

Mattia Colombo

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

AAR	Audio Augmented Reality.
AR	Augmented Reality.
BRIR	Binarual Room Impulse Response.
CNN	Convolutional Neural Network.
DNN	Deep Neural Network.
FDTD	Finite-Difference Time-Domain.
GA	Geometrical Acoustics.
GAN	Generative Adversarial Network.
HAS	Human Hearing System.
HMD	Head-Mounted Display.
HRTF	Head-Related Transfer Function.
IR	Impulse Response.
JND	Just-Noticeable Difference.
LoD	Level of Detail.
NAF	Neural Acoustic Field.
NN	Neural Network.

RIR	Room Impulse Response.
TPU	Tensor Processing Unit.
VE	Virtual Environment.
VR	Virtual Reality.
XR	Extended Reality.

Chapter 1

Advances in Visual-Acoustic Mapping Methods and Sound Rendering Pipelines

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

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

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

points to inform design choices in engineering systems proposed throughout the following Chapters. Discussions around existing work aim at both orienting the reader towards the goal of each domain associated with the thesis component and defining the value of the contributions stemming from this work.

1.1 Introduction

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

1.1.1 Current Trends of Interactions within Immersive Platforms

The domain of immersive acoustic has recently attracted more popularity thanks to the drive of XR platforms towards better and more efficient multimodal interactions. Over the last couple of decades have seen an increase in experiments towards interaction, rendering, visualisation and related fields generating major research interests (Kim et al., 2018).

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

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

users with hearing impairments. Such problems incentivise the field of AAR to develop audio pipelines considering and compensating for acoustic factors of the environment.

1.2 Review of Sound Rendering Pipelines for Immersive Environments

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

1.2.1 Advances in Sound Rendering

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

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

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

In acoustics, it is common to capture an environment adopting measurement techniques

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

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

Schissler, Mückl and Calamia (2021) recently presented a novel method for computing acoustic diffraction in real-time, which can adapt to GA frameworks. They target complex scenes typical of virtual and augmented reality applications. Their approach can overcome the shortcoming of GA techniques in approximating soundscapes, thanks to the ability to incorporate simulated propagation effects into their proposed sound rendering pipeline. The main contribution of their work is a mesh processing system that optimises diffraction simulations for environments expressed through dense geometry. Results gathered from their evaluation show that the technique is comparable to FDTD methods, obtaining a high degree of realism from generated propagation data. Due to the novelty and recency, there is a lack of psychoacoustic characterisation performed on their method, making it hard to

1.2.2 Differentiable Methods for Sound Rendering

Manocha et al. (2020) present a model for simulating sound fields using neural networks without pre-computing the wave field of an acoustic environment, predicting unseen objects

with arbitrary shapes in a VE for sound propagation at interactive rates. They train a geometrical neural network on annotated meshes to infer acoustic data associated with the represented object. Chen et al. (2022) introduce a novel task dependant on this mapping, *visual/acoustic matching*, which produces acoustic stimuli responding to a target space depicted in an image, given an input audio excerpt and an image of the environment in which excerpt propagated. The rapid development in DNN for multi-modal applications has opened new avenues in the field of sound propagation modelling, one of which tapped into visual-acoustic mapping, the process of determining relationships between visual and auditory features in audio-visual or immersive media.

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

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

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

Tang et al. (2020) present a novel scene-aware sound rendering system aimed at rendering audio considering acoustic characteristics of a given room, providing real-time audio effects

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

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

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

Their method takes a triangulated mesh representing the environment, which is fed into a series of graph NN to process vertex and edge information to create a graph encoding of the input scene, simplifying topology information. A modified Generative Adversarial Network (GAN) uses the constructed graph representation to generate an IR by computing a decay curve and optimising acoustic characteristics encoded in the graph representation using a generator and discriminator. Their method is tested on indoor scenes, evaluating T_{60} reverberation, direct-to-reverberant ratio and early decay times. The evaluation shows less than 10% error across all metrics, placing the method amongst one of the pioneering NN-based approaches.

