CHAPTER 1: BACKGROUND

In this chapter, I review work related to the main aspects of this research. I also provide a mathematical background of modal sound synthesis.

1.1 Sound Synthesis

Sound synthesis techniques recreate natural sounds for virtual environments. Sounds are dynamic and can be created by a variety of sound sources. Different types of sound sources produce different types of sounds, so different models are needed. Examples of sound sources that have been modeled are liquids (Langlois et al., 2016; Moss et al., 2010), paper (Schreck et al., 2016), and fire (Chadwick and James, 2011).

In this dissertation, the focus is on sounds created by rigid objects. Strings and drums can be simulated through physical models such as the Karplus-Strong algorithm (Karplus and Strong, 1983) and digital waveguide synthesis (Smith, 1992). Simple objects with known analytical vibration patterns can be simulated through additive synthesis, where individual sine waves are added together to create more complex sounds (van den Doel and Pai, 1996). Arbitrary rigid objects use the same additive synthesis method, but to determine their frequencies of vibration, or *modes of vibration*, discretized models of the objects need to be analyzed first (Morrison and Adrien, 1993; O'Brien et al., 2002). This is referred to as *modal sound synthesis*, consisting of a precomputation step called "modal analysis" and a runtime synthesis step called "modal synthesis". We now review the details of this method and explain the need for accurate damping parameters.

1.1.1 Modal Analysis

When a rigid object is struck, it vibrates in response, though these vibrations may be imperceptible to the eye. As the surface of the object vibrates and deforms, the surrounding air is rapidly compressed and expanded, creating pressure waves which propagate through the environment. Our ears perceive the variation in air pressure as sound. The standard range of human hearing covers sound waves between 20 Hz and 20 kHz. In modal analysis, the shape and material parameters of the object are analyzed to decompose the vibrations into a set of *modes of vibration*. Each mode of vibration describes one independent component of the overall vibration as the object oscillates sinusoidally over time. Each object has a different set of modes depending on the object's shape and material. Vibrations from an impact can roughly be represented as a linear combination of normal modes with different amplitudes, frequencies, and phases.

Modal analysis is often performed numerically, where the object is represented using a discretized model such as a FEM mesh or spring-mass thin-shell system. Regardless of the choice of discretization, we can consider the dynamics of the system as it vibrates using a system of equations:

$$\mathbf{M}\ddot{\mathbf{r}} + \mathbf{C}\dot{\mathbf{r}} + \mathbf{K}\mathbf{r} = \mathbf{f} \tag{1.1}$$

Here, $\mathbf{r}$ is a vector of vertex displacements, where a vector of all zeros represents the object at rest. Since we usually work with three-dimensional objects, an object with n discrete elements would have a $\mathbf{r} \in \mathbb{R}^{3n}$. $\mathbf{f}$ is the vector of forces applied to each element, inducing vibrations. $\mathbf{M}$ is the mass matrix, which describes the distribution of mass throughout the object. $\mathbf{C}$ is the viscous damping matrix, which describes how the velocity of the elements $\dot{\mathbf{r}}$ decays over time. $\mathbf{K}$ is the stiffness matrix, in which the connectivity of the elements is defined. Given these matrices, we can properly simulate the vibration of the object in response to an impulse. $\mathbf{M}$ and $\mathbf{K}$ can be constructed through knowledge of the shape of the object and its material parameters, notably its density, Poisson's ratio, and Young's modulus. The damping matrix $\mathbf{C}$, is not as simple to construct.

Modal analysis examines the eigenvalues and eigenvectors of the system in free vibration, that is, with $f = 0$ after some initial impulse has been applied. Temporarily ignoring damping, we can set up a generalized eigenvalue problem of the form:

$$\mathbf{K}\mathbf{v} = \lambda\mathbf{M}\mathbf{v} \tag{1.2}$$

Finding this eigendecomposition and combining the eigenvectors into a matrix $\mathbf{\Phi}$ allows the matrices $\mathbf{M}$ and $\mathbf{K}$ to be diagonalized. Specifically, the eigenvectors are mass-normalized such that:

$$\mathbf{\Phi}^T\mathbf{M}\mathbf{\Phi} = \mathbf{I} \quad \text{and} \quad \mathbf{\Phi}^T\mathbf{K}\mathbf{\Phi} = \mathbf{\Omega^2} \tag{1.3}$$

The matrix $\mathbf{\Phi}$ can be intuitively described as a matrix that transforms between object space and mode space: each column of $\mathbf{\Phi}$ contains the shape of a normal mode, while $\mathbf{\Phi}^T\mathbf{f}$ converts forces on elements to normal mode amplitudes. The natural undamped frequencies of the system are contained in the diagonal matrix $\mathbf{\Omega}$, while their squares in $\mathbf{\Omega}^2$ are the eigenvalues of the system. We can continue the decoupling by considering the system in mode space $\mathbf{z} = \mathbf{\Phi}^T\mathbf{r}$:

$$\mathbf{\Phi^T M \Phi \ddot{z}} + \mathbf{\Phi^T C \Phi \dot{z}} + \mathbf{\Phi^T K \Phi z} = \mathbf{\Phi^T f} \tag{1.4}$$

$$\mathbf{\ddot{z}} + \mathbf{\Phi^T C \Phi \dot{z}} + \mathbf{\Omega^2 z} = \mathbf{\Phi^T f} \tag{1.5}$$

Equation (1.5) now runs into problems with the damping matrix $\mathbf{C}$. While $\mathbf{\Phi}$ diagonalizes $\mathbf{M}$ and $\mathbf{K}$, if it does not diagonalize $\mathbf{C}$ then the system does not properly decouple and the resulting modes are not linearly independent. The linearly *dependent* modes are called complex modes, and accurately modeling them is much more difficult compared to the linearly *independent* normal modes (Imregun and Ewins, 1995). We must now consider methods for modeling damping behavior and constructing appropriate $\mathbf{C}$ matrices.

1.1.2 Damping Modeling

Damping has long been a concern in analysis of vibrations of buildings and other structures (Nashif et al., 1985; Adhikari and Woodhouse, 2001). There are a number of ways to model material-based damping to varying degrees of accuracy (Woodhouse, 1998; Slater et al., 1993), and standard tests have been designed to consistently measure damping in materials (E756, 2017). Complex models are often required to produce accurate fits to observed damping behavior (Adhikari, 2001).

