CHAPTER 1: INTRODUCTION

1.1 Introduction

User interaction scenarios use virtual environments to achieve diverse goals. Interactive training simulations enable low-risk practice of high-risk tasks, such as performing surgery or piloting an airplane. Immersive story-driven video games let users interact with another environment or become involved in an engaging narrative. Emerging social applications unite multiple users from around the world in one virtual location such that the users actually feel as though they are in the same space.

In these scenarios, a user should be able to forget their presence in the real world and temporarily experience a sense of presence in the virtual environment. If the user remembers the virtual environment is fake, they experience a break in presence. Breaks in presence reduce the emotional weight of the virtual environment, making the environment less effective at its intended goal. Better maintaining a user’s sense of presence can improve the quality of their 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 sounding 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 audial 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. This does not mean that virtual environments must always be realistic to maintain presence, but they must be consistent. Sensory conflict occurs when two senses provide conflicting expectations about one another. Each sense creates expectations for the other senses, and avoiding conflict between senses helps a user retain their sense of presence. 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. Similarly, 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 a user’s expectations about their interactions does not require perfectly realistic virtual environments. As long as the user can establish consistent expectations about the environment, the environment is effective, regardless of how accurate the recreations are compared to the real world. As long as perceptual realism is maintained, the user maintains their sense of presence.

A likely source of sensory conflict is in interaction with objects, which is common in many virtual environments and involves multiple senses. As virtual objects are moved or struck, we expect them to produce impact sounds. To create realistic impact sounds, my work uses methods for physically-based impact sound synthesis, which model vibrations in struck objects. 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, it is important to use accurate material parameters.

An impact sound’s rate of decay largely depends on the material of the struck object: a plastic object has a short-lasting sound while a metal object has a long-lasting sound. Different materials cause different 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.

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, but 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. These acquired object properties can be used to virtually reconstruct, or virtualize, the original object. 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 cohesion between input modalities could better recreate virtual objects with high sensory cohesion.

Evaluation of the methods in this dissertation can be performed through comparisons to ground truth or through perceptual evaluation. With comparisons to ground truth, an error metric determines the difference between our results, prior results, and the ground truth. Perceptual evaluation is more frequently used in my work due to the emphasis on ensuring perceptual realism. To evaluate perceptual realism, I rely on user studies with human subjects. 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 our 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 improving multimodal sensory cohesion.”

In this dissertation, I describe research that improves interaction with virtual objects by addressing limitations of previous work and simplifying virtualization of real-world objects. The contributions presented by myself and my collaborators include:

Damping Modeling for Modal Sound Synthesis: We present a novel method for deriving new material damping models and using those models for modal sound synthesis. We extend modal sound synthesis to support these additional damping models, which are able to express a wider range of damping behaviors than the traditional Rayleigh damping model. We also propose a material parameter estimation method that uses a single input 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 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 use a probabilistic model for observed damping values in recorded sounds, which expressly models multiple external factors affecting the damping values. This method requires no information about the shape of the object or the locations of the impacts, and is able to reduce the effect of external factors to produce more accurate estimates of the material damping parameters in suboptimal recording environments. User evaluation shows that sounds synthesized using our estimated parameters are comparable in realism to those of previous work. Given that our method has significantly fewer requirements about 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 cohesion between senses. In task-based user evaluation, our method improves results over alternative, possibly conflicting representations of detail.
**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 rough voxelized shape estimate. We perform 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 results for this 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 and how my proposed methods address the limitations of previous work.

1.2.1 Interactive Modal Sound Synthesis Using Generalized Proportional Damping

In order to create higher quality modal sound, this research aims to improve methods for damping modeling. For this research area, we (myself and Ming C. Lin) consider more expressive models for the damping behavior of objects. Since the damping rate of a sound tell us about an object’s material, more realistic damping modeling in synthesized sound should be better able to recreate the sounds of real-world materials.

Damping is a complex phenomenon to model, and it can be difficult to determine exactly how vibrations in an object decay. Additionally, 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 simpler models. The most common approach is to assume all damping is viscous and to approximate the decay rate of a material as 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 technique for modeling damping 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). The most general damping model to date that limits vibrations to normal modes is *generalized proportional damping* (GPD) (Adhikari, 2006). These
alternative damping models may be able to improve sound quality by providing a better fit to observed real-world damping.

