ABSTRACT

Title of Dissertation: EFFICIENT ACOUSTIC SIMULATION FOR LEARNING-BASED VIRTUAL AND REAL-WORLD AUDIO PROCESSING

Zhenyu Tang
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Dinesh Manocha
Department of Computer Science

Sound propagation is commonly known to be air pressure perturbations due to vibrating/moving objects. The energy of sound gets attenuated by transmitting in the air over a distance and by being absorbed at other objects' surfaces. Numerous researchers have focused on devising better acoustic simulation methods to model sound propagation in a more realistic manner. The benefits of accurate acoustic simulations include but are not limited to computer-aided acoustic design, acoustic optimization, synthetic speech data generation, and immersive audio-visual rendering for mixed reality. However, acoustic simulation has been underexplored for relevant virtual and real-world audio processing applications. The main challenges in adopting accurate acoustic simulation methods include the tradeoff between accuracy and time-space cost and the difficulties in acquiring and reconstructing acoustic scenes in the real world.

In this dissertation, we propose novel methods to overcome the above challenges by leveraging the inferential power of deep neural networks, and combining them with interactive acoustic simulation techniques. First, we develop a neural network model that can learn the acoustic scattering fields of different objects given their

3D representations as the input. This works facilitates the inclusion of wave acoustic scattering effects in interactive sound rendering applications, which used to be difficult without intensive pre-computation. Second, we incorporate a deep acoustic analysis neural network into the sound rendering pipeline to allow the generation of sounds that are perceptually consistent with real-world sounds. This is achieved by predicting acoustic parameters at run-time from real-world audio samples and optimizing simulation parameters accordingly. Finally, we build a pipeline that utilizes general 3D indoor scene datasets to generate high-quality acoustic room impulse responses and demonstrate the usefulness of the generated data on several practical speech processing tasks. Our results demonstrate that by leveraging state-of-the-art physics-based acoustic simulation and deep learning techniques, realistic simulated data can be generated to enhance sound rendering quality in the virtual world and boost the performance of audio processing tasks in the real world.

EFFICIENT ACOUSTIC SIMULATION FOR LEARNING-BASED VIRTUAL AND REAL-WORLD AUDIO PROCESSING

by

Zhenyu Tang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Dinesh Manocha, Chair/Advisor
Professor Carol Espy-Wilson, Dean's Representative
Professor Ming C. Lin
Professor Ramani Duraiswami
Professor Nirupam Roy

Acknowledgements

The past fives years in my life has been a journey full of discoveries - both in scientific research and in learning my own pursuits, limits, and potential. The completion of this dissertation would not have been possible without the help from many people. First, I would like to express my deepest appreciation to my advisor Prof. Dinesh Manocha, for lending me his invaluable experience and insights in conducting high-quality research. While leading the large GAMMA research group, he never hesitated to provide me the necessary resources, and has always trusted and supported my decisions.

I would like to extend my sincere thanks to my collaborators from both academia and industry. Thank you all for dedicating your time to my preliminary ideas and helping me turn them into meaningful findings that I would not have the bandwidth to achieve alone - Dr. Nicolas Morales, Dr. Dingzeyu Li, Dr. Timothy Langlois, Dr. Nicholas Bryan, Dr. Dong Yu, Dr. Buye Xu, Hsien-Yu Meng, Anton Jeran Ratnarajah, Rohith Aralikatti. I especially want to thank Dr. Atul Rungta, a GAMMA member who discussed acoustic research with me in my early days at graduate school and encouraged me to set forth on this exciting new line of research. I am also grateful to my undergraduate advisor Prof. Hongzhi Wu, for his initial guidance and influence on me, which motivated me into computer science research in the first place.

Doctoral research is known to be physically and mentally demanding, and I am very privileged to be supported by my loving family and friends during this period.

I deeply appreciate the support from my parents, who have selflessly cared for my well-being all the time, even though being thousands of miles apart.

Lastly, a special thank you to my partner Ran for all her kindness and love, who has filled my life with joyful expectations. I am so lucky to have encountered you in this period of my life.

Table of Contents

[1] The work in this chapter has been published in Tang et al. (2019a)

[2]The work in this chapter has been published in Tang et al. (2021)
[3]The work in this chapter has been published in Tang et al. (2019b, 2020, 2022)

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Accurate and efficient simulation of physics has been an important topic for computer science and applied math research. Over past decades, the rapid development of computing hardware and software has facilitated many simulation techniques that transfer theories to professional tools that we use to interpret and predict real-world physics. The application of computer simulation has seen huge success in many fields, including weather forecasting, industrial computer-aided design (CAD), flight simulation for personnel training, digital entertainment, etc.

One research area that has gained increased interest in recent years is efficient acoustic simulation for audio processing. Audio signals corresponding to music, speech, and non-verbal sounds in the real world encode rich information regarding the surrounding environment. Many digital signal processing algorithms and audio deep learning techniques have been proposed to extract information from audio signals. These methods are widely used for different applications such as music information retrieval, automated speech recognition, sound separation and localization, sound synthesis and rendering, etc. Acoustic simulation techniques are often used in audio

processing tasks where real-world audio data is difficult to acquire.

In contrast to realistic visual rendering techniques, which have been the main topic for computer graphics research, acoustic simulation has not been as widely adopted by related applications such as computer games, digital film making, and virtual reality. While state-of-the-art acoustic simulation techniques can add high-fidelity physics-based sounds to these applications, there are still barriers that make them less practical to be used for situations where: (1) it is difficult to faithfully describe the acoustic environment and give accurate inputs to the simulator or (2) there is a strict requirement for computing efficiency. As a result, digital audio in many applications is often post-processed by professionals subjectively, even though they can deviate hugely from physically realistic sounds. However, there are still areas where accurate acoustic simulation is irreplaceable, including but not limited to computer-aided acoustic design, environmental acoustic optimization, and immersive audio-visual rendering for mixed reality. This motivates us to investigate how to use acoustic simulation techniques practically in various virtual and real-world audio processing tasks.

1.2 Challenges and Contributions

In contrast to previous research, which focused on theoretical acoustic simulations, my dissertation research aims to bridge the gap between theoretical methods and their applications in practical audio processing tasks. One challenge is the trade-off between simulation accuracy and time-space cost. Conventional numeric wave solvers based on the first-principal wave equation provides the most accurate results and can be validated with real-world measurements. However, they usually scale poorly with simulation frequency and scene scale, making them unsuitable for large simulations (in number or scale). Another challenge is incorporating synthetic sound in real-world

settings, where the simulated sound needs to be consistent with the recorded sound. This requires the simulator to be *scene-aware*: the sound simulation setups need to align well with the real-world scene. The main difficulty comes from two parts: (1) The real-world scene configurations are not always well known, and they need to be empirically inferred or measured on-site. Prior solutions are either inaccurate or not user-friendly. (2) The wave effects are essential for low-frequency components but are poorly approximated by state-of-the-art real-time acoustic simulators. A large amount of pre-computation time is needed to incorporate results from wave-based solvers.

In this dissertation, we develop a series of algorithms and tools to overcome the above challenges and verify their effectiveness via real-world acoustic benchmarks and subjective listening studies. Our main contributions can be summarized in the following three aspects:

Scene-Aware Audio for Mixed Reality We propose a novel method that allows automatic analysis of real-world acoustics for generating virtual sounds that are perceptually consistent with real-world sounds. This is achieved by training acoustic parameter predictors[1] from a large amount of simulated data in various room environments. The scene analysis can be performed on new real-world scenes on-the-fly while still generating plausible sound rendering that is consistent with the recorded sound in the same environment.

Fast Learning-Based Acoustic Scattering We present a novel approach to approximate the acoustic scattering field of any geometric object using neural networks for interactive sound propagation of highly dynamic scenes[2]. Our approach is general and makes no assumption about the scene or the motion or topology of the objects. We exploit properties of the acoustic scattering field

[1]Code available at https://github.com/GAMMA-UMD/deep-acoustic-analysis
[2]Code available at https://github.com/GAMMA-UMD/Fast3DScattering-release

of objects for lower frequencies and use neural networks to learn this field from geometric representations of the objects.

High-Quality Synthetic Acoustic Datasets We propose methods to simulate high-quality room impulse responses (RIRs) using our physics-based geometric acoustic simulator[3] and a hybrid geometric-acoustic simulation approach. We address the challenges in accurately modeling acoustic phenomena, including occlusion, specular and diffuse reflections, and diffraction and demonstrate the benefits of our method in speech recognition, speech enhancement, key-word spotting, and direction of arrival estimation tasks.

1.3 Organization

The rest of the dissertation is organized as follows: **Chapter 2** gives a comprehensive background and overview of previous research related to topics in this dissertation. **Chapters 3, 4, and 5** present our work on scene-aware audio, fast 3D acoustic scattering, and high-quality acoustic datasets generation, respectively. Then we discuss the limitations, envision several future research directions, and conclude my dissertation in **Chapter 6**.

[3]Code available at https://github.com/GAMMA-UMD/pygsound

Chapter 2

Background and Previous Research

2.1 Room Acoustics

2.1.1 Room Impulse Response

Sound is commonly known to be air pressure perturbations caused by vibrating/moving objects. One conventional way to define a sound signal is by describing the air pressure perturbation (in *Pascal*) as a function of time, denoted as $s(t)$. A sound signal can get attenuated by transmitting in the air over a distance and by being absorbed at other objects' surfaces. A room, or more generally, an acoustic environment, affects any sound signal excited within it before the sound is received by a listener (e.g., human ears or microphones). The transformation from the input signal to the output signal can be characterized by the room impulse response (RIR), which specifies how a signal is delayed and attenuated in a linear time-invariant (LTI) system. If we denote the RIR by $h(t)$, we can write the input-output relationship as

$$s_{out}[t] = s_{in}[t] \circledast h[t], \tag{2.1}$$

where $\circledast$ denotes 1D convolution. More formally, the RIR is defined as the output signal in response to an impulsive input signal represented by the Dirac function $\delta(t)$, which is zero everywhere except at the origin, where it is infinite (Kuttruff, 2016). Conventionally, an RIR can be decomposed into three parts: the direct response, early reflections, and the late reverberation. The direct response is determined by the visibility between the source and listener. Early reflections have stronger energy peaks and follow shortly after the direct response. The late reverberation is the result of high-order reflections and is more random. A typical RIR energy distribution is shown in Figure 2.1.

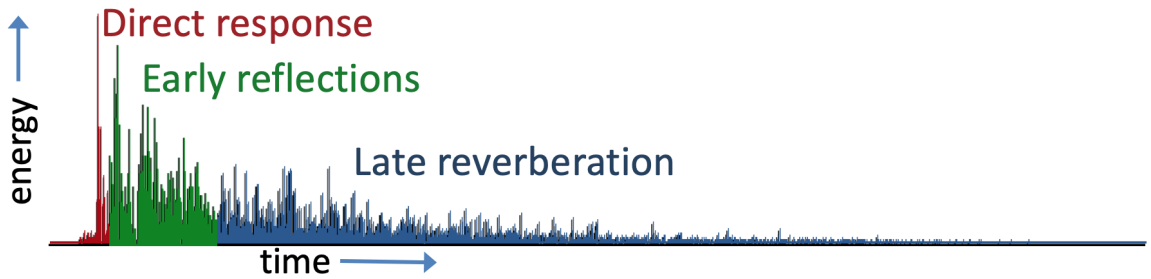


Figure 2.1: Energy distribution of an impulse response in time.

The Fourier transform of the RIR is known as the frequency response of the room, which can reveal the frequency dependence for changes in sound intensity and phase. Once the RIR for a particular source-listener pair in a room is known, it can be used as a digital filter to reproduce any sound signal as if the sound was emitted in the same room.

In terms of recording RIRs in the real world, the most reliable methods involve playing and recording Golay codes (Foster, 1986) or sine sweeps (Farina, 2000) at high signal-to-noise ratios. Also required are fairly high-quality speakers and microphones with flat frequency responses, small harmonic distortion, and little cross-talk. The speaker and microphone should be acoustically separated from surfaces, i.e., they shouldn't be placed directly on tables (else surface vibrations could contaminate the signal). Clock drift between the source and microphone must be accounted for (Bryan

et al., 2010). Alternatively, balloon pops or hand claps have been proposed for easier RIR estimation, but require additional post-processing (Abel et al., 2010; Seetharaman and Tarzia, 2012).

2.1.2 Reverberation Time

The sound signal emitted by any finite-time sound source will eventually drop its amplitude below the human hearing threshold as the signal is absorbed by its propagating medium (i.e., air) and boundaries in the room. Such energy decay is often exponential with respect to time. One acoustic metric commonly used to describe the decay rate is the *reverberation time* (Sabine, 1927), defined as the time interval in which the sound pressure level for an impulse input decays by 60dB from its onset, written as T_{60}. T_{60} can be directly evaluated from a recorded RIR (Karjalainen et al., 2001). Conventional rooms may have reverberation times from 0.3s to 2.0s, mostly depending on the size and furnishing of the room. Extremely large environments and reverberation chambers can have reverberation times up to 10s. In theory, a signal in the free field (e.g., vacuum) will have a T_{60} of 0s. An acoustically treated anechoic chamber can also have its T_{60} close to zero; sounds that are recorded in this condition is often called "dry" or "clean" sounds.

2.1.3 Room Modes

As the sound signal propagates in a room, standing waves can form at discrete resonant frequencies whose wavelength λ satisfies $\lambda = \frac{2L}{n}, n = 1, 2, 3, ...$, where L is the room dimension along some direction (e.g., axial, tangential, oblique). Room modes are the collection of these resonant frequencies and consist of mostly low frequencies below the Schroeder frequency (in Hz) $f_c = 2000\sqrt{\frac{T_{60}}{V}}$ (Schroeder, 1996), where T_{60} is the reverberation time in seconds and V is the volume of the room in m^3. For typical residential rooms, f_c will be lower than 200Hz. At these resonant frequencies, the

sound pressure tends to be significantly modified in different locations in the same room, which can cause problems for accurate sound reproduction. Various methods have been devised to remove room modes using equalization filters (Cecchi et al., 2018).

2.2 Acoustic Simulation

Acoustic simulation can involve the process of sound generation and propagation. In this dissertation, we focus on the sound propagation aspect and refer readers interested in modal sound simulation to Zheng and James (2011).

2.2.1 Wave Acoustics

First, we describe the theoretical foundation of wave acoustic simulations. A scalar acoustic pressure field, $P(\mathbf{x}, t)$, satisfies the inhomogeneous wave equation

$$\frac{\partial^2 P(\mathbf{x}, t)}{\partial t^2} - c^2 \nabla^2 P(\mathbf{x}, t) = f(\mathbf{x}, t), \tag{2.2}$$

where c is the speed of sound, $\mathbf{x}$ is the 3D coordinate, and $f(\mathbf{x}, t)$ is the forcing term, usually representing some driving source signal. An RIR can be obtained by setting $f(\mathbf{x}, t)$ to an impulse signal at a source location $\mathbf{x}_s$, fixing $P(\mathbf{x}, t)$ at the receiver location $\mathbf{x}_r$ and extracting its time-varying component. The wave equation can be solved numerically using the finite-difference time domain (FDTD) (Botteldooren, 1995) method or in the frequency domain using the finite-element (FEM) method (Thompson, 2006), the boundary-element (BEM) method (Wrobel and Kassab, 2003), the adaptive rectangular decomposition (ARD) method (Raghuvanshi et al., 2009), etc. These methods are also referred to as *wave-based methods*. Their computation complexity increases linearly with the size of the environment (surface area or volume) and as a third or fourth power of frequencies. As a result, they are limited to lower

frequencies and offline simulations (Raghuvanshi et al., 2010; Mehra et al., 2013; Yeh et al., 2013).

2.2.2 Geometric Acoustics

When the wavelength of the sound is smaller than the size of the obstacles in the environment, the sound wave can be treated in the form of a ray, which is the key idea of geometric acoustics. Typical geometric acoustic simulation techniques include the image method (Allen and Berkley, 1979), which only models specular reflections, path tracing methods (Taylor et al., 2009, 2012b; Schissler and Manocha, 2016, 2018) based on efficient Monte Carlo path tracing (Kajiya, 1986), and beam or frustum tracing methods (Funkhouser et al., 1998b; Chandak et al., 2008). These techniques are designed to run magnitudes faster than wave acoustic solvers and can be enhanced to simulate low-frequency diffraction effects. This category includes the time-domain Biot-Tolstoy-Medwin (BTM) model, which can be expensive and is also limited to offline computations (Svensson et al., 1999). For interactive applications, commonly used techniques are based on the uniform theory of diffraction (UTD), which is a less accurate frequency-domain model that can generate plausible results in some cases (Tsingos et al., 2001; Taylor et al., 2012a; Schissler et al., 2014). Moreover, the complexity of edge-based diffraction algorithms can increase exponentially with the maximum diffraction order. A more extensive review of geometric acoustic techniques can be found in Liu and Manocha (2020).

2.3 Acoustic Scene Representation

Room acoustics depend on many factors. Room geometry and acoustic materials together can greatly affect how a sound signal is being modified by propagation. In larger and less absorbent rooms, the sound signal can keep travelling for longer

times before vanishing (hence longer T_{60}), and vice versa. For the same room size, a rectangular room would have different room modes if it were differently shaped. In practice, modern 3D vision techniques can be used on commodity devices to construct geometry proxies from a video recording of the real-world scene in the form of dense 3D point clouds or meshes (Zhi et al., 2019; Bloesch et al., 2018).

Acoustic materials are often described in terms of how they react to incoming sound. The complex acoustic impedance indicates how much sound pressure would be generated in response to vibrations in the acoustic medium (e.g., air). This indicator is being used by many wave acoustic solvers but needs to be measured in controlled lab settings (Hiremath et al., 2021), making it less accessible for most materials. In the geometric acoustics context, the absorption and scattering coefficients are more commonly used, though they are also the sources of errors for acoustic simulations (Vorländer, 2013). The absorption coefficient $\alpha \in [0, 1]$ is defined as the fraction of sound energy at a specific frequency that is absorbed by the material. While the measurement of α also requires a reverberation chamber environment, the frequency-dependent absorption coefficients of many commonly seen materials have been measured and compiled as acoustic material databases. The energy that is not absorbed can be further described using the scattering coefficient $s \in [0, 1]$, which represents the fraction of sound that is diffusely reflected (e.g., following Lambertian distribution), while the remaining fraction is specularly reflected (i.e., having high directivity). However, the scattering coefficient is highly relevant to the roughness of the surface (Christensen and Rindel, 2005), and available measured data are few. In theory, bidirectional reflection distribution functions (BRDFs) that are widely used in computer graphics can more accurately describe the interaction between an incoming sound and the material (Mückl and Dachsbacher, 2014), but acoustic BRDFs have not been commonly measured.

Once the room geometry and materials are well-defined, the soundfield can be

simulated for any given sources. Wave-based methods solve the soundfield for the whole space within the acoustic scene, so the sound pressure at specific locations can be evaluated after the simulation finishes. In contrast, geometric methods only simulate results at pre-defined receiver locations but use less memory than wave-based methods. One convenient property for acoustic measurements and simulations is that the location of the source and the receiver can be interchanged without affecting the measured/simulated result according to acoustic reciprocity (Wapenaar, 2019). This is sometimes useful in reducing the number of measurements/simulations in various scenes.

2.4 Audio Processing Applications

The measurement/computation of RIRs and resulting datasets has been used for audio processing applications including but not limited to:

1. Sound Propagation and Rendering: Sounds in nature are produced by vibrating objects and then propagate through a medium (e.g., air) before finally being heard by a listener. Humans can perceive these sound waves in the frequency range of 20Hz to 20KHz (human aural range). There is a large body of literature on modeling sound propagation in indoor scenes using geometric and wave-based methods (Liu and Manocha, 2020; Krokstad et al., 1968; Vorländer, 1989; Funkhouser et al., 1998a; Raghuvanshi et al., 2009; Mehra et al., 2013; Schissler et al., 2014). Wave-based solvers are practical for lower frequencies and limited to static scenes. Geometric methods, widely used in interactive applications, are accurate for higher frequencies. We need automatic software systems that can accurately compute IRs corresponding to human aural range and handle arbitrary 3D models.

2. Deep Audio Synthesis for Videos: Video acquisition has become very common and easy. However, it is difficult to add realistic audio that can be synchronized with

animation in a video. Many deep learning methods have been proposed for such audio synthesis that utilize acoustic impulse responses for such applications (Li et al., 2018; Owens et al., 2016; Zhou et al., 2018)

3. Speech Processing using Deep Learning: IRs consist of many clues related to reproducing or understanding intelligible human speech. Synthetic datasets of IRs have been used in machine learning methods for automatic speech recognition (Malik et al., 2021; Ko et al., 2017; Tang et al., 2020; Ratnarajah et al., 2021), sound source separation (Aralikatti et al., 2021; Jenrungrot et al., 2020), and sound source localization (Grumiaux et al., 2021).

4. Sound Simulation using Machine Learning: Many recent deep learning methods have been proposed for sound synthesis (Hawley et al., 2020; Ji et al., 2020; Jin et al., 2020), scattering effect computation, and sound propagation (Fan et al., 2020b; Meng et al., 2021; Pulkki and Svensson, 2019). Deep learning methods have also been used to compute material properties of a room and acoustic characteristics (Schissler et al., 2017; Tang et al., 2019a).

Some of these will be discussed in more detail in this dissertation. Other applications that have used acoustic datasets include navigation (Chen et al., 2020), floorplan reconstruction (Purushwalkam et al., 2021) and depth estimation algorithms (Gao et al., 2020).

Chapter 3

Scene-Aware Audio for Mixed Reality[1]

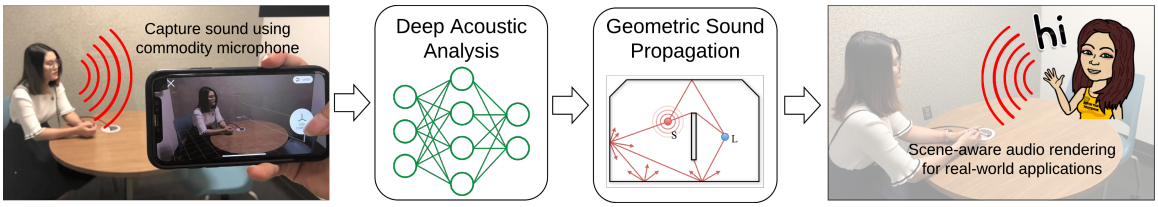


Figure 3.1: Given a natural sound in a real-world room that is recorded using a cellphone microphone (left), we estimate the acoustic material properties and the frequency equalization of the room using a novel deep learning approach (middle). We use the estimated acoustic material properties for generating plausible sound effects in the virtual model of the room (right). Our approach is general and robust, and works well with commodity devices.

3.1 Introduction

Auditory perception of recorded sound is strongly affected by the acoustic environment it is captured in. Concert halls are carefully designed to enhance the sound on stage, even accounting for the effects an audience of human bodies will have on the propagation of sound (Barron, 2010). Anechoic chambers are designed to remove

[1]The work in this chapter has been published in Tang et al. (2019a)

acoustic reflections and propagation effects as much as possible. Home theaters are designed with acoustic absorption and diffusion panels, as well as with careful speaker and seating arrangements (Rizzi et al., 2016).

The same acoustic effects are important when creating immersive effects for virtual reality (VR) and augmented reality (AR) applications. It is well known that realistic sounds can improve a user's sense of presence and immersion (Larsson et al., 2002). There is considerable work on interactive sound propagation in virtual environments based on geometric and wave-based methods (Vorländer, 1989; Schissler and Manocha, 2017; Raghuvanshi and Snyder, 2014c; Cao et al., 2017). Furthermore, these techniques are increasingly used to generate plausible sound effects in VR systems and games, including Microsoft Project Acoustics[2], Oculus Spatializer[3], Steam Audio[4], etc. However, these methods are limited to synthetic scenes where an exact geometric representation of the scene and acoustic material properties are known apriori.

In this chapter, we address the problem of rendering realistic sounds that are similar to recordings of real acoustic scenes. These capabilities are needed for VR as well as AR applications (Conference, 2018), which often use recorded sounds. Foley artists often record source audio in environments similar to the places the visual contents were recorded in. Similarly, creators of vocal content (e.g. podcasts, movie dialogue, or video voice-overs), carefully re-record content made in different environment or with different equipment to match the acoustic conditions. However, these processes are expensive, time-consuming, and cannot adapt to spatial listening location. There is strong interest in developing automatic spatial audio synthesis methods.

For VR or AR content creation, acoustic effects can also be captured with an

[2]https://aka.ms/acoustics
[3]https://developer.oculus.com/downloads/package/oculus-spatializer-unity
[4]https://valvesoftware.github.io/steam-audio

impulse response (IR) – a compact acoustic description of how sound propagates from one location to another in a given scene. Given an IR, it can be convolved with any virtual sound or dry sound to generate the desired acoustic effects. However, recording the IRs of real-world scenes can be challenging, especially for interactive applications. Many times special recording hardware is needed to record the IRs. Furthermore, the IR is a function of the source and listener positions and it needs to be re-recorded as either position changes.

Our goal is to replace the step of recording an IR with an unobtrusive method that works on in-situ speech recordings and video signals and uses commodity devices. This can be regarded as an acoustic analogy of visual relighting (Debevec, 2002): to light a new visual object in an image, traditional image based lighting methods require the capture of real-world illumination as an omnidirectional, high dynamic range (HDR) image. This light can be applied to the scene, as well as on a newly inserted object, making the object appear as if it was always in the scene. Recently, Gardner et al. (2017) and Hold-Geoffroy et al. (2017) proposed convolutional neural network (CNN)-based methods to estimate HDR indoor or outdoor illumination from a single low dynamic range (LDR) image. These high-quality visual illumination estimation methods enable novel interactive applications. Concurrent work from LeGendre et al. (2019) demonstrates the effectiveness on mobile devices, enabling photorealistic mobile mixed reality experiences.