To construct appropriate $\mathbf{C}$ matrices, for sound synthesis purposes we restrict ourselves to *classical* damping with only normal modes, which means all of our damping matrices must be diagonalizable by $\mathbf{\Phi}$. Various damping models have been developed that guarantee only normal modes (Caughey, 1960). These damping models typically have real-valued parameters that vary between materials. In this dissertation, α_j is used to represent these *damping parameters*. However, be aware that each damping model has a different definition of α_j.

The most popular model is Rayleigh damping (Rayleigh, 1896), in which the damping is a linear combination of mass and stiffness:

$$\mathbf{C} = \alpha_1\mathbf{M} + \alpha_2\mathbf{K} \tag{1.6}$$

α_1 and α_2 are the real-valued parameters in this Rayleigh damping model. Rayleigh damping has been, to the best of our knowledge, the only damping model used for sound synthesis in computer graphics.

Caughey and O'Kelly proposed a more general model, now known as *Caughey damping* or a *Caughey series* (Caughey and O'Kelly, 1965), which they proved to be a necessary and sufficient condition for normal modes:

$$\mathbf{C} = \mathbf{M} \sum_{j=0}^{n-1} \alpha_j (\mathbf{M}^{-1}\mathbf{K})^j \tag{1.7}$$

All α_j are real-valued parameters for Caughey damping models. In practice, the series could truncated after a few terms.

For a given damping model, the real-valued parameters α_j are the *damping parameters* which define the damping of each mode. By varying these values, the same object can be made to sound like a wide range of materials. Damping parameters have been shown to be perceptually geometry-invariant for a wide range of geometries under the Rayleigh damping model (Ren et al., 2013a); it is reasonable to assume this holds for other damping models as well. Thus, if damping parameters can be estimated for a metal bowl, synthesizing sound for a solid cube with those parameters will produce a metallic sound. However, the geometry-invariance assumption has only been thoroughly tested on thick, very rigid objects (Ren et al., 2013a), and the assumption may fail for thin-shelled objects (Chadwick et al., 2009), less rigid objects, objects with loosely-coupled points of self-collision, or objects demonstrating nonlinear vibrational behavior.

1.1.3 Modal Synthesis

With these damping models, we have a damping matrix guaranteed to be diagonalizable by $\mathbf{\Phi}$. With the system diagonalized, the free-vibration form is now decoupled into independent second order differential equations:

$$\ddot{z}_i + c_i \dot{z}_i + \omega_{in}^2 z_i = 0 \tag{1.8}$$

c_i is an entry in the diagonalized damping matrix corresponding to the i'th mode of vibration, and is discussed in more detail in Section 1.1.4. These equations each have known analytical solutions as damped sinusoids:

$$z_i(t) = a_i e^{-d_i t} \cos(\omega_{id} t) \tag{1.9}$$

a_i is the amplitude of the sinusoid, while the damping coefficient $d_i = c_i/2$ defines the rate at which the amplitude decreases. ω_{in} in Equation (1.8) is the natural undamped frequency of oscillation, but in the presence of damping we use the *damped* frequency ω_{id}:

$$\omega_{id} = \sqrt{\omega_{in}^2 - d_i^2} \tag{1.10}$$

1.1.4 Obtaining Damping Coefficients

In practice, we do not actually want to perform the matrix operations in the damping models. Through heavy use of Equation (1.3), we can find analytical solutions for how $\mathbf{C}$ is diagonalized and compute c_i in terms of the corresponding eigenvalue ω_{in}^2. The solution for Rayleigh damping is common in modal sound synthesis work:

$$\begin{aligned}
\mathbf{\Phi}^T\mathbf{C}\mathbf{\Phi} &= \mathbf{\Phi}^T\alpha_1\mathbf{M}\mathbf{\Phi} + \mathbf{\Phi}^T\alpha_2\mathbf{K}\mathbf{\Phi} \\
&= \alpha_1\mathbf{I} + \alpha_2\mathbf{\Omega^2} \\
c_i &= \alpha_1 + \alpha_2\omega_{in}^2 \qquad (1.11)
\end{aligned}$$

Caughey damping is slightly more involved, but leads to a fairly intuitive solution:

$$\begin{aligned}
\mathbf{\Phi}^T\mathbf{C}\mathbf{\Phi} &= \mathbf{\Phi}^T\mathbf{M}\sum_{j=0}^{n-1}\alpha_j(\mathbf{M}^{-1}\mathbf{K})^j\mathbf{\Phi} \\
&= \mathbf{\Phi}^{-1}\sum_{j=0}^{n-1}\alpha_j(\mathbf{\Phi}\mathbf{\Omega}^2\mathbf{\Phi}^{-1})^j\mathbf{\Phi} \\
&= \sum_{j=0}^{n-1}\alpha_j\mathbf{\Omega}^{2^j} \\
c_i &= \sum_{j=0}^{n-1}\alpha_j{\omega_{in}^2}^j \qquad (1.12)
\end{aligned}$$

Using these solutions, the damping rates for each mode of vibration can be determined.

1.1.5 Real-Time Synthesis

For real-time synthesis, a preprocessing step is first performed for a given object and material. In this step, the eigendecomposition is performed and the resulting $\mathbf{\Phi^T}$ and each mode's d_i and ω_{id} are saved.

At runtime, an applied force $\mathbf{f}$ is transformed to mode space by $\mathbf{\Phi^T}$, and the resulting vector contains the amplitudes with which to excite each mode. The resulting damped sinusoids can be combined and sampled at 44.1 kHz to produce the sound itself. Tools for performing additive synthesis and modal sound synthesis are plentiful; examples include the Synthesis ToolKit (Cook and Scavone, 1999) and the Faust programming language (Michon et al., 2017; Michon and Smith, 2011).

