CHAPTER 1: INTRODUCTION

1.1 Introduction

Virtual environments have found applications for different user interaction scenarios. Interactive training simulations let users practice high-risk tasks, such as performing surgery or piloting an airplane, with low risk. Immersive story-driven video games let users interact with another environment or involve themselves in an engaging narrative. Emerging social applications let multiple users from around the world unite in one virtual location and feel as though they are in the same space.

In all three scenarios, users should be able to forget their presence in the real world and temporarily experience a *sense of presence* in the virtual environment (Lombard and Jones, 2015; Lee, 2006). If users are reminded that the virtual environment is fake, they experience a *break in presence*, which reduces the emotional weight of the virtual environment and makes it less effective at its intended goal. Thus, avoiding these breaks in presence can improve the quality of users' experiences. Training simulations feel more lifelike, video games convey more powerful emotions, and social interactions with other users flow more naturally.

Virtual environments most commonly recreate input to the senses of sight and hearing. The visual appearances and audio of the real world are relatively easily replaceable with those of a virtual world. A virtual

Figure 1.1: A virtual environment with interactive objects of different shapes and materials. The objects in these scenes should produce realistic impact sounds consistent with their visual appearances.

reality (VR) headset such as the Oculus Rift or the HTC Vive replaces the visual input, while headphones (sometimes built into the VR headset) replace the audio input. Examples of VR-enabled environments are shown in Figure 1.1. Humans also rely heavily on the sense of touch, but virtual environments are limited by current hardware, which cannot effectively recreate complete input for the sense of touch.

Undesirable breaks in presence have many causes; a common cause is violation of a user's expectations about their interactions. Each sense creates expectations for the other senses; *sensory conflict* occurs when senses provide conflicting expectations that violate one another. For example, if a table looks like wood when visually inspected, but sounds like ringing metal when struck, the user's expectations have been violated and a break in presence is likely. For another example, a rough surface will visually have a diffuse scattering of light instead of a sharp reflection, and if the user feels that surface with a stylus, the roughness they feel should match the roughness they see.

Maintaining users' expectations about their interactions does not require perfectly realistic virtual environments. Some studies have found that visual rendering quality has little effect on presence (Zimmons and Panter, 2003) (though this is still an open area of research (Slater et al., 2009)), and other studies have found that virtual environments with deliberately low-fidelity visuals still evoke strong desired emotional responses (Slater et al., 2006b,a). As long as the user can establish consistent expectations about the environment, sensory conflict can be avoided, regardless of the objective real-world accuracy of the recreations. Consistent sensory expectations cause the sensation of *perceptual realism*, protecting a user's sense of presence from sensory conflict.

A common source of sensory conflict is interaction with objects. Objects, such as furniture, tableware, and musical instruments, are common in real and virtual environments, and interaction with objects involves multiple senses. As objects are moved or struck, we expect them to produce *impact sounds*. To create realistic impact sounds, my work uses *modal sound synthesis*, a physically-based method which models vibrations in struck objects (O'Brien et al., 2002). When using physically-based methods for sound synthesis, perceptual realism depends on an object's *material parameters*. The material parameters affecting impact sounds can be collectively referred to as the *audio-material*, in contrast with parameters affecting the visual appearance or haptic texture of the surface. Since users have expectations about how virtual objects should sound from their other senses (Fujisaki et al., 2014), it is important to use accurate material parameters.

The audio-material of a struck object affects the impact sound's rate of decay, *e.g.*, a plastic object has a short-lasting sound while a metal object has a long-lasting sound. Different materials cause different

Figure 1.2: Images of objects and their virtual reconstructions. The top row shows pictures of the real objects, while the bottom row shows manually-constructed meshes and textures modeling the objects. We seek to create realistic multimodal interactions with these objects.

amounts of *damping*, producing different decay rates. *Damping models* are a common approximation that simplify computations by modeling the damping as a function of an object's mass and stiffness. When performing modal sound synthesis, selecting realistic damping rates is important for recreating the sound of the appropriate material.