In terms of audio "relighting" or reproduction, there have been several approaches proposed toward realistic audio in 360° images (Kim et al., 2019), multi-modal estimation and optimization (Schissler et al., 2017), and scene-aware audio in 360° videos (Li et al., 2018). However, these approaches either require separate recording of an IR, or produce audio results that are perceptually different from recorded scene audio. Important acoustic properties can be extracted from IRs, including the reverberation time (T_{60}), which is defined as the time it takes for a sound to decay 60 decibels (Kut-

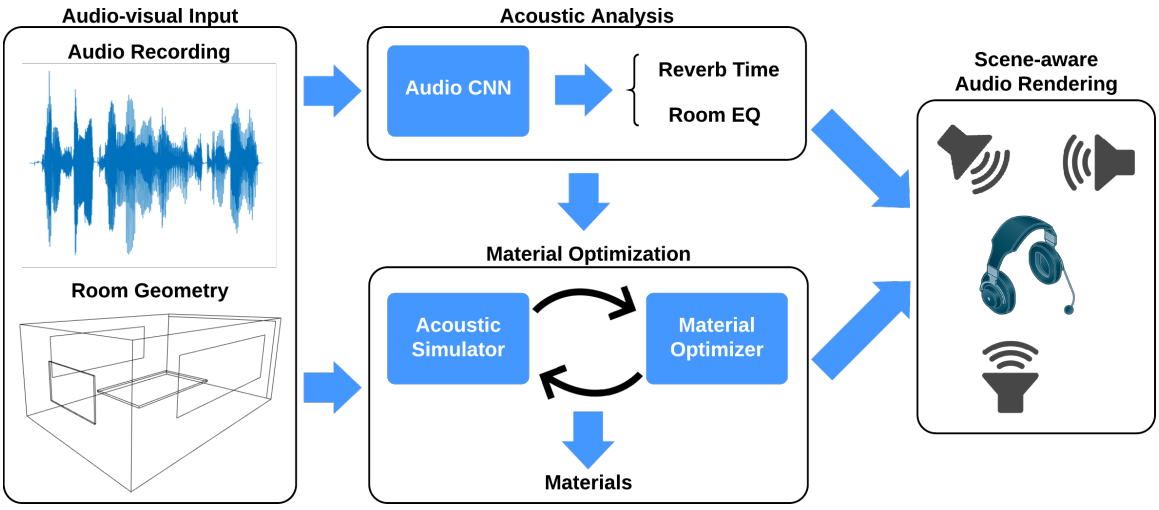


Figure 3.2: **Our pipeline:** Starting with a audio-video recording (left), we estimate the 3D geometric representation of the environment using standard computer vision methods. We use the reconstructed 3D model to simulate new audio effects in that scene. To ensure our simulation results perceptually match recorded audio in the scene, we automatically estimate two acoustic properties from the audio recordings: frequency-dependent reverberation time or T_{60} of the environment, and a frequency-dependent equalization curve. The T_{60} is used to optimize the frequency-dependent absorption coefficients of the materials in the scene. The frequency equalization filter is applied to the simulated audio, and accounts for the missing wave effects in geometrical acoustics simulation. We use these parameters for interactive scene-aware audio rendering (right).

truff, 2016), and the frequency-dependent amplitude level or equalization (EQ) (Hak et al., 2012). This heavy reliance on IRs greatly constrains the wide adoption of audio for immersive applications or video post-production that require realistic acoustic simulation that is calibrated to real-world acoustic scenes.

Main Results: We present novel algorithms to estimate two important environmental acoustic properties from recorded sounds (e.g. speech). Our approach uses commodity microphones and does not need to capture any IRs. The first property is the frequency-dependent T_{60}. This is used to optimize absorption coefficients for geometric acoustic (GA) simulators for audio rendering. Next, we estimate a frequency equalization filter to account for wave effects that cannot be modeled accurately using geometric acoustic simulation algorithms. This equalization step is crucial to ensur-

ing that our GA simulator outputs perceptually match existing recorded audio in the scene. Estimating the equalization filter *without an IR* is challenging since it is not only speaker dependent, but also scene dependent, which poses extra difficulties in terms of dataset collection. For a model to predict the equalization filtering behavior accurately, we need a large amount of diverse speech data and IRs. Our key idea is a novel dataset augmentation process that significantly increases room equalization variation. With robust room acoustic estimation as input, we present a novel inverse material optimization algorithm to estimate the acoustic properties. We propose a new objective function for material optimization and show that it models the IR decay behavior better than (Li et al., 2018). We demonstrate our ability to add new sound sources in regular videos. Similar to visual relighting examples where new objects can be rendered with photorealistic lighting, we enable audio reproduction in any regular video with existing sound with applications for mixed reality experiences. We highlight their performance on many challenging benchmarks.

We show the importance of matched T_{60} and equalization in our perceptual user study §3.5. In particular, our perceptual evaluation results show that: (1) Our T_{60} estimation method is perceptually comparable to all past baseline approaches, even though we do not require an explicit measured IR; (2) Our EQ estimation method improves the performance of our T_{60}-only approach by a statistically significant amount ($\approx$ 10 rating points on a 100 point scale); and (3) Our combined method (T_{60}+EQ) outperforms the average room IR (T_{60} = .5 seconds with uniform EQ) by a statistically significant amount (+10 rating points) – the only reasonable comparable baseline we could conceive that does not require an explicit IR estimate. To the best of our knowledge, ours is the first method to predict IR equalization from raw speech data and validate its accuracy. Our main contributions include:

- A CNN-based model to estimate frequency-dependent T_{60} *and* equalization filter from real-world speech recordings.

- An equalization augmentation scheme for training to improve the prediction robustness.
- A derivation for a new optimization objective that better models the IR decay process for inverse materials optimization.
- A user study to compare and validate our performance with current state-of-the-art audio rendering algorithms. Our study is used to evaluate the perceptual similarity between the recorded sounds and our rendered audio.

3.2 Related Work

Cohesive audio in mixed reality environments (when there is a mix of real and virtual content), is more difficult than in fully virtual environments. This stems from the difference between "Plausibility" in VR and "Authenticity" in AR (Kim et al., 2019). Visual cues dominate acoustic cues, so the perceptual difference between how audio sounds and the environment in which it is seen is smaller than the perceived environment of two sounds. Recently, Li et al. (2018) introduced scene-aware audio to optimize simulator parameters to match the room acoustics from existing recordings. By leveraging visual information for acoustic material classification, Schissler et al. (2017) demonstrated realistic audio for 3D-reconstructed real-world scenes. However, both of these methods still require explicit measurement of IRs. In contrast, our proposed pipeline works with any input speech signal and commodity microphones.

Sound simulation can be categorized into wave-based methods and geometric acoustics. While wave-based methods generally produce more accurate results, it remains an open challenge to build a real-time universal wave solver. Recent advances such as parallelization via rectangular decomposition (Morales et al., 2015), pre-computation acceleration structures (Mehra et al., 2015), and coupling with geometric acoustics (Yeh et al., 2013; Rungta et al., 2018) are used for interactive

applications. It is also possible to precompute low-frequency wave-based propagation effects in large scenes (Raghuvanshi et al., 2010), and to perceptually compress them to reduce runtime requirements (Raghuvanshi and Snyder, 2014a). Even with the massive speedups presented, and a real-time runtime engine, these methods still require tens of minutes to hours of pre-computation depending on the size of the scene and frequency range chosen, making them impractical for augmented reality scenarios and difficult to include in an optimization loop to estimate material parameters. With interactive applications as our goal, most game engines and VR systems tend to use geometric acoustic simulation methods (Vorländer, 1989; Schissler and Manocha, 2017; Cao et al., 2017). These algorithms are based on fast ray tracing and perform specular and diffuse reflections (Savioja and Svensson, 2015). Some techniques have been proposed to approximate low-frequency diffraction effects using ray-tracing (Tsingos et al., 2001; Rungta et al., 2018; Taylor et al., 2012a). Our approach can be combined with any interactive audio simulation method, though our current implementation is based on bidirectional ray tracing (Cao et al., 2017). The sound propagation algorithms can also be used for acoustic material design optimization for synthetic scenes (Morales and Manocha, 2016).

The efficiency of deep neural networks has been shown in audio/video-related tasks that are challenging for traditional methods(Virtanen et al., 2018; Gharib et al., 2018; Hinton et al., 2012; Evers et al., 2016; Sterling et al., 2018). Hershey et al. (2017) showed that it is feasible to use CNNs for large-scale audio classification problems. Many deep neural networks require a large amount of training data. Salamon and Bello (2017) used data augmentation to improve environmental sound classification. Similarly, Bryan (2020) estimates the T_{60} and the direct-to-reverberant ratio (DRR) from a single speech recording via augmented datasets. Tang et al. (2019b) trained CRNN models purely based on synthetic spatial IRs that generalize to real-world recordings. We strategically design an augmentation scheme to address the challenge

of equalization's dependence on both IRs and speaker voice profiles, which is fully complimentary to all prior data-driven methods.

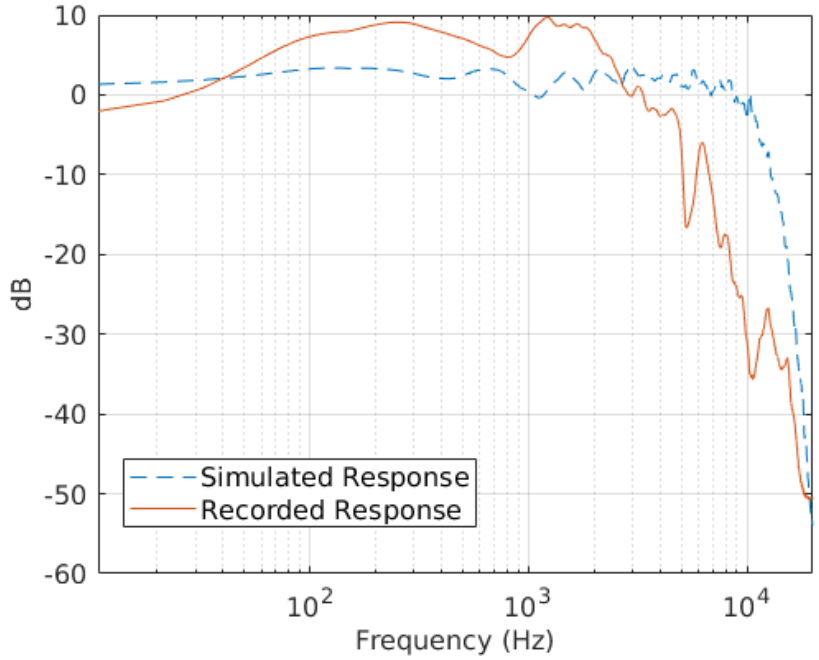


Figure 3.3: The simulated and recorded frequency response in the same room at a sample rate of 44.1kHz is shown. Note that the recorded response has noticeable peaks and notches compared with the relatively flat simulated response. This is mainly caused by room equalization. Missing proper room equalization leads to discrepancies in audio quality and overall room acoustics.

Acoustic simulators require a set of well-defined material properties. The material absorption coefficient is one of the most important parameters (Bork, 2000), ranging from 0 (total reflection) to 1 (total absorption). A material's acoustic properties are correlated with its visual appearance to some extent. For example, a carpet is usually more absorptive for sound than a glass is. This audio-visual correlation enables rough material estimation from visual cues (Schissler et al., 2017). However, despite a non-zero material recognition error, visual information alone does not accurately capture the acoustic property of materials. Prior work shows that 7D (source-listener 3D locations and time) acoustic fields in an environment can be effectively compressed into 6D time-invariant fields using only four selected scalar acoustic metrics with low

reconstruction errors (Raghuvanshi and Snyder, 2014c). This indicates that certain acoustic metrics can be used to guide the modeling of acoustic materials.

When a reference IR is available, it is straightforward to adjust room materials to match the energy decay of the simulated IR to the reference IR (Li et al., 2018). Similarly, Ren et al. (2013) optimized linear modal analysis parameters to match the given recordings. A probabilistic damping model for audio-material reconstruction has been presented for VR applications (Sterling et al., 2019). Unlike all previous methods which require a clean IR recording for accurate estimation and optimization of boundary materials, we infer typical material parameters including T_{60} values and equalization from raw speech signals using a CNN-based model.

Analytical gradients can significantly accelerate the optimization process. With similar optimization objectives, it was shown that additional gradient information can boost the speed by a factor of over ten times (Li et al., 2018; Schissler et al., 2017). The speed gain shown in (Li et al., 2018) is impressive, and we further improve the accuracy and speed of the formulation. More specifically, the original objective function evaluated energy decay relative to the first ray received (the direct sound if there were no obstacles). However, energy estimates can be noisy due to both the oscillatory nature of audio as well as simulator noise. Instead, we optimize the slope of the best fit line of ray energies to the desired energy decay (defined by the T_{60}), which we found to be more robust.

3.3 Deep Acoustic Analysis: Our Algorithm

In this section, we overview our proposed method for scene-aware audio rendering. We begin by providing background information, discuss how we capture room geometry, and then proceed with discussing how we estimate the frequency dependent room reverberation and equalization parameters directly from recorded speech. We follow

by discussing how we use the estimated acoustic parameters to perform acoustic materials optimization such that we calibrate our virtual acoustic model with real-world recordings.

3.3.1 Background

To explain the motivation of our approach, we briefly elaborate on the most difficult parts of previous approaches, upon which our method improves. Previous methods require an impulse response of the environment to estimate acoustic properties (Li et al., 2018; Schissler et al., 2017). Recording an impulse response is a non-trivial task. The most reliable methods involve playing and recording Golay codes (Foster, 1986) or sine sweeps (Farina, 2000), which both play loud and intrusive audio signals. Also required are a fairly high-quality speaker and microphone with constant frequency response, small harmonic distortion and little crosstalk. The speaker and microphone should be acoustically separated from surfaces, i.e., they shouldn't be placed directly on tables (else surface vibrations could contaminate the signal). Clock drift between the source and microphone must be accounted for (Bryan et al., 2010). Alternatively, balloon pops or hand claps have been proposed for easier IR estimation, but require significant post-processing and still are very obtrusive (Abel et al., 2010; Seetharaman and Tarzia, 2012). In short, correctly recording an IR is not easy, and makes it challenging to add audio in scenarios such as augmented reality, where the environment is not known beforehand and estimation must be done interactively to preserve immersion.

Geometric acoustics is a high-frequency approximation to the wave equation. It is a fast method, but assumes that wavelengths are small compared to objects in the scene, while ignoring pressure effects (Savioja and Svensson, 2015). It misses several important wave effects such as diffraction and room resonance. Diffraction occurs when sound paths bend around objects that are of similar size to the wavelength.

Resonance is a pressure effect that happens when certain wavelengths are either reinforced or diminished by the room geometry: certain wavelengths create peaks or troughs in the frequency spectrum based on the positive or negative interference they create.

We model these effects with a linear finite impulse response (FIR) equalization filter (Schafer and Oppenheim, 1989). We compute the discrete Fourier transform on the recorded IR over all frequencies, following Li et al. (2018). Instead of filtering directly in the frequency domain, we design a linear phase EQ filter with 32ms delay to compactly represent this filter at 7 octave bin locations. We then blindly estimate this compact representation of the frequency spectrum of the impulse response as discrete frequency gains, without specific knowledge of the input sound or room geometry. This is a challenging estimation task. Since the convolution of two signals (the IR and the input sound) is equivalent to multiplication in the frequency domain, estimating the frequency response of the IR is equivalent to estimating one multiplicative factor of a number without constraining the other. We are relying on this approach to recognize the a compact representation of the frequency response magnitude in different environments.

3.3.2 Geometry Reconstruction

Given the background, we begin by first estimating the room geometry. In our experiments, we utilize the ARKit-based iOS app MagicPlan[5] to acquire the basic room geometry. A sample reconstruction is shown in Figure 3.4. With computer vision research evolving rapidly, we believe constructing geometry proxies from video input will become even more robust and easily accessible (Zhi et al., 2019; Bloesch et al., 2018).

[5] https://www.magicplan.app/

Figure 3.4: We use an off-the-shelf app called MagicPlan to generate geometry proxy. Input: a real-world room (left); Output: the captured 3D model of the room (right) without high-level details, which is used by the runtime geometric acoustic simulator.

3.3.3 Learning Reverberation and Equalization

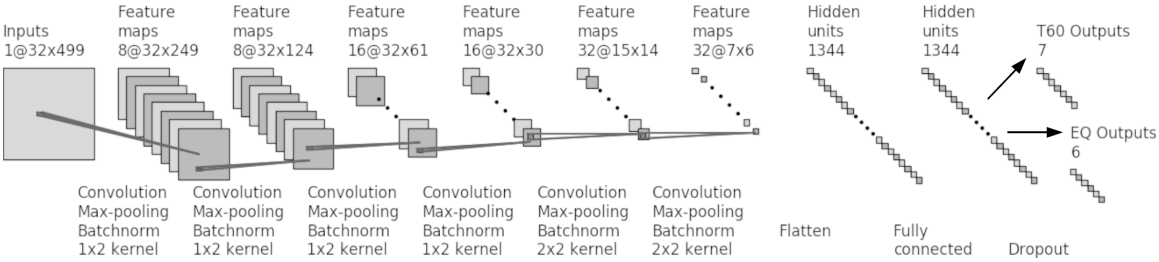


Figure 3.5: Network architecture for T_{60} and EQ prediction. Two models are trained for T_{60} and EQ, which have the same components except the output layers have different dimensions customized for the octave bands they use.

We use a convolutional neural network (Figure 3.5) to predict room equalization and reverberation time (T_{60}) directly from a speech recording. Training requires a large number of speech recordings with known T_{60} and room equalization. The standard practice is to generate speech recordings from known real-world or synthetic IRs (Kim et al., 2017; Doulaty et al., 2017). Unfortunately, large scale IR datasets do not currently exist due to the difficulty of IR measurement; most publicly available IR datasets have fewer than 1000 IR recordings. Synthetic IRs are easy to obtain and can be used, but again lack wave-based effects as well as other simulation deficiencies. Recent work has addressed this issue by combining real-word IR measurements with augmentation to increase the diversity of existing real-world datasets (Bryan, 2020).

This work, however, only addresses T_{60} and DRR augmentation, and lacks a method to augment the frequency-equalization of existing IRs. To address this, we propose an augmentation method in this section. Beforehand, however, we discuss our neural network estimation method for estimating both T_{60} and equalization.

Octave-Based Prediction

Most prior work takes the full-frequency range as input for prediction. For example, one closely related work (Bryan, 2020) only predicts one T_{60} value for the entire frequency range (full-band). However, sound propagates and interacts with materials differently at different frequencies. To this end, we define our learning targets over several octaves. Specifically, we calculate T_{60} at 7 sub-bands centered at $\{125, 250, 500, 1000, 2000, 4000, 8000\}$Hz. We found prediction of T_{60} at the 62.5Hz band to be unreliable due to low SNR. During material optimization, we set the 62.5Hz T_{60} value to the 125Hz one. Our frequency equalization estimation is done at 6 octave bands centered at $\{62.5, 125, 250, 500, 2000, 4000\}$Hz. Note that we will compute equalization relative to the 1kHz band, so we do not estimate it. When applying our equalization filter, we set bands greater than or equal to 8kHz to -50dB. Given our target sampling rate of 16kHz and the limited content of speech in higher octaves, this did not affect our estimation.

Data Augmentation

We use the following datasets as the basis for our training and augmentation.

- ACE Challenge (Eaton et al., 2016): 70 IRs and noise audio;
- MIT IR Survey (Traer and McDermott, 2016): 271 IRs;
- DAPS dataset (Mysore, 2014): 4.5 hours of 20 speakers' speech (10 males and 10 females).

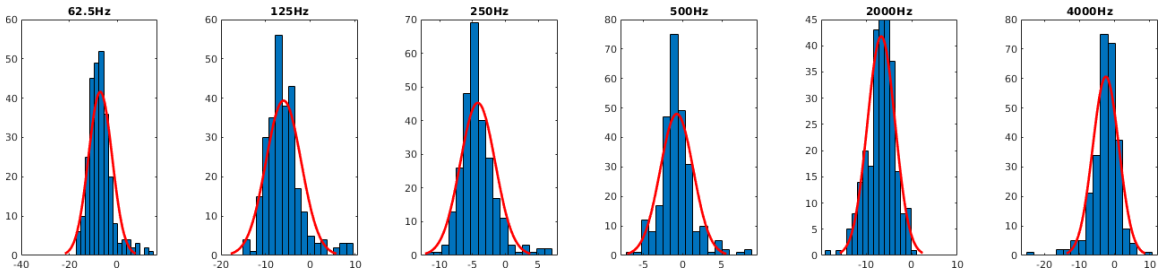
Table 3.1: Dataset composition. The training set and validation set are based on synthetic IRs and the test set is based on real IRs to guarantee model generalization. Clean speech files are also divided in a way that speakers ("f1" for female speaker 1; "m10" for male speaker 10) in each dataset partition are different, to avoid the model learning the speaker's voice signature. Audio files are generated at a sample rate of 16kHz, which is sufficient to cover the human voice's frequency range.

Partition	Noise	Clean Speech	IR
Training set (size: 56.5k)	ACE ambient	f5∼f10, m5∼m10	Synthetic IR (size: 4.5k)
Validation set (size: 19.5k)	ACE ambient	f3, f4, m3, m4	Synthetic IR (size: 1k)
Test set (size: 18.5k)	ACE ambient	f1, f2, m1, m2	MIT survey IR (size: 271)

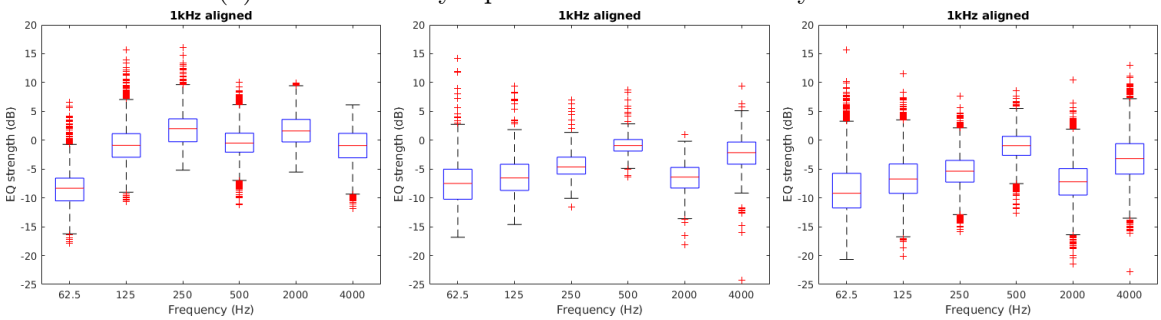
First, we use the method in Bryan (2020) to expand the T_{60} and direct-to-reverberant ratio (DRR) range of the 70 ACE IRs, resulting in 7000 synthetic IRs with a balanced T_{60} distribution between $0.1 \sim 1.5$ seconds. The ground truth T_{60} estimates can be computed directly from IRs can be computed is a variety of ways. We follow the methodology of Karjalainen et al. (2001) when computing the T_{60} from real IRs with a measurable noise floor. This method was found to be the most robust estimator when computing the T_{60} from real IRs in recent work (Eaton et al., 2016). The final composition of our dataset is listed in Table 3.1.

While we know the common range of real-world T_{60} values, there is limited literature giving statistics about room equalization. Therefore, we analyzed the equalization range and distribution of the 271 MIT survey IRs as a guidance for data augmentation. The equalization of frequency bands is computed relative to the 1kHz octave. This is a common practice (Välimäki and Reiss, 2016), unless expensive equipment is used to obtain calibrated acoustic pressure readings.

For our equalization augmentation procedure, we first fit a normal distribution (mean and standard deviation) to each sub-band amplitude of the MIR IR dataset as shown in Figure 3.6. Given this set of parametric model estimates, we iterate through our training and validation IRs. For each IR, we extract its original EQ.



(a) MIT IR survey equalization distribution by sub-band.



(b) Original synthetic IR equalization.

(c) Target (MIT) IR equalization.

(d) Augmented synthetic IR equalization.

Figure 3.6: Equalization augmentation. The 1000Hz sub-band is used as reference and has unit gain. We fit normal distributions (red bell curves shown in (a)) to describe the EQ gains of MIT IRs. We then apply EQs sampled from these distributions to our training set distribution in (b). We observe that the augmented EQ distribution in (d) becomes more similar to the target distribution in (c).

We then randomly sample a target EQ according to our fit models (independently per frequency band), calculate the distance between the source and target EQ, and then design an FIR filter to compensate for the difference. For simplicity, we use the window method for FIR filter design (Smith III, 2008). Note, we do not require a perfect filter design method. We simply need a procedure to increase the diversity of our data. Also note, we intentionally sample our augmented IRs to have a larger variance than the recorded IRs to further increase the variety of our training data.

We compute the log Mel-frequency spectrogram for each four second audio clip, which is commonly used for speech-related tasks (Chen et al., 2018; Eskimez et al., 2018). We use a Hann window of size 256 with 50% overlap during computation of the short-time Fourier transform (STFT) for our 16kHz samples. Then we use 32 Mel-scale bands and area normalization for Mel-frequency warping (Stevens et al., 1937).

The spectrogram power is computed in decibels. This extraction process yields a 32 x 499 (frequency x time domain) matrix feature representation. All feature matrices are normalized by the mean and standard deviation of the training set.

Network Architecture and Training

We propose using a network architecture differing only in the final layer for both T_{60} and room equalization estimation. Six 2D convolutional layers are used sequentially to reduce both the time and frequency resolution of features until they have approximately the same dimension. Each `conv` layer is immediately followed by a rectified linear unit (`ReLU`) (Nair and Hinton, 2010) activation function, 2D max pooling, and batch normalization. The output from `conv` layers is flattened to a 1D vector and connected to a fully connected layer of 64 units, at a dropout rate of 50% to lower the risk of overfitting. The final output layer has 7 fully connected units to predict a vector of length 7 for T_{60} or 6 fully connected units to predict a vector of length 6 for frequency equalization. This network architecture is inspired by Bryan (2020), where it was used to predict full-band T_{60}. We updated the output layer to predict the more challenging sub-band T_{60}, and also discovered that the same architecture predicts equalization well.

For training the network, we use the mean square error (MSE) with the ADAM optimizer (Kingma and Ba, 2014) in Keras (Chollet et al., 2015). The max number of epochs is 500 with an early stopping mechanism. We choose the model with the lowest validation error for further evaluation on the test set. Our model architecture is shown in Figure 3.5.

3.3.4 Acoustic Material Optimization

Our goal is to optimize the material absorption coefficients at the same octave bands as T_{60} of a set of room materials to match the sub-band T_{60} of the simulated sound

with the target predicted in § 3.3.3.