For interactive applications, as a user performs actions to create sounds, sound synthesis algorithms must run fast enough to generate sound in real time. The computation requirements at runtime are proportional to the complexity of the analyzed input shape, making some objects' sounds too slow for real-time applications without optimizations. Vibration modes can be culled based on psychoacoustic principles, for example, humans cannot tell the difference between two frequencies very close to one another, so those modes can be combined into one (Raghuvanshi and Lin, 2006). If an object has any geometric symmetries, these can be exploited to reduce memory usage and caching requirements (Langlois et al., 2014). Synthesis can be done in frequency space to further improve performance (Bonneel et al., 2008). When performing real-time synthesis, vectorization (van Walstijn and Mehes, 2017) and parallelism on CPUs (Bilbao et al., 2013) and GPGPUs (Webb, 2014) are effective, as each mode of vibration can be synthesized independently.

1.1.6 Additional Factors

Modal sound synthesis roughly simulates the sounds produced by rigid, vibrating objects, but in the real world more factors influence the final sound we hear; four such examples are acoustic radiance, sound propagation effects, contacts with other objects, and acceleration sound.

Acoustic radiance is the efficiency of propagation for each mode: depending on the shape of an object some modes radiate in different directions with different strengths (James et al., 2006; Li et al., 2015). More generally, any sound source can be directional, requiring additional simulation considerations (Mehra et al., 2014). Once the vibrations transfer to the surrounding air, sound waves bounce around the environment before reaching a listener's ears.

Sound propagation refers to this propagation of sound waves through air. Propagation can be simulated most realistically with wave-based simulation (Raghuvanshi et al., 2009), though for use them in interactive applications these methods have heavy precomputation and storage requirements (Mehra et al., 2015, 2013; Raghuvanshi et al., 2016, 2010). Geometric methods for sound propagation are less accurate for low frequencies, but faster to compute for interactive applications (Savioja and Svensson, 2015; Chandak et al., 2008; Schissler and Manocha, 2016, 2011), as long as diffraction can be properly simulated (Tsingos et al., 2001; Svensson et al., 1999; Rungta et al., 2018). Hybrid methods use geometric propagation for higher frequencies and wave-based methods for low frequencies heavily affected by wave effects (Hampel et al., 2008; Southern et al., 2011; Yeh et al., 2013). Some work has achieved tight coupling between sound synthesis and propagation (Rungta et al., 2016; Wang et al., 2018).

Contacts with other objects are common as objects rarely float in midair. These contacts with other objects modify the produced sound and can be accounted for with contact models (O'Brien et al., 2002; Zheng and James, 2011). Interactions between a sounding object and a striking tool can be modeled to better simulate the attack of the sound (Avanzini and Rocchesso, 2001; Bilbao et al., 2015). Contact modeling can be exploited to create real objects that vibrate only at desired frequencies. An object can be placed on foam blocks, specifically positioned to damp out the undesired frequencies while leaving the desired frequencies alone (Bharaj et al., 2015).

Continuous interactions between objects, such as sliding and scraping, require additional effort. Fractal noise is a common way of representing the small impacts generated during rolling and scraping (Doel et al., 2001). Ren et al. presented a framework for synthesizing contact sounds between textured objects (Ren et al., 2010). This work introduced a multi-level model for lasting contact sounds combining fractal noise with impulses collected from the normal maps on the surfaces of the objects. However, this application of normal maps to sound generation without similar application to rigid-body dynamics causes noticeable sensory conflict between the produced audio and visible physical behavior.

Acceleration sound is produced when an object is rapidly accelerated through air, and is perceptually noticeable for very small objects such as dice and keys (Chadwick et al., 2012).

1.1.7 Full Physically-Accurate Simulation

To emphasize the restrictions that must be made for real-time sound synthesis, consider the case of dominoes falling on a table, as seen in **??**. A full and physically-correct simulation would need to consider all of the above factors. Normal modes of vibration are simulated by modal sound synthesis, the perceptually-dominant factor that this dissertation focuses on. Complex modes and acoustic radiance would need to be simulated for a complete model of object vibrations. Acceleration noise may be perceptually noticeable for these small dominoes. Accurate contact modeling for this scene would be important for this scene given the stacking structure of the fallen dominoes. Sound propagation would be necessary not just to model the acoustics of the room, but also to model the interactions between objects.

This theoretical full simulation would require tight coupling between each objects' interior deformations, inter-object forces, and the air pressure/velocity fields. Some of this could be achieved with a wave-based simulator (Wang et al., 2018) and accurate contact model (Zheng and James, 2011). However, simulation of these factors is too computationally-intensive for real-time sound synthesis. Therefore, for real-time applications such as virtual environments, we are limited to modal sound synthesis and approximate sound propagation (James et al., 2006).

1.2 Multimodal Interaction with Virtual Objects

Multimodal interaction, in the context of this dissertation, refers to interaction using multiple senses simultaneously. The senses of sight, hearing, and touch are each different *interaction modalities*, which have been independently researched. In this section, I discuss prior work related to texture mapping and each of these additional modalities of interaction. As one of the main contributions in this dissertation is a method for multimodal interaction with *textured* surfaces (**??**), much of the background in this section focuses on surface interactions.

1.2.1 Human Auditory Perception

Since this work focuses on audio, rendering for the sense of hearing has been discussed earlier inSection 1.1. However, as sound is inherently perceptual, studies of human auditory perception provide important clues about perceptually important parameters. Studies have evaluated which parameters humans rely on for material identification, finding that damping rate and frequency (i.e. pitch) are particularly important (Klatzky et al., 2000; McAdams et al., 2010). Studies have tested the discriminability of materials and the generalizability of the Rayleigh damping model (Ren et al., 2013a). Similar studies have focused on perception of object size from sound (Giordano and McAdams, 2006; Grassi, 2005). Material perception is also affected by concurrent visual stimuli (Fujisaki et al., 2014).

1.2.2 Visual Rendering

Realistic visual rendering has been the focus of the computer graphics field for many decades, and photo-realistic visual appearances are possible given talented artists and sufficient computational resources. Many books provide an introduction to the field (Akenine-Moller et al., 2002; Foley et al., 1990). Creating realistic visual appearances in interactive environments in real time is more challenging, but can be accomplished using optimizations.