In order to interact with virtual objects, they must first be created by defining properties such as its shape and material parameters. A common way of creating virtual environments and objects is modeling existing real-world objects. Figure 1.2 shows examples of real-world objects that have been modeled by hand, though this process can be automated to a limited degree. The shape of an object can be acquired through a 3D scan, and the object's material parameters can be acquired through vision or its impact sounds. The acquired properties can be used to *virtualize* the original object, reconstructing a digital version of the object for interaction in virtual environments. However, few methods have attempted to combine these two input modalities (vision and impact sounds) in a single coupled process. An object reconstruction method that ensures consistency between input modalities could better recreate virtual objects with minimal sensory conflict.

In this dissertation, I propose methods for multimodal interaction and object reconstruction. These methods are evaluated through comparisons to ground truth and perceptual experiments. Comparisons to ground truth analyze the difference between my results, prior results, and a referenced ground truth. Perceptual experiments consist of user studies to analyze perception of objects and sounds, and are frequently used in my work due to the emphasis on ensuring perceptual realism. In some user studies, subjects report their opinions on the realism, effectiveness, or similarity of real or synthetic virtual objects. In other user studies, subjects must complete tasks using either my methods or those from previous work. User studies directly evaluate the methods with respect to user expectations, providing insight into the methods' performance in an immersive setting.

1.1.1 Thesis Statement

"Interaction with objects in virtual environments can be made more perceptually realistic by using expressive object material models that account for real-world phenomena and by reducing multimodal sensory conflict."

In this dissertation, I describe research that improves interaction with virtual objects by improving material modeling and object virtualization. The contributions proposed by myself and my collaborators include:

Damping Modeling for Modal Sound Synthesis: We propose a novel method for deriving new material damping models which are able to express a wider range of damping behaviors than the traditional Rayleigh damping model. We extend modal sound synthesis to support these new models, We also propose a method for material parameter estimation that uses a single impact sound to accurately estimate damping parameters for *any* damping model. Perceptual evaluation demonstrates that no single existing damping model best represents the damping behavior of every material, and thus multiple damping models should be considered.

Robust Material Parameter Estimation: We propose a novel method for estimating material damping parameters from recorded impact sounds. We model the observed damping values in recorded sounds with a probabilistic model, which expressly models multiple external factors affecting estimation of material damping. This method requires no information about the shape of the object or the locations of the impacts, and reduces the effect of external factors to produce more accurate estimates of the material damping parameters in suboptimal recording environments. Perceptual evaluation shows that sounds synthesized using our estimated material parameters are comparable in realism to those of previous work and human-tuned sounds. Given that our method places fewer requirements on the inputs, our method significantly reduces manual effort needed to obtain high-quality results.

Multimodal Surface Interaction: We propose a method for using a single texture map as a unified representation of detail for visual rendering, audio rendering, tactile rendering, and physical simulation. Our method runs in real time and allows for multimodal interaction with textured surfaces, while the unified representation ensures consistency between senses. In task-based user evaluation, our method improves results over alternative, conflicting representations of detail. In a comparison study, we identify situations where each of our two texture representations are most effective.

Multimodal Object Classification We propose a method for estimating both an object's material and geometry leveraging both audio and visual input. The method takes as input an impact sound and (optionally) a voxelized estimate of the object's shape. We perform quantitative evaluation on datasets in which the output is a geometry class (such as "chair" or "dresser"), and on datasets in which the output is a specific geometric model (a retrieval task). Our method results in state-of-the-art accuracy for these sets of inputs, while proving competitive against methods using different sets of inputs.

1.2 Main Contributions

In this section, I discuss my primary areas of research.

1.2.1 Interactive Modal Sound Synthesis Using Generalized Proportional Damping

In order to create higher quality modal sound, this research aims to improve modeling of material damping. We (myself and Ming C. Lin) more accurately capture real-word damping behavior by considering more expressive damping models. We apply these expressive models to modal sound synthesis to better recreate the sounds of real-world materials.

Material damping is a complex phenomenon, and is difficult to accurately model. For example, the presence of damping may give rise to *complex* modes of vibration, which are more difficult to model than *normal* modes (Caughey and O'Kelly, 1965). In practice, approximations are used to produce computationally simpler models using only normal modes. The most common approximation is to assume all damping is viscous and that the decay rate of a material is a linear combination of its density and stiffness. This model is referred to as Rayleigh (or linearly proportional) damping, and produces only normal modes. It is the de-facto damping model for modal sound synthesis, but has always been understood to be an approximation for convenience.