Yang et al. (2020) present a method for synthesising RIRs, reproducing perceptually-convincing acoustics of real environments based on a small number of ultrasonic measurements. The method consists of using a loudspeaker and a microphone to record an ultrasonic IRs that can be transformed into an octave-IRs by approximating reflection decay curves. Octave-IRs constructed from Gaussian noise and modelled using estimated

decay curves are combined into a final monoaural response. The authors test the approach in two indoor spaces (a lounge and a classroom) demonstrating that the technique can generate perceptually plausible auditory stimuli and showing potential application for AR platforms. However, the apparatus adopted would require the wearable AR device to be equipped with a recording setup to sample ultrasonic responses.

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

1.2.3 Discussion

The spectrum of sound rendering techniques continues to refine existing methods, increasing their efficiency and applying them to newer platforms and use cases, as well as present novel methods leveraging recent advances in deep learning. Table 1.1 summarises the crucial techniques, reporting advantages and limitations in relation to their employment in real-time auralisation applications.

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

GA-derived methods, such as Schissler and Manocha (2016); Savioja and Svensson (2015b); Schröder (2011)'s, share some inherent limitation of the wider geometrical acoustics family; though there is a deeper understanding of their perceptual impact and are generally easier to scale for platforms with limited computational resources.

Table 1.1: A summary of experimental methods for sound propagation reviewed. These methods vary depending on the task perfomed within the VE, its underline architecture. These are compared based on their inputs and efficiency.

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

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

1.3 Material Recognition for Rendering Tasks

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

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

1.3.1 Supervised Material Recognition Techniques

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

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

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

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

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

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

1.3.2 Unsupervised and Semi-Supervised Alternatives

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

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

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

materials depicted by the input image.

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

1.3.3 Discussion

Considering the broad field of computer vision and focusing on techniques that have a direct application to the retrieval of acoustic characteristics from VEs and assigning properties to environment geometry, the area is generally underdeveloped and has potential for improvements. Table 1.2 summarises relevant experimental methods discussed.

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

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

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

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

Table 1.2: A summary of key techniques for material recognition that can apply to VEs. These are largely based on convolutional neural network layers extracting features from input image data, which virtual cameras can often provide as renders. These methods predict semantic features of materials represented in the input image data, allowing mappings visual features to acoustic characteristics of enviroment geometry.

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

1.4 Human Factors and Perceptual Rendering

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

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

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

1.4.1 Perception of Audio Quality

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

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

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

1.4.2 Psychoacoustic Characterisation of Sound Propagation Methods

In light of the discussion on novel sound rendering pipelines in earlier sections (see Section 1.2.3), there is a rising need for profiling the psychoacoustic factors of simulated auditory stimuli. González-Toledo et al. (2023) present a toolbox providing an acoustic binaural rendering system, exposing parameters that researchers can measure for subjective evaluations. Their proposed framework provides control over listener pose information, sound source spatial information and binaural rendering parameters. In addition, they enable annotations on audio stimuli, allowing participants to save contextual information related to tasks administered. A point of expansion for this method may be the limited reverberation models available: the toolbox would benefit from interfacing with arbitrary sound propagation systems, allowing researchers to test recent experimental advances in sound rendering.

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

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

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

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

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

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

1.4.3 Findings and Limitations

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

with reconstruction of real space surrounding the viewer. The dynamic nature of the platform often presents varying accuracy and precision in recognising and tracking real space and it is crucial to define the error margin in simulated displays before the user notices incoherence in stimuli.

1.5 Conclusions

Overall, the fields of sound rendering and sound propagation in immersive environments have advanced by significant strides into cutting-edge differentiable methods for simulating auditory stimuli. From approximating acoustic characteristics of a soundscape from a single photograph (Singh et al., 2021) to generating thousands of IRs from a 3D representation of a given environment (Ratnarajah et al., 2022), there is now a plethora of methods that can adapt to varying needs of realism and accuracy.

- The rising development of computer vision methods is generating momentum towards novel sound propagation methods
- Novel sound propagation methods could potentially be feasible for AR platforms
- But we need to profile psychoacoustic factors of standard methods before venturing in experimental sound propagation in ar
- Even though GA methods are outdated, they have well-defined requirements and perceptual responses.

Bibliography

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

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

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

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

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.

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

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y., 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arxiv preprint arxiv:1709.06158*.

Chen, C., Gao, R., Calamia, P. and Grauman, K., 2022. Visual acoustic matching. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.18858–18868.