For example, *texture mapping* uses low-resolution 3D triangle meshes with higher-resolution 2D textures to model detailed objects. Normal maps and relief maps are used as representations of fine detail of the surface of objects. Normal maps were originally introduced for the purposes of bump mapping, where they would perturb lighting calculations to make the details more visibly noticeable (Blinn, 1978). Relief mapping uses both depths and normals for more complex shading (Oliveira et al., 2000; Policarpo et al., 2005). Numerous other texture mapping techniques exist as well. Displacement mapping, parallax mapping, and a number of more recent techniques use height maps to simulate parallax and occlusion (Cook, 1984; Kaneko et al., 2001; Tevs et al., 2008). A recent survey goes into more detail about many of these techniques (Szirmay-Kalos and Umenhoffer, 2008). Mapping any of these textures to progressive meshes can preserve texture-level detail as the level-of-detail (LOD) of the mesh shifts (Cohen et al., 1998).

1.2.3 Rigid-Body Simulation

Simulation of the movement and collisions between rigid objects allows virtual environments to simulate gravity and user behavior such as stacking and throwing objects (Featherstone, 2007). Height maps mapped to object surfaces have been used to modify the behavior of simple collisions in rigid-body simulations (Nykl et al., 2013). When height maps are applied to two colliding objects, previous methods can effectively compute and resolve their collision (Otaduy et al., 2004).

1.2.4 Haptic Rendering

Haptics refers to interaction using the sense of touch, and focuses on the textures of surfaces (Loomis and Lederman, 1986). There has been significant work on how humans perceive haptic sensations to recognize shapes and textures (Lederman and Taylor, 1972; Klatzky et al., 1985). In haptic rendering, a 3D object's geometries and textures can be felt by applying forces based on point-contacts with the object (Basdogan et al., 1997; Ho et al., 1999). Complex objects can also be simplified, with finer detail placed in a displacement map and referenced to produce accurate force *and torque* feedback on a probing object (Otaduy et al., 2004). The mapping of both normal and displacement maps to simplified geometry for the purposes of haptic feedback has also been explored (Theoktisto et al., 2010). Dynamic deformation textures, a variant of displacement maps, can be mapped to create detailed objects with a rigid center layer and deformable outer layer. The technique has been extended to allow for 6-degree-of-freedom (DOF) haptic interaction with these deformable objects (Galoppo et al., 2007). A common approach to force display of textures is to apply lateral force depending on the gradient of a height map such that the user of the haptic interface feels more resistance when moving "uphill" and less resistance when moving "downhill" (Minsky et al., 1990; Minsky, 1995).

1.2.5 Integrated Multimodal Interaction

For realistic multimodal interaction, it is important that content is not only rendered well for individual senses, but that each sense is consistent with one another. Between audio and rigid-body simulation, modal sound synthesis can be coupled with physics simulations to couple the movements of objects and their resulting sounds (O'Brien et al., 2002; Zheng and James, 2011). Depth maps can modify contacts between objects, coupling the visual appearances of the objects with their physical movements (Nykl et al., 2013). When multiple objects are in contact, the long-lasting contacts produce continuous sounds which depend heavily on the objects' textures, further coupling motion and sound (Ren et al., 2010). Between audio and haptics, some work has considered the multimodal aspects of touch-enabled interfaces for sound synthesis (Ren et al., 2012). These methods involve only one or two interaction modalities each, and do not use a single representation of surface detail to inform all modalities.

1.3 Auditory Understanding

The inverse of the modal sound synthesis problem is to use impact sounds to understand the objects that created those sounds. In this dissertation, I present multiple methods for estimating properties of real-world objects from recorded sounds. In this section, I will review the broad area of processing sounds to learn something about the sound's source. I will begin with discussion of general concepts and methodology, then focus in on object sounds.

1.3.1 Environmental Sound Classification

One broad way of approaching sound understanding is to classify sounds into descriptive categories. Multiple datasets have been established for evaluating classification of various environmental sounds (Gemmeke et al., 2017; Piczak, 2015b; Salamon et al., 2014). Traditional techniques use a variety of features extracted from sounds, such as Mel frequency spectral coefficients and spectral shape descriptors (Büchler et al., 2005; Cowling and Sitte, 2003). Similar approaches are used to classify an environment based on the sounds heard within it (Barchiesi et al., 2015).

Convolutional neural networks have also been applied to these problems, producing improved results (Piczak, 2015a; Salamon and Bello, 2017). Recently, some interest has been given to exploring the performance of different network structures (Hershey et al., 2017; Huzaifah, 2017). Impact sounds are a specific category of environmental sounds, which contain fewer cues to differentiate them from one another.

1.3.2 Statistical Sound Modeling

Statistical methods have found applications in summarizing and analyzing sound. The late reverberations of sounds in rooms have been modeled as Gaussian noise, whose summary statistics convey properties of the environment (Traer and McDermott, 2016). It has also been found that humans inherently use summary statistics to understand sounds (McDermott et al., 2013). Previous methods for material parameter estimation assume minimal variable effects in estimation of damping rates. In comparison, the probabilistic damping model presented in **??** models recorded impact sounds as inherently stochastic.

1.3.3 Object Understanding Through Sound

I now shift the focus to understanding of objects' impact sounds in particular. A common application is to learn properties of a real-world object in order to *resynthesize* similar sounds in a virtual environment. Some methods use a single recorded sound, then apply modifications to create realistic variety in resynthesized sounds. Deterministic features of a sound can be extracted, then stochastic noise can be added to those features to model slight variations (Serra and Smith, 1990). Alternatively, the modal content of a sound can be extracted, then resynthesized, slightly modifying mode amplitudes to create variations (Lloyd et al., 2011). Other methods use multiple input sounds for a single object, generated by striking the object in known locations. The sounds' spectral content can be interpolated spatially to approximate hit points at new locations (Pai et al., 2001).

However, these methods work by modifying recorded sounds without gaining much fundamental knowledge about the object itself. They do not model the object's material or shape independently from one another. Ideally, sounds from struck real-world objects could be used to recreate the shape and material parameters of the objects. In the rest of this section, I will consider both independent material and shape estimation.

1.3.3.1 Auditory Material Estimation

Material parameters can be estimated experimentally with specialized measurement equipment (E756, 2017), but impact sounds do not require specialized equipment or trained personnel to record. The Young's modulus for small parts of the object can be individually optimized to best match input sounds (Yamamoto and Igarashi, 2016). The most relevant work is that of Ren et al. (Ren et al., 2013b), which performs automatic estimation of material parameters from a single audio sample. This method works by optimizing a synthetic sound to most closely match the recorded sound. The material parameters that optimally match the synthetic and recorded sounds are the most likely material parameters for the real-world object. The estimated parameters can be applied to synthesis of sounds for any object with that material. However, their method is able to estimate damping parameters only for the Rayleigh damping model, which may be limited in its ability to represent diverse materials.