Ray Energy. We borrow notation from Li et al. (2018). Briefly, a geometric acoustic simulator generates a set of sound paths, each of which carries an amount of sound energy. Each material m_i in a scene is described by a frequency dependent absorption coefficient, ρ_i. A path leaving the source is reflected by a set of materials before it reaches the listener. The energy fraction that is received by the listener along path j is

$$e_j = \beta_j \prod_{k=1}^{N_j} \rho_{m^k}, \tag{3.1}$$

where m^k is the material the path intersects on the k^{th} bounce, N_j is the number of surface reflections for path j, and β_j accounts for air absorption (dependent on the total length of the path). Our goal is to optimize the set of absorption coefficients ρ_i to match the energy distribution of the paths e_j to that of the environment's IR. Again similar to (Li et al., 2018), we assume the energy decrease of the IR follows an exponential curve, which is a linear decay in dB space. The slope of this decay line is $m' = -60/T_{60}$.

Objective Function. We propose the following objective function:

$$J(\rho) = (m - m')^2 \tag{3.2}$$

where m is the best fit line of the ray energies on a decibel scale:

$$m = \frac{n \sum_{i=0}^{n} t_i y_i - \sum_{i=0}^{n} t_i \sum_{i=0}^{n} y_i}{n \sum_{i=0}^{n} t_i^2 - \left(\sum_{i=0}^{n} t_i\right)^2}, \tag{3.3}$$

with $y_i = 10\log_{10}(e_i)$, which we found to be more robust than previous methods. Specifically, in comparison with Equation (3) in Li et al. (2018), we see that they try to match the slope of the energies relative to e_0, forcing e_0 to be at the origin on a

dB scale. However, we only care about the energy decrease, and not the absolute scale of the values from the simulator. We found that allowing the absolute scale to move and only optimizing the slope of the best fit line produced a better match to the target T_{60}.

We minimize J using the L-BFGS-B algorithm (Zhu et al., 1997). The gradient of J is given by

$$\frac{\partial J}{\partial \rho_j} = 2(m - m') \frac{n t_i - \sum_{i=0}^{n} t_i}{n \sum_{i=0}^{n} t_i^2 - \left(\sum_{i=0}^{n} t_i\right)^2} \frac{10}{\ln(10) e_i} \frac{\partial e_i}{\partial \rho_j} \tag{3.4}$$

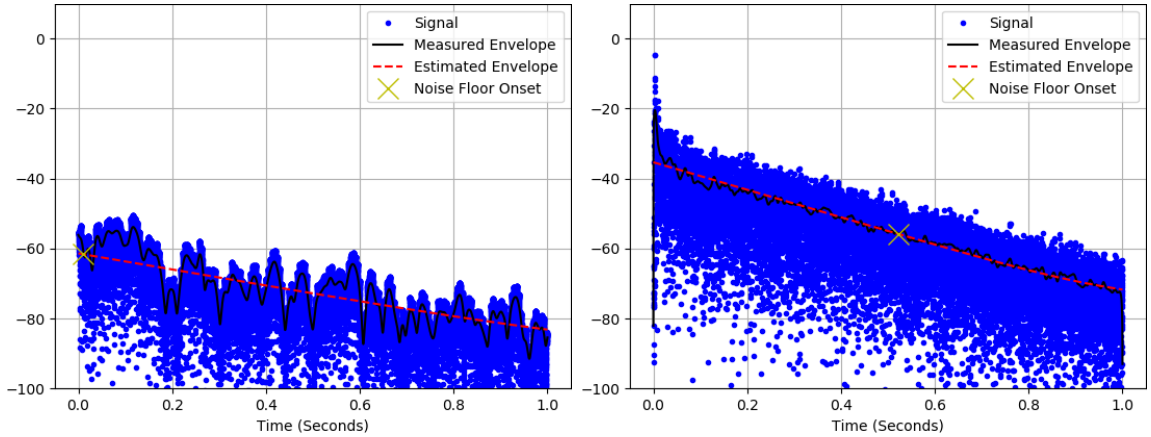
3.4 Analysis and Applications

3.4.1 Analysis

Speed. We implement our system on an Intel Xeon(R) CPU @3.60GHz and an NVIDIA GTX 1080 Ti GPU. Our neural network inference runs at 222 fps on 4-second sliding windows of audio due to the compact design (only 18K trainable parameters). Optimization runs twice as fast with our improved objective function. The sound rendering is based on the real-time geometric bi-directional sound path tracing from Cao et al. (2017).

Sub-band T_{60} prediction. We first evaluate our T_{60} blind estimation model and achieve a mean absolute error (MAE) of 0.23s on the test set (MIT IRs). While the 271 IRs in the test set have a mean T_{60} of 0.49s with a standard deviation (STD) of 0.85s at the 125Hz sub-band, the highest sub-band 8000Hz only has a mean T_{60} of 0.33s with a STD of 0.24s, which reflects a narrow subset within our T_{60} augmentation range. We also notice that the validation MAE on ACE IRs is 0.12s, which indicates our validation set and the test set still come from different distributions. Another error source is the inaccurate labeling of low-frequency sub-band T_{60} as shown in

Figure 3.7, but we do not filter any outliers in the test set. In addition, our data is intended to cover frequency ranges up to 8000Hz, but human speech has less energy in high-frequency range (Titze et al., 2017), which results in low signal energy for these sub-bands, making it more difficult for learning.



(a) 125Hz sub-band. (b) 8000Hz sub-band.

Figure 3.7: Evaluating T_{60} from signal envelope on low and high frequency bands of the same IR. Note that the SNR in the low frequency band is lower than the high frequency band. This makes T_{60} evaluation for low frequency bands less reliable, which partly explains the larger test error in low frequency sub-bands.

Material Optimization. When we optimize the room material absorption coefficients according to the predicted T_{60} of a room, our optimizer efficiently modifies the simulated energy curve to a desired energy decay rate (T_{60}) as shown in Figure 3.8. We also try fixing the room configuration and set the target T_{60} to values uniformly distributed between 0.2s and 2.5s, and evaluate the T_{60} of the simulated IRs. The relationship between the target and output T_{60} is shown in Figure 3.9, in which our simulation closely matches the target, demonstrating that our optimization is able to match a wide range of T_{60} values.

To test the real-world performance of our acoustic matching, we recorded ground truth IRs in 5 benchmark scenes, then use the method in Li et al. (2018), which requires a reference IR, and our method, which does not require an IR, for comparison.

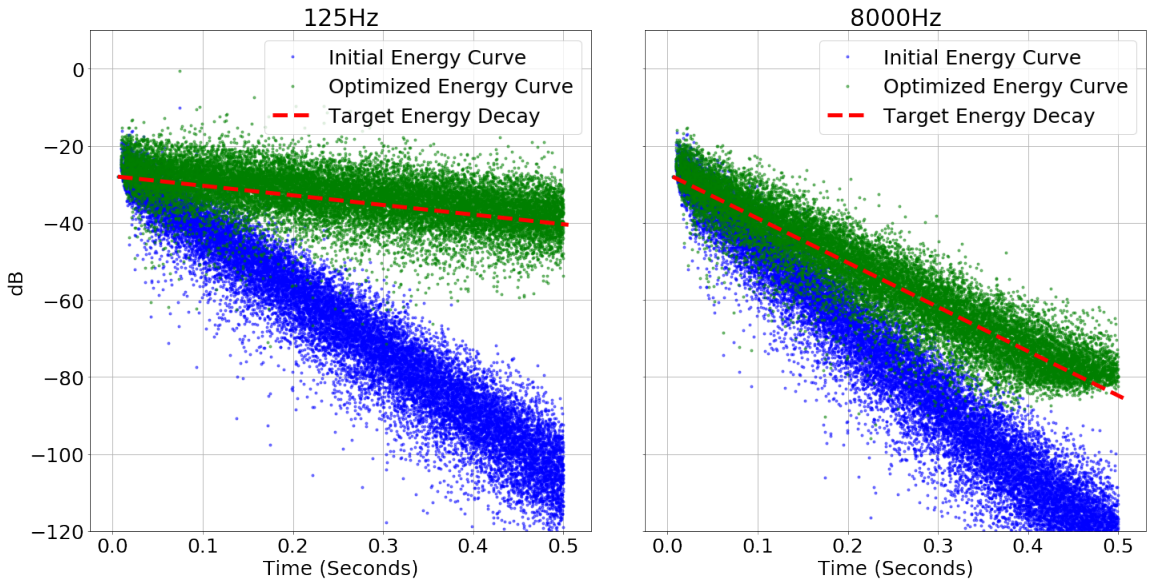


Figure 3.8: Simulated energy curves before and after optimization (with target slope shown).

Benchmark scenes and results are summarized in Table 3.2.

We apply the EQ filter to the simulated IR as a last step. Overall, we obtain a prediction MAE of 3.42dB on our test set, whereas before augmentation, the MAE was 4.72dB under the same training condition, which confirms the effectiveness of our EQ augmentation. The perceptual impact of the EQ filter step is evaluated in §3.5.

3.4.2 Comparisons

We compare our work with two related projects, Schissler et al. (2017) and Kim et al. (2019), where the high-level goal is similar to ours but the specific approach is different.

Material optimization is a key step in our method and Schissler et al. (2017). One major difference is that we additionally compensate wave effects explicitly with an equalization filter. Figure 3.10 shows the difference in spectrogram where the high frequency equalization was not properly accounted for. Our method better replicate the rapid decay in the high frequency range. For audio comparison, please refer to

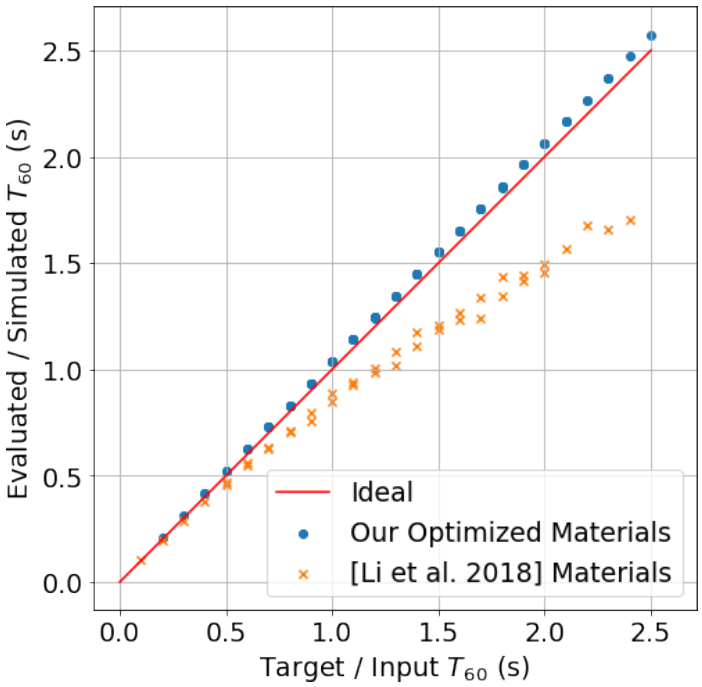


Figure 3.9: Stress test our our optimizer. We uniformly sample T_{60} between 0.2s and 2.5s and set it to be the target. The ideal I/O relationship is a straight line passing the origin with slope 1. Our optimization results matches the ideal line much better than prior optimization method.

our supplemental video[6].

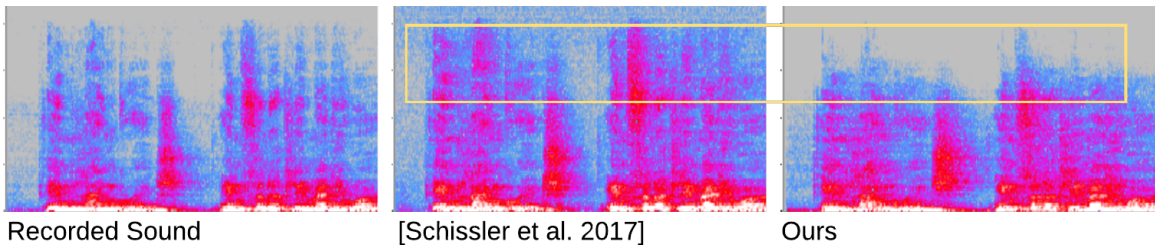
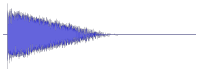
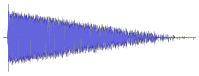
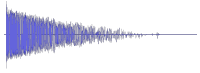
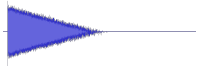
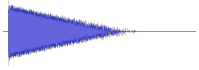
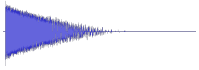
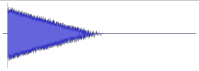
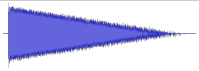
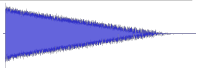


Figure 3.10: We show the effects of our equalization filtering on audio spectrograms, compared with Schissler et al. (2017). In the highlighted region, we are able to better reproduce the fast decay in the high-frequency range, closely matching the recorded sound.

We also want to highlight the importance of optimizing T_{60}. In (Kim et al., 2019), a CNN is used for object-based material classification. Default materials are assigned to a limited set of objects. Without optimizing specifically for the audio objective, the resulting sound might not blend in seamlessly with the existing audio. In Figure 3.11,

[6] https://gamma.umd.edu/pro/sound/sceneaware

Table 3.2: Benchmark results for acoustic matching. These real-world rooms are of different sizes and shapes, and contain a wide variety of acoustic materials such as brick, carpet, glass, metal, wood, plastic, etc., which make the problem acoustically challenging. We compare our method with Li et al. (2018). Our method does not require a reference IR and still obtains similar T_{60} and EQ errors in most scenes compared with their method. We also achieve faster optimization speed. Note that the input audio to our method is already noisy and reverberant, whereas Li et al. (2018) requires clean IR recording. All IR plots in the table have the same time and amplitude scale.

Benchmark Scene				
Size (m^3)	1100 (irregular)	1428 (12x17x7)	72 (4x6x3)	352 (11x8x4)
# Main planes	6	6	11	6
Groundtruth IR (dB scale)				
Li et al. (2018) IR (dB scale)				
Opt. time (s)	29	43	71	46
T_{60} error (s)	0.11	0.23	0.02	0.10
EQ error (dB)	1.50	2.97	3.61	7.55
Ours IR (dB scale)				
Opt. time (s)	13	13	31	20
T_{60} error (s)	0.14	0.12	0.04	0.24
EQ error (dB)	2.26	3.86	3.46	4.62

we show that our method produces audio that matches the decay tail better, whereas (Kim et al., 2019) produces a longer reverb tail than the recorded ground truth.

3.4.3 Applications

Acoustic Matching in Videos Given a recorded video in an acoustic environment, our method can analyze the room acoustic properties from noisy, reverberant recorded audio in the video. The room geometry can be estimated from video (Bloesch et al., 2018), if the user has no access to the room for measurement. During post-processing, we can simulate sound that is similar to the recorded sound in the room. Moreover,

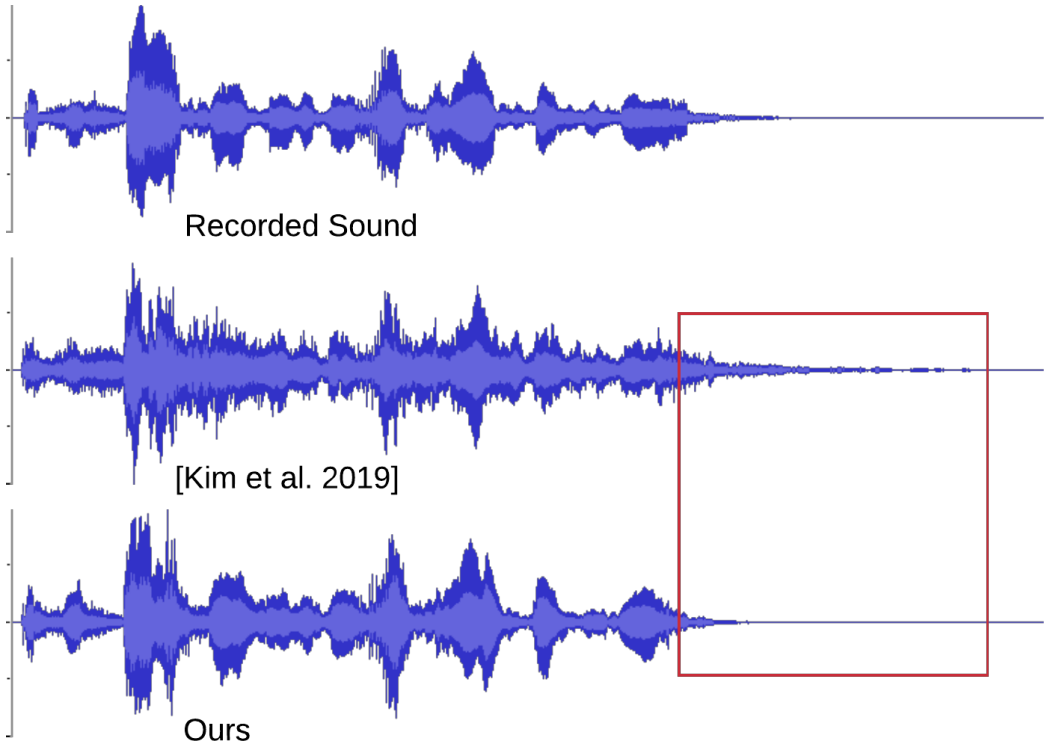


Figure 3.11: We demonstrate the importance on T_{60} optimization on audio amplitude waveform. Our method optimizes the material parameters based on input audio and matches the tail shape and decay amplitude with the recorded sound, whereas the visual-based object materials from Kim et al. (2019) failed to compensate for the audio effects.

virtual characters or speakers, such as the ones shown in Figure 3.1, can be added to the video, generating sound that is consistent with the real-world environment.

Real-time Immersive Augmented Reality Audios Our method works in a real-time manner and can be integrated into modern AR systems. AR devices are capable of capturing real-world geometry, and can stream audio input to our pipeline. At interactive rates, we can optimize and update the material properties, and update the room EQ filter as well. Our method is not hardware-dependent and can be used with any AR device (which provides geometry and audio) to enable a more immersive listening experience.

Real-world Computer-Aided Acoustic Design Computer-aided design (CAD) software has been used for designing architecture acoustics, usually before construction is done, in a predictive manner (Pelzer et al., 2014; Kleiner et al., 1990). But when given an existing real-world environment, it becomes challenging for traditional CAD software to adapt to current settings because acoustic measurement can be tedious and error-prone. By using our method, room materials and EQ properties can be estimated from simple input, and can be further fed to other acoustic design applications in order to improve the room acoustics such as material replacement, source and listener placement (Morales et al., 2019), and soundproofing setup.

3.5 Perceptual Evaluation

We perceptually evaluated our approach using a critical listening test. For this test, we studied the perceptual similarity of a reference speech recording with speech recordings convolved with simulated impulse responses. We used the same speech content for the reference and all stimuli under testing and evaluated how well we can reconstruct the same identical speech content in a given acoustic scene. This is useful for understanding the absolute performance of our approach compared to the ground truth results.

3.5.1 Design and Procedure

For our test, we adopted the multiple stimulus with hidden reference and anchor (MUSHRA) methodology from the ITU-R BS.1534-3 recommendation (Series, 2014). MUSHRA provides a protocol for the subjective assessment of intermediate quality level of audio systems (Series, 2014) and has been adopted for a wide variety of audio processing tasks such as audio coding, source separation, and speech synthesis evaluation (Schoeffler et al., 2015; Cartwright et al., 2016).

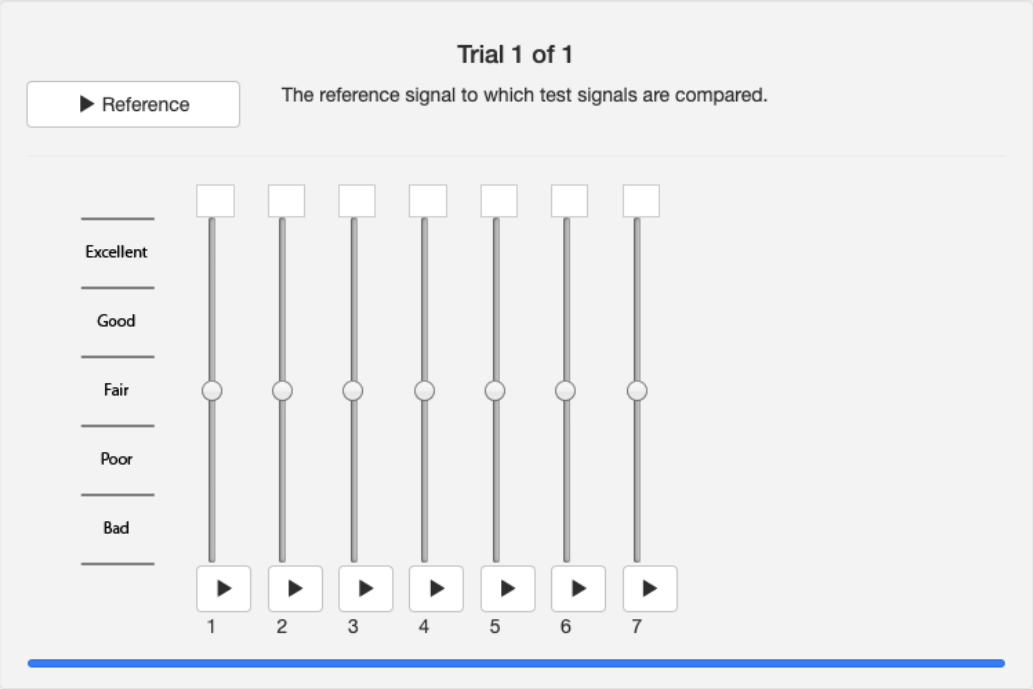


Figure 3.12: A screenshot of MUSHRA-like web interface used in our user study. The design is from Cartwright et al. (2016).

In a single MUSHRA trial, participants are presented with a high-quality reference signal and asked to compare the quality (or similarity) of three to twelve stimuli on a 0-100 point scale using a set of vertical sliders as shown in Figure 3.12. The stimuli must contain a hidden reference (identical to the explicit reference), two anchor conditions – low-quality and high-quality, and any additional conditions under study (maximum of nine). The hidden reference and anchors are used to help the participants calibrate their ratings relative to one another, as well as to filter out inaccurate assessors in a post-screening process. MUSHRA tests serve a similar purpose to mean opinion (MOS) score tests (Series, 2016), but requires fewer participants to obtain results that are statistically significant.

We performed our studies using Amazon Mechanical Turk (AMT), resulting in a MUSHRA-like protocol (Cartwright et al., 2016). In recent years, web-based MUSHRA-like tests have become a standard methodology and have been shown to perform equivalently to full, in-person tests(Schoeffler et al., 2015; Cartwright et al.,

2016).

3.5.2 Participants

We recruited 261 participants on AMT to rate one or more of our five acoustic scenes under testing following the approach proposed by Cartwright et al. (2016). To increase the quality of the evaluation, we pre-screened the participants for our tests. To do this, we first required that all participants have a minimum number of 1000 approved Human Intelligence Task (HITs) assignments and have had at least 97 percent of all assignments approved. Second, all participants must pass a hearing screening test to verify they are listening over devices with adequate frequency response. This was performed by asking participants to listen to two separate eight second recordings consisting of a 55Hz tone, a 10kHz tone and zero to six tones of random frequency. If any user failed to count the number of tones correctly after two or more attempts, they were not allowed to proceed.

3.5.3 Training

After having passed our hearing screening test, each user was presented with a one page training test. For this, the participant was provided two sets of recordings. The first set of training recordings consisted of three recordings: a reference, a low-quality anchor, and a high-quality anchor. The second set of training recordings consisted of the full set of recordings used for the given MUSHRA trail, albeit without the vertical sliders present. To proceed to the actual test, participants were required to listen to each recording in full. In total, the training time was estimated to take approximately two minutes.

3.5.4 Stimuli

For our test conditions, we simulated five different acoustic scenes. For each scene, a separate MUSHRA trial was created. In AMT language, each scene was presented as a separate HIT per user. For each MUSHRA trial or HIT, we tested the following stimuli: hidden reference, low-quality anchor, mid-quality anchor, baseline T_{60}, Baseline T_{60}+EQ, proposed T_{60}, and proposed T_{60}+EQ.

As noted by the ITU-R BS.1534-3 specification (Series, 2014), both the reference and anchors will have a significant effect on the test results, must resemble the artifacts from the systems, and must be designed carefully. For our work, we set the hidden reference as an identical copy of the explicit reference (required), which consisted of speech convolved with the ground truth IR for each acoustic scene. Then, we set the low-quality anchor to be completely anechoic, non-reverberated speech. We set the mid-quality anchor to be speech convolved with an impulse response with a 0.5 second T_{60} (typical conference room) across frequencies, and uniform equalization.

For our baseline comparison, we included two baseline approaches following previous work (Li et al., 2018). More specifically, our Baseline T_{60} leverages the geometric acoustics method proposed by Cao et al. (2017) as well as the materials analysis calibration method of Li et al. (2018). Our Baseline T_{60}+EQ extends this and includes the additional frequency equalization analysis (Li et al., 2018). These two baselines directly correspond to the proposed materials optimization (Proposed T_{60}) and equalization prediction subsystems (Proposed T_{60}+EQ) in our work. The key difference is that we blindly estimate the parameters necessary for both steps *blindly from speech*.

3.5.5 User Study Results

When we analyzed the results of our listening test, we post-filtered the results following the ITU-R BS.1534-3 specification (Series, 2014). More specifically, we excluded assessors if they

- rated the hidden reference condition for $> 15\%$ of the test items lower than a score of 90
- or, rated the mid-range (or low-range) anchor for more than 15% of the test items higher than a score of 90.

Using this post-filtering, we reduce our collected data down to 70 unique participants and 108 unique test trials, spread across our five acoustic scene conditions.