Chen, L., Tang, W., John, N., Wan, T.R. and Zhang, J.J., 2018. Context-aware mixed reality: A framework for ubiquitous interaction. *arxiv preprint arxiv:1803.05541*.

Chen, M., Su, K. and Shlizerman, E., 2023. Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. *Proceedings of the ieee/cvf international conference on computer vision*. pp.7853–7862.

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)*.

Dourado, A., Campos, T.E. de, Kim, H. and Hilton, A., 2019. Edgenet: Semantic scene completion from rgb-d images. *arxiv preprint arxiv:1908.02893*.

Gaur, U. and Manjunath, B.S., 2020. Superpixel embedding network. *Ieee transactions on image processing*, 29, pp.3199–3212. Available from: https://doi.org/10.1109/TIP.2019.2957937.

González-Toledo, D., Molina-Tanco, L., Cuevas-Rodriguez, M., Majdak, P. and Reyes-Lecuona, A., 2023. The binaural rendering toolbox. a virtual laboratory for reproducible research in psychoacoustics. *Forum acusticum*.

Hacihabiboglu, H., De Sena, E., Cvetkovic, Z., Johnston, J. and Smith III, J.O., 2017. Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. *Ieee signal processing magazine*, 34(3), pp.36–54.

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.

Hulusic, V., Harvey, C., Debattista, K., Tsingos, N., Walker, S., Howard, D. and Chalmers, A., 2012. Acoustic rendering and auditory–visual cross-modal perception and interaction. *Computer graphics forum*. Wiley Online Library, vol. 31, pp.102–131.

Jillings, N., Moffat, D., De Man, B. and Reiss, J.D., 2015. Web Audio Evaluation Tool: A browser-based listening test environment. *12th sound and music computing conference*.

Kiechle, M., Storath, M., Weinmann, A. and Kleinsteuber, M., 2018. Model-based learning of local image features for unsupervised texture segmentation. *Ieee transactions on image processing*, 27(4), pp.1994–2007.

Kim, H., Hernaggi, L., Jackson, P.J. and Hilton, A., 2019. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. *2019 ieee conference on virtual reality and 3d user interfaces (vr)*. IEEE, pp.120–126.

Kim, K., Billinghurst, M., Bruder, G., Duh, H.B.L. and Welch, G.F., 2018. Revisiting trends in augmented reality research: A review of the 2nd decade of ismar (2008–2017). *Ieee transactions on visualization and computer graphics*, 24(11), pp.2947–2962.

Lagunas, M., Malpica, S., Serrano, A., Garces, E., Gutierrez, D. and Masia, B., 2019. A similarity measure for material appearance. *arxiv preprint arxiv:1905.01562*.

Li, D., Langlois, T.R. and Zheng, C., 2018. Scene-aware audio for 360 videos. *Acm transactions on graphics (tog)*, 37(4), pp.1–12.

Liang, S., Huang, C., Tian, Y., Kumar, A. and Xu, C., 2023. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arxiv preprint arxiv:2309.15977*.

Liebetrau, J., Nagel, F., Zacharov, N., Watanabe, K., Colomes, C., Crum, P., Sporer, T. and Mason, A., 2014. Revision of rec. itu-r bs. 1534. *Audio engineering society convention 137*. Audio Engineering Society.

Lokki, T. and Grohn, M., 2005. Navigation with auditory cues in a virtual environment. *Ieee multimedia*, 12(2), pp.80–86.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the ieee conference on computer vision and pattern recognition*. USA: IEEE, pp.3431–3440.

Manocha, P., Finkelstein, A., Jin, Z., Bryan, N.J., Zhang, R. and Mysore, G.J., 2020. A differentiable perceptual audio metric learned from just noticeable differences. *arxiv preprint arxiv:2001.04460*, 1(1), p.1.

Manocha, P., Jin, Z., Zhang, R. and Finkelstein, A., 2021. Cdpam: Contrastive learning for perceptual audio similarity. *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp.196–200.

Mansour, N., Marschall, M., May, T., Westermann, A. and Dau, T., 2021. Speech intelligibility in a realistic virtual sound environment. *The journal of the acoustical society of america*, 149(4), pp.2791–2801.

Mao, T., Zhang, W., He, H., Lin, Y., Kale, V., Stein, A. and Kostic, Z., 2018. Aic2018 report: Traffic surveillance research. *Proceedings of the ieee conference on computer vision and pattern recognition workshops*. pp.85–92.