These methods are often limited in their robustness by relying on Rayleigh damping, not accounting for environmental factors, and not using multimodal input. These limitations are addressed as part of this dissertation. Methods that estimate material damping parameters support only the Rayleigh damping model, which I address in **??**. All of these methods assume that properties of the recording environment are known or are assumed to be minimal, which I address in **??**. If both video and audio of the object are available, these methods have no way of using the visual information to improve material estimates, which I address in **??**.

1.3.3.2 Auditory Shape Estimation

The ideal case of using one sound to reconstruct an entire object is known to be underconstrainted (Kac, 1966), but prior research has explored what information can be estimated under different constraints. For example, binary shape attributes such as planarity and mirror symmetry, may be easier to estimate than a full geometric model (Fouhey et al., 2016; Fouhey et al., 2019). Sound can be used as a source of information for deeper understanding of 3D object structure. Little work has been done in this area, and existing methods limit themselves to estimation of shape attributes (Zhang et al., 2017b). Zhang et al. evaluated the ability of the ShapeNet neural network (Aytar et al., 2016) to identify an object shape out of 14 possible shapes, but these were largely shape primitives that were not representative of real-world object geometries (Zhang et al., 2017a). In **??**, I address these limitations with a method that uses multimodal input and evaluates on datasets of shapes more representative of the real world.

1.4 Visual Understanding

Having discussed machine understanding of sound, I now discuss what information can be gleaned through visual means.

1.4.1 Visual Object Reconstruction

An important step in object virtualization is to obtain the 3D shape and visual surface texture of a real-world object. This information can be obtained by reconstructing the object from a series of images looking at the object from multiple angles. Structure from Motion (SFM) (Westoby et al., 2012; Snavely et al., 2006), Multi-View Stereo (MVS) (Goesele et al., 2007; Seitz et al., 2006), and Shape from Shading (Zhang et al., 1999) are classes of techniques for obtaining 3D shape information from a set of 2D images. Although these methods alone do not achieve a segmented representation of the objects within the scene, they serve as a foundation for many algorithms. Bundle adjustment is used to jointly optimize poses when many images are used as input (Triggs et al., 2000). RGB-D depth-based, active reconstruction methods can also be used to generate 3D geometrical models of static (Newcombe et al., 2011; Golodetz* et al., 2015) and dynamic (Newcombe et al., 2015; Dai et al., 2017) scenes using commodity sensors such as the Microsoft Kinect and GPU hardware in real-time.

1.4.1.1 3D Object Datasets

With the rise of data-driven methods for visual understanding, large and well-annotated datasets have become valuable. Thanks to a plethora of 3D scene and object datasets such as BigBIRD(Singh et al., 2014) and RGB-D Object Dataset (Lai et al., 2011), neural network models have been trained to label objects based on their visual representation. 3D ShapeNets (Wu et al., 2015) also provides two sets of object categories for object classification referred to as ModelNet10 and ModelNet40, which are common benchmarks for evaluation (Kanezaki et al., 2016). Scene-based datasets have also been built from RGB-D reconstruction scans of entire spaces, allowing for semantic data such as object and room relationships. For instance, NYU Depth Dataset (Silberman et al., 2012) and SUNCG (Song et al., 2017) enable indoor segmentation and semantic scene completion from depth images.

1.4.2 Multimodal Understanding

It is well known that vision alone is limited in its ability to understand scenes. In this dissertation I focus on using audio as a primary cue for improved object understanding, though many other modalities have also been explored. Here, I will discuss prior work on multimodal understanding.

Additional input modalities may improve results for objects and materials that are difficult to reconstruct. Reflective objects have glare which change in location with the movement of the viewer, while transparent objects make it difficult to determine depth. A time-of-flight camera can correct estimated depth of transparent objects (Tanaka et al., 2017). The dip transform for 3D shape reconstruction (Aberman et al., 2017) uses fluid displacement of an object to obtain shape information.

Sound and video are intrinsically linked modalities for understanding the same scene, object, or event. Using visual and audio information, it is possible to predict the sound corresponding to a visual image or video (Owens et al., 2016b; Aytar et al., 2016). Sound prediction from video has also been specifically explored for impact sounds (Owens et al., 2016a).

Impact sound provides an additional input modality, containing cues about the internal structure of an object. Environmental scene classification is a related task approached through spectral analysis (Büchler et al., 2005) or convolutional neural networks (Piczak, 2015a; Salamon and Bello, 2017), but produces broad classifications of an entire environment. However, no current methods use impact sounds in particular to aid in complete shape reconstruction. One goal of this research is to use impact sounds to help determine the object shape and material in cases where visual methods struggle.

1.4.2.1 Multimodal Fusion

Other works have fused audio and visual cues to better understand objects and scenes. Sparse auditory clues can supplement the ability of random fields to obtain material labels and perform segmentation (Arnab et al., 2015). Neural networks have proven valuable in fusing audio-visual input to emulate the sensory interactions of human information processing (Zhang et al., 2017b). While multimodal methods have succeeded in fusing input streams to capture material and low-level shape properties to aid segmentation, they have not attempted to identify specific object geometries.

Early attempts at multimodal fusion in neural networks focused on increasing classification specificity by combining the individual classification results of separate input streams (Simonyan and Zisserman, 2014). Bilinear modeling can model the multiplicative interactions of differing input types, and has been applied as a method of pooling input streams in neural networks (Tenenbaum and Freeman, 2000; Lin et al., 2015). Bilinear methods have been further developed to reduce complexity and increase speed, while other approaches to modeling multiplicative interactions have also been explored (Gao et al., 2016; Yu et al., 2017; Park et al., 2016). Bilinear methods have not yet been applied to merging audio-visual networks. The ISNN networks I present in **??** take a step towards combined audio-visual object reconstruction with bilinear methods.