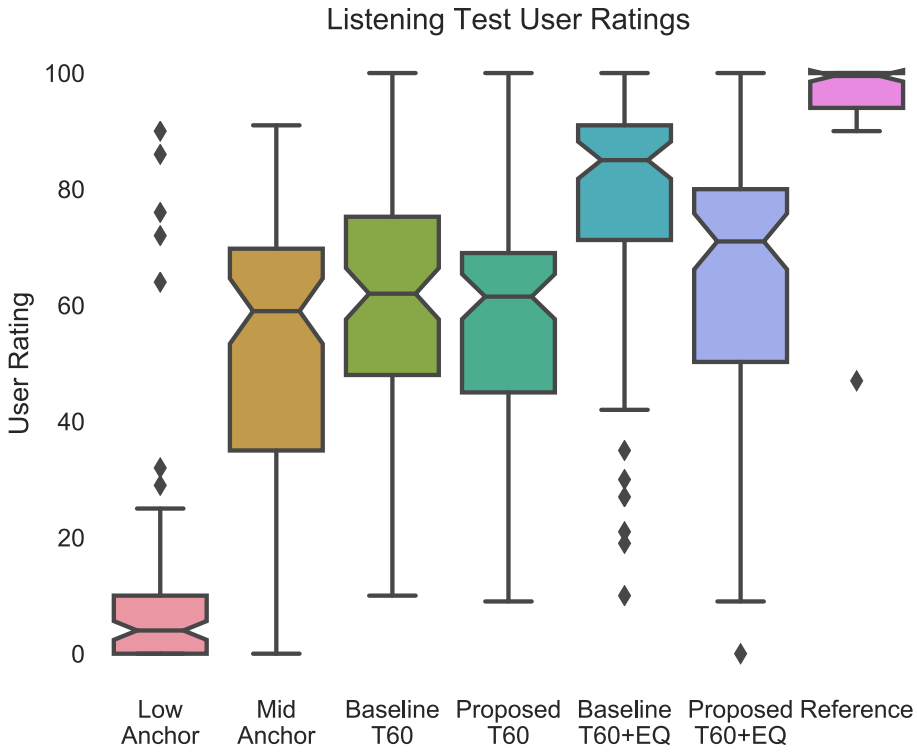


Figure 3.13: Box plot results for our listening test. Participants were asked to rate how similar each recording was to the explicit reference. All recordings have the same content, but different acoustic conditions. Note our proposed T_{60} and T_{60}+EQ are both better than the Mid-Anchor by a statistically significant amount ($approx$10 rating points on a 100 point scale).

We show the box plots of our results in Figure 3.13. The median ratings for each stimulus include: Baseline T_{60} (62.0), Baseline T_{60}+EQ (85.0), Low-Anchor (40.5), Mid-Anchor (59.0), Proposed T_{60} (61.5), Proposed T_{60}+EQ (71.0), Hidden Reference (99.5). As seen, the Low-Anchor and Hidden Reference outline the range of user

scores for our test. In terms of baseline approaches, the Proposed T_{60}+EQ method achieves the highest overall listening test performance. We then see that our proposed T_{60} method and T_{60}+EQ method outperform the mid-anchor. Our proposed T_{60} method is comparable to the baseline T_{60} method, and our proposed T_{60}+EQ method outperforms our proposed T_{60}-only method.

To understand the statistical significance, we perform paired t-tests between stimuli pairs. The p-value between Baseline T_{60} and Proposed T_{60} is 0.09, suggesting that we cannot reject the null hypothesis of identical average scores between prior work (which uses manually measured IRs) and our work. The p-value of Baseline T_{60}+EQ and Proposed T_{60}+EQ, however, is 1.85e-6, suggesting our EQ method has a statistically different average (lower). The p-value of Proposed T_{60} and Proposed T_{60}+EQ, however, is 0.004, suggesting our EQ method does significantly improve performance compared to our proposed T_{60}-only subsystem. We also note that the p-value of the Mid-Anchor and Proposed T_{60}+EQ is 0.0002, suggesting our method is statistically different (higher performing) on average than simply using an average room T_{60} and uniform equalization.

In summary, we see that our proposed T_{60} computation method is comparable to prior work, albeit we perform such estimation directly from a short speech recording rather than relying on intrusive IR measurement schemes. Further, our proposed complete system (Proposed T_{60}+EQ) outperforms both the mid-anchor and proposed T_{60} system alone, demonstrating the value of EQ estimation. Finally, we note our proposed T_{60}+EQ method does not perform as well as prior work, largely due to the EQ estimation subsystem. This result, however, is expected as prior work requires manual IR measurements, which result in perfect EQ estimation. This is in contrast to our work, which directly estimates both T_{60} and EQ parameters from recorded speech, enabling a drastically improved interaction paradigm for matching acoustics in several applications.

3.6 Summary

We present a new pipeline to estimate, optimize, and render immersive audio in video and mixed reality applications. We present novel algorithms to estimate two important acoustic environment characteristics – the frequency-dependent reverberation time and equalization filter of a room. Our multi-band octave-based prediction model works in tandem with our equalization augmentation and provides robust input to our improved materials optimization algorithm. Our user study validates the perceptual importance of our method. To the best of our knowledge, our method is the first method to predict IR equalization from raw speech data and validate its accuracy.

Limitations and Future Work To achieve a perfect acoustic match, one would expect the real-world validation error to be zero. In reality, zero error is only a sufficient but not necessary condition. In our evaluation tests, we observe that small validation errors still allow for plausible acoustic matching. While reducing the prediction error is an important direction, it is also useful to investigate the perceptual error threshold for acoustic matching for different tasks or applications. Moreover, temporal prediction coherence is not in our evaluation process. This implies that given a sliding windows of audio recordings, our model might predict temporally incoherent T_{60} values. One interesting problem is to utilize this coherence to improve the prediction accuracy and can be an interesting future direction.

Modeling real-world characteristics in simulation is a non-trivial task – as in previous work along this line, our simulator does not fully recreate the real world in terms of precise details. For example, we did not consider the speaker or microphone response curve in our simulation. In addition, sound sources are modeled as omnidirectional sources (Cao et al., 2017), where real sources exhibit certain directional patterns. It remains an open research challenge to perfectly replicate and simulate

our real world in a simulator.

Like all data-driven methods, our learned model performs best on the same kind of data on which it was trained. Augmentation is useful because it generalizes the existing dataset so that the learned model can extrapolate to unseen data. However, defining the range of augmentation is not straightforward. We set the MIT IR dataset as the baseline for our augmentation process. In certain cases, this assumption might not generalize well to estimate the extreme room acoustics. We need to design better and more universal augmentation training algorithms. Our method focused on estimation from speech signals, due to their pervasiveness and importance. It would be useful to explore how well the estimation could work on other audio domains, especially when interested in frequency ranges outside typical human speech. This could further increase the usefulness of our method, e.g., if we could estimate acoustic properties from ambient/HVAC noise instead of requiring a speech signal.

Chapter 4

Fast Learning-Based Acoustic Scattering[1]

Figure 4.1: We show the dynamic scenes with various moving objects that are used to evaluate our hybrid sound propagation algorithm. We compute the acoustic scattered fields of each object using a neural network and couple them with interactive ray tracing to generate diffraction and occlusion effects. Our approach can generate plausible acoustic effects in dynamic scenes in a few milliseconds and we demonstrate its benefits for sound rendering in virtual environments.

4.1 Introduction

Interactive sound propagation and rendering are increasingly used to generate plausible sounds that can improve a user's sense of presence and immersion in virtual environments (Larsson et al., 2002; Liu and Manocha, 2020). Recent advances in ge-

[1]The work in this chapter has been published in Tang et al. (2021)

ometric and wave-based simulation methods have lead to integration of these methods into current games and virtual reality (VR) applications to generate plausible acoustic effects, including Project Acoustics (Mic, 2019), Oculus Spatializer (Ocu, 2019), and Steam Audio (Ste, 2018). The underlying propagation algorithms are based on using reverberation filters (Valimaki et al., 2012), ray tracing (Schissler et al., 2014; Schissler and Manocha, 2018), or precomputed wave-based acoustics (Raghuvanshi and Snyder, 2014b).

A key challenge in interactive sound rendering is handling dynamic scenes that are frequently used in games and VR applications. Not only can the objects undergo large motion or deformation, but their topologies may also change. In addition to specular and diffuse effects, it is also important to simulate complex diffracted scattering, occlusions, and inter-reflections that are perceptible (James et al., 2006; Pulkki and Svensson, 2019; Raghuvanshi and Snyder, 2014b). Prior geometric methods are accurate in terms of simulating high-frequency effects and can be augmented with approximate edge diffraction methods that may work well in certain cases (Tsingos et al., 2001; Schissler et al., 2014), though their behavior can be erratic (Rungta et al., 2016). On the other hand, wave-based precomputation methods can accurately simulate these effects, but are limited to static scenes (Raghuvanshi and Snyder, 2014b, 2018). Some hybrid methods are limited to interactive dynamic scenes with well-separated rigid objects (Rungta et al., 2018). Our goal is to design similar hybrid methods that can overcome these restrictions and can generate diffraction and occlusion effects that translate into good perceptual differentiation (Rungta et al., 2016).

Many recent works use machine learning techniques for audio processing, including recovering acoustic parameters of real-world scenes from recordings (Eaton et al., 2016; Genovese et al., 2019; Tsokaktsidis et al., 2019). Furthermore, machine learning methods have been used to approximate diffraction scattering and occlusion effects from rectangular plate objects (Pulkki and Svensson, 2019) and frequency-dependent

loudness fields for 2D convex shapes (Fan et al., 2020a). These results are promising and have motivated us to develop good learning based methods for more general 3D objects.

Main Results: We present a novel approach to approximate the acoustic scattering field of an object in 3D using neural networks for interactive sound propagation in dynamic scenes. Our approach makes no assumption about the motion or topology of the objects. We exploit properties of the acoustic scattering field of objects for lower frequencies and use neural networks to learn this field from geometric representations of the objects.

Given an object in 3D, we use the neural network to estimate the scattered field at runtime, which is used to compute the propagation paths when sound waves interact with objects in the scene. The radial part of the acoustic scattering field is estimated using geometric ray tracing, along with specular and diffuse reflections. Some of the novel components of our work include:

- **Learning acoustic scattering fields:** We use techniques based on geometric deep learning to approximate the angular component of acoustic wave propagation in the wave-field. Our neural network takes the point cloud as the input and outputs the spherical harmonic coefficients that represent the acoustic scattering field. We compare the accuracy of our learning method with an exact BEM solver, and the error on new, unseen objects (as compared to training data). Our empirical results are promising and we observe average normalized reproduction error(Lilis et al., 2010; Betlehem and Abhayapala, 2005) of 8.8% in the pressure fields.

- **Interactive wave-geometric sound propagation:** We present a hybrid propagation algorithm that uses a neural network-based scattering field representation along with ray tracing to efficiently generate specular, diffuse, diffraction, and occlusion effects at interactive rates.

- **Plausible sound rendering for dynamic scenes:** We present the first interactive approach for plausible sound rendering in dynamic scenes with diffraction modeling and occlusion effects. As the objects deform or change topology, we compute a new spherical harmonic representation using the neural network. Compared with prior interactive methods, we can handle unseen objects at real-time, without using precomputed transfer functions for each object.

- **Perceptual evaluation:** We perform a user study to validate the perceptual benefits of our method. Our propagation algorithm generates more smooth and realistic sound and has increased perceptual differentiation over prior methods used for dynamic scenes (Schissler and Manocha, 2017; Rungta et al., 2018).

We demonstrate the performance in dynamic scenes with multiple moving objects and changing topologies. The additional runtime overhead of estimating the scattering field from neural networks is less than 1ms per object on a NVIDIA GeForce RTX 2080 Ti GPU. The overall running time of sound propagation is governed by the underlying ray tracing system and takes few milliseconds per frame on multi-core desktop PC. We also evaluate the accuracy of acoustic scattering fields, as shown in Figure 4.7.

4.2 Related Work

4.2.1 Interactive Sound Rendering in Dynamic Scenes

At a broad level, techniques for dynamic scenes can be classified into reverberation filters, geometric and wave-based methods, and hybrid combinations. The simplest and lowest-cost algorithms are based on artificial reverberators (Valimaki et al., 2012), which simulate the decay of sound in rooms. These filters are designed based on different parameters and are either specified by an artist or computed using scene

characteristics (Tsingos, 2009). They can handle dynamic scenes but assume that the reverberant sound field is diffuse, making them unable to generate directional reverberation or time-varying effects.

Many interactive techniques based on geometric acoustics and ray tracing have been proposed for dynamic scenes (Vorländer, 1989; Taylor et al., 2012a; Schissler and Manocha, 2017). They use spatial data structures along with multiple cores on commodity processors and caching techniques to achieve higher performance. Furthermore, hybrid combinations of ray tracing and reverberation filters (Schissler and Manocha, 2018) have been proposed for low-power, mobile devices. In practice, these methods can handle scenes with a large number of moving objects, along with sources and the listener, but can't model diffraction or occlusion effects well.

Many precomputation-based wave acoustics techniques tend to compute a global representation of the acoustic pressure field. They are limited to static scenes, but can handle real-time movement of both sources and the listener (Raghuvanshi et al., 2010; Mehra et al., 2015). These representations are computed based on uniform or adaptive sampling techniques (Chaitanya et al., 2019). Overall, the acoustic wave field is a complex high-dimensional function and many efficient techniques have been designed to encode this field (Raghuvanshi and Snyder, 2014b, 2018) within 100MB and with a small runtime overhead. A hybrid combination of BEM and ray tracing has been presented for dynamic scenes with well-separated rigid objects (Rungta et al., 2018). A recent *Planeverb* system (Rosen et al., 2020) is able to perform 2D wave simulation at interactive rates and calculate perceptual acoustic parameters that can be used for sound rendering.

4.2.2 Machine Learning and Acoustic Processing

Machine learning techniques are increasingly used for acoustic processing applications. These include isolating the source locations in multipath environments (Fer-

guson et al., 2018) and recovering the room acoustic parameters corresponding to reverberation time, direct-to-reverberant ratio, room volume, equalization, etc. from recorded signals (Eaton et al., 2016; Genovese et al., 2019; Tsokaktsidis et al., 2019; Tang et al., 2019a). These parameters are used for speech processing or audio rendering in real-world scenes. Neural networks have also been used to replace the expensive convolution operations for fast auralization (Tenenbaum et al., 2019), to render the acoustic effects of scattering from rectangular plate objects for VR applications (Pulkki and Svensson, 2019), or to learn the mapping from convex shapes to the frequency dependent loudness field (Fan et al., 2020a). The last method formulates the scattering function computation as a high-dimension image-to-image regression and is mainly limited to convex objects that are isomorphic to spheres. In contrast, our learning-based method can compute a good approximation of the acoustic scattering field of arbitrary objects (e.g. non-convex or non-manifold).

4.3 Acoustic Scattering Preliminary

4.3.1 Helmholtz Equation

We can analyze the acoustic pressure field in the frequency domain by converting $P(\mathbf{x}, t)$ from Equation (2.2) using Fourier transform

$$p(\mathbf{x}, \omega) = \mathcal{F}_t\{P(\mathbf{x}, t)\} = \int_{-\infty}^{\infty} P(\mathbf{x}, t) e^{-j\omega t} dt. \tag{4.1}$$

At each frequency ω the pressure field satisfies the homogeneous Helmholtz wave equation

$$(\nabla^2 + k^2) p(\mathbf{x}, \omega) = 0, \tag{4.2}$$

where $k = \frac{\omega}{c}$ is the wavenumber. We can expand the Laplacian operator in terms of spherical coordinates (r, θ, ϕ) as

$$\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \phi^2} + k^2\right) p = 0. \tag{4.3}$$

The general free-field solution of (4.3) can be formulated as

$$p(\mathbf{x}, \omega) = \sum_{l=0}^{\infty} \sum_{m=-l}^{+l} \left[A_{lm} h_l^{(1)}(kr) + B_{lm} h_l^{(2)}(kr)\right] Y_l^m(\theta, \phi), \tag{4.4}$$

where $h_l^{(1)}$ and $h_l^{(2)}$ are Hankel functions of the first and the second kind, respectively. A_{lm} and B_{lm} are arbitrary constants, $A_{lm} h_l^{(1)}(kr) + B_{lm} h_l^{(2)}(kr)$ together represents the radial part of the solution and the spherical harmonics term $Y_l^m(\theta, \phi)$ represents the angular part of the solution.

4.3.2 Acoustic Wave Scattering

Equation (4.2) describes the behavior of acoustic waves in free-field conditions. When a propagating acoustic wave generated by a sound source interacts with an obstacle (the scatterer), a scattered field is generated outside the scatterer. The Helmholtz equation can be used to describe this scenario:

$$(\nabla^2 + k^2) p(\mathbf{x}, \omega) = -Q(\mathbf{x}, \omega), \quad \forall\, \mathbf{x} \in E, \tag{4.5}$$

where E is the space that is exterior to the scatterer and $Q(\mathbf{x}, \omega)$ represents the acoustic sources in the frequency domain. Common types of sound sources include monopole sources, dipole sources, and plane wave sources. To obtain an exact solution to Equation (4.5), the boundary conditions on the scatterer surface S need to be specified. In this work, we assume all the scattering objects are sound-hard (i.e. all energy is scattered, not absorbed) and therefore use the zero Neumann boundary

condition for all S:

$$\frac{\partial p}{\partial \mathbf{n}(\mathbf{x})} = 0, \quad \forall \mathbf{x} \in S, \tag{4.6}$$

where $\mathbf{n}(\mathbf{x})$ is the normal vector at $\mathbf{x}$. Alternatively, other conditions including the sound-soft Dirichlet boundary condition and the mixed Robin boundary condition (Pierce and Beyer, 1990) can be used to model different acoustic scattering problems. When the boundary conditions are fully defined, the constants in Equation (4.4) can be uniquely determined.

4.3.3 Global and Localized Sound Fields

Sound fields typically refer to the sound energy/pressure distribution over a bounded space as generated by one or more sound sources. The global sound field in an acoustic environment depends on each sound source location, the propagating medium, and any reflections from boundary surfaces and objects. This requires solving the wave equation in the free-field condition and evaluating inter-boundary interactions of sound energy using a global numeric solver (details in § 4.3.1). In this case, the position of all scene objects/boundaries and sound sources needs to be specified beforehand, and any change in these conditions changes the sound field. The exact computation of the global pressure field is very expensive and can take tens of hours on a cluster (Mehra et al., 2013; Raghuvanshi et al., 2010; Raghuvanshi and Snyder, 2014b).

Our goal is to generate plausible sounds in virtual environments with dynamic objects. Therefore, it is important to model the acoustic scattering field (ASF) of each object. The ASFs of different objects are used to represent the localized pressure field, which is needed for diffraction and inter-reflection effects (James et al., 2006; Mehra et al., 2013). At the same time, the sound field in the free space (e.g., the far-field) between two distant objects is approximated using ray tracing, and we do not

compute that pressure field accurately using a wave-solver. In practice, computing the sound field in a localized space for each object in the scene is much simpler and easier to represent than using a global solver (Mehra et al., 2013; Rungta et al., 2018).

4.3.4 Overview

We present a learning method to approximate the ASFs of static or dynamic 3D objects of moderate sizes. In terms of correlation between the object shape and its scattering field, the volume of the scatterer closely relates to its low-order shape characteristics that can be represented by coarse triangle faces, which dominate the low-frequency scattering behaviors; while at high frequencies, this relationship shifts to high-order shape characteristics (i.e., geometrical details). Given the powerfulness of deep learning inference, we hypothesize the scattering sound distribution can be directly learned from the scatterer geometry, without solving the complicated wave equations. The inference speed on a modern GPU far exceeds conventional wave solvers, making deep neural networks suitable for interactive sound rendering applications. Therefore, we propose using appropriate 3D representation of objects to feed a neural network that can learn its corresponding scattered acoustic pressure field. We build and evaluate our method mainly on low frequency sounds and leverage state-of-the-art geometric ray-tracing techniques to handle high frequency sounds.

For each object, we consider a spherical grid of incoming directions and model the plane-waves from each direction of this grid. For each plane wave, our goal is to compute the scattered field for the object on an offset surface of the object. Our geometric deep learning method is used to compute the angular portion of the scattered field (Equation 4.4). If two objects move and are in a touching configuration, our learning algorithm treats them as a one large object and estimates its scattered field. Similarly, we can recompute the scattered field for a deforming object. An overview of our approach is illustrated in Figure 4.2.

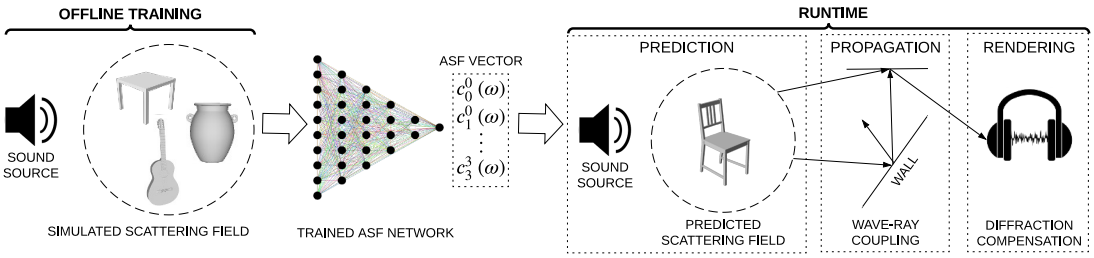


Figure 4.2: **Overview:** Our algorithm consists of the training stage and the runtime stage. The training stage uses a large dataset of 3D objects and their associated acoustic pressure fields computed using an accurate BEM solver to train the network. The runtime stage uses the trained neural network to predict the sound pressure field from a point cloud approximation of different objects at interactive rates.

4.4 Learning-based Sound Scattering

4.4.1 Wave Propagation Modeling

Our approach is designed for synthetic scenes and we assume a geometric representation (e.g., triangle mesh) is given to us. So the acoustic scattering field $p(\mathbf{x}, \omega)$ around the object can be solved numerically (derivation in § 4.3.1 and 4.3.2). In this work, we propose modeling the angular part of the scattering field using our learning based pressure field inference. The radial part is approximated using geometric sound propagation techniques.

Radial Decoupling

Our goal is to determine the scattering field exterior to an object using a wave-solver. This field needs to be compactly encoded for efficient training. As shown in Equation (4.4), acoustic wave propagation in the free-field can be decomposed into radial and angular components. Furthermore, the radial sound pressure in the far-field follows the *inverse-distance law* (Beranek and Mellow, 2012): $p \sim 1/r$, as shown in Figure 4.3. We utilize this property to extrapolate the full ASF from one of its far-field "snapshots" at a fixed radius, so that the full ASF does not need to be stored. Following the inverse-distance law, the sound pressure at any far-field location (r, θ, ϕ)

can be computed as

$$p(r, \theta, \phi, \omega) = \frac{r_{ref}}{r} p(r_{ref}, \theta, \phi, \omega), \tag{4.7}$$

where r_{ref} is the reference distance and only $p(r_{ref}, \cdot, \cdot, \cdot)$ needs to be computed and stored. For brevity, we omit r in following sections.

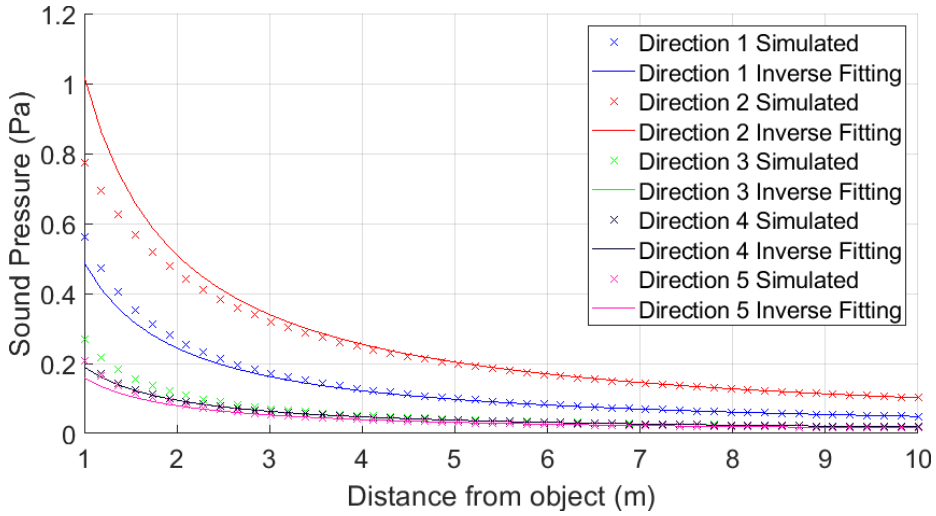


Figure 4.3: **Simulated sound pressure fall-off and inverse-distance law fitted curves:** We calculate the sound pressure around a sound scatterer in our dataset using the BEM solver as reference. We examine the sound pressure from $1m$ to $10m$ scattered along 5 directions ($0°, 72°, 144°, 216°$, and $288°$). We regard the sound pressure value at $10m$ to correspond to far-field condition, and inversely fit the pressure values for distance within $10m$ according to Equation 4.7. We use$r_{ref} = 5m$ is used for generating our ASFs, although other values can be used as well.

Angular Pressure Field Encoding

A spherical field consisting of a fixed number of points (e.g., 642 points evenly distributed on a sphere surface) is obtained by generating an icosphere with 4 subdivisions. Real valued scattered sound pressures are evaluated at these field points during wave-based simulation. Spherical harmonics (SH) can represent a spherical scalar field compactly using a set of SH coefficients; they have been widely used for 3D sound field recording and reproduction (Poletti, 2005). SH function up to order l_{max} has $M = (l_{max} + 1)^2$ coefficients. The angular pressure at the outgoing direction

(θ, ϕ) can be evaluated as

$$p(\theta, \phi, \omega) = \sum_{l=0}^{l_{max}} \sum_{m=-l}^{+l} Y_l^m(\theta, \phi) c_l^m(\omega), \tag{4.8}$$

where $Y_l^m(\theta, \phi)$ are the SH basis functions at degree l and order m, and $c_l^m(\omega)$ are the SH coefficients that encode our angular pressure fields. Increasing the number of coefficients can lead to more challenges because the dimension of our learning target is raised.

4.4.2 Learning Spherical Pressure Fields

We need an appropriate geometric representation for the underlying objects in the scene so that we can apply geometric deep learning methods to compute the ASF. It is important that our approach should be able handle dynamic scenes with moving objects or changing topology. It can be difficult to handle such scenarios with mesh-based representations (Hanocka et al., 2019; Tan et al., 2018; Zheng et al., 2017). For example, (Hanocka et al., 2019) calculates intrinsic geodesic distances for convolution operations, which cannot be applied when one big object breaks into two.

Our approach uses a point cloud representation of the objects in the scene as an input. And we leverage the PointNet (Charles et al., 2017) architecture to regress the spherical harmonics term c_l^m in Equation (4.8). PointNet is a highly efficient and effective network architecture that works on raw point cloud input, and can perform various tasks including 3D object classification, semantic segmentation and our ASF regression. It also respects the permutation invariance of points. We slightly modify its output layers to predict the SH vector as shown in Figure 4.4.