Other damping models are common in material and structural analysis, but have not been thoroughly examined for interactive sound synthesis. Caughey damping is a polynomial extension of the linear Rayleigh damping model (Caughey, 1960; Caughey and O'Kelly, 1965). Generalized proportional damping (GPD) is the most general damping model to date that limits vibrations to normal modes (Adhikari, 2006). These

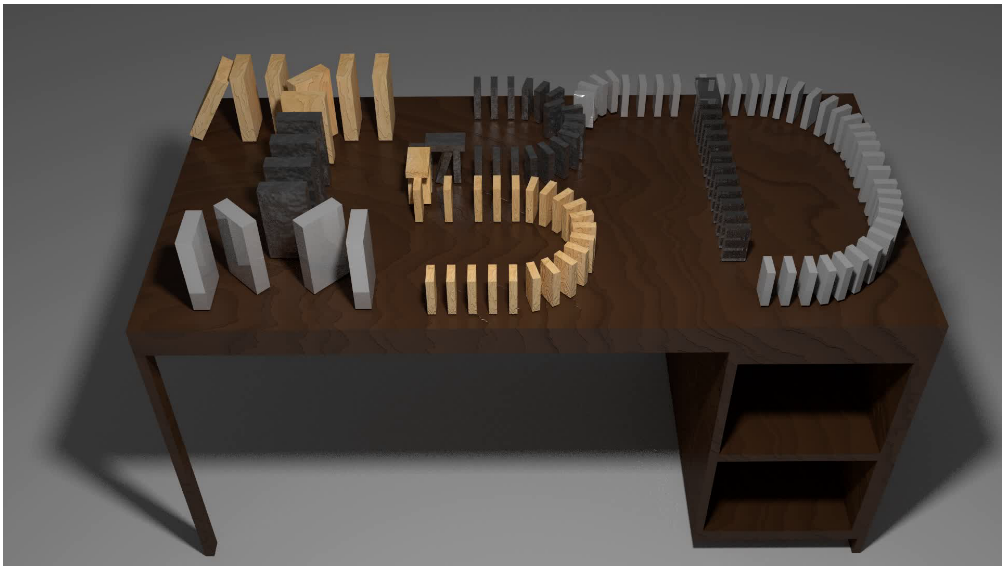

Figure 1.3: A scenario with dominoes made of different materials. Each material uses a set of material parameters estimated from recorded impact sounds, including parameters for a damping model.

alternative damping models may be able to improve sound quality by providing a better fit to observed real-world damping.

We propose a method that employs Generalized Proportional Damping to create alternative damping models for sound synthesis and we propose specific damping models within the larger space of GPD functions. These damping models are more expressive, enabling them to model damping behavior that would be coarsely approximated by the Rayleigh damping model. We also propose a method for estimating the damping parameters of a real-world object using a recorded impact sound as input. This parameter estimation method works for any arbitrary damping model, producing estimates of the parameters specific to that model. We also conduct a user study to evaluate the perceptual differences between multiple damping models. Figure 1.3 shows one scenario with objects creating sound based on materials estimated from recorded impact sounds. More results can be found in **??** and online: `http://gamma.cs.unc.edu/gpdsynth/`.

1.2.2 Audio-Material Reconstruction for Virtualized Reality using a Probabilistic Damping Mode

Recorded impact sounds can be used to estimate damping parameters, but resulting parameters may be inaccurate if the sounds are recorded in noisy and uncontrolled recording environments. This research explores a novel probabilistic damping model for estimating material damping parameters while reducing the

Figure 1.4: A small porcelain plate (left) and a small travertine tile (right) being struck to produce impact sounds. Both objects are supported by a gripping hand. Methods for material damping parameter estimation should be robust to these external damping factors.

confounding effect of the recording environment. My collaborators on this research are Nicholas Rewkowski, Roberta L. Klatzky, and Ming C. Lin.

While recent methods have been able to estimate material damping parameters (Ren et al., 2013), they assume all observed decay in amplitude is due to material damping and not any other source. However, multiple external factors produce effects similar to material damping, causing error in material damping estimates.