BIBLIOGRAPHY

Aberman, K., Katzir, O., Zhou, Q., Luo, Z., Sharf, A., Greif, C., Chen, B., and Cohen-Or, D. (2017). Dip transform for 3D shape reconstruction. *ACM Transactions on Graphics (Special Issue of SIGGRAPH)*, 36(4):Article No. 79.

Adhikari, S. (2001). *Damping models for structural vibration*. PhD thesis, University of Cambridge.

Adhikari, S. and Woodhouse, J. (2001). Identificaion of damping: Part 1, viscous damping. *Journal of Sound and Vibration*, 243(1):43 – 61.

Akenine-Moller, T., Moller, T., and Haines, E. (2002). *Real-Time Rendering*. A. K. Peters, Ltd., Natick, MA, USA, 2nd edition.

Arnab, A., Sapienza, M., Golodetz, S., Valentin, J., Miksik, O., Izadi, S., and Torr, P. H. S. (2015). Joint object-material category segmentation from audio-visual cues. In Xianghua Xie, M. W. J. and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 40.1–40.12. BMVA Press.

Avanzini, F. and Rocchesso, D. (2001). Modeling collision sounds: Non-linear contact force. In *In Proc. COST-G6 Conf. Digital Audio Effects (DAFx-01*, pages 61–66.

Aytar, Y., Vondrick, C., and Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900.

Barchiesi, D., Giannoulis, D., Stowell, D., and Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3):16–34.

Basdogan, C., Ho, C.-H., and Srinivasan, M. A. (1997). A ray-based haptic rendering technique for displaying shape and texture of 3D objects in virtual environments. In *Proc. ASME Dynamic Systems and Control Division*, pages 77 – 84.

Bharaj, G., Levin, D. I. W., Tompkin, J., Fei, Y., Pfister, H., Matusik, W., and Zheng, C. (2015). Computational design of metallophone contact sounds. *ACM Trans. Graph.*, 34(6):223:1–223:13.

Bilbao, S., Hamilton, B., Torin, A., Webb, C., Graham, P., Gray, A., Kavoussanakis, K., and Perry, J. (2013). Large scale physical modeling sound synthesis. In *Proceedings of the Stockholm Musical Acoustics Conference/Sound and Music Computing Conference*.

Bilbao, S., Torin, A., and Chatziioannou, V. (2015). Numerical modeling of collisions in musical instruments. *Acta Acustica united with Acustica*, 101(1):155–173.

Blinn, J. F. (1978). Simulation of wrinkled surfaces. *SIGGRAPH Comput. Graph.*, 12(3):286–292.

Bonneel, N., Drettakis, G., Tsingos, N., Viaud-Delmon, I., and James, D. (2008). Fast modal sounds with scalable frequency-domain synthesis. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)*, 27(3).

Büchler, M., Allegro, S., Launer, S., and Dillier, N. (2005). Sound classification in hearing aids inspired by auditory scene analysis. *EURASIP Journal on Advances in Signal Processing*, 2005(18):387845.

Caughey, T. (1960). Classical normal modes in damped linear dynamic systems. *Journal of Applied Mechanics*, 27(2):269–271.

Caughey, T. and O'Kelly, M. (1965). Classical normal modes in damped linear dynamic systems. *Journal of Applied Mechanics*, 32(3):583–588.

Chadwick, J. N., An, S. S., and James, D. L. (2009). Harmonic shells: A practical nonlinear sound model for near-rigid thin shells. In *ACM SIGGRAPH Asia 2009 Papers*, SIGGRAPH Asia '09, pages 119:1–119:10, New York, NY, USA. ACM.

Chadwick, J. N. and James, D. L. (2011). Animating fire with sound. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11, pages 84:1–84:8, New York, NY, USA. ACM.

Chadwick, J. N., Zheng, C., and James, D. L. (2012). Precomputed acceleration noise for improved rigid-body sound. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2012)*, 31(4).

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

Cohen, J., Olano, M., and Manocha, D. (1998). Appearance-preserving simplification. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pages 115–122, New York, NY, USA. ACM.

Cook, P. R. and Scavone, G. P. (1999). The synthesis toolkit (STK). In *In Proceedings of the International Computer Music Conference*.

Cook, R. L. (1984). Shade trees. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '84, pages 223–231, New York, NY, USA. ACM.

Cowling, M. and Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895 – 2907.

Dai, A., Niessner, M., Zollhöfer, M., Izadi, S., and Theobalt, C. (2017). BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3).

Doel, K. V. D., Kry, P. G., and Pai, D. K. (2001). FoleyAutomatic: Physically-based sound effects for interactive simulation and animation. In *in Computer Graphics (ACM SIGGRAPH 01 Conference Proceedings*, pages 537–544. ACM Press.

E756, A. (2017). Standard test method for measuring vibration-damping properties of materials.

Featherstone, R. (2007). *Rigid Body Dynamics Algorithms*. Springer-Verlag, Berlin, Heidelberg.

Foley, J. D., van Dam, A., Feiner, S. K., and Hughes, J. F. (1990). *Computer Graphics: Principles and Practice (2Nd Ed.)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Fouhey, D. F., Gupta, A., and Zisserman, A. (2016). 3D shape attributes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1516–1524.

Fouhey, D. F., Gupta, A., and Zisserman, A. (2019). From images to 3D shape attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):93 – 106.

Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., and Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision*, 14(4):12.

Galoppo, N., Tekin, S., Otaduy, M. A., Gross, M., and Lin, M. C. (2007). Interactive haptic rendering of high-resolution deformable objects. In *Proceedings of the 2Nd International Conference on Virtual Reality*, ICVR'07, pages 215–233, Berlin, Heidelberg. Springer-Verlag.

Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016). Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Giordano, B. L. and McAdams, S. (2006). Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119(2):1171–1181.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., and Seitz, S. M. (2007). Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

Golodetz*, S., Sapienza*, M., Valentin, J. P. C., Vineet, V., Cheng, M.-M., Arnab, A., Prisacariu, V. A., Kähler, O., Ren, C. Y., Murray, D. W., Izadi, S., and Torr, P. H. S. (2015). SemanticPaint: A Framework for the Interactive Segmentation of 3D Scenes. Technical Report TVG-2015-1, Department of Engineering Science, University of Oxford. Released as arXiv e-print 1510.03727.

Grassi, M. (2005). Do we hear size or sound? Balls dropped on plates. *Perception & Psychophysics*, 67(2):274–284.