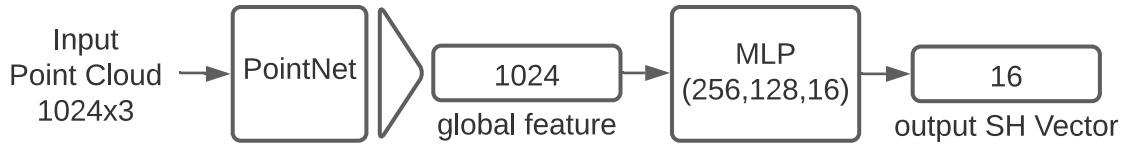


Figure 4.4: **PointNet regression:** Given an input point cloud with $N = 1024$ 3D points, we feed it to the PointNet architecture (Charles et al., 2017) until maxpooling to extract the global feature. Then we use multi-layer perceptrons (MLPs) of layer size 256, 128, and 16 to map the feature to a SH vector of length 16 representing the scattering field.

4.5 Interactive Sound Propagation with Wave-Ray Coupling

In this section, we describe how our learning-based method can be combined with geometric sound propagation techniques to compute the impulse responses for given source and listener positions. Then, we can render them in highly dynamic scenes.

Hybrid Sound Propagation We use a hybrid sound propagation algorithm that combines wave-based and ray acoustics. Each of them handles different parts of wave acoustics phenomena, but they are coupled in terms of incoming and outgoing energies at multiple localized scattering fields. Specifically, our trained neural network estimates the scattering field and is used to compute propagation paths when sound interacts with obstacles in the scene. On the other hand, modeling sound propagation in the air along with specular and diffuse reflections at large boundary surfaces (e.g., walls, floors) is computed using ray tracing methods (Schissler et al., 2014; Schissler and Manocha, 2017; Rungta et al., 2018).

Ray Tracing with Localized Fields Our localized ASFs are represented using SH coefficients. Given the most general ray tracing formulation at a scattering surface, the sound intensity I_{out} of an outgoing direction (θ_o, ϕ_o) from a scattering surface is

given by the integral of the incoming intensity from all directions:

$$I_{out}(\theta_o, \phi_o, \omega) = \int_S I_{in}(\theta_i, \phi_i, \omega) f(\theta_i, \phi_i, \theta_o, \phi_o, \omega) dS, \tag{4.9}$$

where S represents the directions on a spherical surface around the ray hit point, $I_{in}(\theta_i, \phi_i, \omega)$ is the incoming sound intensity from direction (θ_i, ϕ_i), and $f(\theta_i, \phi_i, \theta_o, \phi_o, \omega)$ is the bi-directional scattering distribution function (BSDF) that is commonly used in visual rendering (Pharr et al., 2016). Our problem of acoustic wave scattering is different from visual rendering in two aspects: (1) sound wave scatters around objects, whereas light mostly transmits to visible directions or propagates through transparent materials; (2) BSDFs are point-based functions that depend on both incoming and outgoing directions, whereas our localized scattered fields are region-based functions. Therefore, we replace BSDFs in Equation (4.9) with our localized scattered field $p(\theta, \phi, \omega)$ representation from Equation (4.8). Our choice of a spherical offset surface to model the scattered field also enables us to perform integration over the whole spherical surface in a straightforward manner, since evaluating spherical coordinates is efficient with SH functions. Although $p(\theta, \phi, \omega)$ encodes only the outgoing directions and assumes incoming plane waves to $-x$ direction, one can easily rotate the point cloud to align any incoming direction to the $-x$ direction and use our network to infer $p(\theta, \phi, \omega)$ at that direction. We update Equation (4.9) to

$$I_{out}(\theta_o, \phi_o, \omega) = \int_S I_{in}(\theta_i, \phi_i, \omega) p^2(\theta_i, \phi_i, \omega) dS. \tag{4.10}$$

We use the Monte Carlo integration to numerically evaluate the outgoing scattered intensity:

$$I_{out}(\theta_o, \phi_o, \omega) \approx \frac{1}{N} \sum_{j=1}^{N} \frac{I_{in}(\theta_j, \phi_j, \omega) p^2(\theta_j, \phi_j, \omega)}{Pr(\theta_j, \phi_j)}, \tag{4.11}$$

where N is the number of samples and $Pr(\theta_j, \phi_j)$ is the probability of generating a sample for direction (θ_j, ϕ_j). A uniform sampling over the sphere surface gives $Pr(\theta_j, \phi_j) = \frac{1}{4\pi}$. As N increases, the approximation becomes more accurate.

Diffraction Compensation In wave acoustics, the total sound field at a position can be decomposed into the sum of the free-field sound pressure and the scattered sound field. Similar to (Rungta et al., 2018), we only have computed the scattered sound field up to now. But when the listener is obstructed from the sound source, the traditional ray-tracing algorithm will miss the contribution from the free-field, which will result in a very unnatural phenomenon: the sound would be greatly attenuated by a single obstacle if we only render the scattered sound, whereas in a realistic setup, low-frequency sound should not be attenuated by a small obstacle by much. To address this issue in a ray-tracing context, we propose to approximate sound interference with and without an obstacle depending on an extra visibility check. Specifically, for a sound source from direction (θ_j, ϕ_j) and the listener at (θ_o, ϕ_o), we calculate the sound at the listener position based on whether they are blocked by a scatterer from each other as:

$$I_{out}(\theta_o, \phi_o, \omega) \approx \begin{cases} \frac{1}{N}\sum_{j=1}^{N} \frac{I_{in}(\theta_j,\phi_j,\omega)(1-p^2(\theta_j,\phi_j,\omega))}{Pr(\theta_j,\phi_j)}, \text{if invisible} \\ \frac{1}{N}\sum_{j=1}^{N} \frac{I_{in}(\theta_j,\phi_j,\omega)p^2(\theta_j,\phi_j,\omega)}{Pr(\theta_j,\phi_j)}, \text{if visible} \end{cases} \tag{4.12}$$

Note that the visible case remains the same as Equation (4.11), because the direct response will be automatically accounted for by the original ray-tracing pipeline. Obviously, this implementation is not physically accurate compared with wave acoustic simulations, since additional phase information is missing. However, this formulation will generate more realistic and more smooth sound rendering than prior work that only considers the scattering field, and we verify its benefits through a perceptual evaluation in § 4.7.

4.6 Implementation and Results

In this section, we describe our implementation details and demonstrate the performance on many dynamic benchmarks.

4.6.1 Data Generation

Dataset To generate our learning examples, we choose to use the *ABC Dataset* (Koch et al., 2019). This dataset is a collection of one million general Computer-Aided Design (CAD) models and is widely used for evaluation of geometric deep learning methods and applications. In particular, this dataset has been used to estimate of differential quantities (e.g., normals) and sharp features, which makes it attractive for learning ASFs as well. We sample 100,000 models from the *ABC Dataset* and process them by scaling objects such that their longest dimension is in the range of $[1m, 2m]$. The choice of such an object size limit is not fixed and can depend on the specific problem domain (e.g., size of objects used in applications like games or VR). Because the scattered pressure field is orientation-dependent, we augment our models by applying random 3D rotations to the original dataset to create an equal-sized rotation augmented dataset. To generate accurate labeled data, we use an accurate BEM wave solver, placing a plane wave source with unit strength propagating to the $-x$ direction. The solver outputs the ASF for each object, which becomes our learning target. The dataset pipeline is also illustrated in Figure 4.5.

Mesh Pre-processing The original meshes from the *ABC Dataset* have high levels of details with fine edges of length shorter than $1cm$. Dense point cloud inputs could also be modeled or collected from the real-world scenes with granularity similar to this dataset. However, a high number of triangle elements in a mesh will significantly increase the simulation time of BEM solvers. For wave-based solver, our highest simulation frequency is $1000Hz$, which converts to a wavelength of $34cm$. Therefore,

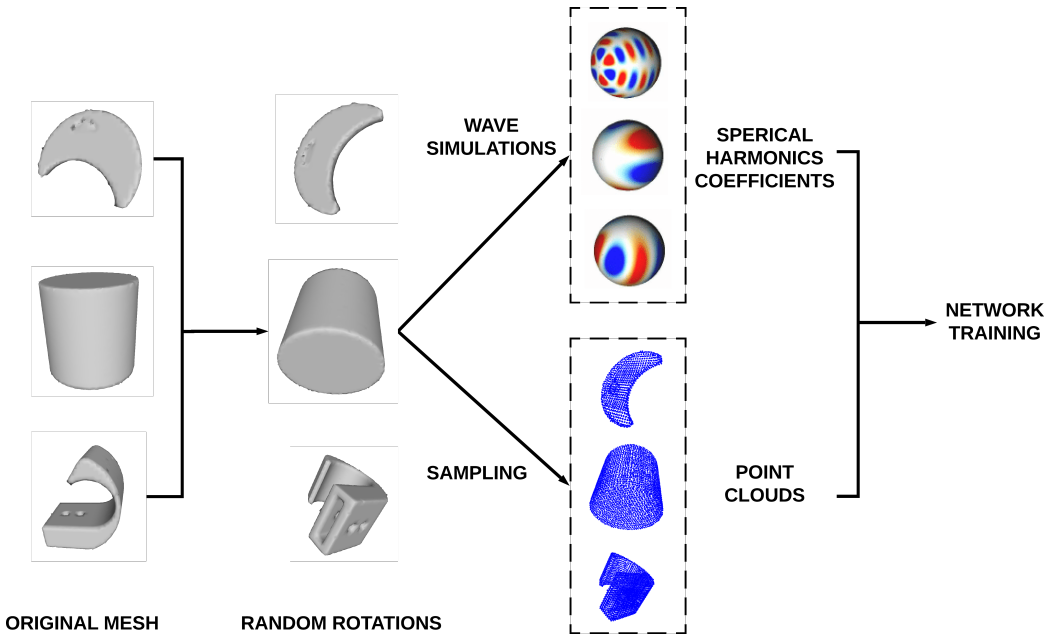


Figure 4.5: **Our dataset generation pipeline for neural network training:** Given a set of CAD models, we apply random rotations with respect to their center of mass to generate a larger augmented dataset and use a BEM solver to calculate the ASFs.

we use the standard procedure of mesh simplification and mesh clustering algorithm from the *vcglib* [2] to ensure that our meshes have a minimum edge length of $1.7cm$, which is 1/20 of our shortest target wavelength. This is sufficient according to the standard techniques used in BEM simulators (Marburg, 2002). Most meshes after pre-processing have fewer than 20% number of elements than the original and the BEM simulation for dataset generation gains over 10× speedup.

BEM Solver We use the *FastBEM Acoustics* software [3] as our wave-based solver. Simulations are run on a Windows 10 workstation that has 32 Intel(R) Xeon(R) Gold 5218 CPU cores with multi-threading. First we use the adaptive cross approximation (ACA) BEM (Kurz et al., 2002) to compute the ASF since it can achieve near $\mathcal{O}(N)$ computational performance for small to medium sized models (e.g., element count $N \leq 100,000$). If it fails to converge within some fixed number of iterations, we use

[2] http://vcg.isti.cnr.it/vcglib/
[3] https://www.fastbem.com/

the conventional and accurate BEM solver. Overall, it takes about 12 days to compute the ASF up to $1000Hz$ frequency of about 100,000 objects from the *ABC Dataset*. The sound pressure field is evaluated at 642 field points that are evenly distributed on the spherical field surface. Next, we use *pyshtools* [4] software (Wieczorek and Meschede, 2018) to compute the spherical harmonics coefficients from the pressure field using least squares inversion.

Reference Field Distance Since the inverse-distance law has increasing error in the near-field of objects, we need to find a suitable distance for computing our reference field. We experimentally simulate the sound pressure fall-off with respect to distance and observe that sound pressure that is $5m$ or further away from the scatterer closely agrees with this far-field approximation (see Figure 4.3). Therefore, we choose to calculate the pressure field on an offset surface $5m$ away from the scatterer's center using a BEM solver (i.e., setting $r_{ref} = 5m$ in Equation 4.7). Note that this choice of $5m$ is not strict or fixed. If higher accuracy along the radial line is desired, multiple locations (especially in the near field) can be sampled during the simulation to interpolate the curve at a higher accuracy. The precomputation time and memory overhead will increase linearly with respect to the number of sampled distance fields.

Max Spherical Harmonics Order We experiment with the number of SH coefficients by projecting our scattered sound pressure fields to SH functions with different orders, as shown in Figure 4.6. Based on this analysis, we choose to use up to a 3rd order SH projection, which yields sufficiently small fitting errors (relative error smaller than 2%) with 16 SH coefficients. This sets the output of our neural network (Section 4.2.3) to be a vector of length 16.

[4] https://shtools.oca.eu/shtools/public/index.html

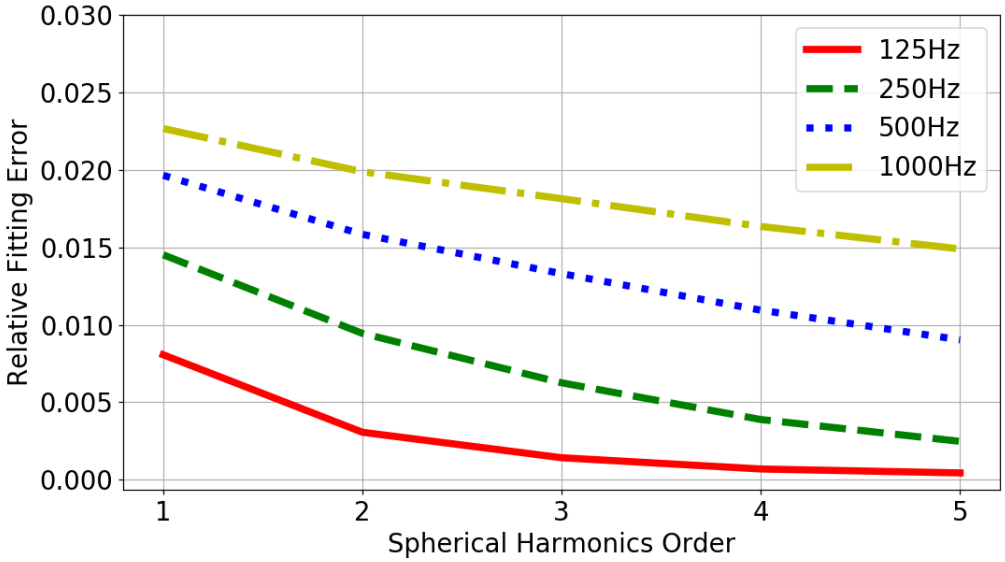


Figure 4.6: **Spherical harmonics approximation of sound pressure fields:** We evaluate different orders of SH functions to fit our pressure fields at 4 frequencies and calculate the relative fitting errors.

4.6.2 Network Training

Our network model is trained on a GeForce RTX 2080 Ti GPU using the *Tensorflow* framework (Abadi et al., 2016). The dataset is split into training set and test set using the ratio 9 : 1. In the training stage, we use Adam optimizer to minimize L_2 norm loss between predicted spherical harmonic coefficients and the groundtruth. In practice, the initial learning rate is set to 1×10^{-3}, which decays exponentially at a rate of 0.9 and clips at 1×10^{-5}. The batch size is set to 128 and typically our network converges after 100 epochs in 8 hours. The number of our trainable parameters is about $800k$.

4.6.3 Runtime System and Benchmarks

We use the geometric sound propagation and rendering algorithm described in (Schissler et al., 2014). Our sound rendering system traces sound rays at octave frequency bands at $125Hz$, $250Hz$, $500Hz$, $1000Hz$, $2000Hz$, $4000Hz$, and $8000Hz$. The direct output from ray tracing for each frequency band is the energy histogram with respect

to propagation delays. We take square root of these responses to compute the frequency dependent pressure response envelopes. Broadband frequency responses are interpolated from our traced frequency bands, and the inverse Fourier transform is used to re-construct the broadband impulse response. In theory, it is possible to encode phase information within a spherical harmonics representation. However, prior auralization research (Kuttruff, 1993) suggests that using a random phase spectrum along with the energy response does not introduce noticeable sound difference during auralization. Therefore, our method does not preserve phase information to keep the system light-weight.

We require that the wall boundaries are explicitly marked in our scenes. As a result, when a ray hits the wall, only conventional sound reflections occur for all frequencies. During audio-visual rendering, when a ray hits a scattering object, we first extend the hit point along its ray direction by $0.5m$ and use it as the scattering region center. We include all the points within a search radius of $1m$ from the region center to generate a point cloud approximation of the scatterer. This point cloud is resampled using furthest point sampling and fed into our neural networks. Our network predicts the ASFs for sound frequencies corresponding to $125Hz$, $250Hz$, $500Hz$ and $1000Hz$. The higher frequencies (i.e., $2000Hz$, $4000Hz$, and $8000Hz$) are handled by conventional geometric ray-tracing with specular and diffuse reflections and it does not use ASFs. Our neural network has small prediction overhead of less than $1ms$ per view on an NVIDIA GeForce RTX 2080 Ti GPU. The interactive runtime propagation system is illustrated in Figure 4.2. Our ray-tracer performs 200 orders of reflections to generate late reverberation.

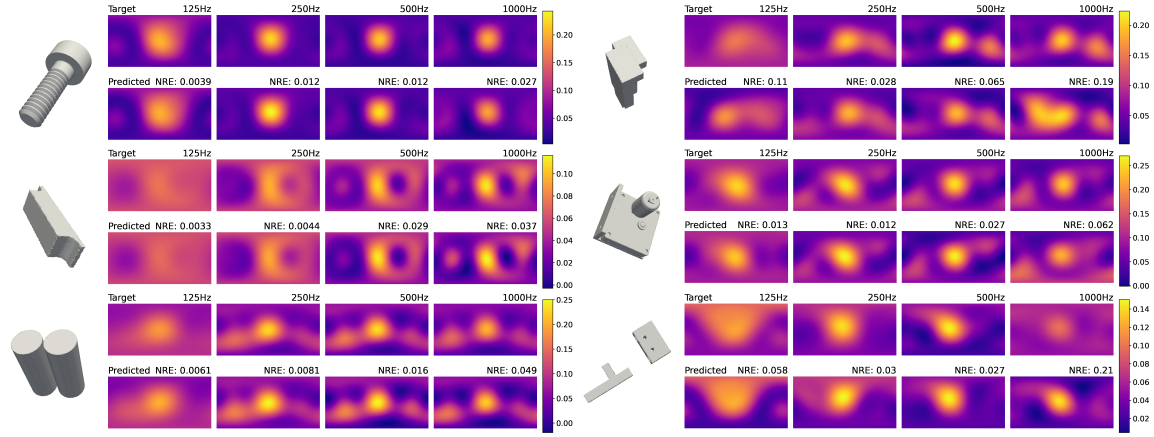
We evaluate the performance of our hybrid sound propagation and rendering algorithms several benchmark scenes shown in Figure 4.1 and Table 4.6.3. They have with varying levels of dynamism in terms of moving objects and are demonstrated in

Scene	Benchmark Description	#Triangle	Frame time
Floor	One static sound scatterer and one static sound source above an infinitely large floor. The listener moves horizontally so that the sound source visibility changes periodically. This is the simplest case where no sound reverberation occurs so as to accentuate the effect of sound diffraction.	4065	$10.65ms$
Sibenik	Two disjoint moving objects are used as scatterers in a church. The two scatterers revolve around each other in close proximity such that there are complicated near-field interactions of sound waves. This scene is a reverberant benchmark.	122798	$6.87ms$
Trinity	Six objects fly across a large indoor room and dynamically generate new composite scatterers or decompose into separate scatterers (i.e., changing topologies). As a result, the total number of separate scattering entities in the scene change and prior methods (Rungta et al., 2018) are not effective. The occluded regions also change dynamically and create challenging scenarios for sound propagation.	386007	$12.95ms$
Havana	Two rotating walls that are generally larger than scatterers in previous benchmarks in a half-open space. We use this benchmark to show that our approach can also handle large static objects, in addition to a large number of dynamic objects. It is an outdoor scene with moderate reverberation.	54383	$6.78ms$

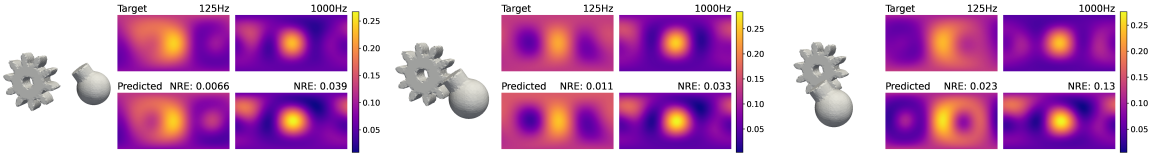
Table 4.1: Runtime performance on our benchmarks. The computation of ASFs takes $\leq 1ms$ per view and most frame time is spent in ray tracing.

our supplemental video[5].

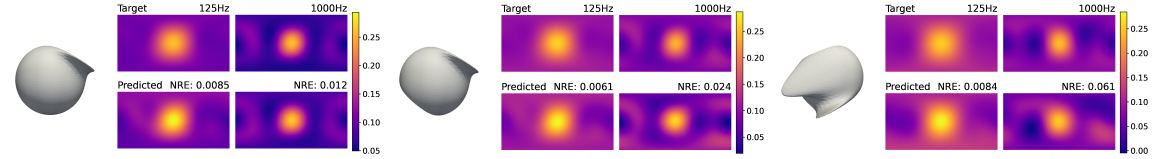
4.6.4 Analysis



(a) ASF of static objects from the unseen test set.



(b) ASF of dynamically moving objects (lowest and highest frequencies). We recompute the ASF at each time instance using our network.



(c) ASFs of a deforming object (lowest and highest frequencies), computed using our network.

Figure 4.7: **Comparing ASF prediction accuracy in latitude-longitude plots:** We highlight the ASFs for different simulation frequencies. For each image block, the left column shows the mesh rendering of the objects. The Lat-Long plots visualize the ASF used in Equation (4.9) by frequency using perceptually uniform colormaps: the top row (*Target*) is the groundtruth ASF computed using a BEM solver on the original mesh; the bottom row (*Predicted*) represents the ASF computed using our neural network based on point-cloud representation. The error metric NRE from Equation (4.13) is annotated above predicted ASFs.

Accuracy Evaluation Our goal is to approximate the acoustic scattering fields of general 3D objects. While there is a preliminary 2D scattering dataset (Fan et al.,

[5] https://gamma.umd.edu/pro/sound/asf

2020a), there are no general or well-known datasets or benchmarks for evaluating such ASFs or related computations. Therefore, we use $10k$ objects from our test dataset to evaluate the performance of our trained network in terms of accuracy. Compared with the original *ABC Dataset*, our test dataset has been augmented in terms of scale and using different orientations to evaluate the performance of our learning method. Since the prediction $p(\theta, \phi, \omega) \in [0, 1]$ from our network is used as the BSDF in Equation (4.9), by fixing ω and varying θ and ϕ, we visualize the field using latitude-longitude plots in Figure 4.7. We use the common normalized reproduction error (NRE) (Lilis et al., 2010; Betlehem and Abhayapala, 2005) to measure the error level of our predicted fields, which is defined as:

$$E(\omega) = \frac{\int_0^{2\pi}\int_0^{\pi} |p^{target}(\theta, \phi, \omega) - p^{predict}(\theta, \phi, \omega)|^2 \mathrm{d}\phi\mathrm{d}\theta}{\int_0^{2\pi}\int_0^{\pi} |p^{target}(\theta, \phi, \omega)|^2 \mathrm{d}\phi\mathrm{d}\theta}. \tag{4.13}$$

We analyze three types of results. **1) Static Objects:** Figure 4.7a shows a subset of CAD objects sampled from our test set, which is from the same distribution as the training set. The average NREs over the entire test set are $4.2\%, 7.6\%, 8.5\%, 10\%$ for $125Hz, 250Hz, 500Hz$, and $1000Hz$ respectively, with an overall NRE of 8.8%. In addition, we show the NRE distribution in Figure 4.8, where we see most test errors are contained below the average NRE. We observe a close visual match in most objects across frequencies. **2) Dynamic Objects:** Figure 4.7b shows an example where two disjoint objects moves in proximity. Such scenarios are not created for the training set. We show the compraison and NREs at the lowest and highest frequencies. **3) Deforming objects:** Figure 4.7c shows an example where a sphere deforms in different parts.

These examples show that our network is able to perform consistently well on a large unseen test set when they are similar to the CAD models in training. Preliminary results on dynamic objects and deforming objects indicate that our network has

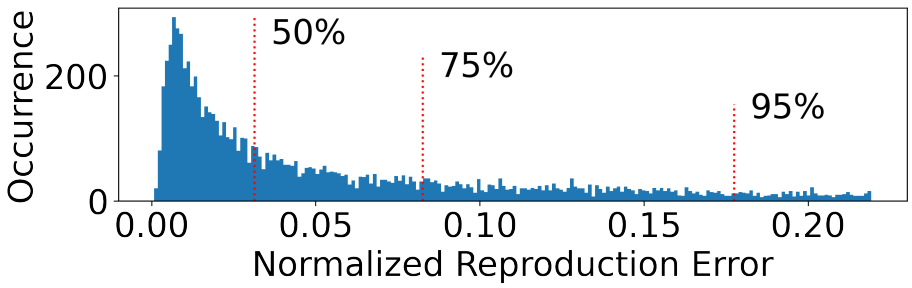


Figure 4.8: **Distribution of test set prediction errors:** We also mark the 50%, 75% and 95% percentiles in the error histogram.

the potential to generalize to more complicated scenarios that are not explicitly modeled during training, although we cannot provide the error bound on these cases. Note that the ASFs are not directly the perceived sound field at specific listener positions - instead they are intermediate transfer functions as one part in the sound rendering pipeline. Therefore, we further demonstrate the perceptual benefits of our predicted ASFs in §4.7 and show that we can reliably generate plausible sound renderings under this error level.

Frequency Growth In theory, our learning-based framework and runtime system can also incorporate wave frequencies beyond $1000Hz$. However, two important factors need to be considered when extending our setup: 1) the wave simulation time increases with the simulation frequency (e.g., between a square and cubic function for an accurate BEM solver); and 2) the ASF becomes more complicated at higher frequencies, which makes it more difficult to be learned or approximated using the same neural network. The per-object simulation time in our experiment is $0.87s, 1.10s, 2.04s, 2.80s$ for $125Hz, 250Hz, 500Hz$, and $1000Hz$, respectively. Note that the simulation time is governed much by the choice of the wave solver, as well as the relevant parameters/strategies used. We pre-processed our meshes according to the highest simulation frequency and used that mesh representation for all frequencies. When a higher frequency needs to be added, the meshes need to have finer

details, meaning more boundary elements will be involved. A frequency-adaptive mesh simplification strategy (Li et al., 2015) can be used to reduce the simulation time at low frequencies. Our network prediction error also grows with the target frequency, but not at a prohibitive rate.