We present a method that employs Generalized Proportional Damping to create alternative damping models for sound synthesis. We describe how more expressive damping models can be integrated into current methods for modal sound synthesis, and propose specific damping models within the larger space of GPD functions. These damping models are more expressive, and are able to model damping behavior that would be coarsely approximated by the Rayleigh damping model. We generalize a prior optimization method previously for damping parameter estimation from recorded impact sounds to estimate parameters for any arbitrary damping model. We also conduct a user study to evaluate the perceptual differences between multiple damping models. Figure 1.3 shows one possible scenario with objects creating sound based on their materials estimated from recorded impact sounds on real objects. Results, including audio samples and videos, can be found online: http://gamma.cs.unc.edu/gpdsynth/.

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

In order to create realistic virtual objects, recorded impact sounds can be used to estimate damping parameters, even if the sounds are recorded in noisy and uncontrolled settings. This research explores a novel
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.

probabilistic damping model for estimating the damping parameters of an object’s material from a set of impact sounds. The researchers are myself, 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 vibrational decay is due to the material damping and not any other source. However, there are multiple external factors that would produce effects similar to damping.

In the real world, an object struck for the purposes of recording either needs to be held by hand or left to rest on another surface. In either case, the interface between the object and its support introduces additional damping, as energy is transferred from the vibrating object to the more stationary 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 factors include complex modes of vibration, room acoustics, and error in the feature extraction step. Standard damping models capture only normal modes of vibration, not complex modes of vibration. These models are approximations, as real objects always have some amount of complex vibrations, causing error in damping estimates. If impact sound recording is performed in an enclosed room, the room acoustics—reflections off walls—extend the length of sounds. 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 and uncontrolled environments with significant impact 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 additional damping due to the recording environment. We explicitly model the external damping factors in a probabilistic damping model. For a given mode of vibration with an estimated frequency, this model provides a probability distribution for possible observable damping values. When multiple impact sounds are provided as input, the damping model’s parameters can be optimized to fit the real-world sounds, providing an estimate of the material’s damping parameters. The optimized parameters can be those of any real-valued damping model, and we propose one additional hybrid model between Rayleigh and power law damping. Our probabilistic model can express stochastic properties of the recording environment, reducing their effect on the estimated parameters.

Our method is more applicable to real-world recordings 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 points would fall along one line representing the damping from the material alone. However, the external factors
cause the points to vary, throwing off a more traditional least squares (LSQ) approach. More results are available online: http://gamma.cs.unc.edu/ProbDampModel/.

1.2.3 Integrated Multimodal Interaction Using Texture Representations

There have been 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 sight, touch, hearing, 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, particularly normal and relief maps, as a unified source of information for all interaction modes.

Haptic rendering, sound rendering, and rigid-body simulation with textured surfaces have been independently explored (Otaduy et al., 2004; Ren et al., 2010), but have not been integrated together cohesively. 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 a virtual environment, a user may see a rough, bumpy surface represented by its texture equivalent mapped to a flat surface. In the real world, objects behave differently when bouncing, sliding, or rolling...
on bumpy or rough surfaces than they do on flat surfaces. The visually complex detail would contrast with simple physical behavior due to the flat surface, causing sensory conflict and breaking the sense of immersion. In order to model such physical behavior, the shape geometry used in 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 effective method for simulation of physical behaviors for rigid objects textured with normal maps. We also propose seamlessly integrated multisensory interaction using normal and relief maps to improve sensory cohesion. By using a single representation of surface detail for each interaction modality, we ensure sensory cohesion for users. See Figure 1.6 for examples of interaction with textured surfaces. A virtual pen is controlled through a haptic device, allowing the 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 the method through perceptual user studies. In these studies, subjects are asked to identify the surface displayed to them, but in some trials certain interaction modalities are removed. When all modalities are present using our method, performance on the task is at its highest, demonstrating that the senses are not in conflict with one another. More results are available 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 maintain sensory cohesion by ensuring consistency between the object’s shape, surface appearance, and material. Object shape and surface appearance have historically been estimated through visual cues. Similarly, material parameters have 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. Knowledge of an object’s external shape provides constraints to the material parameter estimation problem. Therefore, estimation of either shape or material could benefit from using both visual and audio modalities of input.

We propose a method for estimating both an object’s material and shape geometry using combined audio-visual inputs. This work is performed by myself, Justin Wilson, Sam Lowe, and Ming C. Lin. 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 audial 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 model.

We propose a 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 produce a joint estimate of material and geometry.

Evaluation is performed on multiple datasets. The synthetic ModelNet10 and ModelNet40 datasets (Wu et al., 2015) produce classifications to object classes such as “table” or “dresser”. We create synthesized 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. We also present a utility for scene reconstruction in which impact sounds can be recorded to classify and segment objects. More results are available online: http://gamma.cs.unc.edu/ISNN/.
BIBLIOGRAPHY