Hampel, S., Langer, S., and Cisilino, A. P. (2008). Coupling boundary elements to a raytracing procedure. *International Journal for Numerical Methods in Engineering*, 73(3):427–445.

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

Ho, C.-H., Basdogan, C., and Srinivasan, M. A. (1999). Efficient point-based rendering techniques for haptic display of virtual objects. *Presence: Teleoper. Virtual Environ.*, 8(5):477–491.

Huzaifah, M. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *CoRR*, abs/1706.07156.

Imregun, M. and Ewins, D. J. (1995). Complex Modes - Origins and Limits. In *Proceedings of the 13th International Modal Analysis Conference*, volume 2460, page 496.

James, D. L., Barbič, J., and Pai, D. K. (2006). 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.

Kac, M. (1966). Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1–23.

Kaneko, T., Takahei, T., Inami, M., Kawakami, N., Yanagida, Y., Maeda, T., and Tachi, S. (2001). Detailed shape representation with parallax mapping. In *In Proceedings of the ICAT*, pages 205–208.

Kanezaki, A., Matsushita, Y., and Nishida, Y. (2016). RotationNet: Learning object classification using unsupervised viewpoint estimation. *CoRR*, abs/1603.06208.

Karplus, K. and Strong, A. (1983). Digital synthesis of plucked-string and drum timbres. *Computer Music Journal*, 7(2):43–55.

Klatzky, R. L., Lederman, S. J., and Metzger, V. A. (1985). Identifying objects by touch: An "expert system". *Perception & Psychophysics*, 37(4):299–302.

Klatzky, R. L., Pai, D. K., and Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, 9(4):399–410.

Lai, K., Bo, L., Ren, X., and Fox, D. (2011). A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824.

Langlois, T. R., An, S. S., Jin, K. K., and James, D. L. (2014). Eigenmode compression for modal sound models. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2014)*, 33(4).

Langlois, T. R., Zheng, C., and James, D. L. (2016). Toward animating water with complex acoustic bubbles. *ACM Trans. Graph.*, 35(4):95:1–95:13.

Lederman, S. J. and Taylor, M. M. (1972). Fingertip force, surface geometry, and the perception of roughness by active touch. *Perception & Psychophysics*, 12(5):401–408.

Li, D., Fei, Y., and Zheng, C. (2015). Interactive acoustic transfer approximation for modal sound. *ACM Trans. Graph.*, 35(1).

Lin, T., RoyChowdhury, A., and Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457.

Lloyd, D. B., Raghuvanshi, N., and Govindaraju, N. K. (2011). Sound synthesis for impact sounds in video games. In *Symposium on Interactive 3D Graphics and Games*, I3D '11, pages 55–62, New York, NY, USA. ACM.

Loomis, J. M. and Lederman, S. J. (1986). Tactual perception. *Handbook of perception and human performances*, 2:2.

McAdams, S., Roussarie, V., Chaigne, A., and Giordano, B. L. (2010). The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *The Journal of the Acoustical Society of America*, 128(3):1401–1413.

McDermott, J. H., Schemitsch, M., and Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493.

Mehra, R., Antani, L., Kim, S., and Manocha, D. (2014). Source and listener directivity for interactive wave-based sound propagation. *IEEE Transactions on Visualization and Computer Graphics*, 20(4):495–503.

Mehra, R., Raghuvanshi, N., Antani, L., Chandak, A., Curtis, S., and Manocha, D. (2013). Wave-based sound propagation in large open scenes using an equivalent source formulation. *ACM Trans. Graph.*, 32(2):19:1–19:13.

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

Michon, R., Martin, S. R., and Smith, J. O. (2017). Mesh2faust: a modal physical model generator for the faust programming language – application to bell modeling. In *Proceedings of the International Computer Music Conference (ICMC-17)*, Shanghai, China.

Michon, R. and Smith, J. O. (2011). Faust-STK: a set of linear and nonlinear physical models for the faust programming language. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, pages 19–23, Paris, France.

Minsky, M., Ming, O.-y., Steele, O., Brooks, Jr., F. P., and Behensky, M. (1990). Feeling and seeing: Issues in force display. *SIGGRAPH Comput. Graph.*, 24(2):235–241.

Minsky, M. D. R. R. (1995). *Computational Haptics: The Sandpaper System for Synthesizing Texture for a Force-feedback Display*. PhD thesis, Cambridge, MA, USA. Not available from Univ. Microfilms Int.

Morrison, J. D. and Adrien, J.-M. (1993). MOSAIC: A framework for modal synthesis. *Computer Music Journal*, 17(1):45–56.

Moss, W., Yeh, H., Hong, J.-M., Lin, M. C., and Manocha, D. (2010). Sounding liquids: Automatic sound synthesis from fluid simulation. *ACM Trans. Graph.*, 29(3):21:1–21:13.

Nashif, A. D., Jones, D. I., and Henderson, J. P. (1985). *Vibration damping*. John Wiley & Sons.

Newcombe, R. A., Fox, D., and Seitz, S. M. (2015). DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 343–352.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). KinectFusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136.

Nykl, S., Mourning, C., and Chelberg, D. (2013). Interactive mesostructures. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '13, pages 37–44, New York, NY, USA. ACM.

O'Brien, J. F., Shen, C., and Gatchalian, C. M. (2002). Synthesizing sounds from rigid-body simulations. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '02, pages 175–181, New York, NY, USA. ACM.

Oliveira, M. M., Bishop, G., and McAllister, D. (2000). Relief texture mapping. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 359–368, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Otaduy, M., Jain, N., Sud, A., and Lin, M. (2004). Haptic display of interaction between textured models. In *IEEE Visualization Conference*, pages 297–304.

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016a). Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. (2016b). *Ambient Sound Provides Supervision for Visual Learning*, pages 801–816. Springer International Publishing, Cham.

Pai, D. K., Doel, K. v. d., James, D. L., Lang, J., Lloyd, J. E., Richmond, J. L., and Yau, S. H. (2001). Scanning physical interaction behavior of 3D objects. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 87–96. ACM.

Park, E., Han, X., Berg, T. L., and Berg, A. C. (2016). Combining multiple sources of knowledge in deep CNNs for action recognition. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE.

Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Piczak, K. J. (2015b). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1015–1018, New York, NY, USA. ACM.

Policarpo, F., Oliveira, M. M., and Comba, J. a. L. D. (2005). Real-time relief mapping on arbitrary polygonal surfaces. In *Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games*, I3D '05, pages 155–162, New York, NY, USA. ACM.

Raghuvanshi, N., Allen, A., and Snyder, J. (2016). Numerical wave simulation for interactive audio-visual applications. *The Journal of the Acoustical Society of America*, 139(4):2008–2009.

Raghuvanshi, N. and Lin, M. C. (2006). Interactive sound synthesis for large scale environments. In *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games*, I3D '06, pages 101–108, New York, NY, USA. ACM.

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

Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., and Govindaraju, N. (2010). Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM Trans. Graph.*, 29(4):68:1–68:11.

Rayleigh, J. W. S. B. (1896). *The Theory of Sound*, volume 2. Macmillan.

Ren, Z., Mehra, R., Coposky, J., and Lin, M. (2012). Designing virtual instruments with touch-enabled interface. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 433–436. ACM.

Ren, Z., Yeh, H., Klatzky, R., and Lin, M. C. (2013a). Auditory perception of geometry-invariant material properties. *Visualization and Computer Graphics, IEEE Transactions on*, 19(4):557–566.

Ren, Z., Yeh, H., and Lin, M. (2010). Synthesizing contact sounds between textured models. In *Virtual Reality Conference (VR), 2010 IEEE*, pages 139–146.

Ren, Z., Yeh, H., and Lin, M. C. (2013b). Example-guided physically based modal sound synthesis. *ACM Trans. Graph.*, 32(1):1:1–1:16.

Rungta, A., Schissler, C., Mehra, R., Malloy, C., Lin, M., and Manocha, D. (2016). Syncopation: Interactive synthesis-coupled sound propagation. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1346–1355.

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

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

Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 1041–1044, New York, NY, USA. ACM.

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

Schissler, C. and Manocha, D. (2011). GSound: Interactive sound propagation for games. In *Audio Engineering Society Conference: 41st International Conference: Audio for Games*.

Schissler, C. and Manocha, D. (2016). Adaptive impulse response modeling for interactive sound propagation. In *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '16, pages 71–78, New York, NY, USA. ACM.

Schreck, C., Rohmer, D., James, D. L., Hahmann, S., and Cani, M.-P. (2016). Real-time sound synthesis for paper material based on geometric analysis. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '16, pages 211–220, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528.

Serra, X. and Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, pages 746–760, Berlin, Heidelberg. Springer-Verlag.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc.

Singh, A., Sha, J., Narayan, K. S., Achim, T., and Abbeel, P. (2014). BigBIRD: A large-scale 3D database of object instances. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 509–516. IEEE.

Slater, J. C., Keith Belvin, W., and Inman, D. J. (1993). A survey of modern methods for modeling frequency dependent damping in finite element models. In *PROCEEDINGS-SPIE THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING*, pages 1508–1508. SPIE INTERNATIONAL SOCIETY FOR OPTICAL.

Smith, J. O. (1992). Physical modeling using digital waveguides. *Computer Music Journal*, 16(4):74–91.

Snavely, N., Seitz, S. M., and Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846.

Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., and Funkhouser, T. (2017). Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*.

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

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

Szirmay-Kalos, L. and Umenhoffer, T. (2008). Displacement mapping on the gpu — state of the art. *Computer Graphics Forum*, 27(6):1567–1592.

Tanaka, K., Mukaigawa, Y., Funatomi, T., Kubo, H., Matsushita, Y., and Yagi, Y. (2017). Material Classification using Frequency- and Depth-dependent Time-of-Flight Distortion. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 79–88.

Tenenbaum, J. B. and Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Comput.*, 12(6):1247–1283.

Tevs, A., Ihrke, I., and Seidel, H.-P. (2008). Maximum mipmaps for fast, accurate, and scalable dynamic height field rendering. In *Proceedings of the 2008 Symposium on Interactive 3D Graphics and Games*, I3D '08, pages 183–190, New York, NY, USA. ACM.

Theoktisto, V., González, M. F., and Navazo, I. (2010). Hybrid rugosity mesostructures (HRMs) for fast and accurate rendering of fine haptic detail. *CLEI Electron. J.*, pages –1–1.

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

Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. (2000). Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, UK. Springer-Verlag.

Tsingos, N., Funkhouser, T., Ngan, A., and Carlbom, I. (2001). Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 545–552, New York, NY, USA. ACM.

van den Doel, K. and Pai, D. K. (1996). The sounds of physical shapes. *Presence*, 7:382–395.

van Walstijn, M. and Mehes, S. (2017). An explorative string-bridge-plate model with tunable parameters. In *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK.

Wang, J.-H., Qu, A., Langlois, T. R., and James, D. L. (2018). Toward wave-based sound synthesis for computer animation. *ACM Trans. Graph.*, 37(4):109:1–109:16.

Webb, C. J. (2014). *Parallel computation techniques for virtual acoustics and physical modelling synthesis*. PhD thesis, The University of Edinburgh.

Westoby, M., Brasington, J., Glasser, N., Hambrey, M., and Reynolds, J. (2012). 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300 – 314.

Woodhouse, J. (1998). Linear damping models for structural vibration. *Journal of Sound and Vibration*, 215(3):547–569.

Wu, Z., Song, S., Khosla, A., and Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920.

Yamamoto, K. and Igarashi, T. (2016). Interactive physically-based sound design of 3D model using material optimization. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '16, pages 231–240, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Yeh, H., Mehra, R., Ren, Z., Antani, L., Manocha, D., and Lin, M. (2013). Wave-ray coupling for interactive sound propagation in large complex scenes. *ACM Trans. Graph.*, 32(6):165:1–165:11.

Yu, Z., Yu, J., Fan, J., and Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848.

Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. (1999). Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706.

Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J., and Freeman, B. (2017a). Shape and material from sound. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1278–1288. Curran Associates, Inc.

Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J. H., Tenenbaum, J. B., and Freeman, W. T. (2017b). Generative modeling of audible shapes for object perception. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1260–1269.

Zheng, C. and James, D. L. (2011). Toward high-quality modal contact sound. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2011)*, 30(4).