4.7 Perceptual Evaluation

We perceptually evaluate our method using audio-visual listening tests. Our goal is to verify that our method generates plausible sound renderings and identify conditions it may or may not work well. We evaluate three pipelines: 1) Using predicted ASFs and our diffraction handling (ours); 2) Using predicted ASFs and the scattering sound rendering pipeline in diffraction kernels (DK) (Rungta et al., 2018); and 3) Using geometric sound propagation only (GSound) (Schissler and Manocha, 2017). The reason for choosing the two alternatives is that GSound is the state-of-the-art for interactive sound propagation without diffraction modeling. DK is regarded as state of the art hybrid algorithm for interactive sound propagation in dynamic scenes with rigid objects and uses ASFs precomputed by a BEM solver. Since wave-based methods are limited to static scenes, they are not included in our evaluation.

4.7.1 Participants

We performed our studies using Amazon Mechanical Turk[6] (AMT), a popular online crowdsourcing platform that can help data collection. We recruited 71 participants on AMT to take our study. To ensure the quality of our evaluation, we pre-screened our participants for this study. The pre-screening question is designed to test whether the participant has the proper listening device and is in a comfortable listening environment, so that they can tell basic qualitative differences between audios. Specifically,

[6] https://www.mturk.com/

we convolved three impulse responses of reverberation times 0.2s, 0.6s and 1.0s with a 5-second long clean human speech recording to generate three corresponding reverberant speech. The commonly used just-noticeable-difference (JND) of reverberation times is a 5% relative change (ISO, 2009), so in normal conditions we expect a listener to correctly rank our three audios by their reverberation levels. Each participant is asked to listen to the three audios with no time limit, and rank them by their reverberance levels. The initial presentation order of the three audios is randomized for each participant. After pre-screening, our participants consist of 35 males and 16 females, with an average age of 35.9 and a standard deviation of 9.5 years.

4.7.2 Training

As expected, general listeners have varied levels of understanding for sound effects, and we try to diminish this variance to some extent through a quick introduction of sound diffraction. During the training, we provide educational materials about sound diffraction including texts in non-academic language and a short YouTube video showing this phenomenon in the real world (where the sound travels around a pillar while the sound source is invisible). These materials require about one minute to read and watch.

In addition, our participants become familiar with the video playing interface and are asked to adjust their audio playing volume to a comfortable level before the main listening tasks.

4.7.3 Stimuli and Procedure

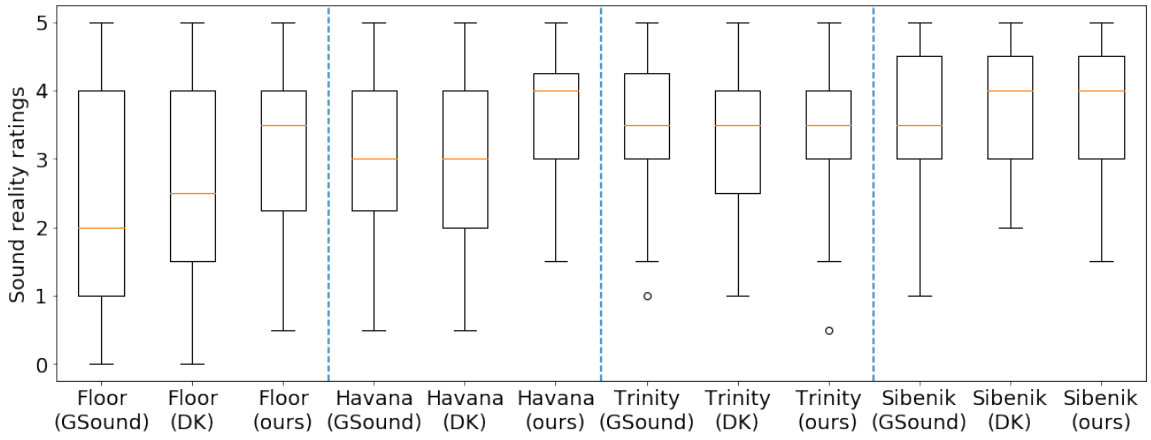
We use the four scenes from benchmarks in §4.6.3 in combination with the three sound rendering pipelines to populate 12 audio-visual renderings that we ask our participants to give ratings on, with no time limit. We present the videos in four pages one after another, each containing only three videos from the same scene (e.g.,

Floor (ours), Floor (DK), and *Floor (GSound)*). The presentation order of the pages, as well as the order of videos within each page, are randomized for each participant. Immediately after each video, participants are asked to give a sound reality rating and a sound smoothness rating. Both ratings range from 0 to 5 stars, with a half-star granularity. Participants are asked to *"give 5 stars for the most realistic and most smooth video and 0 star for the least realistic and smooth"*. Although we believe the perceptual sensitivity can vary among individuals, we expect that participants will be able to recognize cases where unnatural abrupt sound changes occur in response to scene dynamics, and will penalize them in their ratings.

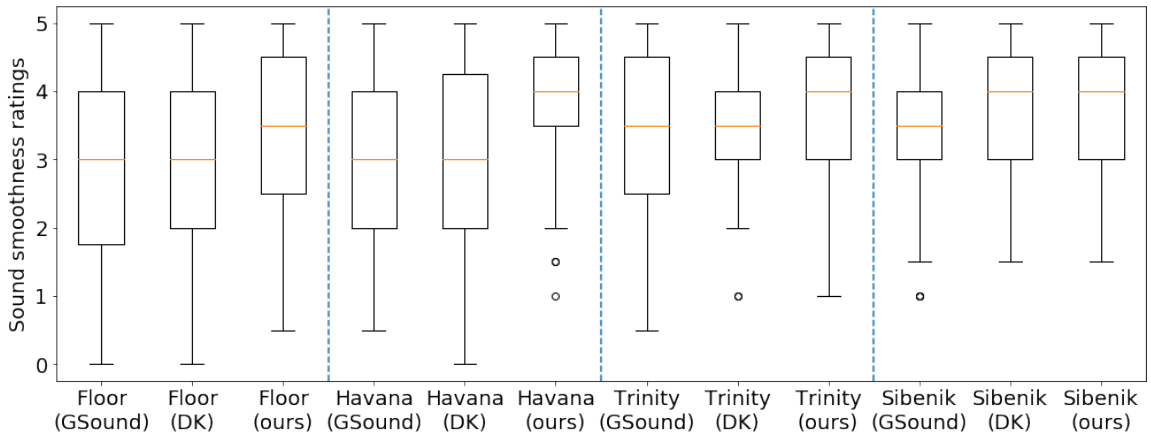
4.7.4 Results

The average study completion time is 13 minutes. We show the box plots of user ratings in Figure 4.9. We are interested in user's rating differences under the 3 test conditions (i.e., *GSound, DK,* and *ours*) on a per scene basis. Therefore, we perform within-group statistical analysis to identify potential significant differences. A significance level of 0.05 is adopted for all results in our discussions.

Sound Reality Ratings First we conduct a non-parametric Friedman test to the ratings given to the 3 rendering conditions, and find significant group differences in *Floor* ($\chi^2 = 10.82, p < 0.01$) and *Havana* ($\chi^2 = 8.27, p = 0.02$), but not in *Trinity* ($\chi^2 = 0.16, p = 0.92$) or *Sibenik* ($\chi^2 = 3.70, p = 0.16$). Note that *Floor* and *Havana* are basically open space scenes with less reverberation, whereas *Trinity* and *Sibenik* are common indoor environments that have a lot of reverberation. Considering that the sound power of reverberation is usually more dominant than diffraction, this result indicates that it is harder to tell the perceptual difference between these rendering pipelines when there is a strong reverberation. To identify the source of differences in *Floor* and *Havana* scenes, we perform post-hoc non-parametric Wilcoxon signed-rank



(a) Sound reality ratings by scene.



(b) Sound smoothness ratings by scene.

Figure 4.9: **Perceptual evaluation results:** User ratings are visualized as box plots. A higher rating means better quality. Results are grouped by benchmark scene and each box represents the rating of a specific rendering pipeline in that scene.

tests with Bonferroni correction (Holm, 1979). We observe that *ours* receives higher ratings than *DK* and *GSound* in both *Floor* ($Z = \{215.0, 144.0\}, p < 0.01$) and *Havana* ($Z = \{254.0, 186.5\}, p < 0.01$). However, there are no significant differences between *GSound* and *DK* in any scene.

Sound Smoothness Ratings Following the same procedure, we perform a Friedman test to the smoothness ratings, and discover that there are significant group differences in *Floor* ($\chi^2 = 10.29, p < 0.01$), *Havana* ($\chi^2 = 7.63, p = 0.02$), and *Sibenik* ($\chi^2 = 12.59, p < 0.01$). Post-hoc Wilcoxon tests show consistent results with real-

ity ratings - we are only able to see a higher smoothness rating of *ours* compared with both *DK* and *GSound* in *Floor* ($Z = \{203.5, 186.0\}, p = 0.01$) and *Havana* ($Z = \{233.5, 127.5\}, p < 0.01$). In *Sibenik*, both *ours* and *DK* receive a higher rating than *GSound* ($Z = \{146.5, 171.0\}, p = 0.01$).

In conclusion, our pipeline receives better perceptual ratings than the other two methods in moderately reverberant conditions, which may not hold in highly reverberant scenes. We have increased perceptual differentiation over the *DK* method. This is due to our better computation of the ASF for dynamic objects which *DK* cannot handle well and our diffraction handling that aligns better with wave acoustic observations.

4.8 Summary

We present a new learning-based approach to approximate the ASFs of objects for interactive sound propagation. We exploit properties of the acoustic scattering field and use a geometric learning algorithm based on point-based approximation. We evaluate the accuracy of our learning method on a large number of objects not seen in the training dataset, also undergoing topology changes. We observe low relative error in our benchmarks. Furthermore, we combine with a ray-tracing based engine for sound rendering in highly dynamic scenes. A perceptual study confirms that our approach generates smooth and realistic sound effects in dynamic environments with increased perceptual differentiation over prior interactive methods.

Our approach has several limitations. These include all the challenges of geometric deep learning in terms of choosing an appropriate training dataset and long training time. It is very hard to provide any rigorous guarantees in terms of error bounds on arbitrary objects. Furthermore, we assume that objects in the scene are sound-hard and do not take into account various material properties. There is a linear

scaling of training time with the number of frequencies and the number of scattering objects, while the simulation time could scale as a cubic function of the frequency. One mitigation is to limit the training to the kind of objects that are frequently used in an interactive application (i.e., customized training).

There are many avenues for future work. It would be useful to take into account the material properties by considering them as an additional object characteristic during training. We would also like to use other techniques from geometric deep learning to improve the performance of our approach. Our runtime ray tracing algorithm could use a different sampling scheme that exploits the properties of ASF. In-person user study using a VR headset or standardized lab listening tests may add more insights to how spatial sound perception is affected by different sound propagation schemes.

Chapter 5

High-Quality Synthetic Acoustic Datasets[1]

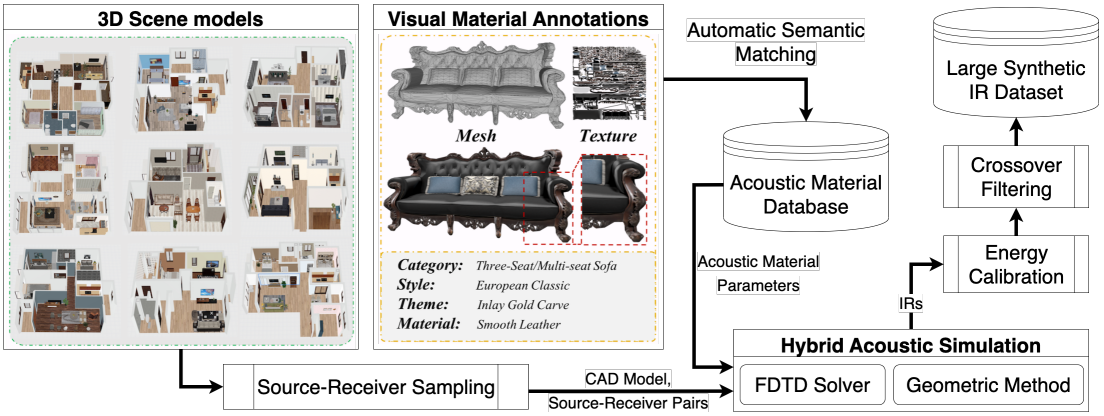


Figure 5.1: Our IR data generation pipeline starts from a 3D model of a complex scene and its visual material annotations (unstructured texts). We sample multiple collision-free source and receiver locations in the scene. We use a novel scheme to automatically assign acoustic material parameters by semantic matching from a large acoustic database. Our hybrid acoustic simulator generates accurate impulse responses (IRs), which become part of the large synthetic impulse response dataset after post-processing.

[1]The work in this chapter has been published in Tang et al. (2019b, 2020, 2022)

5.1 Introduction

Many audio processing tasks have seen rapid progress in recent years due to advances in deep learning and the accumulation of large-scale audio or speech datasets. Not only are these techniques widely used for speech processing, but also acoustic scene understanding and reconstruction, generating plausible sound effects for interactive applications, audio synthesis for videos, etc. A key factor in the advancement of these methods is development and release of audio-related datasets. There are many datasets for speech processing, including datasets with different settings and languages (Park and Mulc, 2019), emotional speech (Tits et al., 2019), speech source separation (Drude et al., 2019), sound source localization (Wu et al., 2018), noise suppression (Reddy et al., 2020), background noise (Reddy et al., 2019), music generation (Briot et al., 2017), etc.

In this chapter, we present a large, novel dataset corresponding to synthetic room impulse responses (IRs). As introduced in § 2.1.1, an IR is regarded as the *acoustical signature* of a system and contains information related to reverberant decay, signal-to-noise ratio, arrival time, energy of direct and indirect sound, or other data related to acoustic scene analysis. These IRs can be convolved with anechoic sound to generate artificial reverberation, which is widely used in the music, gaming and VR applications, as enumerated in § 2.4.

There are some known datasets of recorded IRs from real-world scenes and synthetic IRs (see Table 5.1). The real-world datasets are limited in terms of number of IRs or the size and characteristics of the captured scenes. All prior synthetic IR datasets are generated using geometric simulators and do not accurately capture low-frequency wave effects. This limits their applications.

Main Results: We present a large, accurate acoustic dataset (GWA) of synthetically generated IRs. Our approach is based on using a hybrid simulator that combines a wave-solver based on finite differences time domain (FDTD) method with geometric

Dataset	Type	#IRs	#Scenes	Scene Descriptions	Scene Types	Acoustic Material	Quality
BIU (Hadad et al., 2014)	Rec.	234	3	Photos	Acoustic lab	Real-world	LF, HF
MeshRIR (Koyama et al., 2021)	Rec.	4.4K	2	Room dimensions	Acoustic lab	Real-world	LF, HF
BUT Reverb (Szöke et al., 2019)	Rec.	1.3K	8	Photos	Various sized rooms	Real-world	LF, HF
S3A (Coleman et al., 2020)	Rec.	1.6K	5	Room dimensions	Various sized rooms	Real-world	LF, HF
dEchorate (Di Carlo et al., 2021)	Rec.	2K	11	Room dimensions	Acoustic lab	Real-world	LF, HF
Ko et al. (2017)	Syn.	60K	600	Room dimensions	Empty shoebox rooms	Uniform sampling	HF
BIRD (Grondin et al., 2020)	Syn.	100K	100K	Room dimensions	Empty shoebox rooms	Uniform sampling	HF
SoundSpaces (Chen et al., 2020)	Syn.	16M	101	Annotated 3D model	Scanned indoor scenes	Material database	HF
GWA (ours)	Syn.	2M	18.9K	Annotated 3D model	Professionally designed	Material database	LF, HF

Table 5.1: Overview of some existing large IR datasets and their characteristics. In the "Type" column, "Rec." means *recorded* and "Syn." means *synthetic*. The real-world datasets capture the low-frequency (LF) and high-frequency (HF) wave effects in the recorded IRs. Note that all prior synthetic datasets use geometric simulation methods and are accurate for higher frequencies only. In contrast, we use an accurate hybrid geometric-wave simulator on more diverse input data, corresponding to professionally designed 3D interior models with furniture, and generate accurate IRs corresponding to the entire human aural range (LF and HF). We highlight the benefits of our high-quality dataset for different audio and speech applications.

sound propagation based on path tracing. The resulting IRs are accurate over the human aural range. Moreover, we use a large database of more than 6.8K professional designed scenes with more than 18K rooms with furniture that provide a diverse set of geometric models. We present a novel and automatic scheme for semantic acoustic material assignment based on natural language processing techniques. We use a database of absorption coefficients of $2,042$ unique real-world materials and use a transformer network for sentence embedding. Currently, GWA consists of about 2 million IRs. We can easily use our approach to generate more IRs by either changing the source and receiver positions or using different set of geometric models or materials. The novel components of our work include:

- Our dataset has more acoustic environments than real-world IR datasets by two orders of magnitude.
- Our dataset has more diverse IRs with higher accuracy, as compared to prior synthetic IR datasets.
- The accuracy improvement of our hybrid method over prior methods is evaluated by comparing our IRs with recorded IRs of multiple real-world scenes.

- We use our dataset to improve the performance of deep-learning speech processing algorithms, including automatic speech recognition, speech enhancement, and source separation, and observe significant improvement in accuracy.

5.2 Data Augmentation Preliminary

In this section, we explain the process of audio data augmentation, with an emphasis on speech data, and their use for deep learning tasks. Deep learning theory indicates that having more training examples that have the same data distribution as the test data is crucial to reduce the generalization error of trained models in real test cases (Seltzer et al., 2013). However, the majority of popular speech corpuses were recorded under relatively ideal conditions, i.e. anechoic speech with negligible noise and environmental reverberation. When training models for real-world applications, it is common to distort the clean speech by adding noise and reverberation as a pre-processing step to augment the training data (Kim et al., 2017; Doulaty et al., 2017). In general, speech processing tasks use IR dataset to augment anechoice speech data to create synthetic distant data as the training data, whereas the test data is reverberant data recorded in the real world. In practice, both recorded IRs and synthetic IRs have been used to convolve with the clean speech. Significant improvements in model accuracy have been observed due to this type of data augmentation. When high-quality IR datasets are used, the training set is expected to generalize better on the test data.

Specifically, we can generate distant speech data $x_d[t]$ by convolving anechoic speech $x_c[t]$ with different IRs $h[t]$ and adding environmental noise $n[t]$ (e.g., from noise datasets like BUT ReverbDB (Szöke et al., 2019)) using

$$x_d[t] = x_c[t] \circledast h[t] + n[t]. \tag{5.1}$$

This process is the most common way of reverberant speech data augmentation.

The image method is the current most widely used method in the speech community for generating IRs for speech augmentation (Ko et al., 2017). It is based on the principle of specular reflections where all reflection paths can be constructed by mirroring sound sources with respect to the reflecting plane. We hypothesize that more accurate acoustic simulations (i.e., not only considering specular reflections) can benefit downstream tasks that are trained using the simulated IRs. To verify this, we run various speech processing benchmarks to test a diffuse geometric acoustic simulator we developed (Tang et al., 2020) and compare with an image method simulator. Specifically, we test our geometric simulation with diffuse components against the conventional image method on the automated speech recognition (ASR) task (Table 5.2), the key-word spotting (KWS) task (Table 5.3), as well as the direction-of-arrival (DOA) estimation task (Table 5.4). In all tests, our method has consistently achieved the best performance.

Table 5.2: Character accuracy of ASR systems. Our method has the highest accuracy and outperforms IM by 1.58%.

Model	%
Image Method (IM)	59.96
Our Geometric Simulator	**61.54**

Table 5.3: Equal error rates of KWS systems. Our method has the lowest equal error rate and results in a 21% error reduction relative to that of IM.

Model	%
Image Method (IM)	1.48
Our Geometric Simulator	**1.17**

Table 5.4: Results on the SOFA (Pérez-López and De Muynke, 2018) dataset. First three columns show the percentage of DOA labels correctly predicted within error tolerances, followed by average angular errors, and %-improvement on baseline. Best performance in each column is highlighted in **bold**.

Model	$< 5^\circ$	$< 10^\circ$	$< 15^\circ$	Error	Improv.
Image Method	11.9%	35.9%	73.2%	16.9°	-
Ours	**24.4%**	**66.3%**	**88.2%**	**9.68°**	**43%**

In addition, we are aware that the geometric simulation has the drawback of inaccurate low-frequency modeling due to diffraction and room modes (see § 2.2). This

motivates us to develop a larger dataset with the highest quality synthetic IRs, which model all acoustic phenomena including specular and diffuse reflections, occlusion, diffraction, and low-frequency wave effects.

5.3 Dataset Creation

A key issue in terms of the design and release of an acoustic dataset is the choice of underlying 3D geometric models. Given the availability of interactive geometric acoustic simulation software packages, it is relatively simple to randomly sample a set of simple virtual shoebox-shaped rooms for source and listener positions and generate unlimited simulated IR data. However, the underlying issue is such IR data will not have the acoustic variety (e.g., room equalization, material diversity, wave effects, reverberation patterns, etc.) frequently observed in real-world datasets. We identify several criteria that are important in terms of creating a useful synthetic acoustic dataset: (1) a wide range of room configurations: the room space should include regular and irregular shapes as well as furniture placed in reasonable ways. Many prior datasets are limited to rectangular, shoebox or empty rooms (see Table 1); (2) meaningful acoustic materials: object surfaces should use physically plausible acoustic materials with varying absorption and scattering coefficients, rather than randomly assigned frequency-dependent coefficients; (3) an accurate simulation method that accounts for various acoustic effects, including specular and diffuse reflections, occlusion, and low-frequency wave effects like diffraction. It is important to generate IRs corresponding to the human aural range for many speech processing and related applications. In this section, we present our pipeline for developing a dataset that satisfies all these criteria. An overview of our pipeline is illustrated in Figure 5.1.

5.3.1 Acoustic Environment Acquisition

Acoustic simulation for 3D models requires that environment boundaries and object shapes be well defined and represented as 3D meshes. Simple image-method simulations may only require a few room dimensions (i.e., length, width, and height) and have been used for speech applications, but these methods cannot handle complex 3D indoor scenes. Many techniques have been proposed in computer vision to reconstruct large-scale 3D environments using RGB-D input (Choi et al., 2015). Moreover, they can be combined with 3D semantic segmentation (Dai et al., 2018) to recover category labels of objects in the scene. This facilitates the collection of indoor scene datasets. However, real-world 3D scans tend to suffer from measurement noise, resulting in incomplete/discontinuous surfaces in the reconstructed model that can be problematic for acoustic simulation algorithms. One alternative is to use professionally designed scenes of indoor scenes in the form of CAD models. These models are desirable for acoustic simulation because they have well-defined geometries and the most accurate semantic labels. Therefore, we use CAD models from the 3D-FRONT dataset (Fu et al., 2021), which contains 18,968 diversely furnished rooms in 6,813 irregularly shaped houses/scenes. These different types of rooms (e.g., bedrooms, living rooms, dining rooms, and study rooms) are diversely furnished with varying numbers of furniture objects in meaningful locations. This differs from prior methods that use empty shoebox-shaped rooms (Grondin et al., 2020; Ko et al., 2017), because room shapes and the existence of furniture will significantly modify the acoustic signature of the room, including shifting the room modes in the low frequency. 3D-FRONT dataset is designed to have realistic scene layouts, and has received higher human ratings in subjective studies. Generating audio data from these models allows us to better approximate real-world acoustics.

5.3.2 Semantic Acoustic Material Assignment

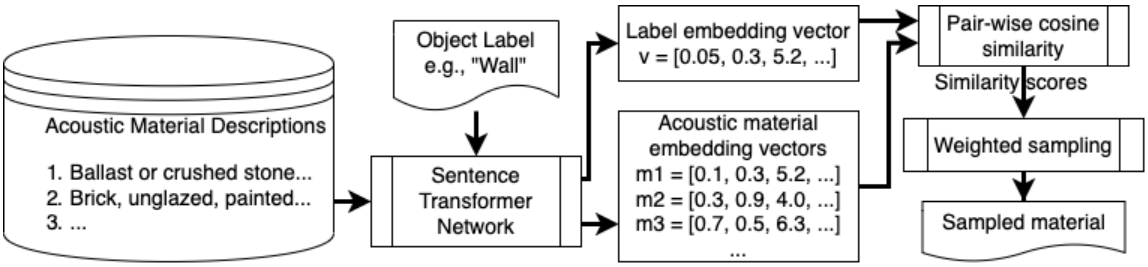


Figure 5.2: Our semantic material assignment algorithm. We use NLP techniques based on sentence embedding along with transformer network to choose absorption coefficients from a database of 2,042 unique materials.

Because the 3D-FRONT dataset also provides object semantics (i.e., object material labels), it is possible to assign more meaningful acoustic materials to individual surfaces or objects in the scene. For example, an object with "window" description is likely to be matched with several types of window glass material from the acoustic material database. *SoundSpaces* dataset (Chen et al., 2020) also utilizes scene labels by using empirical manual material assignment (e.g., acoustic materials of carpet, gypsum board, and acoustic tile are assumed for floor, wall, and ceiling classes), creating a one-to-one visual-acoustic material mapping for the entire dataset. This approach works for a small set of known material types. Instead, we present a general and fully automatic method that works for unknown materials with unstructured text descriptions.

To start with, we retrieve measured frequency-dependent acoustic absorption coefficients for 2,042 unique materials from a room acoustic database (Kling, 2018). The descriptions of these materials do not directly match the semantic labels of objects in the 3D-FRONT dataset. Therefore, we present a method to calculate the semantic similarity between each label and material description using natural language processing (NLP) techniques. In NLP research, sentences can be encoded into definite length numeric vectors known as sentence embedding (Mishra and Viradiya, 2019). One goal of sentence embedding is to find semantic similarities to

identify text with similar meanings. Transformer networks have been very successful in generating good sentence embeddings (Liu et al., 2020) such that sentences with similar meanings will be relatively close in the embedding vector space. We leverage a state-of-the-art sentence transformer model [2] (Reimers and Gurevych, 2019) that calculates an embedding of dimension 512 for each sentence. Next, we calculate the cosine similarity score between each pair of embedding vectors, which can represent the pair-wise semantic distance between each material label and each description in the material database. For each material label in a 3D scene, we assign a set of absorption coefficients from the acoustic database using weighted sampling based on the cosine similarity scores between the 3D-FRONT material label and all descriptions from the material database. This process is illustrated in Figure 5.2. Note that we do not directly pick the material with the highest score because for the same type of material, there are still different versions with different absorption coefficients (e.g., in terms of thickness, brand, painting, etc.). These slightly different descriptions of the same material are likely to have similar semantic distance to the 3D-FRONT material label being examined. We use a probabilistic assignment process that provides balanced sampling among the material database and thereby increase the diversity of our acoustic database.