One external factor is *support damping*. In the real world, an object struck for recording must be supported in some way, *e.g.*, held by hand or left to rest on another surface. The interface between the object and its *support* introduces additional damping, as energy is transferred from the vibrating object to the support. Figure 1.4 shows multiple objects supported by a hand while being struck, altering the produced sound. **??** provides a more in-depth example: depending on how the bowl is held, it produces dramatically different sound.

Other external factors include complex modes of vibration, room acoustics, and error in the feature extraction step. Complex modes of vibration are not captured by standard damping models, which only model normal modes of vibration. Room acoustics—reflections off walls—extend the length of sounds when recording is performed in an enclosed room. Feature extraction steps are common in most damping parameter estimation methods, but even with clean input these steps often introduce their own error.

In realistic, uncontrolled environments with significant effect from external factors, the parameters estimated by current methods are not truly material parameters. Instead, they are parameters modeling *both* the material and the environment used for the recording and thus do not generalize to arbitrary environments.

We propose a practical and efficient method to estimate material damping parameters from recorded impact sounds, while accounting for the external factors present in the recording environment. We explicitly model the external factors using a probabilistic damping model. For a given frequency of vibration and parameters describing the recording environment, this model provides a probability distribution of possible observable damping values. Using multiple impact sounds as input, the probabilistic damping model can be optimized to fit the sounds, providing an estimate of the material damping parameters separately from the external factors. The optimized damping parameters can be those of any real-valued damping model, and we propose one additional hybrid model combining Rayleigh and power law damping.

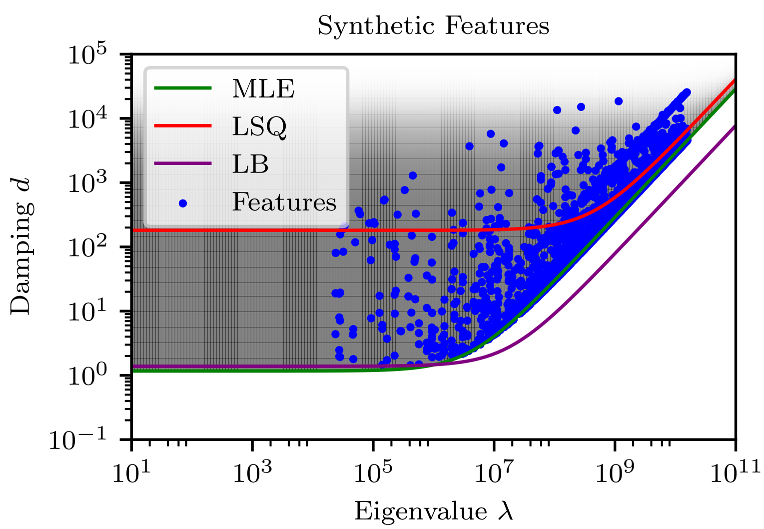

Figure 1.5: Parameter estimation on sound features. Each feature is one extracted mode of vibration, consisting of an eigenvalue λ_i (approximately the square of its frequency) and its corresponding damping coefficient d_i. Estimated Rayleigh damping curves are plotted, with the variation from the curve caused by external factors. Our method is labeled MLE, and provides the ideal fit to the lower bound of the sound features

Our method is more applicable to real-world recordings taken in less controlled environments. The method is fast, requires no prior knowledge about the recorded object, and can use multiple recordings to improve accuracy. Figure 1.5 shows a visual representation of the material damping models as estimated from the real-world sound features shown as points. In the absence of external damping factors, all of these