Mehra, R., Raghuvanshi, N., Savioja, L., Lin, M.C. and Manocha, D., 2012. An efficient gpu-based time domain solver for the acoustic wave equation. *Applied acoustics*, 73(2), pp.83–94.

Mehra, R., Rungta, A., Golas, A., Lin, M. and Manocha, D., 2015. Wave: Interactive wave-based sound propagation for virtual environments. *Ieee transactions on visualization and computer graphics*, 21(4), pp.434–442.

Naef, M., Staadt, O. and Gross, M., 2002. Spatialized audio rendering for immersive virtual environments. *Proceedings of the acm symposium on virtual reality software and technology*. pp.65–72.

Park, S.M. and Kim, Y.G., 2022. A metaverse: Taxonomy, components, applications, and open challenges. *Ieee access*, 10, pp.4209–4251.

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.

Qiu, W. and Yuille, A., 2016. Unrealcv: Connecting computer vision to unreal engine. *European conference on computer vision*. Springer, pp.909–916.

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

Ratnarajah, A., Tang, Z., Aralikatti, R. and Manocha, D., 2022. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. *Proceedings of the 30th acm international conference on multimedia*. pp.924–933.

Reilly, A. and McGrath, D., 1995. Convolution processing for realistic reverberation. *Audio engineering society convention 98*. Audio Engineering Society.

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.

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

Savioja, L. and Svensson, U.P., 2015b. Overview of geometrical room acoustic modeling techniques. *The journal of the acoustical society of america*, 138(2), 08, pp.708–730. https://pubs.aip.org/asa/jasa/article-pdf/138/2/708/13242401/708_1_online.pdf, Available from: https://doi.org/10.1121/1.4926438.

Schissler, C., Loftin, C. and Manocha, D., 2017. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *Ieee transactions on visualization and computer graphics*, 24(3), pp.1246–1259.

Schissler, C. and Manocha, D., 2016. Interactive sound propagation and rendering for large multi-source scenes. *Acm transactions on graphics (tog)*, 36(4), p.1.

Schissler, C., Mückl, G. and Calamia, P., 2021. Fast diffraction pathfinding for dynamic sound propagation. *Acm transactions on graphics (tog)*, 40(4), pp.1–13.

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

Schwartz, G. and Nishino, K., 2019. Recognizing material properties from images. *Ieee transactions on pattern analysis and machine intelligence*, 1(1), p.1.

Series, B., 2014. Method for the subjective assessment of intermediate quality level of audio systems. *International telecommunication union radiocommunication assembly*.

Singh, N., Mentch, J., Ng, J., Beveridge, M. and Drori, I., 2021. Image2reverb: Cross-modal reverb impulse response synthesis. *Proceedings of the ieee/cvf international conference on computer vision (iccv)*. pp.286–295.

Slater, M., Khanna, P., Mortensen, J. and Yu, I., 2009. Visual realism enhances realistic response in an immersive virtual environment. *Ieee computer graphics and applications*, 29(3), pp.76–84.

Somayazulu, A., Chen, C. and Grauman, K., 2023. Self-supervised visual acoustic matching. *arxiv preprint arxiv:2307.15064*.

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*.

Tang, Z., Bryan, N.J., Li, D., Langlois, T.R. and Manocha, D., 2020. Scene-aware audio rendering via deep acoustic analysis. *Ieee transactions on visualization and computer graphics*, 26(5), pp.1991–2001.

Vorländer, M., Schröder, D., Pelzer, S. and Wefers, F., 2015. Virtual reality for architectural acoustics. *Journal of building performance simulation*, 8(1), pp.15–25.

Xia, X. and Kulis, B., 2017. W-net: A deep model for fully unsupervised image segmentation. *arxiv preprint arxiv:1711.08506*.

Yang, J., Barde, A. and Billinghurst, M., 2022. Audio augmented reality: a systematic review of technologies, applications, and future research directions. *journal of the audio engineering society*, 70(10), pp.788–809.

Yang, J., Pfreundtner, F., Barde, A., Heutschi, K. and Sörös, G., 2020. Fast synthesis of perceptually adequate room impulse responses from ultrasonic measurements. *Proceedings of the 15th international audio mostly conference*. pp.53–60.