5.3.3 Geometric-Wave Hybrid Simulation

It is well known that geometric acoustic (GA) methods do not model low-frequency acoustic effects well due to the linear ray assumption (Funkhouser et al., 1998a; Schissler et al., 2014). Therefore, we use a hybrid propagation algorithm that combines wave-based methods with GA. These wave-based methods can accurately model low-frequency wave effects, but their running time increases as the third or fourth power of the highest simulation frequency (Raghuvanshi et al., 2009). Given the

[2] Using pre-trained model at https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

high time complexity of wave-based methods, we also want to use methods that are: (1) highly parallelizable so that dataset creation takes acceptable time on high-performance computing clusters; (2) compatible with arbitrary geometric mesh representations and acoustic material inputs; and (3) open-source so that the simulation pipeline can be reused by the research community. In this chapter, we develop our hybrid simulation pipeline from a CPU-based GA implementation *pygsound* [3] and a GPU-based wave FDTD implementation *PFFDTD* (Hamilton, 2021).

Inputs

The scene CAD models from the 3D-FRONT dataset, each corresponding to several rooms with open doors, are represented in a triangle mesh format. Most GA methods have native support for 3D mesh input. The meshes are converted to voxels to be used as geometry input to the wave-based solver. We randomly sample 1 source and 50 receiver locations in each scene. We perform collision checking to ensure all sampled locations have at least $0.2m$ clearance to any object in the scene.

We assign acoustic absorption coefficients according to the scheme presented in § 5.3.2. These coefficients can be directly used by the GA method and integrated with the passive boundary impedance model used by the wave FDTD method (Bilbao et al., 2015). The GA method also requires scattering coefficients, which account for the energy ratio between specular and diffuse reflections. Such data is less conventionally measured and is not available from the material database in § 5.3.2. It is known that scattering coefficients tend to be negligible (e.g., ≤ 0.05) for low-frequency bands (Cox et al., 2006) handled by the wave method. Therefore, we sample scattering coefficients by fitting a normal distribution to 37 sets of frequency-dependent scattering coefficients obtained from the benchmark data in § 5.4.1, which are only used by the GA method.

[3] https://github.com/roy james/pygsound

Setup

For the GA method, we set $20,000$ rays and 200 maximum depth for specular and diffuse reflections. The GA simulation is intended for human aural range, while most absorption coefficient data is only valid for octave bands from 63Hz to 8,000Hz. The ray-tracing stops when the maximum depth is reached or the energy is below the hearing threshold.

For the wave-based FDTD method, we set the maximum simulation frequency to 1,400Hz. The grid spacing is set according to 10.5 points per wavelength. Our simulation duration is one second since indoor scenes are usually not too large.

Automatic Calibration

Before combining simulated IRs from two methods, one important step is to properly calibrate their relative energies. Southern et al. (2011) describe two objective calibration methods: (1) pre-defining a crossover frequency range near the highest frequency of the wave method and aligning the sound level of the two methods in that range; (2) calibrating the peak level from time-windowed, bandwidth-matched regions in the wave and the GA methods. Both calibration methods are used case-by-case for each pair of IRs. However, the first method is not physically correct, and the second method can be vulnerable when the direct sound is not known, as with occluded direct rays in the GA method. Southern et al. (2013) improved the second method by calculating calibration parameters once in free-field condition using a band-limited source signal.

We use a similar calibration procedure. The calibration source and receivers have a fixed distance $r = 1$ in a large volume with absorbing boundary conditions, and the 90 calibration receivers span a 90° arc to account for the influence of propagation direction along FDTD grids. The source impulse signal is low-pass filtered at a cut-off frequency of 255Hz. When the source signal is a unit impulse, this filtering makes

the source signal essentially the same as the coefficients of the low-pass filter. The simulated band-limited IRs are truncated at twice the theoretical direct response time to further prevent any unwanted reflected wave. The calibration parameter for wave-based FDTD is computed as:

$$\eta_w = \sqrt{\frac{E_s}{E_r}}, \tag{5.2}$$

where E_s is the total energy of the band-limited point source, and E_r is the total energy at the receiver point. For multiple receiver points, η_w takes the average value. During wave-based FDTD calibration, each received signal is multiplied by η_w, and we can calculate the difference between the calibrated signal and the band-limited source signal. As a result, we obtain a very low mean error of $0.50dB$ and a max error of $0.85dB$ among all calibration receivers.

For the GA method, we follow the same procedure though the process is simpler since the direct sound energy is explicit in most GA algorithms (i.e., $\frac{1}{r}$ scaled by some constant). Another calibration parameter η_g is similarly obtained for the GA method. This calibration process ensures that the full-band transmitted energy from both methods will be $E = 1$ at a distance of $1m$ from a sound source, although the absolute energy does not matter and the two parameters can be combined into one (i.e., only use $\eta'_w = \eta_w/\eta_g$ for wave calibration). Figure 5.3 shows an example of simulation results with and without calibration. Without properly calibrating the energies, there will be abrupt sound level changes in the frequency domain, which can create unnatural sound.

Hybrid Combination

Ideally we would want to use the wave-based method for the highest possible simulation frequency. Besides the running time, one issue with FDTD scheme is the rising

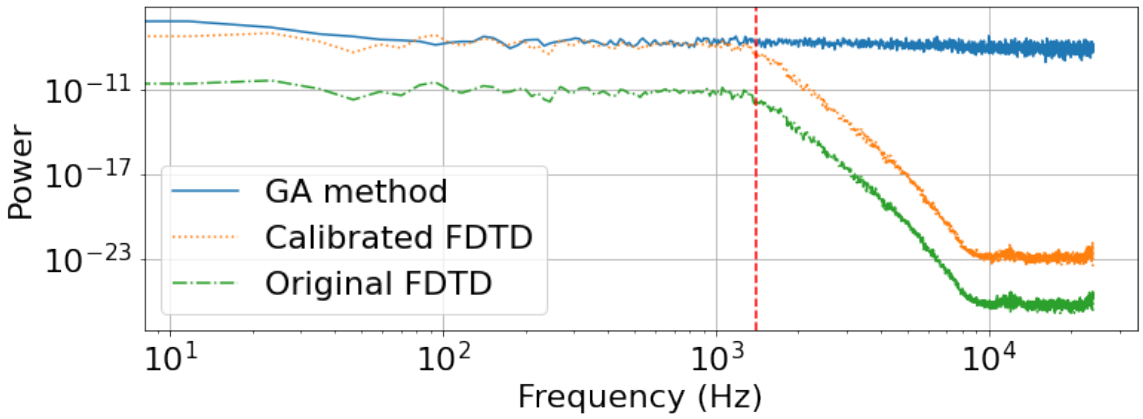
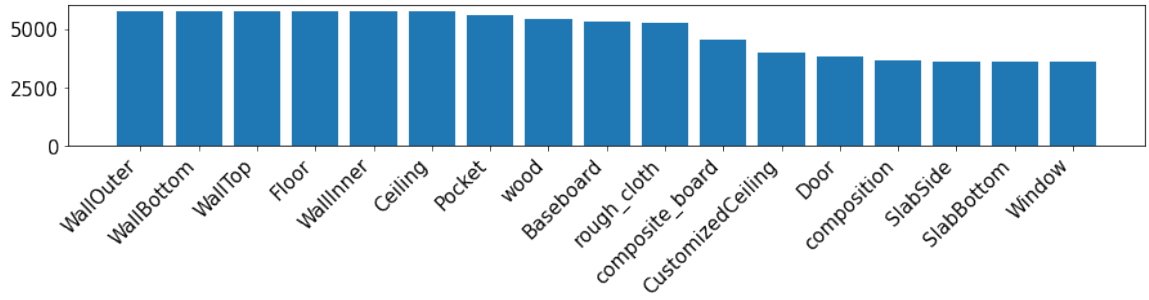


Figure 5.3: Power spectrum comparison between the original wave FDTD simulated IR and the calibrated IR. The vertical dashed line indicates the highest valid frequency of the FDTD method. Our automatic calibration method ensures that the GA and wave-based methods have consistent energy levels so that they can generate high quality IRs and plausible/smooth sound effects.

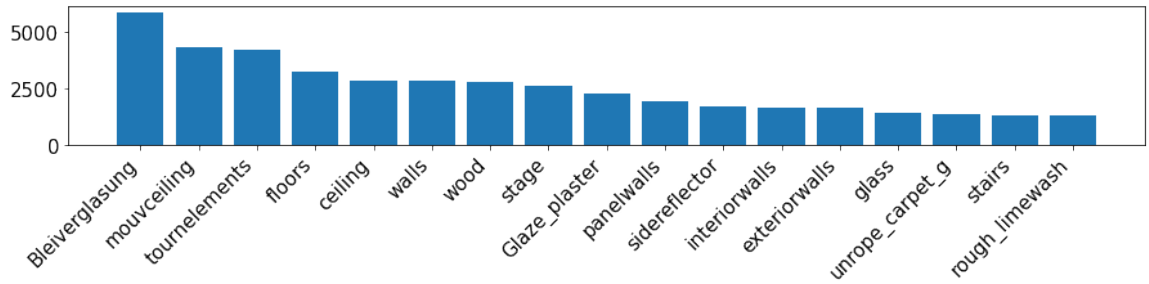
dispersion error with the frequency (Lehtinen, 2003). As a remedy, the FDTD results are first high-pass filtered at a very low frequency (e.g., 10Hz) to remove some DC offset and then low-pass filtered at the crossover frequency to be combined with GA results. We use a Linkwitz-Riley crossover filter (Linkwitz, 1976) to avoid ringing artifacts near the crossover frequency, harnessing its use of cascading Butterworth filters. The crossover frequency in this work is chosen to be $1,400$Hz to fully utilize the accuracy ofwave simulation results. Higher simulation crossover frequencies could be used at the cost of increased FDTD simulation time.

5.3.4 Analysis and Statistics

Runtime The runtime of our hybrid simulator depends on specific computational hardware. We utilize a high-performance computing cluster with 20 Intel Ivy Bridge E5-2680v2 CPUs and 2 Nvidia Tesla K20m GPUs on each node. On a single node, our simulator requires about 800 computing hours for the wave-based FDTD method and about 500 computing hours for the GA method to generate all data. One can roughly estimate the wall time needed by dividing the time above by the number of



(a) Occurrence of top visual material names.



(b) Occurrence of top acoustic material names.

Figure 5.4: We highlight the most frequently used materials in our approach for generating the IR dataset. The acoustic database also contains non-English words, which are handled by a pre-trained multi-lingual language model.

such available compute nodes.

Distributions More than 5,000 scene/house models are used. On average, each scene uses 22.5 different acoustic materials. We assign $1,955$ unique acoustic materials (out of $2,042$) from the material database, and the most frequently used materials are several versions of brick, concrete, glass, wood, and plaster. The occurrence of most frequently used materials are visualized in Figure 5.4.

The distribution of distances between all source and receiver pairs are visualized in Figure 5.5. We also show the relationship between the volume of each 3D house model and the reverberation time for that model in Figure 5.6 to highlight the wide distribution of our dataset. Overall, we have a balanced distribution of the reverberation times in the normal range.

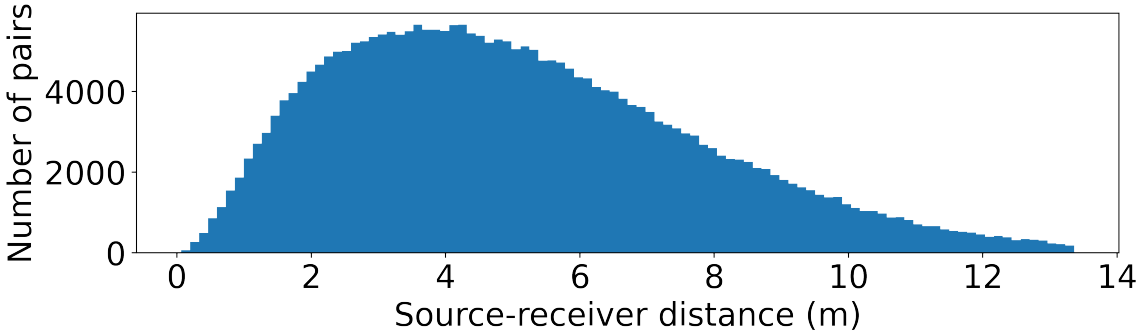


Figure 5.5: Distance distribution between source and receiver pairs in our scene database. No special distance constraints are enforced during sampling except the need to be collision-free from the objects in the scene. The IRs vary based on relative positions of the source and the received in a 3D scene.

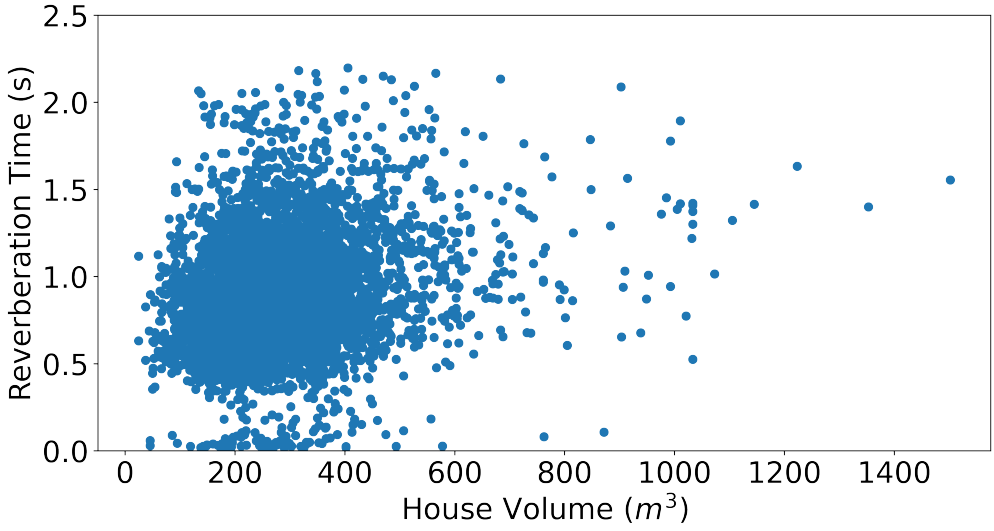


Figure 5.6: Statistics of house/scene volumes and reverberation times. We see a large variation in reverberation times, which is important for speech processing and other applications.

5.4 Acoustic Evaluation

In this section, we evaluate the accuracy of our IR generation hybrid algorithm. We use a set of real-world acoustic scenes that have measured IR data to evaluate the effectiveness and accuracy of our hybrid simulation method.

5.4.1 Benchmarks

Several real-world benchmarks have been proposed to investigate the accuracy of acoustic simulation techniques. A series of three round-robin studies (Vorlander, 1995; Bork, 2000, 2005a,b) have been conducted on several acoustic simulation software systems by providing the same input and then comparing the different simulation results with the measured data. In general, these studies provide the room and material descriptions as well as microphone and loudspeaker specifications including locations and directivity. However, the level of detailed characteristics, in terms of complete 3D models and consistent measured acoustic material properties tend to vary. Previous round-robin studies have identified many issues (e.g., uncertainty in boundary condition definitions) in terms of simulation input definitions for many simulation packages, which can result in poor agreement between simulation results and real-world measurements. A more recent benchmark, the BRAS benchmark (Aspöck et al., 2020), contains the most complete scene description and has a wide range of recording scenarios. We use the BRAS benchmark to evaluate our simulation method. Three reference scenes (RS5-7) are designed as diffraction benchmarks and we use them to evaluate the performance of our hybrid simulator, especially at lower frequencies.

The 3D models of the reference scenes along with frequency-dependent acoustic absorption and scattering coefficients are directly used for our hybrid simulator. We use these three scenes because they are considered difficult for the geometric method

alone (Brinkmann et al., 2019).

5.4.2 Results

We use the room geometry, source-listener locations, and material definitions as an input to our simulation pipeline. Note that the benchmark only provide absorption and scattering coefficients, and no impedance data is directly available for wave solvers. Thus, we only use fitted values rather than exact values. The IRs generated by the GA method and our hybrid method and the measured IRs from the benchmark are compared in the frequency domain in Figure 5.7. In these scenes, the source and receiver are placed on different sides of the obstacle and the semi-anechoic room only has floor reflections. In the high frequency range, there are fewer variations in the measured response, and both methods capture the general trend of energy decay despite response levels not being perfectly matched. This demonstrates that our hybrid sound simulation pipeline is able to generate more accurate results than the GA method for complex real-world scenes.

5.5 Applications

We use our dataset on three speech processing applications that use deep learning methods. Synthetic IRs have been widely used for training neural networks for automatic speech recognition, speech enhancement, and source separation. We evaluate the benefits of generating a diverse and high-quality IRs dataset over prior methods used to generate synthetic IRs.

Far-field speech data is generated according to Equation (5.1) using synthetic IRs. In following test, we use various versions of IR datasets: **GA** (geometric method only), **FDTD** (only up to 1,400Hz), and **GWA** (hybrid method). Then the speech data is used by different training procedure and neural network architectures on different

Table 5.5: Far-field ASR results obtained for the AMI corpus. The best result is marked in **bold**.

IR used	WER[%]↓
None (anechoic speech)	64.2
GA	55.5
GWA (ours)	**54.1**

benchmarks described below.

5.5.1 Automated Speech Recognition

Automatic speech recognition (ASR) aims to convert speech data to text transcriptions. The performance of ASR models is measured by the word error rate (WER), which is the percentage of incorrectly transcribed words in the test data. The AMI speech corpus (Carletta et al., 2005) consisting of 100 hours of meeting recording is used as our benchmark. And we use the *Kaldi* [4] toolbox to run experiments on this benchmark. We randomly select 17,749 IRs out of 2M synthetic IRs in GWA to augment the anechoice training set in AMI, and report the WER on the real-world test set. A lower WER indicates that the synthetic distant speech data used for training is closer to real-world distant speech data. We highlight the improved accuracy obtained using GWA over prior synthetic IR generators in Table 5.5.

5.5.2 Speech Dereverberation

Speech dereverberation aims at converting a reverberant speech signal back to its anechoic version to enhance its intelligibility. We use SkipConvNet (Kothapally et al., 2020), a U-Net based speech dereverberation model. The model is trained on the 100-hour subset of Librispeech dataset (Panayotov et al., 2015). The reverberant input to the model is generated by convolving the clean Librispeech data with our synthetically generated IRs. In addition, we include another synthetic IR dataset,

[4] https://github.com/kaldi-asr/kaldi

Table 5.6: We tabulate the SRMR of the SkipConvNet enhancement model trained using different synthetic IR generation methods. We test the results on real-world reverberant recordings from the VOiCES dataset. Use of our hybrid dataset results in improved accuracy over prior methods.

IR used	**SRMR↑**
None (baseline)	4.96
SoundSpaces (Chen et al., 2020)	7.44
GA	6.01
FDTD	4.78
GWA (ours)	**8.14**

SoundSpaces (Chen et al., 2020) in this comparison. We test the performance of the model on real-world recordings from the VOiCES dataset (Richey et al., 2018). We report the speech-to-reverberation modulation energy ratio (SRMR) over the test set. A higher value of SRMR indicates lower reverberation and higher speech quality. As seen from Table 5.6, our proposed dataset obtains better dereverberation performance as compared to all other datasets.

5.5.3 Speech Separation

We train a model to separate reverberant mixtures of two speech signals into its constituent reverberant sources. We use the Asteroid (Pariente et al., 2020) implementation of the DPRNN-TasNet model (Luo et al., 2020) for our benchmarks. The 100-hour split of the Libri2Mix (Cosentino et al., 2020) dataset is used for training. We test the model on reverberant mixtures generated from the VOiCES dataset. We report the improvement in scale-invariant signal-to-distortion ratio (SI-SDRi) (Roux et al., 2018) to measure separation performance. Higher SI-SDRi implies better separation. As seen from Table. 5.7, our proposed hybrid approach (GWA) outperforms both GA and FDTD for speech separation.

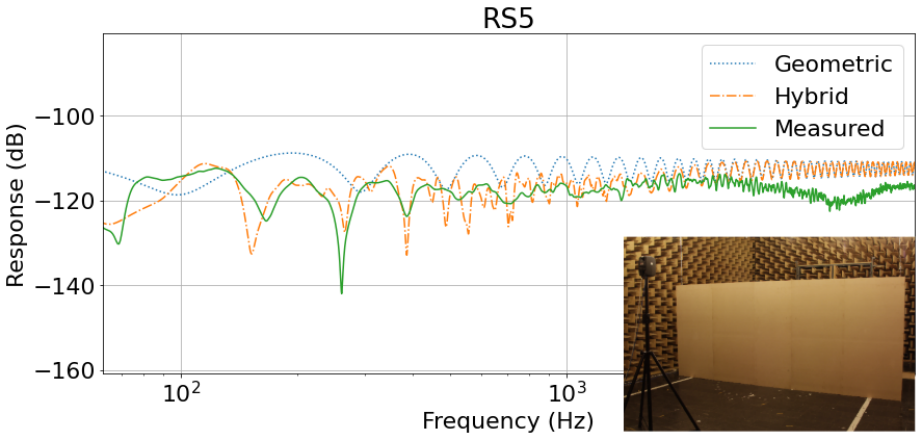
Table 5.7: SI-SDRi values reported for different IR generation methods. We report results separately for the four rooms used to capture the test set (higher is better).

IR used	**SI-SDRi↑**			
	Room 1	**Room 2**	**Room 3**	**Room 4**
GA	2.25	2.55	1.44	2.55
FDTD	2.36	2.43	1.33	2.46
GWA (ours)	**2.94**	**2.76**	**1.86**	**2.91**

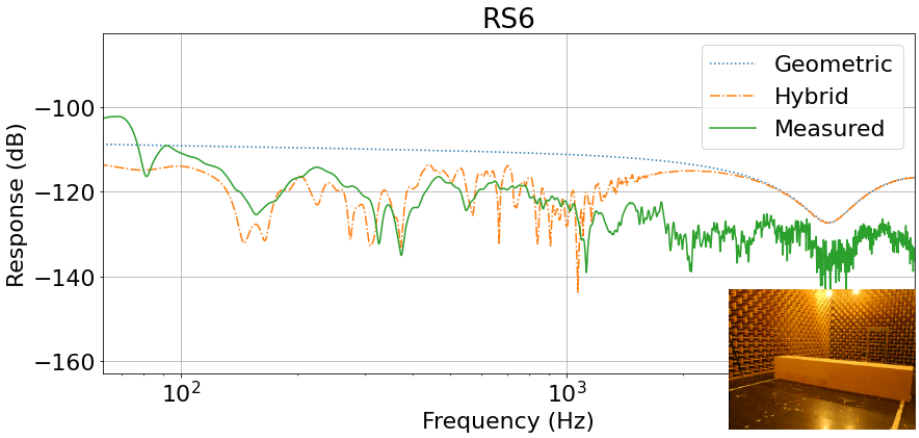
5.6 Summary

We introduced a large new audio dataset of synthetic room impulse responses and the simulation pipeline, which can take different scene configurations and generate higher quality IRs. We demonstrated the improved accuracy of our hybrid geometric-wave simulator on three difficult scenes from the BRAS benchmark. As compared to prior datasets, GWA has more scene diversity than recorded datasets, and has more physically accurate IRs than other synthetic datasets. We also use our dataset with audio deep learning algorithms to improve the performance of speech processing applications.

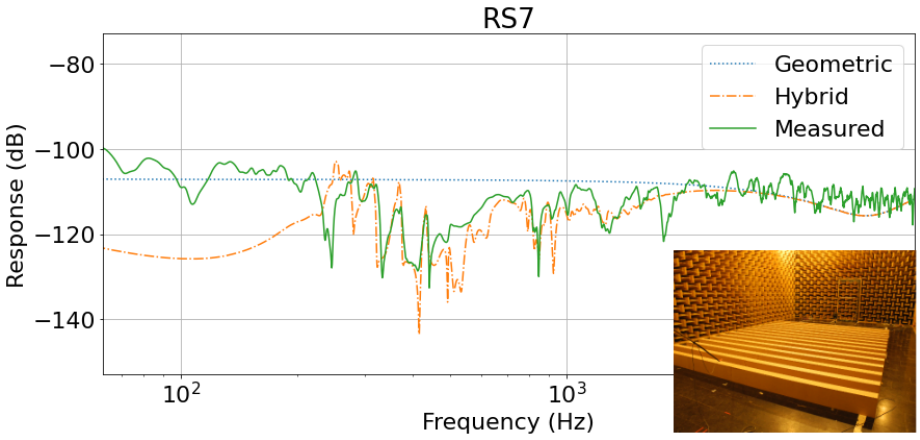
Our dataset only consists of synthetic scenes, and may not be as accurate as real-world captured IRs. In many applications, it is also important to model ambient noise. In the future, we will continue growing the dataset by including more 3D scenes to further expand the acoustic diversity of the dataset. We plan to evaluate the performance of other audio deep learning applications.



(a) RS5: simple diffraction with infinite edge.



(b) RS6: diffraction with infinite body.



(c) RS7: multiple diffraction (seat dip effect)

Figure 5.7: Frequency responses of geometric and hybrid simulations compared with measured IRs in BRAS benchmarks RS5-7 (Aspöck et al., 2020). Images of each setup are attached in the corners of the graph. We notice that the IRs generated using our hybrid method closely match with the measure IRs, as compared to those generated using GA methods. This demonstrates the higher quality and accuracy of our IRs as compared to the ones generated by prior GA methods highlighted in Table 5.1.