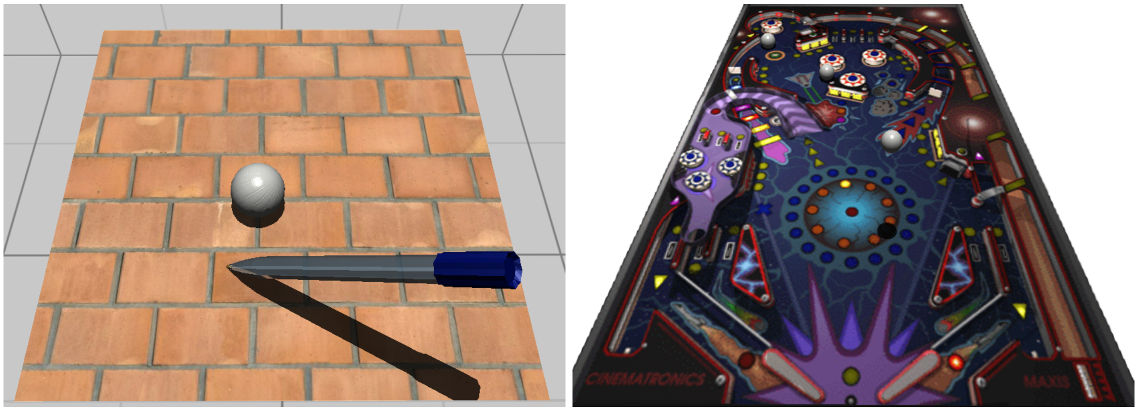

Figure 1.6: A selection of applications based on our method for multimodal interaction: a virtual environment with a normal mapped surface (left) and a pinball game created through a normal mapped surface (right). In both environments, the texture map informs all interaction modalities.

points would fall along one line representing the damping from the material alone. However, external factors cause the points to vary, throwing off a more traditional least squares (LSQ) approach while our method (MLE) provides a better fit. In perceptual evaluation, subjects found that sounds synthesized using parameters estimated with our method were comparable in quality to those of previous work and human hand-tuned parameters. Therefore, our method requires significantly less manual effort to produce high quality results. More results are available in **??** and online: `http://gamma.cs.unc.edu/ProbDampModel/`.

1.2.3 Integrated Multimodal Interaction Using Texture Representations

There have been a few efforts to unify interaction in virtual environments across senses (see **??**). However, they do not clearly consider sensory conflict, nor have any brought together all of visual rendering, haptic rendering, sound rendering, and physical simulation. Sensory conflict is particularly important when considering textured objects, which are often modeled through approximations. In this line of research, we (myself and Ming C. Lin) use texture representations of detail—normal and relief maps—as a unified source of information for all interaction modalities.

Interaction with textured surfaces via haptic rendering, sound rendering, and rigid-body simulation have each been independently explored (Otaduy et al., 2004; Ren et al., 2010), but have not been integrated together consistently. For example, a previous method for sound rendering of contacts with textured surfaces (Ren et al., 2010) displays a pen sliding smoothly across highly bumpy surfaces. While the generated sound from this interaction is dynamic and realistic, the smooth *visual* movement of the pen does not match the texture implied by the sound. In order to minimize sensory conflict, it is critical to present a unified and seamlessly integrated multimodal display to users, ensuring rendering is consistent across the senses of sight, hearing, and touch.

In the real world, objects behave differently when bouncing, sliding, or rolling on bumpy or rough surfaces than they do on flat surfaces. In a virtual environment, a bumpy or rough surface can be represented by its visual texture equivalent mapped to a flat surface. While the surface would appear visually complex, the underlying flat surface would cause simple physical behavior, causing sensory conflict and breaking the sense of presence. In order to model such physical behavior, a physics simulator would require a fine triangle mesh with sufficient surface detail, but in most cases a sufficiently fine mesh is unavailable or would require prohibitive amounts of memory. Since texture maps contain information about the fine detail of a mapped surface, it is possible to use that information to recreate the physical behavior of the fine triangle mesh.

To accomplish this, we propose a new method for simulation of physical behaviors for rigid objects textured with normal maps. We also propose methods for seamlessly integrated multimodal interaction using normal and relief maps. By using a single representation of surface detail across all interaction modalities, we reduce sensory conflict for users. See Figure 1.6 for examples of interaction with textured surfaces. With our method, a virtual pen is controlled through a haptic device, allowing a user to interact with the environment while feeling forces in response. A simulated ball rolls on the surface, its motions affected by both the surface texture and the pen. Contacts between the pen, ball, and surface create physically-based sound, bringing together sight, hearing, touch, and physical simulation.

We evaluate our methods through texture identification and representation comparison user studies. In the texture identification study, subjects were asked to identify the surface displayed to them, but in some trials certain interaction modalities were removed. When all modalities were present using our method, performance on the task was at its highest, demonstrating that the senses were not in conflict with one another. In the representation comparison study, subjects answered questionnaires as they interacted with either normal-mapped or relief-mapped surfaces. When all modalities were present, subjects found the relief-mapped surfaces to be more realistic. More results are available in **??** and online: `http://gamma.cs.unc.edu/MultiDispTexture/`.