Chapter 6

Conclusion

6.1 Summary of Results

In this dissertation, we first investigate novel solutions via acoustic simulation and deep learning to provide high-quality sound rendering in mixed reality settings with fewer limitations than existing vision-based and measurement-based methods. Next, we continue to extend the inferential power of deep neural networks to predict complicated acoustic scattering fields by analyzing object shapes. This becomes the first and the fastest method to generate wave acoustic scattering effects on-the-fly in 3D environments without additional pre-computation for unseen scenes. Finally, we develop a data pipeline that utilizes state-of-the-art geometric and wave acoustic simulators to generate high-quality synthetic impulse response data at scale. Our pipeline can take general 3D model inputs and automatically assign meaningful acoustic materials by semantic matching. The simulation pipeline and dataset can significantly improve the performance of data-driven applications such as deep learning-based speech processing tasks.

Our results have demonstrated that by leveraging state-of-the-art physics-based acoustic simulation and deep learning techniques, realistic simulated data can be

generated to enhance the sound rendering quality in the virtual world and boost the performance of audio processing tasks in the real world.

6.2 Future Work

In the future, I would like to address some limitations mentioned in previous chapters. In the following, I identify several specific future directions.

Just-Noticeable-Difference (JND) in Simulations We have run several perceptual evaluations against other works to verify that the quality of sound rendering from our methods is on par with or better than previous work. However, it is not clear to what extent we want to optimize respective objective functions for acoustic simulations. In other words, how much do factors like accurate material modeling, low-frequency wave simulation, and geometry details affect perceptual listening quality for humans? While some JND metrics have been established for more common acoustic metrics like the T_{60}, less work has been done under the context of acoustic simulations. I believe more rigorous in-lab listening tests with a range of simulation setups will help establish more useful JND metrics for follow-up works.

Curse of Dimensionality Deep learning methods generally suffer from the curse of dimensionality, which means if the dimensionality of the problem being analyzed increases even slightly, the required amount of data will grow exponentially. As a consequence, the time needed to prepare the data and train the model also grows accordingly. This situation applies to deep learning with acoustic problems. As discussed, the soundfield in a room can be affected by the room shape, source and listener positions, acoustic materials/boundary conditions, and medium property (e.g., air temperature). Most of the time we are only able to study a subset of these conditions, as is the case with our deep learning-based acoustic

scattering framework, where we based our analysis entirely on the geometry inputs and ignored the variations in their material properties. While we can expand the training data by adding more dimensions to the simulation setup, techniques like parameter regularization and autoencoders should be considered to mitigate the curse of dimensionality to train a more general acoustic inference model.

Neural Acoustic Fields We have managed to generate a large, high-quality acoustic dataset, and the pipeline allows anyone to expand the dataset to much larger scales if resources permit. This also includes simulating audio data in different hardware (e.g., multi-channel) or software (e.g., spatially encoded) formats. However, there can be infinite amount of data to simulate, and it is unlikely that any one dataset can satisfy all needs. Therefore, one promising direction is to use such a large dataset to learn to construct the acoustic field using neural networks. The same idea has rapidly gained huge success in computer graphics and is known as the neural radiance fields (NeRF) (Mildenhall et al., 2020). While some preliminary work has been done for acoustics (Ratnarajah et al., 2022), dealing with acoustic fields in higher dimensions than radiance fields remains an open and challenging problem.

Bibliography

Steam audio. https://valvesoftware.github.io/steam-audio, 2018.

Microsoft project acoustics. https://aka.ms/acoustics, 2019.

Oculus spatializer. https://developer.oculus.com/downloads/package/oculus-spatializer-unity, 2019.

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

J. S. Abel, N. J. Bryan, P. P. Huang, M. Kolar, and B. V. Pentcheva. Estimating room impulse responses from recorded balloon pops. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

R. Aralikatti, A. Ratnarajah, Z. Tang, and D. Manocha. Improving reverberant speech separation with multi-stage training and curriculum learning. *arXiv preprint arXiv:2107.09177*, 2021.

L. Aspöck, M. Vorländer, F. Brinkmann, D. Ackermann, and S. Weinzierl. Benchmark for room acoustical simulation (bras). *DOI*, 10:14279, 2020.

M. Barron. *Auditorium acoustics and architectural design*. E & FN Spon, 2010.

L. L. Beranek and T. Mellow. *Acoustics: sound fields and transducers*. Academic Press, 2012.

T. Betlehem and T. D. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *The Journal of the Acoustical Society of America*, 117(4): 2100–2111, 2005.

S. Bilbao, B. Hamilton, J. Botts, and L. Savioja. Finite volume time domain room acoustics simulation under general impedance boundary conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):161–173, 2015.

M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison. Codeslam - learning a compact, optimisable representation for dense visual slam. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

I. Bork. A comparison of room simulation software-the 2nd round robin on room acoustical computer simulation. *Acta Acustica united with Acustica*, 86(6):943–956, 2000.

I. Bork. Report on the 3rd round robin on room acoustical computer simulation–part i: Measurements. *Acta Acustica united with Acustica*, 91(4):740–752, 2005a.

I. Bork. Report on the 3rd round robin on room acoustical computer simulation–part ii: Calculations. *Acta Acustica united with Acustica*, 91(4):753–763, 2005b.

D. Botteldooren. Finite-difference time-domain simulation of low-frequency room acoustic problems. *The Journal of the Acoustical Society of America*, 98(6):3302–3308, 1995.

F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl. A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America*, 145(4):2746–2760, 2019.

J. Briot, G. Hadjeres, and F. Pachet. Deep learning techniques for music generation - A survey. *CoRR*, abs/1709.01620, 2017.

N. J. Bryan. Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

N. J. Bryan, J. S. Abel, and M. A. Kolar. Impulse response measurements in the presence of clock drift. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

C. Cao, Z. Ren, C. Schissler, D. Manocha, and K. Zhou. Bidirectional sound transport. *The Journal of the Acoustical Society of America*, 141(5):3454–3454, 2017.

J. Carletta et al. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, page 28–39. Springer-Verlag, 2005. ISBN 3540325492. doi: 10.1007/11677482_3.

M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman. Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 619–623. IEEE, 2016.

S. Cecchi, A. Carini, and S. Spors. Room response equalization—a review. *Applied Sciences*, 8(1):16, 2018.

C. R. A. Chaitanya, J. M. Snyder, K. Godin, D. Nowrouzezahrai, and N. Raghuvanshi. Adaptive sampling for sound propagation. *IEEE transactions on visualization and computer graphics*, 25(5):1846–1854, 2019.

A. Chandak, C. Lauterbach, M. Taylor, Z. Ren, and D. Manocha. Ad-frustum: Adaptive frustum tracing for interactive sound propagation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1707–1722, 2008.

R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.16. URL http://dx.doi.org/10.1109/CVPR.2017.16.

C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020.

L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *The European Conference on Computer Vision (ECCV)*, September 2018.

S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5556–5565, 2015.

F. Chollet et al. Keras. https://keras.io, 2015.

C. L. Christensen and J. H. Rindel. A new scattering method that combines roughness and diffraction effects. In *Forum Acousticum, Budapest, Hungary*, pages 344–352, 2005.

P. Coleman, L. Remaggi, and P. Jackson. S3a room impulse responses, 2020.

A. I. Conference. Audio for virtual and augmented reality. *AES Proceedings*, 2018.

J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent. Librimix: An open-source dataset for generalizable speech separation, 2020.

T. J. Cox, B.-I. Dalenback, P. D'Antonio, J.-J. Embrechts, J. Y. Jeon, E. Mommertz, and M. Vorländer. A tutorial on scattering and diffusion coefficients for room acoustic surfaces. *Acta Acustica united with ACUSTICA*, 92(1):1–15, 2006.

A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018.

P. Debevec. Image-based lighting. *IEEE Computer Graphics and Applications*, 22 (2):26–34, 2002.

D. Di Carlo, P. Tandeitnik, C. Foy, A. Deleforge, N. Bertin, and S. Gannot. dechorate: a calibrated room impulse response database for echo-aware signal processing. *arXiv preprint arXiv:2104.13168*, 2021.

M. Doulaty, R. Rose, and O. Siohan. Automatic optimization of data perturbation distributions for multi-style training in speech recognition. In *Spoken Language Technology Workshop*, 2017.

L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach. Sms-wsj: Database, performance measures, and baseline recipe for multi-channel source separation and recognition. *arXiv preprint arXiv:1910.13934*, 2019.

J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, N. D. Gaubitch, J. Eaton, et al. Estimation of room acoustic parameters: The ace challenge. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(10):1681–1693, 2016.

S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman. Front-end speech enhancement for commercial speaker verification systems. *Speech Communication*, 99:101–113, 2018.

C. Evers, A. H. Moore, and P. A. Naylor. Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10. IEEE, 2016.

Z. Fan, V. Vineet, H. Gamper, and N. Raghuvanshi. Fast acoustic scattering using convolutional neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175. IEEE, 2020a.

Z. Fan, V. Vineet, H. Gamper, and N. Raghuvanshi. Fast acoustic scattering using convolutional neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175. IEEE, 2020b.

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

E. L. Ferguson, S. B. Williams, and C. T. Jin. Sound source localization in a multipath environment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2386–2390. IEEE, 2018.

S. Foster. Impulse response measurement using golay codes. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 929–932. IEEE, 1986.

H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021.

T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 21–32. ACM, 1998a.

T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 21–32. ACM, 1998b.

R. Gao, C. Chen, Z. Al-Halah, C. Schissler, and K. Grauman. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*, pages 658–676. Springer, 2020.

M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev. Blind room volume estimation from single-channel noisy speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 231–235. IEEE, 2019.

S. Gharib, H. Derrar, D. Niizumi, T. Senttula, J. Tommola, T. Heittola, T. Virtanen, and H. Huttunen. Acoustic scene classification: A competition review. In *IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

F. Grondin, J.-S. Lauzon, S. Michaud, M. Ravanelli, and F. Michaud. Bird: Big impulse response dataset. *arXiv preprint arXiv:2010.09930*, 2020.

P. Grumiaux, S. Kitic, L. Girin, and A. Guérin. A survey of sound source localization with deep learning methods. *CoRR*, abs/2109.03465, 2021.

E. Hadad, F. Heese, P. Vary, and S. Gannot. Multichannel audio database in various acoustic environments. In *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317. IEEE, 2014.

C. Hak, R. Wenmaekers, and L. Van Luxemburg. Measuring room impulse responses: Impact of the decay range on derived room acoustic parameters. *Acta Acustica united with Acustica*, 98(6):907–915, 2012.

B. Hamilton. Pffdtd software, 2021. https://github.com/bsxfun/pffdtd.

R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

S. H. Hawley, V. Chatziiannou, and A. Morrison. Synthesis of musical instrument sounds: Physics-based modeling or machine learning. *Phys. Today*, 16:20–28, 2020.

S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

N. Hiremath, V. Kumar, N. Motahari, and D. Shukla. An overview of acoustic impedance measurement techniques and future prospects. *Metrology*, 1(1):17–38, 2021.

Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2017.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

A. ISO. Measurement of room acoustic parameters - part 1. *ISO Std*, 2009.

D. L. James, J. Barbič, and D. K. Pai. Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 987–995. ACM, 2006.

T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman. The cone of silence: speech separation by localization. *arXiv preprint arXiv:2010.06007*, 2020.

S. Ji, J. Luo, and X. Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*, 2020.

X. Jin, S. Li, T. Qu, D. Manocha, and G. Wang. Deep-modal: real-time impact sound synthesis for arbitrary shapes. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1171–1179, 2020.

J. T. Kajiya. The rendering equation. In *ACM SIGGRAPH computer graphics*, volume 20, pages 143–150. ACM, 1986.

M. Karjalainen, P. Antsalo, A. Makivirta, T. Peltonen, and V. Valimaki. Estimation of modal decay parameters from noisy response measurements. In *Audio Engineering Society Convention 110*. Audio Engineering Society, 2001.

C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. In *Interspeech*, 2017.

H. Kim, L. Remaggi, P. Jackson, and A. Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360° images. *Proceedings IEEE VR2019*, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. Kleiner, P. Svensson, and B.-I. Dalenbäck. Auralization: experiments in acoustical cad. In *Audio Engineering Society Convention 89*. Audio Engineering Society, 1990.

C. Kling. Absorption coefficient database, Jul 2018. URL https://www.ptb.de/cms/de/ptb/fachabteilungen/abt1/fb-16/ag-163/absorption-coefficient-database.html.

T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

S. Koch, A. Matveev, Z. Jiang, F. Williams, A. Artemov, E. Burnaev, M. Alexa, D. Zorin, and D. Panozzo. Abc: A big cad model dataset for geometric deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang. Skipconvnet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping. *arXiv preprint arXiv:2007.09131*, 2020.

S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström. Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. *arXiv preprint arXiv:2106.10801*, 2021.

A. Krokstad, S. Strom, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968.

S. Kurz, O. Rain, and S. Rjasanow. The adaptive cross-approximation technique for the 3d boundary-element method. *IEEE transactions on Magnetics*, 38(2):421–424, 2002.

H. Kuttruff. *Room Acoustics*. Taylor & Francis Group, London, U. K., 6th edition, 2016.

K. H. Kuttruff. Auralization of impulse responses modeled on the basis of ray-tracing results. *Journal of the Audio Engineering Society*, 41(11):876–880, 1993.

P. Larsson, D. Vastfjall, and M. Kleiner. Better presence and performance in virtual environments by improved binaural sound rendering. In *Virtual, Synthetic, and Entertainment Audio conference*, Jun 2002. URL http://www.aes.org/e-lib/browse.cfm?elib=11148.

C. LeGendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019.

J. Lehtinen. Time-domain numerical solution of the wave equation. *Feb*, 6:1–17, 2003.

D. Li, Y. Fei, and C. Zheng. Interactive acoustic transfer approximation for modal sound. *ACM Transactions on Graphics (TOG)*, 35(1):1–16, 2015.

D. Li, T. R. Langlois, and C. Zheng. Scene-aware audio for 360° videos. *ACM Trans. Graph.*, 37(4), 2018.

G. N. Lilis, D. Angelosante, and G. B. Giannakis. Sound field reproduction using the lasso. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8): 1902–1912, 2010.

S. H. Linkwitz. Active crossover networks for noncoincident drivers. *Journal of the Audio Engineering Society*, 24(1):2–8, 1976.

Q. Liu, M. J. Kusner, and P. Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

S. Liu and D. Manocha. Sound synthesis, propagation, and rendering: A survey. *arXiv preprint arXiv:2011.05538*, 2020.

Y. Luo, Z. Chen, and T. Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation, 2020.

M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457, 2021.

S. Marburg. Six boundary elements per wavelength: Is that enough? *Journal of computational acoustics*, 10(01):25–51, 2002.

R. Mehra, N. Raghuvanshi, L. Antani, A. Chandak, S. Curtis, and D. Manocha. Wave-based sound propagation in large open scenes using an equivalent source formulation. *ACM Transactions on Graphics (TOG)*, 32(2):19, 2013.

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

H.-Y. Meng, Z. Tang, and D. Manocha. Point-based acoustic scattering for interactive sound propagation via surface encoding. *arXiv preprint arXiv:2105.08177*, 2021.

B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

M. K. Mishra and J. Viradiya. Survey of sentence embedding methods. *International Journal of Applied Science and Computations*, 6(3):592–592, 2019.

N. Morales and D. Manocha. Efficient wave-based acoustic material design optimization. *Computer-Aided Design*, 78:83–92, 2016.

N. Morales, R. Mehra, and D. Manocha. A parallel time-domain wave simulator based on rectangular decomposition for distributed memory architectures. *Applied Acoustics*, 97:104–114, 2015.

N. Morales, Z. Tang, and D. Manocha. Receiver placement for speech enhancement using sound propagation optimization. *Applied Acoustics*, 155:53–62, 2019.

G. Mückl and C. Dachsbacher. Precomputing sound scattering for structured surfaces. In *EGPGV@ EuroVis*, pages 73–80, 2014.

G. J. Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010, 2014.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent. Asteroid: the PyTorch-based audio source separation toolkit for researchers. In *Proc. Interspeech*, 2020.

K. Park and T. Mulc. CSS10: A collection of single speaker speech datasets for 10 languages. *CoRR*, abs/1903.11269, 2019. URL http://arxiv.org/abs/1903.11269.

S. Pelzer, L. Aspöck, D. Schröder, and M. Vorländer. Integrating real-time room acoustics simulation into a cad modeling software to enhance the architectural design process. *Buildings*, 4(2):113–138, 2014.

A. Pérez-López and J. De Muynke. Ambisonics directional room impulse response as a new convention of the spatially oriented format for acoustics. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

M. Pharr, W. Jakob, and G. Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016.

A. D. Pierce and R. T. Beyer. Acoustics: An introduction to its physical principles and applications. 1989 edition, 1990.

M. A. Poletti. Three-dimensional surround sound systems based on spherical harmonics. *Journal of the Audio Engineering Society*, 53(11):1004–1025, 2005.

V. Pulkki and U. P. Svensson. Machine-learning-based estimation and rendering of scattering in virtual reality. *The Journal of the Acoustical Society of America*, 145 (4):2664–2676, 2019.

S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192, 2021.

N. Raghuvanshi and J. Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Trans. Graph.*, 33(4):38:1–38:11, July 2014a. ISSN 0730-0301. doi: 10.1145/2601097.2601184. URL http://doi.acm.org/10.1145/2601097.2601184.

N. Raghuvanshi and J. Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):38, 2014b.

N. Raghuvanshi and J. Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):38, 2014c.

N. Raghuvanshi and J. Snyder. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 37(4):108, 2018.

N. Raghuvanshi, R. Narain, and M. C. Lin. Efficient and accurate sound propagation using adaptive rectangular decomposition. *Visualization and Computer Graphics, IEEE Transactions on*, 15(5):789–801, 2009.

N. Raghuvanshi, J. Snyder, R. Mehra, M. Lin, and N. Govindaraju. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM Trans. Graph.*, 29(4):68:1–68:11, July 2010. ISSN 0730-0301. doi: 10.1145/1778765.1778805. URL http://doi.acm.org/10.1145/1778765.1778805.

A. Ratnarajah, Z. Tang, and D. Manocha. IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition. In *Proc. Interspeech 2021*, pages 286–290, 2021. doi: 10.21437/Interspeech.2021-230.

A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu. FAST-RIR: Fast neural diffuse room impulse response generator. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke. A scalable noisy speech dataset and online subjective test framework, 2019.

C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results, 2020.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Z. Ren, H. Yeh, and M. C. Lin. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1, 2013.

C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni. Voices obscured in complex environmental settings (voices) corpus, 2018.

L. Rizzi, G. Ghelfi, and M. Santini. Small-rooms dedicated to music: From room response analysis to acoustic design. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

M. Rosen, K. W. Godin, and N. Raghuvanshi. Interactive Sound Propagation For Dynamic Scenes Using 2d Wave Simulation. *Computer Graphics Forum*, 2020. ISSN 1467-8659. doi: 10.1111/cgf.14099.

J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. SDR - half-baked or well done? *CoRR*, abs/1811.02508, 2018. URL http://arxiv.org/abs/1811.02508.

A. Rungta, S. Rust, N. Morales, R. Klatzky, M. Lin, and D. Manocha. Psychoacoustic characterization of propagation effects in virtual environments. *ACM Transactions on Applied Perception (TAP)*, 13(4):21, 2016.

A. Rungta, C. Schissler, N. Rewkowski, R. Mehra, and D. Manocha. Diffraction kernels for interactive sound propagation in dynamic environments. *IEEE transactions on visualization and computer graphics*, 24(4):1613–1622, 2018.

W. C. Sabine. *Collected papers on acoustics*. 1927.

J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3): 279–283, 2017.

L. Savioja and U. P. Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015. doi: 10.1121/1.4926438.

R. W. Schafer and A. V. Oppenheim. *Discrete-time signal processing*. Prentice Hall Englewood Cliffs, NJ, 1989.

C. Schissler and D. Manocha. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)*, 36(1):2, 2016.

C. Schissler and D. Manocha. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)*, 36(1):2, 2017.

C. Schissler and D. Manocha. Interactive sound rendering on mobile devices using ray-parameterized reverberation filters. *arXiv preprint arXiv:1803.00430*, 2018.

C. Schissler, R. Mehra, and D. Manocha. High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Transactions on Graphics (TOG)*, 33(4):39, 2014.

C. Schissler, C. Loftin, and D. Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, 2017.

M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre. Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra). In *1st Web Audio Conference*, pages 1–6, 2015.

M. R. Schroeder. The "schroeder frequency"revisited. *The Journal of the Acoustical Society of America*, 99(5):3240–3241, 1996.

P. Seetharaman and S. P. Tarzia. The hand clap as an impulse source for measuring room acoustics. In *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

M. L. Seltzer, Y. Dong, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *IEEE International Conference on Acoustics*, 2013.

B. Series. Recommendation ITU-R BS. 1534-3 method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radio Communication Assembly*, 2014.

P. Series. Methods for objective and subjective assessment of speech and video quality. *International Telecommunication Union Radiocommunication Assembly*, 2016.

J. O. Smith III. *Spectral Audio Signal Processing*. 01 2008.

A. Southern, S. Siltanen, and L. Savioja. Spatial room impulse responses with a hybrid modeling method. In *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

A. Southern, S. Siltanen, D. T. Murphy, and L. Savioja. Room impulse response synthesis and validation using a hybrid acoustic model. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1940–1952, 2013.

A. Sterling, J. Wilson, S. Lowe, and M. C. Lin. Isnn: Impact sound neural network for audio-visual object classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 555–572, 2018.

A. Sterling, N. Rewkowski, R. L. Klatzky, and M. C. Lin. Audio-material reconstruction for virtualized reality using a probabilistic damping model. *IEEE transactions on visualization and computer graphics*, 25(5):1855–1864, 2019.

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

U. P. Svensson, R. I. Fred, and J. Vanderkooy. An analytic secondary source model of edge diffraction impulse responses. *The Journal of the Acoustical Society of America*, 106(5):2331–2344, 1999.

I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876, 2019.

Q. Tan, L. Gao, Y.-K. Lai, J. Yang, and S. Xia. Mesh-based autoencoders for localized deformation component analysis. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha. Scene-aware audio rendering via deep acoustic analysis. *arXiv preprint arXiv:1911.06245*, 2019a.

Z. Tang, J. Kanu, K. Hogan, and D. Manocha. Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks. In *Interspeech*, 2019b.

Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha. Improving reverberant speech training using diffuse acoustic simulation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6969–6973. IEEE, 2020.

Z. Tang, H.-Y. Meng, and D. Manocha. Learning acoustic scattering fields for dynamic interactive sound propagation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 835–844. IEEE, 2021.

Z. Tang, R. Aralikatti, A. Ratnarajah, , and D. Manocha. Gwa: A large geometric-wave acoustic dataset for audio deep learning, 2022.

M. Taylor, A. Chandak, Q. Mo, C. Lauterbach, C. Schissler, and D. Manocha. Guided multiview ray tracing for fast auralization. *IEEE Transactions on Visualization and Computer Graphics*, 18:1797–1810, 2012a.

M. Taylor, A. Chandak, Q. Mo, C. Lauterbach, C. Schissler, and D. Manocha. Guided multiview ray tracing for fast auralization. *IEEE Transactions on Visualization and Computer Graphics*, 18(11):1797–1810, 2012b.

M. T. Taylor, A. Chandak, L. Antani, and D. Manocha. Resound: interactive sound rendering for dynamic virtual environments. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 271–280. ACM, 2009.

R. A. Tenenbaum, F. O. Taminaro, and V. Melo. Room acoustics modeling using a hybrid method with fast auralization with artificial neural network techniques. In *Proc. International Congress on Acoustics (ICA)*, pages 6420–6427, 2019.

L. L. Thompson. A review of finite-element methods for time-harmonic acoustics. *The Journal of the Acoustical Society of America*, 119(3):1315–1330, 2006.

N. Tits, K. El Haddad, and T. Dutoit. Emotional speech datasets for english speech synthesis purpose: A review. In *Proceedings of SAI Intelligent Systems Conference*, pages 61–66. Springer, 2019.

I. R. Titze, L. M. Maxfield, and M. C. Walker. A formant range profile for singers. *Journal of Voice*, 31(3):382.e9 – 382.e13, 2017. ISSN 0892-1997. URL http://www.sciencedirect.com/science/article/pii/S0892199716301096.

J. Traer and J. H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

N. Tsingos. Precomputing geometry-based reverberation effects for games. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

N. Tsingos, T. Funkhouser, A. Ngan, and I. Carlbom. Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 545–552. ACM, 2001.

D. Tsokaktsidis, T. Von Wysocki, F. Gauterin, and S. Marburg. Artificial neural network predicts noise transfer as a function of excitation and geometry. In *Proc. International Congress on Acoustics (ICA)*, pages 4392–4396, 2019.

V. Välimäki and J. Reiss. All about audio equalization: Solutions and frontiers. *Applied Sciences*, 6(5):129, 2016.

V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448, 2012.

T. Virtanen, M. D. Plumbley, and D. Ellis. *Computational analysis of sound scenes and events*. Springer, 2018.

M. Vorländer. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America*, 86(1):172–178, 1989.

M. Vorländer. Computer simulations in room acoustics: Concepts and uncertainties. *The Journal of the Acoustical Society of America*, 133(3):1203–1213, 2013.

M. Vorliander. International round robin on room acoustical computer simulations. In *15th Intl. Congress on Acoustics, Trondheim, Norway*, pages 689–692, 1995.

K. Wapenaar. Unified matrix–vector wave equation, reciprocity and representations. *Geophysical Journal International*, 216(1):560–583, 2019.

M. A. Wieczorek and M. Meschede. Shtools: Tools for working with spherical harmonics. *Geochemistry, Geophysics, Geosystems*, 19(8):2574–2592, 2018.

L. C. Wrobel and A. Kassab. Boundary element method, volume 1: Applications in thermo-fluids and acoustics. *Appl. Mech. Rev.*, 56(2):B17–B17, 2003.

T. Wu, Y. Jiang, N. Li, and T. Feng. An indoor sound source localization dataset for machine learning. In *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, pages 28–32, 2018.

H. Yeh, R. Mehra, Z. Ren, L. Antani, D. Manocha, and M. Lin. Wave-ray coupling for interactive sound propagation in large complex scenes. *ACM Transactions on Graphics (TOG)*, 32(6):165, 2013.

C. Zheng and D. L. James. Toward high-quality modal contact sound. In *ACM SIGGRAPH 2011 papers*, pages 1–12. 2011.

X. Zheng, C. Wen, N. Lei, M. Ma, and X. Gu. Surface registration via foliation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

S. Zhi, M. Bloesch, S. Leutenegger, and A. J. Davison. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018.

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, Dec. 1997. ISSN 0098-3500. URL http://doi.acm.org/10.1145/279232.279236.