1.2.4 Impact Sound Neural Network for Audio-Visual Object Classification

A real-world object reconstructed in virtual reality should minimize sensory conflict by ensuring consistency between the object's shape, surface appearance, and audio-material. Object shape and surface appearance have historically been estimated through visual cues. Similarly, the audio-material can been estimated through audio cues, as I demonstrate in **????**. However, if an object's shape and audio-material are estimated separately through independent methods, sensory conflict may appear. Visual methods for shape reconstruction cannot determine internal object structure (*e.g.*, whether an object is solid or hollow) while audio methods for material estimation are underconstrained (multiple shape/material combinations may produce the same impact sound).

These visual and audio cues can complement one another. Impact sounds provide information about internal object structure that visual methods cannot see. Visual estimates of an object's shape provide constraints to audio-based material parameter estimation. Therefore, estimation of either shape or material could benefit from using both visual and audio modalities of input.

In this research area, my collaborators—Justin Wilson, Sam Lowe, and Ming C. Lin—and I explore the combination of these modalities. We propose a method for estimating both an object's material and shape geometry using combined audio-visual inputs. As a visual input, we use a coarse voxelized shape representation which can be acquired from a rough 3D visual reconstruction or a synthetic dataset such as ModelNet (Wu et al., 2015). As an audio input, we use a single impact sound from the object in question, which can be acquired from a recording of a real-world sound or from modal sound synthesis on a virtual object.

Our method uses a novel neural network architecture, called the Impact Sound Neural Network (ISNN), to process and fuse these two inputs. We present an audio-only network (ISNN-A) for material and geometry classification which uses convolutional layers to process an input sound encoded as a spectrogram. We also present a multimodal network (ISNN-AV) which fuses ISNN-A and VoxNet (Maturana and Scherer, 2015) to jointly produce estimates of material and geometry.

We perform quantitative evaluation on multiple datasets. The synthetic ModelNet10 and ModelNet40 datasets (Wu et al., 2015) produce classifications to object classes such as "table" or "dresser". We synthesize sounds for each ModelNet object, and our ISNN networks obtain higher classification accuracy than baselines for both audio-only and audio-visual inputs. We present a new dataset, RSAudio, consisting of both recorded and synthesized sounds, where each sound classifies to a specific shape geometry. On this dataset and other audio-only impact sound datasets (Arnab et al., 2015; Zhang et al., 2017), our ISNN-A network also outperforms baselines. Finally, we present a utility for scene reconstruction in which impact sounds can be recorded to classify and segment objects. More results are available in **??** and online: `http://gamma.cs.unc.edu/ISNN/`.

BIBLIOGRAPHY

Adhikari, S. (2006). Damping modelling using generalized proportional damping. *Journal of Sound and Vibration*, 293(1-2):156–170.

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.

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.

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.

Lee, K. M. (2006). Presence, Explicated. *Communication Theory*, 14(1):27–50.

Lombard, M. and Jones, M. T. (2015). *Defining Presence*, pages 13–34. Springer International Publishing, Cham.

Maturana, D. and Scherer, S. (2015). VoxNet: A 3D convolutional neural network for real-time object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, page 922 – 928.

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.

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

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. (2013). Example-guided physically based modal sound synthesis. *ACM Trans. Graph.*, 32(1):1:1–1:16.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., and Sanchez-Vives, M. V. (2006a). A virtual reprise of the stanley milgram obedience experiments. *PLOS ONE*, 1(1):1–10.

Slater, M., Khanna, P., Mortensen, J., and Yu, I. (2009). Visual realism enhances realistic response in an immersive virtual environment. *IEEE Computer Graphics and Applications*, 29(3):76–84.

Slater, M., Pertaub, D.-P., Barker, C., and Clark, D. M. (2006b). An experimental study on fear of public speaking using a virtual environment. *CyberPsychology & Behavior*, 9(5):627–633. PMID: 17034333.

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.

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

Zimmons, P. and Panter, A. (2003). The influence of rendering quality on presence and task performance in a virtual environment. In *IEEE Virtual Reality, 2003. Proceedings.*, pages 293–294.